Extracting Disaster Impacts and Impact Related Locations in Social Media Posts Using Large Language Models

Sameeah Noreen Hameed[a], Surangika Ranathunga[a], Raj Prasanna[a], Kristin Stock[a], and Christopher B. Jones[b]

[a]School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand; [b]School of Computer Science and Informatics, Cardiff University, Cardiff, UK

ARTICLE HISTORY
Compiled December 1, 2025

ABSTRACT
Large-scale disasters can often result in catastrophic consequences on people and infrastructure. Situation awareness about such disaster impacts generated by authoritative data from in-situ sensors, remote sensing imagery, and/or geographic data is often limited due to atmospheric opacity, satellite revisits, and time limitations. This often results in geo-temporal information gaps. In contrast, impact-related social media posts can act as "geo-sensors" during a disaster, where people describe specific impacts and locations. However, not all locations mentioned in disaster-related social media posts relate to an impact. Only the impacted locations are critical for directing resources effectively. e.g., *"The death toll from a fire which ripped through the Greek coastal town of #Mati stood at 80, with dozens of people unaccounted for as forensic experts tried to identify victims who were burned alive #Greecefires #AthensFires #Athens #Greece."* contains impacted location *"Mati"* and non-impacted locations *"Greece"* and *"Athens"*. This research uses Large Language Models (LLMs) to identify all locations, impacts and impacted locations mentioned in disaster-related social media posts. In the process, LLMs are fine-tuned to identify only impacts and impacted locations (as distinct from other, non-impacted locations), including locations mentioned in informal expressions, abbreviations, and short forms. Our fine-tuned model demonstrates efficacy, achieving an F1-score of 0.69 for impact and 0.74 for impacted location extraction, substantially outperforming the pre-trained baseline. These robust results confirm the potential of fine-tuned language models to offer a scalable solution for timely decision-making in resource allocation, situational awareness, and post-disaster recovery planning for responders.

KEYWORDS
Social media posts; Disaster response; Large Language Models(LLMs); Locations; Geospatial; Toponym recognition

1. Introduction

Disasters, triggered by natural hazards, are a major challenge that can have a devastating impact on the lives of people and infrastructure. Effective disaster management and response rely heavily on the availability of accurate information, delivered in a timely manner and appropriate format [1]. Authoritative data, derived from sources such as in-situ sensors, remote sensing imagery, and geographic information systems (GIS) are vital for performing accurate damage assessments and coordinating relief operations [2].

CONTACT Sameeah Noreen Hameed Email: snoreen1@massey.ac.nz

However, the acquisition of this sensor-based information is often constrained by real-world limitations, such as atmospheric opacity, which can obscure satellite views, or the inherent revisit time limitations of satellites [3, 4]. These constraints frequently lead to critical geo-temporal gaps, where timely data is unavailable [5, 6]. To address this information deficit and improve situational awareness, studies utilise alternative data sources such as social media text [7]. Social media with its vast user base and real-time communication capabilities, has emerged as a powerful tool for gathering and disseminating critical information during disasters [8]. Platforms such as Twitter, Facebook, and Instagram allow users to share first-hand observations and updates about ongoing events, effectively transforming social media into a network of Volunteered Geographic Information (VGI) [9]. These platforms serve as a "geo-sensor" network, providing valuable real-time data that can complement traditional remote sensing methods and enhance situational awareness for emergency services. Social media platforms are more responsive with a more continuous flow of information during disaster times than official data, which comes in intervals that are often too long [10].

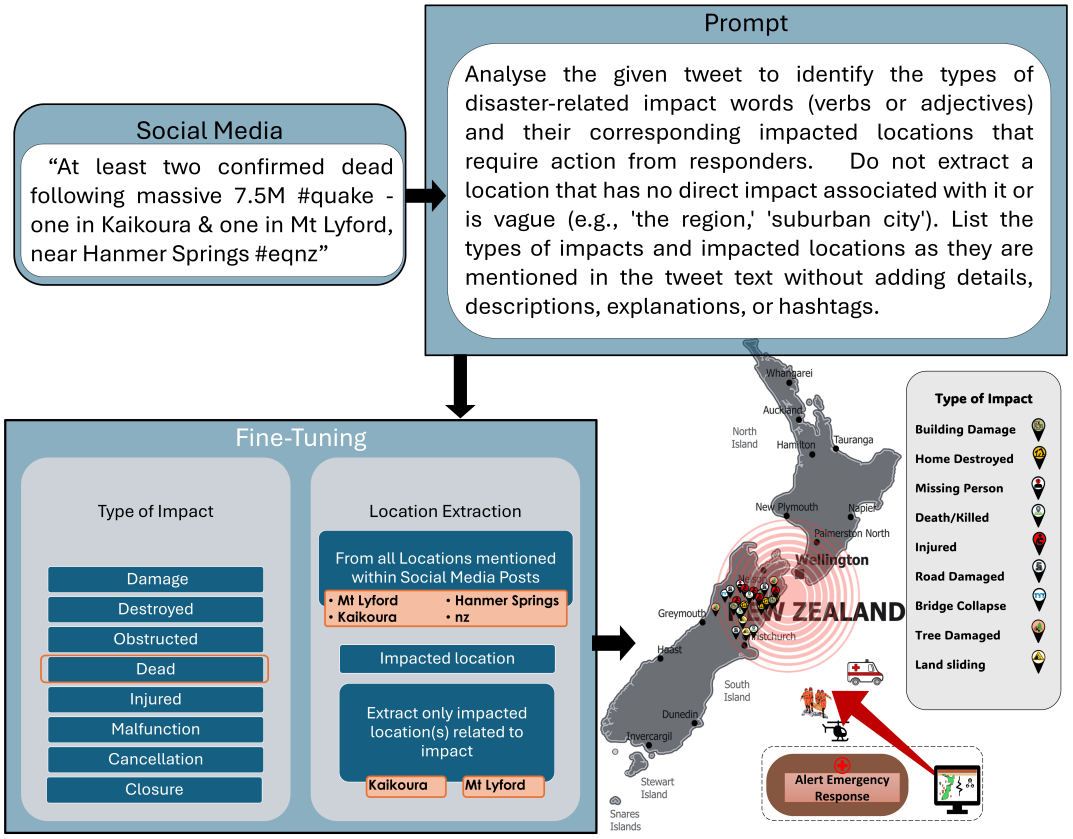

Figure 1.: LLM-based real-time disaster impact mapping from social media data. The process filters non-impact locations to help emergency managers target and prioritise response efforts.

However, the sheer volume of information generated on social media during disasters presents its own set of challenges. One of the key challenges is the identification and extraction of relevant geographic locations, or toponyms (geographically locatable place names), from the vast stream of posts [11]. These locations are critical for mapping

the spatial extent of a disaster and understanding its impact. Situational awareness in disaster scenarios hinges on the ability to accurately identify the locations most affected by the event and the specific nature of the impact at each location [12]. Social media platforms generate a wealth of geospatial information that can be leveraged to geolocate disaster impacts with useful impact-related descriptions, and inform response efforts [13].

Despite the potential of social media as a source of real-time geographic information, existing research in the field of location recognition or place name identification has predominantly focused on identifying a set of locations mentioned in disaster-related posts [14, 11, 15, 16]. In this paper, this set of locations mentioned in disaster-related posts is referred to as **all locations** and Twitter/X posts as **tweets**. The all locations set can also include locations that are not directly related to the disaster's impact, but rather refer to governments, organisations or other facts or events.

For instance, multiple locations are mentioned in the following social media post, *"Jatlaan canal embankment is damaged. It is one of the main canals originating from Mangla dam which irrigates Punjab. #earthquake AJK Mirpur"*. These locations include *Jatlaan canal*, *Mangla dam*, *Punjab*, *AJK*, and *Mirpur*. However, only the *Jatlaan canal* and *AJK Mirpur* are directly impacted, as indicated by the reference to the impact *damaged* and *earthquake*. The presence of multiple locations can lead to confusion and inefficiencies in disaster response, as responders may struggle to identify which location requires immediate attention.

For emergency responders, different impact types require different response strategies [17], such as sending rescue teams with medical aid for the injured versus deploying maintenance teams for infrastructure recovery. Existing disaster classification approaches, such as that developed by Alam et al. [18], categorise posts according to their content. However, they do not identify the impact and the severity, to help emergency responders towards focused relief operations. For example, Alam et al. [18] classify the following post under the infrastructure-and-utility-damage category *"211 houses have been completely destroyed and a further 234 are badly damaged by the devastating #wildfires that ripped through a coastal town east of #Athens, killing at least 80 people. #AthensFires #Greecefires #Greece"*. However, in this example, three distinct impacts: the main impact (wildfires), infrastructure damage and casualties all appear. So, there is a critical need for techniques that can accurately distinguish fine-grained disaster impacts in social media posts. This leads us to our research question: *How can we identify impacts and distinguish locations directly impacted by disasters in social media postings from all locations mentioned, enabling rapid response efforts?*

This research attempts to answer the above question by developing corpora and machine learning (ML)-based methods to accurately identify the types of impact and the directly impacted location(s) mentioned in the social media posts. The study aims to extract disaster impacts and impacted locations that can contribute to situational awareness and effective disaster response efforts during a disaster.

The contribution of this paper is two-fold.

- Firstly, we develop a benchmark corpus that identifies impacted locations and types of impact in disaster-related social media text.
- Secondly, building on the finding that Large Language Models (LLMs) succeeded for all locations recognition, we develop a novel methodology to identify directly impacted locations and types of impact mentioned in disaster-related social media text.

The remainder of this paper is organised into the following sections. Section 2 discusses related work and previous approaches. Section 3 highlights the details of the

dataset used in this research. It is followed by the detailed methodology in Section 4 and the experimental setup in Section 5. The results and their analysis are presented in Section 6. Section 7 discusses the key findings and limitations. Finally, Section 8 concludes the paper and outlines directions for future work.

2. Related Work

2.1. *Impact Information in Disaster Posts*

Disasters such as earthquakes, hurricanes, floods, and wildfires are catastrophic events that disrupt communities and ecosystems, resulting in loss of life, destruction of property, and long-term economic and social impacts [19]. Social consequences include population displacement [20], shortages of essential resources [21], and psychological effects such as anxiety and trauma [22, 23].

While social and economic research examines human and financial impacts [24, 25], the most visible impacts are related to physical damage, and are frequently reported via real-time social media posts [18]. Physical impacts such as casualties, infrastructure damage, and service outages demand urgent attention from emergency responders. These impacts are often captured promptly on social media through user-generated posts containing text, images, or videos of affected areas [26, 27, 28]. Unlike economic losses that unfold over time [29], physical damage information shared online is highly actionable during the early response phase.

Xiao et al. [30] and Shan et al. [10] demonstrated that impact information in social media posts, such as casualties, infrastructure damage, and economic losses, enhances situational awareness and rapid response planning. Despite this, limited research has explored the extraction of fine-grained types of impact information along with locations.

2.2. *Approaches to Location Extraction*

Early research on location recognition in social media used user profile or geo-tag information from Twitter/X to infer the tweet's origin [31, 32]. However, a user's profile location does not necessarily reflect their real-time physical location when a tweet is posted, and profile-based locations can be inaccurate or irrelevant [33]. This led to approaches focusing on identifying locations mentioned in the tweet text itself. Imran et al. [34] explored how points of interest can be identified based on the locations mentioned in the text.

Various approaches and techniques have been used to improve the precision of location recognition strategies and algorithms. Named Entity Recognition (NER) is one of the widely used approaches for location recognition from social media posts [13]. Some of the commonly used NER tools for location recognition include the use of Stanford NER [35] and SpaCy-NER[1]. Both of these tools have been used by researchers for geoparsing, which involves location recognition and resolution from unstructured text into geographic coordinates, in disaster-related tweets [36, 37].

Despite the availability of these NER tools, certain challenges impact their performance in location extraction from social media posts. Some of the major challenges include the use of abbreviations, incorrect spellings, and slang in social media posts [38]. Another challenge with NER tools is that they cannot reliably distinguish between

[1]SpaCy-NER available at: `https://spacy.io/universe/project/video-spacys-ner-model`

locations and similar named entities (e.g., *Washington State* vs. *Washington, D.C.* vs. *George Washington* (person) [39]. Moreover, most of these NER tools are trained to recognise multiple entities (e.g. people and organisations), rather than focusing specifically on location entities, which can reduce their reliability for this task.

To overcome the limitations of publicly available, generic NER tools, researchers have devised mechanisms that explicitly focus on location extraction. These approaches include traditional Machine learning (ML) and Deep Learning (DL) techniques. Early examples of ML approaches, such as Support Vector Machines (SVMs), Naïve Bayes (NB), and Random Forests (RF) have been applied to location extraction from social media, using lexical, grammatical, and geographical features with the Beginning, Inside, Last, Unit-length (BILOU) schema and rule-based methods to enhance performance [40, 41]. DL techniques greatly advanced location extraction by capturing complex patterns in data through a range of Neural Network based methods. These include a Convolutional Neural Network (CNN) [42], the Bidirectional Long Short-Term Memory (Bi-LSTM) model, which enhanced the recognition of rare or unknown locations from Twitter/X [43] and a fusion of CNN and LSTM model for geoparsing city-level geolocation [44].

Transformer models, particularly BERT[45], revolutionised location recognition through contextual embeddings. Suwaileh et al. [11] reported that BERT-based models achieved superior performance in location recognition and disambiguation(differentiating similar locations such as Paris, France vs Paris, Texas) tasks, surpassing all baseline models such as BiLSTM-CRF and SpaCy-NER. Fan et al. [46] proposed a hybrid approach for disaster situational awareness that combined NER for location detection, a location fusion method for refining coordinates, and a fine-tuned BERT model for classifying posts using Hurricane Harvey data. Transformer-based models like BERT set the benchmark with strong contextual understanding; however, their state-of-the-art (SOTA) performance in general entity recognition diminishes in the noisy, time-critical context of disaster response, where they often fail to accurately extract fine-grained, multi-level location information [11, 47].

Recent advances integrate LLMs for fine-grained spatial reasoning. Hu et al. [47] introduced geo-knowledge-guided prompting for detailed location descriptions in disaster texts. Yin et al. [48] use LLM-based techniques to process large-scale textual and visual data for implicit geospatial information, improving the interpretation of location details. Sun et al. [16] developed the annotation tool GALLOC, for labelling and handling detailed, multi-level location descriptions in disaster-related texts. LLMs have moved beyond traditional sequence labelling and shifted towards generative and few-shot learning capabilities, which enable them to perform NER as a text-to-text generation problem [49, 50]. Researchers used various prompting techniques, such as Chain-of-Thought [51] and Retrieval-Augmented Generation (RAG), to enhance performance and deal with hallucinations, where the model invents non-existent locations or geospatial facts [52, 53, 48, 54]. Early prompting-based studies found ChatGPT[2] could perform well in zero to few-shot settings; however, it underperformed relative to fine-tuned transformer baselines on CoNLL-2003 and OntoNotes [55]. Cheng et al. [56] standardised prompts with four elements in few-shot settings for the NER task, along with self-check and correction implementation at the end in output generated by LLMs. Their experiments demonstrated that LLMs still encounter difficulties in technical fields involving specialised terminology in entity recognition. Zaratiana et al. [57] showed that the GLiNER model outperformed LLMs in zero-shot and low-resource NER tasks, due to format sensitivity and lack of structured span control in LLMs. Studies have also shown

[2]OpenAI. ChatGPT. Available at: `https://chatgpt.com/`

that while general-purpose LLMs possess a surprising amount of geospatial knowledge [58], fine-tuning with domain-specific data and incorporating geo-knowledge graphs can significantly improve the accuracy in the location extraction task from disaster-related social media messages [16, 47]. demonstrates how LLMs can integrate textual and visual cues to infer location context explicit location mentions. However, LLMs consistently face challenges due to hallucination [59, 60], sensitivity to prompt formation, and the presence of long contexts [61].

Existing studies have made significant progress using traditional NLP models, transformer-based architectures and more recently, LLMs for identifying location from social media posts. However, these studies focused on identifying all locations mentioned within a post, without considering whether these locations are directly affected by the disaster impact being described. This highlights a critical gap that this research addresses by focusing on the extraction of impacted locations and the impacts to enable actionable situational awareness for disaster response.

2.3. *Classification of Disaster-related Social Media Posts*

The task of extracting disaster-related information from social media text has often been approached as a classification task. Research in this area has focused on two dimensions: (1) whether a post is relevant to the disaster or not, and (2) categorising posts based on the content.

2.3.1. Taxonomy of Disaster-related Categories

Many studies have focused on binary classification of disaster-related posts which distinguishes between relevant/irrelevant or informative/uninformative posts. Imran et al. [62] presented one of the early efforts towards the classification of disaster-related posts. Their framework has two stages: classifying tweets as *informative* or *non-informative*, followed by categorisation into specific types such as *caution/advice*. In addition to work identifying whether a post is relevant to a disaster, or informative, there have been a number of studies that have presented detailed taxonomies [63, 64]. Olteanu et al. [65] begin with a binary classification, which was later extended in [66] to classify posts into six humanitarian categories. Caragea et al. [67] classified SMS messages from the Haiti earthquake into 10 categories. Imran et al. [68] and Alam et al. [18] further developed the disaster-related posts classification on 19 different disasters across the globe. Alam et al. [18] refined the categories defined by them in [68] to include *requests/urgent needs* and a *don't know/can't judge* category for non-English or ambiguous posts. Building on these foundational efforts, recent work has shifted focus from relevance to detecting actionable tweets. For example, Kruspe et al. [69] highlighted the necessity of distinguishing actionable tweets from relevant tweets. Gao et al. [70] propose a knowledge injected prompt learning method rather than using standard classifier pipelines, into actionable categories rather than just *relevant/irrelevant*. Lamsal et al. [71] also highlights tweets classified as *requests or offers* often convey support, generic behavioural guidance, or outdated information, and therefore subdivide this class into four subcategories: *Request, Offer, Irrelevant, Request and Offer*. Table 1 summarises the classification efforts by various researchers.

Table 1.: Classification of disaster-related social media posts

Paper	Classification	Methodology	Techniques
Imran et al. [62]	Informative vs Non-informative, Caution and advice, Information source, Donations, Causalities & damage, Unknown	Used traditional machine learning algorithms (NB, SVM) for classification.	ML Techniques
Ashktorab et al. [63]	Related vs. Not related	Built a pipeline to identify disaster-related tweet relevancy focusing on damage and casualties.	Rule-Based Techniques
Olteanu et al. [65]	Informative vs Uninformative	Manual annotation of crisis tweets followed by lexicon construction using frequent term extraction and distributional similarity.	Lexicon-Based, Semi-Supervised
Parilla-Ferrer et al. [64]	Informative vs Non-informative	Developed a framework to classify tweets based on content relevancy.	ML Techniques
Nguyen et al. [72]	Relevant vs. Irrelevant	Employed Convolutional Neural Networks (CNNs) for tweet classification, achieving higher accuracy.	Deep Learning Techniques
Rathod et al. [73]	Relevant vs. Irrelevant	Utilized Word2Vec-based language embeddings with machine learning models for classification.	ML Techniques
Olteanu et al. [66]	Affected individuals, Infrastructure and utilities, Donations and volunteering, Caution and advice, Sympathy and emotional support, Other useful information	Manually annotated tweets from 26 crisis events into six humanitarian categories using crowdsourcing and expert review.	Manual
Caragea et al. [67]	Medical emergency, People trapped, Shelter needed, Food distribution, Food shortage, Water shortage, Water sanitation, Collapsed structure, Hospital services, Person news	Classified SMS messages from the Haiti earthquake into specified categories based on content.	ML Techniques
Imran et al. [68]	Injured/dead people, Sympathy/ support, Displaced people/ evacuations, Infrastructure damage, Donation needs/offers, Missing/trapped/found people, Caution/advice, Not related, Other useful	Developed a classification strategy based on manually annotated tweets from 19 different disasters.	ML Techniques
Alam et al. [18]	Injured or dead people, Missing/ trapped/found people, Requests/ urgent needs, Displaced people/ evacuations, Rescue/donation/ volunteering, Caution/advice, Sympathy/support, Infrastructure damage, Other useful information, Not related, Don't know/can't judge	Improved previous classifications by splitting and refining categories based on tweet content.	ML Techniques
Suwaileh et al. [74]	Same categories as Alam et al. [18]	Further refined classification strategies for disaster-related tweets.	ML Techniques
Gao et al. [70]	Advice, Threat, Service, Volunteer	Proposed a knowledge-injected prompt learning approach to enhance actionable category detection.	Transformer-based Deep Learning
Lamsal et al. [71]	Request, Offer, Irrelevant, Request and Offer categories	Classification of tweets into actionable categories.	Transformer-based Classification

2.3.2. Techniques for Classifying Disaster Posts

Various techniques have been proposed for automatically categorising social media posts based on the taxonomies mentioned in Section 2.3.1. Some early works, such as Ashktorab et al. [63], used rule-based techniques to classify posts based on content relevancy to disaster keywords. ML approaches were widely used in research that treated the task as a supervised binary classification problem [62, 64] or as a multi-class classification problem for more fine-grained information [67]. These approaches mainly focused on the bag-of-words technique with NB and SVM methods. There have been some efforts using Word2Vec-based language embeddings and machine-learning models for both binary [73] and multi-class classification [68] tasks. Nguyen et al. [72] employed deep learning-based CNNs to classify tweets as *relevant* or *irrelevant*, achieving higher accuracy compared to traditional methods.

Transformer-based architectures have significantly advanced disaster posts classification. Pre-trained language models such as BERT and RoBERTa have been fine-tuned for disaster-related tasks, demonstrating substantial improvements in capturing contextual meaning beyond bag-of-words or Word2Vec embeddings [18, 75]. Models specifically adapted for crisis informatics, such as CrisisBERT [76], leverage domain adaptation techniques to handle the unique vocabulary and context of disaster-related social media.

Beyond BERT-based models, larger generative models and instruction-tuned LLMs have recently been applied for the classification of disaster tweets. Zero-shot and few-shot prompting with LLMs (e.g., GPT-3, Llama, and Mistral) have shown promising performance, reducing dependence on large annotated datasets [77, 78, 79]. McDaniel et al. [80] used GPT-4, Gemini, and Claude in zero-shot classification, showing LLMs outperform traditional supervised methods for both binary and other humanitarian categories. Yin et al. [48] proposed CrisisSense-LLM, which uses instruction fine-tuning for multi-label classification of social media posts, enabling robust detection of multiple humanitarian information types within a single tweet. Imran et al. [81] evaluated the robustness of LLMs across 19 disaster events and linguistic contexts, demonstrating their adaptability while also highlighting challenges with nuanced classes.

Overall, the transition from rule-based approaches to ML, DL, and, most recently, LLMs illustrates a sustained effort to enhance the robustness, generalizability, and timeliness of disaster tweet classification. In particular, LLMs mark a significant shift toward flexible and data-efficient methods that can be quickly adapted and deployed in response to emerging and unforeseen crisis events. However, despite these advancements, most existing approaches focus on broad categorical classifications and do not extract fine-grained information such as the specific impact or the directly impacted locations. This gap highlights the need for models capable of capturing detailed impact-level information from social media text to support more targeted and actionable disaster response.

2.4. *Disaster-Related Social Media Datasets*

Social media platforms such as Twitter, Facebook, and Instagram create a stream of information as soon as an event occurs. Twitter has traditionally been considered the primary medium by researchers [33], due to its real-time nature and ease of sharing short, concise messages. However, since Twitter rebranded to "X" in 2023, major policy changes occurred, particularly the removal of free API access. Ongoing studies continue to use "X" data despite API policy changes for disaster research [82, 83, 84].

Several disaster related social media data sets have been developed, Sit et al. [85] extracted about 20,000 Hurricane Irma tweets and manually annotated 10,000 of them for relevance, informativeness, and situational categories to support disaster response analysis. Similarly, Scheele et al. [86] collected Tweets using keywords related to Hurricane Sandy, during the timeframe of the disaster via the Twitter API. The HumAID dataset [18] is one of the largest collections of disaster-related social media data, comprising tweets from 19 different disaster events, classified into nine categories of disaster-related information. However, this dataset lacks explicit annotations for locations and impact types, limiting its direct applicability to location-specific disaster impact analysis. More recently, Suwaileh et al. [74] used the HumAID dataset for the location mention recognition task (LMR) task. They used manual annotation and synthetic annotation generation techniques to develop a corpus of 20.5k humanly annotated tweets and 57k automatically labelled tweets. They further extended it for the location disambiguation task and termed it the IDRISI-D dataset, which consists of 5,591 manually annotated tweets. The English subset (IDRISI-DE) contains tweets that have one or more locations and was annotated to identify all locations mentioned in the tweet text. Each location was then manually disambiguated. However, a key limitation of IDRISI-DE is that it does not distinguish locations that were directly impacted by the disaster, which is an essential requirement for our task. Also, it does not focus on the types of impacts mentioned in the disaster tweets. Table 2 summarises the key attributes of the above-described data sets.

Table 2.: Key Twitter datasets on disaster management

Reference	**Disaster Event(s)**	**Data Volume**	**Purpose/Task**
Sit et al. [85]	Hurricane Irma (2017)	20,000 labelled tweets	location recognition from tweets
Scheele et al. [86]	Hurricane Sandy (2012)	1,920 tweets labelled informative/uninformative	Geo-tagged tweets based on keyword search
Imran et al. [68]	19 different disasters	50,000 labelled Tweets	Extracting important disaster-related insights
Alam et al. [87]	Hurricane Harvey (2017), Irma (2017), Maria (2017)	50,000 labelled Tweets	Humanitarian categories classification
Alam et al. [18]	19 disasters (2016–2019)	77,000 human-annotated tweets	Human-annotated dataset for humanitarian classification
Imran et al. [88]	COVID-19 pandemic	500 manual and 1,000 crowd-sourced annotations	Sentiment and location recognition
Suwaileh et al. [74]	19 different disasters	5,723 manually annotated tweets, 25,034 automatically labelled	Location extraction
Suwaileh et al. [15]	19 different disasters	3,893 tweets (manually annotated)	Location disambiguation

Further, Ziaullah et al. [89] generated synthetic data, due to the lack of publicly available data, tagging critical infrastructure with impacts, severity, and operational statuses of limited critical infrastructure for Broward County, Florida, USA and Christchurch, New Zealand. It was generated with the help of LLMs and also focuses only on two specific disasters.

Overall, early datasets focused on identifying disaster-related content, including humanitarian categorisation and location extraction. Recent datasets incorporated location disambiguation and automated annotation for scalability. However, none directly distinguish impacted locations or capture detailed impact information. This gap under-

scores the need for a fine-grained dataset with specific impacts to affected locations to enable more accurate and actionable disaster analysis.

3. Corpus Creation for Impacted Location Recognition

3.1. *Data Collection*

For the task of location recognition in disaster-related social media data, we selected the publicly available IDRISI-D English corpus (IDRISI-DE) [90], since it provides all location mentions and categories proposed by Alam et al. [18] across a diverse range of disaster types. However, not all of these categories contribute towards the situational awareness of the disaster for emergency response. For example, tweets related to *rescue volunteering or donation efforts* or those providing *caution and advice* do not provide direct evidence of disaster impacts on people or infrastructure. Similarly, *requests or urgent needs* and *displace people and evacuations* are important for humanitarian coordination, but less informative for emergency response.

To address this, we selected three critical information categories from IDRISI-DE that are highly relevant to disaster response: (1) *injured or dead people*, (2) *infrastructure and utility damage*, and (3) *missing or found people*, which resulted in a reduction of tweets from IDRISI-DE for the annotation. We refer to this newly publicly available dataset as the Disaster Impacted Location Corpus (DILC)[3].

3.2. *Corpus Annotation*

The development of DILC[4] is based on annotating the impact and impacted location mentioned within disaster-related social media posts.

The annotations provide a refined dataset that explicitly identifies disaster impacts and impacted locations, improving the accuracy and relevance of location recognition in disasters. To achieve this, a manual analysis of each tweet was performed by two annotators, who were responsible for identifying only the impact and locations that were directly impacted by the disaster. Additionally, an author acted as a meta-annotator to resolve conflicts and ensure consistency, enhancing the qualitative credibility of the dataset. We crafted annotation guidelines in Appendix A that contain examples of impacted locations versus non-impacted locations and how to identify different types of impacts. We selected BRAT[5] as the annotation tool for the manual annotation process. Each annotator held at least a master's degree and possessed basic knowledge of the task.

Inter-Annotator Agreement (IAA) calculated among annotators based on Cohen's kappa for the impact was 61%, and for impacted locations was 81%. While the kappa score for the impact is modest, this level of agreement is consistent with prior disaster annotation studies where interpreting impact-related information introduces subjectivity [91]. The higher agreement for the impacted locations suggests that they are easier to detect, unlike impacts that often overlap and are expressed ambiguously in social media posts.

[3]DILC available at: `https://github.com/masseygeoinformaticscollaboratory/DILC`

[4]This study was evaluated under the Massey University ethical policies and procedures, was judged to be low risk and has been peer approved with approval number 4000030237)

[5]BRAT is an online environment for collaborative text annotation, available at: https://brat.nlplab.org/index.html

The resulting DILC data set labels both the impact and the impacted location of 19 large disaster events of different types (e.g., wildfires, cyclones, earthquakes, floods, etc.) spanning 11 countries, covering both native and non-native English-speaking regions, and contains 1461 tweets annotated with 3359 types of impact and 1831 impacted locations out of 2649 locations mentioned.

4. Methods for Location Recognition

In this section, we describe the methods used for location recognition, focusing on two distinct tasks: the general all location extraction task and the proposed task of identifying impacted locations. We categorise the location recognition approaches into three main types: (1) Traditional NER Tools (2) Pre-trained Language Models, and (3) Large Language Model (LLM) approaches.

4.1. *Traditional NER Tools*

The first category involved the use of traditional NER tools for location recognition, specifically SpaCy and Flair. Both tools were applied to extract all of the locations mentioned in disaster-related social media posts. SpaCy[6] is a widely used NLP toolkit for NER tasks, and we utilised its *en_core_web_trf* model. It is the largest English model optimised for speed and accuracy in recognising named entities, including locations. SpaCy employs a transition-based dependency parser, combined with CNNs and transformer-based embeddings, enabling efficient context-sensitive token classification. Flair [92], on the other hand, employs contextual string embeddings and character-level language models, offering robust performance in NER tasks. On top of these embeddings, Flair employs a BiLSTM-CRF architecture: the BiLSTM captures dependencies across the sentence in both directions, while the CRF layer ensures consistent sequence labelling. It enables cases where the same word may carry different meanings in different contexts to be distinguished. Flair *v0.15.0* was selected as it is the latest model that integrates character-level language, which allows it to handle noisy text, misspellings, and rare words commonly found in disaster-related social media posts.

4.2. *Pre-trained Language Models*

The second approach leveraged pre-trained language models, including BERT [45], XLM-RoBERTa [93], and GLiNER [94]. Both BERT and XLM-RoBERTa use a bidirectional transformer encoder, which allows them to learn context from both the left and right sides of a token at once. Context-aware representation is particularly effective in identifying locations within disaster-related posts, where location mentions often appear in varied syntactic structures[95]. We employed the fine-tuned BERT-NER model *dslim/bert-base-NER*[7] trained on location data [96]. We also experimented with XLM-RoBERTa-*large-finetuned-conll03-english*[8], a multilingual transformer model with cross-lingual representations, fine-tuned for English NER. It is capable of handling diverse linguistic contexts and offers strong generalisation, making it particularly effective for

[6]SpaCy, available at: https://spacy.io

[7]BERT-NER, available at: https://huggingface.co/dslim/bert-base-NER, last accessed: 3-2-2025

[8]XLM-RoBERTa, available at: `https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-english`

recognising region-specific place names in local languages (e.g., *Alibag, Kaikoura*) commonly found in social media data. GLiNER[9] is a label-aware NER model that is designed to take the target entity labels (e.g., LOC, PER, ORG) as an explicit input during the prediction process, and it leverages pre-trained language encoders like XLM-RoBERTa. We employed GLiNER *2.13* to extract all location entities, as it has demonstrated superior performance over LLMs in zero-shot and low-resource NER tasks [94].

4.3. *Large Language Model (LLM)-based approaches*

The third approach involved the use of LLMs, an advanced form of pre-trained language models that learn contextual representations from large-scale text data using transformer architectures. These pre-trained language models are post-trained (fine-tuned or instruction-tuned) on domain data for task adaptation. In this study, we denote models as either **pre-trained** or **fine-tuned** to clearly indicate their respective training stages and usage contexts. We experimented with several state-of-the-art LLMs for their capacity to identify locations within disaster-related social media content, with a focus initially on their effectiveness in recognising all locations mentioned. These included the following:

- Mistral[10], a 7 billion parameter model that is used for its contextual reasoning capabilities.
- Qwen-2.5[11] with 7 billion parameters, which has strong instruction-following ability, multilingual coverage, and robust reasoning.
- Gemma[12] with 2 billion parameters, which is a lightweight yet powerful model designed for efficiency and multilingual coverage, making it suitable for noisy, user-generated text.
- Phi-3 [13], a medium-sized model with 14 billion parameters optimised for reasoning and factual consistency that enables the model to handle complex linguistic contexts.
- Llama-3[14], 3 billion - 70 billion parameters, which is optimised for understanding and generating human-like text.

These LLMs were tested to identify the best-performing LLM for comparison against traditional NER tools and pre-trained language models.

4.4. *Prompt Engineering*

In our LLMs experiments, we employed several prompt engineering techniques to evaluate their effect on two tasks: all location recognition and impact and impacted location recognition. The prompt types utilised in our research include Basic, Persona, and Chain-of-thought, drawing on established strategies and reasoning-based prompting [97, 98, 51, 61]. For each prompt type, we experimented with zero-shot, one-shot, and 6-shot settings.

We utilised ChatGPT-4 to formulate prompts, leveraging its strong instruction-following and reasoning capabilities to optimise response quality. An iterative process was applied to small sets of tweets to refine prompt instructions. To avoid non-geographic contexts (e.g., "*Greece*" from *#prayforGreece*), we added counter-examples

[9]GLiNER, available at: `https://github.com/urchade/GLiNER`
[10]Mistral, available at:`https://huggingface.co/docs/transformers/en/model_doc/mistral`
[11]Qwen 2.5 available at: `https://arxiv.org/abs/2412.15115v1`
[12]Gemma, available at: `https://deepmind.google/models/`
[13]Phi-3, available at:`https://huggingface.co/docs/transformers/en/model_doc/phi`
[14]Llama 3, available at:`https://huggingface.co/meta-llama`

in the prompt [99], to restrict extraction to valid, standalone locations. For all location recognition, the prompt was modified to track duplicate mentions, which we resolved by enforcing structured output that required each location to be listed with its frequency [56].

The final versions of the refined prompts, incorporating these constraints, are provided in Table 3.

Table 3.: Prompt Design and Examples for identifying all locations in disaster-related social media posts.

Prompt Design for all locations

Basic Prompt — Zero Shot

Full prompt. Accurately identify all the place names (locations mentioned) in the tweet below and only respond with exact place names along with their occurrence frequency in the tweet. Keep the original spellings of place names as they appear in the tweet text and do not add any explanation in the response. Any disaster event reference, organization names, or combined hashtags such as #Keralafloods should not be considered as place names.

Output format:

```
Location mentioned: Location 1 (number of occurrences), Location 2
(number of occurrences), ...
```

Persona — Zero Shot

Full prompt. Act as an NER that recognizes all locations worldwide. Your task is to respond with exact location names along with their occurrence frequency in the tweet. Keep the original spellings of place names as they appear in the tweet text and do not add any explanation in the response. Any disaster event reference, organization names, or combined hashtags such as #Keralafloods should not be considered as place names.

Output format:

```
Location mentioned: Location 1 (number of occurrences), Location 2
(number of occurrences), ...
```

Chain-of-Thought — Zero Shot

Full prompt. Basic Prompt + Think step-by-step and then provide only the final formatted answer.

Output format:

```
Location mentioned: Location 1 (number of occurrences), Location 2
(number of occurrences), ...
```

Examples — Location Recognition

#E1 — Tweet

"#keralafloods #chengannur My parents are still stranded in Madavana, Pandanad, Chengannur. No food or water has reached theme yet. Please help."

```
Locations mentioned: Chengannur (2), Madavana (1), Pandanad (1)
```

#E2 — Tweet
RT @0000000000000 When you need where you want im ready for help , #Yunanistan #Greece #PrayForGreece #PrayForAthens
`Locations mentioned: Yunanistan (1), Greece (1)`

#E3 — Tweet
RT @0000000000000000 Kashmir #Eaethquke 4 deaths and 100 injured #Mirpur #Pakistan administered #Kashmir
`Locations mentioned: Kashmir (2), Mirpur (1), Pakistan (1)`

#E4 — Tweet
"Thinking of Kaikoura, those trapped on State highway 1, Nelson & Wellington affected by the 7.5 quakes in North Canterbury @0000000 #eqnz"
`Locations mentioned: Kaikoura (1), highway 1 (1), Nelson (1), Wellington (1), Canterbury (1)`

#E5 — Tweet
"The African Union (AU) has released US$350,000 of emergency funding to Mozambique, Zimbabwe and Malawi in the aftermath of cyclone Idai. Mozambique, as the worst affected country, will receive US$150,000 of that money."
`Locations mentioned: Mozambique (2), Malawi (1), Zimbabwe (1)`

#E6 — Tweet
"655-unit oilsands work camp near Fort McMurray destroyed by wildfire"
`Locations mentioned: Fort McMurray (1)`

The prompts for extracting the impact and the impacted location were also developed through an iterative refinement process. Table 4 shows the refined impacts and impacted locations prompts.

Table 4.: Prompt Design and Examples for identifying disaster-related impacts and impacted locations in social media posts.

Prompt Design for Impact and Impacted Location

Basic Prompt — Zero Shot

Full prompt. Analyse the given tweet to identify the types of disaster-related impact words (verbs or adjectives) and their corresponding impacted locations that require action from responders. Do not extract a location that has no direct impact associated with it or is vague (e.g., the region, suburban city). List the types of impacts and impacted locations as they are mentioned in the tweet text without adding details, descriptions, explanations, or hashtags.

Output format:
`Types of Impact: <comma-separated list>`
`Impacted Location: <comma-separated list>`

Persona — Zero Shot

Full prompt. Act as an Emergency responder. Your task is to extract disaster-related impact words (verbs or adjectives) and their corresponding impacted locations from tweets that require action from responders. Do not extract a location that has no direct impact associated with it or is not explicit (e.g., "the region"). List the types of impacts and impacted locations as they are mentioned in the tweet text without adding details, descriptions, explanations or hashtags.

Output format:

```
Types of Impact: <comma-separated list>
Impacted Location: <comma-separated list>
```

Chain-of-Thought — Zero Shot

Full prompt. Think step by step:+ Basic Prompt

Output format:

```
Types of Impact: <comma-separated list>
Impacted Location: <comma-separated list>
```

Examples — Impact and Impacted Location

#E1 — Tweet
"More than 2500 people died in Uttarakhand flood no one bats an eye! Over 250 people lose their lives in Kerala floods and everyone loses their minds! Fun Fact: North East India is currently also flooded!"
`Types of Impact:` died, lose their lives, flooded
`Impacted Location:` Uttarakhand, Kerala, North East India

#E2 — Tweet
"#Greece PM Tsipras: I accept political responsibility for the tragedy in Mati. He calls on ministers to do the same. - And? That's all? No resignation for at least 87 dead people?"
`Types of Impact:` dead
`Impacted Location:` Mati

#E3 — Tweet
"At least 300 more people are feared dead in #Chimanimani & #Chipinge due to the devastating effects of #CycloneIdai that swept through the country over the weekend, leaving thousands homeless & property damaged."
`Types of Impact:` feared dead, homeless, property damaged
`Impacted Location:` Chimanimani, Chipinge

#E4 — Tweet
"RT @0000000000000 #ItalyEarthquake Death toll in Amatrice alone now at 86, reports Bill Neely, NBC News"
`Types of Impact:` Death
`Impacted Location:` Amatrice

#E5 — Tweet
"St. Johns River at Main Street Bridge is in minor flood stage right now as high tide approaches."
`Types of Impact:` flood
`Impacted Location:` St. Johns River at Main Street Bridge

#E6 — Tweet
"For those who doubted @00000000000000 video of the bus stuck in #Chipinge I saw it,

it stuck in mushy tar, the road is broken so it gonna be stuck for a while »> Zimbabweans struggle with storm floods | Al Jazeera #CycloneIdai"
`Types of Impact:` CycloneIdai, stuck, broken
`Impacted Location:` Chipinge

For impacts, we constrained outputs to core verbs or adjectives (e.g., *dead*), avoiding descriptive (e.g., *people dead*) or numerical phrases (e.g., *100 dead*). For impacted locations, we excluded vague or contextual terms (e.g., *surrounds*) and restricted extraction to explicit, named locations.

We chose "Act as an emergency responder" in the persona prompt so that the model is immediately guided to think like someone whose primary concern is identifying actionable disaster-related information. Emergency responders don't want general descriptions or background context. They need direct, operationally relevant details, and this framing keeps the extraction task tightly focused.

4.5. *Fine-tuned Large Language Models(LLMs)*

In addition to using the pre-trained version of the LLMs, we fine-tuned the best-performing model (Llama) identified in our evaluations in Section 6. Llama 3.3 70b was infeasible due to GPU resource constraints. To overcome this limitation, we chose to fine-tune a more accessible model, Llama 3.2 3b, specifically using the persona prompt variant. Figure 2 shows the approach followed for the LLMs finetuning.

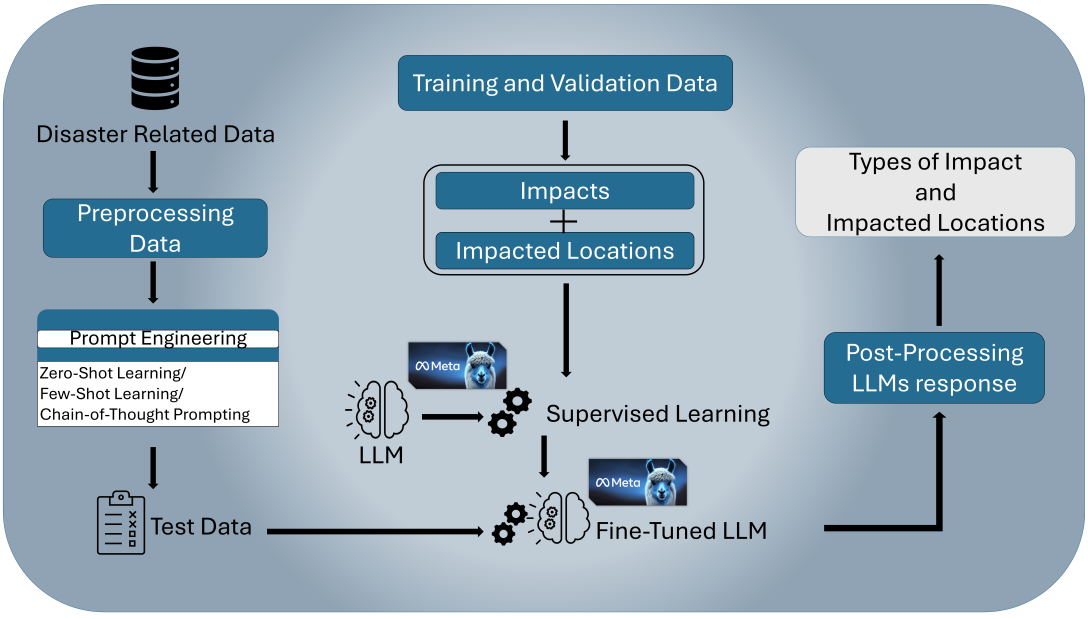

Figure 2.: Workflow for fine-tuning of LLMs

This also reflects a practical design goal of our research, to explore methods that can operate effectively in lower-resourced environments and deliver results swiftly, which is relevant for real-time disaster impact and impacted location mapping, where computational efficiency and timely deployment are important. The fine-tuning process employed LoRA (Low-Rank Adaptation) to efficiently adapt the most critical parameters of the

LLMs while freezing the majority of model weights. This method inserts trainable low-rank matrices into specific transformer layers, allowing the model to focus on learning task-specific knowledge without the computational cost of full fine-tuning for our task.

4.6. *Proposed Approach*

LLMs sometimes generate content not present in the input data, which can lead to the inclusion of extraneous locations not found in the original text, known as hallucination [100, 101]. LLMs do not follow a token-level labelling approach, unlike traditional NER tasks [102, 50]. Existing studies have sought to reduce hallucinations through retrieval-augmented generation (RAG), rubric-based self-evaluation, and internal verification techniques such as chain-of-verification or Yes/No self-checking [53, 103, 104, 59, 52].

In our experiments, LLMs sometimes misrepresented the frequency of locations present in the original text, outputting a single instance even when multiple occurrences existed and vice versa. To address these challenges, we developed a novel post-processing approach designed to mitigate hallucinations and ensure accurate frequency representation, illustrated in Fig. 3 for the all locations task.

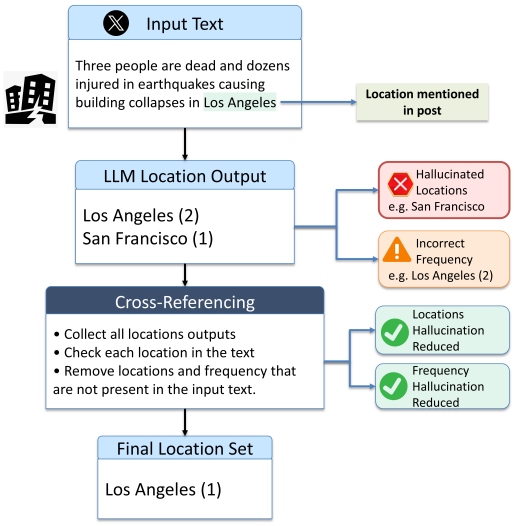

Figure 3.: Post-Processing of LLMs response

The post-processing removes hallucinated outputs by validating all predicted entries against the original text. Unlike LLM-based self-verification, this approach directly compares model outputs with the source text to ensure factual accuracy. In the all location recognition task, post-processing involves checking both the presence and frequency of each predicted location. If a location or its frequency is fabricated by the LLM, either the location is removed or the frequency reduced to match that in the tweet, in the model's output. For the impacted-location task, post-processing also verifies that all predicted impacts and impacted locations appear in the original text; however, it does not perform frequency checking. This approach is applied to reduce hallucinations and improve the reliability of LLM outputs, rather than to modify the model's actual predictions.

5. Experimental Setup

This section outlines the experimental framework. We designed three experiments to address different aspects of the task, ranging from identifying all locations in disaster-related posts to distinguishing between impacted and non-impacted locations and the impacts. The details of all **Experiments 1, 2, and 3**, including models used and specific objectives, are summarised in Table 5.

We identified the best-performing parameter settings through initial experimentation. The settings for the pre-trained LLM models were configured with temperature = 0 to ensure reproducible outputs, essential for factual extraction and top p = 0.9 to capture subtle linguistic variations in all experiments. The Llama 3.2 3b model, implemented using the Unsloth framework, was used for fine-tuning. Key hyper parameters were set to the following: learning rate 2e-4, batch size 8, LoRA rank 16, and LoRA alpha 16. The maximum number of training steps was set to 80, based on early observations of the model.

The models were evaluated based on their precision, recall, and F1-score in identifying the correct locations, impacted locations or impacts (depending on the experiment, see Table 5)within the social media posts. Additionally, we assessed the models' ability to reduce false positives, particularly in distinguishing non-impacted locations from those that were directly affected by the disaster. There can be one or more predicted locations that are impacted, and these may fully or partially match the actual location mentions. So, our evaluation is based on a classification-based evaluation framework where if predicted and actual locations match, it is considered a True Positive (TP), if the location predicted by the model was not the actual location, it is assigned as a False Positive (FP), and if the actual location is not predicted at all, we consider it as a False Negative (FN), consistent with prior disaster-domain evaluations [15]. In cases where multiple locations are present, each predicted location is evaluated separately. If one location is correctly predicted, it contributes to TP, while other incorrectly predicted or missing locations contribute to FP or FN, respectively.

Table 5.: Overview of experiments, models, and purposes

Exp.	**Name**	**Models Used**	**Purpose**
1	All Locations Recognition	Gemma-2 Mistral Phi-3 Llama-3 Qwen-2.5	Identify the best performing open access LLM for recognition of all locations from disaster-related tweets.
2	NER and LLM Comparison	SpaCy Flair GLiNER XLM-RoBERTa BERT-NER Llama-3.3 70b	Compare pre-trained language models and traditional NER tools with the best-performing LLM to establish baseline performance for the all locations recognition task.
3	Impacted Location and Impact Type extraction	Llama-3.3 70b-(pre-trained) Llama-3.2 3b (fine-tuned)	Evaluate the LLM pre-trained model for identifying impacted locations and impacts from disaster-related tweets, and assess improvements after fine-tuning on the complete disaster dataset. Furthermore, both the all-disaster and disaster-specific fine-tuned models were tested on unseen disaster types, which were not included in fine-tuning, to evaluate domain specialisation.

6. Results and Analysis

The goal of **Experiment 1** was to identify the best performing open access model for the all locations task, to help us select which model to use for subsequent experiments. The results are presented in Table 6, which indicates that Llama 3.3-70b outperforms other state-of-the-art LLMs with an F1-score of 0.77, while Qwen2.5-7b, Gemma2-9b, Mistral-7b and phi3-14b achieved F1-scores ranging from 0.60 to 0.73. Thus, Llama 3.3 70b was selected for Experiment 2.

In **Experiment 2**, we compared traditional NER tools and pre-trained NER models with the best-performing LLM to determine which approach performs best at identifying every location mentioned in disaster-related social media posts, regardless of whether the location was impacted or not. Table 7 indicates SpaCy performed the best among all pre-trained NER models, with an F1-score of 0.81. Both Flair and GLiNER followed, each exhibiting a performance of F1-score 0.79. However, the Llama 3.3 70b model outperforms all NER tools, with F1-scores between 0.84 and 0.86 depending on prompt. The superior performance of Llama 3.3 70b over pre-trained NER models confirms the dominance of LLMs for NER through their vast contextual knowledge and instruction-following capability. These results align with recent studies that show LLMs outperform traditional NER models in different domains [47, 105, 78]. Table 7 results also demonstrate that the proposed post-processing approach significantly enhances the effectiveness of the LLMs in Experiments 2.

Table 6.: Experiment 1: Initial comparison of different LLMs on a small dataset without Post-Processing (-PP)

Model	**Version**	**Prompt_Type**	**Precision**	**Recall**	**F1-score**
			-PP	**-PP**	**-PP**
Qwen2.5	7b	Basic zero-shot	0.66	0.83	0.73
Phi3	14b	Basic zero-shot	0.65	0.71	0.68
Gemma2	9b	Basic zero-shot	0.59	0.86	0.70
Mistral	7b	Basic zero-shot	0.49	0.79	0.60
Llama3.3	70b	Basic zero-shot	0.66	0.92	**0.77**

Table 7.: Experiment 2: Performance Comparison of all Location Recognition Models with Post-Processing (PP) and without the Post-Processing (-PP)

Model	**Version**		**Precision**		**Recall**		**F1-score**	
GLiNER	gliner 2.13			0.78		0.80		0.79
Spacy	en_core_web_trf			0.85		0.78		0.81
Flair	flair 15.0			0.78		0.80		0.79
XLM-RoBERTa	large-finetuned-conll03-english			0.62		0.79		0.70
BERT NER	dslim/bert-base-NER			0.67		0.77		0.72
		Prompt_Type	**-PP**	**PP**	**-PP**	**PP**	**-PP**	**PP**
Llama3	3.3, 70B	Basic zero-shot	0.66	0.78	0.92	0.92	0.77	0.85
Llama3	3.3, 70B	Basic one-shot	0.68	0.80	0.93	0.92	0.78	**0.86**
Llama3	3.3, 70B	Basic 6-shot	0.69	0.81	0.91	0.90	0.78	0.85
Llama3	3.3, 70B	Persona zero-shot	0.67	0.79	0.91	0.91	0.77	0.84
Llama3	3.3, 70B	Persona one-shot	0.67	0.80	0.92	0.92	0.78	0.85
Llama3	3.3, 70B	Persona 6-shot	0.68	0.80	0.92	0.91	0.78	0.85
Llama3	3.3, 70B	COT zero-shot	0.67	0.79	0.92	0.91	0.78	0.85
Llama3	3.3, 70B	COT one-shot	0.62	0.76	0.92	0.92	0.74	0.84

We evaluated the Llama 3.3 70b pre-trained model using various prompt engineering strategies of zero-shot to 6-shot with Basic, Persona, and Chain-of-Thought (COT) prompts. The Basic zero-shot Prompt achieved an F1-score of 0.85, and Basic One-Shot Prompt achieved the highest F1-score of 0.86 in identifying all locations. This suggests that when adding more examples, or even using alternate example configurations (which were also tested), the performance remains lower in the 6-Shot prompt compared to One-Shot. This observation is consistent with the findings of Zhao et al. [106], who noted that few-shot prompts often introduce label bias due to imbalanced examples, whereas carefully calibrated one-shot prompts can avoid this bias and yield more accurate and reliable outputs. XLM-RoBERTa, which is a multi-lingual model based on the RoBERTa architecture, demonstrated lower performance than Flair and GLiNER with a F1-score of 0.70. Although Kopanov [58] found XLM-RoBERTa performed well for general NER tasks, its reduced performance in the disaster specific location recognition task suggests that it may lack the specialised contextual understanding required for such domain-specific tasks, due to its pre-training on general-purpose datasets. BERT-NER, although trained on location data, achieved a low F1-score of 0.72 due to its limited contextual understanding to reason beyond token-level boundaries, and difficulty in handling the informal and noisy language often present in disaster-related tweets.

The most successful LLM was fine-tuned and applied with persona-based training with 6-shots and post-processing to identify impacted locations and types of impact within the nuanced language of disaster-related social media posts in **Experiment 3**. Table 8 shows the results of impacts and impacted locations extraction task. On the DILC, the pre-trained Llama 3.2 3b model achieved an F1 score of 0.49 for impacts and 0.73 for impacted locations, whereas the variant that was fine-tuned on 68% randomly selected training data and 20% randomly selected testing data without overlapping improved 0.69 for impacts while maintaining a stable 0.74 for impacted locations.

We further evaluated disaster-specific subsets of tweets written in English from two countries (2017 Hurricane Harvey[15] from the United States and the 2016 Kaikoura earthquake from New Zealand[16]) where English is an official language; and two countries (the 2019 Pakistan earthquake[17] and the 2018 Greece wildfires[18]) where it is not, and the benefits of fine-tuning became more pronounced. These disaster specific subsets were derived from the DILC dataset and were selected based on the diversity of disaster types, the volume of available tweets for training and testing, and the variation between native and non native English speaking countries. For each target disaster, a corresponding domain-specific dataset was created, i.e, DILC-E (which contains all tweets that reference an earthquake in the DILC), DILC-W (wildfires tweets from DILC), and DILC-H (hurricane tweets from DILC). From each of these data sets, the tweets that refer to the particular disaster that is being tested were excluded. That is, tweets referring to Hurricane Harvey were removed from DILC-H, and similarly, tweets referring to the 2018 Greece Wildfires were removed from DILC-W. Two variations of DILC-E were formulated, each excluding tweets for the respective disaster. That is, DILC-EN excludes tweets from the 2016 Kaikoura Earthquake and DILC-EP excludes tweets from the 2019 Pakistan Earthquake. Details of fine-tuning and test datasets and their sizes are provided in Table 8

Two fine-tuning settings were applied. First, an all-disaster model was fine-tuned on tweets from multiple disaster types, excluding the target disaster (e.g., DILC without

[15] https://en.wikipedia.org/wiki/Hurricane_Harvey

[16] https://en.wikipedia.org/wiki/2016_KaikÅDura_earthquake

[17] https://en.wikipedia.org/wiki/2019_Kashmir_earthquake

[18] https://en.wikipedia.org/wiki/2018_Attica_wildfires

2016 Kaikoura Earthquake tweets (NZ)) and then tested on the excluded disaster's tweets. Second, a disaster-specific model was fine-tuned on tweets of the same disaster type (e.g., DILC-EN excluding 2016 Kaikoura Earthquake tweets) and then tested on the target event's tweets. Both models were evaluated on unseen disaster data excluded from their respective fine-tuning sets to assess domain specialisation and generalisation across disaster contexts.

For impact extraction, the best performance was observed on the Pakistan Earthquake dataset, where fine-tuning on earthquake data yielded an F1 score of 0.86, compared to 0.83 when fine-tuned on the entire DILC data set (excluding the Pakistan Earthquake tweets). This highlights the effectiveness of disaster-specific fine-tuning in capturing disaster-related linguistic cues for the detection of impacts. This reflects the tendency for some impacts to be specific to certain disaster types (for example, an earthquake might cause a bridge to collapse, but is less likely to cause fire-related impacts).

For impacted location extraction, the highest performance was obtained on the Greece Wildfire dataset, where the model fine-tuned on all disaster data achieved an F1 score of 0.77. This suggests that exposure to a diverse range of disaster-related texts enhances the model's ability to generalise across unseen disaster scenarios when detecting impacted locations.

On all disasters (DILC), the fine-tuned 3b model also surpasses both the pre-trained 3b (0.49; +0.20) and the larger pre-trained 70b (0.60; +0.09). These results further support the evidence that disaster-specific fine-tuning can improve impact extraction. In contrast, disaster-specific fine-tuning does not generally improve impacted location extraction, with negligible performance of NZ (+0.01) and Harvey (-0.01), and notable performance drops particularly for the case of Pakistan (-0.06) and Greece (-0.12).

The decline of the impacted location in disaster-specific fine-tuning is affected by both the size of the fine-tuning corpus and the number of location mentions, influencing model performance. In the test datasets, ratios of impacted locations to all location mentions vary: New Zealand includes 70/98 (71.4 %); Harvey 163/215 (75.8 %); Pakistan 121/244 (49.6 %) and Greece 343/508 (67.5 %), thereby increasing ambiguity in identifying impacted locations that describe the impacts. Another challenge arises from hierarchical location mentions, such as city, province, and country co-occur within the same tweet, making it difficult for the model to distinguish the most specific impacted location. We also observe that the model sometimes extracts impacted locations with added descriptive context (e.g., *Athens* vs *regions surrounding Athens*) or only partially captures locations (e.g.,*Marlborough Sounds* vs *Marlborough*). Additionally, the model sometimes produced hallucinatory responses where the impact is mistakenly identified as the impacted location.

Overall, the LLM models outperformed traditional NER models that are pre-trained on location entities for the all location recognition task. Tailored prompting strategies and effective post-processing further improved the prediction of the all locations recognition task. The comparative evaluation of the all-disaster and disaster-specific fine-tuned models across different disaster datasets demonstrates distinct strengths. The disaster-specific fine-tuning yielded superior performance in identifying impacts, reflecting the advantages of domain-focused learning. In contrast, the all-disaster fine-tuning achieved the best results in impacted location extraction, indicating that exposure to diverse disaster contexts enhances understanding of location information. However, the comparatively lower performance in disaster-specific fine-tuning also underscores the need for further methodological advances, such as incorporating location resolution and hierarchical place representations, to achieve robust and reliable extraction of impacted locations.

Table 8.: Experiment 3: Performance comparison of pre-trained and fine-tuned Llama models for impacted-location and impact-type extraction (Persona 6-shots)

				Dataset Size (Tweets)		Impact			Location		
Disaster	**Model**	**Finetuning**	**Dataset**	**Fine-tune**	**Test**	**P**	**R**	**F1**	**P**	**R**	**F1**
All Disasters	Llama 3.3	No	DILC	-	300	0.66	0.54	0.60	0.62	0.89	0.73
	Llama 3.2	No	DILC	-	300	0.56	0.44	0.49	0.65	0.83	0.73
	Llama 3.2	Yes	DILC	997	300	0.71	0.66	**0.69**	0.77	0.70	**0.74**
Kaikoura Earthquake (NZ)	Llama 3.2	Yes	DILC ex Kaikoura Earthquake	1401	60	0.67	0.72	0.69	0.67	0.66	0.66
	Llama 3.2	Yes	DILC-EN	252	60	0.68	0.78	0.73	0.68	0.66	0.67
Hurricane Harvey (USA)	Llama 3.2	Yes	DILC ex Hurricane Harvey	1316	145	0.65	0.66	0.66	0.74	0.77	0.76
	Llama 3.2	Yes	DILC-H	255	145	0.70	0.66	0.68	0.75	0.75	0.75
Pakistan Earthquake (PK)	Llama 3.2	Yes	DILC ex Pakistan Earthquake	1372	89	0.92	0.76	0.83	0.62	0.50	0.55
	Llama 3.2	Yes	DILC-EP	223	89	0.93	0.79	**0.86**	0.54	0.46	0.49
Greece WildFire (GR)	Llama 3.2	Yes	DILC ex Greece Wildfire	1149	312	0.71	0.65	0.68	0.78	0.77	**0.77**
	Llama 3.2	Yes	DILC-W	70	312	0.79	0.73	0.76	0.66	0.63	0.65

7. Discussion

The results reveal varying strengths and limitations of LLMs in handling diverse disaster types, information categories, and linguistic complexities. One recurring challenge was span variation, which often led to partial matches in both the impacts and the impacted locations. That is, models frequently predicted only the base form of an impact or impacted location, while the gold standard included an extended form (e.g., *Jatlaan* vs. *Jatlaan Canal*, *found dead* vs. *dead*), or vice versa. Future work should therefore employ span-based metrics and consider overlap-based measures such as Soft Jaccard to provide partial credit for nearly missed predictions. Additional difficulties arose from granularity mismatches, such as when multiple locations were mentioned in hierarchical order (e.g., *Jhelum, Pakistan* two location levels, but only the city *Jhelum* represents the directly impacted area.) or when locations of different levels mentioned separately (e.g., *#Athens #Greece*), often leading to confusion in selecting the appropriate level of specificity. These inconsistencies reflect the high variability and complexity of location usage in disaster contexts.

8. Conclusion and Future Work

Social media provides real-time, ground-level information during disasters, often revealing critical details about affected locations and impacts. This study introduced LLMs to extract types of impacts and impacted locations from disaster-related posts. For this purpose we developed the DILC Corpus, based on 19 major natural disasters spanning 11 countries, covering both native and non-native English speaking regions.

The results show that LLMs, particularly Llama, outperform conventional NER approaches when guided by context-aware prompts and disaster-specific data. Despite challenges such as span variation and granularity mismatches, fine-tuning led to improvements across disaster types. The findings highlight the value of domain adaptation and structured post-processing for reliable extraction of actionable information. Overall, this work demonstrates the potential of LLMs to transform unstructured social media content into structured insights for enhanced situational awareness and decision support in disaster response.

In the future, we aim to extend the current framework by incorporating multilingual datasets, thereby improving the models' adaptability across diverse linguistic and cultural contexts. In parallel, geocoding extracted locations for precise resource targeting and developing granular impact information on severity and affected entities (e.g., *buildings, roads, people*) will further elevate the practical utility of LLMs in disaster preparedness, response, and recovery [107]. By addressing these directions, LLMs can be further advanced into robust, end-to-end systems for comprehensive disaster information management.

Funding

This project was partially supported by QuakeCoRE, a New Zealand Tertiary Education Commission-funded Centre of Research Excellence: QuakeCoRE publication number 1108 and partially by Massey University.

References

[1] Zhijie Sasha Dong, Lingyu Meng, Lauren Christenson, and Lawrence Fulton. Social media information sharing for natural disaster response. *Natural Hazards*, 107 (3):2077–2104, July 2021. ISSN 1573-0840. . URL `https://doi.org/10.1007/s11069-021-04528-9`.

[2] Sajjad Ahadzadeh and Mohammad Reza Malek. Earthquake Damage Assessment Based on User Generated Data in Social Networks. *Sustainability*, 13(9):4814, January 2021. ISSN 2071-1050. . URL `https://www.mdpi.com/2071-1050/13/9/4814`. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[3] Fatemeh Zakeri and Gregoire Mariethoz. Synthesizing long-term satellite imagery consistent with climate data: Application to daily snow cover. *Remote Sensing of Environment*, 300:113877, 2024.

[4] Maria Sigopi, Cletah Shoko, and Timothy Dube. Advancements in remote sensing technologies for accurate monitoring and management of surface water resources in africa: an overview, limitations, and future directions. *Geocarto International*, 39(1):2347935, 2024.

[5] Firouz A Al-Wassai and NV Kalyankar. Major limitations of satellite images. *arXiv preprint arXiv:1307.2434*, 2013.

[6] Farhad Samadzadegan, Ahmad Toosi, and Farzaneh Dadrass Javan. A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: current status and prospects. *International journal of remote sensing*, 46(3):1327–1402, 2025.

[7] Jirapa Vongkusolkit and Qunying Huang. Situational awareness extraction: a comprehensive review of social media data classification during natural hazards. *Annals of GIS*, 27(1):5–28, January 2021. ISSN 1947-5683. . URL `https://doi.org/10.1080/19475683.2020.1817146`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19475683.2020.1817146.

[8] Gabriele Scalia, Chiara Francalanci, and Barbara Pernici. CIME: Context-aware geolocation of emergency-related posts. *GeoInformatica*, 26(1):125–157, January 2022. ISSN 1573-7624. . URL `https://doi.org/10.1007/s10707-021-00446-x`.

[9] Daniel Sui and Michael Goodchild. The convergence of gis and social media: challenges for giscience. *International journal of geographical information science*, 25(11):1737–1748, 2011.

[10] Siqing Shan, Feng Zhao, and Yigang Wei. Real-time assessment of human loss in disasters based on social media mining and the truth discovery algorithm. *International Journal of Disaster Risk Reduction*, 62:102418, 2021.

[11] Reem Suwaileh, Tamer Elsayed, Muhammad Imran, and Hassan Sajjad. When a disaster happens, we are ready: Location mention recognition from crisis tweets. *International Journal of Disaster Risk Reduction*, 78:103107, 2022.

[12] Hannes Seppänen and Kirsi Virrantaus. Shared situational awareness and information quality in disaster management. *Safety Science*, 77:112–122, August 2015. ISSN 0925-7535. . URL `https://www.sciencedirect.com/science/article/pii/S0925753515000764`.

[13] Cheng Zhang, Wenlin Yao, Yang Yang, Ruihong Huang, and Ali Mostafavi. Semiautomated social media analytics for sensing societal impacts due to community disruptions during disasters. *Computer-Aided Civil and Infrastructure Engineering*, 35(12):1331–1348, 2020. ISSN 1467-8667. . URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12576`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mice.12576.

[14] Abhinav Kumar and Jyoti Prakash Singh. Deep neural networks for location reference identification from bilingual disaster-related tweets. *IEEE Transactions on Computational Social Systems*, 11(1):880–891, 2022.

[15] Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. IDRISI-D: Arabic and English Datasets and Benchmarks for Location Mention Disambiguation over Disaster Microblogs. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors, *Proceedings of ArabicNLP 2023*, pages 158–169, Singapore (Hybrid), December 2023. Association for Computational Linguistics. . URL `https://aclanthology.org/2023.arabicnlp-1.14`.

[16] Kai Sun, Yingjie Hu, Kenneth Joseph, and Ryan Zhenqi Zhou. Galloc: a geoannotator for labeling location descriptions from disaster-related text messages. *International Journal of Geographical Information Science*, pages 1–31, 2025.

[17] Saad Mazhar Khan, Imran Shafi, Wasi Haider Butt, Isabel de la Torre Diez, Miguel Angel López Flores, Juan Castanedo Galán, and Imran Ashraf. A systematic review of disaster management systems: approaches, challenges, and future directions. *Land*, 12(8):1514, 2023.

[18] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda 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 (ICWSM)*, volume 15, pages 933–942, 2021.

[19] Muhammad T Chaudhary and Awais Piracha. Natural disasters—origins, impacts, management. *Encyclopedia*, 1(4):1101–1131, 2021.

[20] Michael K. Lindell and Carla S. Prater. Assessing community impacts of natural disasters. *Natural Hazards Review*, 4(4):176–185, November 2003. ISSN 1527-6988, 1527-6996. .

[21] J Rexiline Ragini, PM Rubesh Anand, and Vidhyacharan Bhaskar. Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 42:13–24, 2018.

[22] J. Lee, D. Perera, T. Glickman, and L. Taing. Water-related disasters and their health impacts: A global review. *Progress in Disaster Science*, 8:100123, 2020. .

[23] G.E.C. Charnley, I. Kelman, K.A.M. Gaythorpe, and K.A. Murray. Traits and risk factors of post-disaster infectious disease outbreaks: A systematic review.

Scientific Reports, 11(1):5616, 2021. .

[24] Mariana Arcaya, Emily J Raker, and Mary C Waters. The social consequences of disasters: Individual and community change. *Annual Review of Sociology*, 46: 229–249, 2020. .

[25] Stephane Hallegatte. The indirect cost of natural disasters and an economic definition of macroeconomic resilience. *World Bank policy research working paper*, (7357), 2015.

[26] C. Zhang, W. Yao, Y. Yang, R. Huang, and A. Mostafavi. Semiautomated social media analytics for sensing societal impacts due to community disruptions during disasters. *Computer-Aided Civil and Infrastructure Engineering*, 35(12): 1331–1348, 2020. .

[27] Q. Huang and Y. Xiao. Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, 4(3):1549–1568, 2015. .

[28] J. Xie and T. Yang. Using social media data to enhance disaster response and community service. In *2018 International Workshop on Big Geospatial Data and Data Science (BGDDS)*, pages 1–4. IEEE, 2018. .

[29] R. Berariu, C. Fikar, M. Gronalt, and P. Hirsch. Understanding the impact of cascade effects of natural disasters on disaster relief operations. *International Journal of Disaster Risk Reduction*, 12:350–356, 2015. .

[30] Yu Xiao, Qunying Huang, and Kai Wu. Understanding social media data for disaster management. *Natural Hazards*, 79(3):1663–1679, December 2015. ISSN 1573-0840. . URL `https://doi.org/10.1007/s11069-015-1918-0`.

[31] B. Hecht, L. Hong, B. Suh, and E.H. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246, May 2011.

[32] Yohei Ikawa, Maja Vukovic, Jakob Rogstadius, and Akiko Murakami. Location-based insights from the social web. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 1013–1016, New York, NY, USA, May 2013. Association for Computing Machinery. ISBN 978-1-4503-2038-2. . URL `https://dl.acm.org/doi/10.1145/2487788.2488107`.

[33] K. Stock. Mining location from social media: A systematic review. *Computers, Environment and Urban Systems*, 71:209–240, 2018.

[34] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38, 2015.

[35] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. . URL `https://aclanthology.org/P05-1045`.

[36] R. Dutt, K. Hiware, A. Ghosh, and R. Bhaskaran. Savitr: A system for real-time location extraction from microblogs during emergencies. In *Companion proceedings of the the web conference 2018*, pages 1643–1649, April 2018.

[37] J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web*, pages 1017–1020, May 2013.

[38] Wenzhong Liu and Xiaohui Cui. Improving named entity recognition for social media with data augmentation. *Applied Sciences*, 13(9):5360, 2023.

[39] Wei Liang Seow, Iti Chaturvedi, Amber Hogarth, Rui Mao, and Erik Cambria. A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, 58(10):1–87, 2025.

[40] Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 43–52, Gold Coast Queensland Australia, July 2014. ACM. ISBN 978-1-4503-2257-7. . URL https://dl.acm.org/doi/10.1145/2600428.2609582.

[41] T.B.N. Hoang and J. Mothe. Location extraction from tweets. *Information Processing & Management*, 54(2):129–144, 2018.

[42] A. Kumar and J.P. Singh. Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33:365–375, 2019.

[43] Zi Chen, Badal Pokharel, Bingnan Li, and Samsung Lim. Location Extraction from Twitter Messages Using a Bidirectional Long Short-Term Memory Neural Network with Conditional Random Field Model. In Cédric Grueau, Robert Laurini, and Lemonia Ragia, editors, *Geographical Information Systems Theory, Applications and Management*, pages 18–30, Cham, 2021. Springer International Publishing. ISBN 978-3-030-76374-9. .

[44] Rhea Mahajan and Vibhakar Mansotra. Predicting Geolocation of Tweets: Using Combination of CNN and BiLSTM. *Data Science and Engineering*, 6(4):402–410, December 2021. ISSN 2364-1541. . URL https://doi.org/10.1007/s41019-021-00165-1.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. . URL https://aclanthology.org/N19-1423.

[46] Chao Fan, Fangsheng Wu, and Ali Mostafavi. A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access*, 8:10478–10490, 2020.

[47] Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou, and Kenneth 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):2289–2318, November 2023. ISSN 1365-8816. . URL https://doi.org/10.1080/13658816.2023.2266495. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658816.2023.2266495.

[48] Wenping Yin, Yong Xue, Ziqi Liu, Hao Li, and Martin Werner. Llm-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of hurricane harvey. *International Journal of Applied Earth Observation and Geoinformation*, 137:104423, 2025.

[49] Dhananjay Ashok and Zachary C. Lipton. PromptNER: Prompting For Named Entity Recognition, June 2023. URL http://arxiv.org/abs/2305.15444. arXiv:2305.15444.

[50] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. GPT-NER: Named Entity Recognition via Large Language Models, October 2023. URL http://arxiv.org/abs/2304.10428.

arXiv:2304.10428.

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[52] Shaan Dhuliawala, Dragomir R Radev, and Kalpesh Krishna. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2306.03868*, 2023. URL https://arxiv.org/abs/2306.03868.

[53] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[54] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18 (6):186357, 2024.

[55] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, 2023.

[56] Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. A novel prompting method for few-shot NER via LLMs. *Natural Language Processing Journal*, 8:100099, September 2024. ISSN 2949-7191. . URL https://www.sciencedirect.com/science/article/pii/S2949719124000475.

[57] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, June 2024. Association for Computational Linguistics. . URL https://aclanthology.org/2024.naacl-long.300.

[58] Kalin Kopanov. Comparative Performance of Advanced NLP Models and LLMs in Multilingual Geo-Entity Detection. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, AICCONF '24, pages 106–110, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400716928. . URL https://dl.acm.org/doi/10.1145/3660853.3660878.

[59] Zhe Ji, Nayeon Lee, Rita Frieske, Tao Yu, Dan Su, Yan Xu, Yujia Zhang, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38, 2023. .

[60] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

[61] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[62] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. *Iscram*, 201(3):791–801, 2013.

[63] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. Tweedr: Mining twitter to inform disaster response. In *International Conference on Information Systems for Crisis Response and Management*, 2014. URL https:

//api.semanticscholar.org/CorpusID:6633523.

[64] Beverly Parilla-Ferrer, Proceso Fernandez, and Jaime IV. *Automatic Classification of Disaster-Related Tweets*. De La Salle University Publishing House, Quezon City, Philippines, December 2014.

[65] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 376–385, 2014.

[66] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 994–1009, 2015.

[67] Cornelia Caragea, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H Tapia, Lee Giles, Bernard J Jansen, and John Yen. Classifying Text Messages for the Haiti Earthquake. 2011.

[68] M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages, 2016.

[69] Anna Kruspe, Julia Kersten, and Friederike Klan. Detection of actionable tweets in crisis events. *Natural Hazards and Earth System Sciences*, 21(6):1825–1845, 2021.

[70] Wang Gao, Changrui Zheng, Xun Zhu, Hongtao Deng, Yuwei Wang, and Gang Hu. Knowledge-injected prompt learning for actionable information extraction from crisis-related tweets. *Computers and Electrical Engineering*, 118:109398, 2024.

[71] R. Lamsal, M. Read, and S. Karunasekera. "Actionable Help" in crises: A novel dataset and resource-efficient models for identifying request and offer social media posts. *arXiv preprint arXiv:2502.16839*, 2025.

[72] {Dat Tien} Nguyen, {Kamela Ali} {Al Mannai}, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, pages 632–635. AAAI press, 2017. Publisher Copyright: © Copyright 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 11th International Conference on Web and Social Media, ICWSM 2017 ; Conference date: 15-05-2017 Through 18-05-2017.

[73] Jaydev Rathod, Girish Rathod, Pooja Upadhyay, and Puja Vakhare. Disaster Tweet Classification using ML. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 523–527, May 2022. . URL https://ieeexplore.ieee.org/abstract/document/9792836.

[74] Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. Idrisi-re: A generalizable dataset with benchmarks for location mention recognition on disaster tweets. *Information Processing & Management*, 60(3):103340, 2023.

[75] Rabindra Lamsal, Maria Rodriguez Read, and Shanika Karunasekera. Crisistransformers: Pre-trained language models and sentence encoders for crisis-related social media texts. *Knowledge-Based Systems*, 296:111916, 2024.

[76] Junhua Liu, Trisha Singhal, Lucienne TM Blessing, Kristin L Wood, and Kwan Hui Lim. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 133–141, 2021.

[77] Xingyu Zhu, Feifei Dai, Xiaoyan Gu, Bo Li, Meiou Zhang, and Weiping Wang. Gl-ner: Generation-aware large language models for few-shot named entity recognition. In *International Conference on Artificial Neural Networks*, pages 433–448. Springer, 2024.

[78] Le Xiao, Yunfei Xu, and Jing Zhao. LLM-DER:A Named Entity Recognition Method Based on Large Language Models for Chinese Coal Chemical Domain, September 2024. URL `http://arxiv.org/abs/2409.10077`. arXiv:2409.10077.

[79] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

[80] Emma L McDaniel, Samuel Scheele, and Jeffrey Liu. Zero-shot classification of crisis tweets using instruction-finetuned large language models. In *2024 IEEE International Humanitarian Technologies Conference (IHTC)*, pages 1–7. IEEE, 2024.

[81] Muhammad Imran, Abdul Wahab Ziaullah, Kai Chen, and Ferda Ofli. Evaluating robustness of llms on crisis-related microblogs across events, information types, and linguistic features. In *Proceedings of the ACM on Web Conference 2025*, pages 5117–5126, 2025.

[82] Gamze Aktuna and Şevkat Bahar-Özvarış. Investigating the aftermath of the türkiye 2023 earthquake: exploring post-disaster uncertainty among syrian migrants using social network analysis with public health approach. *Frontiers in public health*, 11:1204589, 2023.

[83] A. Patel et al. Geolocation extraction from twitter data for earthquake response mapping. *arXiv preprint*, 2025. URL `https://arxiv.org/abs/2503.16509`.

[84] Mohammadsepehr Karimiziarani and Hamid Moradkhani. Social response and disaster management: Insights from twitter data assimilation on hurricane ian. *International journal of disaster risk reduction*, 95:103865, 2023.

[85] M.A. Sit, C. Koylu, and I. Demir. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma. In *Social Sensing and Big Data Computing for Disaster Management*, pages 8–32. Routledge, 2020.

[86] Christopher Scheele, Manzhu Yu, and Qunying Huang. Geographic context-aware text mining: Enhance social media message classification for situational awareness by integrating spatial and temporal features. *International Journal of Digital Earth*, 14(11):1721–1743, 2021.

[87] Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *Proc. of ISCRAM, Rochester, USA*, 2018.

[88] Muhammad Imran, Umair Qazi, and Ferda Ofli. Tbcov: Two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels. *Data*, 7(1), 2022. ISSN 2306-5729. . URL `https://www.mdpi.com/2306-5729/7/1/8`.

[89] Abdul Wahab Ziaullah, Ferda Ofli, and Muhammad Imran. Monitoring Critical Infrastructure Facilities During Disasters Using Large Language Models, April 2024. URL `http://arxiv.org/abs/2404.14432`. arXiv:2404.14432 [cs].

[90] Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. IDRISI-D: English and arabic location mention disambiguation resources and tools over disaster tweets. In *Proceedings of ArabicNLP 2023*, pages 158–169, Singapore (Hybrid), December 2023. Association for Computational Linguistics. . URL `https://aclanthology.org/2023.arabicnlp-1.14/`.

[91] Sarthak Khanal, Maria Traskowsky, and Doina Caragea. Identification of fine-grained location mentions in crisis tweets. *arXiv preprint arXiv:2111.06334*, 2021.

[92] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59, 2019.

[93] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, 2020. URL `https://arxiv.org/abs/1911.02116`.

[94] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, June 2024. Association for Computational Linguistics. . URL `https://aclanthology.org/2024.naacl-long.300`.

[95] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[96] Navjeet Kaur, Ashish Saha, Makul Swami, Muskan Singh, and Ravi Dalal. Bertner: A transformer-based approach for named entity recognition. In *2024 15th international conference on computing communication and networking technologies (ICCCNT)*, pages 1–7. IEEE, 2024.

[97] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[98] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352. Association for Computational Linguistics, 2021.

[99] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[100] Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, pages 1089–1099. PMLR, 2020.

[101] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.

[102] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics, 2016. .

[103] Sewon Min, Mike Lewis, Daniel Khashabi, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. Factscore: Fine-grained factual evaluation of generation. *arXiv*

preprint arXiv:2302.01343, 2023. URL https://arxiv.org/abs/2302.01343.

[104] Sheng Chern, Prakhar Bhargava, Kush Raina, Tong Zhang, and Yasin Altun. Evaluating factual consistency in text generation with external tools. *arXiv preprint arXiv:2305.14355*, 2023. URL https://arxiv.org/abs/2305.14355.

[105] Motasem S Obeidat, Md Sultan Al Nahian, and Ramakanth Kavuluru. Do llms surpass encoders for biomedical ner? In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pages 352–358. IEEE, 2025.

[106] Zhiting Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*, 2021. URL https://arxiv.org/abs/2102.09690.

[107] Sophie Francis, Kristin Stock, Sameeah Noreen Hameed, Harikamba Yandamuri, Dukang Li, Raj Prasanna, Federico Liberatore, Christopher B. Jones, Emma Hudson-Doyle, Nishantha Medagoda, Richard Mowll, Aaron Tregoweth, Gavin Treadgold, and Aaron Waterreus. Annotating and extracting disaster impacts from social media with named entity recognition. Poster presented at the ISCRAM Asia Pacific Conference, November 2022. URL https://www.dropbox.com/scl/fo/nie4jw80eel4yghbw7906/AJtvvF36GegbpZH1XeV3py8/Posters/57%20-%20Sophie?dl=0&e=1&preview=ISCRAM_poster_final_copy.pptx&rlkey=dqkwf1fqt29s7m7boy1y3gx01&subfolder_nav_tracking=1. 7–9 November 2022.

Appendix A. Annotation Guidelines

Annotation Guidelines

Everyone tags 300 common tweets and 612 individually assigned using the following scheme:

Basic Tags:

Disaster impact information:

- **Type of impact**
  - We are looking for text that identifies the type of damage or impact that the item has encountered. For example, this could include malfunction, closure, or physical damage. Impact type (e.g., "death", "injury", "damage to buildings", "misplaced") is even mentioned as a hashtag.
  - Negative statements about impacts (road not damaged) should also be tagged, in which case both words (not damaged) should be tagged. If the tweet contains "destroy and damage" together in a sentence, it should be "destroyed" and "damage" as separate impacts.
  - In case of "death" vs "death toll", keep only death as the impact.
  - Main disasters, Flash floods, wildfires, and earthquakes, etc should be highlighted as the type of impact when associated with the locations.

Example Annotations:

- "many lost their family/crops" → Annotate **"loss" as a type of impact**
- The death toll is rising to 247" → Annotate **"death"** as impact
- At least 17 homes on Black Creek were destroyed. 47 sustained major damage. By @00000000000" → Annotate **destroyed, damage as " impact "**

- **Location information:**
- **Impacted location**
  - Only tag items that describe the location of the impact
  - Identify text that describes the location of the item damaged and the disaster impact area. This could be a city, town or a smaller location area.

    For Example, *"Jatlaan canal embankment is damaged. It is one of the main canal originating from Mangla dam which irrigates Punjab. #earthquake AJK Mirpur"* contains multiple place names. However, only the "Jatlaan canal" is directly impacted, where 'damaged' is the impact and 'canal embankment' is the item affected.

  - Impacted Location mentions within the post, as either misspelt or as an abbreviation/short form (NZ for New Zealand) must also be tagged.
  - Impacted locations mentioned in hashtags such as #dallas #texas will also be tagged without including a hashtag sign. Impacted locations mentioned in hashtags, along with other words, will not be tagged, such as #prayfordallas
  - Buildings such as hotels should be items affected unless it is part of the location, for example, "DoubleTree Hotel Orlando Airport".

Which Tweets to Annotate:

The 1525 will be drawn from the following three classes in the IDRISI data sets for annotation:
-infrastructure and utilities damage
-injured or dead people
-Missing-or-found
Each person should do:

- 300 in common from all disasters (all people tag the same 300 tweet
- After the first 300 tweet annotations, we calculate IAA.
- The rest of the 612 tweets can be annotated afterwards.

Meta-Annotator will use Python code to randomly select 300 entries and distribute the same set among annotators. After addressing any corrections and meeting the requirements, the remaining 612 entries will be distributed for annotation. Once these entries are annotated, the Inter-Annotator Agreement (IAA) will be calculated.

Tagging tool:
Brat tag annotating instructions
Tag types for the data.

For entity tagging
Tagging Entities
You will use the Brat interface to highlight and tag parts of the text as specific entities.
Steps for tagging an entity:

1. **Select the Text**: Highlight the portion of the text that you want to tag as an entity.
2. **Choose the Entity Type**: Once the text is selected, a menu will appear where you can choose the entity type. The available entity types for this project are:
    - **impacttype**
    - **impactedlocation**
3. **Save the Tag**: After selecting the entity type, click to confirm. The tagged entity will appear as highlighted in the text, with the appropriate label attached.
4. **Delete/change Tag**: Double-click the entity label, a dialogue box will appear with the list of labels, select the delete option to delete.

Save the Entities: Once the entities are tagged, confirm the action by saving it.