CURD: Context-aware Relevance and Urgency Determination

Ademola Adesokan
aaadfg@mst.edu
Missouri University of Science and Technology
Rolla, Missouri, USA

Sanjay Madria
madrias@mst.edu
Missouri University of Science and Technology
Rolla, Missouri, USA

ABSTRACT

During emergencies where time is of the essence, efficient management of disasters depends on swiftly recognizing relevant and urgent information from online platforms like X (Twitter), which is imperative for augmenting established response frameworks, such as the 911 emergency system. This paper introduces CURD, a Context-aware Relevance and Urgency Determination system designed to enhance the efficiency of disaster response. The system addresses two critical challenges: filtering out irrelevant data and assessing the urgency of relevant information. Our approach includes a multi-level annotation process for event type, relevancy, and an urgency annotation algorithm that significantly improves information extraction accuracy and efficiency. CURD_{dl}, our classifier, uses a deep learning pipeline architecture with a combination of transformer models, a convolution layer, and custom attention mechanisms to classify disaster-related tweets into multiclass-event type, binary-relevance-and-urgency categories, and rank urgent ones based on significance. Experimental results show that our best baseline classifiers for all three tasks achieved $\geq$ 88% F1 and accuracy, and $\geq$ 94%. AUC. Our models also outperformed models from related works in all metrics, validating the effectiveness of CURD in prioritizing response messages that will facilitate decision-making and resource allocation in disaster scenarios. CURD annotated dataset and code are available on GitHub[1].

CCS CONCEPTS

• **Social and professional topics** → **Systems development**; • **Information systems** → **Data cleaning**; **Social networking sites**; **Data analytics**; • **Applied computing** → **Annotation**; • **Computing methodologies** → *Lexical semantics*; *Information extraction*; *Machine learning*;

KEYWORDS

Data Annotation, Social Media, Emergency Management, Relevancy, Urgency

ACM Reference Format:
Ademola Adesokan and Sanjay Madria. 2024. CURD: Context-aware Relevance and Urgency Determination. In *36th International Conference on Scientific and Statistical Database Management (SSDBM 2024), July 10–12, 2024, Rennes, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3676288.3676299

[1]https://github.com/abdul0366/CURD

SSDBM 2024, July 10–12, 2024, Rennes, France

ACM ISBN 979-8-4007-1020-9/24/07
https://doi.org/10.1145/3676288.3676299

1 INTRODUCTION

Climate change has exacerbated the difficulties caused by natural disasters, including strain on infrastructure and resources, complex emergencies, agricultural impacts, migration and displacement, and increased health risks. As a result, advanced technological interventions are necessary for effective response and recovery services. Research shows that hurricanes are becoming more intense, leading to unparalleled flooding, destruction, and loss of lives and properties, as evidenced by Hurricane Ike [38] and, more recently, Hurricane Ida [34]. Additionally, climate change has increased the severity and frequency of wildfires, such as those in Australia [32]and California [39], presenting significant challenges to traditional firefighting methods. This calls for the development of innovative strategies and technologies for early detection and swift response.

In disaster management, response services require innovative and adaptive approaches to address the changing circumstances of different events. The increasing significance of social media, especially X (previously called Twitter), has been recognized in disaster management. Researchers like [10] and [23] leverage X data for predictive analytics and situational awareness using Machine Learning (ML) techniques. Furthermore, [22]'s work on sentiment analysis during the COVID-19 pandemic demonstrates the value of social media insights for crisis communication and service improvement. Recent research in disaster management has made significant strides in social media data analysis for crisis response. Studies like [14], [24], and [25] highlight the utility of ensemble learning, a multimodal strategy, and the combination of NLP and ML in improving the efficiency of disaster response, emphasizing the importance of advanced computational techniques in this field.

In this study, we tackle two key challenges in disaster management: discerning vital information from irrelevant data and evaluating the urgency of information. (1) The first challenge involves a detailed analysis and filtering process that allows disaster management authorities to focus on crucial data, saving time and resources. This enables expedited decision-making processes, which is crucial for allocating resources, identifying the immediate needs of affected individuals, and assigning appropriate first responders. (2) The second challenge involves prioritizing information based on urgency. Decision-makers in disaster response scenarios must determine which messages require immediate attention and action and which can be deferred.

Addressing the above challenges is necessary to improve disaster response. Hence, we propose a framework for better information management by defining 'relevance' and 'urgency' in the context of disaster response.

- **Relevancy** refers to how well the information aligns with the current needs and objectives of the response effort. Relevant information helps disaster management authorities to concentrate their resources and efforts more effectively by prioritizing information that directly contributes to ongoing rescue, relief, and recovery operations.
- **Urgency** refers to the immediate need for action or attention that specific information demands during disaster response. For instance, an urgent tweet or piece of information, such as one reporting people trapped under a collapsed building following an earthquake, requires a swift response to prevent further harm, save lives, or address critical disaster situations.

Considering the outlined challenges and definitions, we present the following arguments:

- During a disaster, there is a likelihood that extracted disaster data may not be directly relevant. Relevance is determined by considering the context, meaning, and potential impact of the message on the current situation and its applicability.
- Moreover, it is also important to recognize that while certain data, such as infrastructure damage reports (damage to bridges or roads), may be relevant to disaster response efforts, it may not always be of immediate concern when compared to life-threatening situations that require immediate attention, such as rescue operations.

The conventional way of annotation in disaster management is to evaluate data against pre-defined criteria and classify it into its appropriate defined group. Studies by [17], [29], [26], and [33] have refined various methods over time, contributing unique perspectives and methodologies to the field. In this work, we introduced **CURD** (**C**ontext-aware **R**elevance and **U**rgency **D**etermination), a novel approach that uses multi-level annotation to enhance the accuracy and efficiency of identifying critical information from the large volume of data generated during disaster scenarios, particularly from X. We propose a four-stage process as the primary aim of CURD: (1) Initially, the stage of categorizing data into specific, low-level disaster-related labels is crucial for organizing incoming information into manageable and meaningful categories. This filtering and structuring of data lays the groundwork for more detailed analysis in subsequent stages. (2) In the second stage of classification, the information labeled in the initial stage is categorized as either 'relevant' or 'not relevant' to the disaster response phase. This step assumes that not all disaster-related information is important for immediate response efforts. The criteria for relevance include the need for the information, the potential impact on response efforts, and the applicability of the information to the current phase of response. (3) The third stage of the CURD method is the urgency determination of relevant information, which classifies all relevant information as either urgent or not urgent for efficient resource allocation. (4) Our final stage involves ranking tweets based on their importance in critical disaster response phrases. We compute the urgency scores of the tweets and sort them based on their level of significance, enabling us to swiftly identify, prioritize, and act upon critical information within large datasets.

Our work has yielded the following contributions:

(1) We utilized the Automatic Content Extraction (ACE) standard [36] to annotate a total of 47,621 natural disaster tweets, resulting in the identification of 29 unique event types. Our transformer-based event type classifier achieved best result with CURD_{dl} – RoBERTa model, achieving 89% accuracy and 90% Fleiss Kappa inter-annotator agreement.
(2) Post event type annotation, we categorized the tweets based on their relevance (relevant/not relevant). Furthermore, based on the relevant class, we also use a combination of BERT-based embeddings, K-Means clustering, and critical event-type labels to determine their urgency.
(3) We developed a customized multiclass and binary classifier model, CURD_{dl}, which utilized pre-trained transformer models (BERT, BERTweet, RoBERTa, DistilBERT, and XLNet) [16], as well as convolution layer, custom attention layer, and fully connected layers. Our best model, CURD_{dl}-RoBERTa, achieved 90% accuracy and an AUC-ROC of 97%.
(4) We introduced a novel method for computing urgency scores based on individual or the combination of disaster response terms, which allowed us to rank the classified urgent tweets and provide valuable assistance to disaster response efforts.

2 RELATED WORKS

Disaster management evolves to utilize social media data for efficient response strategies, marked by innovative research addressing technical and contextual obstacles. Transfer learning for crisis urgency detection is discussed in [7], highlighting the adaptability of existing models to disasters' unpredictable nature. This approach enables quick decision-making. [5] expands on this by addressing Arabic language data processing and integrating NLP methods with cultural nuances, leading to more inclusive disaster response tools. Recent works stress the importance of urgency detection and relevance classification in disaster response. [41] presents the cross-topic relevance embedding aggregation that enhances relevance classification accuracy, especially in data-limited and topic-specific disaster contexts. [21] utilizes a neural network model that merges text and image data from social media to enhance disaster response. The multimodal approach effectively identifies informative content during crises by combining LSTM networks for text and VGG-16 networks for image processing. The connection among researchers in disaster response is crucial for developing effective strategies and advancing knowledge. [11] focuses on using ML to identify urgent tweet requests during hurricanes, improving disaster response efficiency, and setting a precedent for applying ML to other emergencies. Studies on online news comments' classification address misinformation [37]. A low-supervision urgency detection system bridges these studies [20], showcasing AI's potential to enhance resource-efficiency and accessibility in urgency detection. Disaster response research necessitates diverse methodologies and approaches, as demonstrated by multiple studies, including [9], [29], [26], [35], and [13]. These studies cover various aspects of crisis management, such as political context relevance classification [9], the role of social networking sites in disaster relief [29], [26], and situational awareness [35] and linguistic aspects of urgency detection in crisis communication [13].

The range of approaches discussed highlights disaster response research's complex and multifaceted nature, ensuring the development of comprehensive and adaptable strategies for different crises. Advanced algorithms like BERT and XLNet have shown superior performance in handling complex, unstructured data during crises. These algorithms are particularly effective in categorizing and understanding data in disaster scenarios, as noted in [31].

All of these works employ direct annotation of tweets to relevance and urgency. In contrast, our method utilizes multilevel annotation, which involves annotating tweets with event-type labels to determine relevance and urgency indirectly via these labels, rather than directly annotating the tweets themselves. This event-type annotation is used to capture all information initially, preventing information loss by utilizing predefined labels from previous works.

3 OUR APPROACH

Our methodology, depicted in Figure 1, comprises four sequential phases: event classification, relevance classification, urgency classification, and ranking of urgent tweets. This systematic approach guarantees comprehensive analysis and categorization of tweets, enabling the effective identification and prioritization of crucial disaster response information.

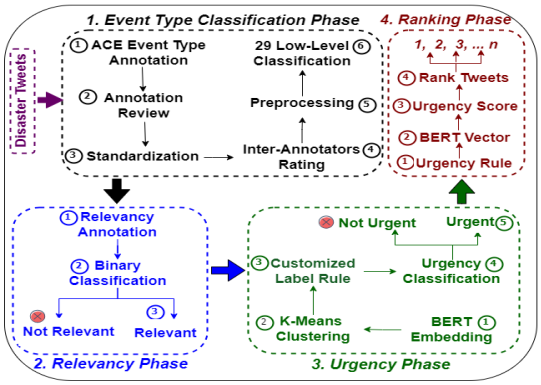


Figure 1: CURD Approach

3.1 Dataset Description

Our research examines tweets about natural disasters, specifically earthquake, and hurricanes, using datasets from the CrisisNLP [18] and UNT Library [30]. The dataset consists of 47,621 tweets, with 2,902 labeled as gold data and 44,719 as silver data. Notably, the gold dataset includes 1,518 tweets about the Nepal earthquake. In contrast, Hurricane Harvey, Hurricane Sandy, and Hurricane Odile have 1,000, 200, and 183 tweets, respectively. We split the dataset into training (80%) and testing (20%) sets for our experiments.

3.2 Event Type Phase

Our process for annotating event types involves label annotation, reviewing annotations, standardizing class labels, preprocessing, and classification. This ensures thorough and standardized analysis.

3.2.1 ***Event Type Annotation.*** Our annotation process uses the ACE methodology developed by the ACE 2005 initiative [36] to automatically extract data elements like entities and relationships from text. The ACE framework consists of 8 event types, 33 sub-event types, triggers, and arguments, but we modified it for our task by only using its standard elements. We did not use the predetermined event types. To accommodate disaster-related data, our annotation process, conducted by three graduate students over six months, used free-form labeling and allowed for varied interpretations and more nuanced labels. By initially annotating all tweets into 92 event types, we established a foundation for further standardization in our study. We avoided the overly specific traditional methods of disaster data annotation, opting instead for free-form labeling via ACE to prevent information loss from excessive specificity.

3.2.2 ***Annotation Review.*** Our annotation review process involves a collaborative approach where annotators actively review and verify each other's work. This method is commonly used in data annotation and labeling tasks, where the annotators not only focus on their individual tasks but also examine their peers' annotations to guarantee accuracy and consistency. After reviewing another's work, an annotator can either agree or express disagreement. If there is disagreement, the annotator must provide specific reasons for their viewpoint. This step is essential because it facilitates dialogue and a collaborative effort to reach a consensus and standardize annotations. Our primary goal is to ensure the production of high-quality, consistent, and accurately labeled data.

3.2.3 ***Label Standardization.*** To standardize the categorization of data, we have developed a method that reduces variability and overlapping in classifying events. During the free-form phase, labels are in their rawest and most varied form. For example, 92 distinct labels were obtained for "Event type". To simplify the categorization process, we have grouped these labels into more specific and encapsulating categories. For instance, "personal matters" is an umbrella label that encompasses specific categories such as "personal account", "personal concern", and "personal view". This initial normalization reduces the number of labels under "Event type" from 92 to 52. Further refinement of these categories results in a more concise set of labels [3]. Following this phase, the "Event type" labels are reduced to 29 ***(admiration, appreciation, business, casualty, climate & environmental issues, communication, damage, die, disaster preparedness, education, empathy, health, humanitarian assistance, immigration, information dissemination, inquiry, life, memories, news, others, personal matters, politics, resources, safety, spiritual, sport, transportation, travel, warning)***. This approach ensures that the labels are relatively specific while avoiding redundancy and operational challenges.

3.2.4 ***Inter-Annotator Rating.*** After standardizing the annotations, we assessed the agreement between different annotators using the inter-annotator agreement process. Each annotator scored their agreement on a scale of 1 for agreement and 0 for disagreement. We calculated the consensus level using the Fleiss Kappa statistical measure [15], which yielded a high consistency rate of 0.90, 0.92, and 0.95 for event types, relevancy, and urgency, respectively. This indicates a considerable agreement among the annotators, demonstrating the reliability and accuracy of the annotations.

3.3 Dataset Preprocessing

We perform preprocessing and cleaning of the disaster tweets dataset to ensure the accuracy and reliability of deep learning applications. This step involves removing duplicates and normalizing text to treat variations uniformly ('earthquake' and 'EARTHQUAKE' become earthquake). We also eliminate extraneous information like links and special characters and tokenize the tweet text into individual words for detailed word frequency, distribution, and co-occurrence analysis [1]. Additionally, we remove stop words to reduce noise and emphasize significant terms that reflect the nature of disaster contexts [2].

3.4 Relevancy Phase

Determining relevance during disaster response is a critical step in identifying actionable insights from vast data. Our relevance annotation methodology, depicted in Figure 1, is based on a binary framework that labels data as either relevant or not relevant. Previous research has employed direct annotation of tweets with specific keywords or criteria to determine relevance. For example, [17] categorized tweets as informative or non-informative, while [27] and [8] used the CrisisLex26 dataset by [28] to segregate tweets based on their usefulness in response and recovery scenarios. However, these methods may be too narrow and could result in insufficient information for response efforts.

In contrast to existing literature, our methodology is novel and straightforward for determining relevance in disaster tweets. We utilize low-level annotated class labels that categorize the event type, as described in Section 3.2. By converting the 29 multiclass labels into binary annotations, we assess whether tweets and their respective labels are relevant to the disaster response. We define relevance as the immediate usefulness of the information in aiding or understanding the disaster response efforts. We have identified and classified the most relevant class labels based on our analysis of the tweets' semantics and their relevance to disasters grounded on the 29 standardized annotated class labels for event type annotation in Section 3.2.3, along with the provided explanations. As a result, we have 23 class labels for the relevant class, which is a total of 26,091 tweets for relevancy annotation out of the 47,621 annotated event types. The remaining tweets belong to six class labels that form the "not relevant" category. We have briefly explained why the event type labels were annotated into 'relevant' class in Table 2.

Our approach to disaster tweet relevance judgment involves a rigorous process that distinguishes between relevant and irrelevant tweets based on comprehensive guidelines developed during event-type annotation. This approach is crucial as it allows us to identify actionable insights from disaster tweets, vital for response efforts. Our guidelines are designed to capture tweets that might unintentionally be deemed irrelevant but hold significance in response scenarios. The guiding criteria for relevant class annotation include event-type labels that contain one or more of the following:

(1) **Disaster Keywords:** we classify tweets as relevant if they contain keywords like 'earthquake', 'flood', or 'injured.'
(2) **Response Actions:** we seek disaster response actions such as rescues, appeals, and aid.
(3) **Entities Involved:** our focus entails organizations, locations, and groups actively involved in disaster response, such as NGOs, government entities, and impacted communities.
(4) **Critical Assumptions:** we evaluate various assumptions to measure the immediate effects of the event on disaster response, the informational value of the content, and the support provided to affected individuals. This involves assessing whether the event directly contributes to response, recovery, or mitigation efforts, offers valuable information for situational awareness, safety, or aid distribution, and provides support, whether emotional, physical, or resource-based, to those affected by the disaster.

Our contribution to the relevancy annotation phase in Figure 1 lies in the innovative application of low-level event types and criteria and in our diligent commitment to avoiding arbitrary annotation. By adopting a multi-faceted approach, we enhance disaster response efforts by offering an accurate, comprehensive, and sensitive method for relevance classification. This attention to detail is crucial in improving the effectiveness and responsiveness of disaster management strategies.

3.5 Urgency Phase

The urgency part of Phase 3 in Figure 1 comprises two steps: (1) The Urgency Annotation Phase, which includes clustering and labeling tweets with customized rule-based tags, and (2) The Binary Classification of Urgency. Our strategy for urgency, as outlined in Section 1, highlights the pressing need for immediate action in emergency situations. This necessitates timely responses to mitigate harm, safeguard lives, and manage critical circumstances. Our methodology builds upon [11]'s work, which filtered out animal-related tweets during Hurricane Harvey. Furthermore, [31] employed MTurk workers to gauge urgency levels by examining tweet elements like exclamation marks, key verbs, and calls to action words such as 'rescue'. Our approach aligns with [26], which used annotations including calls for help, location information, and various action prompts. Our unique approach, however, encompasses all entities, including animals and situations, if they provide crucial information for disaster response services. During this phase, We concentrate on relevant tweets from Section 3.4, disregarding 'not relevant' ones that lack urgency or significance in disaster response. This approach emphasizes tweets that provide significant value in disaster response and urgent action contexts.

3.5.1 ***Urgency Annotation.*** We employ a multi-faceted approach to annotating urgency that involves capturing nuanced context using BERT embeddings and identifying inherent patterns in the data through unsupervised methods like K-Means clustering [19]. This is further refined with customized event-type labels to achieve a comprehensive and accurate urgency annotation.

Algorithm 1 is designed to process, analyze, and annotate tweet data in emergency situations. The process begins with the use of BERT embedding, a transformer model that captures the context of words in tweets, allowing for efficient processing of the text data. The high-dimensional BERT embeddings are then reduced using principal component analysis (PCA) for simplified data visualization and processing. In the Clustering phase, we group the data into clusters based on their PCA-reduced features. This step helps

Algorithm 1 Urgency Annotation

Input: Relevant Labeled Tweets
Output: Clustered Tweets with Urgency Annotated Labels
1: Load dataset: Relevant Labeled Tweets
2: Tokenization and Embedding with BERT:
  a. Initialize BERT tokenizer and model
for each tweet in all relevant classified tweets **do**
    b. Apply BERT embeddings
    c. Store embeddings
end for
3: Stacking Multiple Numpy Arrays into a Single Array:
  a. embeddings ← Stack embeddings
4: Dimensionality Reduction:
  a. Initialize PCA with 2 components
  b. X_reduced ← Fit and transform embeddings with PCA
5: Clustering:
  a. Initialize K-Means with 2 clusters
  b. clusters ← Fit and predict X_reduced with K-Means
  c. Store clusters
6: t-SNE transformation:
  a. Initialize t-SNE with 2 components
  b. X_tsne ← Fit and transform embeddings with t-SNE
7: Calculate the Silhouette Score:
  a. Average Silhouette ← Calculate silhouette score for X_reduced and clusters
  b. Output Average Silhouette
8: Annotate Urgency:
  a. Define customized event type labels
for each label in all event type labels **do**
    b. Apply customized event type labels to annotate urgency
    c. Store urgent/not urgent result
end for

in identifying patterns within the tweet corpus. In the t-SNE transformation stage, we utilize t-SNE for dimensionality reduction and visualization. This allows us to understand the distribution and separation of clusters. The silhouette score, computed for the K-Means clustering, helps us evaluate the effectiveness of the clustering and measure how well texts are positioned within their clusters. We defined custom event-type labels for the urgent label (casualty, damage, die, health, humanitarian assistance, communication, resources, warning, safety) in the Annotate Urgency step, which enhanced the coherence of our clustered tweets into 'Urgent' and 'Not Urgent' categories. This annotation is crucial for prioritizing or filtering tweets based on urgency, particularly in disaster response or urgent public communications.

3.6 Event Type, Relevancy and Urgency Classifier

Our classification system employs a suite of five transformer models, including BERT, BERTweet, DistilBERT, XLNet, and RoBERTa, to capture the semantics and context of our dataset through binary classification for relevancy and urgency, and multiclass classification for event type. By integrating these models, we have created a highly accurate classification system that utilizes the unique strengths of each transformer model to ensure a nuanced and robust understanding of the textual data. This approach is crucial for precise classification of event type, relevancy, and urgency, allowing our classifier to perform optimally.

The CURD$_{dl}$ classifier model, shown in Figure 2, is a deep learning pipeline that processes disaster tweets and generates final-ready classification output for our tasks. The step-by-step algorithm for our classifier can be found in Section 3.6.1. Hence, we briefly describe the role of each block in Figure 2:

(1) **Tokenize Tweet:** This stage involves tokenizing tweets into sequences of tokens using a tokenizer that converts words into numerical IDs, with a limit of 512 tokens for processing by Transformer models like BERT and DistilBERT. Note that none of our tweets exceed half of the token size.
(2) **Pretrained Transformers:** CURD$_{dl}$ model uses embeddings from five different pre-trained transformer models with 12 layers and 768 embedding sizes, which can be fine-tuned for various tasks like classification.
(3) **Conv1d:** Our classifier incorporates a Conv1d layer with 768 input channels, corresponding to the model's output, and 64 output channels, featuring 3 kernel sizes and 1 padding.
(4) **Custom Attention Layer:** Our model computes attention scores using a learnable weight matrix and bias to focus on relevant parts of the input sequence, capturing context and disaster response terms with a custom attention layer having 768 input dimensions.
(5) **Fully Connected Layer 1:** The first dense layer following the attention layer is this block, which combines the 768-dimensional output from transformer models like BERT with the 64-dimensional output from the convolutional layer to create a 256-dimensional feature vector.
(6) **Fully Connected Layer 2:** This is the second dense layer of the model, which takes the 256-dimensional input from the first fully connected layer and maps it to the final output size of 29, corresponding to the number of unique classes for the event type classification task. The layer also maps the 256 features to 1 output for the binary class classification.
(7) **Activation Function:** We apply activation functions to CURD$_{dl}$, for binary classification, a sigmoid function outputs probabilities between 0 and 1. For multiclass classification, a softmax function outputs a probability distribution over the class labels.
(8) **Output Labels:** The output of the model is the final result for multiclass classification, and for binary settings, it is a single probability after applying the sigmoid function. The logits are passed to the loss function during training.

Our CURD$_{dl}$ model, depicted in Figure 2 and detailed in algorithm 2, builds on transformer models like BERT by [12]. Our model differs from standard transformer architectures by introducing several enhancements and modifications that make it particularly adept at analyzing disaster tweets. The model's unique combination of transformer embedding, custom attention mechanism, convolution layer integration, and fully connected layers sets it apart. The integration of a custom attention layer enables the model to focus more intently on vital parts of the input sequence, capturing nuances in disaster tweets that standard transformer models might overlook. The convolution layer extracts local features from the transformer output, resulting in a more comprehensive analysis of text data. Finally, the model utilizes fully connected layers for the final classification task, allowing it to learn more complex relationships and patterns from the combined features, offering a more novel approach than standard transformer models.

3.6.1 ***CURD$_{dl}$ Classification Algorithm*** . Algorithm 2 presents a comprehensive method for constructing an advanced model that classifies disaster tweets based on their event type, relevance, and

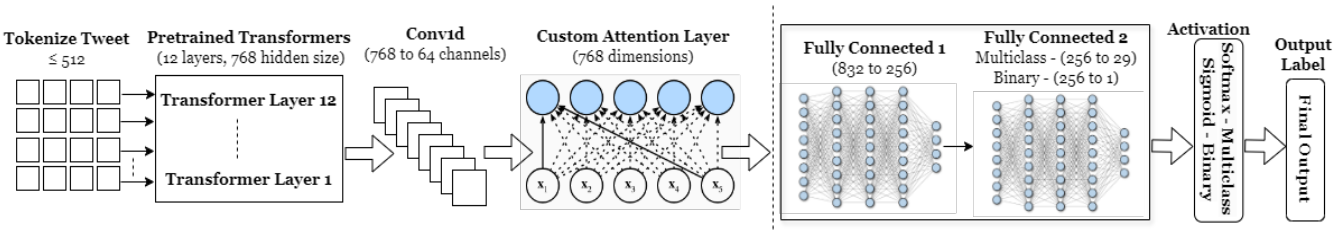


Figure 2: CURD$_{dl}$ Event Type, Relevancy and Urgency Classifier ($_{\text{dl}}$ - denote *deep learning*)

Algorithm 2 Event Type, Relevancy and Urgency Classifier

Input: Preprocessed train and test data
Output: Trained and evaluated tweets classification model
1: Define Custom Attention Layer:
a. Create CustomAttention class inheriting from nn.Module
b. Initialize weights and bias
c. Define forward pass for attention computation
2: Define the Transformer Model:
a. Create MyModel class inheriting from nn.Module
b. Initialize transformer, convolutional layer, custom attention, and fully connected layers
c. Define forward pass combining transformer, a convolution layer, and attention outputs
3: Data Preparation:
a. Create TweetDataset class inheriting from Dataset
b. Define initialization, length, and item retrieval methods
c. Tokenize tweets and prepare model inputs
4: Load Preprocess Data:
a. Load preprocessed training and testing datasets
b. Encode labels as binary/multiclasss
5: Create Datasets and DataLoader:
a. Initialize transformer tokenizer
b. Create TweetDataset instances for training and testing
c. Create DataLoader instances for datasets
6: Define Optimizer and Loss Function:
a. Define Adam optimizer
b. Define BCELoss/CrossEntropyLoss
7: Training Loop:
for each epoch **do**
a. Set model to training mode
b. Initialize loss and accuracy counters
for each batch in train_loader **do**
i. Forward pass
ii. Compute loss
iii. Backward pass and optimization
iv. Update counters
end for
end for
8: Model Evaluation:
a. Set model to evaluation mode
b. Initialize prediction and label lists
for each batch in test_loader **do**
i. Forward pass
ii. Store outputs and labels
end for
c. Convert predictions to binary/multiclass
d. Output classification report and AUC-ROC score

urgency. The model leverages the strength of transformer models, a convolution layer, and an attention mechanism and has several components, each serving a specific purpose. The Custom Attention Layer in line 1 is designed to weigh specific parts of the input sequence, such as critical words for disaster response, allowing the model to focus more heavily on these inputs. This layer computes attention scores through a learnable weight matrix and bias, which are vital in enabling dynamic focus on relevant parts of the tweet. This enhances the model's ability to understand context and meaning. Line 2 incorporates a pre-trained transformer model, a convolution layer, a Custom Attention Layer, and Fully Connected Layers. The transformer provides global context, the convolution layer extracts local features from the transformer output, and the Custom Attention Layer focuses on essential features within the tweet. The Fully Connected Layers are used for the final classification. This combination enables the model to capture both global and local textual features, enhanced by the attention mechanism's focus on relevant parts of the tweet. Data preparation involves creating a dataset compatible with PyTorch's DataLoader for efficient batch processing in line 3. The TweetDataset class tokenizes tweets and prepares them for input into the model, ensuring that the input data is in a format that the model can effectively process. The DataLoader in PyTorch in line 4 organizes data into batches for training and testing the model. This improves computational efficiency by efficiently managing data loading for each training and testing batch. In line 6, the initialization of the neural network model is crucial for utilizing model training and inference. The optimizer (Adam) and the loss function (Binary Cross-Entropy) guide how the model learns from the training data for binary class; and Cross-Entropy for multiclass. However, Our multiclass model combines *nn.LogSoftmax()* and *nn.NLLLoss()* in one class, effectively applying a softmax activation to the output layer and then computing the negative log-likelihood loss. The optimizer updates model weights, and the loss function measures the model's performance on the binary/multiclass classification task. The training loop in line 7 is designed to train the model for 10 epochs; we stopped at 10 to avoid overfitting, thus updating weights based on the loss function. Iterative training allows the model to learn from the data, thereby improving its accuracy and performance. As shown in line 8, model evaluation assesses the model's performance on the test dataset using evaluation metrics such as precision, recall, F1, accuracy, and AUC-ROC score. These metrics offer a comprehensive understanding of the model's effectiveness, accuracy, and areas for improvement.

3.7 Urgency Ranking Phase

Algorithm 3 evaluates tweets according to their urgency, using BERT embeddings and cosine similarity. It begins by loading urgent tweets in line 1 and then, in line 3, it identifies and assigns weights to grouped and individual keywords to determine their significance in determining tweet urgency. The process of defining keywords during preprocessing emphasizes the importance of both individual and combined critical response words. High-urgency keywords, such as 'stranded,' 'trapped,' and 'injured,' are given more weight

Algorithm 3 Tweets Ranking by Urgency Score

Input: Dataset of Urgent Tweets
Output: Ranked Tweets with Urgency Scores
1: Load Dataset: Urgent Data
2: Load Pre-trained BERT Model:
  a. Initialize BERT tokenizer and model
3: Define Keywords and Weights:
  a. Define grouped and individual keywords with corresponding weights
4: Define BERT Vector Computation:
  a. Compute BERT vector for a tweet
5: Define Group Vector Computation:
  a. Compute average vector for a group of keywords
6: Define Urgency Score Computation:
  a. Compute urgency score for a tweet
7: Compute Urgency Scores:
  a. Apply urgency score computation to each tweet in the dataset
8: Rank Tweets Based on Urgency Scores:
  a. Sort tweets by urgency scores in descending order

than medium-urgency keywords like 'hit,' 'quake,' and 'killed,' as well as low-urgency keywords like 'donate,' 'donation,' and 'safety.' The algorithm prioritizes tweets containing combinations of words indicating higher urgency, ensuring that tweets with compounded critical terms are prioritized over those with less critical urgent content. The algorithm evaluates the standalone urgency of individual keywords, such as 'stranded,' to ensure effective prioritization of urgent tweets. This context-aware approach is crucial in disaster scenarios where timely responses can make a significant difference.

The model computes BERT embeddings for each tweet to capture the semantic meaning in a high-dimensional space. It also computes an average vector for groups of keywords, which is a representative vector that aids in comparing with individual tweet vectors. The algorithm's key step is calculating urgency scores in lines 6-7. Urgency scores are determined by comparing tweet vectors with keyword vectors using cosine similarity, with weights applied to highlight the significance of each keyword category. Tweets' urgency scores are determined by a quantitative measure that assesses their alignment with crucial keywords. This step is vital for categorizing and evaluating every tweet based on its content. The eighth line involves ranking the tweets based on their urgency score, which is computed using the algorithm. This ranking is crucial in disaster response, as it allows for prioritization of tweets based on their urgency. Algorithm 3 provides a comprehensive and systematic approach to identifying critical information in large datasets, especially in emergency situations.

4 EXPERIMENTAL RESULTS

We evaluated the effectiveness of our methodologies introduced in Section 3 through various experiments, ranging from event type to urgency classification. We evaluated five transformer models for event types, relevance, and urgency and assessed their performance using AUC-ROC, accuracy, precision, recall, and F1 score (taking into account class imbalance). Our evaluations considered factors such as model complexity, hyperparameter tuning, and class label distribution, showcasing our models' effectiveness and versatility for disaster applications.

To ensure robust training and testing, we conducted our experiments on a high-end gaming desktop with a Ryzen 9 5950X processor, NVIDIA GeForce RTX 3090 GPU, 128 GB of RAM, and CUDA 11.8. This powerful machine allows reliable and consistent model evaluation, as evidenced by the effectiveness of our approaches.

4.1 CURD Annotated Tweet Samples for Our 2-Arguments

Table 1 provides a comprehensive analysis of disaster-related tweets and effectively meets the challenges and objectives outlined in Section 1 of the study. This is achieved through several key features.

The relevancy part of Table 1 addresses the challenge of filtering out irrelevant data, categorizing tweets as either "Relevant" or "Not Relevant." This helps in separating critical information from less pertinent data. For instance, the first tweet from the top of Table 1 is labeled as "Not Relevant" as it holds lower significance in the context of immediate disaster response. This supports the argument that even if a tweet is disaster-related, it might not be relevant.

Table 1's urgency classification directly responds to the challenge of prioritizing information based on urgency. Tweets are classified as "Urgent" or "Not Urgent," which assists decision-makers in identifying which messages necessitate immediate action. For example, the second tweet from the top is deemed "Relevant" but not "Urgent," underscoring its significance in disaster management. This aligns with the argument that relevancy might not depict urgency.

4.2 Relevancy Annotation - Event Type to Binary Relevance

In the context of disaster response, Table 2 categorizes various event types to binary relevancy, along with brief details to emphasize their relevance. A collective understanding of these event types is essential as it offers a complete insight into a disaster's impact. Moreover, they play a pivotal role in facilitating response efforts and extending support to affected communities. Thus, they are instrumental in developing an effective and informed disaster management strategy.

4.3 Baseline Result - Urgency Annotation

Figure 3 shows the use of unsupervised machine learning techniques to uncover patterns in tweets and classify them based on urgency.

In Figure 3a, principal component analysis was used to reduce the two-dimensional space of tweet embeddings produced by a BERT model, resulting in two well-defined clusters. The purple cluster exhibited a high degree of homogeneity, while the yellow cluster demonstrated greater diversity. The K-Means algorithm successfully separated the tweets into distinct groups, as seen by the clear boundary between the clusters. PCA Feature 1 and 2 captured the primary variations in the tweet embeddings, providing a meaningful basis for the clusters. The silhouette score of 0.48 confirmed the presence of a moderately strong cluster structure, indicating the effectiveness of our clustering approach.

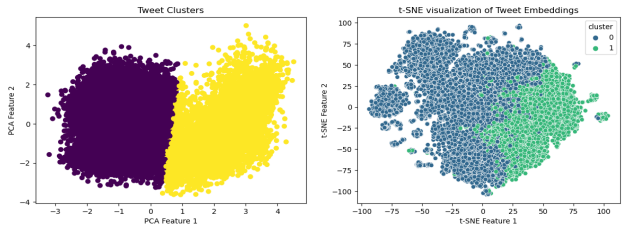
In Figure 3b, t-SNE is used to visualize tweet embeddings. This non-linear technique reduces the dimensionality of high-dimensional data while adeptly preserving its local structure. Unlike PCA, t-SNE uncovers clustering patterns that may not be readily visible. The t-SNE visualization of the dataset shows two clusters with a smooth transition of points between them, indicating the presence of distinct groups and a continuum of tweet embeddings connecting

Table 1: Assumptions - This tweet sample includes information on Event Type, Relevance, Urgency, and Urgency Score. The first tweet falls under the event type category 'admiration,' which is related to disasters, but is classified as 'Not Relevant.' Similarly, the second tweet, which is disaster-related, is categorized as 'Relevant' but falls under the 'Not Urgent' class.

Tweet	Event Type	Relevancy	Urgency
1. Reading about all the help extended to Nepal by different countries, seems like this world is still a good place to live #earthquake	admiration	Not Relevant	Not Urgent
2. TV News Covers Harvey: Weather Channel, ABC, Others Track Chaotic Hurricane	climate & env. issues	Relevant	Not Urgent

Event type	Relevancy Explanation
Casualty	Relevant regarding the human cost of the disaster.
Climate & Envi. Issues	Relevant for environmental-related discussion.
Communication	Relevance in coordinating disaster response services.
Damage	Relevance for assessing the impact of the disaster.
Die	Relevant as it involves the loss of life via the disaster.
Disaster Preparedness	Relevance for mitigating current/future disaster impact.
Education	Relevant for increasing disaster awareness and safety.
Empathy	Relevance as it provides moral support to those affected.
Health	Relevance for direct impact on people's well-being.
Humanitarian Asst.	Relevance as it pertains to aid and relief efforts.
Info. Dissemination	Relevance for the spread of critical information.
Inquiry	Relevance for active information-seeking behavior.
Memories	Relevance for insights from past disasters.
News	Relevant for factual reporting and updates.
Personal Matters	Relevant for personal disaster impacts/insights.
Politics	Relevance for discussing political intervention and help.
Resources	Relevant for discussing necessary supplies and aids.
Safety	Relevance for protecting life and well-being.
Spiritual	Relevant as a form of moral support.
Sport	Relevant as it discusses the impact on local events.
Transportation	Relevant for information on logistics and movement.
Travel	Relevant for those affected by/responding to the disaster.
Warning	Relevant for immediate safety/preventative measures.

Table 2: Event Types that constitute our Relevant Class



(a) PCA-based Tweet Clusters **(b)** t-SNE Tweet Embeddings

Figure 3: Urgency Tweets Annotation

these groups. This nuanced depiction contrasts with PCA's more segregated cluster portrayal. Notably, the t-SNE plot reveals some degree of cluster overlap or intermingling at the boundaries, a hallmark of t-SNE's ability to capture subtle relationships within the data. In the t-SNE feature space, the axes do not correspond to explicit features but function as a map of similarities, with the proximity of points reflecting their co-occurrence in the original complex dataset. This quality makes t-SNE a powerful tool for showing the complex topology of data embeddings, providing a deeper understanding of the intrinsic patterns in our dataset.

Subsequently, we also employ a quantitative method to pinpoint the closest neighbors within our urgent tweets by leveraging an index that signifies the location of a tweet in an embedding matrix. Each row of the matrix corresponds to the BERT-generated embeddings of a specific tweet. As shown in Figure 4, our approach entails retrieving the top five most similar tweets to a randomly selected one. To accomplish this, we calculate the cosine similarity between the chosen tweet's embedding and all other tweets' embeddings in the dataset. Cosine similarity measures the cosine of the angle between two vectors, serving as an indicator of similarity. We then sort these similarity scores to identify the indices corresponding to the highest values, which represent the closest neighbors. Importantly, we exclude the randomly selected tweet from this list, as it would naturally be the most similar to its own embedding. Furthermore, we describe a process for randomly selecting a tweet from the dataset and employing the nearest neighbors function to identify and display its nearest neighbors, as demonstrated in Figure 4. These tweets share similar contexts, as indicated by the embeddings' similarity, as determined by the model.

4.3.1 ***Top Similar Tweets To Selected Tweet.*** Upon analyzing the nearest neighbors for urgency annotation, our cluster of tweets shares thematic relevance with the response and recovery efforts following the disaster tweets. Figure 4 depicts a randomly selected tweet, which highlights the activation of systems by social media companies to verify the safety of individuals affected by the disaster. The proximity of the nearest neighbors in the feature space suggests a similarity in the content features of these tweets, which could be attributed to the use of common keywords such as "#NepalEarthquake."

Selected Tweet: @TeklaPerry: Social media giants activate systems designed to let people know if loved ones are safe #NepalEarthquake http://t.co/MBBmVJMkuj

Nearest Neighbors:

1. @ievaluate My company launched the #NepalRecoveryFund to get money to vetted community-based orgs: http://t.co/1cG4pirs4L #NepalEarthquake
2. I hope that someday soon @ReactionHousing will be helping people stricken by disasters. #NepalEarthquake http://t.co/QjnVlzXajE
3. An amazing blog by @RichendaG + @kylevermeulen: The challenges facing #NFP orgs responding to the #NepalEarthquake http://t.co/DGio9nSppp
4. Pls. RT: Great primer from @USAID on how to help and stay informed on #NepalQuake: http://t.co/BBzkcyO5oU
5. RT @OxfamAmerica: Help rush life-saving aid to #NepalEarthquake survivors http://t.co/JPAG0mjZxX

Figure 4: Top-5 Nearest Tweets to a Randomly Selected Tweet for Urgency Annotation

The top-5 tweets selected by our model for the nearest neighbors in Figure 4 include tweets that appeal for donations to recovery funds, suggest innovations for disaster relief, discuss the challenges faced by non-profit organizations in response efforts, and share information on how to help and stay informed about the earthquake's

aftermath. Notably, these tweets revolve around philanthropic efforts, providing resources and sharing information relevant to the earthquake's impact and response initiatives.

This grouping of tweets signifies that the embeddings, likely calculated by the BERT model, successfully capture the semantic similarity between these messages. The relatedness of these tweets depends upon their context and content, which is centered around the Nepal earthquake relief efforts. The model's ability to cluster these similar tweets through embeddings is a testament to the effectiveness of BERT in understanding and capturing the nuances of language use in social media content.

4.4 Baseline Result - Event Type Classification

The results of our event type classification, presented in Table 3, offer a comprehensive overview of the performance of the five transformers models we utilize in the CURD$_{dl}$ classifier model.

Our experimental findings in Table 3 demonstrate that CURD$_{dl}$ – RoBERTa is the most effective model for event-type classification, achieving an accuracy rate of 89%, which is the highest among the five transformers models we used in the CURD$_{dl}$ model. This model also exhibits a balanced performance across precision, recall, and weighted F1 score, all at 0.89. CURD$_{dl}$ robustness makes it suitable for different pre-trained transformers and disaster-related tasks where accurate event-type classification is crucial to accommodate all information into different class labels.

Models	Precision	Recall	F1	Accuracy
CURD$_{dl}$ – BERT	0.88	0.88	0.88	0.88
CURD$_{dl}$ – BERTweet	0.88	0.88	0.88	0.88
CURD$_{dl}$ – RoBERTa	**0.89**	**0.89**	**0.89**	**0.89**
CURD$_{dl}$ – DistilBERT	0.88	0.88	0.88	0.88
CURD$_{dl}$ – XLNet	0.81	0.84	0.82	0.84

Table 3: Baseline Models for Event Type Classification

The analysis provides valuable insights that are important for future research and applications. The results indicate that CURD$_{dl}$ – RoBERTa is the top choice for tasks requiring high accuracy in event-type classification. However, CURD$_{dl}$ – BERT, CURD$_{dl}$ – BERTweet, and CURD$_{dl}$ – DistilBERT are also suitable alternatives, especially when computational efficiency is a priority. This analysis serves as a reference point for future research using the 29 event-type annotation class labels, highlighting the importance of selecting the appropriate model based on specific task requirements and the potential need to improve performance in more challenging classification categories.

4.5 Baseline Result – Relevancy

Our findings on relevance classification in Table 4 demonstrate the high performance of our models in classifying tweets as relevant or non-relevant. With precision, recall, F1 score, and accuracy scores mainly exceeding 0.90, the models exhibit robustness in accurately distinguishing between the two categories, which is crucial in identifying critical information from social media during disasters.

CURD$_{dl}$ – BERTweet and CURD$_{dl}$ – RoBERTa models exhibit negligible superior performance, with an accuracy of 0.91 and F1 scores of 0.91. CURD$_{dl}$ – RoBERTa's AUC score of 0.96 indicates its superior ability to differentiate between classes, suggesting its suitability for relevance classification. These models effectively sift through large volumes of social media data to classify crucial information.

Models	Precision	Recall	F1	Accuracy	AUC
CURD$_{dl}$ – BERT	0.89	0.90	0.89	0.90	0.95
CURD$_{dl}$ – BERTweet	0.91	0.91	0.91	0.91	0.95
CURD$_{dl}$ – RoBERTa	**0.91**	**0.91**	**0.91**	**0.91**	**0.96**
CURD$_{dl}$ – DistilBERT	0.90	0.90	0.90	0.90	0.95
CURD$_{dl}$ – XLNet	0.90	0.90	0.90	0.90	0.94

Table 4: Performance Evaluation for Relevancy Classification

4.6 Baseline Result – Urgency

The results in Table 5 reveal that 3 models (CURD$_{dl}$ – BERTweet, CURD$_{dl}$ – BERT, and CURD$_{dl}$ – RoBERTa) have same scores that outperform CURD$_{dl}$ – DistilBERT and CURD$_{dl}$ – XLNet models in classifying tweets based on their urgency, with 1% precision, recall, F1, and accuracy of 0.88, and 0.94 AUC. This suggests that these models have a slightly better ability to distinguish urgent tweets from non-urgent ones.

Models	Precision	Recall	F1	Accuracy	AUC
CURD$_{dl}$ – BERT	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**
CURD$_{dl}$ – BERTweet	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**
CURD$_{dl}$ – DistilBERT	0.87	0.87	0.87	0.87	0.92
CURD$_{dl}$ – RoBERTa	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**
CURD$_{dl}$ – XLNet	0.87	0.87	0.87	0.87	0.94

Table 5: Evaluation for Urgency Classification Models

The results indicate that CURD$_{dl}$ – BERT and CURD$_{dl}$ – RoBERTa are suitable for urgency classification, while CURD$_{dl}$ – BERTweet's performance is worth noting, given its special pre-training and tuning for X data. The choice of the CURD$_{dl}$ model may also depend on factors such as pre-trained data and parameters, where CURD$_{dl}$ – RoBERTa offers a specific advantage.

Our dataset size was reduced to 18,911 from over 26,000 in the relevance annotation as we only used the relevant class to determine the message's urgency. This is evident in the urgency classification results compared to the relevance classification despite using the same classifier for both tasks.

Notably, the five models in Tables 3, 4, and 5 show minimal differences in evaluation metrics, demonstrating the robustness and versatility of CURD$_{dl}$. This consistency is largely due to the roles of other components of the classifier like the customize layer.

4.7 Urgency Ranking by Score

The urgency ranking as depicted in Table 6 enhances the prioritization process by providing urgency scores. Higher scores, as observed in tweets regarding the disaster response tweets, indicate greater urgency based on severity, context, significance, and need for response services during a disaster.

In Table 6, the urgency ranking is determined by the severity of an event that measures the immediate human impact. Tweets

Table 6: Urgency Ranking - Evaluated the urgency of tweets based on severity, context, significance, and the requirement for response services.)

Tweet	Event Type	Relevancy	Urgency	Urgency Score
1. #NepalEarthquake | 40 school children injured in #WestBengal Live updates	casualty	Relevant	Urgent	258.89
2. A children's hospital in Texas is evacuating ten of its sickest and smallest patients out of Hurricane Harvey	casualty	Relevant	Urgent	258.07
3. #BREAKING: #Sandy now 40 miles from New Jersey as winds water swell #SandyNJ	damage	Relevant	Urgent	257.89

detailing casualties, such as Tweet 1 of Table 6 reporting 40 school children injured in the Nepal earthquake, signal a high level of urgency. Such situations necessitate swift medical and rescue operations. The context of the tweet amplifies this urgency, particularly when vulnerable groups are mentioned. The Nepal earthquake and Tweet 2 of the Texas hospital evacuation tweets highlight situations involving at-risk populations like school children and sick infants, respectively, underscoring the gravity of these situations.

The impact of a disaster tweet is reflected in its potential to affect numerous lives and communities. A tweet mentioning a large number of injuries, for instance, not only points to a considerable event but also to the ripple effect it has on the community. The response services' needs are closely tied to this. Tweets indicating urgent situations, like the injuries of children or the evacuation of critically ill infants from Tweet 1 and 2, respectively, call for immediate and well-coordinated medical and logistical support.

On the other hand, tweets that serve more as alerts, like Tweet 3 about Hurricane Sandy approaching New Jersey, provide valuable information for preparedness and preventive measures. Though they convey urgency, they typically demand a different level of urgency and response than those reporting actual incidents with established casualties or ongoing crises. This evaluative approach to tweet content ensures that the most critical situations receive prompt attention and resources required amidst disaster scenarios.

4.8 Comparison with Other Works (Relevancy and Urgency)

We compared our best-performing model (CURD_{dl}) for relevancy and urgency, respectively, with results from other state-of-the-art-models from related studies in the field of disaster management for tweet classification. To enhance the coherence of the comparison, we categorized it into two types, direct and indirect comparison.

4.8.1 ***Direct Comparison with State-Of-The-Art (SOTA) Models with their Corresponding Annotated Datasets.*** We compare our CURD_{dl} model with other SOTA models where we use publicly accessible datasets to train and test our model through direct comparisons. Our direct comparison includes works from [8], which employed non-neural (SVMs) and neural methods (CNN and Dual-CNN) for relevance classification and information type classification on the CrisisLexT26 dataset, which includes 26 crisis events from 2012 and 2013. Our approach is also compared to [4], which used domain adaptation with adversarial training and graph-based semi-supervised learning on tweets from the 2015 Nepal Earthquake and the 2013 Queensland Floods.

To provide a thorough evaluation, we utilized precision, recall, and F1 scores, as well as AUC scores from related studies. Each model in the table is named based on its authors first letter, the year of the study, and the model's name for multiple models. Our models are distinguished by 'Re' for relevance and 'Ur' for urgency. The abbreviations and corresponding classification tasks are: [4] - (F-NEQ and F-QFL), [8] - (B-SVM and B-DCNN), [6] - A-2014, [40] - Y-2017, and [31] - P-2023. This analysis not only places our work in the broader context, but also highlights our methodologies' advancements and unique contributions.

Direct Comparison - Relevancy					
Models	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
F–QFL	0.93	0.94	0.94	-	0.92
CURD_{dl}–Re	**0.97**	**0.97**	**0.97**	**0.97**	**0.99**
F–NEQ	0.65	0.65	0.65	-	0.65
CURD_{dl}–Re	**0.77**	**0.76**	**0.76**	**0.76**	**0.85**
B–SVM	0.87	0.74	0.79	-	-
B–DCNN	0.86	0.76	0.80	-	-
CURD_{dl}–Re	**0.93**	**0.93**	**0.93**	**0.93**	**0.94**
Direct Comparison - Urgency					
B–SVM	0.64	0.60	0.62	-	-
B–CNN	0.63	0.59	0.61	-	-
CURD_{dl}–Ur	**0.86**	**0.86**	**0.86**	**0.86**	**0.93**

Table 7: Direct Comparison with SOTA Models

Our analysis in Table 7 reveals that the CURD_{dl} model outperformed all state-of-the-art-models (F–QFL, B–SVM, and B–DCNN) in relevancy and urgency classification, showcasing its exceptional, contextual understanding of disaster-related tweets.

4.8.2 ***Indirect Comparison with SOTA Models and Datasets.*** Indirect comparisons involve studies that are related to our task but do not make their datasets publicly available. For instance, Ashktorab et al. [6] developed logistic regression models to classify tweets for mentions of human or infrastructure damage. Yang et al. [40] created SVM classifiers to identify rescue requests in tweets during Hurricane Harvey. Lastly, Powers et al. [31] proposed a binary relevance and urgency classification scheme for Hurricane Harvey tweets and developed both neural and non-neural ML models. By comparing our approach with these studies, we can better understand the strengths and weaknesses of our models, as well as potential areas for improvement.

In the indirect comparison analysis presented in Table 8, CURD_{dl} exhibited outstanding performance compared to models such as

Indirect Comparison - Relevancy					
Models	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
A–2017	0.78	0.57	0.65	0.86	0.88
Y–2017	0.61	0.79	0.69	0.93	-
P–2023	0.78	0.79	0.78	0.78	0.78
$CURD_{dl}$–Re	**0.91**	**0.91**	**0.91**	**0.91**	**0.96**
Indirect Comparison - Urgency					
P–2023	0.67	0.71	0.68	0.77	0.71
$CURD_{dl}$–Ur	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**

Table 8: Indirect Comparison with SOTA Models

A–2017, Y–2017, and P–2023. These models were impressive in accuracy and AUC, highlighting the advancements of our approach in relevancy and urgency tasks. This comparison, especially with P–2023, underscores the potential of our models in real-world disaster response applications. The results from both comparison types affirm our models' effectiveness in complex disaster tweet classification. Notably, the outstanding performance of $CURD_{dl}$ in terms of accuracy and AUC highlights its potential for practical implementations. This comparative study contributes to ongoing research, providing valuable insights for future endeavors.

4.9 Ablation Study

Accurately categorizing tweets is crucial for efficient aid provision in disaster response. In order to achieve this objective, we conducted an ablation study to determine the impact of our additional design features on the pre-trained transformer and the $CURDdl_{dl}$ – RoBERTa model's functionality. Our best hyperparameter combination from ablation experiments is a learning rate of 2e-5 and a batch size of 32. We also noticed there were negligible differences with or without preprocessing techniques during the ablation.

Figure 5a displays the accuracy and F1 scores for event-type classification. The $CURD_{dl}$ – RoBERTa model attains 89 in both metrics. Without the convolutional layer, the scores decrease to 65 and 66, and removing both convolutional and custom layers results in the lowest scores of 57 and 56. Figure 5b demonstrates the $CURD_{dl}$ – RoBERTa model's performance in relevancy classification using accuracy, F1 score, and AUC as three metrics. The model achieves impressive results with all layers intact, including 91% accuracy, 78% F1 score, and 87% AUC. These results highlight the significance of these layers in identifying relevant disaster communications. Figure 5c demonstrates the urgency of handling tweets during disasters. The $CURD_{dl}$ – RoBERTa model scores 87 in accuracy, F1, and AUC, while models without custom layers perform worse, emphasizing the importance of advanced feature extraction for determining urgency.

The inclusion of custom attention and convolutional layers boosts $CURD_{dl}$ – RoBERTa's classification abilities, and their removal decreases performance, highlighting their vital role in accurate disaster response classification.

5 CONCLUSION AND FUTURE WORK

This study presents CURD, an innovative approach for annotating, classifying, and prioritizing social media content during disasters. Our annotation process is a three-way systematic labeling (1) 29 event type, (2) binary relevance, and (3) binary urgency. This method effectively tackles the challenges of determining event type, relevance, and urgency in disaster-related tweets. Our classifier ($CURD_{dl}$) leverages a combination of transformer-based models, a convolution layer, and custom attention layers, resulting in exceptional filtering performance and assessing classification tasks. The experimental evaluations highlight the potential of CURD in enhancing real-time disaster management and response while computing urgency scores for urgent tweets. The findings of this study highlight the significance of utilizing advanced deep learning techniques to process large social media data for crucial applications such as disaster response.

In future work, we plan to use multilingual and cross-cultural adaptation, where we will expand the CURD model's capabilities to process and interpret data from multiple languages and cultural contexts, thereby increasing its global applicability in diverse disaster scenarios. We also plan to integrate with other social media platforms, which will expand the scope to analyze data from various social media platforms other than X to gather more comprehensive public responses and information during disaster.

ACKNOWLEDGMENTS

This research received support from the NSF - USA CNS-2219615, 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] Ademola Adesokan and Sanjay Madria. 2023. NeuEmot: Mitigating Neutral Label and Reclassifying False Neutrals in the 2022 FIFA World Cup via Low-Level Emotion. In *2023 IEEE International Conference on Big Data (BigData)*. 578–587. https://doi.org/10.1109/BigData59044.2023.10386146

[2] Ademola Adesokan, Sanjay Madria, and Long Nguyen. 2023. Hatemotweet: Low-level emotion classifications and spatiotemporal trends of hate and offensive COVID-19 tweets. *Social Network Analysis and Mining* 13 (2023), 136. https://doi.org/10.1007/s13278-023-01132-6

[3] Ademola Adesokan, Sanjay Madria, and Long Nguyen. 2023. TweetACE: A Fine-grained Classification of Disaster Tweets using Transformer Model. In *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. 1–9. https://doi.org/10.1109/AIPR60534.2023.10440656

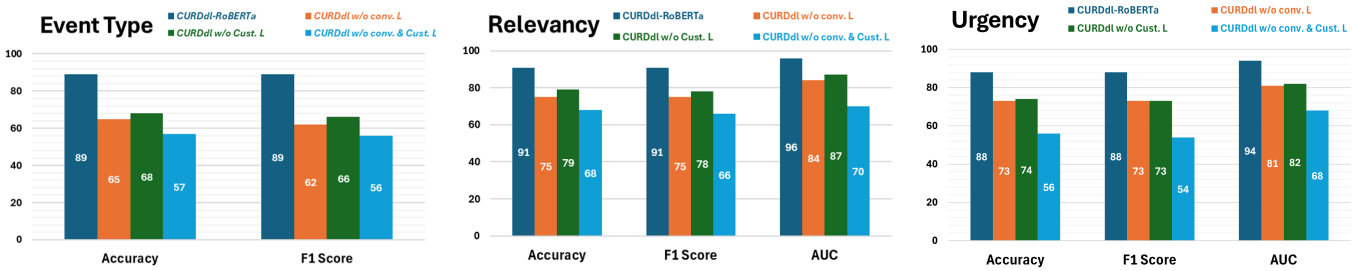
[4] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain Adaptation with Adversarial Training and Graph Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, Melbourne, Australia, 1077–1087.

[5] Abdullah M. Alkadri, Abeer ElKorany, and Cherry A. Ezzat. 2023. An Integrated Framework for Relevance Classification of Trending Topics in Arabic Tweets. *International Journal of Advanced Computer Science and Applications* 14, 7 (2023).

[6] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining Twitter to inform disaster response. In *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*. 269–272.

[7] Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Gaëtan Chevalier, and Laurent Leygue. 2022. Are Sudden Crises Making me Collapse? Measuring Transfer Learning Performances on Urgency Detection. In *Proceedings of the 19th International Conference on Information Systems for Crisis Response and Management*. ISCRAM, Tarbes, France.

[8] Gregoire Burel and Harith Alani. 2018. Crisis event extraction service (CREES): Automatic detection and classification of crisis-related content on social media. (2018).



(a) Event Type (b) Relevancy (c) Urgency

Figure 5: Visual Representations of CURD Ablation Study

[9] Denis Eka Cahyani and Alfan Wiguna Putra. 2021. Relevance Classification of Trending Topic and Twitter Content Using Support Vector Machine. In *2021 International Seminar on Application for Tech. of Info. and Comm. (iSemantic)*. IEEE.

[10] Enrique Cano-Marin, Marçal Mora-Cantallops, and Salvador Sánchez-Alonso. 2023. Twitter as a predictive system: A systematic literature review. *Journal of Business Research* 157 (2023). https://doi.org/10.1016/j.jbusres.2022.113561

[11] Ashwin Devaraj, Dhiraj Murthy, and Aman Dontula. 2020. Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction* 51 (2020), 101757.

[12] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1.

[13] Laurenti Enzo, Bourgon Nils, Farah Benamara, Mari Alda, Véronique Moriceau, and Courgeon Camille. 2022. Speech acts and Communicative Intentions for Urgency Detection. In *11th Joint Conference on Lexical and Computational Semantics (* SEM 2022)*. Association for Computational Linguistics.

[14] Hafiz Budi Firmansyah, Jesús Cerquides, and Jose Luis Fernandez-Marquez. 2022. Ensemble Learning for the Classification of Social Media Data in Disaster Response. In *International Conference on Info. Sys, for Crisis Response and Mgt.* https://api.semanticscholar.org/CorpusID:253448364

[15] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (1971). https://doi.org/10.1037/h0031619

[16] HuggingFace. 2020. Transformers 2.4.0 documentation. https://huggingface.co/transformers/v2.4.0/pretrained_models.html. Accessed 27-10-2023.

[17] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Extracting Information Nuggets from Disaster-Related Messages in Social Media. In *ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management*. https://dblp.org/rec/conf/iscram/ImranECDM13.html

[18] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*. 1638–1643.

[19] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. An efficient k-means clustering algorithms: Analysis and implementation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, 7 (2002). https://doi.org/10.1109/TPAMI.2002.1017616

[20] Mayank Kejriwal and Peilin Zhou. 2020. On detecting urgency in short crisis messages using minimal supervision and transfer learning. *Social Network Analysis and Mining* 10, 58 (2020).

[21] Abhinav Kumar, Jyoti Prakash Singh, Yogesh K. Dwivedi, and Nripendra P. Rana. 2022. A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research* 319 (2022), 791–822.

[22] Xinwei Li, Mao Xu, Wenjuan Zeng, Ying Kei Tse, and Hing Kai Chan. 2023. Exploring customer concerns on service quality under the COVID-19 crisis: A social media analytics study from the retail industry. *J. of Retailing and Consumer Services* 70 (2023). https://doi.org/10.1016/j.jretconser.2022.103157

[23] Valerio Lorini, Emanuele Panizio, and Carlos Castillo. 2022. SMDRM: A Platform to Analyze Social Media for Disaster Risk Management in Near Real Time. In *ICWSM Workshops*. https://api.semanticscholar.org/CorpusID:249651281

[24] Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and P Jayadev. 2021. Multi-modal classification of Twitter data during disasters for humanitarian response. *Journal of ambient intelligence and humanized computing* 12 (2021), 10223–10237. https://doi.org/10.1007/s12652-020-02791-5

[25] Viyom Mittal, Hongmiao Yu, and K. K. Ramakrishnan. 2022. FUSED: Fusing Social Media Stream Classification Techniques for Effective Disaster Response. In *Proceedings - 1st Workshop on Cyber Phys. Sys. for Emergency Resp., CPS-ER 2022*. https://doi.org/10.1109/CPS-ER56134.2022.00013

[26] Xenia Yasmin Zia Gutierrez Morales. 2010. *Networks to the Rescue: Tweeting Relief and Aid During Typhoon Ondoy*. Georgetown Univ. https://www.amazon.in/Networks-Rescue-Tweeting-Relief-Typhoon/dp/1248974336

[27] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. 2016. Applications of online deep learning for crisis response using social media information. arXiv preprint arXiv:1608.03902.

[28] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the 8th international AAAI conference on weblogs and social media (ICWSM'14)*. No. CONF.

[29] Avijit Paul. 2015. *Identifying Relevant Information for Emergency Services from Twitter in Response to Natural Disaster*. Ph. D. Dissertation. Queensland University of Technology. https://eprints.qut.edu.au/89220/1/Avijit_Paul_Thesis.pdf

[30] Mark Edward Phillips. 2017. Hurricane Harvey Twitter dataset. https://digital.library.unt.edu/ark:/67531/metadc993940/. Accessed January 29, 2018.

[31] Courtney J. Powers, Ashwin Devaraj, Kaab Ashqeen, Aman Dontula, Amit Joshi, Jayanth Shenoy, and Dhiraj Murthy. 2023. Using artificial intelligence to identify emergency messages on social media during a natural disaster: A deep learning approach. *International Journal of Info. Mgt. Data Insights* 3 (2023), 100164.

[32] Francisco J. Pérez-Invernón, Francisco J. Gordillo-Vázquez, Heidi Huntrieser, and Patrick Jöckel. 2023. Variation of lightning-ignited wildfire patterns under climate change. *Nature Comm.* 14, 1 (2023). https://doi.org/10.1038/s41467-023-36500-5

[33] Brett W. Robertson, Matthew Johnson, Dhiraj Murthy, William Roth Smith, and Keri K. Stephens. 2019. Using a Combination of Human Insights and 'Deep Learning' for Real-Time Disaster Communication. *Progress in Disaster Science* 2 (2019). https://doi.org/10.1016/j.pdisas.2019.100030

[34] James A. Smith, Mary Lynn Baeck, Yibing Su, Maofeng Liu, and Gabriel A. Vecchi. 2023. Strange Storms: Rainfall Extremes From the Remnants of Hurricane Ida (2021) in the Northeastern US. *Water Resources Research* 59, 3 (2023). https://doi.org/10.1029/2022WR033934

[35] Luke S Snyder, Yi-Shan Lin, Morteza Karimzadeh, Dan Goldwasser, and David S Ebert. 2020. Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 558–568. https://doi.org/10.1109/TVCG.2019.2934614

[36] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Prog. Theor. Phys. Suppl.* 110 (2006).

[37] Hongxia Wei, Yingyuan Xiao, Wenguang Zheng, and Chen Dong. 2021. News-Comment Relevance Classification Algorithm Based on Feature Extraction. In *2021 International Conf. on Big Data Analysis and Computer Sci. (BDACS)*. IEEE.

[38] Chaoran Xu, Benjamin T Nelson-Mercer, Jeremy D Bricker, Meri Davlasheridze, Ashley D Ross, and Jianjun Jia. 2023. Damage Curves Derived from Hurricane Ike in the West of Galveston Bay Based on Insurance Claims and Hydrodynamic Simulations. *International Journal of Disaster Risk Science* (2023), 1–15. https://doi.org/10.1007/s13753-023-00524-8

[39] Kamini Yadav, Francisco J. Escobedo, Alyssa S. Thomas, and Nels G. Johnson. 2023. Increasing wildfires and changing sociodemographics in communities across California, USA. *International Journal of Disaster Risk Reduction* 98 (2023), 104065. https://doi.org/10.1016/j.ijdrr.2023.104065

[40] Zhiwei Yang, Linh Hoang Nguyen, John Stuve, Guofeng Cao, and Fei Jin. 2017. Harvey flooding rescue in social media. In *Proceedings of the IEEE Intl. Conf. on Big Data*. IEEE, 2177–2185.

[41] Jiawei Yong. 2019. A Cross-Topic Method for Supervised Relevance Classification. In *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*. ACL, Hong Kong, 147–152.