Harnessing Prompt-based Large Language Models for Disaster Monitoring and Automated Reporting from Social Media Feedback

Riccardo Cantini[a], Cristian Cosentino[a], Fabrizio Marozzo[a,*], Domenico Talia[a], Paolo Trunfio[a]

[a]*DIMES, University of Calabria, Italy*

Abstract

In recent years, social media has emerged as one of the main platforms for real-time reporting of issues during disasters and catastrophic events. While great strides have been made in collecting such information, there remains an urgent need to improve user reports' automation, aggregation, and organization to streamline various tasks, including rescue operations, resource allocation, and communication with the press. This paper introduces an innovative methodology that leverages the power of prompt-based Large Language Models (LLMs) to strengthen disaster response and management. By analyzing large volumes of user-generated content, our methodology identifies issues reported by citizens who have experienced a disastrous event, such as damaged buildings, broken gas pipelines, and flooding. It also localizes all posts containing references to geographic information in the text, allowing for aggregation of posts that occurred nearby. By leveraging these localized citizen-reported issues, the methodology generates insightful reports full of essential information for emergency services, news agencies, and other interested parties. Extensive experimentation on large datasets validates the accuracy and efficiency of our methodology in classifying posts, detecting sub-events, and producing real-time reports. These findings highlight the practical value of prompt-based LLMs in disaster response, emphasizing their

*Corresponding author

Email addresses: rcantini@dimes.unical.it (Riccardo Cantini), ccosentino@dimes.unical.it (Cristian Cosentino), fmarozzo@dimes.unical.it (Fabrizio Marozzo), talia@dimes.unical.it (Domenico Talia), trunfio@dimes.unical.it (Paolo Trunfio)

flexibility and adaptability in delivering timely insights that support more effective interventions.

Keywords: Social media, Events detection, Natural disasters, Catastrophic events, Crisis computing, Disaster management, Mass emergencies

1. Introduction

Social media platforms have become essential tools for understanding human dynamics due to their widespread use [1]. Each post contains a large amount of information encompassing various aspects, including the discussed topic, the nature of the post (opinion-based or news-based), the sentiment and opinions conveyed, and more [2, 3]. When this vast amount of information is aggregated and analyzed at scale, it can provide valuable insights into business trends, consumer behaviors, and real-time events, offering a deeper understanding of emerging patterns and societal dynamics [4]. Consequently, in recent years, researchers and companies have increasingly utilized advanced machine learning techniques to extract knowledge from social media data, including trends, sentiments, and social behaviors in real-time.

Even within the study of disasters and catastrophic events, social media remains one of the primary sources of real-time updates on the analysis of these events [5]. While advanced techniques for classifying and aggregating social media content have been refined over the years, there is still an urgent need to enhance automation, aggregation, and organization of issues reported by citizens to streamline various tasks, including rescue operations, resource allocation, and communication with the press. Large Language Models (LLMs) can play a fundamental role in these tasks. Not only do they serve as powerful textual analysis tools for classifying and geolocating user posts, but they also act as information generators, improving data presentation and explanation. Their ability to understand linguistic context aids in fast decision-making without compromising accuracy, crucial for early detection of sub-events and automatic report generation during crises [6]. However, this approach also faces limitations, such as the challenge of defining specific filters to prevent irrelevant responses or hallucinations in critical situations [7].

In this paper, we propose an innovative methodology for gathering and analyzing citizen-reported issues following a catastrophic event, aiming to generate a comprehensive summary detailing the main issues reported. Ini-

tially, we systematically collect relevant posts from social media, focusing on those originating from the affected area of the disaster. Leveraging prompt-based LLMs (i.e., ChatGPT from OpenAI, Gemini from Google AI, Llama from Meta AI, and Command from Cohere), we classify and localize posts to identify and aggregate geolocated citizen feedback, including sub-events like water main breaks, structural damage, falling debris, and other issues affecting community life. These models were specifically chosen for their capability to leverage in-context learning and operate effectively in zero-shot and few-shot settings [8], where minimal labeled data is available. This feature is crucial for scenarios where rapid analysis is required, and comprehensive training phases may be infeasible. Compared to traditional models like BERT, which rely heavily on extensive training datasets, prompt-based models are more flexible and efficient for timely and adaptable response generation. Subsequently, we utilize these LLMs to generate a detailed and advanced summary. This automated process enables us to produce well-structured reports encompassing street-level details and broader geographic classifications, incorporating factual data sourced from localized citizen feedback, along with their sentiments and concerns. Comprehensive testing on extensive datasets, with variations in prompts and the use of zero- and few- shot approaches, has verified the precision and effectiveness of our approach in classifying and localizing posts, identifying user feedback, and generating informative reports. Such reports are crucial for facilitating rescue operations, enhancing situational awareness of damage in the area, and supporting timely and effective response efforts. Overall, our findings demonstrate the practical value of prompt-driven LLMs in disaster response, enabling faster and more informed interventions by equipping responders with detailed, automatically generated reports. This allows for timely action that can effectively mitigate the impact of disasters.

The structure of this paper is as follows. Section 2 discusses related work and compares our methodology with existing research. Section 3 describes the proposed methodology. Section 4 discusses the achieved results. Finally, Section 5 concludes the paper.

2. Related work

In recent years, social media has assumed an increasingly central role in rescue operations and disaster management, emerging as an indispensable tool in the timely resolution of critical situations. Indeed, their ubiquity

and widespread adoption can facilitate communication by reducing response times and improving the coordination of rescue operations. Nonetheless, information gathered from social media may not be easily exploited for providing immediate assistance to those affected by traumatic events and managing critical situations, posing the need for proper approaches to collect and organize it effectively. In the following sections, the state-of-the-art related to disaster management using social media data will be discussed, with a focus on the use of LLMs for extracting information from textual data. Furthermore, the main techniques presented in the literature for extracting geographical information from texts will be described, given the crucial importance of geographic information for aiding rescue activities.

2.1. Disaster management from social media data

Numerous studies have recently focused on leveraging social media to enhance the efficiency of organizing emergency response operations. These studies analyze the primary challenges associated with using social media data in disaster contexts, including the complexity of processing vast amounts of data promptly, the presence of unwanted or false information, and difficulties in collecting data that document various phases of a disaster. [9, 10, 3, 11]. Further investigations have explored the complexities related to analyzing social posts during large-scale emergencies, with a focus on various aspects such as coordinated management of evacuation operations [12], data integration from diverse sources, including satellite imagery [13], and an in-depth understanding of information dissemination dynamics during such events [14]. In this context, the integration of potential heterogeneous data sources - from social media to IoT-generated data - as well as the use of powerful big data analytics tools such as Spark and Kafka, can lead to an improvement in the efficiency and effectiveness of emergency management processes [15].

Natural disasters such as earthquakes have attracted significant attention among critical situations addressed by disaster management-related studies in the literature. For instance, the EARS system [16] analyzes data streams from Twitter to detect seismic events and assess their impact on people and infrastructure. Additionally, initiatives like LastQuake [17], developed in collaboration with the European Mediterranean Seismological Center (EMSC), focus on providing visual information about perceived seismic events and gathering user feedback on the main shocks. Other studies centered on the prediction and management of urban floods [18, 19], which affect urban or densely populated areas. These floods can be caused by various fac-

tors, including heavy rainfall, ineffective stormwater drainage, malfunctioning drainage infrastructure, or natural events such as sea level rise. These events can cause significant property damage, and disruptions in transportation and daily activities, as well as pose a risk to people's safety.

A key concept related to disaster management is represented by the *sub-event*, i.e., a specific event, related to a broader critical situation such as an earthquake or a hurricane, which occurs in a specific location. Examples are building or bridge collapses, gas pipe ruptures, and floods. Several investigations have explored the discovery of sub-events through social media data, employing diverse approaches ranging from supervised to unsupervised techniques. Most supervised techniques rely on weighted graphs [20] and neural networks to identify, classify, and summarize sub-events in social media data [21, 22, 23]. Although they may yield satisfactory results, these methods typically require large amounts of labeled data, which is often costly and time-consuming to produce, involving expert annotation to ensure quality. This labeling process is particularly challenging in rapidly evolving environments like social media, where maintaining updated datasets is difficult, hindering the effectiveness of the obtained results. For this reason, several studies have turned their attention to unsupervised sub-event detection approaches. Most approaches in this category use clustering algorithms applied to social media data, leveraging textual features, extracted from text and hashtags, and classical measures such as the cosine similarity [23]. These methods can automatically group similar content without requiring predefined labels, making them particularly useful for discovering emerging or unforeseen sub-events. However, while unsupervised techniques can reduce the burden of manual labeling, they often require fine-tuning of parameters and may still present difficulties in accurately capturing different types of sub-events, especially when the data is noisy or ambiguous.

A different class of approaches relies on topic modeling, employing classical algorithms like LDA (Latent Dirichlet Allocation) and HDP (Hierarchical Dirichlet Processes), which extract underlying events by analyzing the semantic representations of documents [24]. Another significant contribution is the SEDOM-DD methodology [25], which focuses on detecting sub-events as consequences of a disaster, using a spatial clustering algorithm to identify specific geographic areas involved.

A complementary perspective on leveraging social media data in disaster management is offered by Lei et al. [26], who examined the potential of Online Social Networks (OSNs) for environmental monitoring. Their work high-

lights how social media posts can serve as a form of human-centric sensing, offering insights into environmental conditions that complement traditional sensor data. Although OSNs may sometimes reflect subjective perceptions rather than objective measurements, they provide valuable information about public awareness and sensitivity to environmental phenomena, enriching the understanding of how people experience and respond to such events.

2.2. Using Large Language Models for information extraction

With the advent of AI-powered conversational agents enhanced by the latest advanced LLMs, the extraction of insights and knowledge from data has significantly improved in efficiency and usability. Endowed with natural language processing (NLP) and machine learning capabilities, these agents, commonly referred to as chatbots, naturally interpret user requests through prompts, generating textual responses carrying relevant information and insights.

Large Language Models (LLMs) can be broadly categorized into two main types: decoder-only models, such as ChatGPT, which are primarily generative and prompt-based, and encoder-only models, like BERT, which are typically used for tasks such as text classification and information retrieval [27]. Numerous studies have recently leveraged the capabilities and flexibility of LLMs in various tasks like question-answering and report generation by utilizing techniques like fine-tuning and in-context learning. *Fine-tuning* involves adjusting a model on a smaller, task-specific dataset to improve performance in a particular domain. In-context learning, on the other hand, enables models to utilize information directly from prompts, allowing them to adapt to new tasks without additional training. Prompt-based *zero- and few-shot* approaches, which are a form of *in-context* learning, enable models to perform tasks with minimal examples or even no examples during training, relying instead on carefully crafted prompts that guide the model's output [8]. Additionally, varying prompt structures and formulations can significantly impact the model's effectiveness, allowing for adaptation to different contexts and task requirements. While fine-tuning provides a compelling alternative to complex prompt engineering and few-shot learning, by allowing models to achieve high precision with less reliance on intricate prompting, in-context learning remains crucial for analyzing continually changing data. A hybrid approach, where the model is initially fine-tuned on accumulated, domain-specific data, and then adapted through in-context learning for dynamic, task-specific data, allows it to effectively handle applications involving both

stable historical data and evolving information. Furthermore, *Retrieval-Augmented Generation* (*RAG*) has emerged as a powerful technique that enhances the relevance and accuracy of LLM responses by retrieving pertinent information from external knowledge sources, which may be either static or dynamically updated [28]. This approach enables RAG to generate responses that are informed by specific sources, extending the model's capabilities beyond its predefined training data.

In the literature, Gilson et al. [29] assessed the performance of ChatGPT in answering questions related to the United States Medical Licensing Examination (USMLE) Step 1 and 2, finding comparable performance to that of a third-year medical student. Guo et al. [30] curated the Human ChatGPT Comparison Corpus (HC3), intending to compare responses generated by ChatGPT with those provided by humans in various sectors such as finance, medicine, and psychology. Bang et al. [31] evaluated the effectiveness of ChatGPT compared to other large language models, used in a zero-shot fashion, and fine-tuned models for various NLP tasks, also evaluating reasoning abilities and hallucination issues. Authors in [32] highlighted the importance of few-shot learning in utilizing LLMs, particularly gpt-3.5, to proactively rephrase potential hate speech.

Focusing more specifically on information extraction from large datasets for report generation, Messina et al. [33] proposed a deep learning-based approach for the automatic generation of reports from medical images. The ability to compose a brief report on an X-ray was examined by combining deep learning algorithms for image analysis and natural language processing techniques for report writing. In Wang et al. [34], controlled text generation from tables is addressed, aiming at creating natural language descriptions for highlighted sections of a table, being robust to changes in table layout. Regarding RAG, this approach is particularly valuable in domains where factual accuracy is crucial, such as disaster response and medical reporting, where LLMs must produce reliable and contextually accurate outputs. For example, studies on emergency triage improvement have demonstrated that RAG-enhanced LLMs can significantly improve the precision and efficiency of emergency response processes [35].

As LLMs become increasingly used in natural language processing and understanding tasks, effective metrics are essential for their fair and accurate evaluation. Recently, the concept of *LLM-as-a-Judge* has emerged, where LLMs serve as evaluators to dynamically assess artifacts generated by these models. Such an approach can better align with human preferences and

user needs, but also poses challenges related to potential biases and internal inconsistencies [36, 37, 38]. In this context, traditional metrics such as BLEU and ROUGE remain valuable, yet these new self-evaluation techniques allow for a more nuanced assessment of output quality and consistency, enhancing model effectiveness in complex application scenarios [39, 40].

2.3. Automatic geographic location identification from text

In recent years, the widespread use of social media during disasters has witnessed requests for help and the sharing of information [41, 42, 43]. A key aspect in effectively utilizing social media posts for enhancing disaster management is the extraction of the related location, which allows for precisely targeted interventions [44, 45]. The extraction of geographical information from the information present in the text can be accurately carried out by leveraging current LLMs capabilities, which can address the lack of geotagging and location metadata. Indeed, previous studies have primarily focused on geotagged locations, i.e., those tagged in tweets [46, 47], overlooking locations described within the content of tweets. Specifically, social users may provide descriptions of locations within the content of posts, without necessarily using geotags. Moreover, the current location of a Twitter user may not necessarily correspond to that of the victim. Therefore, it becomes essential to extract locations described within the content of social media messages.

Previous studies have tackled the extraction of locations from the content of social media messages by treating locations as specially named entities. To identify positions in tweets, researchers have employed pre-trained Named Entity Recognition (NER) tools, such as Stanford NER and SpaCy NER [48, 49]. With the advancements in deep learning, the NeuroTPR model emerged as a refinement of a Bidirectional Long Short-Term Memory (BiLSTM) architecture for extracting locations from social media messages [50]. Recently, new approaches leveraging transformers like BERT have been also introduced [42, 51] further enhancing geolocation performance. Besides NER-based tools and models, other works are present in the literature, focusing on directly detecting geospatial descriptions in natural language. As an example, in [52] the authors fused geo-knowledge of location descriptions and a Generative Pre-trained Transformer (GPT), thus obtaining a geo-knowledge-guided GTP model that can accurately extract location descriptions from disaster-related social media posts. In addition, Suwaileh et al. proposed a BERT-based model for Location Mention Recognition (LMR) from social media posts related to a just-occurred disaster, experimenting

with zero- and few-shot settings, as well as mono-, cross-, and multilingual scenarios [42].

2.4. Contributions of our work

The proposed methodology employs prompt-based LLMs to enhance disaster response by efficiently identifying user-reported issues via in-context learning, also enabling precise geolocation. This is crucial for rapid response and optimal resource allocation during critical situations to prioritize emerging issues, aiding timely interventions for hazardous conditions and infrastructure problems. Additionally, our approach underscores the importance of data aggregation and synthesis, furnishing stakeholders across various sectors with comprehensive insights into the nature, magnitude, and spatio-temporal distribution of critical events. This information equips emergency services with the necessary insights to devise well-informed response strategies while aiding news agencies and other stakeholders in effectively addressing social challenges.

The contributions of this research can be summarized as follows: (*i*) we conduct a comprehensive examination of different LLMs utilized in fine-tuning, zero-shot, and few-shot modes for various tasks such as classification and geographic information extraction. We compare the transfer learning abilities of approaches like fine-tuning and in-context learning, discussing their advantages and disadvantages; (*ii*) we introduce an approach to geolocating posts that leverages the capabilities of LLMs in the Named Entity Recognition (NER) task to enhance the localization process conducted in zero-shot mode; (*iii*) we produce structured and automated reports that summarize user reports gathered from social channels regarding problems or requests for assistance arising during natural disasters.

3. Methodology

The proposed methodology aims to collect issues reported by citizens from social media platforms for enhancing disaster management and facilitating targeted intervention. Specifically, user feedback in the aftermath of a catastrophic event is utilized to generate a detailed report that summarizes the main issues and critical situations, thus enabling a user-centric data-driven approach. The classification of user posts, the identification of user feedback (sub-events), and the generation of the report are all accomplished by exploiting prompt-based LLMs, such as ChatGPT from OpenAI, Gemini from

Google AI, and Llama from Meta AI. The proposed methodology, whose execution flow is depicted in Figure 1, comprises three distinct phases:

1. Disaster-related posts are gathered from social media.
2. Retrieved posts are classified and geolocated using prompt-based LLMs, to identify issues reported by citizens.
3. Classified geo-located posts are leveraged to write an information-rich report through prompt-based LLMs.

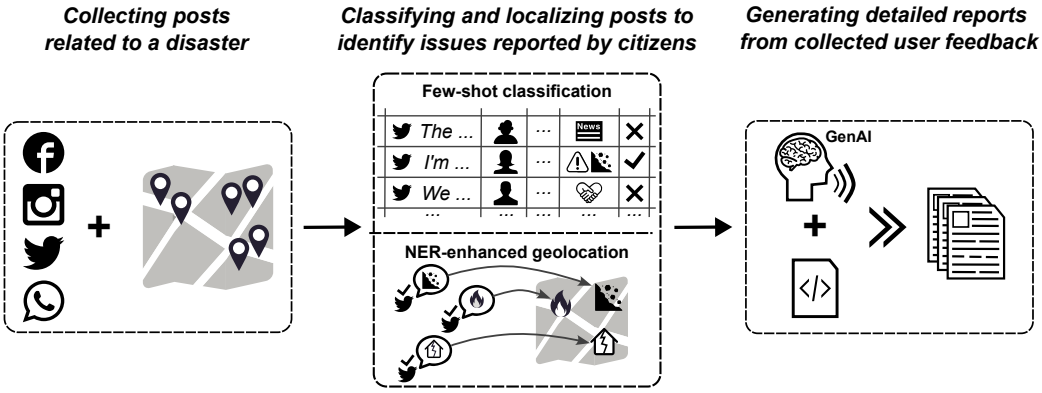


Figure 1: Execution flow of the proposed methodology, encompassing three main steps: (*i*) disaster-related posts are gathered from social media; (*ii*) retrieved posts are classified and geolocated using LLMs; (*iii*) processed posts are used to write an informative report.

The initial phase of our methodology — *collecting posts related to a disaster* — involves systematically gathering relevant social media posts related to a catastrophic event. This process begins by aggregating all posts originating from the affected area. We initiate this extraction by conducting targeted searches on social media platforms, employing keywords or geographical metadata associated with the disaster. The significance of collected posts is further underscored by their connection to users residing in the affected region. To ensure the dataset's appropriateness for subsequent phases, we apply a filtering mechanism to select only relevant posts for the event under analysis. This preliminary phase facilitates a focused and efficient analysis in the subsequent stages of classification, localization, and report generation.

In the second phase of our methodology — *classifying and localizing posts to identify issues reported by citizens* — we categorize posts to discern the diverse problems and issues highlighted by users. Leveraging the capabilities of LLMs, posts are classified into various categories including cautionary

warnings, expressions of sympathy and support, requests for urgent assistance, observations on infrastructure, and utility issues. The classification is performed by crafting a specific prompt to guide a language model in generating the desired output. An example prompt is displayed below, which may come with the sole description of the task (zero-shot) or include examples in the LLM's context that can aid classification (few-shot).

"Classify each post into one of the following categories based on their description: [$label_1$*:* $description_1$*,* $label_2$*:* $description_2$*,...,* $label_n$*:* $description_n$]. For each class, here there are some correct classification examples: + [$text_1$*:* $class_1$*,* $text_2$*:* $class_1$*,...,* $text_k$*:* $class_1$] + ... + [$text_1$*:* $class_n$*,* $text_2$*:* $class_n$*,...,* $text_k$*:* $class_n$]. Provide only the post_id and the classification label separated by a comma without any explanation: [$post_1$*,* $post_2$*,... ,* $post_m$]".

In terms of geolocating textual posts lacking explicit geographical metadata, we employed a geolocation approach enhanced by Named Entity Recognition (NER). By conducting NER analysis on the textual content of a post, we can identify a wide array of information, isolating that pertaining to *locations* within the text. Geographical-related information is then provided to a LLM, prompting it to solve the Location Mention Recognition (LMR) task. This process enables us to reconstruct the location (or locations) with varying levels of detail, such as street, neighborhood, city, region, and state, if deductible. The overall goal is to identify and aggregate all geolocated posts expressing citizen feedback, including sub-events such as water main breaks, structural damage to bridges, cases of falling debris, and similar issues impacting the community. An example prompt is displayed below:

"Considering the following tweets, extract the following geographical information: state, zip code, city, and other geographical information (e.g., street and/or district). You have to consider only the text of the tweets, and their associated location entities, extracted through Named Entity recognition. Provide only the values of the tweet id and the required information separated by a comma without any explanation: [$tweet_1$*,* $tweet_2$*, ...,* $tweet_m$], [loc_1*,* loc_2*, ...,* loc_m]".

In Table 1, four examples of user posts related to Hurricane Harvey, which made landfall in Texas and Louisiana in August 2017, are presented. Each post is classified, and the presence of geographic information is indicated.

These examples are provided as generic illustrations to demonstrate how the methodology processes and enriches typical disaster-related posts.

ID	Post	Class	Location
1	News Update: Buffalo Bayou water levels rising in Houston, TX. Please stay informed and safe.	News from a website	Houston, Texas (TX)
2	Enormous gratitude to the Red Cross volunteers and everyone for their incredible support.	Support message	-
3	Unable to reach our grandmother on Cactus Street in Fulton. Can anyone local provide information on the situation there?	Help request	Cactus St., Fulton, Texas (TX)
4	Just in: People reportedly trapped inside a collapsed building in Rockport, TX. Please stay safe.	User-reported issue	Rockport, Texas (TX)

Table 1: Examples of four user posts with their classification and identified locations.

The first post features a newspaper story, while the second solicits donations and support for citizens affected by the disaster. The third post requests assistance, and the fourth reports a problem. Geographic information is successfully extracted from the first, third, and fourth posts. Our methodology focuses on posts similar to the last one, which both report problems and provide crucial location details, aggregating them to generate comprehensive insights about the critical event under consideration.

In the final phase of our methodology - *generating detailed reports from collected user feedback* - we streamline the process of creating comprehensive reports by leveraging prompt-based LLMs through their API interface. This approach exploits the power of LLMs to automate the generation of structured reports, enhancing both efficiency and accuracy. Through the API, programmers can seamlessly interact with these models, providing specific data and queries to guide the automated production of informative reports. This process ensures a clear and organized presentation, avoiding the spread of scattered data that may be difficult to interpret. The resulting detailed report is marked by its thoroughness, summarizing all user reports while offering the flexibility to aggregate information at various levels of detail, from street-level specifics to broader geographic classifications like neighborhoods, cities, regions, and states. In addition to factual data, the report captures users' emotions and concerns, providing a nuanced understanding of the emotional landscape following the disaster. This aggregate analysis not only provides a comprehensive summary but also extracts higher-level

insights from a wide range of specific details. By employing generative language models in this phase, our methodology ensures the prompt delivery of easily understandable text, enabling informed decision-making and response strategies based on a comprehensive understanding of the critical situation on the ground.

4. Experimental Results

As outlined in Section 3, this work pursues a three-fold objective: (i) identify all issues reported by citizens on social media concerning a disaster event; (ii) geolocate non-geotagged posts containing in-text geographical information (e.g. street, and city name); and (iii) generate comprehensive reports detailing spatiotemporal information about the multitude of sub-issues that have emerged, offering a comprehensive understanding of the impact of the disaster event.

The experimental evaluation is organized as follows. In Section 4.1, we elaborate on the dataset used in our experiments and the preprocessing steps applied to it. Subsequently, in Section 4.2 and 4.3, we detail the performance of prompt-based LLMs in classifying posts based on a standardized classification scale encompassing nine classes [53], including *caution_and_advice*, *sympathy_and_support*, and others, utilizing different prompts, and employing both zero-shot and few-shot approaches. Moving forward, Section 4.4 describes how we binarized classification to discern citizen-reported issues from other online content related to the event under consideration, while also comparing the performance with other types of LLMs, such as BERT models. Following this, Section 4.5 examines the capability of LLMs in identifying geolocation information within texts, assessing the benefits of the proposed integration of NER-based information into the zero-shot geolocation process. Lastly, in Section 4.6, we show the process of generating information reports, focusing on the prompts utilized and the resultant outcomes.

4.1. Data Collection and Preprocessing

In the field of disaster research, several datasets containing posts written by users on social platforms about catastrophic events have been published over the years. Some of these events, such as hurricanes, are predictable and can be easily identified and followed on social media platforms using specific keywords or hashtags. Other events, such as earthquakes, are unpredictable and are often tracked using systems that continuously monitor

generic keywords or hashtags (e.g., earthquake or #earthquake). In our study, we opted to utilize the HumAID (Human-Annotated Disaster Incidents Data) dataset [53]. This repository comprises over 77,000 labeled tweets extracted from a pool of 24 million tweets generated during 19 major real-life disasters spanning from 2016 to 2019, encompassing hurricanes, earthquakes, fires, and floods. The classified tweets were categorized into 11 different labels, that are:

1. *caution_and_advice*: notices issued or revoked;
2. *sympathy_and_support*: tweets containing prayers, thoughts, and emotional support;
3. *requests_or_urgent_needs*: reports of urgent needs or supplies such as food, water, clothing, money, medicine, or blood;
4. *infrastructure_and_utility_damage*: reports of damage to buildings, roads, bridges, power lines, communication poles, or vehicles;
5. *rescue_volunteering_or_donation_effort*: reports of any rescue, volunteering, or donation efforts, including safe transport, evacuation, medical or food assistance, shelters, monetary or service donations, etc.;
6. *not_humanitarian*: if the tweet does not convey information related to humanitarian aid;
7. *displaced_people_and_evacuations*: people who have changed residence due to the crisis, even temporarily (including evacuations);
8. *injured_or_dead_people*: reports of people injured or killed as a result of the disaster;
9. *missing_or_found_people*: reports of missing or found people after a catastrophic event.
10. *dont_know_cant_judge*: where there's insufficient information to decide.
11. *other_relevant_information*: details that do not fit in the other classes, but are still relevant.

This extensive set of classified data is particularly interesting as it likely comprises posts authored by users in locations affected during one of the disasters under consideration [53]. This allows the collection of valuable information from eyewitnesses or individuals directly involved in critical circumstances. In our analysis, we excluded data classified as *dont_know_cant_judge* and *other_relevant_information*, considered outliers, reducing the classification problem from 11 classes to 9. Moreover, since the classified dataset is

unbalanced, we created a balanced dataset of tweets with 200 instances for each class (1,800 in total) using random undersampling with deduplication.

As anticipated earlier, besides the multi-class setting, we also considered a binary version of the classification problem, which is useful for distinguishing between user-reported issues and other content irrelevant to our analysis. To this aim, we aggregated posts into two general categories: citizen-reported issues (*sub-events*) and other content. In the first category we include tweets labeled as *infrastructure_and_utility_damage*, *displaced_people_and_evacuations*, *injured_or_dead_people*, or *missing_or_found_people*. The remaining five classes are grouped into the second category. Also in this case, we balanced the dataset, reaching 1,600 instances (800 instances for the sub-events and 800 for other contents).

It is essential to emphasize that preprocessing data before analysis is crucial, as highlighted in previous research on social media analysis. These studies underscore the importance of cleaning and filtering posts, removing irrelevant or noisy content, and retaining only the most relevant and reliable information. For example, Belcastro et al. [54] considered only posts with a clear indication of the user's residence and statistically validated the collected sample with official data, while Cantini et al. [55] filtered out posts generated by bots or automated accounts to improve result quality. In Shu et al. [56], detecting false information in the dataset was critical to achieving accurate analysis results, and Aman et al. [57] trained a large language model-based algorithm to detect disinformation.

In the field of disaster monitoring, handling noisy and ambiguous documents, integrating information from multiple sources, discarding false information, and ensuring accurate geolocation are some of the critical factors to be addressed for effective disaster management. For example, robust filtering mechanisms are essential to prevent the inclusion of irrelevant, misleading, or false information. When dealing with noisy and ambiguous documents, it is important to employ advanced filtering and text-cleaning techniques to refine the data before analysis, ensuring that only relevant and clear information is retained. For integrating information from multiple sources, techniques such as data fusion and cross-referencing should be employed to combine data from various inputs, enhancing the reliability and completeness of the information. When discarding false information, it is necessary to filter out such data as early as possible, or at least partially if complete filtering is not feasible, to make LLMs more robust and capable of effectively detecting and handling misinformation, especially in accuracy-critical scenarios. Ensur-

ing accurate geolocation involves utilizing named entity recognition (NER) and spatial analysis techniques to correctly identify locations mentioned in posts, while filtering out incorrect or ambiguous references that could lead to misinterpretation of the situation on the ground. However, in the case study analyzed here, these preprocessing steps, while generally important, were less critical because the dataset we used is manually classified and thus results are ready to be analyzed.

4.2. Zero-shot classification using LLMs

In our preliminary experiments, we utilized prompt-based LLMs in zero-shot mode to classify tweets into the nine different predefined classes outlined in the previous section. Prompt-based LLMs offer greater flexibility and adaptability compared to traditional models like BERT, as they rely on carefully crafted prompts rather than extensive task-specific training. In this context, zero-shot mode enables these prompt-based LLMs to effectively tackle tasks they have not encountered previously, without requiring specific training for those tasks [58].

We began by utilizing Chat-GPT (OpenAI) in its *gpt-3.5-turbo* version via API, while varying both the prompt and the temperature value (ranging from 0 to 1). Prompt engineering, involving adjustments in both the content and structure of the prompt, allows us to explore the nuances of language model responses and optimize performance. The temperature value affects the randomness and diversity of responses, with higher values resulting in more varied outputs, while lower values lead to more focused responses. The initial prompt (`Prompt_1`) follows a simple structure, serving as our baseline:

> `Prompt_1`: "Classify each tweet as [$label_1$, $label_2$,... , $label_n$]. Each tweet is identified by its own tweet_id. Provide only the tweet_id and the classification label separated by a comma without any explanation: [$tweet_1$, $tweet_2$,... , $tweet_m$]".

In our definition, the term [$label_1$, $label_2$,..., $label_n$] denotes the set of nine labels for annotation by ChatGPT. We explicitly instructed ChatGPT to refrain from including explanations to expedite label generation without additional context, thereby reducing both processing time and annotation costs. To reduce the volume of requests made to ChatGPT, we provided a set of tweets as input, generating in output a list of tweet_ids along with their labels. It is worth noticing that the number of tweets that can be analyzed

in each prompt depends on the maximum context length of the model itself, defining its capacity in terms of words. Typically, ChatGPT models have a maximum token limit of approximately 4,096 tokens for gpt-3.5 models.

In the second prompt (`Prompt_2`), our goal was to broaden ChatGPT's understanding of the context associated with each label by providing detailed and clear descriptions for each of them.

> `Prompt_2`: "Classify each tweet into one of the following categories based on their description: [$label_1$*:* $description_1$*,* $label_2$*:* $description_2$*,...,* $label_n$*:* $description_n$]. Each tweet is identified by its own tweet_id. Provide only the tweet_id and the classification label separated by a comma without any explanation: [$tweet_1$*,* $tweet_2$*,... ,* $tweet_m$]".

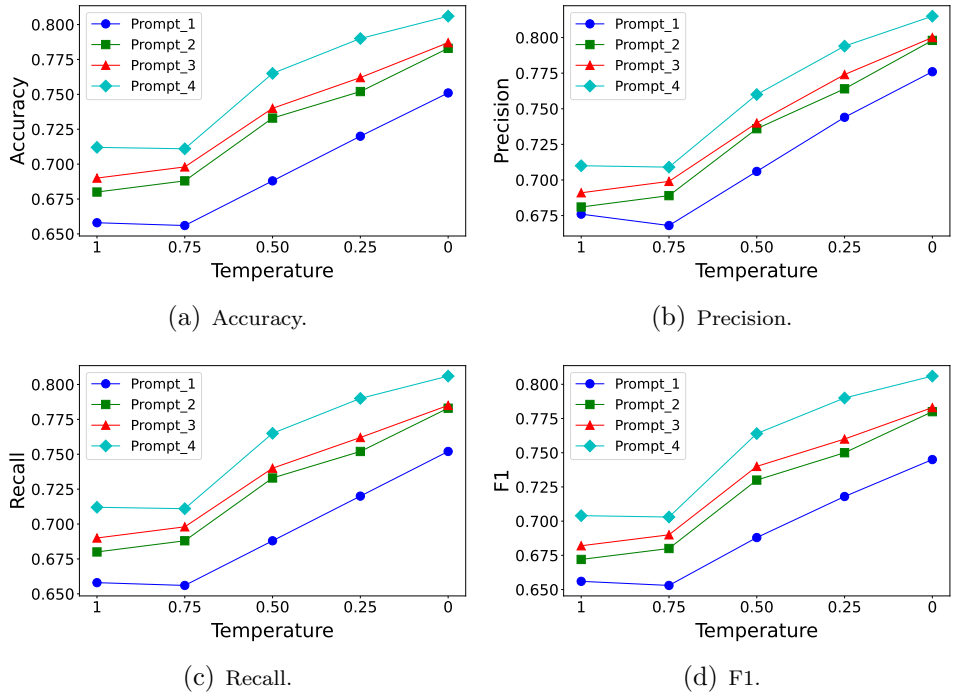
In `Prompt_3`, we assigned a specific role to ChatGPT, directing it to leverage its expertise in data annotation and categorization, and then instructing it about the task to be performed by using `Prompt_2`. By providing a clear role-playing context, this approach offers ChatGPT clear direction and purpose in the annotation process, aligning with the decision-making approach commonly used by human annotators.

> `Prompt_3`: "Act as a data annotator. You will apply your expertise in data annotation and labeling to analyze and categorize complex datasets, ensuring accurate and meaningful annotations for training machine learning models. Your role will involve understanding the specific annotation requirements, utilizing annotation tools and techniques to annotate data points, collaborating with domain experts to clarify ambiguous cases, and delivering high-quality annotated datasets. " + [*Prompt_2*].

In the latest variation of the prompt (`Prompt_4`), we employ a similar role description mechanism as in `Prompt_3`, albeit in a more concise and straightforward format.

> `Prompt_4`: "Act as a helpful data annotator. As a data annotator, your role is to provide precise and accurate labels for the given data. " + [*Prompt_2*].

Figure 2 analyzes the results obtained by applying all the described prompts, also exploring a range of temperature values from 0 to 1, with increments of 0.25. In particular, `Prompt_2` outperformed `Prompt_1` by offering



(a) Accuracy. (b) Precision.

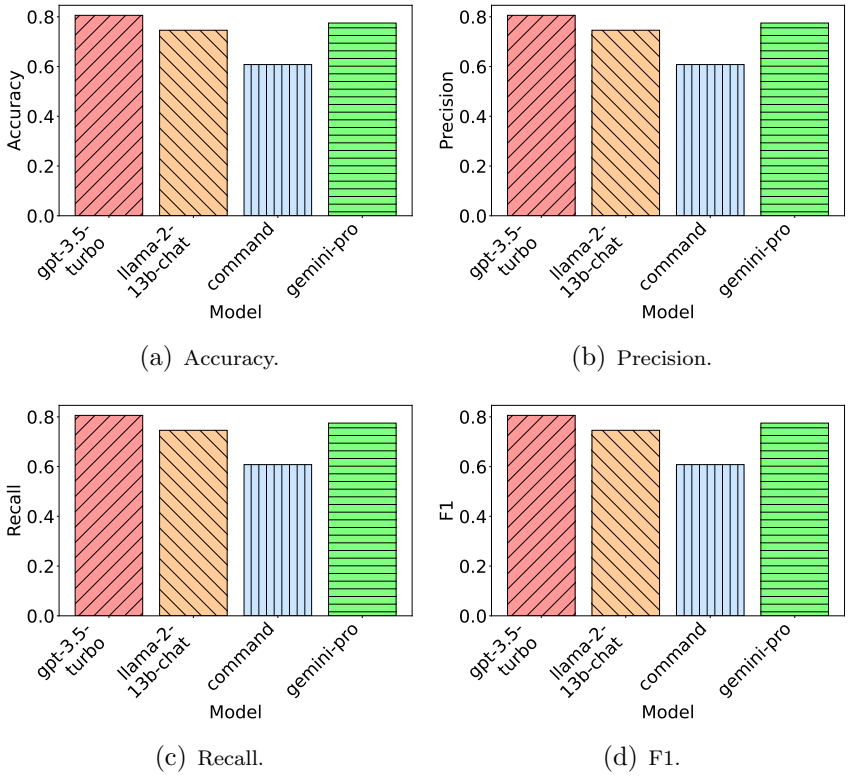
(c) Recall. (d) F1.

Figure 2: Impact of different temperature values and prompt templates on zero-shot classification performance.

detailed and clear descriptions for each label, thus providing valuable context associated with each classification. In addition, `Prompt_3` and `Prompt_4`, which provide further context to ChatGPT by assigning it a specific role (i.e., *data annotator*) led to increased performance. In particular, between these two, `Prompt_4` emerged as the best option, due to the provision of a more concise and clear role to ChatGPT, avoiding unnecessary noisy information.

A consistent trend is observed regarding temperature: as the temperature decreases, there is a corresponding increase in the values for all scores. Specifically, between 1 and 0.75, there is a relatively flat behavior, whereas all scores increase as the temperature decreases further in all the prompts analyzed. This suggests that higher variability (higher temperature) corresponds to lower scores in such a classification setting, while stronger determinism (lower temperature) leads to better performance.

After determining the optimal prompt and temperature settings (i.e., `Prompt_4` and temperature equal to 0), we compared the performance achieved



(a) Accuracy. (b) Precision.

(c) Recall. (d) F1.

Figure 3: Zero-shot classification performance using various prompts and temperatures.

with ChatGPT against three other alternative prompt-based LLMs, specifically *llama-2-13b-chat* by Meta AI, *command* by Cohere, and *gemini-pro* by Google AI. The results of this comparison are shown in Figure 3.

Remarkably, the *gpt-3.5-turbo* model outperforms other models in all experiments. This superiority is due to its advanced capabilities in understanding and processing natural language, as well as differences in inner structure, and fine-tuning or optimization strategies. However, this performance comes at the expense of a large number of parameters, leading to significant energy and resource consumption. Therefore, the choice of the right LLMs depends on balancing performance needs with resource constraints in the specific application scenario, also considering additional aspects such as model openness and ease of use.

4.3. Few-shot classification using LLMs

After the zero-shot phase, following the principles of few-shot prompting [59], we strengthened LLMs' input by incorporating annotated examples, aiming to improve their understanding of the task and generalization abilities. Specifically, we constructed our prompts by including k examples of tweets for classification. In these experiments, we utilized `Prompt_4` as it yielded the best results in zero-shot mode, also fixing the temperature value at 0. In our experiments, we kept the number of examples relatively low, by varying k between 1, 3, 5 and 8, to avoid overly complex requests while ensuring better model adaptability across diverse context window sizes.

`Prompt` k`-shot`: [*Prompt_4*] + "For each class, in the following there are some correct classification examples:" + [$text_1$*:* $class_1$, $text_2$*:* $class_1$,..., $text_k$*:* $class_1$] + ... + [$text_1$*:* $class_n$, $text_2$*:* $class_n$,..., $text_k$*:* $class_n$].

The prompt is refined by incorporating a variable number of examples, contingent on the value of k, which is determined by the number of shots considered for each class. Following this, the prompt provides the LLM with k times the number of example classes. Table 2 illustrates the results, demonstrating the effect of increasing examples on the final classification. Overall, there's a consistent rise in all scores as the number of examples considered for each class increases. However, this increase gradually flattens until it almost stabilizes when 8 shots are used.

Model	Num. of shots	Prompt	Temp.	Accuracy	Precision	Recall	F1
gpt-3.5-turbo	0	Prompt4	0	0.806	0.815	0.806	0.806
gpt-3.5-turbo	1	Prompt4	0	0.809	0.820	0.809	0.811
gpt-3.5-turbo	3	Prompt4	0	0.810	0.819	0.810	0.811
gpt-3.5-turbo	5	Prompt4	0	0.812	0.827	0.812	0.813
gpt-3.5-turbo	8	Prompt4	0	0.814	0.832	0.814	0.814

Table 2: Performance achieved by gpt-3.5-turbo in the k-shot classification task.

Similarly to the approach used for models in zero-shot mode, we compared the performance achieved by *gpt-3.5-turbo* with other prompt-based models such as *llama-2-13b-chat*, *command*, and *gemini-pro*. Notably, as shown in Table 3, the *gpt-3.5-turbo* model outperformed all other LLMs across all considered metrics.

Model type	Model	Number of shots	Accuracy	Precision	Recall	F1	Trainable parameters	Training time
Decoder-based	command	8	0.618	0.626	0.618	0.612	-	-
	llama-2-13b-chat	8	0.750	0.751	0.750	0.749	-	-
	gemini-pro	8	0.755	0.758	0.755	0.755	-	-
	gpt-3.5-turbo	**8**	**0.814**	**0.822**	**0.814**	**0.814**	-	-
Encoder-based	albert	-	0.781	0.822	0.781	0.774	11,684,353	871 s
	bert	-	0.827	0.851	0.827	0.822	109,483,009	994 s
	bertweet	-	0.848	0.859	0.848	0.848	134,900,737	938 s
	distilbert	-	0.832	0.849	0.832	0.830	66,363,649	456 s
	roberta	-	0.855	0.864	0.855	0.853	124,646,401	1025 s

Table 3: Performance comparison between encoder- and decoder-based models in the k-shot classification task. All encoder-based models are used in their *base-uncased* version.

In addition to the prompt-based LLMs considered, which are primarily *decoder-based* models, we also evaluated *encoder-based* models that do not accept prompts for direct interaction. This decision was made to explore the potential benefits and trade-offs associated with different architectural designs and utilization approaches. Both categories derive from the Transformer architecture [60]: the former comprises GPT-like autoregressive models optimized for causal language modeling tasks, while the latter encompasses BERT-like models that leverage blocks of transformer encoders to generate semantically rich representations of the input in a latent space. While decoder-based models, such as those used in our approach, support k-shot prompting without requiring additional training, encoder-based models typically need fine-tuning for specific downstream tasks. To effectively compare these models, we established a training set comprising 20% of the data to fine-tune the encoder-based models.

This comparative analysis revealed an improvement in model performance with fine-tuning encoder-based models compared to k-shot prompting with decoder-based ones. However, it is important to recognize that fine-tuning involves a more substantial time commitment than the immediate use of prompt-based models. Additionally, it requires considerable resources and energy for parameter updates, as well as the collection of a high-quality representative dataset for model adaptation. Conversely, the immediacy of in-context learning with prompt-based models make them a good option for quickly adapting to contextual nuances from a limited set of examples, allowing for commendable levels of accuracy relevant to our analytical efforts in a shorter amount of time.

4.4. Sub-Event classification using LLMs

Here we evaluate the performance of LLMs in distinguishing between user-reported issues (i.e., *sub-events*) and *other content*, in a binary classification task. As mentioned in Section 4.1, the *sub-events* class includes tweets labeled as *infrastructure_and_utility_damage*, *displaced_people_and_evacuations*, *injured_or_dead_people*, or *missing_or_found_people*, while all tweets belonging to different classes are grouped into the *other content* category.

The *sub-events* class therefore encompasses infrastructure and utility damage, including reports of flooded roads, damaged buildings, or disrupted power lines, signifying the extent of physical destruction and potential hazards within affected areas. Displaced people and evacuations highlight the urgent need for relocating residents from hazardous zones to safety, often accompanied by the establishment of emergency shelters and evacuation orders. Reports of injured or deceased individuals underscore the human toll of disasters, necessitating prompt medical response and aid distribution. Similarly, the identification of missing or found individuals, such as those separated from their families or rescued from dangerous conditions, emphasizes the importance of search and rescue efforts and community support networks.

In Table 4, the results obtained for the binary classification task are presented. Even in this case, we compared decoder-based models with encoder-based ones, fine-tuned on the binary task under consideration. As seen before, *gpt-3.5-turbo* turns out to be the best decoder-based model among those tested, with a fair improvement compared to the results obtained in the nine-class problem. Encoder-based models were always found to be slightly superior to decoder-based ones but required specific training data and time. However, as highlighted previously, it is not always possible to have a sufficient amount of data and resources to conduct a robust training phase. Therefore, despite the *encoder-based* models boasting higher accuracy, it might be preferable to adopt the best *decoder-based* model. This choice is driven by the need for quicker and more timely utilization, especially in scenarios where the amount of available data may be limited, and the training phase could be compromised.

4.5. Geographic Location Identification using LLMs

In addition to determining whether a post refers to sub-events that occurred during a disaster event, it is also crucial to geolocate posts that lack explicit geolocation metadata, based on the information present in the text. This operation is essential for increasing the number of geolocated posts and,

Model type	Model	Number of shots	Accuracy	Precision	Recall	F1	Trainable parameters	Training time
Decoder-based	command	8	0.626	0.635	0.626	0.622	-	-
	llama-2-13b-chat	8	0.756	0.760	0.756	0.754	-	-
	gemini-pro	8	0.764	0.768	0.764	0.764	-	-
	gpt-3.5-turbo	**8**	**0.818**	**0.824**	**0.818**	**0.816**	-	-
Encoder-based	albert	-	0.921	0.922	0.921	0.921	11,684,353	797 s
	bert	-	0.936	0.936	0.936	0.936	109,483,009	841 s
	bertweet	-	0.906	0.907	0.906	0.906	134,900,737	871 s
	distilbert	-	0.933	0.933	0.933	0.933	66,363,649	428 s
	roberta	-	0.930	0.931	0.930	0.930	124,646,401	854 s

Table 4: Performance comparison between encoder- and decoder-based models in the sub_event binary classification task. All encoder-based models are used in their *base-uncased* version.

consequently, generating reports focused on date and place (e.g., a specific city on a given day).

We initially adopted a zero-shot approach using the *gpt-3.5-turbo* model for direct geolocation from textual information contained in the text. The initial prompt, used as our baseline, is reported in the following box.

`Prompt geolocation`: "Considering the following tweets, extract the following geographical information: {state, zip_code, city, and other_geographical_information}. You have to consider only the text of the tweets. In many cases, the information is not provided in the tweet, so provide null values. Provide only the values of the tweet id and the required information separated by a comma without any explanation: [$tweet_1$, $tweet_2$,... , $tweet_m$]".

Afterward, we strengthened the geolocation prompt by providing additional context in the form of geographic-related information. Specifically, we introduced a NER-enhanced geolocation approach, which consists of a two-step process. Firstly, Named Entity Recognition (NER) [61] techniques are leveraged, focused on identifying and classifying named entities within text into predefined categories, such as names, locations, dates, and other specific terms. Afterward, location information is extracted and integrated into the zero-shot geolocation prompt described previously. The NER-enhanced geolocation prompt is presented in the following box:

`Prompt NER-enhanced geolocation`: "Considering the following tweets, extract the following geographical information: state, zip code, city, and other

geographical information (e.g., street and/or district). You have to consider only the text of the tweets, and their associated location entities, extracted through Named Entity recognition. Provide only the values of the tweet id and the required information separated by a comma without any explanation: [$tweet_1$, $tweet_2$,... , $tweet_m$], [loc_1, loc_2,... , loc_m]".

In our experimental evaluation, we compared *gpt-3.5-turbo* with other prompt-based models such as *llama-2-13b-chat*, *command*, and *gemini-pro* for the zero-shot geolocation task. Given the strings S_1 and S_2, indicating the exact location (e.g., *Cactus St., Fulton, Texas*) and the extracted one, we used the following metrics:

- *Jaro similarity* [62]: it measures the similarity between two strings S_1 and S_2. The score is normalized so that 0 corresponds to no similarity, while 1 indicates an exact match. It is calculated as:

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

  where m is the number of matching characters; $|S_1|$, $|S_2|$ is the size of strings; t is half the number of transpositions.

- *Jaccard similarity* [63]: it determines the similarity between two strings S_1 and S_2 as the number of common tokens over the total amount of tokens:

$$J(S_1, S_2) = \frac{|\text{token}(S_1) \cap \text{token}(S_2)|}{|\text{token}(S_1) \cup \text{token}(S_2)|}$$

  where token(S) represents the set of tokens in the string S.

- *Cosine similarity* [64]: it measures the similarity between two non-zero vectors of an inner product space as the cosine of the angle between them. Given to strings S_1 and S_2, it is defined as:

$$\text{cosine\_similarity}(S_1, S_2) = \frac{\bar{S}_1 \cdot \bar{S}_2}{\|\bar{S}_1\| \cdot \|\bar{S}_2\|}$$

  where $\bar{S}$ represents the vector representation of the string S; $\bar{S}_1 \cdot \bar{S}_2$ represents the scalar product between them; $\|\bar{S}\|$ denotes the Euclidean norm (or length) of the vector $\bar{S}$.

Table 5 presents the results attained in the zero-shot modality, using the different LLMs under consideration. As already noted for event classification, the *gpt-3.5-turbo* model demonstrates superior performance across all evaluated metrics, achieving the highest average score for the zero-shot geolocation task. We then extended our analysis to assess the performance of the introduced NER-enhanced geolocation strategy, testing the integration of *gpt-3.5-turbo* — the best performer in the zero-shot setting — with several NER techniques [61], including SpaCy[1], CoreNLP[2], and BERT-base-NER[3]. We also evaluated a GPT-only approach, where also the NER task is performed using *gpt-3.5-turbo* in zero-shot mode.

Model	Strategy	NER technique	jaro sim.	jaccard sim.	cosine sim.	avg score
llama-2-13b-chat	zero-shot	-	0.737	0.464	0.622	0.608
command	zero-shot	-	0.740	0.430	0.696	0.622
gemini-pro	zero-shot	-	0.782	0.556	0.782	0.706
gpt-3.5-turbo	**zero-shot**	-	**0.878**	**0.666**	**0.897**	**0.813**
gpt-3.5-turbo	NER-enhanced zero shot	spacy	0.840	0.596	0.759	0.732
gpt-3.5-turbo	NER-enhanced zero shot	coreNLP	0.851	0.630	0.785	0.755
gpt-3.5-turbo	NER-enhanced zero shot	bert-base-NER	0.883	0.677	0.858	0.806
gpt-3.5-turbo	**NER-enhanced zero shot**	**gpt-3.5-turbo**	**0.898**	**0.738**	**0.858**	**0.831**

Table 5: Performance metrics for Geographic location identification for zero-shot and NER-enhanced geolocation strategies.

The results, shown in Table 5, demonstrate the performance improvements brought by the proposed NER-enhanced strategy over simple zero-shot geolocation. Furthermore, by analyzing the different combinations with NER techniques, we found *gpt-3.5-turbo* itself to be the best option for extracting NER-based geographical information to be integrated as additional context into the geolocation process.

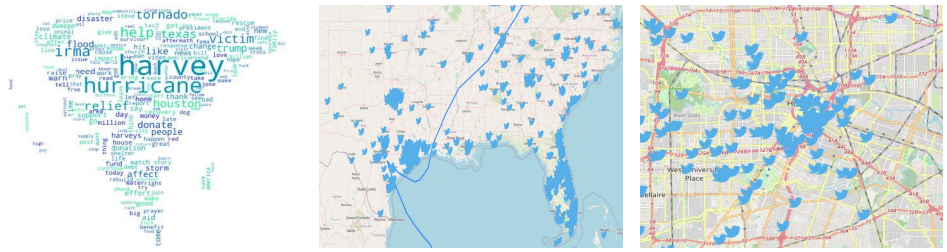
4.6. Disaster Reporting using LLMs

After using LLMs to classify tweets to search for reports of sub-events by users, and geolocate them starting from the information contained in the text, our methodology aims to create timely reports on a disastrous event. Specifically, by leveraging prompt-based models, our approach ensures the

[1]https://spacy.io/api/entityrecognizer
[2]https://stanfordnlp.github.io/CoreNLP/
[3]https://huggingface.co/dslim/bert-base-NER

timely delivery of easily understandable reports, enabling informed decision-making and response strategies based on a comprehensive understanding of the situation on the ground. The objective is to continuously monitor a disastrous event and generate reports for each affected city.

We chose Hurricane Harvey as a case study from the HumAID (Human-Annotated Disaster Incidents Data) dataset [53]. It comprises tweets related to Hurricane Harvey, a Category 4 storm that struck Texas in 2017, causing approximately USD 200 billion in damages and claiming at least 82 lives, as reported by the Texas Department of Public Safety. The dataset consists of approximately 6.7 million tweets collected between August 25, 2017, and September 5, 2017, using specific keywords such as "Hurricane Harvey," "Harvey," and "Harvey 2017" as outlined in [25]. Figure 4 depicts word cloud representations illustrating various aspects of Hurricane Harvey, including its trajectory and the geographical distribution of related tweets. It also highlights the localization of tweets within the city of Houston, one of the most affected areas.



(a) Word cloud representation from Twitter of Hurricane Harvey.

(b) Tracking Hurricane Harvey and geolocated tweets related to that event.

(c) Geolocated tweets in the Houston area.

Figure 4: Exploring social media interactions during Hurricane Harvey.

As done in previous analyses, we have defined some prompts and tested them on different prompt-based LLMs, such as OpenAI's gpt-3.5-turbo, Meta AI's llama-2-13b-chat, Cohere's command, and Google AI's gemini-pro. Specifically, we asked one of these models to behave like a report writer, as specified below:

`"system"`: "Act as an adept report writer creating reports based on a set of tweets."

Then, we defined a series of prompts to define the different parts of our report. In particular, given a city, and given all the posts classified as sub-events and geolocalized in that city, we asked to define a title, an introduction, and an informative section with reference to the most relevant tweets, as defined below:

`"title_prompt"`: "Generate a concise and captivating English title for the report on the city [c] during the disaster [d], within a maximum of 10 words. Avoid the use of words with hashtags(#). Use the following tweets as input: [$tweet_1$, $tweet_2$,... , $tweet_m$]".

`"introduction_prompt"`: "The title of the report is"+[*response_title_prompt*]+ "Generate a concise and engaging English introduction for the report on the city [c] during the disaster [d]. It must be interesting and engaging, capture the reader's attention, and present key information concisely. It must be a single block of text (paragraph), without carriage returns, without the use of a colon (:). Use the following tweets as input: [$tweet_1$, $tweet_2$,... , $tweet_m$]".

`"content_prompt"`: "The title of the report is"+[*response_title_prompt*]+ "The introduction of the report is "+[*response_introduction_prompt*]+ "Generate a concise and engaging English section to analyze the issues reported by citizens on the city [c] during the disaster [d]. Use colons (:) only at the beginning, do not use them in text content. In the text, add references to the tweets in the dataset used to describe detected issues. Use the following tweets as input: [$tweet_1$, $tweet_2$,... , $tweet_m$]".

For the case study considered, we generated a report for all major cities affected by the hurricane using each of the four prompt-based LLMs evaluated. Among these cities, we find Houston, Rockport, Dallas, Washington, Corpus Christi, Austin, San Antonio, Galveston, Texas, Port Arthur, League City, New Orleans, and Victoria. The number of tweets transmitted to a model depends on the ability of the model to handle a certain number of input tokens (words).

Using a concrete example, let us consider a text excerpt related to the city of Houston generated with gpt-3.5-turbo (a similar report can be produced with other models). In the report, references to tweets used by ChatGPT to generate the content are included. This shows that the model actually relies on the tweets provided in the construction of the reports, thus excluding the risk of generating unreliable information.

"response_title": "Harvey's Impact on Houston: A City's Resilience Tested"

"response_introduction": Amidst the deluge of Hurricane Harvey's wrath, Houston stood as a city tested by nature's fury [...]

"response_content": During Hurricane Harvey, Houston was engulfed by devastating floods, with many streets and neighborhoods submerged in torrential water. Intersections like the one between I-45 and N. Main Street became impassable [1], as did areas like Allen Parkway and Studemont St & Memorial Ct, which were completely submerged [2]. In other parts of the city, such as Clay Rd & Queenston, rescue operations were frequent to aid those trapped in their homes [3]. Major arteries like Almeda Genoa Rd., from Windmill Lakes Blvd. to Rowlett Rd., were also affected, making road circulation extremely perilous [4]. The floods also hit the city's iconic green spaces, such as Buffalo Bayou and Hermann Park, causing severe damage [5] [6]. Residents reported hazardous situations, such as high water on Sterrett St. and the need for immediate evacuations in places like 425 Shane St [7]. Many people were stranded in their homes or public places, requiring urgent rescue. For instance, 80 people were trapped in a church located at 5547 Cavalcade St [8]. Three women were reported trapped at 8130 Snowden St [9], while a couple in the 5th district reported people trapped in their church at 3401 Liberty Rd [10]. A mother was trapped at 5400 Bayou Dr. and needed immediate assistance [11], while an abandoned Husky was found tied to a trailer at 2105 Matilda St [12]. A group of elderly and disabled individuals was stranded in waist-high water at 8802 Banting St [13]. Requests for rescue include a newborn and their family trapped at 9 Woodstone St [14], elderly individuals trapped at 1230 Havner Lane [15], and dogs chained up along with a bird in urgent need of rescue at 805 Pennsylvania Ave, SouthHouston [16]. Residents of Melrose Park need a safe shelter for their pets [17]. During the emergency, several locations were designated as shelters for those in need. For example, NRG Park is now open as an emergency shelter [18], while Klein Oak High School, located at 22603 Northcrest Dr., has opened its doors to provide shelter [19]. The Convention Center, accessible by taking the Polk St exit from 59 north, has been set up as a refuge [20], while veterans and medical personnel have set up a large mobile pharmacy at Paige St/Mckinney St [21]. First Baptist Church, at 4422 Lauder St, is providing food and essential supplies [22]. [...]

[1] "Houston HoustonFlood the intersection of I-45 & N. Main Street."
[2] "houstonflood ANYONE gets info on extent of buffalo bayou flooding near

Allen Parkway and Studemont St & Memorial Ct."
[3] "Major flooding at Clay Rd & Queenston in west Houston. Lots of rescues going on for ppl trapped back in the neighborhood."
[4] "Streets Flooded: Almeda Genoa Rd. from Windmill Lakes Blvd. to Rowlett Rd. HurricaneHarvey Houston.
[5] Buffalo Bayou Park at Houston Texas ALMOST destroyed by Harvey HarveyStorm HoustonFlood HurricaneHarvey houwx txwx.
[6] Just found out that my childhood park, Hermann Park is under water ... HoustonFlood HoustonStrong.
[7] Fast moving water abt 5 ft high on Sterret St. Stay safe and avoid high water houstonflood Harvey.
[...]

The report provides a detailed and insightful analysis of Hurricane Harvey's impact on Houston, showcasing the resilience and challenges faced by the city in the aftermath of the devastating storm. The introduction sets the tone by highlighting the city's resilience tested by nature's fury, while the content delves into the immediate and long-term consequences of the disaster. The report effectively utilizes the classification of tweets and their correct localization to provide precise and relevant information, painting a comprehensive picture of the event's aftermath. It addresses various aspects of the impact, including infrastructure damage, environmental concerns, community resilience, and the emotional toll on residents. Overall, the report offers a thorough examination of Harvey's impact on Houston, demonstrating the power of data-driven analysis in understanding and addressing the challenges posed by natural disasters.

4.6.1. Evaluation of reports generated by different LLMs models

Here, we evaluate the performance of the four LLMs considered in our study — OpenAI's gpt-3.5-turbo, Meta AI's llama-2-13b-chat, Cohere's Command, and Google AI's gemini-pro — in generating reports for four cities affected by Hurricane Harvey in 2017, i.e., Houston, Dallas, Rockport, and Corpus Christi. First, we employ the TextDescriptive library [65] to evaluate the reports produced by each model, focusing on linguistic aspects such as readability, coherence, quality, and complexity [66]. Second, following approaches used in related studies [67], we utilize an LLM, such as ChatGPT, to assess the reports based on criteria like informativeness, coherence, quality, and attributability.

	Prompt-based LLMs			
	llama-2-13b-chat	*command*	*gemini-pro*	*gpt-3.5-turbo*
Readability	14.19	14.41	15.06	15.51
Quality	0.14	0.15	0.17	0.18
Coherence	0.75	0.77	0.79	0.83
Complexity	3.73	3.98	4.57	4.89

Table 6: Evaluation scores of reports using the TextDescriptives library.

Table 6 presents the average scores obtained from the TextDescriptives library for the generated reports. Below is a summary of each criterion and the corresponding results:

- *Readability*: assessed using the Coleman-Liau index, which estimates the U.S. school grade level required to comprehend a text. Reports generated by ChatGPT require a higher grade level compared to those from other LLMs. Similar trends are observed across other readability indices, indicating that ChatGPT's reports demand a more advanced linguistic understanding.

- *Quality*: measured using text repetitiveness metrics, specifically the fraction of duplicated n-gram characters. This metric reflects the proportion of characters within a document contained in repeated n-grams. Reports generated by all models show similar percentages, implying comparable levels of informative content.

- *Coherence*: evaluated based on cosine similarity between sentences, with embeddings derived from the average vector representation of words calculated through Latent Semantic Analysis. ChatGPT's reports generally exhibit higher coherence, reflected in greater similarity values compared to Llama, Command, and Gemini.

- *Complexity*: measured using text entropy, which indicates the level of randomness or unpredictability, with higher values representing greater linguistic diversity and complexity. ChatGPT's reports show the highest complexity, characterized by diverse language use, followed by those generated by Gemini-pro, Command, and Llama.

The analysis indicates that while ChatGPT outperforms in readability, coherence, and complexity, models like Gemini-pro, Command, and Llama

offer similar informativeness with simpler structures. The choice of model depends on whether the priority is sophisticated language or clear, essential content.

	Prompt-based LLMs			
	llama-2-13b-chat	*command*	*gemini-pro*	*gpt-3.5-turbo*
Informativeness	3.5	4.5	4.25	4.5
Quality	3.5	4.25	4.25	4.5
Coherence	4.25	4.5	4.25	4.5
Attributability	3.25	3.75	4	4.25
Overall preference	3.5	4	4.25	4.5

Table 7: Evaluation scores for reports assessed by ChatGPT, rated on a scale from 1 (worst) to 5 (best).

Table 7 shows the average scores obtained from reports assessed using ChatGPT as an evaluator. For each city, we provided ChatGPT with a generated report from one of the LLMs and the corresponding set of tweets used to produce the report (selected by our methodology) and asked it to rate the following aspects on a scale from 1 (worst) to 5 (best):

- *Informativeness*: assesses how well the report conveys crucial details from the original data. ChatGPT and Command score the highest, indicating particularly detailed reports. Gemini follows with slightly lower scores, while Llama underperforms, indicating less detailed content.

- *Quality*: measures the clarity and readability of the report. ChatGPT achieves the highest score, reflecting well-structured and clear reports. Command and Gemini are close behind, while Llama scores lowest, suggesting room for improvement in report clarity.

- *Coherence*: evaluates the report's logical flow and organization. All models achieve similar scores, likely due to the report structure being generated from multiple prompts.

- *Attributability*: checks if the information in the report can be traced back to the original tweets. ChatGPT scores highest, demonstrating strong attribution capabilities. Gemini follows, while Llama and Command score lower, indicating challenges in information traceability.

- *Overall Preference*: synthesizes the evaluation of how effectively the report conveys the main ideas. Gemini and ChatGPT receive the highest scores, excelling in clear and concise communication. Command performs well, while Llama scores lowest, indicating a need for improvement in presenting key ideas.

Overall, the evaluation highlights that while all models exhibit strengths across different aspects, ChatGPT consistently outperforms others in informativeness, clarity, and attribution, with Gemini and Command following closely in terms of overall report quality and coherence, while Llama lags behind, particularly in clarity and detail.

5. Conclusion

In recent years, social media emerged as a crucial platform for real-time reporting during disasters and catastrophic events. While significant progress has been made in collecting and classifying such information, there remains an urgent need to improve the automation, aggregation, and organization of user reports to simplify various tasks, including rescue operations, resource allocation, and communication with the press. This paper introduces an innovative methodology that fully leverages the power of Large Language Models to strengthen disaster response and management. Specifically, we focus on prompt-based LLMs that, using zero- or few-shot prompts, can ensure faster, more effective, and better-coordinated disaster relief efforts.

By analyzing large volumes of user-generated content, our methodology identifies issues reported by citizens who have experienced a disaster, such as collapsed buildings, broken gas pipelines, and damaged homes. It also localizes all posts containing references to geographic information in the text, improving localization and consequently the aggregation of posts that occurred nearby. Utilizing these localized citizen-reported issues, the system aggregates information, grouping it by location and date, to generate reports packed with essential information for emergency services, news agencies, and other interested parties. Extensive experimentation on large datasets validates the accuracy and efficiency of our methodology in detecting secondary events and producing real-time reports. These findings highlight the effectiveness of prompt-based LLMs in disaster response, demonstrating their ability to accurately classify posts, identify relevant user feedback, and generate informative reports without the need for additional training on specific datasets.

Looking ahead, future work will focus on refining the methodology to improve its scalability, adaptability, and robustness across various disaster scenarios and geographic regions. A key challenge is balancing model reasoning capabilities with inference speed in disaster management contexts. To address this, we propose research directions aimed at enhancing model reasoning while maintaining efficiency. Additionally, optimizing the methodology to provide accurate, real-time information is crucial for practical application in disaster response and management scenarios. We also acknowledge the importance of integrating robust filtering mechanisms and evaluating LLM performance in handling noisy and ambiguous data. Future work will explore detailed mechanisms for filtering noisy documents, managing ambiguous geolocation data, and assessing how LLMs deal with deliberately inserted false information to ensure accurate and reliable outputs in complex disaster settings. Effective filtering will not only enable faster analysis but also lead to more accurate results by relying on high-quality data. Overall, the ongoing development and refinement of such approaches hold promise for improving preparedness, response, and resilience to increasingly frequent and severe natural disasters.

Acknowledgment

We acknowledge financial support from "National Centre for HPC, Big Data and Quantum Computing", CN00000013 - CUP H23C22000360005, and from "FAIR – Future Artificial Intelligence Research" project - CUP H23C22000860006.

References

[1] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, S. Nerur, Advances in social media research: Past, present and future, Information Systems Frontiers 20 (2018) 531–558.

[2] L. Belcastro, F. Branda, R. Cantini, F. Marozzo, D. Talia, P. Trunfio, Analyzing voter behavior on social media during the 2020 us presidential election campaign, Social Network Analysis and Mining 12 (1) (2022) 83.

[3] R. Cantini, C. Cosentino, I. Kilanioti, F. Marozzo, D. Talia, Unmasking covid-19 false information on twitter: A topic-based approach with bert,

in: International Conference on Discovery Science, Springer, 2023, pp. 126–140.

[4] L. Belcastro, R. Cantini, F. Marozzo, Knowledge discovery from large amounts of social media data, Applied Sciences 12 (3) (2022) 1209.

[5] C. Castillo, Big crisis data: social media in disasters and time-critical situations, Cambridge University Press, 2016.

[6] H. T. Otal, E. Stern, M. A. Canbaz, Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration, in: 2024 IEEE Conference on Artificial Intelligence (CAI), IEEE, 2024, pp. 851–859.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (12) (2023) 1–38.

[8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (9) (2023) 1–35.

[9] T. H. Nazer, G. Xue, Y. Ji, H. Liu, Intelligent disaster response via social media analysis a survey, ACM SIGKDD Explorations Newsletter 19 (1) (2017) 46–59.

[10] Z. Wang, X. Ye, Social media analytics for natural disaster management, International Journal of Geographical Information Science 32 (1) (2018) 49–72.

[11] O. D. Apuke, B. Omar, Fake news and covid-19: modelling the predictors of fake news sharing among social media users, Telematics and Informatics 56 (2021) 101475.

[12] C. Slamet, A. Rahman, A. Sutedi, W. Darmalaksana, M. A. Ramdhani, D. S. Maylawati, Social media-based identifier for natural disaster, in: IOP conference series: materials science and engineering, Vol. 288, IOP Publishing, 2018, p. 012039.

[13] N. Said, K. Ahmad, M. Riegler, K. Pogorelov, L. Hassan, N. Ahmad, N. Conci, Natural disasters detection in social media and satellite imagery: a survey, Multimedia Tools and Applications 78 (2019) 31267–31302.

[14] R. Dong, L. Li, Q. Zhang, G. Cai, Information diffusion on social media during natural disasters, IEEE transactions on computational social systems 5 (1) (2018) 265–276.

[15] S. A. Shah, D. Z. Seker, S. Hameed, D. Draheim, The rising role of big data analytics and iot in disaster management: recent advances, taxonomy and prospects, IEEE Access 7 (2019) 54595–54614.

[16] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, M. Tesconi, Ears (earthquake alert and report system) a real time decision support system for earthquake crisis management, in: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, 2014, pp. 1749–1758.

[17] R. Bossu, F. Roussel, L. Fallou, M. Landès, R. Steed, G. Mazet-Roux, A. Dupont, L. Frobert, L. Petersen, Lastquake: From rapid information to global seismic risk reduction, International journal of disaster risk reduction 28 (2018) 32–42.

[18] Y. Li, F. B. Osei, T. Hu, A. Stein, Urban flood susceptibility mapping based on social media data in chengdu city, china, Sustainable Cities and Society 88 (2023) 104307.

[19] L. Lin, C. Tang, Q. Liang, Z. Wu, X. Wang, S. Zhao, Rapid urban flood risk mapping for data-scarce environments using social sensing and region-stable deep neural network, Journal of Hydrology 617 (2023) 128758.

[20] P. Meladianos, C. Xypolopoulos, G. Nikolentzos, M. Vazirgiannis, An optimization approach for sub-event detection and summarization in twitter, in: Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40, Springer, 2018, pp. 481–493.

[21] D. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, P. Mitra, Robust classification of crisis-related data on social networks using convolutional neural networks, in: Proceedings of the international AAAI conference on web and social media, Vol. 11, 2017, pp. 632–635.

[22] Z. Wang, Y. Zhang, A neural model for joint event detection and summarization., in: IJCAI, 2017, pp. 4158–4164.

[23] G. Bekoulis, J. Deleu, T. Demeester, C. Develder, Sub-event detection from twitter streams as a sequence labeling problem, arXiv preprint arXiv:1903.05396 (2019).

[24] C. Xing, Y. Wang, J. Liu, Y. Huang, W.-Y. Ma, Hashtag-based sub-event discovery using mutually generative lda in twitter, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 30, 2016.

[25] L. Belcastro, F. Marozzo, D. Talia, P. Trunfio, F. Branda, T. Palpanas, M. Imran, Using social media for sub-event detection during disasters, Journal of Big Data 8 (79) (2021).

[26] P. Lei, G. Marfia, G. Pau, R. Tse, Can we monitor the natural environment analyzing online social network posts? a literature review, Online Social Networks and Media 5 (2018) 51–60.

[27] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).

[28] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[29] A. Gilson, C. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, How well does chatgpt do when taking the medical licensing exams? the implications of large language models for medical education and knowledge assessment, medRxiv (2022) 2022–12.

[30] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is chatgpt to human experts? comparison corpus, evaluation, and detection, arXiv preprint arXiv:2301.07597 (2023).

[31] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al., A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, arXiv preprint arXiv:2302.04023 (2023).

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[33] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andia, C. Tejos, C. Prieto, D. Capurro, A survey on deep learning and explainability for automatic report generation from medical images, ACM Computing Surveys (CSUR) 54 (10s) (2022) 1–40.

[34] F. Wang, Z. Xu, P. Szekely, M. Chen, Robust (controlled) table-to-text generation with structure-aware equivariance learning, arXiv preprint arXiv:2205.03972 (2022).

[35] M. Yazaki, S. Maki, T. Furuya, K. Inoue, K. Nagai, Y. Nagashima, J. Maruyama, Y. Toki, K. Kitagawa, S. Iwata, et al., Emergency patient triage improvement through a retrieval-augmented generation enhanced large-scale language model, Prehospital Emergency Care (2024) 1–7.

[36] S. Shankar, J. Zamfirescu-Pereira, B. Hartmann, A. Parameswaran, I. Arawjo, Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 2024, pp. 1–14.

[37] H. Wei, S. He, T. Xia, A. Wong, J. Lin, M. Han, Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates, arXiv preprint arXiv:2408.13006 (2024).

[38] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, B. Li, Alert: A comprehensive benchmark for assessing large language models' safety through red teaming, arXiv preprint arXiv:2404.08676 (2024).

[39] T. A. van Schaik, B. Pugh, A field guide to automatic evaluation of llm-generated summaries, in: Proceedings of the 47th International ACM

SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2832–2836.

[40] T. Hu, X.-H. Zhou, Unveiling llm evaluation focused on metrics: Challenges and solutions, arXiv preprint arXiv:2404.09135 (2024).

[41] V. V. Mihunov, N. S. Lam, L. Zou, Z. Wang, K. Wang, Use of twitter in disaster rescue: lessons learned from hurricane harvey, International Journal of Digital Earth 13 (12) (2020) 1454–1466.

[42] R. Suwaileh, T. Elsayed, M. Imran, H. Sajjad, When a disaster happens, we are ready: Location mention recognition from crisis tweets, International Journal of Disaster Risk Reduction 78 (2022) 103107.

[43] B. Zhou, L. Zou, A. Mostafavi, B. Lin, M. Yang, N. Gharaibeh, H. Cai, J. Abedin, D. Mandal, Victimfinder: Harvesting rescue requests in disaster response from social media with bert, Computers, Environment and Urban Systems 95 (2022) 101824.

[44] S. R. Hiltz, A. L. Hughes, M. Imran, L. Plotnick, R. Power, M. Turoff, Exploring the usefulness and feasibility of software requirements for social media use in emergency management, International journal of disaster risk reduction 42 (2020) 101367.

[45] L. Belcastro, D. Carbone, C. Cosentino, F. Marozzo, P. Trunfio, Enhancing cryptocurrency price forecasting by integrating machine learning with social media and market data, Algorithms 16 (12) (2023) 542.

[46] J. P. De Albuquerque, B. Herfort, A. Brenning, A. Zipf, A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management, International journal of geographical information science 29 (4) (2015) 667–689.

[47] Z. Wang, X. Ye, M.-H. Tsou, Spatial, temporal, and content analysis of twitter for wildfire hazards, Natural Hazards 83 (2016) 523–540.

[48] R. Dutt, K. Hiware, A. Ghosh, R. Bhaskaran, Savitr: A system for real-time location extraction from microblogs during emergencies, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1643–1649.

[49] M. Karimzadeh, S. Pezanowski, A. M. MacEachren, J. O. Wallgrün, Geotxt: A scalable geoparsing system for unstructured text geolocation, Transactions in GIS 23 (1) (2019) 118–136.

[50] J. Wang, Y. Hu, K. Joseph, Neurotpr: A neuro-net toponym recognition model for extracting locations from social media messages, Transactions in GIS 24 (3) (2020) 719–735.

[51] C. Berragan, A. Singleton, A. Calafiore, J. Morley, Transformer based named entity recognition for place name extraction from unstructured text, International Journal of Geographical Information Science 37 (4) (2023) 747–766.

[52] Y. Hu, G. Mai, C. Cundy, K. Choi, N. Lao, W. Liu, G. Lakhanpal, R. Z. Zhou, K. Joseph, Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages, International Journal of Geographical Information Science 37 (11) (2023) 2289–2318.

[53] F. Alam, U. Qazi, M. Imran, F. Ofli, Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks, in: Proceedings of the International AAAI Conference on Web and social media, Vol. 15, 2021, pp. 933–942.

[54] L. Belcastro, R. Cantini, F. Marozzo, D. Talia, P. Trunfio, Learning political polarization on social media using neural networks, IEEE Access 8 (1) (2020) 47177–47187.

[55] R. Cantini, F. Marozzo, D. Talia, P. Trunfio, Analyzing political polarization on social media by deleting bot spamming, Big Data and Cognitive Computing 1 (6) (2022).

[56] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD explorations newsletter 19 (1) (2017) 22–36.

[57] M. Aman, Large language model based fake news detection, Procedia Computer Science 231 (2024) 740–745.

[58] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[59] A. Parnami, M. Lee, Learning from few examples: A summary of approaches to few-shot learning, arXiv preprint arXiv:2203.04291 (2022).

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[61] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE transactions on knowledge and data engineering 34 (1) (2020) 50–70.

[62] M. A. Jaro, Unimodal search for fixed strings, ACM Transactions on Mathematical Software 15 (3) (1989) 332–340.

[63] P. Jaccard, Nouvelles recherches sur la distribution florale dans le jura suisse et dans ses contrées voisines, Bulletin de la Société Vaudoise des Sciences Naturelles 44 (2) (1908) 375–452.

[64] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1973.

[65] L. Hansen, L. R. Olsen, K. Enevoldsen, Textdescriptives: A python package for calculating a large variety of metrics from text, Journal of Open Source Software 8 (84) (2023) 5153.

[66] R. Cantini, C. Cosentino, F. Marozzo, Multi-dimensional classification on social media data for detailed reporting with large language models, in: 20th International Conference on Artificial Intelligence Applications and Innovations, 2024, pp. 100–114.

[67] G. Adams, A. Fabbri, F. Ladhak, E. Lehman, N. Elhadad, From sparse to dense: Gpt-4 summarization with chain of density prompting, arXiv preprint arXiv:2309.04269 (2023).