Information Processing and Management 60 (2023) 103340

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets☆

Reem Suwaileh [a,*], Tamer Elsayed [a], Muhammad Imran [b]

[a] Computer Science and Engineering Department, Qatar University, Qatar
[b] Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University (HBKU), Qatar

ARTICLE INFO

Keywords:
Location mention recognition
Twitter
Geolocation
Disaster management
Dataset
Domain generalizability
Geographical generalizability

ABSTRACT

While utilizing Twitter data for crisis management is of interest to different response authorities, a critical challenge that hinders the utilization of such data is the scarcity of automated tools that extract geolocation information. The limited focus on Location Mention Recognition (LMR) in tweets, specifically, is attributed to the lack of a standard dataset that enables research in LMR. To bridge this gap, we present IDRISI-RE, a large-scale human-labeled LMR dataset comprising around 20.5k tweets. The annotated location mentions within the tweets are also assigned location types (e.g., country, city, street, etc.). IDRISI-RE contains tweets from 19 disaster events of diverse types (e.g., flood and earthquake) covering a wide geographical area of 22 English-speaking countries. Additionally, IDRISI-RE contains about 56.6k automatically-labeled tweets that we offer as a *silver* dataset. To highlight the superiority of IDRISI-RE over past efforts, we present rigorous analyses on reliability, consistency, coverage, diversity, and generalizability. Furthermore, we benchmark IDRISI-RE using a representative set of LMR models to provide the community with baselines for future work. Our extensive empirical analysis shows the promising generalizability of IDRISI-RE compared to existing datasets. We show that models trained on IDRISI-RE better tackle domain shifts and are less susceptible to change in geographical areas.

1. Introduction

During emergencies and natural disasters, social media platforms, such as Twitter, receive information pertinent to situational awareness, urgent needs, and reports of damages to infrastructure (Lorini et al., 2021). Information shared by locals or eyewitnesses often contains the locations of damaged sites or areas with particular urgent needs such as food and medicine. As response authorities need to predict incidents over fine spatial and temporal resolutions for successful response, Twitter data is invaluable for observing situational reports and managing response activities when combined with geolocation information (Grace, Kropczynski, & Tapia, 2018; Hu & Wang, 2020; Kropczynski et al., 2018; Pettet et al., 2022; Reuter, Ludwig, Kaufhold, & Spielhofer, 2016), either at fine-grained (e.g., street, city, or neighborhood) or coarse-grained (e.g., province, district, or country) levels.

There are impressive real use cases of employing the extracted geolocation information from Twitter by relief organizations.[1] For instance, the Ushahidi platform[2] was exploited to map geotagged tweets during the Port-au-Prince earthquake in Haiti in 2010

☆ This document is the results of the Graduate Sponsorship Research Award (GSRA) funded by the Qatar National Research Fund (a member of Qatar Foundation).

* Corresponding author.
E-mail addresses: rs081123@qu.edu.qa (R. Suwaileh), telsayed@qu.edu.qa (T. Elsayed), mimran@hbku.edu.qa (M. Imran).

[1] www.hsdl.org/?abstract&did=805223

[2] www.ushahidi.com/

https://doi.org/10.1016/j.ipm.2023.103340
Received 27 September 2022; Received in revised form 25 February 2023; Accepted 26 February 2023
Available online 16 March 2023

for better management. It was also used to map damages and requests during Typhoon Haiyan in 2013. More interestingly, Fairfax County in Virginia, US, explored the usefulness of employing the Geofeedia platform to monitor and aggregate data from various social media platforms including Twitter. According to a survey of emergency officials, geotagging and grouping social media content on a map is perceived as one of the top required features in an information processing system during emergencies (Hiltz et al., 2020). However, as users tend to set imprecise geotags to their tweets, Twitter announced that they are removing the geotagging feature in tweets, in June 2019.[3] This change had increased the importance of developing automatic tools for extracting geolocation information.

Nevertheless, there is no public unified evaluation framework with all essential components, including annotated datasets, diverse open-source (or public) baselines, and fair evaluation metrics. In fact, the absence of large and generalizable LMR datasets makes the comparison difficult between the existing LMR models. Additionally, the existing English LMR tweet datasets are either nonpublic non-disaster-specific (Al Emadi, Abbar, Borge-Holthoefer, Guzman, & Sebastiani, 2017; Das & Purves, 2020; Inkpen, Liu, Farzindar, Kazemi, & Ghazi, 2015; Ji, Sun, Cong, & Han, 2016; Kumar & Singh, 2019; Li & Sun, 2014, 2017; Sultanik & Fink, 2012; Zhang & Gelernter, 2014), or disaster-specific but suffer from several limitations (Al-Olimat, Thirunarayan, Shalin, & Sheth, 2018; Dutt, Hiware, Ghosh, & Bhaskaran, 2018; Fernández-Martínez, 2022; Gelernter & Balaji, 2013; Hu & Wang, 2020; Hu, Zhou, Li, et al., 2022; Khanal, Traskowsky, & Caragea, 2022; Middleton, Middleton, & Modafferi, 2014; Molla & Karimi, 2014; Wallgrün, Karimzadeh, MacEachren, & Pezanowski, 2018) such as the limited size, the confined domain and geographical coverage, the absence of location type annotations, among others.

In this work, we focus on the Location Mention Recognition (LMR) task that aims at extracting location mentions (LMs) from the textual content of tweets. To address the aforementioned drawbacks, we introduce IDRISI-RE,[4] the largest manually-labeled (*gold* version) and automatically-labeled (*silver* version) tweet dataset for LMR comprising 19 disaster events, whose tweets are labeled to identify both toponyms and their geographical types. It covers disaster incidents that occurred in 22 English-speaking countries.

To demonstrate the *domain* and *geographical* generalizability of IDRISI-RE, we empirically answer the following research questions. In comparison to existing datasets, can an LMR model that is trained on IDRISI-RE generalize to:

- Unseen events of the *same* disaster type? (**RQ1**)
- Unseen events of *different* disaster types? (**RQ2**)
- Unseen events that happen in the *same geographical* areas? (**RQ3**)
- Unseen events that happen in *different geographical* areas? (**RQ4**)

Our rigorous empirical analysis demonstrates that IDRISI-RE is the best domain and geographically generalizable LMR Twitter dataset for the disaster management domain, compared to all public datasets of its kind. We also found that both the geographical coverage and the data size are the top influencers on the generalizability of the LMR datasets. Additionally, IDRISI-RE shows a descent reliability level, and reasonable geographical, domain, temporal, and location granularity coverage. Furthermore, a thorough experimental evaluation of a representative set of LMR models show that BERT_{LMR} model is the state-of-the-art LMR model over IDRISI-RE dataset.

The contributions of this work are as follows:

- We present IDRISI-RE, the largest *manually-labeled* publicly-available English LMR dataset of about 20.5k tweets (gold version) for the LMR task.[5] It covers diverse disaster types and geographic areas around the globe. We also release the largest *automatically*-labeled LMR dataset (silver version) constituting 57k tweets.
- We annotate the extracted location mentions in IDRISI-RE into coarse- and fine-grained location types to enable building more accurate LM recognition and disambiguation models, and to allow finer evaluation and comparison between LMR models.
- We benchmark the IDRISI-RE using diverse and representative LMR models to establish a set of baselines for the interested community.
- We empirically demonstrate that IDRISI-RE is the best *domain*- and *geographically*- generalizable dataset for LMR compared to the existing datasets.

The remainder of the paper is organized as follows. We discuss the related work in Section 2. We define the LMR task in Section 3. We then list the design objectives of IDRISI-RE dataset in Section 4. We discuss the dataset creation and annotation processes in Section 5. We then analyze IDRISI-RE for reliability, consistency, coverage, and diversity, and discuss its limitations in Section 6. We benchmark IDRISI-RE with diverse and representative LMR models and discuss the results in Section 7. We further empirically study its *domain* and *geographical* generalizability and answer the related research questions in Section 8. To show the value of IDRISI-RE for the research community, we discuss several use cases in Section 9. We finally conclude and list a few future directions in Section 10.

[3] https://twitter.com/TwitterSupport/status/1141039841993355264

[4] Named after Muhammad Al-Idrisi, who is one of the pioneers and founders of advanced geography: https://en.wikipedia.org/wiki/Muhammad_al-Idrisi. The "R" refers to the recognition task and the "E" refers to the English language.

[5] https://github.com/rsuwaileh/IDRISI/

Table 1
Comparison between IDRISI-RE and the existing NER and LMR datasets. "*" indicates the disaster-related datasets, entirely or partially. "–" indicates the information that we could not obtain.

Dataset	# Twt	# LM (unique)	Annotation	LM Type	Public
Twitter NER datasets					
Ritter, Clark, Etzioni, et al. (2011)	2,400	276 (193)	In-house	×	✓
Liu, Zhang, Wei, and Zhou (2011)	12,245	–	In-house	×	×
Li et al. (2012)	7,750	–	In-house	×	×
Gelernter and Zhang (2013) *	4,488	2,866 (–)	Translation	×	×
WNUT2017 (Derczynski, Nichols, van Erp, & Limsopatham, 2017) *	2,296	773 (559)	In-house	×	✓
BTC (Derczynski, Bontcheva, & Roberts, 2016) *	9,551	3,114 (1,295)	In- & Crowd	×	✓
General Twitter LMR datasets					
Sultanik and Fink (2012)	500	99 (–)	In-house	–	×
Zhang and Gelernter (2014)	956	1,393 (779)	In-house	–	×
Inkpen et al. (2015)	6,000	4,369 (–)	In-house	–	×
Ji et al. (2016)	3,611	1,542 (–)	In-house	–	×
Li and Sun (2014, 2017)	3,570	2,056 (906)	Automatic	–	×
Kumar and Singh (2019)	5,107	3,230 (–)	In-house	–	×
Disaster-specific Twitter LMR datasets					
GEL (Gelernter & Balaji, 2013) *	3,987	–	In-house	✓	×
MID (Middleton et al., 2014) *	3,996	2,030 (451)	In-house	×	✓
ALTA (Molla & Karimi, 2014) *	3,003	4,854 (1,704)	Crowd	×	✓
OLM (Al-Olimat et al., 2018) *	4,500	5,323 (1,619)	In-house	×	✓
DUT (Dutt et al., 2018) *	1,000	~100 (–)	In-house	×	×
GeoCorpora (Wallgrün et al., 2018) *	6,648	3,100 (1,119)	Crowd	×	✓
HU1 (Hu & Wang, 2020) *	1,000	2,139 (989)	In-house	✓	✓
HU3 (Hu, Zhou, Li, et al., 2022) *	3,000	3,530 (1,351)	In-house	×	✓
FGLOCTweet (Fernández-Martínez, 2022) *	9,435	5,958 (3,457)	Automatic	×	✓
KHAN (Khanal et al., 2022) *	9,339	9,655 (1,639)	Crowd	✓	✓
IDRISI-RE					
Gold *	20,514	21,879 (3,830)	Crowd	✓	✓
Silver *	56,682	67,576 (2,675)	Automatic	✓	✓

2. Related work

In this section, we review the Twitter Named Entity Recognition (NER) datasets that were employed in LMR studies, the general Twitter LMR datasets, and the disaster-specific Twitter LMR datasets. We discuss their characteristics and limitations while comparing them to IDRISI-RE. Table 1 summarizes the existing NER and LMR datasets.

2.1. Twitter NER datasets

As LMR is a subtask of NER by definition, different studies explored the effectiveness of the off-the-shelf NER tools for LMR task or retrained their LMR models using NER datasets (Hoang & Mothe, 2018; Lingad, Karimi, & Yin, 2013; Suwaileh, Elsayed, Imran, & Sajjad, 2022; Wang, Hu, & Joseph, 2020). The data availability is a key obstacle for this line of research. To elaborate, only half of the Twitter NER datasets presented in Table 1 are public (Derczynski et al., 2016, 2017; Ritter, Clark, Mausam, & Etzioni, 2011). The main drawback of Ritter (Ritter, Clark, Etzioni, et al., 2011) and WNUT2017 (Derczynski et al., 2017) datasets is the small number of *Location* entities. The Broad Twitter Corpus (BTC) (Derczynski et al., 2016) is the largest public Twitter NER dataset employed for the LMR task. It offers around 2,852 *Location* entities only. Nevertheless, the disaster-specific datasets are empirically preferable over the general-purpose datasets, e.g., BTC (Derczynski et al., 2016), for training LMR models in the disaster domain Suwaileh, Imran, Elsayed, and Sajjad (2020). We note here that there is a burgeoning literature on NER models and datasets (Hu, Zhou, Li, et al., 2022; Nasar, Jaffry, & Malik, 2021; Yadav & Bethard, 2018), but we solely list the datasets that were exploited in the LMR research. For an exhaustive list of Twitter NER models and datasets, Hu, Zhou, Li, et al. (2022) evaluated the effectiveness and efficiency of 27 approaches over 26 public datasets (many of them are adopted from NER studies) from different data domains (e.g., social media posts and newswire).

2.2. General Twitter LMR datasets

The LMR problem is of interest to many domains such as emergency management (refer to Section 2.3), traffic monitoring (Alkouz & Al Aghbari, 2020; Paule, Sun, & Moshfeghi, 2019; Shang, Zhang, Youn, & Wang, 2022), POIs recommendation (Zhang et al., 2017; Zhao, Zhao, King, & Lyu, 2017), geographical text analysis and retrieval (Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012; Purves, Clough, Jones, Hall, & Murdock, 2018), to name a few. There are a few existing general English LMR tweet datasets (refer to the second group of datasets in Table 1). These datasets are not event-centric, however, they are useful to evaluate the generalizability of the LMR models to different domains other than the disaster domain. Although the size of these datasets is fair and the annotation is done by in-house annotators, none of them is public due to the issue of third-party copyrights.

2.3. Disaster-specific Twitter LMR datasets

Despite the abundance of disaster datasets that are made available by academic researchers (Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013; Imran, Mitra, & Castillo, 2016; Olteanu, Vieweg, & Castillo, 2015), a few of them are labeled for the LMR task. In this section, we review both disaster-specific Twitter Recognition and Disambiguation (refer to definitions in Section 3) datasets as the latter could be used to develop and evaluate LMR models.

Generally, the existing LMR disaster-specific datasets (refer to the third group in Table 1) are limited in size with the largest being FGLOCTweet dataset which constitutes only 9435 *automatically-labeled* tweets and 3457 unique LMs (Fernández-Martínez, 2022). All other *manually-labeled* LMR datasets contain around half of the unique LMs in the FGLOCTweet dataset. For example, KHAN being the largest human-annotated dataset contains only 1639 unique LMs. Our IDRISI-RE dataset offers 3830 *manually-labeled* and *automatically-labeled* LMs. Additionally, the existing LMR datasets suffer from the confined domain and geographical coverage. For instance, in all public natural disaster event-centric English LMR datasets, five flood events happened in Australia, India, the UK, and the US (ALTA, KHAN, and OLM), 3 hurricane events happened in the US (ALTA and KHAN), and two earthquake events happened in New Zealand (MID) and Nepal (KHAN).

Another group of LMR datasets are the keyword-based datasets. For example, the GeoCorpora dataset was collected using general disaster keywords such as "earthquake", "flood", "fires", among others. The geographical coverage of LMs is dominated by the United States (42%), and United Kingdom (12%). The remaining LMs cover other different countries. Albeit reasonable diversity, the small number of data per disaster type forms the main barrier to exploiting these datasets. FGLOCTweet dataset was also collected by tracking general disaster keywords such as "earthquake", "car accident", "bombing attack", etc. Unfortunately, we could not obtain its geographical distribution for comparison. The major weakness of the keyword-based datasets is the information loss associated with limiting the sample to tweets containing specific keywords that may not appear in many informative tweets.

Moreover, a few datasets contain location type annotations of LMs (Gelernter & Balaji, 2013; Hu & Wang, 2020; Khanal et al., 2022; Middleton et al., 2014). GEL (Gelernter & Balaji, 2013) contains 4,000 tweets collected during the 2011 Christchurch Earthquake in New Zealand. The LMs are annotated into four categories including street, building, toponym, and abbreviation. However, the GEL dataset is not available to the research community. MID dataset (Middleton et al., 2014) contains three types of locations including "admin", "building", and "transport". However, it suffers from annotation issues such as incompleteness (Middleton, Kordopatis-Zilos, Papadopoulos, & Kompatsiaris, 2018). KHAN dataset (Khanal et al., 2022) offers categories of locations that requires further annotations for lower-level location types. For example, all fine-grained locations (e.g., buildings, landmarks) are labeled into one category called "lan". We note that the major issue in KHAN dataset is the noisiness associated with its broad definition of LMs. For example, ambiguous locations are labeled such as "cities", "college", "earth", "elsewhere", "at home", "inside", to list a few. HU1 dataset (Hu & Wang, 2020) is also labeled for location types but it contains solely 1000 tweets sampled from Hurricane Harvey 2017 event, thus it is limited in size, disaster domain, and geographical coverage. To overcome this shortcoming in existing datasets, we collect location type annotations for all LMs in IDRISI-RE.

Furthermore, during disasters, new LMs within the affected areas emerge in the Twitter stream which demands a long temporal coverage of data during the entire disaster period. The OLM dataset is the only one that we could analyze its temporal coverage because other datasets do not contain the IDs (MID), or the event notion is ignored (keyword-based datasets) when they are collected or released. Using the tweets that we managed to crawl at the time of this writing, we found that the OLM dataset misses critical periods of the disaster events, especially during the Chennai Floods, 2015. To elaborate, the flood happened between 8 Nov–14 Dec 2015,[6] but the covered period in the dataset is between 2–4 Dec 2015. While constructing IDRISI-RE, we ensure that the sampled tweets cover the critical periods of disaster events.

While an LMR dataset could cover all relevant topics discussed during the respective disaster, it has to mainly contain informative tweets that are useful for the response authorities. Unfortunately, solely the ALTA dataset is labeled for relevance and the nonpublic GEL dataset is labeled for informativeness. When constructing IDRISI-RE, we aim to select events that are already filtered for relevance and contain informative tweets.

A key issue of the existing LMR datasets is the inconsistency of the "location mention" definition between and within datasets. Indeed, the guidelines used to train annotators are rarely discussed (Wallgrün et al., 2018; Zhang & Gelernter, 2014). We release our annotation task instructions that articulate our "location mention" definition and further elaborate on this issue in Section 6.1. Furthermore, the tweets' metadata is required for incorporating multimodal features for LMR systems such as utilizing different types of social networks or simulating real-time processing, among others. Nevertheless, the public datasets lack some metadata. For instance, the MID dataset does not provide tweet identifiers for all events, the ALTA dataset releases only the tweet identifiers, and OLM and GeoCorpora datasets release some of the tweet attributes such as identifiers and text but not timestamps, user identifiers, etc. The FGLOCTweet dataset contains only the BIO-scheme textual annotations. While the full JSON data of the original collection is available on request, one has to replicate the preprocessing (e.g., removing duplicates and near duplicates) to get the final version.

[6] https://en.wikipedia.org/wiki/2015_South_India_floods

3. Task overview

During disaster events, response authorities rely on situational information to operate (e.g., make decisions or take actions). However, with the abundance of information captured on Twitter, the response authorities need automatic tools to extract informative and reliable tweets that locate incidents and requests. In this work, we consider the *Location Mention Recognition* (LMR) task that aims to *automatically extract toponyms (places or location names) from text*.

To distinguish the LMR from other tasks, we emphasize that the LMR task aims at removing geo/non-geo ambiguity of tokens in text. It is also known as *location extraction* or *geoparsing* in the literature. Differently, the Location Mention Disambiguation (LMD), which is a consecutive task for LMR, aims at removing geo/geo ambiguity between candidate LMs extracted by LMR systems. The LMD task is also known as *location resolution*, *location linking* (looking up a geo-positioning database), or *geocoding* (assigning geo-coordinates to LMs) in the literature.

There are two task setups for LMR. The first recognizes toponyms without their types, denoted as "type-less recognition", and the second distinguishes between types of LMs (e.g., country, city, and street), denoted as "type-based recognition". The latter better serves the development and evaluation of geolocation processing systems in light of the responders' needs. It enables a variety of downstream tasks (e.g., crisis maps) at different location granularity, in addition to being crucial for accurately disambiguating the toponyms.

4. Design objectives

Creating LMR datasets for practical, event-centric, and fine-grained evaluation requires identifying a set of characteristics that guide the dataset construction efforts. Grounded on our review of past efforts (refer to Section 2), we introduce the set of characteristics that we anticipate can form an optimal LMR dataset in the following:

- O1. ***Geographical coverage***: The naming conventions of places vary from one country to another which decisively affects the performance of LMR models. The wider the geographical coverage of an LMR dataset, the more naming conventions it captures. While constructing IDRISI-RE, we aim to capture various naming conventions by annotating disaster events that cover a large range of English-speaking countries.
- O2. ***Domain coverage***: At the onset of disaster events, acquiring training data is impractical and expensive. Alternatively, an acceptable performing LMR model could be trained using previous disasters of the same type (i.e., in-domain data) (Suwaileh et al., 2022). As such an approach is infeasible due to the limited domain coverage of existing datasets, we aim to cover a variety of disaster types with larger number of tweets when constructing IDRISI-RE.
- O3. ***Location type annotations***: The location types (e.g., cities, streets, POIs, etc.) allow customizing the downstream applications to meet the responders' needs, such as generating crisis maps at different granularities. Additionally, the evaluation per location type shows the weaknesses and strengths of LMR models based on the responders' preferences. When constructing IDRISI-RE, we aim to annotate the LMs into location types to overcome the deficiency of existing LMR datasets.
- O4. ***Large-scale***: Learning models, in particular deep neural networks, are data hungry. Models trained on a large number of training examples tend to yield higher performance and generalize to unseen data. However, most of the existing LMR datasets are limited in size (refer to Table 1). We aim to overcome this shortcoming while creating IDRISI-RE by annotating larger number of LMs compared to the existing datasets.
- O5. ***Temporal coverage***: As new LMs emerge in Twitter stream during disaster events, longer temporal coverage of the disaster events is demanded to provide geographical-aware situational reports to responders throughout the disaster event. While existing datasets do not show reasonable temporal coverage of disaster events, we aim to overcome this issue while creating IDRISI-RE.
- O6. ***Relevance and informativeness***: An LMR dataset has to contain informative and actionable tweets to support effective disaster management. Unlike existing datasets, in constructingIDRISI-RE, we extend a dataset that is already labeled for informativeness. This simulates the expected input to the LMR models in real-world information processing systems for disaster management.

We emphasize here that all of these objectives together constitute a generalizable LMR dataset and should be collectively achieved to eliminate any barrier against establishing an effective and fair evaluation framework for LMR. We elaborate on how IDRISI-RE achieves these objectives throughout the paper.

5. Dataset annotation

Two main factors guided the choice of our underlying dataset. First, while responders look for *informative posts on Twitter*, the tweets become more invaluable in the presence of geographical context (Hiltz et al., 2020). Second, *the likelihood of LMs occurrence* increases during events (Kitamoto & Sagara, 2012). Consequently, we selected an *event-centric* dataset that is already labeled for humanitarian categories to simulate the deployment phase of LMR models in real-world information processing systems for disaster management. We analyzed multiple existing disaster-related tweet datasets and selected HumAID (Alam, Qazi, Imran, & Ofli, 2021) for its geographical wide coverage and disaster domain diversity.

We carried out two annotation versions, namely the *gold* annotations using crowdsourced human-labeling, and the *silver* annotations using a learned model. Before creating our pool of tweets for the gold annotations, we dropped the less informative

Table 2
Tweet and Location Mention statistics of IDRISI-RE dataset.

Version	Tweets	$\text{Tweets}_{\|LM\|=0}$	LMs (uniq)
IDRISI-RE_*gold*	20,514	5,723	21,879 (3,830)
IDRISI-RE_*silver*	56,682	25,034	43,404 (2,675)

classes (to relief authorities) of HumAID including *sympathy and support*, *not humanitarian*, *do not know or cannot judge*, and *other relevant information*. The gold annotations contain only tweets that belong to one of the following humanitarian categories: caution and advice, displaced people and evacuations, infrastructure and utility damage, injured or dead people, missing or found people, requests or urgent needs, and rescue volunteering or donation effort.

Using our overall cost budget of $4300, we estimated a maximum of 21k tweets to label for the *gold version* of the dataset. Using this upper bound estimate, we equally sampled a representative number of tweets from each of the 19 disaster events. This led us to randomly sample a maximum of 1300 tweets per event. As some events contain fewer tweets that fall within the humanitarian categories of interest (inherited from HumAID dataset) than our sample size per event, we included all their tweets in the sample. We used stratified sampling to inherit the distribution of the humanitarian classes from the HumAID dataset. We show the number of tweets of the *gold* and the *silver* versions in the "Tweets" column in Table 2.

5.1. Gold annotations

To collect the gold LMR annotations, we used the *Appen* crowdsourcing platform[7] due to its cost efficiency in labeling large datasets. In the annotation task, the textual content of tweets is automatically tokenized by the platform using the SpaCy NLP tool.[8] The workers were asked to (1) highlight the location spans in text (one token or more) that we refer to as location mentions (LMs), and (2) assign the most accurate location types from a predefined list of fine- and coarse-grained location types for the potential LMs. The location types include "Continent", "County", "State", "City/town", "District", "Island", "Neighborhood", "Road/street", "Human-made POI" (features that are built by humans such as schools and hospitals), "Natural POI" (features that are part of the land such as rivers and seas), and "Other locations" (when LMs do not fall in any of the previous types). We provided detailed annotation guidelines for annotators with examples to clearly articulate our definition of location mentions.[9]

Following Appen's recommendation, we randomly picked around 88 tweets for quality control. For workers to be eligible to begin and continue working on the annotation task, their annotation accuracy (i.e., trust score) should not fall below 70% while performing the task. To increase the reliability of the final annotations, we configured the task to collect three annotations per tweet; however, if the agreement level is below a minimum confidence of 80%,[10] we allowed dynamically-collecting up to five more workers to annotate the tweet, achieving a maximum of eight annotations per tweet. We ran our crowdsourcing task for around three weeks and collected annotations for 20,527 tweets from all the disaster events.

To decide the final set of gold LMs, we selected the text spans that received at least two votes from annotators, regardless of the agreement on their location types. Moreover, as the nature of the annotation task allows overlapping annotated spans, we favored the overlapped span with the maximum number of votes by annotators. In case of ties, we selected the longest span. To ensure the quality of labels, we deleted all annotated spans of length equals to or longer than 70% of the length of the original tweet text as we considered them spam or human errors. As a result, we dropped around 13 annotations from all events.

As for the location types, while we cannot prevent the human errors in the crowdsourced annotations, we rely on two factors to increase the reliability of the location type annotations: (i) the local annotators' agreement on the types assigned to a potential LM, and (ii) the global distribution of the types assigned by all annotators to the occurrences of the potential LM within the event's tweets. We achieved the former factor via majority voting. We employed the latter in case of ties. Moreover, we plan to extend the IDRISI-RE dataset for the LMD task in which annotators correct the location types of LMs (whenever needed) while disambiguating them.

Table 2 shows the final number of tweets (column "Tweets"), the number of tweets with no LMs (column "$\text{Tweets}_{|LM|=0}$"), and the total number of annotated LMs with the unique LMs in parentheses (column "LMs (uniq)").

5.2. Silver annotations

Thus far, we discussed the process of acquiring gold annotations using human workers. To increase the size of the dataset beyond our limited budget, we automatically amplify the size of IDRISI-RE by using an automatic-labeler, that is the best performing LMR model on the *gold* annotations (refer to Section 7). More specifically, we trained a BERT-based model using the entire *gold* annotations of IDRISI-RE (all events combined). We then ran the resulted model on the tweets that were not sampled for the *gold* annotations from all the 19 disaster events, including the tweets that belong to the low informative classes. Out of this process, we

[7] http://success.appen.com
[8] https://spacy.io/
[9] https://github.com/rsuwaileh/IDRISI/tree/main/annotation_guidelines
[10] The confidence level is computed by adding up the confidence scores of the contributed workers.

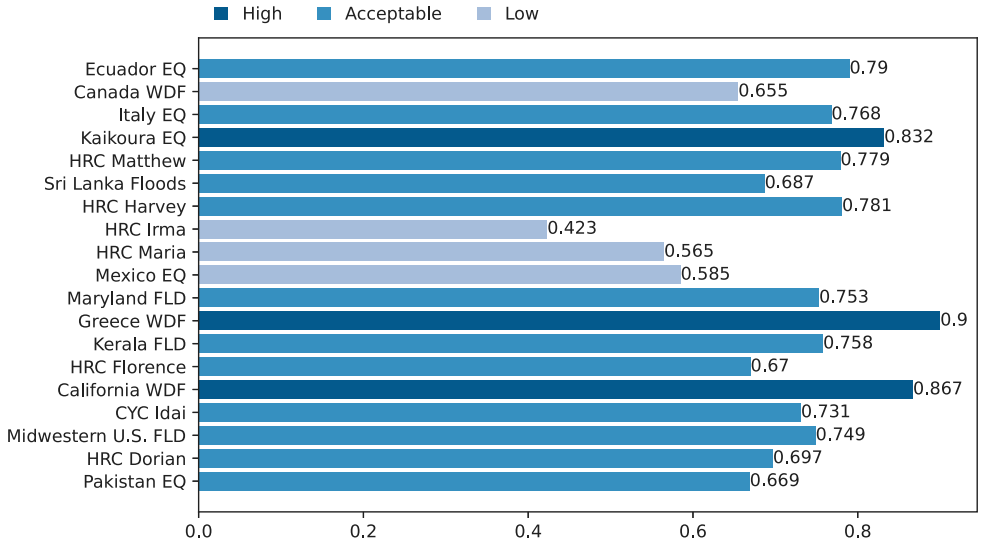

Fig. 1. $k-\alpha$ for IDRISI-RE per disaster event.

constructed the largest automatically-labeled LMR dataset comprising 56,682 tweets. We denote this version as *silver* to imply its level of reliability, and report its statistics in Table 2. We anticipate it to be useful for training better performing LMR models and supporting research on advanced learning techniques (e.g., transfer learning and domain adaptation).

6. Dataset description and quality

In this section, we present a thorough evaluation of IDRISI-RE in terms of reliability, consistency, coverage, and diversity in Sections 6.1 and 6.2, respectively. We further discuss its limitations in Section 6.3.

6.1. Reliability and consistency

To evaluate the quality of IDRISI-RE, we computed the Inter-Annotator Agreement (IAA) that quantifies the reliability of annotations. We further compared the LM definition against the existing LMR datasets.

Annotation Quality: We computed Krippendorff's alpha ($k-\alpha$) (Krippendorff, 1970) to measure the reliability of annotations. Unlike Fleiss Kappa, $k-\alpha$ does not require a fixed number of votes per example. We have two types of annotations: location mentions (*LOC*) and location types (*TYPE*). Due to the class imbalance in token-level classes (having dominant non-LOC tokens compared to the LOC tokens), computing the $k-\alpha$ for the *LOC* annotation is unreasonable, as we will get an almost full agreement (a score of 1), since annotators highly agree on non-LOC tokens. Thus, we only report $k-\alpha$ for *TYPE* annotation (which implicitly encodes the *LOC* annotation). Fig. 1 shows the $k-\alpha$ per disaster event in IDRISI-RE. We only consider the LMs that received two votes or more. As a result, annotators achieve approximately 71.5% IAA across all disaster events, showing an acceptable reliability. Overall, the IAA shows that the annotations are highly reliable for three events, acceptable for twelve events, and of low quality for four events.

LM Definition across Datasets: Table 3 compares the definition of LMs in the public disaster-specific LMR datasets. Columns "Hashtags", "Mentions", "URLs", and "Location Expressions (LEs)" refer to whether these tokens and expressions are considered LMs or not in the corresponding datasets. Table 4 presents example tweets from IDRISI-RE to articulate our LM definition and distinguish it from other datasets. In the existing LMR datasets, an LM can be a substring of a hashtag (tokens start with "#"), however, in IDRISI-RE we only consider a hashtag as a potential LM if it is entirely an LM (e.g., Tweet #1 in Table 4). The locations within user mentions (tokens start with "@") are considered LMs in the ALTA and KHAN datasets, while they are ignored in all other datasets. Although user mentions could indicate the location of incidents that the tweet discusses, we do not consider them as LMs in IDRISI-RE, because they typically refer to organizations or people entities, not locations. We follow the same intuition for URLs. Furthermore, in ALTA, OLM, KHAN, and FGLOCTweet datasets, the LEs and addresses are annotated as a whole, but in IDRISI-RE we differentiate between LMs and LEs; an LE has to be broken down into its locational units. This is mainly because our focus in the LMR task is to detect geographical units. Detecting the LEs as a whole requires an additional text processing layer. For example, in Tweet #3, the annotators have to label "Mohra-Saang" and "Jatlan" separately as two LMs, not the entire expression "Mohra-Saang, a village 1 km away from Jatlan". We follow the same intuition for the full addresses and routes. For instance, in Tweet #4, the consecutive LMs have to be labeled independently.[11]

[11] The annotation guidelines used to construct IDRISI-RE are available in the GitHub repository.

Table 3
Comparison between IDRISI-RE and the existing LMR dataset in the annotation guidelines for the special cases of Location Mentions.

Dataset	Hashtags	Mentions	URLs	LEs
MID (Middleton et al., 2014)	✓	✓	×	×
ALTA (Molla & Karimi, 2014)	✓	✓	✓	✓
OLM (Al-Olimat et al., 2018)	✓	×	×	✓
GeoCorpora (Wallgrün et al., 2018)	✓	×	×	×
HU1 (Hu & Wang, 2020)	✓	×	×	×
HU3 (Hu, Zhou, Li, et al., 2022)	✓	×	×	×
KHAN (Khanal et al., 2022)	✓	✓	×	✓
FGLOCTweet (Fernández-Martínez, 2022)	✓	✓	×	✓
IDRISI-RE	✓	×	×	×

Table 4
Example tweets from IDRISI-RE dataset. The single-underlined and double-underlined LMs represent the undesired and desired LMs, respectively, in our annotation guidelines.

#	Tweet text
1	To all my followers please RT: Where to #Donate to #Mexico #Earthquake Victims - @nytimes #PrayForMexico
2	Stay safe @california Camp Fire burns over 6700 structures and 9 dead become the most destructive fire in #California history. A state of emergency was declared in @ButteCounty in response to the growing …
3	Mohra-Saang, a village 1km away from Jatlan #Earthquake has been levelled. Not a single house left in the village. 3 confirmed dead so far, More than hundred injured. Road that leads to village is no more functional.
4	Flooding. roadway closed in #SilverSpring on Sligo Crk Pkwy Both NB/SB between Piney Branch Rd and Maple Ave #DCtraffic

6.2. Coverage and diversity

In this section, we discuss how IDRISI-RE satisfies the properties presented in Section 4.

Geographical Coverage: To ensure that IDRISI-RE can train generalizable models that are effective in future disaster events, it has to cover different naming conventions of locations that are used in different countries (refer to **O1** in Section 4). The disaster events in IDRISI-RE are indeed geographically-spread over several countries across continents, including Canada, Colombia, Cuba, Dominican Republic, Ecuador, Greece, Haiti, India, Italy, Madagascar, Malawi, Mexico, Mozambique, New Zealand, Pakistan, Peru, Puerto Rico, Sri Lanka, The Bahamas, Turks and Caicos Islands, The United States, and Zimbabwe.

Domain Coverage: To remedy the lack of diversity in disaster types (refer to **O2** in Section 4), IDRISI-RE has to cover the frequently-happening natural disaster events in the English-speaking countries during the past decade (between 2010–2019) that are earthquakes, floods, hurricanes, cyclones, and wildfires (Alam, Joty, & Imran, 2018; Imran et al., 2013, 2016; Nguyen, Ofli, Imran, & Mitra, 2017; Olteanu, Castillo, Diaz, & Vieweg, 2014; Olteanu et al., 2015). IDRISI-RE contains diverse events including six hurricanes, five earthquakes, four floods, three wildfires, and one cyclone.

Location Types Coverage: To support advanced development and finer evaluation of LMR models, we labeled IDRISI-RE for fine- and coarse-grained location types (refer to **O3** in Section 4). Fig. 2 shows the distribution of the location types per disaster event in IDRISI-RE. HRC, EQK, FLD, CYC, and FIR refer to Hurricanes, Earthquakes, Floods, Cyclones, and Wildfires, respectively. The coarse-grained LMs (e.g., Country, State, and City) dominate IDRISI-RE by approximately 89%. Upon further analysis, we found the key factor that explains the dominance of coarse-grain LMs is the HumAID dataset creation method. HumAID was collected by tracking relevant keywords to the disaster events which are usually the name of the coarse-grained impacted areas. Indeed, these coarse-grained LMs are less challenging to detect by annotators. Consequently, we could not prevent annotators from detecting them nor reduce their frequency in the dataset. Furthermore, annotators are more likely to disagree on fine-grained LMs, hence the annotations of potential fine-grained locations are more probable to be discarded when we had initially selected the gold annotations from the crowdsourced data. To mitigate this issue, we provided the *location type annotations* that allows researchers to evaluate the LMR models at different location granularity. We also reported the number of unique LMs for all datasets in Table 1 showing that IDRISI-RE contains the maximum number of unique LMs (3830 LMs). Figs. 7–10 show the distribution of the top 15 LMs per disaster event in Appendix A.

Temporal Coverage: Ideally, the event-centric datasets should span over the entire period of disaster event to allow the response authorities to efficiently operate during all phases of the disaster events (refer to **O5** in Section 4). The events in IDRISI-RE were crawled two days before and two days after their peak incidents (Alam et al., 2021).

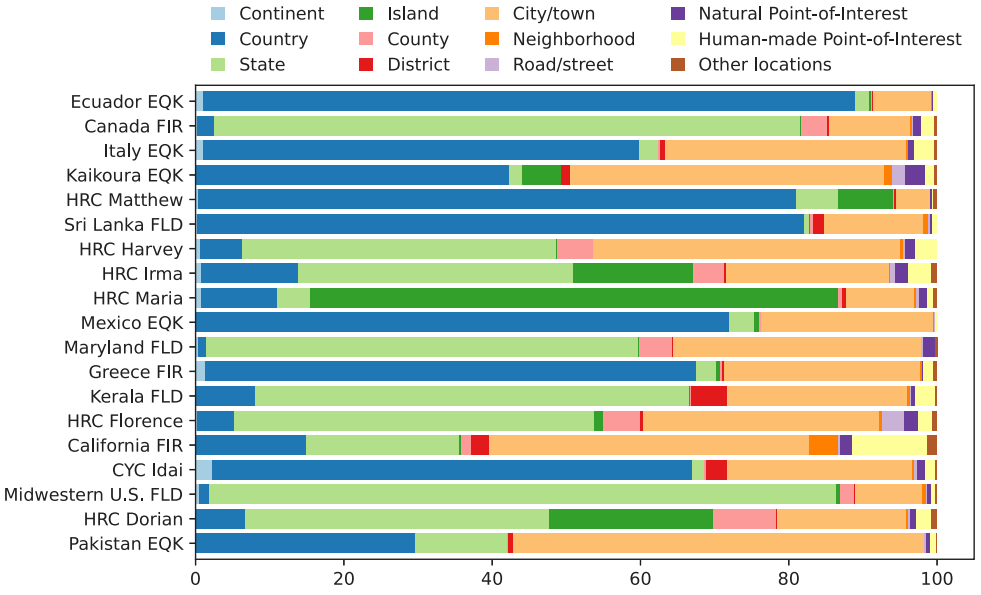

Fig. 2. Distribution of location types in IDRISI-RE. HRC, EQK, FLD, CYC, and FIR refer to Hurricanes, Earthquakes, Floods, Cyclones, and Wildfires, respectively.

6.3. Limitations

Our thorough analysis show that there are shortcomings in the annotations of IDRISI-RE that we discuss here.

- ***Underrepresented Fine-grained LMs***: The fine-grained LMs in IDRISI-RE form solely 9.77% of the LMs in the dataset. Although the skewed distribution is common in the existing datasets, it shapes a limitation when developing LMR models used by downstream applications that rely on fine-grained locations.
- ***Human Errors***: There are some human errors in the crowdsourced annotations that occurred due to the difficulty of the task.
  - In some cases, annotators fail to distinguish between *Location* and *Organization* entities. For example, the "Red Cross" is commonly mentioned as an organization, not the location of its offices, but some of its occurrences are labeled as LMs in IDRISI-RE.
  - Although the annotation guidelines state clearly not to highlight the locations that are mentioned as descriptions within the context of the tweet, this was a confusing case to the annotators.

  We plan to overcome this issue as part of a future Location Mention Disambiguation (LMD) annotation that aims to removing geo/geo ambiguity between candidate LMs.
- ***Temporary Locations***: Temporary facilities (i.e., medical camps, shelters, etc.) are constructed during emergencies to provide resources and support for the affected people. However, these facilities could be disassembled (e.q., quarantine centers) once the emergency is over. Additionally, the names of some locations could change during emergencies, such as allocating a specific school as a shelter and giving it a new expressive name (e.g., "main shelter"). Once the disaster event is over, the school will return to providing its original services. The difficulty of detecting and disambiguating these temporary locations is due to the need for comprehending their context. Although these locations are important for the affected people and response authorities, not all of them are labeled in IDRISI-RE.

7. Benchmarking experiments

To provide baselines for the LMR task, we benchmark IDRISI-RE dataset for different task, data, and disaster domain setups. As for the task setup, we experiment with *type-based* and *type-less* recognition (refer to Section 3). We also use two data setups: *(i) random* and *(ii) time-based*. In the random setup, we ignore tweets' timestamps and randomly select train, development, and test examples. Whereas, in the time-based setup, the data is chronologically ordered before splitting it into training, development, and test sets. We further study the performance of the best model under *(i) in-domain* setup where training and test events are of the same disaster type, and *(ii) cross-domain* setup where training and test events are of different disaster types.

7.1. Learning models

We select a representative set of LMR models for benchmarking IDRISI-RE as described below:

- ***CRF*** (Lafferty, McCallum, & Pereira, 2001): The Conditional Random Fields (CRF) is a competitive probabilistic tagging algorithm, which can be used as a standalone tagger (Li & Sun, 2014, 2017) or integrated into an LMR system (Das & Purves, 2020; Ji et al., 2016; Wang et al., 2020; Xu et al., 2019). We used the *crfsuite* library[12] to train a CRF model using word-level syntactic features, including the identity, suffix, shape, and POS tags. Additionally, we used words contextual tokens such as adjacent words and their syntactical features.
- ***BERT***$_{LMR}$ (Suwaileh et al., 2022): This represents a fine-tuned version of the pre-trained BERT model for the LMR task.
- ***GPNE*** (Hu et al., 2021): This is an *unsupervised* LMR model that specifically shows better detection performance for location mentions within the disaster-hit areas.
- ***GPNE***$_2$ (Hu, Zhou, Sun, et al., 2022): An enhanced version for GazPNE that employs an LMD module to improve the LMR accuracy. It exploits Stanza NER model to accelerate recognition and detect hard LMs. This system uses synthesized training data extracted from gazetteers. Hence, we cannot retrain/finetune it using our data. We use its public CLSTM trained model.
- ***NTPR***$_O$ (Wang et al., 2020): A neural-based toponym recognition tool trained on recurrent neural networks. We run the **o**riginal trained model that is made public by the authors.
- ***NTPR***$_R$: A **r**etrained *NTPR*$_O$ model from scratch on IDRISI-RE dataset per event. We do not tune the hyperparamters and adopt the values used by the authors (Wang et al., 2020).
- ***NTPR***$_F$: A **f**ine-tuned *NTPR*$_O$ on IDRISI-RE dataset per event. Similar to *NTPR*$_R$, we do not tune the hyperparamters.
- ***LORE*** (Martínez & Periñán-Pascual, 2020): an untrainable rule-based recognition model (Martínez & Periñán-Pascual, 2020). We run the original application that is made public by the authors.
- ***nLORE*** (Fernández & Periñán-Pascual, 2021): A deep learning-based model that exploits LORE's rule-based features for recognition. We run the **o**riginal trained application that is made public by the authors. We could not retrain this model or fine-tune it since it is not open source.[13]

7.2. Hyperparameter tuning

During training, we tune the hyperparameters of the BERT$_{LMR}$ model, including the sequence length, the batch size, the number of training epochs, and the learning rate as recommended by Devlin, Chang, Lee, and Toutanova (2019). We experiment with different batch sizes (i.e., 8, 16, 32), the number of epochs (i.e., 2, 3, 4), and learning rates (i.e., 5E−5, 3E−5, 2E−5).

For the CRF-based models, we experiment with five training algorithms, namely Gradient Descent using the L-BFGS method (LBFGS), Stochastic Gradient Descent with L2 regularization term (L2EG), Averaged Perceptron (AP), Passive Aggressive (PA), and Adaptive Regularization Of Weight Vector (AROW). For *LBFGS*, we tune the coefficients for L1 and L2 regularization parameters. For the *L2EG*, we tune the coefficient for L2 regularization and the initial value of the learning rate used for calibration. For *AP*, we tune the epsilon parameter that determines the condition of convergence. For *PA*, we tune the strategy for updating feature weights and the sensitivity parameter that determines whether errors are considered in the objective function. For *AROW*, we tune the initial variance of every feature weight and the tradeoff between loss function and changes of feature weights (gamma). We tune the regularization parameters for values between 0.05 and 1 with a step value of 0.05. We tune the initial learning rate and epsilon using values $\{1 \times 10^i | i \in [2, 6]\}$. The *PA* sensitivity parameter is boolean and the updating strategy includes three types: without slack variables, type I, or type II. We tune the variance and gamma parameters of *AROW* algorithm for values $\{2^{-i} | i \in [0, 3]\}$.

7.3. Evaluation measures

To evaluate the LMR models, we compute the harmonic mean (F_1 score) of Precision (P) and Recall (R). We evaluate LMR models on entity-level rather than token-level. Our evaluation differs from *seqeval*[14] in three aspects: (1) it evaluates per tweet and report the average performance, (2) it rewards the models when they correctly predict no LMs for a single tweet, and (3) it accepts BILOU-like format or JSON formats.

7.4. Benchmarking results

Type-less LMR: In this setup, the LMR models are only required to recognize LMs, regardless of their types. Tables 5 and 6 present the F_1 results of all LMR models over all events. We also report the detailed results, including precision and recall, with the best hyper-parameters in Appendix B for the BERT$_{LMR}$ and CRF models. On average, the BERT$_{LMR}$ model exhibits a compelling performance against all other *type-less* LMR models, for both *random* and *time-based* scenarios. On average, the NTPR$_F$ and NTPR$_R$ models, except the NTPR$_O$ model, show the second-best performance followed by the CRF model for the *random* data setup. In some cases where the NTPR$_F$ and NTPR$_R$ models show the best performance, their absolute results are slightly better than the BERT model. In contrast, the NTPR$_O$ performance is better than the CRF model under the *time-based* data setup. The CRF model's average score on the *time-based* is around 17% lower than the *random* data setup. The LORE and nLORE models exhibit modest performance compared to the other

[12] https://sklearn-crfsuite.readthedocs.io/
[13] It will not be open source in the near future as per the authors.
[14] https://pypi.org/project/seqeval/

Table 5
The F_1 results for the LMR models on IDRISI-RE for the *type-less* LMR task setup and the *Random* data setup.

Event	CRF	$BERT_{LMR}$	GPNE	$GPNE_2$	$NTPR_O$	$NTPR_R$	$NTPR_F$	LORE	nLORE
Ecuador Earthquake	0.866	**0.953**	0.242	0.741	0.840	0.920	0.921	0.653	0.632
Canada Wildfires	**0.732**	**0.732**	0.435	0.683	0.718	0.708	0.727	0.619	0.647
Italy Earthquake	0.558	**0.880**	0.730	0.214	0.828	0.851	0.863	0.200	0.167
Kaikoura Earthquake	0.878	**0.912**	0.594	0.730	0.787	0.906	0.896	0.711	0.756
Hurricane Matthew	0.890	**0.941**	0.141	0.923	0.862	0.915	0.929	0.857	0.882
Sri Lanka Floods	0.856	**0.917**	0.421	0.692	0.654	0.908	0.894	0.735	0.548
Hurricane Harvey	0.810	**0.906**	0.397	0.738	0.788	0.891	0.898	0.672	0.798
Hurricane Irma	0.773	**0.835**	0.369	0.713	0.704	0.814	0.801	0.651	0.735
Hurricane Maria	0.864	**0.925**	0.479	0.779	0.708	0.881	0.865	0.712	0.815
Mexico Earthquake	0.860	**0.929**	0.783	0.759	0.885	0.886	0.902	0.715	0.727
Maryland Floods	0.809	**0.890**	0.754	0.817	0.794	0.869	0.879	0.487	0.737
Greece Wildfires	0.839	0.927	0.792	0.730	0.807	**0.935**	0.929	0.694	0.686
Kerala Floods	0.725	**0.887**	0.664	0.480	0.718	0.863	0.873	0.430	0.441
Hurricane Florence	0.667	**0.755**	0.466	0.535	0.553	0.742	0.738	0.572	0.531
California Wildfires	0.870	**0.920**	0.728	0.760	0.750	0.914	0.905	0.669	0.702
Cyclone Idai	0.892	**0.925**	0.240	0.824	0.716	0.885	0.897	0.472	0.736
Midwestern U.S. Floods	0.904	**0.944**	0.680	0.785	0.772	0.929	0.920	0.706	0.716
Hurricane Dorian	0.820	**0.878**	0.589	0.757	0.760	0.870	0.858	0.616	0.722
Pakistan Earthquake	**0.879**	0.877	0.379	0.770	0.712	0.834	0.849	0.587	0.639
Average	0.815	**0.891**	0.520	0.707	0.756	0.869	0.871	0.619	0.664

Table 6
The F_1 results for the LMR models on IDRISI-RE for the *type-less* LMR task setup and the *Time-based* data setup.

Event	CRF	$BERT_{LMR}$	GPNE	$GPNE_2$	$NTPR_O$	$NTPR_R$	$NTPR_F$	LORE	nLORE
Ecuador Earthquake	0.716	0.916	0.164	0.703	0.854	**0.920**	0.874	0.640	0.563
Canada Wildfires	0.644	**0.767**	0.094	0.696	0.719	0.726	0.721	0.608	0.612
Italy Earthquake	0.504	**0.842**	0.357	0.276	0.777	0.770	0.768	0.234	0.232
Kaikoura Earthquake	0.755	0.896	0.169	0.693	0.769	**0.911**	0.879	0.723	0.731
Hurricane Matthew	0.790	0.944	0.045	0.862	0.866	0.936	**0.945**	0.872	0.892
Sri Lanka Floods	0.740	0.904	0.215	0.679	0.753	**0.919**	0.903	0.756	0.599
Hurricane Harvey	0.599	**0.894**	0.111	0.739	0.820	0.885	0.857	0.659	0.800
Hurricane Irma	0.538	**0.825**	0.111	0.722	0.683	0.805	0.813	0.668	0.732
Hurricane Maria	0.768	**0.904**	0.195	0.733	0.707	0.894	0.861	0.723	0.789
Mexico Earthquake	0.798	**0.911**	0.339	0.734	0.815	0.865	0.884	0.694	0.722
Maryland Floods	0.648	0.845	0.428	0.792	0.794	0.833	**0.892**	0.483	0.663
Greece Wildfires	0.778	**0.883**	0.389	0.767	0.777	**0.883**	0.842	0.706	0.684
Kerala Floods	0.638	**0.923**	0.273	0.575	0.786	0.909	0.888	0.530	0.553
Hurricane Florence	0.465	**0.784**	0.130	0.499	0.562	0.721	0.734	0.614	0.526
California Wildfires	0.832	**0.906**	0.300	0.800	0.764	0.882	0.874	0.715	0.766
Cyclone Idai	0.696	**0.898**	0.169	0.789	0.660	0.866	0.863	0.469	0.727
Midwestern U.S. Floods	0.792	**0.949**	0.440	0.789	0.819	0.927	0.930	0.746	0.754
Hurricane Dorian	0.470	0.862	0.137	0.767	0.791	0.833	**0.864**	0.548	0.639
Pakistan Earthquake	0.723	**0.836**	0.089	0.736	0.669	0.814	0.777	0.605	0.620
Average	0.679	**0.878**	0.219	0.703	0.757	0.858	0.851	0.631	0.663

baselines. GPNE shows poor performance than other baselines. However, its new release ($GPNE_2$) outperforms LORE and nLORE under the *random* and *time-based* data setups and the CRF model in the *time-based* data setup.

Type-based LMR: In this setup, the LMR models are required to recognize the LMs and predict their types simultaneously. Table 7 showed the results of all models over IDRISI-RE dataset. The CRF model is a strong competitor to the $BERT_{LMR}$ model under the *type-based* and shows comparable performance for many events in both *random* and *time-based* settings.

7.5. Domain transfer

We use "domain" to refer to the domain of the target dataset, which is always a specific disaster type, e.g., flood. We study the domain transfer within IDRISI-RE dataset in two setups: "in-domain", where the source and target sets are of the same disaster type, and (ii) "cross-domain", where the disaster type of source and target sets are different.

Experimental Setups: We use the $BERT_{LMR}$ model as it shows the best F_1 scores. We use the *random* data setup for both *type-less* and *type-based* task setups. We tune the hyperparameters of the model (refer to Section 7.2) for each transfer setup over the

Table 7
The F_1 results for the LMR models on IDRISI-RE for the *type-based* LMR task setup.

Data setup	Random		Time-based	
Event	CRF	BERT_{LMR}	CRF	BERT_{LMR}
Ecuador Earthquake	0.932	**0.939**	0.910	**0.926**
Canada Wildfires	**0.853**	0.733	**0.865**	0.771
Italy Earthquake	**0.906**	0.890	**0.881**	**0.881**
Kaikoura Earthquake	0.879	**0.909**	0.875	**0.899**
Hurricane Matthew	0.901	**0.919**	0.899	**0.952**
Sri Lanka Floods	0.910	**0.925**	0.897	**0.912**
Hurricane Harvey	0.906	**0.909**	**0.914**	0.895
Hurricane Irma	**0.906**	0.833	**0.893**	0.823
Hurricane Maria	0.882	**0.924**	0.890	**0.897**
Mexico Earthquake	0.838	**0.913**	0.880	**0.911**
Maryland Floods	0.751	**0.892**	**0.873**	0.805
Greece Wildfires	0.896	**0.925**	0.886	**0.887**
Kerala Floods	**0.880**	**0.880**	0.857	**0.919**
Hurricane Florence	**0.879**	0.772	**0.889**	0.778
California Wildfires	0.907	**0.909**	**0.902**	**0.902**
Cyclone Idai	0.877	**0.900**	0.852	**0.895**
Midwestern U.S. Floods	0.917	**0.936**	0.920	**0.944**
Hurricane Dorian	**0.875**	0.858	**0.865**	0.852
Pakistan Earthquake	0.820	**0.894**	0.780	**0.828**
Average	0.880	**0.883**	**0.880**	0.878

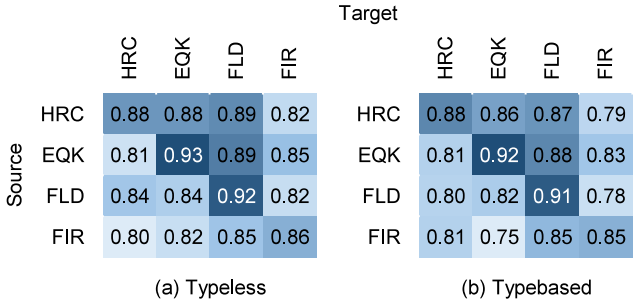

Fig. 3. The F_1 results for the domain transfer experiments within IDRISI-RE. HRC, EQK, FLD, and FIR refer to Hurricanes, Earthquakes, Floods, and Wildfires, respectively.

development sets (same events as the training/source sets). IDRISI-RE covers four disaster types, namely, hurricane, earthquake, flood, and wildfire. A transfer data setup is composed of source-target pair, resulting in 16 setups.

Experimental Results: Fig. 3 illustrates the F_1 scores of the model over the test sets. Below, we elaborate on the results per domain setup:

- ***In-Domain***: As expected, the best results appear on the diagonal, which represents the in-domain setup, for both *type-less* and *type-based* LMR. The high performance shows the advantage of using IDRISI-RE for training LMR models at the onset of disaster events of the same types as the ones offered by IDRISI-RE.
- ***Cross-Domain***: Interestingly, the model achieved a minimum of 80% and 75% of F_1 score for the *type-less* and *type-based* LMR task setups, respectively. This reasonably good performance shows the promising advantage of using IDRISI-RE for training LMR models at the onset of disaster events of different types than the ones offered by IDRISI-RE.

To this end, we confirm that training on IDRISI-RE dataset could generate reasonably performing models in the range of 80% and 75% of F_1 score for the *type-less* and *type-based* LMR, respectively.

8. Generalizability

Generalization allows learning algorithms to identify features and patterns that are universal and not specific to one situation, event, or geographical area. The basic building block that is required to obtain a model's generalizability is its training dataset. However, most existing datasets lack characteristics that are essential to achieve better generalizability. To overcome these

issues, IDRISI-RE dataset is designed to cover data events that span broader geographical locations and cover multiple disaster types/domains, including floods, earthquakes, hurricanes, etc. To this end, we compare the performance of models trained on IDRISI-RE with models trained on seven public datasets, namely, OLM, MID, GeoCorpora (GEO), KHAN, HU1, HU3, and FGLOCTweet (refer to Section 2). We did not use the ALTA dataset because the tweets are not mapped to their corresponding disaster events. This missing mapping prevented us from grouping the tweets by disaster domain and geographical area which is required for running the generalizability experiments.

For all generalizability experiments, we use our BERT_{LMR} model, as it exhibits the best performance in the benchmarking experiments for the *type-less* task setup (refer to Section 7); from hereafter, we refer to it as "the model". We define the *source dataset* as the dataset (or the combination of datasets) used to *train* the model, and the *target dataset* as the dataset used to *test* it. All the experiments are designed using fairness practices that we list below:

- We use the standard training and test splits of the respective data setups for training and testing the model, unless indicated.
- We use the default values of hyperparameters of the model from Hugging Face Transformers,[15] to avoid biasing the model towards any of the datasets.
- We mitigate the influence of training data size on the model performance when comparing different datasets by normalizing the size across all sets. Specifically, we divide the training set, after combining events, into n tweet subsets of the same size as the smallest training set. We apply the size normalization to the training sets of size 70% larger than the smallest training set. We then run n experiments, one for each subset, and report the average performance. We also report the results without size normalization and mark the respective runs with " * ".
- We limit our experiments to only the *random* data setup and the *type-less* task setup; only KHAN and HU1 datasets are labeled for location types. KHAN is labeled for location categories that are higher in granularity compared to IDRISI-RE, which requires manual mapping of annotations. HU1 contains more branched types which requires mapping to common types with IDRISI-RE. It is also limited in size and confined in both domain and geographical aspects, hence it is inadequate for drawing conclusions for generalizability.

8.1. Domain generalizability

We use "domain" to refer to the domain of the target dataset, which is always a specific disaster type, e.g., flood. To this end, we define the *domain* generalizability as *the ability of the model trained on disaster events of a specific domain (source) to generalize and perform well when tested on unseen disaster events (target) of the same domain (denoted as "in-domain" setup) or a different domain (denoted as "cross-domain" setup)*.

8.1.1. Experimental setups

When a dataset contains multiple events of the same type, we randomly choose one of the events as *target* (test set), and the remaining events (combined) as *source* (training set). This is considered a *zero-shot* learning setup in terms of the specific events. We note that *all* of our reported experiments are under zero-shot learning (experiments, where the training and test sets include the same event, are hidden/greyed in the figures). Hence, we use the test splits of Hurricane Dorian 2019, Midwestern US Floods 2019, Puebla Mexico Earthquake 2017, Greece Wildfires 2018, and Louisiana Floods 2016, as the IDRISI.HRC, IDRISI.FLD, IDRISI.EQK, IDRISI.FIR, and OLM.FLD target/test sets, respectively. All remaining events are used for training (only their standard training splits). Table 11 in Appendix C shows the detailed setups for all source and target sets. We follow the same data partitioning method for the event-centric datasets including OLM, MID, HU1, and HU3. We note that the event context is discarded in the released KHAN dataset, hence we manually categorized tweets into their respective events using the tracking hashtags that are made public by the authors. We ended up using only Hurricane Michael 2018 event, since the other events have very few tweets in the order of tens, which is inadequate for training the model (Suwaileh et al., 2022). For the keyword-based datasets, GEO and FGLOCTweet, we split the tweets based on the domains that overlap with IDRISI-RE (earthquake, fire, and flood). For that, we used the tracking keywords used in crawling the dataset to extract matching tweets for each domain Fernández-Martínez (2022), Wallgrün et al. (2018). We excluded the hurricane tweets from FGLOCTweet dataset due to the small size of relevant tweets (only 13 tweets). We then partition each domain's tweets into 70% training, 10% development, and 20% test. We split the GEO dataset because there are no standard splits released for the community. We also split FGLOCTweet dataset since its standard splits become unbalanced after categorizing the tweets by their disaster domain. Furthermore, we also train the model using *IDRISI.ALL* and *GEO.ALL* training sets to show the performance of models trained on all source/training domains for each respective dataset.

8.1.2. Results and discussion

In this section, we discuss the observations we made on the model's performance, analyze the results, and answer the related *domain* generalizability research questions: can an LMR model that is trained on IDRISI-RE generalize to:

- Unseen events of the same disaster type? (**RQ1**)
- Unseen events of different disaster types? (**RQ2**)

[15] https://github.com/huggingface/transformers/blob/a7d73cfdd497d7bf6c9336452decacf540c46e20/src/transformers/training_args.py#L124

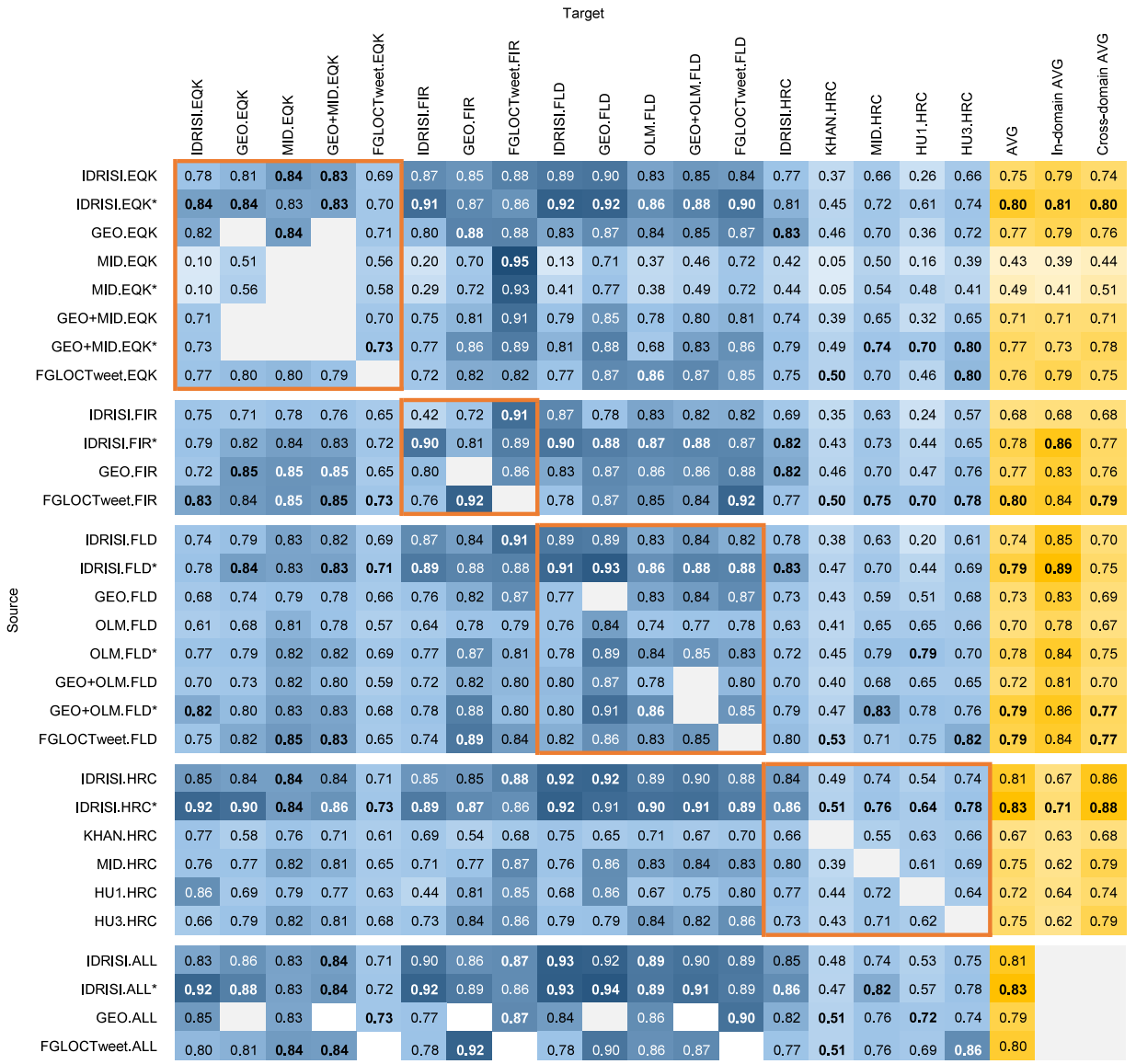

Fig. 4. The F_1 results of the domain generalizability experiments of IDRISI-RE against existing datasets. The best results per column are **boldfaced** column-wise, per disaster domain. EQK, FIR, FLD, and HRC refer to Earthquake, Wildfire, Flood, and Hurricane, respectively.

In-domain: To address **RQ1**, we study the domain generalizability of IDRISI-RE within the same disaster type for source and target sets. The sub-matrices marked in "orange" borders in Fig. 4 presents the F_1 results for the *in-domain* experiments. The "AVG" and "In-domain AVG" columns show the average over *all* and *in-domain* test sets, respectively. We make the following observations:

- *Inconsistent yet reasonable average performance of IDRISI.<domain> source sets*: The models trained on *IDRISI.EQK* consistently outperform *MID.EQK* per target set and on *in-domain* average. Unexpectedly, augmenting the size of source data by merging *GEO.EQK* and *MID.EQK* source sets (*Geo+MID.EQK*) does not improve the performance on majority of the target sets (12 out of 15 sets). The *GEO.EQK* source set alone and *FGLOCTweet.EQK* show better average performance, but both are comparable with *IDRISI.EQK* when looking at the in-domain average performance. Training on *IDRISI.FIR* source set is the worst compared to the other datasets. Further failure analysis is required to understand the reason behind this low performance. The models

trained on *IDRISI.FLD* outperform the ones trained on *GEO.FLD*, *OLM.FLD*, and *GEO+OLM.FLD* as per the *in-domain* average and the total average. The models trained on *IDRISI.HRC* are significantly better than the ones trained on *MID.HRC*, *KHAN.HRC*, *HU1.HRC*, and *HU3.HRC*.

- *Superior performance of IDRISI.<domain>* source sets*: Generally, using IDRISI-RE dataset without size normalization generates the top performing LMR models per target set for all domains and on *in-domain* average. In particular, over all domains, the *IDRISI.<domain>** sources sets consistently generate better models compared to *IDRISI.<domain>* sources sets. These results emphasize the need for large training data to build superior models.
- *Geographical vicinity affects the model performance*: We found that the geographical vicinity of the source and target sets is a potential factor on improving performance. For instance, we found that 40% of the LMs in *GEO.EQK* source set are in the United States, while the events in *IDRISI.EQK* training set happened in Ecuador, Italy, New Zealand, and Pakistan. Having the *IDRISI.EQK* test set containing tweets about an event that happened in Mexico, it is apparent that training on *GEO.EQK* generates a superior model than training on *IDRISI.EQK*.

To answer **RQ1**, we show that IDRISI-RE dataset generates the best domain generalizable models per domain, compared to the other LMR datasets. The only exception is *GEO.EQK* that shows *comparable* performance to *IDRISI.EQK**.

Cross-domain: To address **RQ2**, we study the domain generalizability of IDRISI-RE within different disaster types for source and target sets. Fig. 4 presents the F_1 results for the different setups. The "AVG" and "Cross-domain AVG" columns indicate the average over *all* and *cross-domain* (cells outside the orange boxes) test sets, respectively. We make the following observations:

- *Inferior performance of MID, KHAN, HU1, and HU3 source sets*: Training on these source sets leads to the lowest average performance across all test sets. Upon investigation, we found that the location distribution in Christchurch Earthquake (*MID.EQK*), for example, is highly skewed; the location mention "Christchurch" constitutes approximately 49.7% and 53.8% of the total number of LMs in the training and test sets, respectively. Moreover, around 68% of the tweets in the dataset have no LMs. In KHAN.HRC, "Florida" appears in around 20% and 19% in the training and test sets respectively, and the 10 most frequent LMs constitute 42% and 40% of the training and test sets respectively. For this reason, we believe that these two datasets are inadequate for training generalizable LMR models.
- *Competitive performance of FGLOCTweet.<domain> source sets*: In general, these source sets exhibit better performance compared to *IDRISI.<domain>*, in FIR and FLD domains. Upon investigation, we found that, unlike the *FGLOCTweet.EQK* source set that US dominates its top 20 LMs (constituting 46% of the LMs in the dataset), both *FGLOCTweet.FIR* and *FGLOCTweet.FLD* source sets are more geographically diverse. For example, the 20 most frequent LMs in the *FGLOCTweet.FIR* source set constitute 20%–22% for each of the US, UK, and China. The *FGLOCTweet.FLD* source set is more geographically diverse, containing the top 3 LMs: Jakarta (7%), Indonesia (4%), and Venice (4%).
- *Superior performance of IDRISI.<domain>* source sets*: For those models, we do not apply size normalization. They show better performance compared to their antonymic source sets (*IDRISI.<domain>*). They generate the best LMR models on average (both "AVG" and "Cross-domain" columns) for EQK and HRC domains. They also generate comparable performing models on average for FIR and FLD domains, compared to the best source sets, *FGLOCTweet.FIR* and *FGLOCTweet.FLD* (exhibits slight lower performance by approximately 2.5%).

To answer **RQ2**, training on IDRISI-RE can produce domain-generalizable LMR models with F_1 of 80%, 77%, 75%, and 88%, for EQK, FIR, FLD, and HRC domains, respectively, on *cross-domain* average. Other datasets show inferior *cross-domain* average performance on EQK and HRC domains. However, *FGLOCTweet.FIR* exhibits better yet comparable performance to *IDRISI.FIR**. Similarly, *GEO+OLM.FLD** and *FGLOCTweet.FLD* show comparable performance to *IDRISI.FLD**.

Overall performance: We emphasize the superior performance of IDRISI-RE dataset in the domain generalizability by highlighting a few points:

- Although *GEO.<domain>* and *FGLOCTweet.<domain>* show competitive performance to *IDRISI.<domain>* per disaster domain, they exhibit lower overall performance than IDRISI-RE (*GEO.ALL* and *FGLOCTweet.ALL* versus *IDRISI.ALL*).
- We note that part of the geographical coverage of IDRISI-RE is held out for the *IDRISI.<domain>* target/test set, hence it does not appear in the source/training sets of *IDRISI.ALL*. Thus, merging the held-out data into training could improve the results further.
- As the size of IDRISI-RE is one of the advantages that distinguishes it from the existing datasets, training on *IDRISI.ALL** indeed generates LMR models that surpass the ones trained on *GEO.ALL*, *FGLOCTweet.ALL*, and *IDRISI.ALL*, on average.

8.2. Geographical generalizability

We use "geographical area" to refer to the country where the disaster of the target dataset happened, e.g., the United States. To this end, we define the *geographical* generalizability as *the ability of the model trained on a specific geographical area (source) to generalize and perform well when tested on an unseen disaster event in the same or different geographical area (target)*.

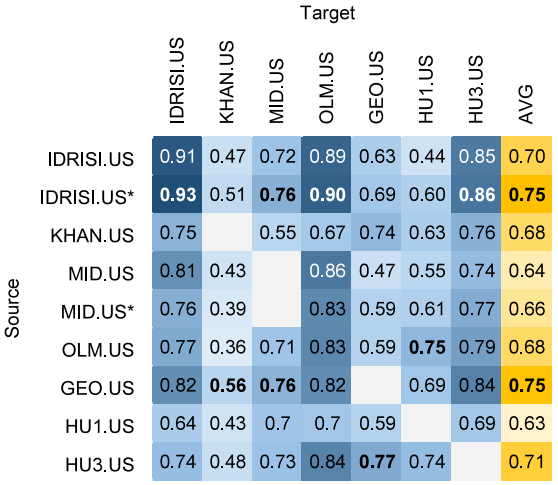

Fig. 5. The geographical inter-generalizability F_1 results for IDRISI-RE for the *geographical few-shot learning*. The blue color scale is global for the entire matrix. The best results per column are **boldfaced**. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

8.2.1. Experimental setups

To study whether IDRISI-RE can generalize to unseen events that happened in the same or different geographical areas, we train the model using the data of the common countries between IDRISI-RE and the existing datasets (OLM, MID, GEO, and KHAN), namely, India (IN), New Zealand (NZ), and the United States (US). We note that *all* of our reported experiments are under zero-shot learning (experiments where the training and test sets include the same event are hidden/greyed in the figures).

8.2.2. Results and discussion

In this section, we address two research questions: can an LMR model that is trained on IDRISI-RE generalize to:

- Unseen events that happen in the same geographical areas? (**RQ3**)
- Unseen events that happen in different geographical areas? (**RQ4**)

Geographical Generalizability within the same country: Fig. 5 presents the F_1 scores for the geographical generalizability experiments for the events that occurred in the United States. We limit our experiments to events that happened in the United States because it is the only country covered by all the public datasets. We found that training on *IDRISI.US* generates higher performing LMR models compared to *KHAN.US*, *MID.US*, *MID.US**, *OLM.US*, and *HU1.US*, on average. While *HU3.US* outperforms *IDRISI.US*, *IDRISI.US** outperforms it significantly by approximately 5.3%. This improvement confirms the important role of the size of source data that IDRISI-RE offers to the community. Additionally, *IDRISI.US** beats *GEO.US* over 4 out of 7 target sets, but *GEO.US* beats *IDRISI.US** over only 2 target sets. Nevertheless, both are comparable on average.

To answer **RQ3**, we conclude that the models trained on IDRISI-RE exhibit an acceptable F_1 average score of 0.75. They achieve the best performance on 4 out of 7 target sets, compared to the models trained on the other source sets.

Geographical Generalizability across countries: Fig. 6 shows the F_1 results of the models trained under the *geographical zero-shot learning*, where the source and target data are sampled from events that happened in *different* countries. Looking at the results, we find that training on IDRISI-RE is significantly better than training on MID, KHAN, and OLM datasets for all geographical areas (*IDRISI.*<*country*> vs. *MID.*<*country*>, *KHAN.*<*country*>, and *OLM.*<*country*>), on average. Additionally, IDRISI-RE outperforms GEO data over most of the test sets. The poor performance of *GEO.US* requires further investigation. *HU1.US* and *HU3.US* exhibit a way lower scores compared to *IDRISI.US* and *IDRISI.US***. However, *HU3* is more comparable to *IDRISI* for *IN* and *NZ* geographical areas. The high performance of *HU3.IN* on *HU3.NZ* and *HU3.US* leads to best average score, yet comparable to *IDRISI.IN*. Similarly, the high performance of *HU3.NZ* on *HU3.IN* and *HU3.US* leads to comparable average against *IDRISI.NZ*.

To answer **RQ4**, it is quite evident that training on IDRISI-RE generates the best performing LMR models that can reasonably generalize to events that happened in different geographical areas. The performance of models generated by other datasets is rather poor in most cases.

9. Implications

Compared to the public datasets, IDRISI-RE is the largest in size, domain diverse, geographically representative, temporally representative, and informative. It also contains annotations for different coarse- (e.g., country, city) and fine-grained (e.g., street, POI) location types. The extensive empirical generalizability analysis showed that IDRISI-RE is the best *domain* and *geographical*

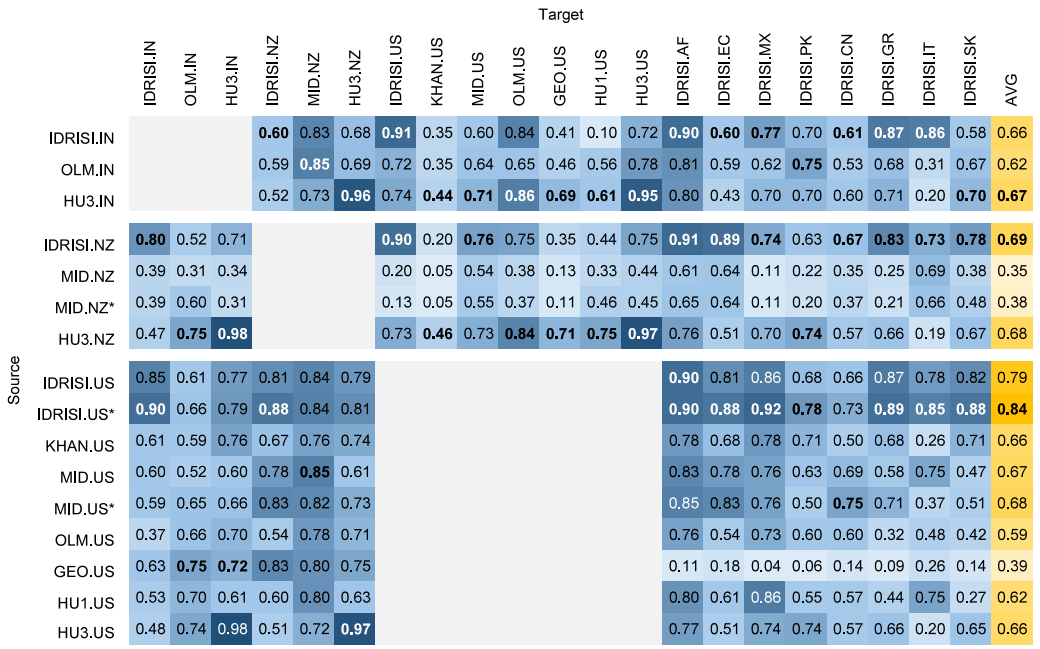

Fig. 6. The geographical inter-generalizability F_1 results for IDRISI-RE for the *geographical zero-shot learning*. IN, NZ, and the US refer to India, New Zealand, and the United States, respectively. The blue color scale is global for the entire matrix. The best results per geographical area per column are boldfaced.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

generalizable LMR dataset. All these advantages of IDRISI-RE cultivate the basis for empowering research on LMR in the domain of disaster response and management, specifically, and other domains. In this section, we describe the theoretical (Section 9.1), practical (Section 9.2), and research (Section 9.3) implications of releasing IDRISI-RE.

9.1. Theoretical implications

While responders need to obtain all useful information that supports managing emergencies effectively and efficiently, the geographical context enables better understanding of the development of disaster events and the behavior of the affected people at the events' onset. For example, the geographical information is very useful in creating diverse crisis maps (refer to Section 9.3). Making IDRISI-RE public enables developing and evaluating generalizable LMR models that better tackle domain shifts and are less susceptible to changes in geographical areas. Such models should be ready for deployment for any future disaster events. To ensure generalizability, IDRISI-RE is designed to meet six objectives that we elaborate on their value in the following:

1. ***Geographical coverage***: Deploying geographically generalizable LMR models at the onset of disaster events happening anywhere on earth requires data that covers broad geographical areas. While IDRISI-RE covers 22 English-speaking countries, it supports response authorities from anywhere in the world to incorporate the geographical context while drawing situational-awareness, assessing impact, managing resources, and deploying relief plans. Hence, the responders gain a better understanding of the disaster events and the behavior of the impacted people, at different location granularity.
2. ***Domain coverage***: Similarly, building a domain generalizable LMR model that is ready for deployment at the onset of the disaster events of any type (e.g., flood, earthquake) requires training it on data that is collected during diverse disaster events. The domain diversity of IDRISI-RE enables the geographical-aware management of disaster events of any type.
3. ***Location type annotations***: Effective geographical-aware management of disaster events is deemed attainable when the needs of different response authorities, in terms of location granularity, are met. While IDRISI-RE offers not only LM annotations but also location type annotations, it enables the development and evaluation of robust LMR models that aid drawing situational-awareness, assessing the disaster impact, managing resources, and deploying relief plans, *at different location granularity*.
4. ***Large-scale***: The trainable LMR models, especially the deep learning-based models, require large training datasets to perform accurately. Thus, IDRISI-RE, being the largest and most generalizable LMR dataset, it supports the responders to better understand the disaster events and the behavior of the impacted people.

5. ***Temporal coverage***: As IDRISI-RE covers the critical periods of the disaster events, it helps different response authorities to better understand the disaster events and the behavior of affected people during different disaster phases (pre-disaster, during disaster, and post-disaster).
6. ***Relevance and informativeness***: Providing the geographical context to only informative content, after discarding noise, is of a high priority to aid the response authorities in understanding the updates of the disaster events on the ground. As IDRISI-RE solely contains informative tweets, it provides more realistic data for training the LMR models that are ready for direct integrating in real-world information processing systems for disaster management.

Moreover, while all these design factors are important, the conclusions we drew when answering the RQ1-4 emphasize the influence of the geographical coverage and data size for creating generalizable LMR datasets. To elaborate, the geographical vicinity of the source and target sets is a potential factor on improving the LMR performance even when disaster domains are segregated (refer to Sections 8.1.2 and 8.2 for details). Additionally, the large size is a key advantage of the training datasets which allows generating more robust LMR models (refer to *IDRISI.ALL* and *IDRISI.ALL** results in Section 8.1.2 for details).

9.2. Practical implications

Using IDRISI-RE enables the deployment of different surveillance and decision-support systems during disaster events that are used by different response authorities. These systems employ the underlying applications discussed in Section 9.1 and generate reports at different location granularity for different phases of the disaster. These reports could be in a form of real-time crisis maps that we briefly elaborate on a few types of them, below.

Situational awareness maps: These maps support the response authorities in understanding the development of the disaster, identifying the critical incidents, and detecting the hotspots of damages and vulnerable people.

Impact assessment maps: Mapping and identifying the most impactful incidents such as infrastructure damage, power outage, facilities closure, among other, helps response authorities manage relief activities and plan for recovery.

Eyewitnesses maps: Locating eyewitnesses and first responders is needed to connect people in need with the first responders (e.g., first aid treatment performers). Furthermore, getting authentic situational information is a critical task that can be achieved by communicating with eyewitnesses who are nearby the locations of incidents.

Resources maps: Resources include facilities (e.g., shelters), funding (e.g., donations), and supplies (e.g., food and water), to list a few. Locating such resources is important to identify places of shortage, adequacy, or abundance of resources and redistribute them based the need.

Population mobility maps: Evacuating the vulnerable people away from the affected areas requires monitoring their movement to consequently study the resource allocation and recovery plans. When exploiting Twitter for disaster relief activities, the essential step to constructing all these maps is to extract toponyms from the text. IDRISI-RE can be utilized to build automatic domain and geographical generalizable LMR models that perform at acceptable accuracy levels.

9.3. Research implications

IDRISI-RE enables research in different computational tasks, such as event/incident detection, relevance filtering, and geolocation tasks, to name a few. In addition to that, as IDRISI-RE dataset covers different types of disaster events, we anticipate it to essentially support transfer learning and domain adaption research. Below we briefly elaborate on a few tasks.

Event/incident detection: Detecting disaster events/incidents facilitates timely prevention and mitigation activities (Pettet et al., 2022). Fortunately, people tend to mention where events/incidents take place when they report them (Hu & Wang, 2020). Harnessing the relation between the occurrence (e.g., peaks) of LMs in tweets and the likelihood of events and incidents happening can aid early prediction and detection. For example, Sankaranarayanan, Samet, Teitler, Lieberman, and Sperling (2009) and Watanabe, Ochi, Okabe, and Onai (2011) had proposed content analysis of tweets by extracting locations for event/incident detection.

Relevance filtering: A key barrier to exploiting social media for crisis management is the noisiness of data which necessities the need for automatic relevance filtering methods (Lorini et al., 2021). Prior studies show that the geographical references in social media messages could indicate their relevance and informativeness (De Albuquerque, Herfort, Brenning, & Zipf, 2015; Vieweg, Hughes, Starbird, & Palen, 2010). Kaufhold, Bayer, and Reuter (2020) achieved the best performance when they incorporated location-related features in their rapid classification model. Thus, we anticipate IDRISI-RE to be useful for relevance filtering models.

Geolocation applications: Several geolocation applications are required, e.g., (1) detecting and disambiguating LMs in tweets, (2) predicting tweet location (Ozdikis, Ramampiaro, & Nørvåg, 2019), (3) inferring user location (Luo, Qiao, Li, Ma, & Liu, 2020), and (4) modeling user movement (Wu et al., 2022). While all these tasks are crucial for crisis management, the LMR task, in particular, plays an essential role in tackling all of them using text-based techniques (Zheng, Han, & Sun, 2018). For instance, combining

extracted entities (e.g., LMs) from tweets and their relations inferred from a Knowledge-base leads to a noticeable improvement on the *user location prediction* model (Miyazaki, Rahimi, Cohn, & Baldwin, 2018).

Displacement monitoring: A terrible consequence of crises is the internal and cross-border displacement. By early May 2019, the number of displaced people reached about 41.3 million due to conflicts and violence.[16] Extracting the location mentions from tweets shared by refugees would give some clues about the routes they are using or planning to use. Therefore, IDRISI-RE supports modeling the patterns of people displacement.

Geographical retrieval: The geographical information retrieval (GIR) systems are concerned with extracting spatial information alongside the relevant multimodal data to the user information need (Purves et al., 2018). IDRISI-RE serves the GIR retrieval techniques that rely on detecting locations and spatial references in queries and documents (García-Cumbreras, Perea-Ortega, García-Vega, & Ureña-López, 2009). The large size of IDRISI-RE dataset provides a promising resource for augmenting spatial information of tweets for geographical indexing and retrieval over the Twitter streams. Additionally, as IDRISI-RE is characterized by its wide geographical coverage, we anticipate it to be a representative resource for Geographical retrieval.

10. Conclusion

We introduced IDRISI-RE, a large-scale *Location Mention Recognition* Twitter dataset comprising around 20k human-labeled and 57k machine-labeled tweets from 19 disaster events, including floods, earthquakes, hurricanes, wildfires. The annotations include spans of location mentions in tweets' content and their geographical types such as country, state, city, street. The dataset events cover countries across continents, including the United States, Canada, Italy, India, Pakistan, Mozambique, and Malawi, among others. Additionally, we benchmark IDRISI-RE using both traditional and deep learning models, offering competitive baselines for future LMR development. We further studied the *domain* and *geographical* generalizability of IDRISI-RE against LMR English datasets under fair comparison setups and reached nuanced conclusions that IDRISI-RE is the most generalizable LMR dataset. The reliability, consistency, coverage, diversity, and generalizability analyses show the robustness of IDRISI-RE that empowers research on LMR. For future work, we plan to extend the annotations for the LMD task. We further plan to explore different transfer learning, domain adaptation, and active learning techniques to tackle the LMR task.

CRediT authorship contribution statement

Reem Suwaileh: Conceptualization, Methodology, Software, Validation, Data curation, Formal analysis, Investigation, Writing – original draft, Visualization, Funding acquisition, Project administration. **Tamer Elsayed:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. **Muhammad Imran:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Data availability

I have shared the link to my data in the manuscript.

Acknowledgments

This work was made possible by the Graduate Sponsorship Research Award (GSRA) #GSRA5-1-0527-18082 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Appendix A. Location mention distribution

Figs. 7–10 show the distribution of top 15 frequent location mentions in IDRISI-RE dataset per disaster event.

[16] https://www.internal-displacement.org/global-report/grid2019/

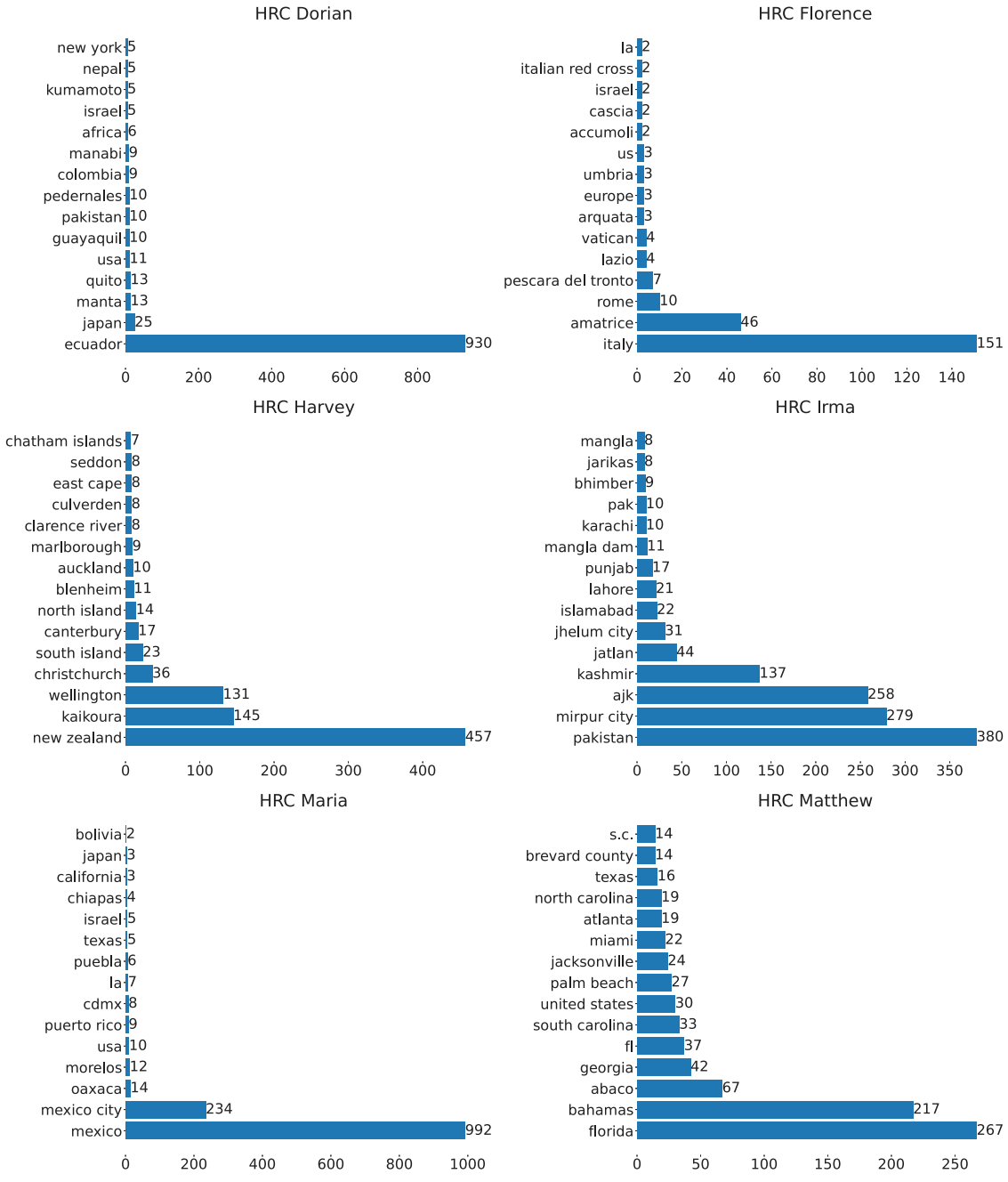

Fig. 7. The distribution of top 15 location mentions in IDRISI-RE per hurricane event.

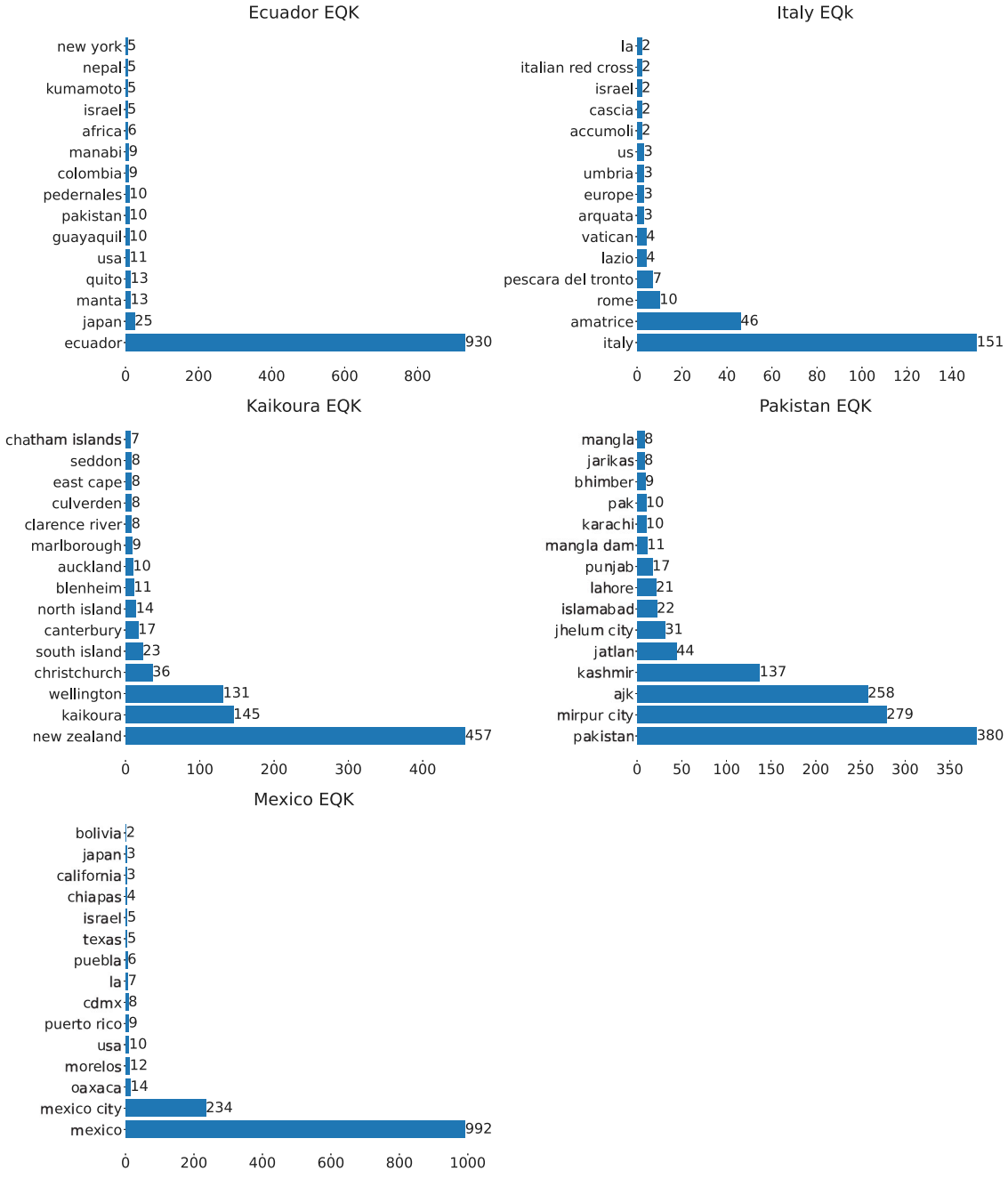

Fig. 8. The distribution of top 15 location mentions in IDRISI-RE per earthquake event.

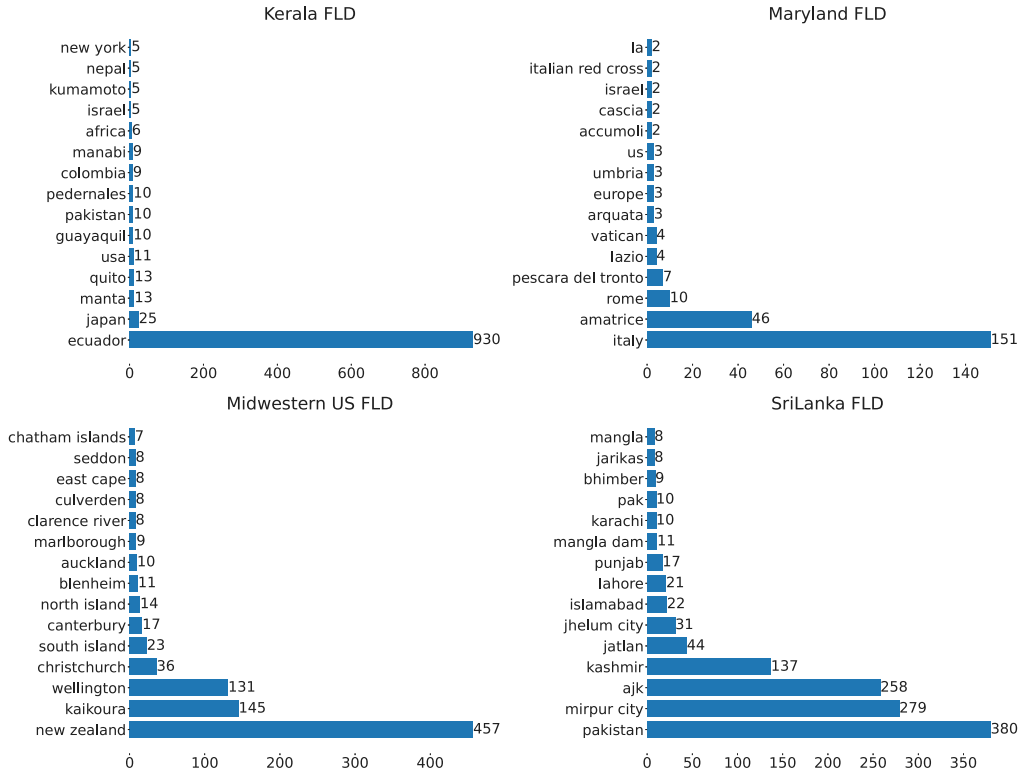

Fig. 9. The distribution of top 15 location mentions in IDRISI-RE per flood event.

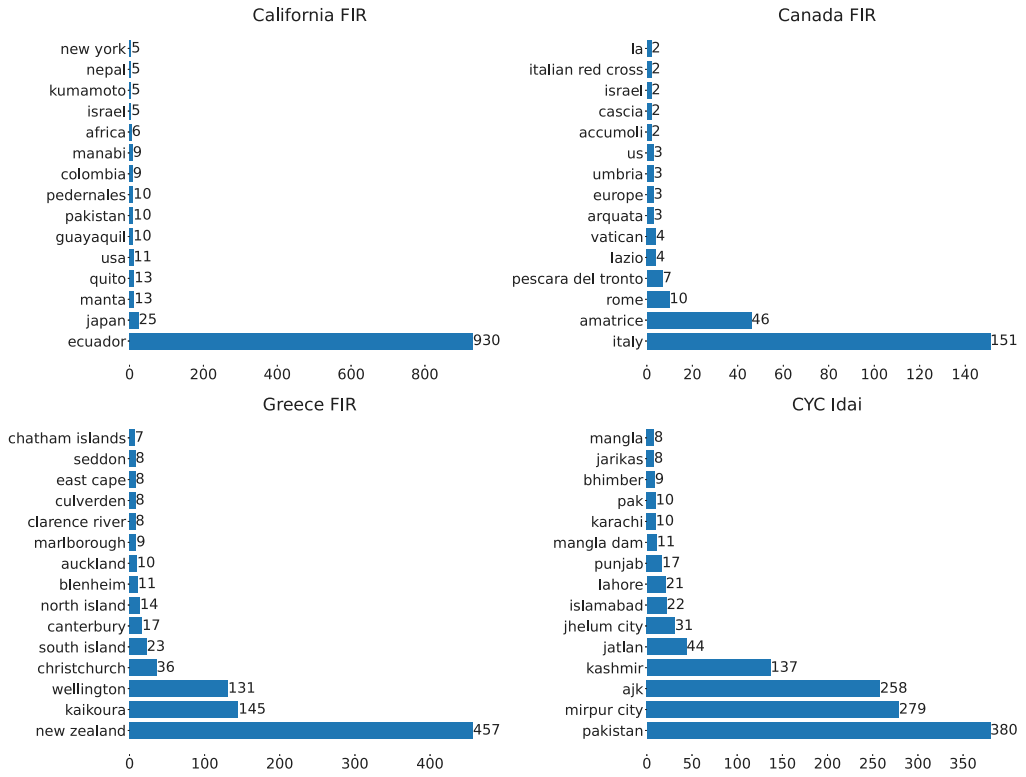

Fig. 10. The distribution of top 15 location mentions in IDRISI-RE per wildfire/cyclone event.

Appendix B. Detailed fine-tuning results and best hyper-parameters

Table 8 shows the best hyper-parameters and detailed results of the BERT-based LMR models for both type-less and type-based recognition. Tables 9 and 10 show the best hyper-parameters and detailed results of the CRF LMR models for both type-less and type-based recognition, respectively.

Table 8
The best hyper-parameters and results for the BERT-based model over IDRISI-RE under *Type-less* LMR. e, bs, lr, and sl refer to the hyper-parameters, number of epochs, batch size, learning rate, and sequence length, respectively.

Event	Random						Time-based					
	e	bs	lr	P	R	F1	e	bs	lr	P	R	F1
Type-less												
Ecuador Earthquake	4	32	4e−5	0.960	0.958	0.953	4	32	4e−5	0.923	0.921	0.916
Canada Wildfires	4	8	4e−5	0.733	0.749	0.732	4	8	4e−5	0.768	0.779	0.767
Italy Earthquake	3	8	3e−5	0.881	0.886	0.880	3	8	3e−5	0.840	0.849	0.842
Kaikoura Earthquake	3	8	3e−5	0.914	0.919	0.912	3	8	3e−5	0.912	0.893	0.896
Hurricane Matthew	4	8	5e−5	0.948	0.945	0.941	4	8	5e−5	0.949	0.956	0.944
Sri Lanka Floods	3	16	4e−5	0.921	0.929	0.917	3	16	4e−5	0.904	0.918	0.904
Hurricane Harvey	4	8	5e−5	0.919	0.902	0.906	4	8	5e−5	0.900	0.893	0.894
Hurricane Irma	4	8	3e−5	0.843	0.839	0.835	4	8	3e−5	0.829	0.833	0.825
Hurricane Maria	2	8	4e−5	0.932	0.926	0.925	2	8	4e−5	0.913	0.909	0.904
Mexico Earthquake	4	8	3e−5	0.932	0.932	0.929	4	8	3e−5	0.919	0.913	0.911
Maryland Floods	3	16	5e−5	0.895	0.901	0.890	3	16	5e−5	0.900	0.838	0.845
Greece Wildfires	3	8	5e−5	0.935	0.934	0.927	3	8	5e−5	0.897	0.895	0.883
Kerala Floods	4	32	5e−5	0.897	0.893	0.887	4	32	5e−5	0.927	0.934	0.923
Hurricane Florence	4	8	4e−5	0.773	0.755	0.755	4	8	4e−5	0.801	0.785	0.784
California Wildfires	3	16	3e−5	0.923	0.930	0.920	3	16	3e−5	0.914	0.906	0.906
Cyclone Idai	3	8	4e−5	0.932	0.927	0.925	3	8	4e−5	0.911	0.900	0.898
Midwestern U.S. Floods	4	8	5e−5	0.948	0.957	0.944	4	8	5e−5	0.946	0.961	0.949
Hurricane Dorian	4	8	5e−5	0.874	0.893	0.878	4	8	5e−5	0.865	0.872	0.862
Pakistan Earthquake	3	32	4e−5	0.876	0.902	0.877	3	32	4e−5	0.830	0.878	0.836
Type-based												
Ecuador Earthquake	2	8	3e−5	0.951	0.940	0.939	4	32	4e−5	0.941	0.922	0.926
Canada Wildfires	3	8	4e−5	0.733	0.749	0.733	4	8	4e−5	0.772	0.780	0.771
Italy Earthquake	3	8	4e−5	0.894	0.894	0.890	3	8	3e−5	0.879	0.888	0.881
Kaikoura Earthquake	4	16	5e−5	0.914	0.916	0.909	3	8	3e−5	0.918	0.895	0.899
Hurricane Matthew	4	32	5e−5	0.931	0.923	0.919	4	8	5e−5	0.955	0.963	0.952
Sri Lanka Floods	4	8	5e−5	0.929	0.933	0.925	3	16	4e−5	0.911	0.925	0.912
Hurricane Harvey	4	16	4e−5	0.921	0.905	0.909	4	8	5e−5	0.898	0.896	0.895
Hurricane Irma	2	8	5e−5	0.847	0.831	0.833	4	8	3e−5	0.827	0.828	0.823
Hurricane Maria	2	8	5e−5	0.936	0.924	0.924	2	8	4e−5	0.910	0.895	0.897
Mexico Earthquake	2	16	4e−5	0.921	0.914	0.913	4	8	3e−5	0.918	0.914	0.911
Maryland Floods	3	8	4e−5	0.906	0.894	0.892	3	16	5e−5	0.851	0.795	0.805
Greece Wildfires	3	16	3e−5	0.927	0.940	0.925	3	8	5e−5	0.899	0.899	0.887
Kerala Floods	4	8	5e−5	0.891	0.885	0.880	4	32	5e−5	0.926	0.927	0.919
Hurricane Florence	3	16	4e−5	0.795	0.774	0.772	4	8	4e−5	0.792	0.781	0.778
California Wildfires	4	32	4e−5	0.913	0.919	0.909	3	16	3e−5	0.918	0.900	0.902
Cyclone Idai	3	32	4e−5	0.906	0.906	0.900	3	8	4e−5	0.905	0.900	0.895
Midwestern U.S. Floods	4	8	5e−5	0.944	0.948	0.936	4	8	5e−5	0.944	0.957	0.944
Hurricane Dorian	4	16	5e−5	0.857	0.871	0.858	4	8	5e−5	0.864	0.860	0.852
Pakistan Earthquake	4	8	5e−5	0.899	0.908	0.894	3	32	4e−5	0.819	0.868	0.828

Table 9
The best hyper-parameters and results for CRF model over IDRISI-RE for *Type-less* LMR. The column "Algo." refers to the training algorithm of CRF. The "HP1" and "HP2" refer to the tuned hyper-parameters with respect to the algorithm.

Event	Algo.	HP1	HP2	P	R	F1
Random data setup						
Ecuador Earthquake	lbfgs	c1=0.95	c2=0.95	0.890	0.842	0.866
Canada Wildfires	ap	epsilon=0.001		0.635	0.864	0.732
Italy Earthquake	lbfgs	c1=0.1	c2=0.1	0.547	0.569	0.558
Kaikoura Earthquake	lbfgs	c1=0.15	c2=0.15	0.872	0.884	0.878
Hurricane Matthew	lbfgs	c1=0.95	c2=0.95	0.925	0.857	0.890
Sri Lanka Floods	ap	epsilon=0.01		0.835	0.878	0.856
Hurricane Harvey	lbfgs	c1=0.15	c2=0.15	0.816	0.804	0.810
Hurricane Irma	lbfgs	c1=0.25	c2=0.25	0.843	0.714	0.773
Hurricane Maria	lbfgs	c1=0.15	c2=0.15	0.858	0.869	0.864
Mexico Earthquake	arow	variance=0.1	gamma=0.5	0.838	0.884	0.860
Maryland Floods	arow	variance=0.1	gamma=0.25	0.750	0.878	0.809
Greece Wildfires	lbfgs	c1=0.25	c2=0.25	0.757	0.941	0.839
Kerala Floods	ap	epsilon=1e-5		0.705	0.745	0.725
Hurricane Florence	ap	epsilon=0.001		0.660	0.673	0.667
California Wildfires	lbfgs	c1=0.5	c2=0.5	0.885	0.855	0.870
Cyclone Idai	lbfgs	c1=0.65	c2=0.65	0.899	0.885	0.892
Midwestern U.S. Floods	lbfgs	c1=0.6	c2=0.6	0.914	0.894	0.904
Hurricane Dorian	lbfgs	c1=0.55	c2=0.55	0.856	0.787	0.820
Pakistan Earthquake	lbfgs	c1=0.35	c2=0.35	0.872	0.885	0.879
Time-based data setup						
Ecuador Earthquake	arow	variance=0.25	gamma=0.25	0.933	0.933	0.932
Canada Wildfires	ap	epsilon=0.01		0.853	0.853	0.853
Italy Earthquake	arow	variance=1	gamma=0.125	0.906	0.906	0.906
Kaikoura Earthquake	arow	variance=0.5	gamma=0.25	0.880	0.880	0.879
Hurricane Matthew	arow	variance=1	gamma=0.1	0.902	0.905	0.901
Sri Lanka Floods	arow	variance=1	gamma=0.1	0.911	0.911	0.910
Hurricane Harvey	arow	variance=0.1	gamma=0.125	0.906	0.906	0.906
Hurricane Irma	lbfgs	c1=0.95	c2=0.95	0.906	0.906	0.906
Hurricane Maria	arow	variance=0.16	gamma=0.5	0.883	0.883	0.882
Mexico Earthquake	arow	variance=0.25	gamma=0.16	0.839	0.839	0.838
Maryland Floods	lbfgs	c1=0.85	c2=0.85	0.754	0.759	0.751
Greece Wildfires	arow	variance=0.5	gamma=0.1	0.895	0.901	0.896
Kerala Floods	arow	variance=0.125	gamma=0.5	0.880	0.881	0.880
Hurricane Florence	arow	variance=1	gamma=0.16	0.879	0.879	0.879
California Wildfires	arow	variance=1	gamma=0.125	0.908	0.908	0.907
Cyclone Idai	arow	variance=0.125	gamma=0.5	0.877	0.879	0.877
Midwestern U.S. Floods	lbfgs	c1=0.9	c2=0.9	0.920	0.923	0.917
Hurricane Dorian	arow	variance=0.16	gamma=0.5	0.875	0.875	0.875
Pakistan Earthquake	arow	variance=1	gamma=0.125	0.821	0.822	0.820

Table 10
The best hyper-parameters and results for CRF model over IDRISI-RE for *Type-based* LMR. The column "Algo." refers to the training algorithm of CRF. The "HP1" and "HP2" refer to the tuned hyper-parameters with respect to the algorithm.

Event	Algo.	HP1	HP2	P	R	F1
Random data setup						
Ecuador Earthquake	lbfgs	c1=0.8	c2=0.8	0.735	0.698	0.716
Canada Wildfires	lbfgs	c1=0.7	c2=0.7	0.597	0.699	0.644
Italy Earthquake	ap	epsilon=1e-5		0.534	0.477	0.504
Kaikoura Earthquake	lbfgs	c1=0.95	c2=0.95	0.856	0.675	0.755
Hurricane Matthew	lbfgs	c1=0.2	c2=0.2	0.774	0.808	0.790
Sri Lanka Floods	lbfgs	c1=0.4	c2=0.4	0.681	0.811	0.740
Hurricane Harvey	lbfgs	c1=0.9	c2=0.9	0.677	0.537	0.599
Hurricane Irma	lbfgs	c1=0.4	c2=0.4	0.586	0.497	0.538
Hurricane Maria	lbfgs	c1=0.25	c2=0.25	0.782	0.754	0.768
Mexico Earthquake	arow	variance=0.1	gamma=0.5	0.828	0.770	0.798
Maryland Floods	lbfgs	c1=0.55	c2=0.55	0.796	0.547	0.648
Greece Wildfires	lbfgs	c1=0.7	c2=0.7	0.770	0.786	0.778
Kerala Floods	lbfgs	c1=0.55	c2=0.55	0.633	0.642	0.638
Hurricane Florence	arow	variance=0.16	gamma=0.5	0.373	0.617	0.465
California Wildfires	lbfgs	c1=0.7	c2=0.7	0.861	0.804	0.832
Cyclone Idai	lbfgs	c1=0.95	c2=0.95	0.784	0.626	0.696
Midwestern U.S. Floods	lbfgs	c1=0.9	c2=0.9	0.794	0.791	0.792
Hurricane Dorian	lbfgs	c1=0.85	c2=0.85	0.621	0.378	0.470
Pakistan Earthquake	lbfgs	c1=0.55	c2=0.55	0.706	0.742	0.723
Time-based data setup						
Ecuador Earthquake	arow	variance=0.1	gamma=0.16	0.910	0.912	0.910
Canada Wildfires	lbfgs	c1=0.05	c2=0.05	0.865	0.865	0.865
Italy Earthquake	arow	variance=0.5	gamma=0.16	0.881	0.881	0.881
Kaikoura Earthquake	arow	variance=0.16	gamma=0.125	0.875	0.874	0.875
Hurricane Matthew	lbfgs	c1=0.15	c2=0.15	0.901	0.903	0.899
Sri Lanka Floods	arow	variance=1	gamma=0.25	0.900	0.896	0.897
Hurricane Harvey	ap	epsilon=0.01		0.914	0.914	0.914
Hurricane Irma	lbfgs	c1=0.55	c2=0.55	0.893	0.893	0.893
Hurricane Maria	arow	variance=1	gamma=0.1	0.890	0.890	0.890
Mexico Earthquake	arow	variance=0.1	gamma=1	0.881	0.882	0.880
Maryland Floods	arow	variance=1	gamma=0.16	0.875	0.878	0.873
Greece Wildfires	arow	variance=0.25	gamma=0.5	0.886	0.890	0.886
Kerala Floods	arow	variance=1	gamma=0.1	0.857	0.859	0.857
Hurricane Florence	arow	variance=0.1	gamma=0.5	0.889	0.889	0.889
California Wildfires	arow	variance=1	gamma=0.125	0.903	0.903	0.902
Cyclone Idai	arow	variance=1	gamma=0.1	0.852	0.854	0.852
Midwestern U.S. Floods	lbfgs	c1=0.35	c2=0.35	0.924	0.925	0.920
Hurricane Dorian	arow	variance=0.5	gamma=0.1	0.865	0.866	0.865
Pakistan Earthquake	arow	variance=0.16	gamma=0.16	0.781	0.782	0.780

Appendix C. Detailed data setups for generalizability experiments

Tables 11 and 12 show the detailed data setups for the *domain* and *geographical* generalizability experiments, respectively.

Table 11
The data setups/splits of the domain generalizability experiments. EQK, FLD, CYC, HRC, and FIR refer to Earthquake, Flood, Cyclone, Hurricane, and Fire, respectively.

Tweet set	Train	Test	Train	Test	Train	Test	Train	Test
	IDRISI.EQK		MID.EQK		GEO.EQK		GEO+MID.EQK	
Ecuador EQK 2016	✓							
Italy EQK 2016	✓							
Kaikoura EQK 2016	✓							
Pakistan EQK 2019	✓							
Puebla Mexico EQK 2017		✓						
ChristChurch EQK 2011			✓	✓			✓	✓
Geocorpora EQK					✓	✓	✓	✓
	IDRISI.FLD		OLM.FLD		GEO.FLD		GEO+OLM.FLD	
Sri Lanka FLD 2017	✓							
Maryland FLD 2017	✓							
Kerala FLD 2018	✓							
CYC Idai 2019	✓							
Midwest. US FLD 2019		✓						
Chennai FLD 2015			✓				✓	
Houston FLD 2016			✓				✓	
Louisiana FLD 2016				✓				✓
Geocorpora FLD					✓	✓	✓	✓
	IDRISI.HRC		MID.HRC					
HRC Matthew 2016	✓							
HRC Harvey 2017	✓							
HRC Irma 2017	✓							
HRC Maria 2017	✓							
HRC Florence 2018	✓							
HRC Dorian 2019		✓						
HRC Sandy 2012			✓	✓				
	IDRISI.FIRE		GEO.FIRE					
Canada FIRE 2016	✓							
California FIRE 2018	✓							
Greece FIRE 2018		✓						
Geocorpora FIRE			✓	✓				

Table 12
The data setups for the geographical generalizability experiments. US, IN, NZ, IT, CA EC, MX, CR, and PK are the 2-char ISO country codes for the United States, India, New Zealand, Italy, Canada, Ecuador, Mexico, Greece, and Pakistan, respectively. AF refers to Africa continent and the countries covered are Mozambique, Zimbabwe, Malawi, and Madagascar.

Tweets	Train	Test	Train	Test	Train	Test
	IDRISI.US		OLM.US		MID.US	
HRC Matthew 2016	✓					
HRC Harvey 2017	✓					
HRC Irma 2017	✓					
HRC Maria 2017	✓					
HRC Florence 2018	✓					
HRC Dorian 2019	✓					
Maryland FLD 2018	✓					
California FIRE 2018	✓					
Midwest. US FLD 2019		✓				
Houston FLD 2016			✓			
Louisiana FLD 2016				✓		
HRC Sandy 2012					✓	✓
	IDRISI.IN		OLM.IN			
Kerala FLD 2018	✓	✓				
Chennai FLD 2015			✓	✓		
	IDRISI.NZ		MID.NZ			
Kaikoura EQK 2016	✓	✓				
ChristChurch EQK 2011			✓	✓		
	IDRISI.IT					
Italy EQK 2016		✓				
	IDRISI.CA					
Canada FIRE 2016		✓				
	IDRISI.EC					
Ecuador EQK 2016		✓				
	IDRISI.SK					
Srilanka FLD 2017		✓				
	IDRISI.MX					
Puebla Mexico EQK 2017		✓				
	IDRISI.CR					
Greece FIRE 2018		✓				
	IDRISI.PK					
Pakistan EQK 2019		✓				
	IDRISI.AF					
CYC Idai 2019		✓				

References

Al Emadi, N., Abbar, S., Borge-Holthoefer, J., Guzman, F., & Sebastiani, F. (2017). QT2S: A system for monitoring road traffic via fine grounding of tweets. In *Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1* (pp. 456–459).

Al-Olimat, H., Thirunarayan, K., Shalin, V., & Sheth, A. (2018). Location name extraction from targeted text streams using gazetteer-based statistical language models. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1986–1997). Association for Computational Linguistics.

Alam, F., Joty, S., & Imran, M. (2018). Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 1077–1087). Melbourne, Australia: Association for Computational Linguistics.

Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). HumAID: Human-annotated disaster incidents data from Twitter. In *15th international conference on web and social media* (pp. 933–942).

Alkouz, B., & Al Aghbari, Z. (2020). SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks. *Information Processing & Management*, *57*(1), Article 102139.

Das, R. D., & Purves, R. S. (2020). Exploring the potential of Twitter to understand traffic events and their locations in Greater Mumbai, India. *IEEE Transactions on Intelligent Transportation Systems*, *21*(12), 5213–5222.

De Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, *29*(4), 667–689.

Derczynski, L., Bontcheva, K., & Roberts, I. (2016). Broad Twitter Corpus: A diverse named entity recognition resource. In *Proceedings of the 26th international conference on computational linguistics: Technical papers* (pp. 1169–1179).

Derczynski, L., Nichols, E., van Erp, M., & Limsopatham, N. (2017). Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd workshop on noisy user-generated text* (pp. 140–147). Copenhagen, Denmark: Association for Computational Linguistics.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4171–4186).

Dutt, R., Hiware, K., Ghosh, A., & Bhaskaran, R. (2018). SAVITR: A system for real-time location extraction from microblogs during emergencies. In *Companion proceedings of the the web conference 2018* (pp. 1643–1649).

Fernández, N. J., & Periñán-Pascual, C. (2021). nLORE: A linguistically rich deep-learning system for locative-reference extraction in tweets. In *Intelligent environments 2021: Workshop proceedings of the 17th international conference on intelligent environments, vol. 29* (p. 243). IOS Press.

Fernández-Martínez, N. J. (2022). The FGLOCTweet Corpus: An English tweet-based corpus for fine-grained location-detection tasks. *Research in Corpus Linguistics*, *10*(1), 117–133.

García-Cumbreras, M. Á., Perea-Ortega, J. M., García-Vega, M., & Ureña-López, L. A. (2009). Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Information Processing & Management*, *45*(5), 605–614.

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *Geoinformatica*, *17*(4), 635–667.

Gelernter, J., & Zhang, W. (2013). Cross-lingual geo-parsing for non-structured data. In *Proceedings of the 7th workshop on geographic information retrieval* (pp. 64–71). ACM.

Grace, R., Kropczynski, J., & Tapia, A. (2018). Community coordination: Aligning social media use in community emergency management. In *Proceedings of the 15th ISCRAM conference* (pp. 609–620).

Hiltz, S. R., Hughes, A. L., Imran, M., Plotnick, L., Power, R., & Turoff, M. (2020). Exploring the usefulness and feasibility of software requirements for social media use in emergency management. *International Journal of Disaster Risk Reduction*, *42*, Article 101367.

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

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (2012). Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st international conference on world wide web* (pp. 769–778).

Hu, X., Al-Olimat, H., Kersten, J., Wiegmann, M., Klan, F., Sun, Y., et al. (2021). GazPNE: Annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. *International Journal of Geographical Information Science*, 310–337.

Hu, Y., & Wang, J. (2020). How do people describe locations during a natural disaster: An analysis of tweets from hurricane harvey. *Leibniz International Proceedings in Informatics, LIPIcs*, *177*.

Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., et al. (2022). Location reference recognition from texts: A survey and comparison. arXiv preprint arXiv:2207.01683.

Hu, X., Zhou, Z., Sun, Y., Kersten, J., Klan, F., Fan, H., et al. (2022). GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models. *IEEE Internet of Things Journal*, 16259–16271.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on world wide web companion* (pp. 1021–1024). International World Wide Web Conferences Steering Committee.

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 1638–1643). Paris, France: European Language Resources Association (ELRA).

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2015). Detecting and disambiguating locations mentioned in Twitter messages. In *Computational linguistics and intelligent text processing* (pp. 321–332). Cham: Springer International Publishing.

Ji, Z., Sun, A., Cong, G., & Han, J. (2016). Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th international conference on world wide web* (pp. 1271–1281). International World Wide Web Conferences Steering Committee.

Kaufhold, M. A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, *57*(1), Article 102132.

Khanal, S., Traskowsky, M., & Caragea, D. (2022). Identification of fine-grained location mentions in crisis tweets. In *Proceedings of the language resources and evaluation conference* (pp. 7164–7173). Marseille, France: European Language Resources Association.

Kitamoto, A., & Sagara, T. (2012). Toponym-based geotagging for observing precipitation from social and scientific data streams. In *Proceedings of the ACM multimedia 2012 workshop on geotagging and its applications in multimedia* (pp. 23–26). New York, NY, USA: Association for Computing Machinery.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70.

Kropczynski, J., Grace, R., Coche, J., Halse, S., Obeysekare, E., Montarnal, A., et al. (2018). Identifying actionable information on social media for emergency dispatch. In *ISCRAM Asia Pacific 2018: Innovating for resilience – 1st international conference on information systems for crisis response and management Asia Pacific* (pp. 428–438).

Kumar, A., & Singh, J. P. (2019). Location reference identification from tweets during emergencies: A deep learning approach. *International Journal of Disaster Risk Reduction*, *33*, 365–375.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th international conference on machine learning* (pp. 282–289).

Li, C., & Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 43–52). ACM.

Li, C., & Sun, A. (2017). Extracting fine-grained location with temporal awareness in tweets: A two-stage approach. *Journal of the Association for Information Science and Technology*, *68*(7), 1652–1670.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., et al. (2012). Twiner: named entity recognition in targeted Twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 721–730). ACM.

Lingad, J., Karimi, S., & Yin, J. (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web* (pp. 1017–1020).

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies-Volume 1* (pp. 359–367). Association for Computational Linguistics.

Lorini, V., Castillo, C., Peterson, S., Rufolo, P., Purohit, H., Pajarito, D., et al. (2021). Social media for emergency management: Opportunities and challenges at the intersection of research and practice. In *18th international conference on information systems for crisis response and management* (pp. 772–777).

Luo, X., Qiao, Y., Li, C., Ma, J., & Liu, Y. (2020). An overview of microblog user geolocation methods. *Information Processing & Management*, *57*(6), Article 102375.

Martínez, N. J. F., & Periñán-Pascual, C. (2020). Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets. *RAEL: revista electrónica de lingüística aplicada*, *19*(1), 136–163.

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, *36*(4), 1–27.

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17.

Miyazaki, T., Rahimi, A., Cohn, T., & Baldwin, T. (2018). Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP workshop W-NUT: The 4th workshop on noisy user-generated text* (pp. 7–16). Brussels, Belgium: Association for Computational Linguistics.

Molla, D., & Karimi, S. (2014). Overview of the 2014 ALTA shared task: identifying expressions of locations in tweets. In *Proceedings of the Australasian Language Technology Association workshop 2014* (pp. 151–156).

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys*, *54*(1), 1–39.

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 569–576). ACM.

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 376–385.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). 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* (pp. 994–1009). ACM.

Ozdikis, O., Ramampiaro, H., & Nørvåg, K. (2019). Locality-adapted kernel densities of term co-occurrences for location prediction of tweets. *Information Processing & Management*, *56*(4), 1280–1299.

Paule, J. D. G., Sun, Y., & Moshfeghi, Y. (2019). On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing & Management*, *56*(3), 1119–1132.

Pettet, G., Baxter, H., Vazirizade, S. M., Purohit, H., Ma, M., Mukhopadhyay, A., et al. (2022). Designing decision support systems for emergency response: Challenges and opportunities. In *Proceedings of the first workshop on Cyber Physical Systems for Emergency Response (CPS-ER) colocated with CPS-IOT week 2022* (pp. 30–35).

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). *Geographic information retrieval: Progress and challenges in spatial search of text*. Now Foundations and Trends.

Reuter, C., Ludwig, T., Kaufhold, M. A., & Spielhofer, T. (2016). Emergency services' attitudes towards social media: A quantitative and qualitative survey across Europe. *International Journal of Human-Computer Studies*, *95*, 96–111.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1524–1534). Association for Computational Linguistics.

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1524–1534).

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th Acm sigspatial international conference on advances in geographic information systems* (pp. 42–51).

Shang, L., Zhang, Y., Youn, C., & Wang, D. (2022). SAT-Geo: A social sensing based content-only approach to geolocating abnormal traffic events using syntax-based probabilistic learning. *Information Processing & Management*, *59*(2), Article 102807.

Sultanik, E. A., & Fink, C. (2012). Rapid geotagging and disambiguation of social media text via an indexed gazetteer. In *Proceedings of the 9th ISCRAM conference, vol. 12* (pp. 1–10).

Suwaileh, R., Elsayed, T., Imran, M., & Sajjad, H. (2022). When a disaster happens, we are ready: Location Mention Recognition from crisis tweets. *International Journal of Disaster Risk Reduction*, Article 103107.

Suwaileh, R., Imran, M., Elsayed, T., & Sajjad, H. (2020). Are we ready for this disaster? Towards location mention recognition from crisis tweets. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6252–6263).

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1079–1088).

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, *32*(1), 1–29.

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

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 2541–2544).

Wu, J., Hu, R., Li, D., Ren, L., Hu, W., & Xiao, Y. (2022). Where have you been: Dual spatiotemporal-aware user mobility modeling for missing check-in POI identification. *Information Processing & Management*, *59*(5), Article 103030.

Xu, C., Pei, J., Li, J., Li, C., Luo, X., & Ji, D. (2019). DLocRL: A deep learning pipeline for fine-grained location recognition and linking in tweets. In *Proceedings of the world wide web conference* (pp. 3391–3397).

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2145–2158). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, *2014*(9), 37–70.

Zhang, C., Zhang, K., Yuan, Q., Peng, H., Zheng, Y., Hanratty, T., et al. (2017). Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th international conference on world wide web* (pp. 361–370).

Zhao, S., Zhao, T., King, I., & Lyu, M. R. (2017). Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web companion* (pp. 153–162).

Zheng, X., Han, J., & Sun, A. (2018). A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1652–1671.