KeyMinES: Extracting Minimal Keyphrases for Sub-events in Disaster Situations

Ademola Adesokan
Computer Science
Missouri University of Science and Technology
Rolla, United States
aaadfg@mst.edu

Sanjay Madria
Computer Science
Missouri University of Science and Technology
Rolla, United States
madrias@mst.edu

***Abstract*—The substantial volume of unstructured social media data generated during disasters often conceals critical information. Developing efficient methods to extract actionable insights from this data can significantly enhance emergency response and resource allocation. However, existing methods, primarily reliant on supervised learning, encounter challenges such as dependence on labeled data, limited adaptability, and scalability. To overcome these limitations, we present KeyMinES, an unsupervised model that extracts minimal keyphrases—bigrams and tokens—from social media data to identify and classify critical sub-events. Our approach integrates semantic and grammar-based reconstruction to ensure that the extracted keyphrases are both grammatically correct and contextually meaningful. Through clustering, we group these reconstructed sub-events, enabling the identification of patterns and offering actionable insights for decision-makers. Our experimental results, attained through quantitative and qualitative evaluations, demonstrate that KeyMinES outperforms baseline methods, achieving higher F1 scores and providing a scalable and cost-effective solution. Our ablation study reveals that combining bigram+token enhances sub-event detection compared to using only bigram or token, capturing both contextual relationships and granular details, thereby leading to more accurate identification of critical sub-events. This model holds significant potential for various stakeholders, including emergency responders and humanitarian organizations, by improving the extraction of actionable insights during disasters.**

***Index Terms*—Sub-event detection, Disaster management, Keyphrase extraction, Unsupervised learning, Social media data.**

I. Introduction

Identifying sub-events and extracting pertinent information from major disaster events is paramount in ensuring a thorough, efficient, and adaptable response [1]. These procedures enrich situational awareness, bolster decision-making, and ensure effective resource distribution, all of which are crucial for handling a disaster's immediate and enduring repercussions. By concentrating on these pivotal elements, disaster management endeavors are more precise, prompt, and triumphant in preserving lives and mitigating harm.

In times of crisis, social media platforms such as X (formerly known as Twitter) have become essential for emergency response, serving as a valuable support system when traditional methods like 911 dispatch calls are overwhelmed [2]. As critical events unfold, individuals on the ground regularly share real-time information on these platforms. 911 dispatchers utilize this data to evaluate the situation and efficiently deploy the appropriate emergency services, including police, fire, or medical personnel, to the scene. However, the sheer volume of tweets can challenge first responders in making timely decisions. Furthermore, X informal and unstructured content nature often complicates extracting pertinent information. Various stakeholders have distinct information requirements. For instance, first responders rely on detailed situational awareness such as *evacuation help, flood warnings, blood donations,* and *roof damage* to make well-informed decisions that prioritize life-saving actions while facilitating efficient resource allocation during a crisis. In contrast, policymakers concentrate on broader aspects such as Infrastructure and Utilities *(e.g., roads and power lines)*, Population Displacement *(e.g., the number of displaced individuals and available shelters)*, Humanitarian Assistance *(e.g., distribution of aid and medical supplies)*, and Environmental Impact *(e.g., pollution levels and ecological damage)* to develop comprehensive policies that promote sustainable recovery and enhance resilience against future disasters.

In this work, we characterize an ***event*** as a catastrophic disaster, such as an earthquake or hurricane, resulting in widespread devastation. These occurrences encompass numerous significant ***sub-events,*** such as *children drowned, floodwater rising, school collapsed, pets trapped, power blackout,* and *babies safety*. For example, during the Hurricane Harvey Event, a tweet *"Be Aware: Most damage will include POWER OUTAGES, trees falling, debris accumulation, severe wind. #HurricaneHarvey"*. In this instance, *"Power Outages," "Trees Falling," "Debris Accumulation,"* and *"Severe Wind"* are potential sub-events associated with Hurricane Harvey Event. The majority of studies in disaster event and sub-event classification have conventionally employed supervised learning techniques using labeled data [3]–[6]. However, this approach faces significant limitations when applied to social media data during disasters, including dependence on labeled data, limited adaptability, bias, scalability issues, and higher resource demands. In contrast, unsupervised learning offers greater flexibility and scalability, making it better suited for handling large volumes of unlabeled data, uncovering hidden patterns, and adapting to the dynamic and unpredictable nature of disaster scenarios, thus leading to more effective and timely insights. This is supported by research from Rudra et al. [7] and Arachie et al. [8] utilizing unsupervised techniques to identify sub-events within larger events in a disaster scenario.

Our approach extends the work of Rudra et al. [7] and Arachie et al. [8]. Rudra et al. [7] represent sub-events primarily as noun-verb pairs. Although this method effectively captures many but not all sub-events. Arachie et al. [8] improve upon this approach by integrating phrase extraction and ranking candidate sub-events using a crisis ontology[1] while filtering out noise and irrelevant information. However, the restriction of their ontology to only 62 crisis terms limits its ability to capture, rank and cluster a wide range of sub-events, and their work also did not discuss the most prominent phrase or grammar combinations in their dataset. For instance, the tweet *"Expecting major or record flooding in most major area rivers, per State Operations Briefing in Austin, Tx. #hurricaneharvey"* highlights specific information about the severity and nature of the flooding, with *"major flooding"* as a candidate sub-event. This phrase does not

[1] http://observedchange.com/moac/ns/

adhere to the noun-verb pair structure outlined by [7], as it utilizes an "Adjective/Past Participle + Main Verb" structure (*"major"* as an adjective modifying *"flooding,"* a gerund, or a nominal form derived from a verb). Furthermore, the method of Arachie et al. [8] may not effectively capture this sub-event due to the limited number of crisis-related terms in the crisis ontology, which restricts its ability to identify and rank diverse sub-events like *"major flooding."* Both methods are limited by their focus on specific grammatical structures, thereby impeding their effectiveness in capturing sub-events expressed through more varied phrase structures, such as those involving adjectives and modifiers. The extracted sub-events can provide a better description and meaning if the phrase is 3-word, for example, *"Expecting major flooding."*

In order to tackle the aforementioned challenges and proficiently extract robust and meaningful sub-events, we propose **KeyMinES**, an innovative and simple method for **E**xtracting **Min**imal **Key**phrases for **S**ub-events in Disaster Situations, which recognizes the most critical two- or three-keyword expressions by selecting sub-phrases (at the bigram and token level) that is closely similar to a disaster event and effectively depict it. This process entails merging a candidate bigram phrase (the bigram with the highest similarity score to the tweet) with a token-based ranking (the token with the highest similarity score to the tweet) to form the candidate sub-event keyphrase. Our approach involves integrating semantic and grammar-based reconstruction rules to standardize these extracted keyphrases and confirm their adherence to proper English grammar, thus enhancing human understanding and interpretability in the context of real-world events. Furthermore, we group the grammatically constructed sub-events (combined candidate bigram phrases and token-level words) into sub-event clusters for high-level understanding. These clusters provide a comprehensive overview of the crisis and aid in recognizing critical sub-events.

The key contributions of our work are outlined as follows:

- We introduce KeyMinES, a scalable, unsupervised model for extracting minimal keyphrases that efficiently identifies and classifies sub-events during disasters. By combining bigram keyphrases, extracted using KeyBERT[2], with tokens through tokenization and similarity scoring, the model enhances phrase reconstruction, resulting in improved detection of critical sub-events and better situational awareness.
- We present a novel method for trigger classification, which categorizes keyphrase/token according to their grammatical roles and reconstructs them into coherent sub-events for enhanced human comprehension. Through our novel semantic and grammar-based rules, KeyMinES ensures that keyphrases are grammatically accurate and contextually relevant, thereby enhancing the clarity and usability of crisis information for stakeholders.
- We group similarly reconstructed sub-events into clusters, enabling the identification of patterns and the generation of actionable insights for disaster response. In contrast to methods constrained by predefined ontologies, KeyMinES dynamically identifies sub-events through similarity scoring and clustering. This approach offers adaptability across various disaster contexts and facilitates the better prioritization of response efforts.
- An intensive evaluation of KeyMinES demonstrates its superiority over baseline methods in identifying and clustering sub-events, as evidenced through both quantitative and qualitative analyses. Additionally, an ablation study was conducted, highlighting the effectiveness of combining bigrams and tokens, with the bigram+token model achieving significantly higher F1-scores. This model offers a scalable, cost-effective solution for extracting actionable insights, benefiting key stakeholders such as emergency responders and crisis management teams during disasters.

The subsequent sections of this paper are organized as follows: Section II provides a review of relevant literature on event and sub-event detection, focusing on supervised, unsupervised, and semi-supervised techniques in disaster management. Section III outlines the KeyMinES model, encompassing keyphrase extraction, trigger classification, and clustering methods. In Section IV, we present the dataset, baseline approaches, and evaluation metrics utilized in our experiments. Section V delves into the experimental results, offering a comprehensive analysis of the model's performance and a comparison with baseline methods. Section VI provides an in-depth discussion of the findings, emphasizing the trade-offs between immediacy and contextual details in sub-event extraction and highlighting the implications for disaster response while underlining non-disaster applicability. Finally, Section VII concludes by summarizing key insights and suggesting directions for future work.

II. Related Work

Identifying smaller sub-events facilitates more precise monitoring and analysis, leading to a better comprehension of the event's development and potential impact. Recent studies [25]–[27] illustrate the crucial role of social media in disseminating information during crises. However, the substantial volume of content and irrelevant information pose a significant challenge in accurately discerning situational awareness details. Therefore, expeditiously acquiring reliable information from credible sources is essential for making informed decisions during emergencies [28], [29]. Consequently, researchers have prominently emphasized evaluating, categorizing, and analyzing social media content to derive actionable insights [30], [31]. An essential element during disasters is the identification of smaller critical events during major emergency [22], [23]. Several studies have delved into the detection of events and sub-events from tweets using diverse methodologies, including supervised approaches [3], [18] and unsupervised techniques [20], [21]. Our comprehensive review explores various approaches to event and sub-event detection, ranging from supervised to unsupervised methods.

Several studies in the supervised domain have contributed to the advancement of event detection and classification through innovative approaches and methodologies. Balali et al. [17] developed the COfEE Event ontology to address the challenge of event-type detection. The ontology integrates expert domain knowledge with a data-driven approach to enhance event extraction from text. It expands traditional frameworks by encompassing various event categories, including environmental issues, cyberspace, criminal activity, and natural disasters. Similarly, Dahou et al. [16] proposed an event detection framework that employs MobileBERT, a transformer-based model, for feature extraction to enhance the extraction of relevant crisis information from social media. They refined feature selection using a modified sparrow search algorithm (SSA) combined with manta ray foraging optimization (MRFO), demonstrating the effectiveness of their approach on various real-world datasets. Additionally, Adesokan et al. [3] also introduced TweetACE, a method that includes dual annotation of tweets for both event and sub-event types, with proper argument and their roles for accurate annotation and reliability in disaster management contexts. These studies set new benchmarks and identify areas for further refinement in handling disaster-related data.

Semi-supervised methods have been found to be effective in event and sub-event detection, especially in scenarios with limited labeled data. Sirbu et al. [15] employed an approach that extends the FixMatch algorithm to integrate text and image data. This multimodal strategy enhances the identification of relevant events and sub-events by leveraging multiple data modalities. Similarly, Wang and Wang [14] utilized a BERT-based semi-supervised domain adaptation model for multimodal disaster event classification, incorporating textual, image, and image description features to enhance classification performance across various disaster domains. This method leverages both linguistic and visual information, increasing the model's robustness in diverse disaster scenarios. Furthermore, Zou et al. [13] introduced a technique that utilizes both labeled and unlabeled data to detect and classify fine-grained disaster-related information, particularly useful in rapidly

[2] https://maartengr.github.io/KeyBERT

evolving events with limited labeled data, enhancing early disaster response and situational awareness.

Despite their effectiveness, supervised and semi-supervised methods rely heavily on high-quality labeled data for training, which can be expensive and time-consuming [12]. Moreover, these methods often face challenges in adapting to new domains because of their dependence on predefined lexical or syntactic features, limiting their generalizability [11]. In contrast, unsupervised methods offer greater scalability and adaptability, as they can identify patterns without the need for labeled data or extensive feature engineering, which makes them more versatile across diverse domains.

Using unsupervised methods, Chowdhury et al. [10] implemented a biclustering technique to cluster tweets and words from microblog posts during disasters. They also implemented a ranking mechanism to determine the most relevant sub-events based on frequency, part-of-speech tagging, and informativeness, thereby improving the accuracy of disaster response and management. Similarly, Belcastro et al. [4] introduced SEDOM-DD, a method that utilizes social media data to detect sub-events during disasters by systematically processing user posts, including collection, filtering, geolocation enrichment, and clustering of sub-events. This approach accurately identified geographic areas and sub-event types across natural and synthetic datasets. In a different approach, Cao et al. [9] presented HISEvent, an unsupervised social event detection framework that utilizes a hierarchical and incremental structural entropy minimization approach to detect social events from message graphs without relying on labeled data or a predefined number of events. Their method demonstrated superior performance compared to existing graph neural network (GNN)-based methods.

Our study both expands upon and differs from the research conducted by Rudra et al. [7] and Arachie et al. [8] concerning identifying and summarizing sub-events from microblogs in the context of disasters. Rudra et al. [7] employed a linguistic approach by using a dependency parser to represent nouns as entities (e.g., "building," "house"), and verbs that denote actions related to these entities (e.g., "collapsed," "burning"). The extracted NV pairs, which indicate potential sub-events, were ranked according to co-occurrence frequency using the Szymkiewicz-Simpson overlap score. Furthermore, a discounting factor was applied for the refinement of less frequent but relevant sub-events. The ranked NV pairs were then used to create concise summaries through an Integer Linear Programming (ILP) technique tailored to address the informational needs of disaster responders. Likewise, Arachie et al. [8] introduced an unsupervised learning framework for analyzing social media posts, particularly tweets, to detect sub-events during large-scale disasters. Their approach involved extracting noun-verb pairs and keyphrases, followed by semantic embedding and ranking against a crisis-specific ontology to eliminate extraneous and irrelevant information, showcasing the superior performance in accurately identifying and summarizing critical disaster response and management sub-events. The most relevant sub-events were clustered to group similar occurrences, thereby enhancing situational awareness and assisting emergency responders and policymakers.

In contrast to other approaches, such as those by Rudra et al. [7] and Arachie et al. [8], our method, KeyMinES, introduces several key innovations. Whereas Rudra et al. [7] predominantly utilize noun-verb structures, KeyMinES focuses on extracting concise yet highly pertinent keyphrases that closely depict disaster events. The selection process encompasses both bigram phrases and individual tokens, utilizing their similarity scores to generate candidate sub-event keyphrases. Additionally, we implement semantic and grammar-based reconstruction rules to ensure these keyphrases' coherence and grammatical correctness, thereby bolstering their interpretability and practical relevance. Unlike Arachie et al. [8], our approach avoids reliance on a predetermined ontology, which can limit flexibility because of its restricted set of crisis terms. Instead, we dynamically identify the most pertinent sub-events through similarity scoring and clustering, thereby enhancing adaptability to diverse disaster contexts. Furthermore, KeyMinES introduces a novel trigger classification process, categorizing extracted keyphrases based on their grammatical roles, an aspect previously unexplored in related studies. By clustering reconstructed sub-events into meaningful groups, our method furnishes a more nuanced comprehension of disaster events, thereby offering actionable insights to assist decision-making processes during crises.

III. Our Approach

In this section, we describe our approach for minimal keyphrase detection (via a combination of bigram and token) and trigger identification from disaster tweets and explain how we utilize part-of-speech tagging to guide these sub-events and restructure their grammatical form to conform to English grammar and identify the candidate sub-event. We subsequently group these candidate-reconstructed phrases into unique clusters of sub-events. Our approach is structured into several components, spanning from preprocessing to the clustering stage, as shown in Figure 1, and we briefly describe each of the components as follows:

1. Preprocessing

Our preprocessing step involves cleaning and preprocessing the raw tweet data for analysis. This includes normalizing to lowercase, converting emojis to text, and removing stopwords, punctuation, and other non-essential parts of the tweet (like special characters). Our preprocessing applies a series of transformations to the input tweet T to generate a cleaned tweet T':

$$T' = \text{Clean}(T)$$

Where:

$$\text{Clean}(T) = T - (\text{stopwords, punctuation, and special characters})$$

Our preprocessing ensures that only the most relevant parts of the tweet are considered for further analysis, reducing noise and improving the accuracy of subsequent steps like keyphrase extraction [19].

2. keyphrases Extraction (Bigrams)

To extract keyphrases, we used KeyBERT[3], which uses BERT embeddings to extract relevant keyphrases (bigrams) from the preprocessed tweet. KeyBERT leverages the power of BERT embeddings, making it more effective in capturing the most contextually relevant keyphrases.

The method generates a substantial number of bigrams, as two-word combinations vary widely, with many candidates failing to qualify as sub-events. To eliminate noisy pairs, we focus on those most relevant to the tweet based on their similarity score. This approach differs from [8], which selects noun-verb pairs that occur multiple times. We assert that frequent word occurrences do not mean the word is important to the tweet [34]. This relevance is measured by the cosine similarity[4] between the tweet embedding and the bigram embedding, and this approach significantly minimizes the quantity of potential sub-events while ensuring that no critical sub-events are overlooked.

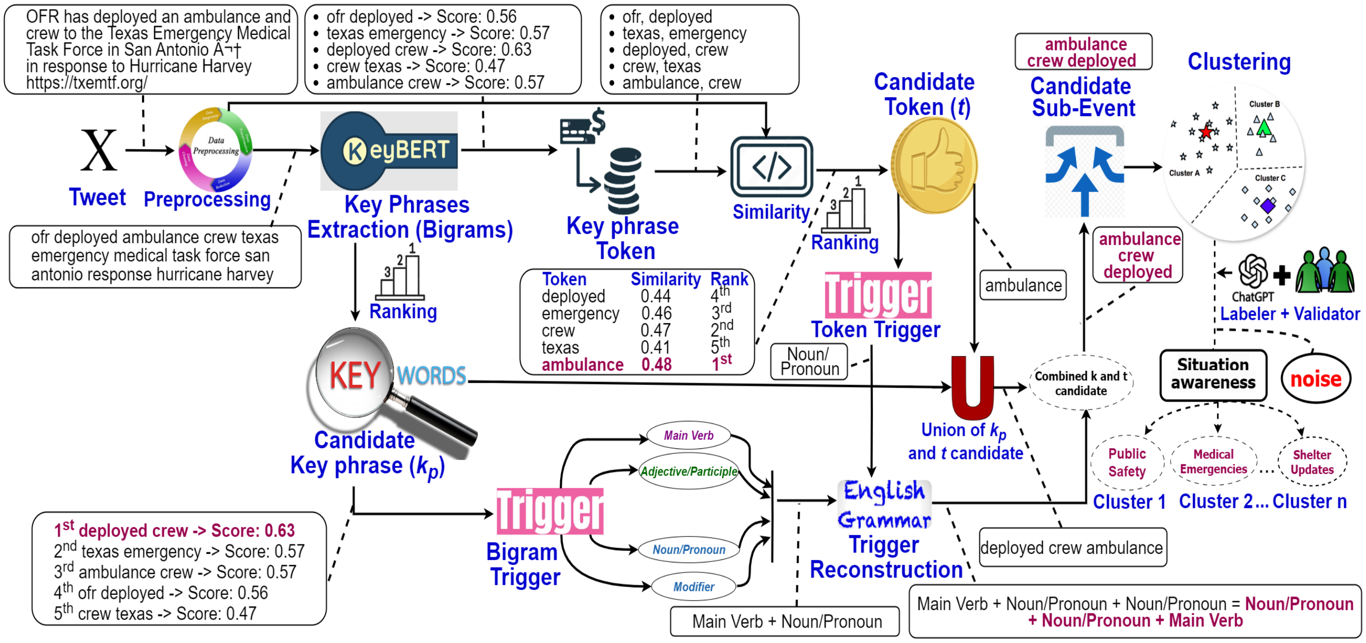
Given a tweet sentence S, KeyBERT generates embedding vectors $\mathbf{v}_i$ for each potential bigram P_i and then scores these bigrams based on their relevance, which we represent as:

$$\text{Score}(P_i) = \text{Relevance}(P_i|S)$$

An example of bigram extraction and scores for Tweet # ***"OFR has deployed an ambulance and crew to the Texas Emergency Medical Task Force in San Antonio in response to Hurricane Harvey https://txemtf.org/"*** are as follows: ***"deployed crew" (0.63)***, *"texas emergency" (0.57)*, *"ambulance crew" (0.57)*, *"ofr deployed" (0.56)*, and ***"crew texas" (0.47)***.

[3] https://maartengr.github.io/KeyBERT

[4] https://scikit-learn.org/cosine_similarity


OFR has deployed an ambulance and crew to the Texas Emergency Medical Task Force in San Antonio Â¬† in response to Hurricane Harvey https://txemtf.org/
• ofr deployed -> Score: 0.56
• texas emergency -> Score: 0.57
• deployed crew -> Score: 0.63
• crew texas -> Score: 0.47
• ambulance crew -> Score: 0.57
• ofr, deployed
• texas, emergency
• deployed, crew
• crew, texas
• ambulance, crew
X
Tweet
Data Preprocessing
Preprocessing
KeyBERT
Key Phrases Extraction (Bigrams)
Key phrase Token
Similarity
Ranking
Candidate Token (t)
ambulance crew deployed
Candidate Sub-Event
Clustering
Cluster A
Cluster B
Cluster C
ofr deployed ambulance crew texas emergency medical task force san antonio response hurricane harvey
Ranking
Trigger
Token Trigger
ambulance
ambulance crew deployed
ChatGPT
Labeler + Validator
Noun/ Pronoun
KEY WORDS
Candidate Key phrase (k_p)
U
Union of k_p and t candidate
Combined k and t candidate
Situation awareness
noise
Public Safety
Medical Emergencies
Shelter Updates
Cluster 1
Cluster 2... Cluster n
1st deployed crew -> Score: 0.63
2nd texas emergency -> Score: 0.57
3rd ambulance crew -> Score: 0.57
4th ofr deployed -> Score: 0.56
5th crew texas -> Score: 0.47
Trigger
Bigram Trigger
Main Verb
Adjective/Participle
Noun/Pronoun
Modifier
English Grammar Trigger Reconstruction
deployed crew ambulance
Main Verb + Noun/Pronoun
Main Verb + Noun/Pronoun + Noun/Pronoun = Noun/Pronoun + Noun/Pronoun + Main Verb

Token	Similarity	Rank
deployed	0.44	4th
emergency	0.46	3rd
crew	0.47	2nd
texas	0.41	5th
ambulance	**0.48**	**1st**

Fig. 1: Our Approach (KeyMinES).

The extraction of keyphrases (bigram) identifies the most relevant concepts in the tweet, which are crucial for understanding the main topics and actions described in the tweets.

3. Keyphrase Token and Similarity

To extract the tokens from the bigram keyphrase, we break down the extracted bigram keyphrases into individual tokens, and each token is scored based on its importance to the tweet (e.g., ***"ambulance crew"*** becomes ***"ambulance" [0.48], "crew" [0.47])***.

If P_b is a bigram keyphrase, we tokenize it into a set of tokens $T = \{t_1, t_2\}$, and each token t_i is assigned a similarity score using Sentence-Transformer *(all-MiniLM-L6-v2)*[5] to generate contextual embeddings that capture the nuanced meanings of phrases better than traditional methods like bag-of-words or TF-IDF.

$$\text{Similarity}(t_i|S) = \frac{\mathbf{v}_S \cdot \mathbf{v}_{t_i}}{\|\mathbf{v}_S\| \|\mathbf{v}_{t_i}\|}$$

The tokenization and scoring allow the model to identify which individual words within the bigram keyphrases are most important, aiding in selecting the best triggers.

4. Trigger Classification (Bigram Keyphrase/Token)

The foundation of our method, KeyMinES, is based on using triggers to classify and represent phrases effectively. As defined in Automatic Content Extraction [32] guideline, a trigger is a word or phrase within an event's scope that distinctly conveys the event's occurrence. It is the most representative element that signifies the event's presence within its context, typically a sentence (e.g., Example Tweet # *"**Emergency** crews **rushed** to **rescue** **victims** **trapped** under **collapsed** **buildings**, while the **intensifying** **storm** **continued** to **flood** **streets** and **homes**, leaving many **devastated** and **stranded**"* comprises multiple triggers as shown in Table [I]). The selection of a trigger is guided by the word that most clearly and unambiguously reflects the event being discussed. In this study, we categorize triggers into four types, as shown in Table [I], reflecting the components necessary for constructing grammatically correct English sentences [33].

For KeyMinES trigger classification, each bigram keyphrase P_b and token t_i is classified using the spaCy[6] to get the part-of-speech (POS) tags for each trigger. We selected spaCy over other POS taggers for trigger classification due to its precision, pace, and ease of use, which are crucial for handling large volumes of data streams like those from X [35], which is essential for the large influx of data streams like X data.

The classification of each P_b or t_i into a trigger class C_j is mapped based on its POS:

$$C_j = \text{POS}(P_b \text{ or } t_i)$$

Our tokens' classification into trigger classes enables the model to identify the grammatical roles these tokens play in the tweet, which is crucial for subsequent reconstruction and clustering. Classifying tokens into grammatical triggers based on their part of speech allows the model to retain the structural meaning of the sentences, which is critical for accurate sub-event detection and clustering.

5. Union of Candidate bigram keyphrases—k_p and token—t

The union of bigram keyphrases (k_p) and token (t) candidates involves combining the bigram keyphrase and token, which merge the extracted keyphrase (k_p—"deployed crew") and token (t—"ambulance") are merged as "deployed crew ambulance" to form a sub-event. Some of these unions contain redundant tokens, which are then processed to eliminate redundancy (e.g., "deployed ambulance ambulance" will become "deployed ambulance").

If k_p is the keyphrase and t is the token candidate, the union U is represented as:

$$U = k_p \cup t$$

Merging the bigram keyphrases with tokens ensures that the extracted phrases are grammatically coherent and semantically meaningful.

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[6] https://spacy.io/usage/linguistic-features#pos-tagging

TABLE I: Trigger Classes and Descriptions in Disaster Events

s/n	Trigger Class	Definition	Sample Trigger (Tweet #)	Sample Trigger Description (Tweet #)
1.	Main Verb	Indicates the primary action that directly describes the event.	*"rushed," "rescue," "continued," "flood"*	These verbs depict actions related to the disaster, such as *emergency response* and *flooding.*
2.	Noun/Pronoun	Identifies the subject or object involved, especially when referring to the event or its occurrence.	*"crews," "victims," "buildings," "storm," "streets," "homes"*	The nouns identify the entities involved or affected, such as *"crews"* (responders), *"victims"* (those needing help), *"buildings"* and *"homes"* (impacted structures), and *"storm"* (the disaster itself).
3.	Adjective/Past Participle	Describes properties or conditions, particularly when depicting a state resulting from a sub-event.	*"trapped," "collapsed," "devastated," "stranded"*	The adjectives and past participles convey the disaster's impact, describing conditions like people being *"trapped"* or buildings *"collapsed,"* and highlighting the aftermath with terms like *"devastated"* and *"stranded."*
4.	Modifier	Provides extra details about the action or state, highlighting an ongoing event or its result.	*"Emergency," "Intensifying"*	Modifiers like *"Emergency"* specify the type of *"crews,"* emphasizes the storm's intensity, and *"intensifying"* indicates the worsening severity of the storm.

6. Trigger Reconstruction

Our reconstruction rules ensure that the merged phrases are reconstructed based on predefined grammatical rules, such as reversing the order of words in specific phrases, as shown in Algorithm 1. Applying reconstruction rules helps to standardize the combined candidate bigram keyphrase and token (e.g., "deployed crew ambulance" is reconstructed to "ambulance crew deployed"), ensuring grammatical consistency across similar events. Phrase reconstruction ensures that the final phrase conforms to English grammar, which improves its interpretability and alignment with real-world events.

In Algorithm 1, U is the merged phrase, and R is a reconstruction rule, the reconstructed phrase $R(U)$ is represented as:

$$R(U) = \text{Apply Rule}(U)$$

7. Clustering

In Figure 1, our final step involves clustering the reconstructed phrases (sub-events) to quickly identify and categorize various low-level information elements, such as damage reports, rescue needs, or infrastructure status, and prioritize response efforts. Initially, we processed the sub-events using a pre-trained Sentence-Transformer model to generate dense vector embeddings. This model is well-suited for the task as it can create semantically meaningful embeddings from text, which is essential for identifying the underlying structure in social media data. The resulting embeddings, which represent the semantic content of the sub-events, were concatenated to create a comprehensive matrix for clustering.

To effectively cluster similar sub-events or actions based on their embeddings, we employ the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm[7], which is suitable for our disaster-related tweets sub-events clustering task because of its ability to identify clusters with arbitrary shapes without requiring the number of clusters to be specified in advance. Moreover, it is adept at handling noise and outliers commonly found in social media data. In our effort to determine the optimal parameters for DBSCAN, we focus on two key parameters: *eps* (the greatest distance at which two samples are regarded as neighbors) and *min-samples* (the smallest number of samples within a neighborhood needed to establish a core point). In determining the optimal *eps* value, we utilized the k-nearest neighbors algorithm with k set to 385, considering $MinPts = D + 1$, where D represents the dimensionality of the embeddings. We chose D to be 384 to align with the embedding size of the *all-MiniLM-L6-v2* model.

DBSCAN clusters the embedding vectors $\mathbf{v}_i$ of the reconstructed phrases based on density:

$$\text{Cluster}(S) = \{\mathbf{v}_i : \text{Density}(\mathbf{v}_i, \epsilon, \text{MinPts}) \geq \text{threshold}\}$$

We generate a k-distance graph by sorting the distances to the 385th-nearest neighbor for each point. Selecting an appropriate *eps* value involves visually inspecting the 'elbow' point on this graph, which signifies a noticeable change in the slope. This heuristic method helps balance including points in clusters and identifying outliers or noise.

To construct the final cluster, we apply DBSCAN on the sub-event embeddings with the *eps* parameter set to 0.25 (derived from the k-distance graph) and *min-samples* set to 385. The selection of the cosine metric to measure similarity was based on its effectiveness in handling high-dimensional text data. This metric quantifies distances between embeddings using the formula:

$$\text{sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$$

Clustering enables the identification of patterns or sub-events within our dataset, which is essential for deriving actionable insights and improving situational awareness in emergency management. Our clustering method assigned each sub-event to a cluster based on density, while points in low-density regions were classified as outliers or noise (-1). Points that did not belong to any cluster were also identified and examined for unique content.

IV. EXPERIMENTS AND EVALUATION

In this section, we describe the dataset used for experimentation, the baseline methods for comparison, and the evaluation metrics employed to assess the performance of our KeyMinES model.

A. Dataset

For our KeyMinES approach's experimentation and evaluation, we used X (formerly known as Twitter) data from Hurricane Harvey, a Category 4 storm that caused widespread devastation in Texas in 2017[8]. This hurricane is recognized as one of the most destructive disasters in U.S. history, with damages amounting to nearly USD 200 billion [36].

The dataset utilized in this study was obtained from CrisisNLP[9], which provides a large set of tweet IDs for unlabeled tweets in addition to a smaller set of labeled tweets. In our tweet retrieval process, we initially compared the tweet IDs from CrisisNLP with those accessible on Kaggle[10], resulting in the successful retrieval of over 399,000 tweets. We acquired the remaining tweets using Apify[11], using the labeled tweets provided. Subsequently, the unlabeled tweets were merged with the labeled data. Table II provides detailed information about the dataset used in this study.

[7] https://scikit-learn.org/cluster.DBSCAN.html

[8] https://www.weather.gov

[9] https://crisisnlp.qcri.org

[10] https://www.kaggle.com

[11] https://apify.com/

Algorithm 1: Trigger Reconstruction

```
Input: Corpus $D$ containing required fields - $K$ ('Keyphrase
        (Bigram)'), $T_k$ ('Token'), $B$ ('Bigram Trigger Class'), $T_c$
        ('Token Trigger Class')
Output: Lists - $M_P$ (MergedPhrases), $M_T$ (MergedTriggers), $R_P$
        (ReconstructedPhrases), $R_T$ (ReconstructedTriggers)
Initialize empty lists:;
$M_P \leftarrow \emptyset$;
$M_T \leftarrow \emptyset$;
$R_P \leftarrow \emptyset$;
$R_T \leftarrow \emptyset$;
for each tweet $t \in D$ do
    Define variables for merging:;
    $U_1 \leftarrow t[K]$;
    $U_2 \leftarrow t[T_k]$;
    $T_1 \leftarrow t[B]$;
    $T_2 \leftarrow t[T_c]$;
    Merge phrases and triggers:;
    $U \leftarrow U_1 + " " + U_2$
      $U \leftarrow " ".join(\text{UNIQUE}(U.split(" ")))T \leftarrow T_1 + " + " + T_2$
      $T \leftarrow " + ".join(\text{UNIQUE}(T.split(" + ")))$
    Apply reconstruction rules $R$ to merged phrase $U$;
    if $T =$ "Main Verb + Noun/Pronoun" then
        $R(U) \leftarrow \text{REVERSE}(U)$;
        $R(T) \leftarrow$ "Noun/Pronoun + Main Verb";
    else if $T =$ "Adjective/Past Participle + Noun/Pronoun" then
        $R(U) \leftarrow \text{REVERSE}(U)$;
        $R(T) \leftarrow$ "Noun/Pronoun + Adjective/Past Participle";
    end
    else if $T =$ "Modifier + Noun/Pronoun" then
        $R(U) \leftarrow \text{REVERSE}(U)$;
        $R(T) \leftarrow$ "Noun/Pronoun + Modifier";
    end
    else if $T =$ "Noun/Pronoun + Main Verb" then
        $R(U) \leftarrow U\ R(T) \leftarrow T$;
    end
    else if $T =$ "Main Verb + Adjective/Past Participle" then
        $R(U) \leftarrow \text{REVERSE}(U)$;
        $R(T) \leftarrow$ "Adjective/Past Participle + Main Verb";
    end
    else if $T =$ "Modifier + Main Verb" then
        $R(U) \leftarrow \text{REVERSE}(U)$;
        $R(T) \leftarrow$ "Main Verb + Modifier";
    end
    else if $T =$ "Noun/Pronoun + Modifier" then
        $R(U) \leftarrow \text{REVERSE}(U)$;
        $R(T) \leftarrow$ "Modifier + Noun/Pronoun";
    end
    else
        $R(U) \leftarrow U\ R(T) \leftarrow T$
    end
    Append results to respective lists:;
    $M_P \leftarrow M_P \cup \{U\}$;
    $M_T \leftarrow M_T \cup \{T\}$;
    $R_P \leftarrow R_P \cup \{R(U)\}$;
    $R_T \leftarrow R_T \cup \{R(T)\}$;
end
return $M_P$, $M_T$, $R_P$, $R_T$;
```

Label \ Dataset Size	Original Size	Selected Size	Experimental Size
Labeled (Info./Not Info.)	4,000	4,000	4,000
Unlabeled	6,664,349	4,600,000	795,461

TABLE II: Dataset description

B. Baseline Approaches

We evaluated our method against the sub-event detection techniques that Rudra et al. [7] and Arachie et al. [8] proposed. [7] used a noun-verb combination to identify sub-events and applied a distinctive ranking approach based on the overlap score. Additionally, they introduced a reduction factor to lessen the influence of less frequent sub-events, yielding superior results compared to baseline methods. [8] expanded on the noun-verb pair approach by integrating phrase extraction and employing a crisis ontology for sub-event ranking. Our method was compared to their most effective model, integrating noun-verb pairs, extracted phrases, and ranking using the MOAC ontology.

To ensure consistency, we adopted similar pre-processing steps as the baseline studies. However, instead of relying on POS tags for sub-event extraction, we utilized bigram and token-level similarity computations to pinpoint the most relevant sub-events, providing a more refined approach than those of [7] and [8].

C. Evaluation Metric

In evaluating the effectiveness of our approach compared to the baseline for different top-k values,we generate an ordered list of sub-events, selecting the top-k sub-events (represented by bigrams) to verify their occurrence in the labeled tweets. Precision and recall are assessed across varying values of k. Additionally, F1 scores are provided at various k values, and the Receiver Operating Characteristic (ROC) curve is analyzed.

1) The F1-score effectively balances precision and recall by measuring both the accuracy and completeness of identifying informative tweets and the capability to capture all relevant tweets from the dataset.
    - Precision denotes the proportion of correctly identified informative tweets from the total retrieved tweets.
    - Recall is the proportion of correctly identified informative tweets relative to the dataset's total number of informative tweets.
2) The ROC curve evaluates the model's ability to distinguish between informative and non-informative disaster-related tweets.
    - A true positive is defined as a tweet containing at least one top-k sub-event (bigram+token), whereas a tweet lacking such sub-events is considered a false negative.

V. Result

The effectiveness of the sub-event categorization scheme hinges on two factors: firstly, the effectiveness of detecting sub-events within the data, and secondly, the relevance of the categories in reflecting the event categories found in the dataset. We evaluate these factors through: A) quantitative and B) qualitative assessments.

A. Quantitative Evaluation

In our quantitative evaluation, we assess the effectiveness of our method by comparing it to baseline approaches. Our focus lies in evaluating the retrieval of informative tweets from annotated datasets by the top-ranked sub-events. This evaluation encompasses two aspects: i) unfiltered and ii) filtered.

i) Unfiltered Sub-events

To ensure a fair comparison with [7] and [8], we utilized a dataset of 795,461 distinct Hurricane Harvey tweets, presented in Table 1. Our bigram keyphrases correspond to the noun-verb pairs used by [7] and [8], while our reconstructed phrases (candidate sub-events) align with the extracted candidate sub-events in these studies.

Table III compares the performance of our approach (KeyMinES) with the noun-verb pair baseline regarding bigram phrases. Furthermore, we assess the effectiveness of our approach in identifying candidate sub-events relative to the baselines. KeyMinES identified

TABLE III: Bigram Keyphrase and Candidate Sub-event Identification Performance. Our bigram Keyphrase is equivalent to noun-verb pair in [7] and [8]. Our (bigram+token) is referred to as the candidate sub-event, which is similar to the Noun-Verb pair in [7] and MOAC+NV+Phrases in [8]'s work.

Models	**Bigram Keyphrase**	**Candidate Sub-event**
Rudra et al. [7]	769,670	769,670
Arachie et al. [8]	769,670	796,792
KeyMinES (Ours)	791,482	840,825

791,482 unique bigram phrases, outperforming [7] and [8]. Additionally, KeyMinES identified 840,825 candidate sub-events, surpassing the 796,792 sub-events found by [8] and the 769,670 identified by [7].

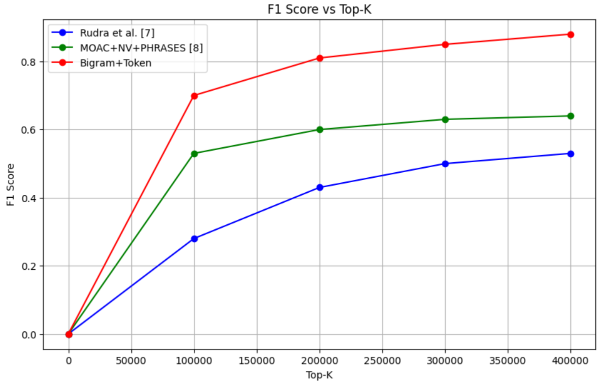


Fig. 2: Comparison of F1 Score vs Top-K Sub-events using our method (bigram+token) with Rudra et al. [7] and Arachie et al. [8] on the unfiltered dataset.

The performance of our method (bigram+token) was evaluated at various top-k values (ranging from 50K to 400K) using the F1 score, as shown in Figure 2. Our results consistently show that bigram+token outperforms the approaches of [7] and [8] across all top-k sub-event levels. This indicates that our method is more effective in accurately identifying and classifying sub-events within disaster-related tweets. Additionally, the F1-score improves as the number of top-k sub-events increases, suggesting that extracting more high-quality sub-events generally enhances performance. However, the rate of improvement slows once the number exceeds 200,000.

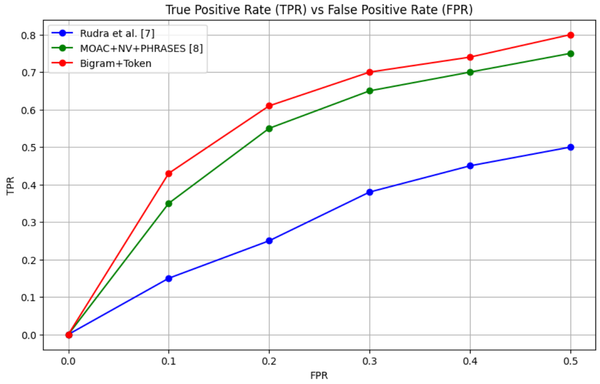


Fig. 3: Performance of bigram+token in TPR across all FPR levels for identifying disaster-related sub-events.

In Figure 3, the ROC curve illustrates the True Positive Rate (TPR) against the False Positive Rate (FPR) for bigram+token and the baseline methods (Rudra et al. 2018; Arachie et al., 2020) for the Hurricane Harvey dataset. bigram+token consistently achieves a higher TPR across all FPR levels, demonstrating its ability to identify relevant sub-events while minimizing false positives. This highlights the effectiveness of our method in accurately capturing disaster-related information from tweets, outperforming the baseline methods. The ability of bigram+token to maintain a high TPR while controlling FPR is critical for avoiding false alarms that could overwhelm responders and hinder disaster response efforts. These findings underscore that advanced models like bigram+token can significantly improve the efficiency of automated disaster response tools compared to other approaches.

ii) Filtered Sub-events

In Figure 4, we compare our approach with baseline methods that employ a filtered approach, restricting consideration only to bigrams (comparable to noun-verb pairs in the baseline) that appear more than once in the tweet corpus. After filtering, our method reduced the number of bigrams in Table IV 3 to 15,892, while bigram+token identified a total of 72,520 sub-events. This performance significantly exceeded that of [8], identifying 3,187 noun-verb pairs and 30,309 sub-events. For evaluation and comparison with the baseline, we focused on the top 30 K sub-events, as illustrated in Figure 3. The results consistently show that our filtering approach outperforms [8] across all top-k values, notably achieving a 20% improvement in the F1-score at the top-30K level.

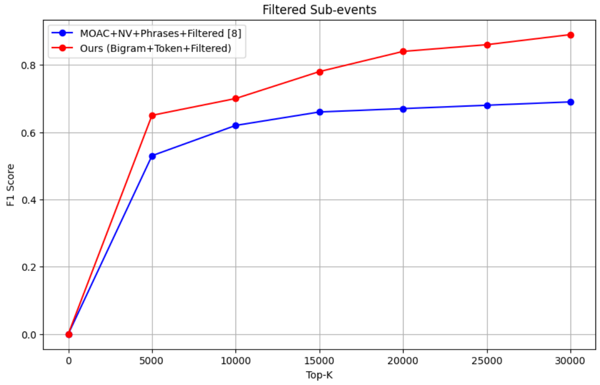


Fig. 4: F1 Score comparison of our bigram+token filtering approach with the baseline across all top-k values.

TABLE IV: Filtered Bigram Phrase and Sub-event Detection Comparison. Our Bigram Phrase (filtered) in this case is equivalent to the filtered noun-verb pair in [8]. Our (bigram+token+filtered) is referred to as the candidate sub-event, which is similar to MOAC+NV+Phrases+filtered in [8]'s work.

Models	**Bigram Phrase**	**Candidate Sub-event**
Arachie et al. [8]	3,187	30,309
KeyMinES (Ours)	15,892	72,520

B. *Qualitative Evaluation*

In addition to the quantitative analysis, we evaluated our methods qualitatively, concentrating on four key aspects: 1) sub-event identification, 2) criticality and time-sensitivity, 3) immediacy and context, and 4) cluster categories.

*1) **Sub-Event Identification**:* [8] emphasizes the importance of effective sub-event identification methods that capture significant and diverse sub-events. In our study, we evaluated the efficacy of our approach (bigram+token) by comparing the top 10 ranked sub-events with those identified by [7] and [8] using a crisis ontology for ranking. Employing Arachie et al.'s filtering approach, we selected unique sub-events based on their similarity scores and ranked them in descending order.

Table V displays the comparison, revealing that our approach (bigram+token) identified a broader range of essential sub-events compared to the baselines. While [7] and [8] captured some relevant sub-events such as "people lost" and "Coldwell impacted" in [7] and "road blocked" and "hundreds trapped" in [8], our method extracted key sub-events that better represent urgent and significant information. For instance, the sub-event "victims buzzfeed" identified by our approach did not reflect critical incidents. The sub-events identified through our method offer actionable insights for first responders, providing information on the most demanding needs during a crisis.

TABLE V: Comparison of Top 10 Sub-events Identified by Our Approach (bigram+token) and Baseline Methods. [8] approach in this Table represents MOAC-NV+Phrases+filtered.

Rudra et al. [7]	**Arachie et al. [8]**	**2-word (bigram+token)**	**3-word (bigram+token)**
foxnews flooding	feeding centers	children drowned	ambulance crew deployed
victims buzzfeed	road blocked	residents trapped	urgent blood donations
flotus donated	shortage fuel	building collapsed	ill babies safety
redcross serving	price gouging	floodwater rising	devastating hurricane flooding
spca need	hundreds trapped	emergency response	apocalypse weather warning
mullins flooding	shelter supplies	evacuations help	clinics closed today
coldwell impacted	drug shortage	ems care	emergency animal shelters
sentedcruz impacted	infectious disease	firefighter paramedic	relief aid support
peoples lost	medical equipment	flood warning	needs food donations
hurr impacted	water contamination	power blackout	apartment roof damage

2) Criticality and Time-Sensitivity in Disaster Response: We posit that adequate disaster information must be concise, specific, and capable of providing clear indicators of critical situations and urgent needs to enable a timely response, even though only some tweets contain relevant sub-events. To evaluate this, we classified the top 10 ranked sub-events into four levels: (i) most critical and time-sensitive, (ii) moderately critical but important, (iii) less time-sensitive but necessary for recovery, and (iv) not critical or time-sensitive.

i) Table V illustrates that our approach (bigram+token) identifies ***highly critical and time-sensitive sub-events***, such as "children drowned," "residents trapped," "building collapsed," "emergency response," "evacuations help," and "EMS care," which describe life-threatening situations or urgent needs requiring immediate action to prevent loss of life or further harm. Similarly, [8] identifies time-sensitive sub-events like "road blocked" and "shortage fuel," highlighting urgent logistical issues crucial for facilitating rescue efforts and supply distribution.

ii) For ***moderately critical but important sub-events***, our approach (bigram+token) includes phrases such as "floodwater rising," "flood warning," "power blackout," and "firefighter paramedic," signaling conditions that, while not immediately life-threatening, could escalate without proactive intervention. In contrast, [8] identify sub-events like "feeding centers," "shelter supplies," and "drug shortage," reflecting ongoing needs critical for sustained disaster response but not requiring the same urgency as life-saving actions.

iii) In the category of ***less time-sensitive but necessary*** for recovery, [8] include phrases such as "price gouging" and "water contamination," which, while not urgent, remain critical for long-term recovery efforts once immediate threats are managed.

iv) Lastly, Table V also presents sub-events that are ***neither critical nor time-sensitive***, such as "foxnews flooding," "victims buzzfeed," and "flotus donated" from [7]. These sub-events provide context or reference peripheral issues but do not contribute directly to high-priority, immediate disaster response efforts.

3) Immediacy VS Context: The successful dissemination of information during a disaster relies on three critical factors: i) providing clear and specific details to facilitate immediate action [37] ii) understanding the context of the situation [38] iii) effectively communicating necessary information while balancing quick decision-making with strategic planning needs [39].

Our evaluation of the bigram+token approach reveals that it produces two-word sub-events when the token is part of the bigram (e.g., "ambulance deployed") and three-word sub-events when the token differs from the bigram (e.g., "ambulance crew deployed"). Phrases in Table V illustrates that two-word sub-events convey more direct actions or situations, while three-word sub-events provide additional context or details about the event.

i) Clarity and Specificity: are crucial for enabling rapid decision-making amid a disaster. Our findings from Table V indicate that two-word sub-events are more concise, offering clear and immediate indicators of situations or required actions (e.g., "children drowned" or "building collapsed"). This conciseness makes them particularly useful for facilitating quick understanding and response planning. On the other hand, three-word sub-events offer more detail and context, although this may sometimes reduce the sense of urgency. For example, "devastating hurricane flooding" provides more context than "floodwater rising," helping responders better assess the event type and severity.

ii) Contextual Understanding: is essential for fully comprehending the scope and impact of a disaster. In general, three-word sub-events provide greater context, offering more detailed insights into the nature of the situation, as demonstrated in Table V. For instance, "ambulance crew deployed" signifies an emergency response and specifies the service type being used. While two-word sub-events are more direct, they may sometimes lack the contextual richness that can aid in comprehensive planning and coordination (e.g., "flood warning" vs. "devastating hurricane flooding").

iii) Effectiveness in Communication: Effective communication is crucial for a prompt disaster response. Our observations from Table V suggest that three-word sub-events are more effective when detailed communication is feasible, providing responders with a better understanding of the situation and the underlying factors. However, two-word sub-events are ideal for rapid communication and decision-making, especially when speed is paramount, such as during the immediate phases of disaster response.

4) Assessment of Cluster Quality: In the final phase of our approach, as depicted in Figure 1, we employ DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to cluster similar candidate sub-events into distinct groups. Unlike methods requiring a predefined number of clusters, DBSCAN forms clusters based on data point density, determined by the parameters ϵ and MinPts discussed in Section 3. This results in 79 clusters, excluding noisy data.

To assess the quality of these clusters, we utilized two resources: 1) ChatGPT and 2) human assessors. Our human evaluators, comprising three students, and GPT-4 (ChatGPT) were employed to ensure quality and accuracy. Each cluster was assessed by randomly sampling five tweets and their corresponding sub-events from the 79 clusters. Initially, ChatGPT was prompted to label the clusters based on thematic similarities among sub-events. Subsequently, human assessors rated the clusters using a three-point scale—agree (≥ 0.7), partially agree (0.4–0.7), or disagree (< 0.4)—based on whether the assigned label accurately described the cohesiveness and homogeneity of the sampled sub-events. The inter-rater reliability was measured using Fleiss' Kappa [40] due to its simplicity and effectiveness, yielding a kappa value of 0.92, indicating strong agreement among raters.

TABLE VI: Cluster Label and Sub-events with Human Validation

Cluster ID	**GPT Cluster Label**	**Sub-events**	**Human Validation**
21	First Responder Support	"ambulance crew deployed", "rescue responders", "help needed"	Agree
49	Emergency Evacuations	"evacuate early", "evacuation warnings", "evacuation mandatory"	Agree
3	Public Safety	"stay safe", "texas safe", "public safety concerns"	Agree
65	Medical Emergencies	"hospital evacuees", "medical care", "sick babies evacuated"	Agree
0	Entertainment Gossip	"swift taylor", "kanye say", "celebrity news"	Partial Agree

Table VI presents the results from five randomly selected clusters, demonstrating that KeyMinES clusters consistently generated cohesive and homogeneous groups of disaster-related tweets. Additionally, the presence of noise or outliers, which were segregated from meaningful clusters, was observed. For instance, in cluster 0 (Table VI), labeled as "Partial Agree," assessors noted that while the cluster was generally relevant, sub-events such as "swift taylor" and "kanye," though related to entertainment, lacked clear relevance to gossip, indicating weaker ties within the cluster. This showcases the effectiveness of DBSCAN in separating noise from true sub-event clusters.

C. Ablation Study

The ablation study, depicted in Figure 5, assesses our approach's performance using different modules—bigram, token, and bigram+token—across various top-k sub-event values, with the F1 score as the evaluation metric. The results reveal that the bigram+token model consistently outperforms the individual bigram and token models, achieving an F1 score of nearly 90% for the top 400,000 sub-events. This improvement emphasizes the effectiveness of combining bigrams and tokens to accurately capture keyphrases in disaster-related tweets.

The superior performance of the bigram+token module reflects its ability to merge complementary information from bigrams and tokens, leading to a more nuanced understanding of tweet context. bigrams capture relationships between word pairs, such as "ambulance crew" or "texas emergency," essential for identifying main themes in disaster tweets. However, bigrams alone sometimes lack the granularity provided by tokens, which can refine the extraction of individual actions or entities, allowing for a more robust detection of sub-events. In contrast, the bigram module outperforms the token module but falls short of the bigram+token model. While bigrams identify essential phrases, they may overlook the finer details captured by tokens. Tokens, representing individual words, exhibit the lowest performance when used in isolation, as they often require more context provided by bigrams. Although tokens are valuable for identifying specific entities or actions, they do not fully capture the relational dynamics found in phrases like "crew deployed."

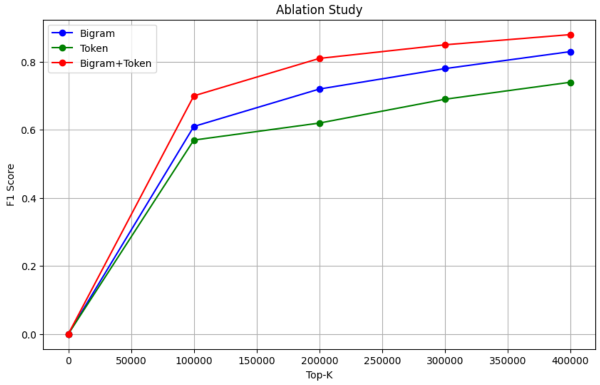


Fig. 5: Ablation study comparing F1 scores for Bigram, Token, and bigram+token models across different Top-K sub-event values.

This study highlights the significant advantage of the bigram+token approach, particularly as the number of sub-events increases. At the 200,000 sub-event mark, bigram+token surpasses the bigram model by approximately 10% in F1 score, demonstrating its critical role in disaster management tasks. By leveraging both context and granularity, this approach ensures accurate sub-event identification even within large datasets, making it highly effective for managing noisy tweet data during disaster events.

VI. Discussion

The study focuses on the challenges of identifying sub-events for situational awareness in datasets, using Hurricane Harvey as a case study. It emphasizes the potential of AI for social good, especially in disaster management, where accurate sub-event identification can improve real-time decision-making.

An important observation is the trade-off between immediacy and contextual detail. As depicted in Table V, two-word sub-events are more effective for urgent and actionable situations, providing concise and direct information about immediate needs. On the other hand, three-word sub-events offer a more comprehensive context, which is particularly useful in disaster response planning and coordination phases, aiding strategic decision-making.

The accuracy of a sub-event classification and phrase reconstruction is crucial for maintaining grammatical coherence and semantic clarity. Proper trigger identification ensures that extracted keyphrases retain their intended meaning. For instance, "shortage fuel" in Arachie's model in Table V disrupts the natural word order, diminishing clarity and intelligibility. Ensuring proper reconstruction (fuel shortage) is vital to maintaining context relevance, especially in critical settings like emergency response.

The approach is beneficial for various stakeholders as it eliminates the need for resource-intensive supervised learning methods that require labeled data and significant training costs. It can be applied in multiple disaster management phases, such as planning, mitigation, response, and recovery, delivering insights without the operational complexities of traditional AI models.

Our sub-event detection and classification approach (KeyMinES) has diverse applications in and outside of disaster management, offering advantages to various stakeholders, such as:

- **Emergency Responders:** enables the identification of critical sub-events for prompt response and efficient resource allocation during disasters.
- **Crisis Management Teams:** analyzing social media data enhances situational awareness and real-time decision-making.
- **Government Agencies:** provides actionable insights from disaster-related social media content, assisting public safety initiatives and policy planning.
- **Humanitarian Organizations:** aids in identifying needs and assessing damages to effectively coordinate relief efforts.
- **Marketing and Advertising Teams:** facilitates the extraction of critical insights from social media data to understand customer sentiments and trends.
- **Journalists and Media Organizations:** assists in tracking trending topics and significant events for accurate and timely reporting.
- **Brand Managers:** supports monitoring brand mentions and public relations by detecting sub-events related to brand reputation management.

VII. Conclusion and Future Work

This paper introduces KeyMinES, an unsupervised model developed to extract minimal keyphrases from social media data to identify and classify sub-events in disaster scenarios. The approach combines bigram and token-level representations, significantly improving the detection and understanding of critical sub-events. By clustering these sub-events, actionable insights are provided to support decision-makers during emergencies. Our evaluations, comprising quantitative and qualitative analyses, show that KeyMinES outperforms baseline methods in accuracy and scalability. These results underscore the model's potential to enhance situational awareness and improve the efficiency of disaster response efforts for various stakeholders, including emergency responders and crisis management teams.

Future research will focus on sourcing additional datasets for comparison with supervised and semi-supervised approaches. We also plan to detect and classify the urgency level of extracted sub-events based on tense and aspect-level information, enabling more precise prioritization during disaster responses. Furthermore, we aim to identify appropriate first responders by extracting their arguments and roles in various emergency scenarios, ensuring that the appropriate teams are dispatched for each specific situation. This will further enhance the model's ability to provide actionable insights and improve the overall effectiveness of disaster management initiatives.

Acknowledgement

Our gratitude goes to the assessors who contributed to validating the models' output. This research received support from the Kummer Institute for Student Success, Research, and Economic Development at Missouri University of Science and Technology through the Kummer Innovation and Entrepreneurship Doctoral Fellowship and the NSF—USA CNS-2219615

References

[1] S. R. Chowdhury, S. Basu, and U. Maulik, "A survey on event and subevent detection from microblog data towards crisis management," *International Journal of Data Science and Analytics*, vol. 14, no. 4, pp. 319-349, 2022.

[2] A. Adams, "Beyond social media push strategies: Incorporating citizen-developed social media pull tactics to supplement disaster communication and response," in *Case Studies in Disaster Response*, Butterworth-Heinemann, 2024, pp. 201-211.

[3] A. Adesokan, S. Madria, and L. Nguyen, "TweetACE: A Fine-grained Classification of Disaster Tweets using Transformer Model," in *2023 IEEE Appl. Imagery Pattern Recognition Workshop*, Sep. 2023, pp. 1-9.

[4] L. Belcastro, F. Marozzo, D. Talia, P. Trunfio, F. Branda, T. Palpanas, and M. Imran, "Using social media for sub-event detection during disasters," *Journal of Big Data*, vol. 8, no. 1, pp. 1-22, 2021.

[5] S. H. Ro, Y. Li, and J. Gong, "A Machine learning approach for Post-Disaster data curation," *Advanced Engineering Informatics*, vol. 60, p. 102427, 2024.

[6] A. Adesokan, S. Madria, and L. Nguyen, "HatEmoTweet: Low-level emotion classifications and spatiotemporal trends of hate and offensive COVID-19 tweets," *Soc. Net Anal. & Min.*, vol. 13, no. 1, p. 136, 2023.

[7] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran, "Identifying sub-events and summarizing disaster-related information from microblogs," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Jun. 2018, pp. 265-274.

[8] C. Arachie, M. Gaur, S. Anzaroot, W. Groves, K. Zhang, and A. Jaimes, "Unsupervised detection of sub-events in large scale disasters," in *Proceedings Of The AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 354-361, Apr. 2020.

[9] Y. Cao, H. Peng, Z. Yu, and S. Y. Philip, "Hierarchical and incremental structural entropy minimization for unsupervised social event detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 8255-8264, Mar. 2024.

[10] S. R. Chowdhury, S. Basu, and U. Maulik, "Disastrous event and sub-event detection from microblog posts using bi-clustering method," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 161-170, 2022.

[11] A. Edouard, "Event detection and analysis on short text messages," Ph.D. dissertation, Université Côte D'Azur, 2017.

[12] X. Sun, "Event Detection and Evolution in Online Social Networks," Ph.D. dissertation, University of Leicester, 2024.

[13] H. P. Zou, C. Caragea, Y. Zhou, and D. Caragea, "Semi-supervised few-shot learning for fine-grained disaster tweet classification," in *Proceedings of the 20th International ISCRAM Conference*, May 2023.

[14] J. Wang and K. Wang, "Bert-based semi-supervised domain adaptation for disastrous classification," *Multimedia Systems*, vol. 28, no. 6, pp. 2237-2246, 2022.

[15] I. Sirbu, T. Sosea, C. Caragea, D. Caragea, and T. Rebedea, "Multimodal semi-supervised learning for disaster tweet classification," in *Proceedings of the 29th International Conference on Computational Linguistics*, Oct. 2022, pp. 2711-2723.

[16] A. Dahou, A. Mabrouk, A. A. Ewees, M. A. Gaheen, and M. Abd Elaziz, "A social media event detection framework based on transformers and swarm optimization for public notification of crises and emergency management," *Technological Forecasting and Social Change*, vol. 192, p. 122546, 2023.

[17] A. Balali, M. Asadpour, and S. H. Jafari, "COfEE: A comprehensive ontology for event extraction from text," *Computer Speech & Language*, vol. 101702, 2024.

[18] A. Dhiman and D. Toshniwal, "An approximate model for event detection from twitter data," *IEEE Access*, vol. 8, pp. 122168-122184, 2020.

[19] A. Adesokan and S. Madria, "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)*, Dec. 2023, pp. 578-587.

[20] P. K. Garg, R. Chakraborty, and S. K. Dandapat, "OntoDSumm: ontology-based tweet summarization for disaster events," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 2724-2739, 2023.

[21] H. Hettiarachchi, M. Adedoyin-Olowe, J. Bhogal, and M. M. Gaber, "WhatsUp: An event resolution approach for co-occurring events in social media," *Information Sciences*, vol. 625, pp. 553-577, 2023.

[22] R. Kumar, R. Sinha, S. Saha, and A. Jatowt, "Extracting the Full Story: A Multimodal Approach and Dataset to Crisis Summarization in Tweets," *IEEE Transactions on Computational Social Systems*, 2024.

[23] M. C. Mason, S. Iacuzzi, G. Zamparo, and A. Garlatti, "How do stakeholders co-create value in a service ecosystem? Insight from mega-events," *Management Decision*, vol. 62, no. 13, pp. 398-425, 2024.

[24] Z. Rezaei, B. Eslami, M. A. Amini, and M. Eslami, "Event detection in twitter by deep learning classification and multi label clustering virtual backbone formation," *Evolutionary Intelligence*, vol. 16, no. 3, pp. 833-847, 2023.

[25] N. Noor, R. Okhai, T. B. Jamal, N. Kapucu, Y. G. Ge, and S. Hasan, "Social-media-based crisis communication: Assessing the engagement of local agencies in Twitter during Hurricane Irma," *International Journal of Information Management Data Insights*, vol. 4, no. 2, p. 100236, 2024.

[26] P. Zhang, H. Zhang, and F. Kong, "Research on online public opinion in the investigation of the "7–20" extraordinary rainstorm and flooding disaster in Zhengzhou, China," *International Journal of Disaster Risk Reduction*, vol. 105, p. 104422, 2024.

[27] Z. Zhou, X. Zhou, Y. Chen, and H. Qi, "Evolution of online public opinions on major accidents: Implications for post-accident response based on social media network," *Expert Systems with Applications*, vol. 235, p. 121307, 2024.

[28] P. Režek and B. Žvanut, "Towards optimal decision making in mass casualty incidents management through ICT: A systematic review," *International Journal of Disaster Risk Reduction*, p. 104281, 2024.

[29] H. Riddell, "Investigating social media users' preferences of content and sourcing during a crisis," *Communication and the Public*, p. 20570473241254175, 2024.

[30] E. S. Apostol, C. O. Truică, and A. Paschke, "ContCommRTD: A distributed content-based misinformation-aware community detection system for real-time disaster reporting," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[31] F. Vitiugin and H. Purohit, "Multilingual Serviceability Model for Detecting and Ranking Help Requests on Social Media during Disasters," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, pp. 1571-1584, May 2024.

[32] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The automatic content extraction (ACE) program-tasks, data, and evaluation," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004, vol. 2, no. 1, pp. 837-840.

[33] S. Egamov and V. Karimova, "Parts of speech and sentence structure in English grammar," *Galaxy International Interdisciplinary Research Journal*, vol. 10, no. 7, pp. 156-160, 2022.

[34] M. Brysbaert, P. Mandera, and E. Keuleers, "The word frequency effect in word processing: An updated review," *Current Directions in Psychological Science*, vol. 27, no. 1, pp. 45-50, 2018.

[35] R. P. Nair and M. G. Thushara, "Investigating Natural Language Techniques for Accurate Noun and Verb Extraction," *Procedia Computer Science*, vol. 235, pp. 2876-2885, 2024.

[36] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *Proceedings of the international AAAI conference on web and social media*, vol. 12, no. 1, Jun. 2018.

[37] I. Dallo, M. Stauffacher, and M. Marti, "Actionable and understandable? Evidence-based recommendations for the design of (multi-) hazard warning messages," *International Journal of Disaster Risk Reduction*, vol. 74, p. 102917, 2022.

[38] N. Ashish, R. Eguchi, R. Hegde, C. Huyck, D. Kalashnikov, S. Mehrotra, P. Smyth, and N. Venkatasubramanian, "Situational awareness technologies for disaster response," in *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*, pp. 517-544, 2008.

[39] Z. Ivetić, J. Tošić, and L. Miletić, "Successful communications as an element of effective management of emergency situations," 2023.

[40] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, 1971, doi: 10.1037/h0031619.