Detecting Actionable Requests and Offers on Social Media During Crises Using LLMs

Ahmed El Fekih Zguir
Qatar Computing Research Institute,
Hamad Bin Khalifa University, Doha, Qatar
azguir@hbku.edu.qa

Ferda Ofli
Qatar Computing Research Institute,
Hamad Bin Khalifa University, Doha, Qatar
fofli@hbku.edu.qa

Muhammad Imran
Qatar Computing Research Institute,
Hamad Bin Khalifa University, Doha, Qatar
mimran@hbku.edu.qa

ABSTRACT

Natural disasters often result in a surge of social media activity, including requests for assistance, offers of help, sentiments, and general updates. To enable humanitarian organizations to respond more efficiently, we propose a fine-grained hierarchical taxonomy to systematically organize crisis-related information about requests and offers into three critical dimensions: *supplies*, *emergency personnel*, and *actions*. Leveraging the capabilities of Large Language Models (LLMs), we introduce Query-Specific Few-shot Learning (QSF Learning) that retrieves class-specific labeled examples from an embedding database to enhance the model's performance in detecting and classifying posts. Beyond classification, we assess the actionability of messages to prioritize posts requiring immediate attention. Extensive experiments demonstrate that our approach outperforms baseline prompting strategies, effectively identifying and prioritizing actionable requests and offers.

Keywords

Large language models, Disaster management, Taxonomy, Social media, Query-Specific Few-shot Learning

INTRODUCTION

Social media platforms have become vital during crises, as they enable real-time communication and coordination during emergencies. During past events like Hurricane Dorian, Hurricane Harvey, and COVID-19, platforms like Twitter (now X) and Facebook played a crucial role, allowing people to request help and organize assistance quickly (Mihunov et al., 2020; Olawale, 2018). For emergency responders and humanitarian organizations, this stream of information presents an opportunity to enhance situational awareness, allocate resources more efficiently, and provide aid where it is most needed. However, much of the data, including emotional expressions and commentary from those outside the affected area, is irrelevant to immediate response needs. To fully harness the value of social media during disasters, it is essential to develop effective methods for identifying the most relevant and actionable information amidst the noise (He et al., 2017).

This study focuses on identifying actionable social media posts categorized as "requests" (messages seeking assistance) and "offers" (messages providing assistance). Actionability, as defined by Zade et al. (2018), refers to posts containing sufficient contextual information to assess urgency, time, and the specific location for the required assistance or available resources. Previous efforts to identify such posts have primarily focused on high-level categorization into requests and offers, with limited attention to fine-grained distinctions based on specific types of assistance. Moreover, while supervised machine learning models trained on human-labeled data are ideal for such tasks, they are often impractical in disaster scenarios due to the time, resources, and effort required for data

annotation. Furthermore, adapting these models to new events or categories demands additional training data, which is challenging to obtain during emergencies.

To address these limitations, this study focuses on fine-grained identification of requests and offers without relying on extensive human-labeled data. We propose a fine-grained hierarchical taxonomy that organizes crisis-related information along three critical dimensions: *supplies*, *emergency personnel*, and *actions*. This taxonomy was developed using a *"top-down"* approach informed by documents, guidelines, and expertise from humanitarian organizations, resulting in a structured framework comprising 1,093 elements across six levels of granularity. Unlike previous *"bottom-up"* approaches that rely on text mining and topic modeling techniques (Durham et al., 2023), our taxonomy ensures consistency, practicality, and alignment with real-world disaster management needs.

A significant challenge in prior research has been the oversimplification of social media classification tasks as single-label multi-class problems, which fail to account for the multifaceted nature of disaster-related posts. For instance, a tweet like, *"We're cooking meals for displaced families tonight. DM me if you want to help or donate. #CaliforniaFires,"* contains both an offer (meals for displaced families) and a request (volunteers to help). To address this, we frame the problem as a multi-label multi-class classification task, where a post can be labeled as both a request and an offer, and each post can belong to multiple other fine-grained categories.

Additionally, existing approaches often lack the granularity required for effective disaster response. For example, the category *"medical supplies"* can encompass diverse items such as bandages, oxygen tanks, or mobility aids (e.g., wheelchairs), which are critical to differentiate during a crisis. Traditional research has also predominantly focused on tangible supplies, overlooking requests for actions (e.g., search and rescue) or specific personnel (e.g., military support during riots). Our taxonomy addresses this gap by preserving granular and actionable information, enabling a more comprehensive analysis of social media posts.

To implement this taxonomy, we leverage Large Language Models (LLMs) with various prompting strategies. Additionally, we propose a Query-Specific Few-shot Learning (QSF learning) approach supported by Retrieval-Augmented Generation (Lewis et al., 2020). This method enables the classification of posts into fine-grained categories without extensive labeled data and outperforms several baselines on both real-world and synthetic data. By framing the problem as three distinct multi-label, multi-class classification tasks corresponding to supplies, actions, and emergency personnel, our framework provides a robust and scalable solution for detecting actionable requests and offers on social media data during disasters. We provide the dataset, taxonomy, and other related resources at the following URL: https://crisisnlp.qcri.org/requests_offers/.

RELATED WORK

The growing reliance on social media during natural disasters has spurred significant research into identifying and categorizing actionable information, particularly requests and offers, to aid humanitarian efforts. Early research primarily focused on traditional machine learning techniques for classifying disaster-related posts. For instance, Purohit et al. (2014) developed and released labeled datasets and regular expressions to identify requests and offers on Twitter. Their work used cascading SVM classifiers, prioritizing precision over recall, to classify tweets into categories like money, shelter, and medical supplies. However, their approach was limited to posts containing either requests or offers, without accounting for posts that included both. Building on this, Nazer et al. (2016) incorporated additional features, such as URLs and hashtags, and used decision tree classifiers to enhance classification performance. Similarly, Devaraj et al. (2020) employed GloVe word vectors (Pennington et al., 2014) to distinguish urgent posts from non-urgent ones, demonstrating the evolving sophistication of feature engineering in this domain. Taking a step further, Basu et al. (2022) presented a utility-driven model for optimized resource allocation in a post-disaster scenario, based on information extracted from microblogs in real time.

The introduction of transformer architectures (Vaswani et al., 2017) marked a paradigm shift in natural language processing (NLP), enabling efficient processing of long text sequences through attention mechanisms. This breakthrough paved the way for models like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), which established new benchmarks in text classification. Fine-tuning these pre-trained models on disaster-related tasks became a popular approach (Howard & Ruder, 2018). For instance, Seeberger and Riedhammer (2022) fine-tuned BERT to classify disaster-related tweets into actionable categories. Prompt engineering and few-shot prompting, introduced with GPT-3, further reduced dependence on large labeled datasets, showcasing the potential for effective performance with minimal examples (Brown et al., 2020).

More recent work has leveraged multiple pre-trained transformer models to address the complexity of disaster-related tasks. For example, L. Zhou et al. (2022) employed BERT, RoBERTa, and XLNet to classify tweets across various disaster-related categories, outperforming traditional machine learning methods. Ziaullah et al. (2024) highlighted the zero-shot capabilities of large language models (LLMs) for monitoring critical infrastructure during emergent

disasters. Furthermore, Lamsal et al. (2024) introduced crisis-specific fine-tuned transformers (*CrisisTransformers*) to classify tweets into requests and offers, demonstrating significant improvements over earlier approaches.

Several studies have emphasized the importance of structured taxonomies for organizing disaster-related information. RweetMiner (Ullah et al., 2021) introduced a formal framework for identifying and categorizing "rweets" (request tweets) into sub-types such as medical, food, and shelter, using machine learning classifiers with high precision. Similarly, Basu et al. (2017) analyzed WhatsApp messages during the 2015 Nepal earthquake to curate resource requirements and delays, demonstrating the value of taxonomy-driven approaches for disaster preparedness. More recently, Durham et al. (2023) employed text mining and topic modeling techniques, such as latent Dirichlet allocation, to develop a *bottom-up* taxonomy from tweets. In contrast, our work adopts a *top-down* approach, leveraging humanitarian guidelines and expertise to define a fine-grained hierarchical taxonomy. This taxonomy captures three critical dimensions—supplies, emergency personnel, and actions—providing a robust framework for classifying posts into actionable categories.

Fine-grained classification has proven essential for improving resource allocation during crises. For instance, Basu et al. (2019) experimented with supervised and unsupervised models for identifying resource needs and availabilities, emphasizing the importance of granular classifications when high-quality training data is available. Ullah et al. (2021) categorized tweets into sub-types such as medical, food, and shelter using machine learning classifiers with high precision. Similarly, Zhang et al. (2021) employed a topic model-based framework to identify the spatial distribution of demand for relief supplies, while Dutt et al. (2019) proposed a methodology to match resource needs and availabilities, considering resource type, quantity, and geographical proximity. However, earlier works often simplified the problem to single-label classification, overlooking the complexity of posts containing both requests and offers. For example, a post offering food while simultaneously requesting volunteers exemplifies the need for multi-label classification.

Highlighting the need for multilingual support, Vitiugin and Purohit (2024) introduced MulTMR, a multilingual serviceability model leveraging knowledge distillation with task-related and behavior-guided teacher models to detect and rank help requests on social media. Their approach, validated across multiple languages and disaster events, demonstrated substantial performance improvements in multilingual scenarios. Similarly, Lamsal et al. (2024) proposed CReMa, a systematic framework integrating textual, temporal, and spatial features for cross-lingual identification and matching of requests and offers. Their multilingual embedding space and crisis-specific pretrained model significantly advanced performance benchmarks and highlighted the importance of cross-lingual analysis in disaster response.

In addition to other challenges, earlier research has also emphasized the need for actionable intelligence tailored to responders' roles. Zade et al. (2018) highlighted issues like information overload and misinformation in integrating social media data into disaster response. They proposed shifting from general situational awareness to actionable intelligence, aligning with our redefinition of *actionability*. Our approach prioritizes posts with sufficient context to drive direct actions, addressing gaps in traditional urgency-based classifications and supporting humanitarian organizations in effective decision-making during crises.

METHODOLOGY

Our goal is to develop a robust approach capable of identifying any predefined categories of requests or offers while enabling rapid deployment without requiring large amounts of labeled data or supervised model training. To this end, we design a comprehensive taxonomy comprising three key dimensions (supplies, actions, emergency personnel). We then propose a Query-Specific Few-shot Learning (QSF learning) approach leveraging Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to construct few-shot prompts for message classification. Our goal is not only to improve classification performance but also to assess how these improvements generalize across different LLMs. To ensure broad applicability, we evaluate our approach using multiple instruction-tuned LLMs of varying sizes and architectures, including Llama 3 8B, Llama 3.1 8B (Grattafiori et al., 2024), Gemma 2 9B (Team et al., 2024), Mistral 7B v0.2 (Jiang et al., 2023), and GPT-4o mini. While all models are tested on the full range of baseline prompts and our proposed solution, GPT-4o mini—being a paid API—was evaluated exclusively on our solution (QSF Learning) to benchmark its performance relative to the other models. This diverse set of models allows us to systematically examine whether our findings generalize across LLMs with different capacities and training backgrounds. Next, we provide details of our methodology.

Taxonomy Generation

Most prior works, including the work by Lamsal et al. (2024), build on the taxonomy by Purohit et al. (2014), which categorizes resources as tangible supplies or services requested or offered during disasters, such as monetary

donations, volunteer work, shelter, clothing, and medical supplies. While valuable, these taxonomies face two key limitations: *(i)* They group distinct resource types—tangible supplies (e.g., medical supplies, clothing) and intangible services (e.g., volunteer work)—failing to capture their unique characteristics and complicating accurate categorization. *(ii)* They often omit critical resources frequently highlighted during disasters. For example, social media posts during Hurricane Harvey in 2017 frequently requested bottled water and baby formula, while the 2020 Beirut explosion saw significant demand for dust masks and personal protective equipment—resources absent in existing taxonomies. These gaps hinder comprehensive disaster response and underscore the need for more fine-grained categorization.

Furthermore, previous research overlooks two critical types of requests and offers: "actions" and "emergency personnel." Many social media posts during disasters highlight urgent actions, such as search and rescue operations, debris clearance, medical aid, food distribution, and crowd control. For example, posts during the 2015 Nepal earthquake frequently requested search and rescue teams for locating survivors (Bleiker, 2015), while the 2023 Türkiye-Syria earthquake emphasized coordinated debris removal and emergency medical care (OCHA, 2023b). Beyond actions, messages contain requests for trained personnel, e.g., firefighters, medical professionals, and law enforcement. For instance, the California wildfires saw appeals for firefighting reinforcements, and the Ebola outbreak required infectious disease specialists and emergency nurses.

To address these gaps, we propose a taxonomy with three main branches: **supplies**, **actions**, and **emergency personnel**.

- **Supplies:** tangible resources such as medical supplies, shelter, food, water, hygiene products, and more.
- **Actions:** represent tasks like search and rescue, medical care, debris clearance, and food distribution.
- **Emergency personnel:** refers to trained responders, including paramedics, firefighters, structural engineers, and specialized volunteers.

Next, we utilized official online resources, including situation reports, guidelines, and articles from organizations such as UN OCHA, UNDP, FEMA, Red Cross, and UNHCR (OCHA, 2023a; UNHCR, 2023; UNICEF, 2023). The selection of these three dimensions—**supplies**, **actions**, and **emergency personnel**—is based on a thorough manual evaluation of the collected documents. Through careful analysis, we observed that the majority of disaster response activities naturally cluster around these three core areas: tangible resources (which we categorize as supplies), human responders and teams (personnel), and required operational activities (actions). This observation reflects the real-world practices and language used by humanitarian agencies, ensuring that our taxonomy is aligned with operational workflows and sufficiently comprehensive to capture the essential elements of disaster response.

Information from 20 such sources was processed using GPT-4o, a state-of-the-art language model, to generate and augment categories within the three main branches. Table 1 outlines the prompts used to create the different taxonomy levels. The model suggested categories and sub-categories at different depths of the taxonomy underwent thorough human review, with adjustments made as needed. In the prompts, we ensure that each level provides increasingly fine-grained information about its parent category, incorporating synonyms, regional variations, and linguistic nuances. For example, the category "bandages" was expanded to include terms such as "Band-Aids," "adhesive strips," and "plasters," reflecting diverse social media expressions. The final taxonomy comprises 1,093 elements distributed as follows at different depth levels: Level-1: 3, Level-2: 33, Level-3: 129, Level-4: 635, Level-5: 271, and Level-6: 22. Figure 1 shows a partial view of our taxonomy, highlighting the root branches, all categories at depth two that are directly under the root, and an expansion of some selected branches.

Task Definition and Classifiers

Task Definition

The goal of our task is to process an input message (e.g., a tweet) and extract key structured information relevant to crisis response. Specifically, for each input message, we aim to generate an 8-element tuple capturing its essential attributes. The tuple consists of:

- **Type**: A list indicating the type of message, such as `"request"`, `"offer"`, or `"other"`.
- **Actions (r)**: A list of requested actions (e.g., `"Search and Rescue"`).
- **Supplies (r)**: A list of requested supplies (e.g., `"Medical"`, `"Clothing and Warmth"`).
- **Personnel (r)**: A list of requested personnel (e.g., `"Medical and Health Teams"`).
- **Actions (o)**: A list of offered actions (e.g., `"Infrastructure Repair and Debris Clearance"`).
- **Supplies (o)**: A list of offered supplies (e.g., `"Money"`).
- **Personnel (o)**: A list of offered personnel (e.g., `"Medical and Health Teams"`).
- **Actionability**: A boolean value indicating whether the message contains actionable information.

Table 1. Prompts used to generate and expand the taxonomy

Extraction Prompt: Extracting Relevant Terms	**Level 2 Prompt: Creating the Hierarchy (Per Category)**
<context> Shared Context (given below) **</context>** **<instruction>** You are provided with multiple documents from humanitarian organizations that contain terms related to disaster relief. These terms can broadly fall under three categories: actions, emergency personnel, and supplies. Extract all relevant terms and group them based on which category they fall under. **</instruction>**	**<context>** Shared Context **</context>** **<instruction>** You are provided with a list of terms that fall under the category (actions, supplies, or personnel). Your task is to group the terms into distinct, non-overlapping groupings within this category. Ensure that the groupings cover all terms and are logical, clear, and comprehensive. **</instruction>**
Level 3 & 4 Prompt: Refining the Combined Taxonomy	**Levels 5 & 6 Prompt: Expansion**
<context> Shared Context **</context>** **<instruction>** You are provided with a taxonomy that organizes terms under actions, emergency personnel, and supplies. Your task is to refine and improve this hierarchy by reorganizing or expanding branches as needed. Different branches may vary in depth, but ensure the taxonomy remains logical, comprehensive, and well-organized. **</instruction>**	**<context>** Shared Context **</context>** **<instruction>** You are provided with a detailed taxonomy. Your task is to review and expand leaf terms by: 1. Adding synonyms for terms that are commonly referred to by different names. Make sure to include terms used both in social media and by humanitarian organizations. 2. Adding subcategories or breaking down elements where justified. Only make expansions where they are necessary to improve clarity, usability, or completeness. Keep the taxonomy concise and avoid unnecessary additions. **</instruction>**
Shared context: In this task, you are assisting in the development of a taxonomy to classify and organize requests and offers made during disaster scenarios. The goal is to create a structured, top-down taxonomy based on a corpora of documents sourced from humanitarian organizations. These documents include situation reports, needs assessments, articles, websites, and guidelines which describe various forms of needs and offers.	

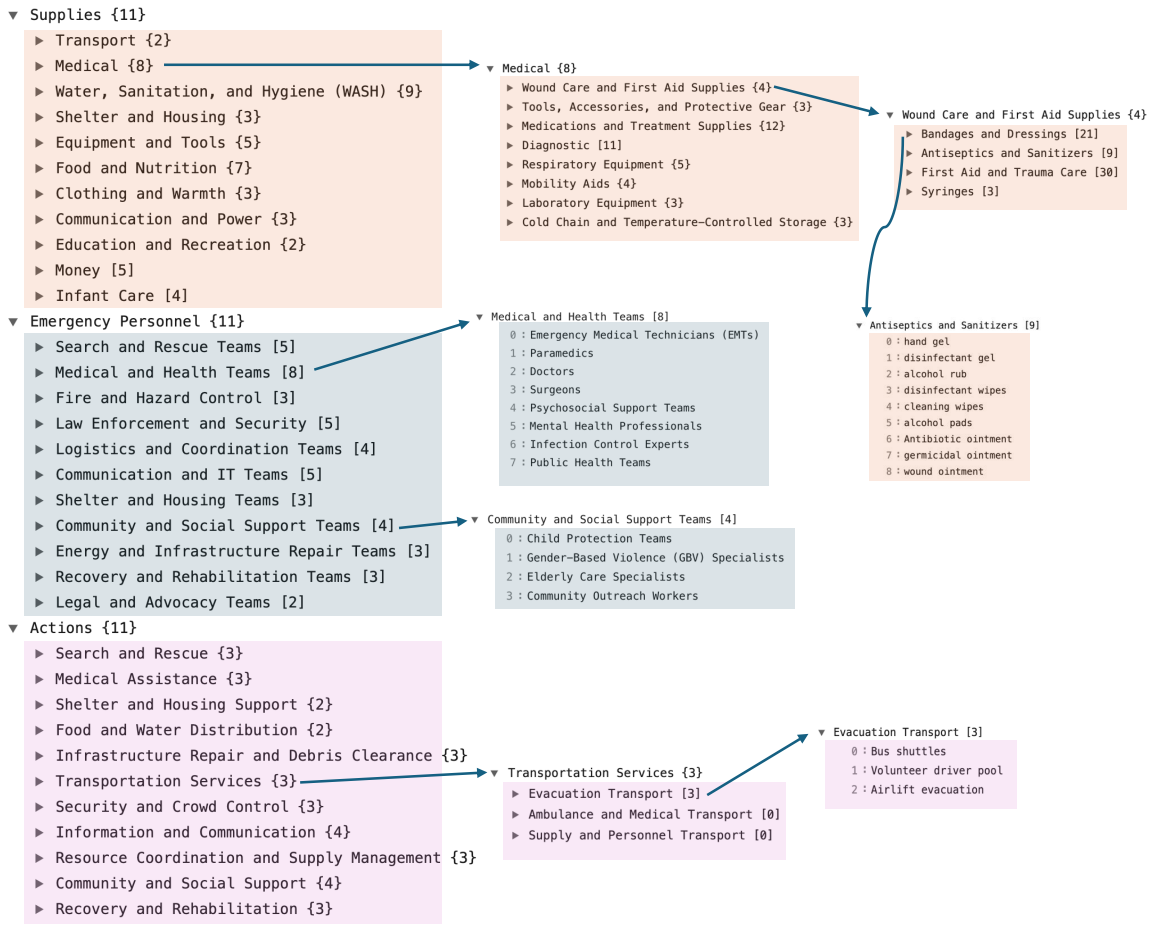


Figure 1. Request and Offer Taxonomy: A partial representation

Here, `(r)` denotes *request* and `(o)` denotes *offer*. All elements of the tuple are treated as multi-label classification tasks, except for Actionability, which is framed as a binary classification problem.

Building upon prior work (Zade et al., 2018), we define actionability as: *Information related to a crisis that either requires a response or constitutes an offer, where details such as time, location, urgency, and specific needs (e.g., quantities or actions) determine its usefulness.* For any category without relevant information, the output is either an empty list (for list-based fields) or `False` (for actionability).

Formally, let T be the set of input messages, and $t \in T$ an individual message. We define a mapping function f as:

$$f(t) = (\text{Type}, A_r, S_r, P_r, A_o, S_o, P_o, \text{Actionability})$$

where:

- Type: A subset of {`"request"`, `"offer"`, `"other"`}.
- A_r: Set of requested actions ($A_r \subseteq A$).
- S_r: Set of requested supplies ($S_r \subseteq S$).
- P_r: Set of requested personnel ($P_r \subseteq P$).
- A_o: Set of offered actions ($A_o \subseteq A$).
- S_o: Set of offered supplies ($S_o \subseteq S$).
- P_o: Set of offered personnel ($P_o \subseteq P$).
- Actionability: A boolean value (`True` or `False`).

Here, A, S, and P represent predefined sets of target labels for actions, supplies, and personnel, respectively. Each set contains 11 distinct labels, derived from our taxonomy at depth 2.

Baseline classifiers

We approach the classification of crisis-related messages using LLMs, as illustrated in Figure 2. Our method relies on prompt engineering to guide the model's output, starting with a baseline prompt and progressively refining it by incorporating additional contextual information. Each refinement results in a new classifier, designed to evaluate the impact of various prompting strategies. These classifiers differ in the level of detail they provide, specifically in terms of taxonomy depth and the inclusion of labeled examples (few-shots). The baseline prompt follows this structure:

- **Instruction:** In this section, we explain the LLM's role as an advanced AI trained to classify social media posts related to crises, emphasizing the use of taxonomy and its knowledge of natural disasters.
- **Taxonomy:** Provides the taxonomy, defining the labels the model should output.
- **Context:** Discusses social media's critical role during disasters, categorizing posts into three types: Request, Offer, and Other. It highlights the importance of identifying urgent, actionable messages, using in-context learning principles (Y. Zhou et al., 2024).
- **Output Format:** Specifies the output as a JSON object with structured labels.

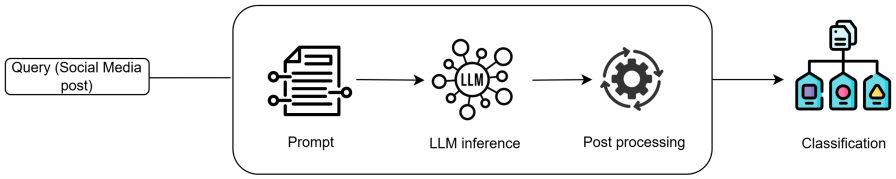


Figure 2. Diagram of baseline (BL) classifiers. Each BL has a different prompt in the diagram.

The baseline prompt, used with BL 1 (baseline classifier 1), can be seen below, which served as the foundation for subsequent variants detailed as follows:

- **BL 1 (Baseline Classifier 1):** Uses the baseline prompt with taxonomy limited to depth 2 and no few-shot examples.
- **BL 2:** Extends the taxonomy to depth 3, providing a more detailed hierarchical structure to improve label comprehension.
- **BL 3:** Incorporates few-shot prompting by adding a small set of labeled examples to the prompt, improving task understanding as shown in prior work (Brown et al., 2020).

- **BL 4:** Combines BL 2 and BL 3, integrating both a detailed taxonomy and few-shot examples.
- **BL 5:** Builds on BL 4 by adding chain-of-thought (CoT) prompting (Kojima et al., 2022). This classifier addresses ambiguous cases with detailed explanations and requires step-by-step reasoning for each classification decision, including actionability.

Baseline Prompt for Classification

<**Instruction**>
You are an advanced AI trained to label social media posts related to crisis situations, specifically natural disasters. Your goal is to label the given posts. You make use of the vast knowledge that you have about natural disasters, your knowledge of social media posts of people during those disasters, the provided taxonomy and information mentioned below to label the data.
</**Instruction**>
<**Context**>
Natural disasters, such as hurricanes, wildfires, earthquakes, floods, tornadoes, landslides, etc., have significant impacts on communities. During these events, social media platforms (such as Twitter, Facebook, and Instagram) become primary channels of sharing and receiving information. This information can be broadly categorized into **Request**, **Offer**, or **Other**. Note that some posts are both **Request** and **Offer** at the same time ... (context was omitted to fit on a page)
</**Context**>
<**Taxonomy**>
{Taxonomy till depth 2 here}
</**Taxonomy**>
<**Output formats**>
This is the output format when the type is either **Request** or **Offer** or both:

```
{
    "text": "The social media post text here",
    "type": ["Request" | "Offer" | "Offer", "Request"],
    "action_request": [...],
    "personnel_request": [...],
    "supplies_request": [...],
    "action_offer": [...],
    "personnel_offer": [...],
    "supplies_offer": [...],
    "actionability": true | false
}
```

This is the output format when the type is **Other**:

```
{
    "text": "The social media post text here",
    "type": ["Other"]
}
```

</**Output formats**>
<**Task**>
Your task is to label the following social media post based on the taxonomy and rules mentioned above. Only output the JSON dictionary and nothing else.
</**Task**>

Query-Specific Few-Shot Learning Approach

Prior studies show that including labeled examples (few-shots) in prompts improves LLM performance on classification tasks. However, in a multi-class setting, adding an equal number of shots to all classes often leads to a decline in performance, as demonstrated by Imran et al. (2025). Building on this insight, we introduce a query-specific few-shot learning strategy (QSF learning) using Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to retrieve relevant, query-specific examples dynamically for each input message.

For each message, we compute its embedding using OpenAI's `text-embedding-3-small` model and retrieve the $k/2$ most similar labeled examples from a pre-built embedding database. This database contains examples from previous crisis events. To maintain variability and prevent bias toward specific classes, we also include $k/2$ randomly selected examples. These examples are then appended to the prompt. The detailed steps of the QSF Learning algorithm can be seen in Figure 3 and are as follows:

1. **Top-k retrieval:** Embeddings for each input are computed using OpenAI's `text-embedding-3-small` model, and the $k/2$ most relevant examples are retrieved using cosine similarity.
2. **Random sampling:** $k/2$ random examples are added for variability and to avoid over-fitting.
3. **Mapping and appending:** Retrieved embeddings are mapped to their labeled examples and appended to the prompt.
4. **Prompt creation:** A tailored prompt combines structured sections with retrieved examples.
5. **Inference, cleanup, and evaluation:** The same steps as BL 1–5 are applied to process and evaluate outputs.

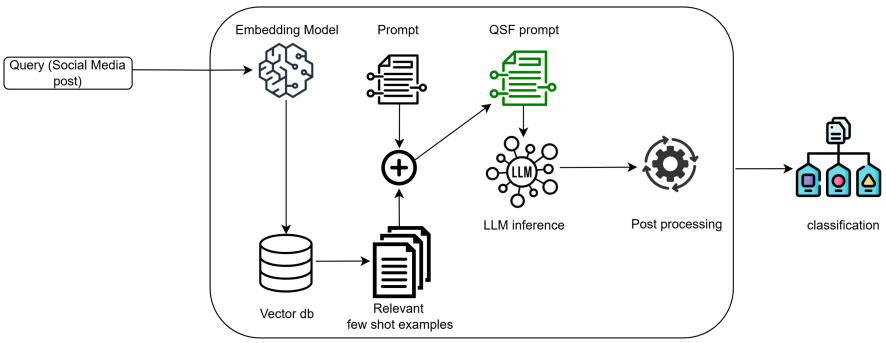


Figure 3. A high-level overview of the Query-Specific Few-shot Learning approach

DATA AND EVALUATION

Synthetic Data Generation

To evaluate the proposed methodology, we sought ground-truth data aligned with our detailed taxonomy. However, to the best of our knowledge, no such fine-grained dataset currently exists. As a result, we opted to generate synthetic data (tweets) to mimic social media posts during disasters. To achieve this, we tested various data generation methods, ultimately combining them to produce a diverse, high-quality dataset with unique elements and accurate labels. We used GPT-4o (with a May 2023 knowledge cutoff) as our data generation model.

Our initial data generation approach involved a simple prompt containing the classes from depth 2 of our taxonomy. In this approach, we asked the model to generate 100 tweets and their labels simultaneously. However, the generated data lacked naturalness and appeared robotic, as shown below:

"Urgently need water, sanitation, and hygiene (WASH) at Westside Park. Please help!"
"Offering water, sanitation, and hygiene (WASH) at St. Andrews Church. Available anytime."
"Urgently need medical supplies at Westside Park."

We observed that the model overly relied on the provided taxonomy labels, resulting in repetitive and unnatural outputs. Additionally, asking the model to generate 100 labeled examples in a single prompt led to repetitive text, as the model got "lazy" and started stitching examples by merely substituting labels from the taxonomy in the same text. We observed significant improvements when we seeded the prompt with real disaster event names. This way, we ask the model to use its knowledge of the events mentioned. For instance, when focusing on the Türkiye-Syria earthquake of 2023, the generated data appeared more realistic:

"This is so heartbreaking. Entire neighborhoods flattened in Kahramanmaraş. Stay strong. #TurkeySyriaEarthquake"

Table 2. Data generation prompt evaluation results

Prompt	Avg. Examples	Mean Similarity	Max Similarity
P1	38	0.4169	0.8154
P2	40	0.4022	0.7617
P3	68	0.3660	0.8189
P4	70	0.3727	0.8302

We further redesigned the prompt structure to address other issues, e.g., related to the length of the tweets, realisticness, etc., through iterative rounds of prompt engineering. The final structure included detailed instructions, context, output format, and enhanced instructions, formatted as follows:

Synthetic Data Generation Prompt

<Instruction>
You are an advanced AI trained to generate realistic and diverse synthetic social media posts related to crisis situations, specifically natural disasters. Your goal is to create posts that closely mimic real-life data while ensuring creativity, uniqueness, and variability. You make use of the vast knowledge that you have about natural disasters and social media posts of people during those disasters to generate the data.
</Instruction>
<Context>
{same context as baseline prompt for classification}
</Context>
<EnhancedInstructions>
The data that you generate should adhere to the following guidelines:

- All generated data must be unique. Avoid creating similar posts with minor variations (e.g., changing only the place or disaster name).
- Data should look as realistic as possible, simulating posts created by humans during a natural disaster.
- Posts should be diverse, including noisy data, partial data, and complete data.
- Incorporate realistic elements such as links, phone numbers, hashtags, grammatical mistakes, abbreviations, etc., to enhance authenticity.
- Integrate references to actual past events to ground the data in reality. Use the following disasters as reference points: {list of chosen disasters}.
- Ensure a balance of content types (**Request**, **Offer**, and **Other**) in the generated posts.

</EnhancedInstructions>
<OutputFormat>
The output should be a JSON object containing 100 generated posts in the following format:

```
{
    "posts": [
        "generated text 1 here",
        "generated text 2 here",
        ...
        "generated text n"
    ]
}
```

</OutputFormat>
<Task>
Your task is to generate N realistic social media posts based on the context and instructions provided above. Use the same context as the baseline prompt for classification. Each post should adhere to the guidelines, reflect a variety of disaster-related content, and maintain uniqueness.
</Task>

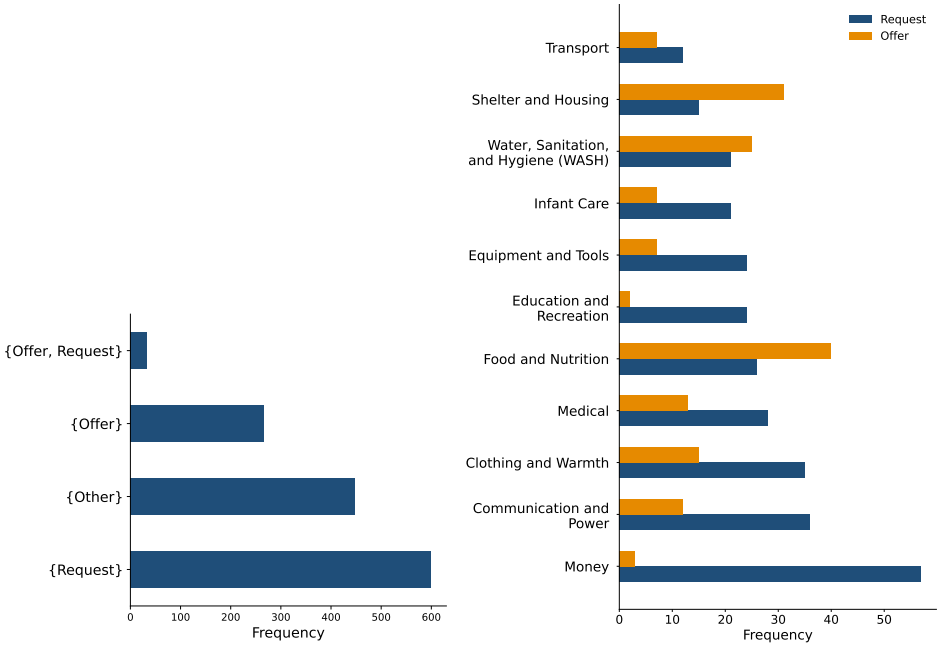
To test the best data generation approach through prompting, in total, we created four prompts based on the above-mentioned structure, as follows:

- **Prompt 1:** Taxonomy + labeling (data and labels generated together).
- **Prompt 2:** No taxonomy + labeling.
- **Prompt 3:** Taxonomy (data generation only).
- **Prompt 4:** No taxonomy (data generation only).

We instructed the LLM to generate 75 examples per prompt, using a temperature setting of 0.8 to enhance variability. Each prompt was executed 10 times to evaluate which prompting strategy produced better results, particularly in terms of generating more unique tweets. We calculated the mean and maximum similarity of the generated texts and recorded the number of messages produced. Specifically, we calculated the cosine similarity between all pairs of embeddings of generated examples. OpenAI's `text-embedding-3-small` model was used to create the embeddings. While the model was instructed to generate 75 examples, the output varied across runs. Table 2 summarizes the averaged metrics across the 10 iterations.

The results indicate that generating and labeling simultaneously (Prompts 1 and 2) produces fewer examples due to token limitations and results in higher similarity between outputs. This aligns with prior research showing that breaking complex tasks into smaller chunks improves the performance of LLMs (Khot et al., 2023).

Based on these findings, we selected Prompt 4 (no taxonomy, data generation only) for 80% of the data and Prompt 3 (taxonomy, data generation only) for 20% to balance realism and coverage. We focused on disasters such as the Pakistan floods (2014), Türkiye-Syria earthquakes (2023), Australian bushfires (2019), Hurricane Maria (2017), and Haiti earthquake (2010). From an initial pool of 2,000 generated examples, duplicates (pairs with cosine similarity > 0.925) and unnatural outputs were removed, resulting in 1,346 high-quality examples.



(a) Type **(b) Supplies**

Figure 4. Distribution of the synthetic generated data across (a) Type and (b) Supply categories

The generated data was labeled using few-shot learning, guided by human-labeled examples. Human annotators reviewed 40% of the data (538 examples) and found fewer than 10% mislabeled, which were manually corrected. Specifically, for the message "type" task, only 10 examples were mislabeled, and for other categories (*actions, supplies, personnel*), 50 examples had minor issues with extra or missing labels. Posts labeled as type "Other" (neither a request nor an offer) were excluded from further labeling for actions, supplies, personnel, and actionability categories. Figures 4 & 5 show the distribution of the generated data for the message types, supplies, actions, and emergency personnel for both requests and offers. As for actionability, from the 899 posts that were either requests or offers, 748 were actionable while 151 were not actionable.

Real-World Data

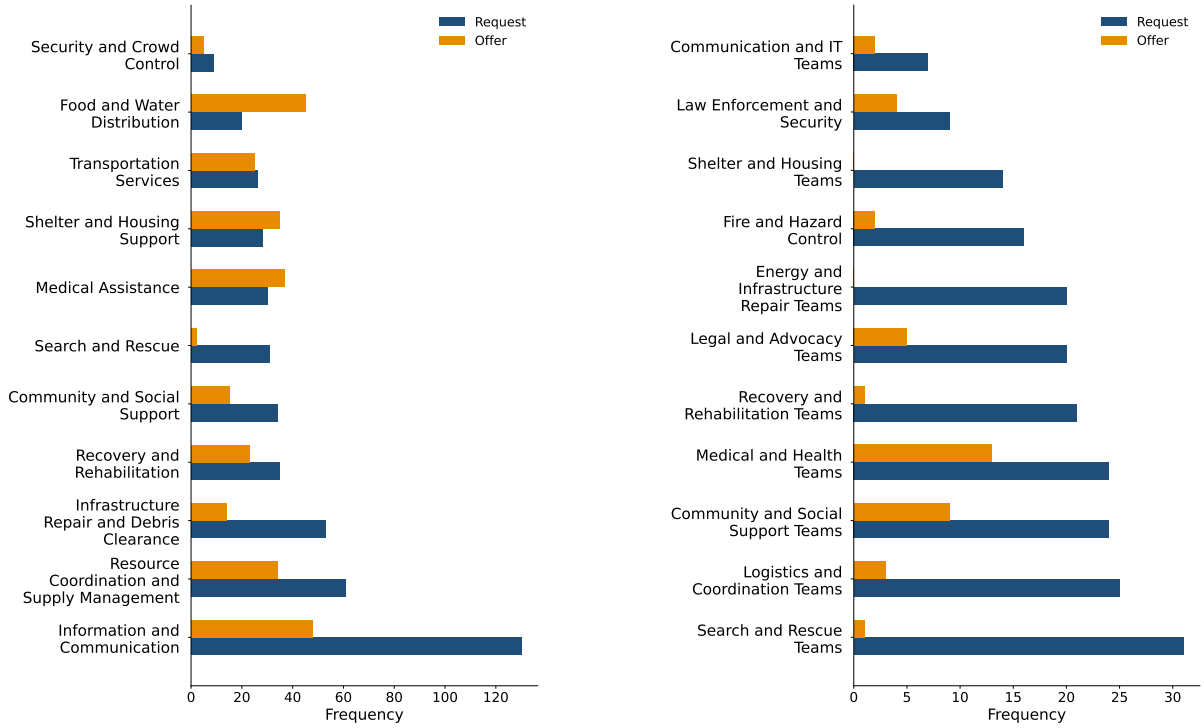
While synthetic data enabled us to create a controlled and balanced dataset aligned with our detailed taxonomy, it was essential to evaluate the robustness of our approach on real-world data, where posts are inherently noisier, less structured, and often incomplete. To this end, we utilized an existing dataset of disaster-related tweets collected and originally annotated by Purohit et al. (2014) and later improved by Lamsal et al. (2024). Their dataset contains tweets labeled as either requests or offers during the Hurricane Sandy disaster, providing a solid foundation for real-world evaluation.

From this dataset, we randomly sampled 300 tweets and manually annotated them according to our taxonomy. Each tweet was reviewed to assign the relevant labels across all applicable categories, including type, supplies, actions, personnel, and actionability. Importantly, this subset of real-world data allowed us to test the model's performance in scenarios where noise and incomplete information are prevalent. For example, consider the following tweet:

> *#LiveWire Game Donated $10,000 to Hurricane Sandy Voters: The rapper wanted to help storm victi... http://t.co/EWkLK4ph #LiveWireRecords*

Here, part of the message is truncated, and the text lacks explicit mentions of key elements such as location or clear action verbs—common challenges encountered in authentic social media posts. By incorporating such examples, we ensured that the model's ability to generalize extends beyond synthetic, well-formed data and can handle the ambiguity and noise characteristic of real-world platforms like Twitter.

Due to the labor-intensive nature of manual labeling, especially given the multi-label setup and the breadth of categories, we limited the annotation to 300 examples. Nevertheless, this subset proved sufficient to validate that our methodology remains effective even when applied to naturally occurring, less structured data.



(a) Actions **(b) Emergency Personnel**

Figure 5. Distribution of the synthetic generated data across (a) Action and (b) Emergency Personnel categories

Evaluation metrics

We split the synthetic dataset into 50% training (673 examples) and 50% evaluation (673 examples). Examples used for few-shot prompting in BL 3, 4, and 5 are strictly from the training set. For the QSF Learning method, relevant examples are retrieved exclusively from the training set to ensure no data leakage into the evaluation set. For the real-world dataset, we use a training set of 107 examples and a test set of 200 examples. We evaluate the classification performance using micro F1-scores across all multi-label tasks: type, supplies, actions, and personnel. Micro averaging aggregates contributions of all classes globally, treating each instance-label pair equally, making it particularly well-suited for multi-label classification tasks. For the binary classification task of actionability, we report macro F1-scores to account for both classes equally, regardless of class distribution.

RESULTS AND DISCUSSION

Our task includes one binary classification task (i.e., actionability) and seven multi-label classification tasks. Despite careful prompt design, the model occasionally produces problematic outputs, such as assigning labels outside the taxonomy. Figure 6 shows the distribution of the errors from models (Llama 3) at the classification/inference time. BL 1 and 2 exhibit a substantial number of errors (359 and 410, respectively), whereas BL 3 through 6 show a significant reduction. This sharp decline indicates that adding few-shot examples to the prompts significantly enhances the LLM's ability to generate taxonomy-compliant outputs. BL 3 and 4, which include few-shot examples, produce similar counts, as do BL 5 and our QSF learning technique. The difference between BL 4 and 5 can be attributed to the inclusion of chain-of-thought prompting in BL 5, which encourages step-by-step reasoning and results in better alignment with the taxonomy.

In addition to errors, the model sometimes provides broken responses that cannot be processed and evaluated even after post processing. Only BL 3 and 4 result in such responses, with BL 3 resulting in 38 broken responses, and BL 4 resulting in 8 broken responses. Both prompts lack chain-of-thought prompting, indicating that while few-shot examples reduce mislabels, incorporating structured reasoning further stabilizes the output quality.

Table 3 and 4 present evaluation results for our multi-task, multi-label classification tasks on both synthetic and real-world data. We run our evaluation on multiple models, namely: Llama 3 8B, Llama 3.1 8B, Gemma 2 9B, Mistral 7B v0.2, and GPT-4o mini. For the *Type* task, baseline prompts (BL1–BL5) exhibit gradual improvements, with BL5 consistently outperforming earlier versions across all models. This is likely due to the relative simplicity

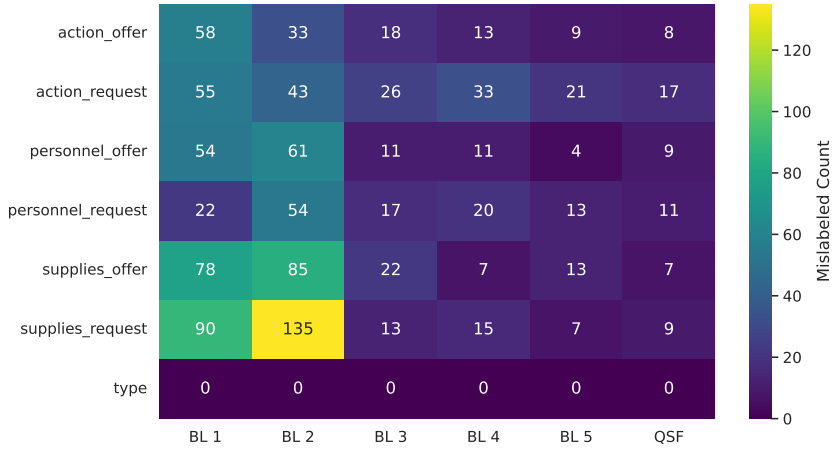


Figure 6. Error distribution during inference on synthetic data using Llama 3

of the *Type* task and its limited label space (*request*, *offer*, *other*). However, incorporating QSF learning yields the highest F1-scores across most models. For instance, on synthetic data, Llama 3.1 improves from 0.86 (BL5) to 0.89, while GPT-4o mini achieves a peak F1-score of 0.92. On real-world data, although absolute F1-scores are lower due to increased complexity, the same pattern persists. Models such as Gemma 2 and Llama 3 benefit from a 2-3% improvement over their best baselines, further confirming the advantage of QSF Learning for this task.

For the *Supplies* task, baseline prompts improve steadily from BL1 to BL5, particularly when few-shot prompting is introduced. Given the task's complexity, involving 11 labels and 2048 label combinations (2^{11}), static prompts alone achieve moderate performance. The QSF Learning consistently leads to the highest scores across both synthetic and real-world data. For example, in the synthetic setting, Llama 3.1 improves from 0.82 (BL5) to 0.84, while GPT-4o mini achieves an F1 score of 0.88. A similar trend is observed in real data, where the QSF classifier ("QSF") consistently boosts model performance to around 0.83–0.85, suggesting that context-aware prompting is particularly effective in providing the necessary contextual variety to handle high-variability multi-label tasks.

The *Actions* task follows a similar trajectory. Baseline prompts struggle, especially in real data, with BL1 and BL2 achieving F1 scores as low as 0.20–0.30 across models. However, each subsequent enhancement results in gradual gains. The most notable improvements are observed with QSF learning, which increases F1 scores by up to 15% compared to the strongest static baseline. For instance, on synthetic data, Mistral improves from 0.62 (BL5) to 0.77, while on real-world data, Llama 3's performance increases from 0.42 (BL5) to 0.57. This indicates that tailoring examples dynamically allows the model to better capture complex action patterns.

The *Personnel* task presents significant challenges across both datasets, reflected by generally lower F1 scores. In real-world data, baseline prompts achieve particularly poor performance, often under 0.30. Despite this, our QSF learning approach consistently yields the highest improvements. For example, Llama 3.1 improves from 0.27 (BL5) to 0.47, and Gemma 2 increases from 0.25 to 0.38. On synthetic data, the trend is similar, with performance gains of approximately 10–15% across models. These results suggest that dynamically selecting query-relevant few-shot examples is crucial for addressing class imbalance and label sparsity in the *Personnel* task.

Examining the detailed taxonomy reveals that *supplies* are the most granular dimension, followed by *actions* and *personnel*. This reflects the humanitarian organization corpora used to build the taxonomy, which primarily emphasize *supplies*. The taxonomy itself contains 946 supply-related elements, compared to 90 for actions and 57 for personnel, contributing to richer coverage in that dimension.

Furthermore, supplies are typically mentioned explicitly in social media posts, making them easier to classify. In contrast, references to actions or personnel are often implicit, requiring the model to infer intent, which increases prediction difficulty. This challenge is compounded by class prevalence: supplies dominate both our synthetic and real datasets, while actions and personnel are less frequently labeled.

In our real-world dataset (307 labeled posts: 107 for retrieval, 200 for testing), posts labeled with supplies significantly outnumber those mentioning actions or personnel. Consequently, the embedding database used in our QSF classifier is denser and more representative for supplies, providing the model with more relevant examples. These factors—the taxonomy's granularity, explicit mentions, and higher class frequency—collectively explain the stronger classification performance observed for supplies.

Task	Model	BL1	BL2	BL3	BL4	BL5	QSF
Type	Mistral	0.73	0.74	0.80	0.81	0.87	0.87
	Llama 3.1	0.71	0.72	0.82	0.82	0.86	0.89
	Gemma 2	0.73	0.74	0.85	0.83	0.88	0.86
	Llama 3	0.75	0.77	0.84	0.82	0.85	0.89
	GPT-4o mini	–	–	–	–	–	0.92
Supplies	Mistral	0.62	0.57	0.79	0.82	0.82	0.81
	Llama 3.1	0.57	0.58	0.75	0.77	0.82	0.84
	Gemma 2	0.69	0.63	0.79	0.75	0.75	0.79
	Llama 3	0.38	0.41	0.76	0.78	0.79	0.79
	GPT-4o mini	–	–	–	–	–	0.88
Actions	Mistral	0.49	0.51	0.55	0.57	0.62	0.77
	Llama 3.1	0.47	0.48	0.55	0.59	0.71	0.77
	Gemma 2	0.49	0.53	0.61	0.61	0.69	0.73
	Llama 3	0.40	0.49	0.54	0.57	0.64	0.75
	GPT-4o mini	–	–	–	–	–	0.77
Personnel	Mistral	0.44	0.40	0.46	0.49	0.54	0.64
	Llama 3.1	0.45	0.41	0.51	0.52	0.62	0.70
	Gemma 2	0.51	0.53	0.57	0.56	0.59	0.66
	Llama 3	0.31	0.35	0.46	0.47	0.54	0.69
	GPT-4o mini	–	–	–	–	–	0.72
Actionability	Mistral	0.64	0.63	0.70	0.67	0.72	0.67
	Llama 3.1	0.54	0.56	0.50	0.44	0.54	0.59
	Gemma 2	0.62	0.64	0.68	0.67	0.78	0.49
	Llama 3	0.68	0.60	0.60	0.64	0.58	0.81
	GPT-4o mini	–	–	–	–	–	0.60

Table 3. Results (F1-scores) for all models and baselines on synthetic data.

The *Actionability* task shows mixed results. On synthetic data, the improvements due to our QSF learning approach vary by model. For instance, Llama 3.1 sees a moderate increase from 0.54 (BL5) to 0.59, while Llama 3 shows a significant gain from around 0.58 to 0.81. However, Gemma 2's QSF score (0.49) is lower than its best baseline (0.78), indicating that the benefits of QSF learning might depend on how well the model's underlying representation aligns with the task. In real-world data, though, our method generally yields better F1-scores i.e., 0.66 for Llama 3.1 and 0.70 for Gemma 2, which points to its potential to enhance recall for less frequent classes.

These findings reveal key insights into both model behavior and prompt design in multi-task, multi-label classification. Across both synthetic and real-world datasets, QSF learning consistently delivers the most substantial performance gains, particularly for tasks with complex label structures such as *Actions* and *Personnel*. By tailoring contextual examples to each input, this technique enables models to better handle label imbalance and variability, leading to superior generalization. While static prompt enhancements (BL1–BL5) result in incremental improvements, they plateau quickly, especially in granular tasks. Interestingly, the results show that QSF learning not only improves performance across tasks but also narrows the gap between models of different sizes and capabilities. High-performing models like GPT-4o mini benefit from this technique, but smaller models such as Llama 3.1 and Gemma 2 also achieve competitive results when QSF learning is applied. This indicates that prompt engineering, particularly dynamic few-shot approaches, plays a pivotal role in bridging the performance disparity between models, allowing even smaller models to generalize effectively in complex, real-world scenarios.

LIMITATIONS AND FUTURE WORK

In this section, we outline the key limitations and biases of our study. One primary source of bias stems from the dataset used. While a portion of our evaluation is conducted on a real-world dataset comprising 300 manually labeled tweets, the relatively small size of this dataset limits its ability to fully capture the variability, noise, and evolving nature of social media posts during crises. To address the scarcity of large-scale curated datasets, we also utilized synthetically generated data produced using GPT-4o. Although GPT-4o provides diverse and partially

Task	Model	BL1	BL2	BL3	BL4	BL5	QSF
Type	Mistral	0.51	0.56	0.63	0.65	0.69	0.74
	Llama 3.1	0.64	0.66	0.74	0.73	0.72	0.75
	Gemma 2	0.55	0.58	0.69	0.67	0.74	0.77
	Llama 3	0.54	0.54	0.75	0.73	0.74	0.77
	GPT-4o mini	–	–	–	–	–	0.72
Supplies	Mistral	0.58	0.53	0.83	0.84	0.79	0.83
	Llama 3.1	0.41	0.45	0.63	0.69	0.80	0.82
	Gemma 2	0.76	0.57	0.80	0.79	0.71	0.85
	Llama 3	0.41	0.32	0.79	0.81	0.81	0.83
	GPT-4o mini	–	–	–	–	–	0.85
Actions	Mistral	0.21	0.22	0.37	0.36	0.43	0.54
	Llama 3.1	0.23	0.21	0.36	0.39	0.38	0.55
	Gemma 2	0.34	0.31	0.48	0.49	0.50	0.50
	Llama 3	0.29	0.26	0.41	0.43	0.42	0.57
	GPT-4o mini	–	–	–	–	–	0.61
Personnel	Mistral	0.07	0.08	0.00	0.19	0.21	0.21
	Llama 3.1	0.06	0.07	0.19	0.17	0.27	0.47
	Gemma 2	0.04	0.14	0.22	0.19	0.25	0.38
	Llama 3	0.06	0.00	0.19	0.22	0.22	0.31
	GPT-4o mini	–	–	–	–	–	0.35
Actionability	Mistral	0.60	0.53	0.49	0.45	0.47	0.58
	Llama 3.1	0.43	0.47	0.47	0.40	0.60	0.66
	Gemma 2	0.50	0.51	0.69	0.64	0.63	0.70
	Llama 3	0.45	0.48	0.52	0.54	0.45	0.47
	GPT-4o mini	–	–	–	–	–	0.72

Table 4. Results (F1-scores) for all models and baselines on real-world data.

grounded examples, it may not entirely replicate the informal language, emerging trends, or unpredictability present in real-world streams.

Another limitation is the focus solely on English-language posts, which excludes the multilingual nature of crisis communication in many global contexts. Expanding to multilingual datasets is an important direction for enhancing the generalizability of our approach.

Additionally, while the taxonomy introduces valuable structure and granularity to the classification process, its evaluation has primarily been qualitative. A more systematic, quantitative assessment of the taxonomy's depth, completeness, and impact on classification performance would provide deeper insights and potentially inform refinements to better support crisis response applications.

For future work, we recommend expanding the real-world dataset, incorporating multilingual posts, and exploring alternative data generation strategies to improve robustness. Further, developing quantitative metrics to assess the taxonomy's design and conducting controlled experiments to evaluate its effect on classification outcomes will be key to enhancing its utility. Incorporating optimization techniques, such as fine-tuning or multi-agent frameworks, may also offer additional performance gains.

CONCLUSION

In this work, we introduced a fine-grained hierarchical taxonomy and a dynamic few-shot prompting technique to improve the detection of actionable requests and offers in social media posts during natural disasters. Our taxonomy organizes crisis-related information related to requests and offers into three core dimensions: *supplies*, *emergency personnel*, and *actions*. By leveraging the capabilities of Large Language Models, we demonstrated through extensive experiments that our approach significantly outperforms baseline prompting methods in accurately identifying and prioritizing actionable content. These contributions provide a valuable framework for enhancing the efficiency of humanitarian organizations in crisis management and rapid response. Future work will focus on expanding the approach to diverse disaster scenarios, integrating real-world data, and incorporating advancements in next-generation LLMs to further refine performance and adaptability.

REFERENCES

Basu, M., Bit, S. D., & Ghosh, S. (2022). Utilizing microblogs for optimized real-time resource allocation in post-disaster scenarios. *Social Network Analysis and Mining*, *12*, 1–20.

Basu, M., Ghosh, S., Jana, A., Bandyopadhyay, S., & Singh, R. (2017). Resource mapping during a natural disaster: A case study on the 2015 nepal earthquake. *International journal of disaster risk reduction*, *24*, 24–31.

Basu, M., Shandilya, A., Khosla, P., Ghosh, K., & Ghosh, S. (2019). Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations. *IEEE Transactions on Computational Social Systems*, *6*(3), 604–618.

Bleiker, C. (2015, April). Social media after nepal earthquake. https://www.dw.com/en/nepal-searching-for-missing-loved-ones-with-google/a-18411530

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*.

Devaraj, A., Murthy, D., & Dontula, A. (2020). Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction*, *51*, Art. no. 101757. https://doi.org/10.1016/j.ijdrr.2020.101757

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Durham, J., Chowdhury, S., & Alzarrad, A. (2023). Unveiling key themes and establishing a hierarchical taxonomy of disaster-related tweets: A text mining approach for enhanced emergency management planning. *Information*, *14*(7), 385. https://doi.org/10.3390/info14070385

Dutt, R., Basu, M., Ghosh, K., & Ghosh, S. (2019). Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities. *Information Processing & Management*, *56*(5), 1680–1697.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., & Ma, Z. (2024, July). The llama 3 herd of models. https://arxiv.org/abs/2407.21783

He, X., Lu, D., Margolin, D., Wang, M., Idrissi, S. E., & Lin, Y.-R. (2017). The signals and noise: Actionable information in improvised social media channels during a disaster. *Proceedings of the 2017 ACM on web science conference*, 33–42.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. https://doi.org/10.18653/v1/P18-1031

Imran, M., Ziaullah, A. W., Chen, K., & Ofli, F. (2025). Evaluating robustness of llms on crisis-related microblogs across events, information types, and linguistic features. *arXiv preprint arXiv:2412.10413*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b. https://arxiv.org/abs/2310.06825

Khot, T., Trivedi, H., Sabharwal, A., & Clark, P. (2023). Decomposed prompting: A modular approach for solving complex tasks. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*. https://arxiv.org/abs/2205.11916

Lamsal, R., Read, M., Karunasekera, S., & Imran, M. (2024). Crema: Crisis response through computational identification and matching of cross-lingual requests and offers shared on social media. *IEEE Transactions on Computational Social Systems*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mihunov, V. V., Lam, N. S. N., Zou, L., Wang, Z., & Wang, K. (2020). Use of twitter in disaster rescue: Lessons learned from hurricane harvey. *International Journal of Digital Earth*.

Nazer, T. H., Morstatter, F., Dani, H., & Liu, H. (2016). Finding requests in social media for disaster relief. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1410–1413. https://doi.org/10.1109/ASONAM.2016.7752424

OCHA. (2023a). Response planning and coordination. https://www.unocha.org

OCHA. (2023b). Turkey/Syria: Earthquakes - feb 2023. https://reliefweb.int/disaster/eq-2023-000015-tur

Olawale, S. (2018). Social media and crisis management: A review and analysis of existing studies. *LAÜ Sosyal Bilimler Dergisi*, *9*(2), 199–215.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Purohit, H., Castillo, C., Diaz, F., Sheth, A., & Meier, P. (2014). Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, *19*(1). https://doi.org/10.5210/fm.v19i1.4848

Seeberger, P., & Riedhammer, K. (2022). Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning. *Proceedings of the Workshop on Natural Language Processing for Positive Impact (NLP4PI)*, 70–78. https://aclanthology.org/2022.nlp4pi-1.9.pdf

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., … Andreev, A. (2024). Gemma 2: Improving open language models at a practical size. https://arxiv.org/abs/2408.00118

Ullah, I., Khan, S., Imran, M., & Lee, Y.-K. (2021). Rweetminer: Automatic identification and categorization of help requests on twitter during disasters. *Expert Systems with Applications*, *176*, 114787.

UNHCR. (2023). Emergency response and guidelines. https://www.unhcr.org

UNICEF. (2023). Humanitarian action reports and guidelines. https://www.unicef.org

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vitiugin, F., & Purohit, H. (2024). Multilingual serviceability model for detecting and ranking help requests on social media during disasters. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 1571–1584.

Zade, H., Shah, A., Imran, M., & Ostermann, F. O. (2018). From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–18.

Zhang, T., Shen, S., Cheng, C., Su, K., & Zhang, X. (2021). A topic model based framework for identifying the distribution of demand for relief supplies using social media data. *International Journal of Geographical Information Science*, *35*(11), 2216–2237.

Zhou, L., Wu, X., Xu, Z., & Fujita, H. (2022). VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. *Computers, Environment and Urban Systems*, *95*, 101824. https://doi.org/10.1016/j.compenvurbsys.2022.101824

Zhou, Y., Li, J., Xiang, Y., Yan, H., Gui, L., & He, Y. (2024). The mystery of in-context learning: A comprehensive survey on interpretation and analysis. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14365–14378.

Ziaullah, A. W., Ofli, F., & Imran, M. (2024). Monitoring critical infrastructure facilities during disasters using large language models. *Proceedings of the International ISCRAM Conference*. https://doi.org/https://doi.org/10.59297/755e8b64