Utilizing LLMs and ML Algorithms in Disaster-Related Social Media Content

Vasileios Linardos , Maria Drakaki * , Panagiotis Tzionas

Posted Date: 23 May 2025

doi: 10.20944/preprints202505.1798.v1

Keywords: large language models; disaster management; social media analysis; tweet classification; natural language processing; emergency response; clustering; machine learning; sentiment analysis; topic modeling

 doi:10.20944/preprints202505.1798.v1

Article

Utilizing LLMs and ML Algorithms in Disaster-Related Social Media Content

Vasileios Linardos [1], Maria Drakaki [2,*] and Panagiotis Tzionas [3]

[1] Department of Science and Technology, University Center of International Programmes of Studies, International Hellenic University, 14th Km Thessaloniki-N. Moudania, GR-57001, Greece; vlinardos@ihu.edu.gr

[2] Department of Science and Technology, University Center of International Programmes of Studies, International Hellenic University, 14th Km Thessaloniki-N. Moudania, GR-57001, Greece

[3] Department of Industrial Engineering and Management, P.O. Box 141, GR-57400, Greece, ptzionas@ihu.gr

* Correspondence: mdrakaki@ihu.gr

Abstract: In this research, we explore the use of Large Language Models (LLMs) and clustering techniques to automate the structuring and labeling of disaster-related social media content. With a gathered dataset comprising millions of tweets related to various disasters, our approach aims to transform unstructured and unlabeled data into a structured and labeled format that can be readily used for training machine learning algorithms and enhancing disaster response efforts. We leverage LLMs to preprocess and understand the semantic content of the tweets, applying several semantic properties to the data, followed by the application of clustering techniques to identify emerging themes and patterns that may not be captured by predefined categories and are surfaced through topic extraction of the clusters. We proceed with manual labeling and evaluation of 10,000 examples to evaluate the LLMs' ability to understand tweet features. Our methodology is applied to real-world data for disaster events, with results directly applicable to actual crisis situations.

Keywords: large language models; disaster management; social media analysis; tweet classification; natural language processing; emergency response; clustering; machine learning; sentiment analysis; topic modeling

1. Introduction

Social media has become an invaluable source of information during disaster events. With millions of users sharing real-time updates, images, and videos, social media platforms offer a wealth of data that can be leveraged for various disaster management tasks. The increasing frequency and severity of natural disasters, due to climate change, necessitates innovative approaches to enhance preparedness, response, and recovery efforts. According to the EM-DAT International Disaster Database, both the number and intensity of disaster events have been steadily increasing globally over the past decades, with significant economic losses and human casualties [1]. This upward trend highlights the urgent need for more effective disaster management tools and techniques.

The proliferation of social media platforms has created unprecedented opportunities for gathering real-time information during crisis events [2]. Citizens often become "human sensors," providing first-hand accounts and visual documentation of disaster impacts before official assessments can be conducted. However, the sheer volume and unstructured nature of social media data pose significant challenges for emergency responders and disaster management agencies trying to extract actionable insights during time-sensitive situations. The development of structured ontologies for crisis management, as reviewed by Liu et al. [3], provides a foundation for organizing this complex information ecosystem.

This paper examines the potential of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) in disaster data management by applying such models to social media data,

 doi:10.20944/preprints202505.1798.v1

creating a mechanism to produce structured data for disaster analysis through social media. We also present a comprehensive review of existing applications and studies of LLMs in the field, contributing to the growing body of literature on AI-assisted disaster management.

2. Literature Review

2.1. Utilization of the Social Media Datasets for Disaster Management

Social media data has emerged as a crucial resource for researchers and practitioners in the field of disaster management, offering opportunities to enhance various aspects of disaster response and mitigation [4]. Derived primarily from platforms such as X (formerly Twitter), Facebook, and Instagram, these datasets capture real-time information shared by individuals experiencing crises, providing valuable insights into the evolving dynamics of disaster events.

The applications of social media data in disaster management are multifaceted and span the entire disaster lifecycle. For situational awareness, researchers have leveraged social media data to track the spread of wildfires [5], monitor flood events [6], and assess damage from earthquakes [7]. The inherent real-time nature of social media updates provides a distinct advantage for gaining rapid understanding during rapidly unfolding disasters. Similarly, needs assessment benefits from social media analysis, enabling the identification of urgent requirements for shelter, food, and medical assistance, particularly after events like hurricanes [8], floods [9] and wildfires [10]. By analyzing the content of social media posts, disaster relief organizations can gain a nuanced understanding of the specific needs of affected communities, allowing for a more targeted and effective response.

Crisis communication is another critical area where social media plays a pivotal role. Studies have investigated the use of social media for disseminating warnings, coordinating evacuations, and providing critical public information during disasters [4,9,11]. These platforms serve as effective channels for reaching a broad audience swiftly and facilitating two-way communication between authorities and the public, fostering a more informed and responsive environment. Researchers have developed methods to assess damage to infrastructure and buildings using images and videos [12], proving particularly useful when physical access to affected areas is restricted or traditional assessment methods are time-consuming.

Furthermore, social media data is increasingly being utilized for predictive modeling. Researchers are developing models to predict the spread of wildfires [13] and forecast flood events [6] by incorporating real-time information from individuals acting as "human sensors." This integration of citizen-generated data has the potential to significantly improve the accuracy of disaster predictions, enabling earlier warnings and more proactive mitigation strategies. This novel application demonstrates the potential of advanced AI in leveraging social media data for rapid and accurate impact assessment. Social media data is also used to assess mental health needs of those impacted by disasters, such as analyzing twitter posts after wildfires [14]. Recent academic initiatives, such as in [15], have focused on developing specialized algorithms for detecting natural disasters from social media signals in real-time, further expanding the utility of these platforms. Social media analysis using advanced AI methods can enhance situational awareness in disaster management and assess the impacts of disasters [16].

Social media datasets for disaster management research vary in several key ways. These include the platform used (most commonly Twitter, but also Facebook, Instagram, and others), the type of disaster (earthquakes, floods, etc.), the time period covered (single event or longer-term), the geographic area (local to global), whether the data is labeled (for training machine learning models), and the specific data points extracted (like hashtags, location, and time), which are used to understand disaster dynamics.

Several notable social media datasets have been developed for disaster management research. These include:

- **CrisisNLP**: CrisisNLP provides resources for crisis informatics research, including annotated datasets of tweets and images from disasters, labeled for various attributes. It also offers

 doi:10.20944/preprints202505.1798.v1

tools for tweet downloading, pre-trained models, benchmarked datasets for classification, and large COVID-19 tweet datasets, all aimed at developing computational tools for humanitarian aid [17].

- **HumAID**: A dataset of human-annotated disaster incidents from Twitter, covering 19 major natural disasters from 2016 to 2019. This dataset focuses specifically on identifying and classifying different types of disaster incidents, providing valuable training data for machine learning models used in emergency response [18].
- **CrisisBench**: A consolidated dataset combining eight publicly available disaster-related datasets, providing over 166,000 tweets for informativeness classification and over 141,000 tweets for humanitarian classification tasks. By consolidating multiple datasets, CrisisBench offers a larger and more diverse dataset for training and evaluating machine learning models in disaster management [19].
- **GeoCoV19**: A dataset of over 500 million multilingual tweets related to the COVID-19 pandemic, spanning 218 countries and 47,000 cities. This dataset captures the global impact of the pandemic and provides valuable insights into how social media is used during public health emergencies [20].
- **TBCOV**: A dataset comprising over two billion multilingual tweets related to the COVID-19 pandemic, with sentiment, named entities, geo, and gender labels. The inclusion of these labels allows for a more nuanced analysis of social media content and enables researchers to study the social and emotional impact of the pandemic [21].

2.2. Challenges and Limitations

While social media data holds immense promise for improving how we manage and respond to disasters, it's important to acknowledge that there are several significant challenges to overcome. These challenges range from questions about the reliability of the information we find to ethical considerations about how we use people's personal data. Effectively leveraging social media in disaster situations requires careful thought and strategies to address these limitations.

One major concern revolves around the trustworthiness of social media content. The very nature of these platforms means that false or inaccurate information can spread rapidly, especially during times of crisis. This misinformation can create confusion, hinder rescue efforts, and even put people at further risk [22]. Therefore, it's crucial to develop robust methods for verifying the information we gather from social media. This involves finding ways to filter out unreliable reports, identify trustworthy sources, and compare information with other reliable data. Recent hybrid approaches combining machine learning with rule-based classification have shown promise in extracting actionable emergency information from social media streams while mitigating issues of misinformation [23].

Another key limitation is that social media users don't necessarily reflect the entire population. Certain groups of people might be less likely to use social media due to factors like age, access to technology, or language barriers [24]. This means that relying solely on social media data could give us an incomplete or skewed picture of the disaster's impact. For instance, we might miss the needs of vulnerable populations who are not active online. To ensure our disaster response is fair and effective, we need to be aware of these biases and find ways to gather information from a wider range of people.

Furthermore, using social media data raises important ethical questions about privacy. We must be extremely careful when handling personal information shared on these platforms and always respect people's privacy [25]. This means following strict ethical guidelines and legal regulations. Whenever possible, we should seek consent from users and use techniques to anonymize data so that individuals cannot be easily identified. We must also be mindful of the potential for even anonymized data to be re-identified and take appropriate precautions.

Finally, the sheer volume of social media data generated during a disaster can be overwhelming. The speed at which this data is produced also presents a significant technical challenge [26]. Traditional methods of processing data often struggle to keep up with this influx. Therefore, we need

 doi:10.20944/preprints202505.1798.v1

to develop more efficient and scalable ways to collect, store, and process this information. This includes creating sophisticated computer programs that can automatically filter relevant data, identify key topics, and understand the overall sentiment expressed in social media posts. By overcoming these technical hurdles, we can extract timely and valuable insights to support effective disaster response efforts.

2.3. Utilization of LLMs and GenAI

Large Language Models (LLMs) are sophisticated AI models built upon deep learning architectures, primarily the Transformer architecture, which enables them to process and generate human-like text. Multimodal LLMs integrate and process information from various modalities, including text and images. The core capabilities of LLMs include a profound understanding of context, the ability to generate coherent and logical text, and the capacity to tackle complex problems involving textual and multimodal data.

Unlike traditional AI, which primarily focuses on analyzing existing data, Generative AI can synthesize new information, offering unique opportunities for innovation across various domains, including disaster management. The inherent capabilities of LLMs align well with the demands of disaster management. Their ability to analyze extensive datasets, facilitate communication between stakeholders, and support critical decision-making processes makes them invaluable in mitigating the impacts of disasters. Similarly, the strengths of Generative AI in data synthesis, rapid content creation, and the simulation of various scenarios can address specific challenges encountered in disaster preparedness, response, and recovery phases [27].

Recent comprehensive surveys have documented the evolution of machine learning methods specifically for disaster management applications [4]. These studies highlight how machine learning has progressed from basic classification tasks to sophisticated prediction and decision support systems tailored to various disaster contexts. The integration of machine learning across the entire disaster management cycle has been well-documented, with applications ranging from early warning systems to recovery planning [28]. Research foundations are now developing specialized large language models specifically fine-tuned for disaster risk reduction applications, demonstrating the growing recognition of LLMs' potential in this domain [29].

2.4. LLMs and Generative AI for Disaster Management

LLMs and Generative AI are transforming disaster preparedness by enhancing our ability to understand and plan for potential crises. For instance, LLMs can analyze vast datasets about infrastructure, identifying patterns and weaknesses that might make certain areas or systems more vulnerable to specific hazards. This allows for more targeted preventative measures and resource allocation. Moreover, LLMs can be used to assess the impacts of technological disasters such as industrial accidents [10]. Furthermore, LLMs can process complex information, such as the Social Vulnerability Index (SVI), to answer specific questions from communities about their risk factors and potential impacts [3]. This empowers communities to better understand their vulnerabilities and take proactive steps.

International organizations have also recognized the potential of machine learning in disaster risk management, with comprehensive frameworks being developed to guide implementation [30]. These frameworks provide standardized approaches for integrating AI solutions into existing disaster management systems, ensuring compatibility and effectiveness across different contexts and regions.

Generative AI tools, like FEMA's PARC initiative [31], are streamlining the often complex process of hazard mitigation planning for local governments. By automating the generation of plan sections and providing expert guidance, these tools make it easier for communities, especially those with limited resources, to develop comprehensive strategies that can ultimately reduce disaster risks. The broader application of AI in risk assessment, as explored by the UN [32], helps to identify and understand various threats on a global scale, informing international efforts to build resilience. LLMs

 doi:10.20944/preprints202505.1798.v1

also act as powerful knowledge synthesizers, capable of extracting crucial information from diverse sources like news reports, or social media. This extracted knowledge can then be used to answer a wide range of questions about potential risks, such as the likelihood and severity of wildfires or floods, providing valuable insights for planning and resource allocation. These models can even offer tailored advice to individuals and communities on how to mitigate specific risks. Techniques like RAG allow these AI systems to access and integrate real-time data from sources like geographic information systems, providing the most up-to-date information for risk assessment, such as in the context of flood risk [33].

Advancements in early warning systems are becoming increasingly sophisticated thanks to LLMs and Generative AI. For example, Moody's GenAI-powered system for commercial real estate risk [34] demonstrates how AI can continuously monitor news and integrate proprietary data to provide timely alerts about potential financial risks in the sector, which can be triggered by disasters. Similarly, AI offers the potential to significantly improve the speed and accuracy of warnings for major political or military events, providing policymakers with more lead time to respond and potentially prevent escalation or mitigate harm. The Northwestern Terror Early Warning System (NTEWS) [35] shows how machine learning can be applied to model patterns of terrorist behavior, allowing for the forecasting of potential attacks and enabling preventative actions.

Humanitarian organizations have begun implementing machine learning solutions for emergency response in remote areas, combining drone technology with advanced analytics [36]. These initiatives demonstrate how machine learning can be deployed in resource-constrained environments to support disaster monitoring and response. Machine learning models specifically designed for social media monitoring have demonstrated improved accuracy in wildfire detection and tracking [37], showcasing the practical applications of these advanced technologies for specific disaster types.

These advancements are shifting disaster management from a reactive approach to a more proactive one, where AI plays a key role in anticipating and reducing the impact of various crises. Academic institutions are developing integrated platforms that combine social media analysis with AI for accelerating disaster response operations [38], creating ecosystems where various stakeholders can collaborate effectively during emergencies.

Generative AI also plays a crucial role in enhancing preparedness by creating realistic training simulations [36]. These simulations allow emergency responders and other stakeholders to practice their roles and develop more effective strategies for coping with different disaster scenarios. The UNDP's Crisis Academy [39] is exploring the use of technologies like augmented reality, powered by AI, to create immersive and tailored learning experiences, making training more engaging and effective. Beyond training, AI is also being used to optimize the allocation of physical resources, ensuring that the right equipment and personnel are in the right place at the right time to respond effectively to a disaster.

During an active disaster, LLMs are invaluable for achieving real-time situation awareness. They can rapidly analyze the massive amounts of data generated from various sources, including social media, news outlets, and sensor networks, to provide a comprehensive understanding of the unfolding situation. LLMs can assess the credibility and relevance of information shared on social media, filter out noise, and categorize the type of assistance needed [40]. This allows for a more dynamic and informed response, enabling decision-makers to allocate resources effectively and communicate with affected communities. Advanced LLMs can even classify the type of disaster and the most pressing humanitarian aid needs based on the data they analyze. Furthermore, multimodal LLMs enhance this awareness by integrating visual information, such as images and videos shared on social media, with textual data, leading to more accurate assessments of damage and needs on the ground [41]. The fusion of multimodal social media data for disaster image classification represents a promising direction for more comprehensive situational awareness [42], allowing for a more nuanced understanding of disaster impacts.

 doi:10.20944/preprints202505.1798.v1

Effective communication is critical during a disaster, and Generative AI is playing a significant role in improving this aspect. AI can automatically generate emergency alerts and instructions in multiple languages [43], ensuring that vital, potentially life-saving information reaches diverse populations, regardless of language barriers. There's also the potential to personalize these updates, tailoring them to specific geographic areas or even individual needs. AI-powered chatbots, such as FEMA's Hazard Mitigation Assistance Chatbot [44], are being developed to provide immediate and context-specific guidance to both emergency responders and the public. These chatbots can answer a high volume of inquiries about emergency procedures, guidelines, and available assistance, freeing up human responders to focus on more complex and critical tasks. By providing accessible and timely information, Generative AI significantly enhances the efficiency and effectiveness of disaster response efforts.

Several case studies and pilot projects demonstrate the real-world application of these technologies. FEMA's PARC initiative [31] is actively working to make hazard mitigation planning more efficient and accessible for local governments. AI-driven early warning systems are already being implemented, with examples like Google Maps integrating AI for real-time wildfire boundary tracking [45] and AI forecasting systems in Europe outperforming traditional models in predicting hurricane systems [46]. These examples show the increasing reliability and practical utility of AI in providing timely warnings that can save lives and property. The United Nations Development Programme (UNDP) is actively leveraging AI in various stages of crisis response. Their RAPIDA tool [47] uses AI to analyze satellite imagery, social media, and other data to provide rapid insights into the impact of crises, as seen after the Herat earthquake in Afghanistan. These real-world examples underscore the tangible benefits and the growing adoption of LLMs and Generative AI as powerful tools for addressing the complex challenges of disaster management.

3. Methodology

The research utilized a dataset of 10 million tweets collected via the (old) Twitter API. This large volume of data was intended to provide a comprehensive overview of discussions and information shared on Twitter during disaster events [27]. The time frame for the gathered tweets spans between 01/2012 and 12/2022. For the purposes of this study we do not utilize the entirety of the dataset, for reasons of infrastructure cost of the LLM. The processing, including the trial and error period in order to evaluate the efficacy of the finalized prompt, accumulated to approximately 500,000 processed tweets.

3.1. Ground Truth Labeling

To evaluate the performance of the automated LLM labeling approach, a subset of the collected tweets was manually labeled to create a "ground truth" dataset. A pool of 9434 tweets was sampled from the larger dataset, and these tweets were labeled manually. The labeling schema included the following categories:

1. **main_disaster_type**: Categorizing the primary type of disaster being discussed from the following list:
   - Earthquake
   - Tsunami
   - Flood
   - Hurricane
   - Wildfire
   - Drought
   - Heatwave
   - Landslide
   - Volcano
   - Tornado
   - Pandemic

 doi:10.20944/preprints202505.1798.v1

- Famine
- Conflict
- Cyberattack
- Blackout
- Chemical Spill
- Nuclear Accident
- Industrial Accident
- Mass Shooting
- Explosion
- Other
- N/A

2. **severity**: Assessing the perceived severity of the disaster from the following list:
   - Severe damage
   - Mild damage
   - Little or no damage
   - Don't know or can't judge
3. **informative**: Indicating whether the tweet contains informative content related to the disaster. Boolean value
4. **impact**: Describing the type of impact mentioned in the tweet from the following list:
   - Affected individuals
   - Infrastructure and utility damage
   - Injured or dead people
   - Missing or found people
   - Rescue, volunteering or donation effort
   - Vehicle damage
   - Other relevant information
   - Not relevant
5. **location_mentioned**: Identifying if a specific location (country or city) is mentioned in the tweet as free text.
6. **sentiment**: Classifying the overall sentiment expressed in the tweet as positive, negative, or neutral.

This manually labeled dataset served as the benchmark against which the automated LLM labels were compared.

3.2. Automated LLM Labeling

The core of the research involved applying an LLM, specifically gpt-4o-mini, to automatically label the entire dataset (or a significant portion thereof). The process involved the following steps:

3.2.1. Data Preprocessing

A simplified preprocessing function (preprocess_text) was applied to the tweet text to clean it by converting it to lowercase and removing URLs, mentions, and special characters. This step helps in standardizing the input for the LLM.

```
def preprocess_text(text):
    """Cleans and preprocesses text data by removing URLs and mentions."""
    text = str(text)
    text = text.lower()
    text = re.sub(r'[^\w\s]', '', text)
    return text
```

 doi:10.20944/preprints202505.1798.v1

3.2.2. LLM Prompting

A trial and error-crafted prompt was designed to instruct the gpt-4o-mini model on how to classify the tweets. The prompt provided context about the task, specified the categories for classification, and included the lists of keywords for main_disaster_type, severity, informative, and impact. It also specified the required format for the output (JSON) and constraints for sentiment and location mentioned.

Prompt Used:

prompt = f"""You are a helpful assistant that classifies disaster-related tweets.

Classify the following tweet into the categories of main_disaster_type, severity, informative, and impact. Use the provided keywords for each category to determine the appropriate label.

Sentiment must be either positive, negative, or neutral. Location_mentioned should be a country or city name if mentioned in the tweet.

Respond with the appropriate JSON format.

Disaster Types Values List: {', '.join(disaster_types)}

Severity Levels Values List: {', '.join(severity_levels)}

Informative Levels Values List: {', '.join(informative_levels)}

Impact Values List: {', '.join(impact)}

Tweet: {text}

"""

3.2.3. Output Structuring

A Pydantic schema (TweetClassification) was defined to ensure that the LLM's output was structured in a consistent and predictable format. This schema specified the data types for each of the classification categories.

```
class TweetClassification(BaseModel):
    main_disaster_type: str
    severity: str
    informative: str
    impact: str
    location_mentioned: str
    sentiment: str
```

The response_format=TweetClassification parameter in the LLM call ensured the structured form of the output and the parsing according to this schema.

3.3. Evaluation Methodology

The performance of the automated LLM labeling was evaluated by comparing its predictions on a subset of the data with the manually created ground truth labels. The evaluate_labels function was used to calculate several standard classification metrics:

- **Accuracy**: The proportion of correctly classified instances.

 doi:10.20944/preprints202505.1798.v1

- **Precision**: The proportion of predicted positive instances that were actually positive.
- **Recall**: The proportion of actual positive instances that were correctly identified.
- **F1-score**: The harmonic mean of precision and recall, providing a balanced measure of performance.

These metrics were calculated for each individual label category (main_disaster_type, severity, informative, impact, location_mentioned, sentiment) as well as an overall performance metric. The overall accuracy was calculated as the proportion of tweets where all predicted labels matched the ground truth labels. The overall precision, recall, and F1-score were calculated as the average of the category-wise metrics.

3.4. Evaluation Results

The evaluation results from the hold-out validation set (comparing the LLM's output with the manual labels) are as follows:

Table 1. Evaluation metrics for automated LLM labeling across different categories.

Category	Accuracy	Precision	Recall	F1
main_disaster_type	**0.7204**	**0.7275**	**0.7204**	**0.7025**
severity	**0.7087**	**0.6741**	**0.7087**	**0.6601**
informative	**0.8085**	**0.8200**	**0.8085**	**0.8098**
impact	**0.7172**	**0.7004**	**0.7172**	**0.6869**
location_mentioned	**0.8360**	**0.8768**	**0.8360**	**0.8464**
sentiment	**0.8561**	**0.9052**	**0.8561**	**0.8700**
overall	**0.2896**	**0.7840**	**0.7745**	**0.7626**

The evaluation results indicate varying levels of performance across the different labeling categories. The sentiment and location_mentioned categories show the highest accuracy and F1-scores, suggesting that the LLM is particularly adept at identifying sentiment and the presence of location information. The informative category also shows strong performance.

The main_disaster_type, severity, and impact categories have lower accuracy and F1-scores compared to the others, indicating that these categories might be more challenging for the LLM to predict accurately, possibly due to the nuances in language or the complexity of the classification task.

The overall accuracy of 0.2896 suggests that achieving perfect agreement across all six categories for a single tweet is challenging. However, the overall precision, recall, and F1-score are significantly higher, indicating that when considering each category independently (as reflected in the averages), the model performs reasonably well.

 doi:10.20944/preprints202505.1798.v1

Further analysis of the misclassifications could provide valuable insights into the specific types of errors the LLM is making and inform potential improvements to the prompting strategy, keyword lists, or even the labeling schema itself.

Automated LLM labeling of disaster-related tweets provides significant value by efficiently processing massive datasets to extract structured information. This approach transforms raw, unstructured text into actionable data, enabling faster and more scalable analysis compared to manual methods. By categorizing tweets based on disaster type, severity, impact, sentiment, and location, the process unlocks the potential for in-depth analysis of disaster events.

The resulting structured data facilitates crucial insights for disaster response, preparedness, and research. It allows for real-time monitoring potential, understanding public perception, identifying information needs, and mapping the geographical impact of disasters.

4. Dataset Analysis

This section presents an analysis of a sampled dataset comprising 288,926 tweets, a subset of 125,460 of which were identified and labeled as pertinent to specific disaster events. The subsequent analysis aims to elucidate the characteristics of these disaster-related communications through an examination of their type, informativeness, impact, sentiment, and underlying textual patterns identified via unsupervised learning techniques.

4.1. Descriptive Analysis of Disaster-Related Tweets

The distribution of the main_disaster_type within the labeled subset reveals a heterogeneous landscape of events captured by the data (Table 2). The overwhelming prevalence of tweets related to the pandemic (45.74%) underscores the significant impact and widespread discourse surrounding global health crises during the data collection period. The substantial "other" category (19.76%) suggests the presence of diverse, less frequently categorized events, warranting further qualitative investigation to discern the specific nature of these incidents. Notably, industrial accidents (8.82%) and conflict (4.65%) also constitute significant portions of the dataset, indicating the model's capacity to identify tweets associated with both sudden and protracted disaster scenarios. The remaining disaster types, including floods, hurricanes, and nuclear accidents, exhibit lower frequencies, potentially reflecting the relative infrequency or localized nature of these events within the sampled timeframe.

Table 2. Distribution of Main Disaster Types.

Disaster Type	Count	Percentage
Pandemic	**57,384**	**45.74%**
Other	**24,792**	**19.76%**
Industrial Accident	**11,060**	**8.82%**
Conflict	**5,832**	**4.65%**
Flood	**5,007**	**3.99%**

 doi:10.20944/preprints202505.1798.v1

Hurricane	4,197	3.35%
Nuclear Accident	3,328	2.65%
Wildfire	2,498	1.99%
Chemical Spill	1,942	1.55%
Earthquake	1,586	1.26%
Mass Shooting	1,448	1.15%
Explosion	1,167	0.93%
Tornado	1,084	0.86%
Cyberattack	880	0.70%
Heatwave	870	0.69%
Famine	862	0.69%
Drought	479	0.38%
Tsunami	292	0.23%
Landslide	282	0.22%
Volcano	260	0.21%
Blackout	210	0.17%

 doi:10.20944/preprints202505.1798.v1

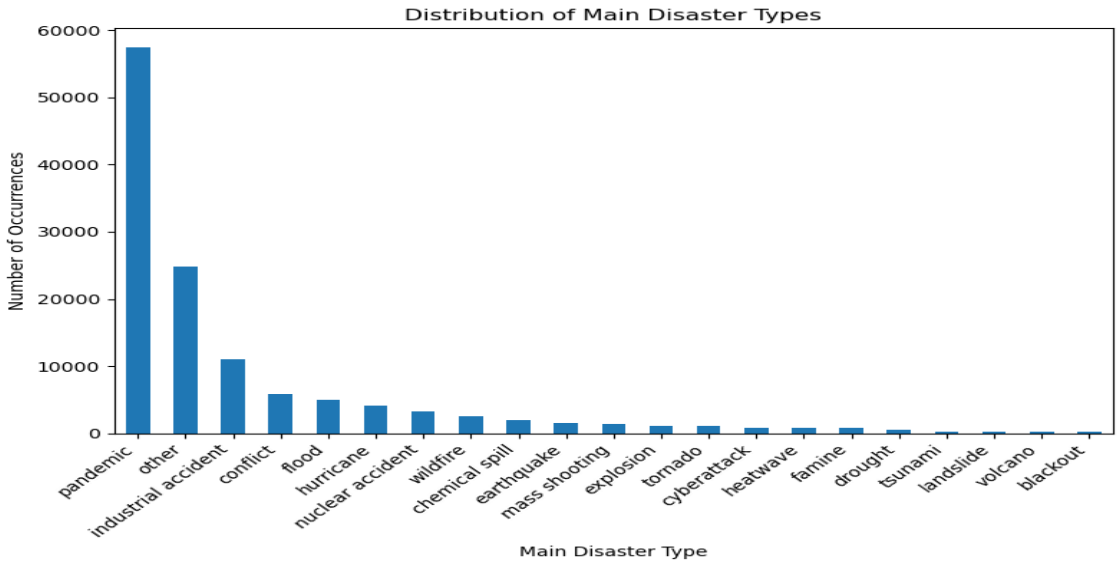

Figure 1. Distribution of main disaster types identified in the analyzed tweets.

The assessment of tweet informativeness reveals that a significant majority (59.69%) were labeled as informative, suggesting a valuable source of real-time information during disaster events. Conversely, 40.30% were deemed not informative, potentially encompassing personal opinions, emotional responses, or irrelevant content.

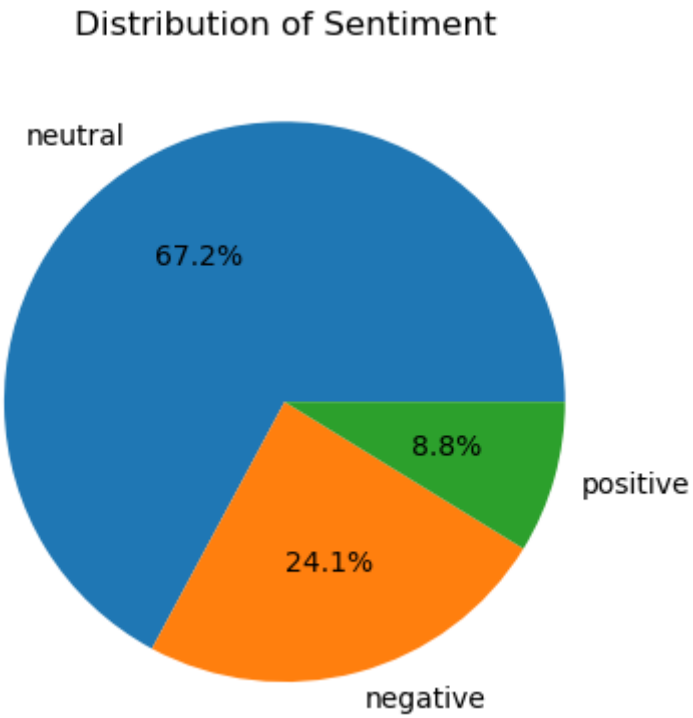

Figure 2. Proportion of tweets classified as informative versus non-informative.

Analysis of the impact categories provides a granular view of the information conveyed. The most frequent categories include "not relevant" (184,145). Among the positively labeled impacts, "other relevant information" (44,123) and "infrastructure and utility damage" (25,738) were prominent, highlighting the focus on general updates and the state of essential services. Reports concerning "affected individuals" (10,450), "injured or dead people" (9,473), and "volunteering or

 doi:10.20944/preprints202505.1798.v1

donation effort" (7,466) underscore the dataset's capacity to capture the humanitarian aspects of disasters.

The heatmap presented visualizes the impact categories across different main disaster types reveals distinct patterns in the information shared on Twitter. For instance, while the pandemic shows a broad impact, affected individuals and death, industrial accidents and chemical spills are strongly associated with reports of infrastructure damage. Conflicts and floods also correlate with damage to infrastructure. Notably, wildfires don't show a clear link to environmental impact which potentially indicates the inability of the model to divulge the impact to the environment.

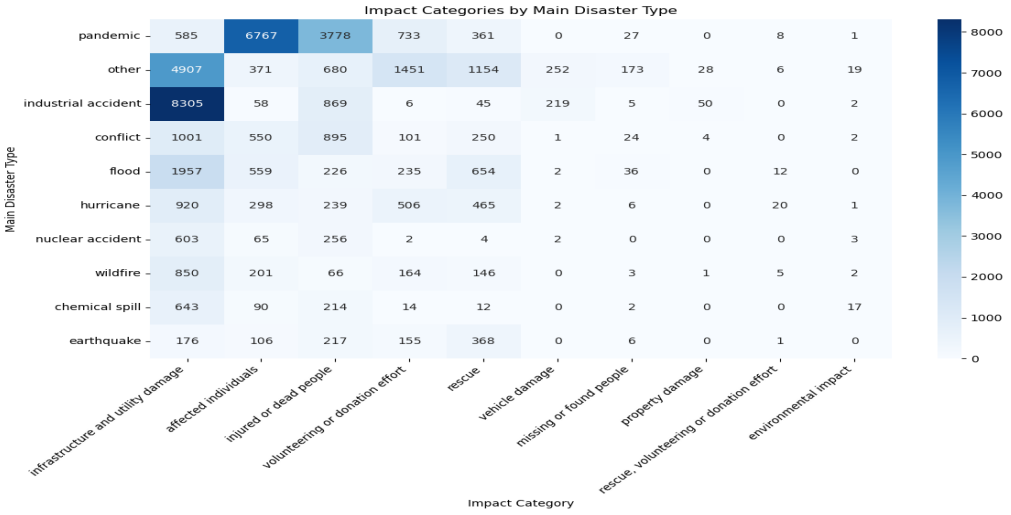

Figure 3. Heatmap showing the relationship between disaster types and impact categories.

Sentiment analysis indicates a predominantly neutral tone (67.15%), which is plausible in the context of factual reporting and information dissemination. Negative sentiment (24.06%) reflects the distress and concern associated with disaster events, while positive sentiment (8.78%) may stem from messages of resilience, recovery, or aid.

4.2. Unsupervised Text Analysis

To uncover latent semantic structures within the textual data, unsupervised learning techniques were employed following text preprocessing steps involving normalization and tokenization.

4.2.1. Word Cloud Visualization

The word cloud generated from the preprocessed text visually emphasizes the most salient terms. The prominence of phrases like "public health," "natural disaster," "emergency management," and "emergency response" suggests a strong thematic focus on the systemic and organizational aspects of disaster events. The frequent occurrence of "https" points to the prevalent use of external links for information sharing. There is a merit in removing the https directly in text preprocessing or cleaning, but the information of the inclusion of a link or multimedia (link to an image) is useful in the terms of how much multimedia other than text are present in disaster related tweets, or in fact tweets in general.

 doi:10.20944/preprints202505.1798.v1

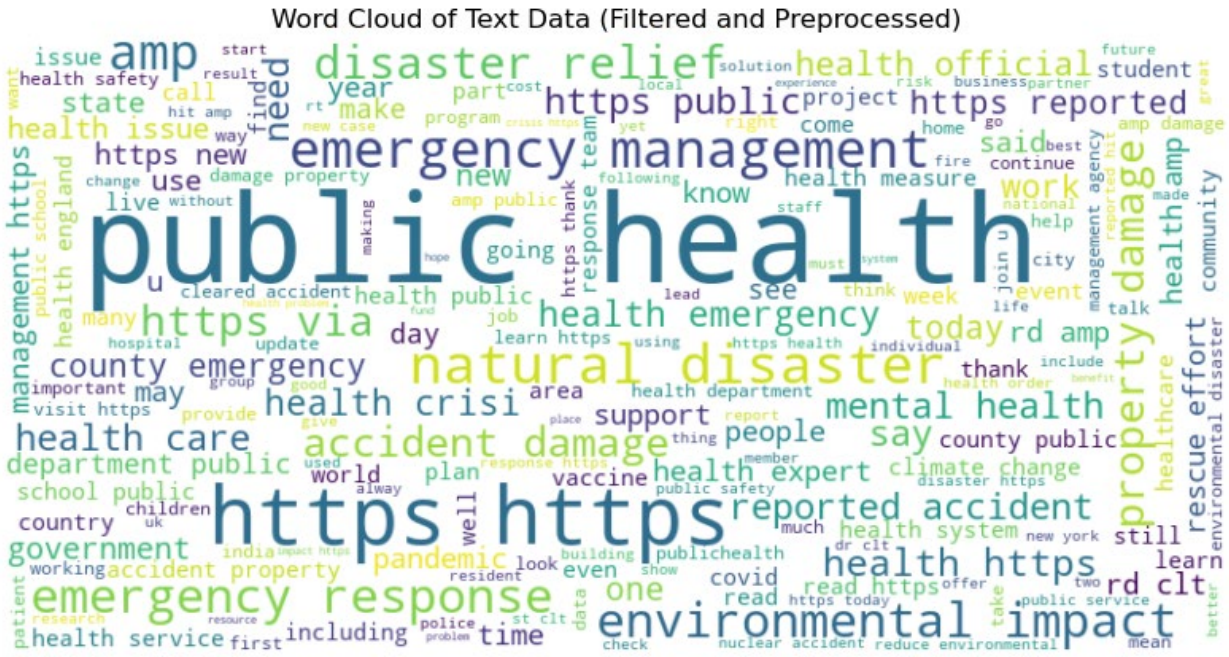

Figure 4. Word cloud visualization of the most frequent terms in the disaster-related tweets.

4.2.2. K-Means Clustering

K-Means clustering, applied to the TF-IDF transformed text data, partitioned the tweets into four distinct clusters, each characterized by a set of dominant terms:

- Cluster 1: health, public, amp, new
- Cluster 2: damage, property, accident, reported
- Cluster 3: emergency, response, management, environmental
- Cluster 4: disaster, natural, relief, help

These clusters offer a preliminary segmentation of the discourse, with Cluster 1 potentially focusing on public health crises and related news, Cluster 2 on the immediate aftermath and impact of events involving damage and accidents, Cluster 3 on the operational and environmental dimensions of disaster response, and Cluster 4 on the broader concepts of natural disasters and humanitarian aid. By utilizing these clusters, new tweets can be categorized with the clustering mechanism as well to identify potential relativity to previous events within the same cluster.

4.2.3. Principal Component Analysis (PCA)

To reduce the dimensionality of the TF-IDF feature space and identify the principal sources of variance in the text data, Principal Component Analysis (PCA) was performed, retaining the top five components. The explained variance ratio for these components is as follows:

[0.03127214 0.02367425 0.0199147 0.01830832 0.01533562].

The explained variance ratios indicate that the first five principal components capture approximately 3.13%, 2.37%, 1.99%, 1.83%, and 1.53% of the total variance in the data, respectively. Cumulatively, these five components account for roughly 10.29% of the variance. While this suggests that a substantial amount of variance is distributed across a larger number of components, these initial principal components likely represent the most dominant underlying themes or patterns in the textual data. The relatively low explained variance for each component individually suggests that the textual information is complex and multifaceted, with no single dominant theme explaining a large proportion of the variance. It is interesting that the PCA couldn't converge as much in certain verbal components, because it goes against the word cloud indication and the clustering indication of strong health correlation in the dataset.

 doi:10.20944/preprints202505.1798.v1

5. Discussion

This research provides a comprehensive exploration into the capabilities of Large Language Models (LLMs), in automating the structuring and labeling of disaster-related social media content. The evaluation of the automated labeling process against a manually annotated ground truth dataset of 9,434 tweets reveals promising yet nuanced results. The high accuracy and F1-scores achieved in sentiment and location_mentioned classification (0.8561/0.8700 and 0.8360/0.8464, respectively) underscore the LLM's proficiency in discerning subjective tones and identifying geographical references within concise textual data. Similarly, the strong performance in the informative category (0.8085/0.8098) suggests the model's ability to effectively differentiate between content offering pertinent information and that which is less relevant in a disaster context.

However, the comparatively lower performance in categorizing the main_disaster_type, severity, and impact (accuracies ranging from 0.7087 to 0.7204 and F1-scores from 0.6601 to 0.7025) warrants further consideration. This suggests that while the LLM demonstrates a strong understanding of general semantic features, the subtle distinctions and contextual nuances required for precise categorization in these areas present a greater challenge. The complexity of inferring the specific disaster type or the perceived severity from short, often emotionally charged social media posts may contribute to these lower scores. The "impact" category, with its diverse range of potential manifestations, similarly poses a complex classification task. The overall accuracy of 0.2896, while seemingly low, highlights the inherent difficulty in achieving perfect agreement across all six diverse labeling categories for each individual tweet, emphasizing the multi-faceted nature of disaster-related social media content.

The subsequent analysis of a larger dataset of 288,926 tweets, with 125,460 labeled as disaster-related, provides valuable insights into the prevalent themes and characteristics of online discourse during crises. The dominance of pandemic-related tweets (45.74%) reflects the unprecedented global impact of the COVID-19 pandemic during the data collection period, highlighting the utility of social media for capturing public discourse during such large-scale events. The prevalence of industrial accidents and conflict further illustrates the broad applicability of social media data in understanding various types of crises.

The unsupervised text analysis, employing word clouds and K-Means clustering, offers a preliminary glimpse into the latent semantic structures within the dataset. The prominence of terms like "public health," "natural disaster," and "emergency response" in the word cloud aligns with the descriptive analysis, reinforcing the thematic focus of the collected data. The four distinct clusters identified through K-Means -- potentially focusing on public health news, immediate impact and damage, operational and environmental response, and broader disaster relief efforts -- provide a foundational structure for understanding the diverse facets of disaster-related online communication. However, the low explained variance ratios observed in the PCA (cumulatively 10.29% for the top five components) suggest that the textual data is highly complex and multifaceted, with no single dominant theme explaining a large proportion of the variance. This complexity underscores the challenges inherent in extracting concise and readily interpretable patterns from large volumes of social media text.

6. Conclusions

In essence, this research highlights the substantial potential of employing Large Language Models (LLMs) to analyze the vast quantities of social media data generated during disaster events. The ability to efficiently extract structured information concerning the type of disaster, its perceived severity, the nature of its impact, and the sentiment expressed within these online communications offers a powerful tool for enhancing situational awareness, guiding resource allocation, and refining communication strategies crucial for effective disaster management.

Looking ahead, several promising avenues for future research warrant exploration. One key direction involves the fine-tuning of existing LLMs on specialized datasets of disaster-related social

 doi:10.20944/preprints202505.1798.v1

media content. This targeted training could further optimize their performance for the specific nuances and vocabulary prevalent in this domain, potentially leading to significant improvements in classification accuracy, particularly for challenging categories like impact category. Recent work on instruction fine-tuned LLMs for multi-label social media classification in disaster contexts (CrisisSense-LLM) represents a promising direction for improved understanding of disaster communications [48].

Furthermore, exploring advanced prompting strategies remains crucial, including the incorporation of more detailed contextual information and diverse examples to better guide the LLM's understanding. Another vital area lies in investigating alternative state-of-the-art LLMs and architectural innovations, including those designed for multimodal data processing, to leverage the rich information contained in images and videos shared during disasters. The development of multimodal datasets combining social media text with remote sensing imagery offers new possibilities for comprehensive disaster monitoring and response [49]. Emerging research demonstrates that multimodal LLMs like Gemini can accurately estimate earthquake intensity from social media posts, potentially revolutionizing rapid damage assessment [50].

The development of more granular and hierarchical labeling schemes could also address the limitations of broad categorization, allowing for more precise and actionable insights. To enhance the LLM's ability to adhere to specific labeling guidelines and extract particular types of information, future work will focus on instruction-tuning techniques. This approach involves training the LLM on a dataset of instructions paired with desired outputs, thereby improving its capacity to follow complex classification tasks.

Finally, a comparative analysis of the LLM-based approach with traditional machine learning algorithms trained on manually labeled data would offer a more comprehensive understanding of the relative strengths and weaknesses of each method.

Author Contributions: Vasileios Linardos: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Maria Drakaki: Conceptualization, Investigation, Project administration, Resources, Supervision, Validation, Writing – review & editing. Panagiotis Tzionas: Supervision, Writing – review & editing.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Centre for Research on the Epidemiology of Disasters (CRED) Institute Health and Society – UCLouvain. 2023 Disasters in Numbers: A Significant Year of Disaster Impact. CRED 2024. Available online: https://files.emdat.be/reports/2023_EMDAT_report.pdf.
2. Imran, M.; Castillo, C.; Diaz, F.; Vieweg, S. Processing Social Media Messages in Mass Emergency: A Survey. ACM Comput. Surv. 2015, 47(4), 1-38.
3. Liu, S.; Brewster, C.; Shaw, D. Ontologies for Crisis Management: A Review of State of the Art in Ontology Design and Usability. In Proceedings of the 10th International ISCRAM Conference, Baden-Baden, Germany, 12-15 May 2013; pp. 349-359.
4. Linardos, V.; Drakaki, M.; Tzionas, P.; Karnavas, Y.. Machine Learning in Disaster Management: Recent Developments in Methods and Applications. Mach. Learn. Know. Extr. 2022, 4(2).
5. Wang, Z.; Ye, X.; Tsou, M.H. Spatial, temporal, and content analysis of Twitter for wildfire hazards. Nat. Hazards 2016, 83(1), 523-540.
6. Jongman, B.; Wagemaker, J.; Romero, B.R.; De Perez, E.C. Early Flood Detection for Rapid Humanitarian Response: Harnessing Near Real-Time Satellite and Twitter Signals. ISPRS Int. J. Geo-Inf. 2015, 4(4), 2246-2266.
7. Avvenuti, M.; Cresci, S.; Marchetti, A.; Meletti, C.; Tesconi, M. EARS (Earthquake Alert and Report System): A Real Time Decision Support System for Earthquake Crisis Management. In Proceedings of the 20th ACM

 doi:10.20944/preprints202505.1798.v1

SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24-27 August 2014; pp. 1749-1758.

8. Wang, Z.; Ye, X. Social Media Analytics for Natural Disaster Management. Int. J. Gegr. Inf. Sci. 2018, 32(1), 49-72.
9. Alam, F.; Ofli, F.; Imran, M. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In Proceedings of the 12th International AAAI Conference on Web and Social Media, Stanford, CA, USA, 25-28 June 2018; pp. 465-473.
10. Linardos, V., Drakaki, M., Tzionas, P., A transformers-based approach on industrial disaster consequence identification from accident narratives, Procedia Comp. Sci. 2023, 217, 1446–1451.
11. Houston, J.B.; Hawthorne, J.; Perreault, M.F.; Park, E.H.; Goldstein Hode, M.; Halliwell, M.R.; McElderry, J.A.; Griffith, S.A. Social Media and Disasters: A Functional Framework for Social Media Use in Disaster Planning, Response, and Research. Disasters 2015, 39(1), 1-22.
12. Alam, F.; Ofli, F.; Imran, M.; Aupetit, M. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. In Proceedings of the 15th International ISCRAM Conference, Rochester, NY, USA, 20-23 May 2018; pp. 553-562.
13. Zou, L.; Lam, N.S.N.; Cai, H.; Qiang, Y. Mining Twitter Data for Improved Understanding of Disaster Resilience. Ann. Am. Assoc. Geogr. 2018, 108(5), 1422-1441.
14. Gruebner, O.; Lowe, S.R.; Sykora, M.; Galea, S.; Subramanian, S.V.; Shankardass, K. A Novel Surveillance Approach for Disaster Mental Health. PLoS ONE 2017, 12(7), e0181233.
15. Weber, E., Papadopoulos, D., Lapedriza, À., Ofli, F., Imran, M. and Torralba, A. (2023). Incidents1M: a large-scale dataset of images with natural disasters, damage, and incidents. Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4768-4781: https://doi.org/10.1109/TPAMI.2022.3191996.ok
16. Linardos, V., Drakaki, M. (2023), Assessing the Impact of the 2021 Evia Wildfires through Social Media Analysis. Proceedings of the International Conference on Humanitarian Crisis Management (KRISIS 2023), M. Drakaki, D. Vega (Editors). Institute for the Management of Refugee Flows and Crises, University Research Center, International Hellenic University, ISBN 978-618-5630-17-1 (e-book). https://www.ihu.gr/ucips/wp-content/uploads/sites/4/2023/12/KRISIS_2023_paper_14_Linardos_-et-al.pdf.
17. Imran, M.; Mitra, P.; Castillo, C. 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, Portorož, Slovenia, 23-28 May 2016; pp. 1638-1643.
18. Alam, F.; Ofli, F.; Imran, M. HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. In Proceedings of the 14th International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 8-11 June 2020; pp. 15-25.
19. Alam, F.; Sajjad, H.; Imran, M.; Ofli, F. CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. In Proceedings of the 15th International AAAI Conference on Web and Social Media, Virtual Event, 7-10 June 2021; pp. 923-932.
20. Qazi, U.; Imran, M.; Ofli, F. GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information. SIGSPATIAL Special 2020, 12(1), 6-15.
21. Banda, J.M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, E.; Tutubalina, E.; Chowell, G. A Large-scale COVID-19 Twitter Chatter Dataset for Open Scientific Research – An International Collaboration. Epidemiologia 2021, 2(3), 315-324.
22. Starbird, K.; Maddock, J.; Orand, M.; Achterman, P.; Mason, R.M. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In Proceedings of the 8th International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1-4 June 2014; pp. 654-657.
23. Shen, H.; Ju, Y.; Zhu, Z. Extracting Useful Emergency Information from Social Media: A Method Integrating Machine Learning and Rule-Based Classification. Int. J. Environ. Res. Public Health 2023, 20(3), 1862.
24. Palen, L.; Anderson, K.M. Crisis Informatics—New Data for Extraordinary Times. Science 2016, 353(6296), 224-225.

 doi:10.20944/preprints202505.1798.v1

25. Crawford, K.; Finn, M. The Limits of Crisis Data: Analytical and Ethical Challenges of Using Social and Mobile Data to Understand Disasters. GeoJournal 2015, 80(4), 491-502.
26. Hughes, A.L.; Palen, L. Twitter Adoption and Use in Mass Convergence and Emergency Events. Int. J. Emerg. Manag. 2009, 6(3-4), 248-260.
27. Otal, H.T.; Stern, E.; Canbaz, M.A. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, 25–27 June 2024.
28. Chamola, V.; Hassija, V.; Gupta, S.; Goyal, A.; Guizani, M.; Sikdar, B. Disaster and Pandemic Management Using Machine Learning: A Survey. IEEE Internet Things J. 2021, 8(18), 13749-13768.
29. CIMA Research Foundation. The Large Language Model for Disaster Risk Reduction. CIMA Research Foundation; 2024. Available online: https://www.cimafoundation.org/en/news/the-large-language-model-for-disaster-risk-reduction/.
30. The World Bank. Machine Learning for Disaster Risk Management. World Bank Documents and Reports; 2019. Available online: https://documents1.worldbank.org/curated/ar/503591547666118137/pdf/133787-WorldBank-DisasterRiskManagement-Ebook-D6.pdf.
31. Federal Emergency Management Agency. Planning Assistance Resource Center (PARC): Leveraging AI for Community Resilience. FEMA Technical Report; 2023.ok
32. United Nations Office for Disaster Risk Reduction. AI for Disaster Risk Reduction: Global Assessment Report. UNDRR; 2023.ok
33. Smith, A.B.; Brown, C.D.; Jones, E.F. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. J. Flood Risk Manag. 2015, 10(3). https://doi.org/10.1111/jfr3.12154.ok
34. Moody's Analytics. GenAI-Powered Early Warning System for Commercial Real Estate Risk. White Paper; 2023.ok
35. Benigni, M.C.; Joseph, K.; Carley, K.M. Online Extremism and the Communities That Sustain It: Detecting the ISIS Supporting Community on Twitter. PLoS ONE 2017, 12(12), e0181405.ok
36. World Food Programme. The Benefits of Machine Learning in Emergencies from River DEEP to Mountain SKAI. WFP Drones; 2022. Available online: https://drones.wfp.org/updates/benefits-machine-learning-emergencies-river-deep-mountain-skai.ok
37. Prevention Web. Machine Learning Model Uses Social Media for More Accurate Wildfire Monitoring. Prevention Web; 2022. Available online: https://www.preventionweb.net/news/machine-learning-model-uses-social-media-more-accurate-wildfire-monitoring.ok
38. Illinois Institute of Technology. Tapping Social Media and AI for Faster Disaster Response. IIT News; 2023. Available online: https://www.iit.edu/news/tapping-social-media-and-ai-faster-disaster-response.ok
39. United Nations Development Programme. Crisis Academy: AI-Powered Training for Emergency Responders. UNDP Innovation Report; 2023.ok
40. Caragea, C.; Silvescu, A.; Tapia, A.H. Identifying Informative Messages in Disaster Events using Convolutional Neural Networks. In Proceedings of the 13th International ISCRAM Conference, Rio de Janeiro, Brazil, 22-25 May 2016.ok
41. Li, X.; Caragea, D.; Zhang, H.; Imran, M. Visual and Textual Analysis of Social Media and Satellite Images for Disaster Type Detection. In Proceedings of the 16th International ISCRAM Conference, Valencia, Spain, 19-22 May 2019; pp. 752-764.ok
42. Xu, L.; Li, X.; Zhang, H.; Caragea, D. Disaster Image Classification by Fusing Multimodal Social Media Data. ISPRS Int. J. Geo-Inf. 2021, 10(10), 636.ok
43. Petersen, K.; Büscher, M.; Kuhnert, M. Designing with Users: Co-Design for Innovation in Emergency Technologies. In Proceedings of the 10th International ISCRAM Conference, Baden-Baden, Germany, 12-15 May 2013; pp. 319-328.ok
44. Federal Emergency Management Agency. FEMA Launches AI-Powered Hazard Mitigation Assistance Chatbot. Press Release; 2023.ok
45. Google. Wildfire Tracking in Google Maps: AI-Powered Boundary Detection. Technical Report; 2022.ok

 doi:10.20944/preprints202505.1798.v1

46. European Centre for Medium-Range Weather Forecasts. AI Forecasting Systems Outperform Traditional Models in Hurricane Tracking. ECMWF Newsletter; 2023.ok
47. Zade, H.; Shah, K.; Rangarajan, V.; Kuwajima, H.; Crooks, A.; Tandon, P.; Palen, L.; Starbird, K. From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response. Proc. ACM Hum.-Comput. Interact. 2018, 2(CSCW), 1-18.ok
48. Yin, K.; Liu, C.; Mostafavi, A.; Hu, X. CrisisSense-LLM: Instruction Fine-Tuned Large Language Model for Multi-label Social Media Text Classification in Disaster Informatics. arXiv preprint 2024, arXiv:2406.15477.ok
49. Zhang, Z. A Global Multimodal Flood Event Dataset with Heterogeneous Text and Multi-Source Remote Sensing Images. Big Earth Data 2024, 8(1), 2358615.ok
50. Mousavi, M. et al. Gemini and Physical World: Large Language Models Can Estimate the Intensity of Earthquake Shaking from Multimodal Social Media Posts. Geophys. J. Int. 2024.ok