QATAR UNIVERSITY

COLLEGE OF ENGINEERING

LOCATION MENTION PREDICTION FROM DISASTER TWEETS

BY

REEM ALI SUWAILEH

A Dissertation Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Doctorate of Philosophy in Computer Science

June 2023

COMMITTEE PAGE

The members of the Committee approve the Dissertation of

Reem Ali Suwaileh defended on 22/05/2023.

Dr. Tamer Elsayed
Dissertation Supervisor

Dr. Muhammad Imran
Dissertation Co-Supervisor

Prof. Doina Caragea
Committee Member

Prof. Cagatay Catal
Committee Member

Dr. Abdelkarim Erradi
Committee Member

Prof. Fayçal Bensaali
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

Suwaileh, Reem, Ali., Doctorate : June : 2023, Doctorate of Philosophy in Computer Science

Title: Location Mention Prediction from Disaster Tweets

Supervisor of Dissertation: Dr. Tamer Elsayed.

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 and resolve geolocation information. This dissertation focuses on the Location Mention Prediction (LMP) problem that consists of Location Mention Recognition (LMR) and Location Mention Disambiguation (LMD) tasks. Our work contributes to studying two main factors that influence the robustness of LMP systems: (i) the dataset used to train the model, and (ii) the learning model. As for the training dataset, we study the best training and evaluation strategies to exploit existing datasets and tools at the onset of disaster events. We emphasize that the size of training data matters and recommend considering the data domain, the disaster domain, and geographical proximity when training LMR models. We further construct the public IDRISI datasets, the largest to date English and first Arabic datasets for the LMP tasks. Rigorous analysis and experiments show that the IDRISI datasets are diverse, and domain and geographically generalizable, compared to existing datasets. As for the learning models, the LMP tasks are understudied in the disaster management domain. To address this, we reformulate the LMR and LMD modeling and evaluation to better suit the requirements of the response authorities. Moreover, we introduce competitive and state-of-the-art LMR and LMD models that are compared against a representative set of baselines for both Arabic and English languages.

DEDICATION

To my parents.

ACKNOWLEDGMENTS

First and foremost, Alhamdulillah (praise be to God)!

I would like to express my sincere gratitude to my supervisors, Dr. Tamer Elsayed and Dr. Muhammad Imran for their enduring support and much-appreciated guidance throughout my dissertation. I owe them too much for the researcher who I am today. Jazakum ALLAH Khairan!

I am grateful to the members of bigIR group for their support, academic and life advice, and being models in enthusiasm and ingenuity. Warmest thanks goes out to Fatima Haouari and Maram Hasanain in particular for the joyful times in research. I would also like to thank all positive and kind people at Qatar University, especially the ones who always smile :)!

I extend my heartfelt gratitude to my parents for their unconditional love, support, and continuous prayers. I also thank my siblings Aisha, Yousuf, Mohamed, Wafa, Salah (and his family), and Nabil (and his family) for their continuous support, love, and prayers <3! I cannot forget my lovely supporters, all my relatives. Special thanks go out to Hana, Hind, Saba, Aisha, Zienab, Balqees, Arwa, Aljoori, Aisha, Noof, and Shooq. I am greatly thankful to all my friends for the adventures, gatherings, support, love, and prayers. Special thanks to Noura, Faiza, Waseema, Shatha, Safa, and Laila. Throughout my college years, I have met special CS colleagues who are more than colleagues that I would like to thank. Special thanks go out to Tooba, Linah, Sara, Nada, and Rahma.

Finally, this dissertation 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.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

During disaster and emergency events, georeferenced information is crucial for response authorities to identify impacted areas and population segments, and plan relief operations. Concretely, *crisis* (or *disaster) maps* are invaluable tools for drawing up-to-date situational awareness [1] and for taking actions that should ideally be performed within the first 48 to 72 hours of a disaster event [2]. Situational awareness is concerned with capturing the status on the ground, such as the inundated areas, landslides, electricity outages, or resource needs. Actionable information, on the other hand, refers to requests that demand actions from a particular response authority [3]. Examples include offering resources (e.g., food, water, and shelter), rescuing trapped or injured individuals, and rebuilding or repairing damaged infrastructure (e.g., bridges, roads, hospitals). The geographical context raises the value of situational and actionable information in several ways. First, responders must determine whether the reported incidents and requests are within their jurisdiction; otherwise, they redirect them to the responsible authorities [3]. Second, responders use geolocation information to locate incidents, resources, and requests for timely decision-making and response. Indeed, the geolocation information makes requests actionable as it specifies *where* the help is needed. Third, responders need to assess the overall impact of the disaster event at different granularity (e.g., state or city). For these use cases, crisis maps greatly help response authorities.

Crisis mapping refers to the real-time acquisition, analysis, and visualization of relevant information during a crisis [4]. Examples of crisis maps are (1) spatial situational awareness maps, (2) hotspot maps of causalities, damages, and resources, (3) first responders (i.e., eyewitnesses) and resources maps (e.g., food and shelters), (4)

population mobility maps (e.g., evacuations), or (5) impact assessment maps, among others (refer to Section 6.2.2 for further elaboration). The process of crisis mapping leverages multiple heterogeneous data sources, such as mobile and web technologies, hotlines, volunteering, crowdsourcing, and physical surveys. Nevertheless, relying on traditional sources to perform real-time crisis mapping during large-scale disaster events becomes challenging. Several studies demonstrated the effectiveness of non-traditional data sources such as social networking sites and remote sensing to acquire real-time crisis information [5]. Twitter, particularly during disaster events, has been proven to be a useful information source to gather time-sensitive situational and actionable data directly posted by the affected people [6]. Notably, effective social sensing via crisis mapping depends on the availability and quality of geolocation information on Twitter [7].

What makes Twitter content invaluable is the fine- and coarse-grained locations of incidents and needs that are reported by eyewitnesses [8]–[11]. There are different successful real-world examples of exploiting Twitter for disaster response. For instance, the Ushahidi platform [12] was deployed to map geotagged tweets during the Port-au-Prince earthquake 2010 in Haiti [13]. It was also used for Typhoon Haiyan 2013 in Southeast Asia to map damages and requests. Furthermore, Fairfax County in Virginia, US, is a case in point. It employed the Geofeedia platform to monitor and aggregate data from various social media platforms, including Twitter. It also took part in releasing the "National Capital Region News and Information" portal that offers geotagging capabilities for crisis communication and management.

Response agencies' requirements to use Twitter for situational awareness or mapping tasks vary. In a participatory design workshop, participants from different

agencies (e.g., police officers, firefighters, paramedics, among others) provided a set of example tweets for what they look for on Twitter to respond during disaster events. Most of these tweets contain fine-grained LMs such as intersections and buildings [8]. Other studies have emphasized the need for coarse-grained locations in planning relief activities and assessing the disaster impact by emergency managers [9]–[11].

Nevertheless, Twitter announced removing the geotagging feature in tweets in June 2019[1] as users often set imprecise geotags to their tweets which necessitates the development of automated geolocation information extraction tools. This dissertation addresses this need to enable drawing useful situational awareness reports and actionable requests from Twitter in the disaster management domain. Different computational tasks were defined in the literature over Twitter for geolocation information extraction, including *User Location Prediction*, *Home Location Prediction*, *User Movement Modeling*, *Next Location Prediction*, *Tweet Location Predicting*, *Locational Focus Prediction*, and *Location Mention Prediction*. However, we mainly focus on the *Location Mention Prediction* (LMP) task because it is vital in tackling all other geolocation computational tasks. The LMP task aims to (1) extract Location Mentions (LMs) from the textual content of tweets, known as *Location Mention Recognition* (LMR), and (2) disambiguate them using toponyms from geo-positioning databases (i.e., gazetteers), known as *Location Mention Disambiguation* (LMD).

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

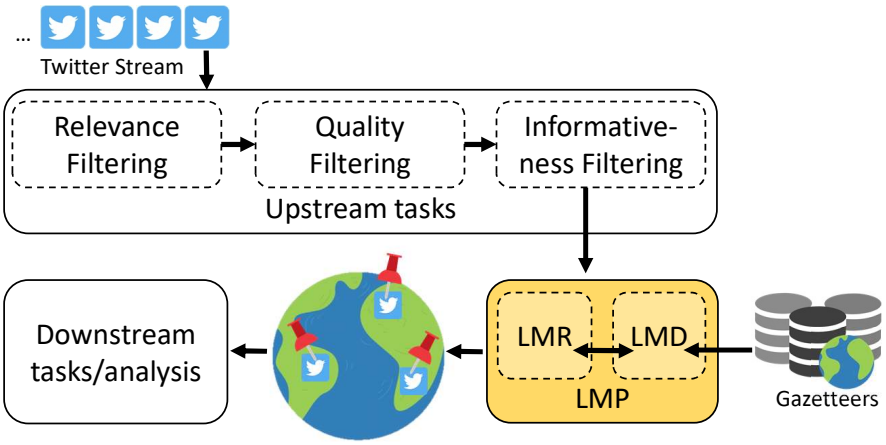


Figure 1.1. Locating the LMP task in the disaster response pipeline

We provide an overview of the LMP problem in Section 1.1. We further discuss the challenges associated with utilizing geolocation information from Twitter for crisis management in Section 1.2. Next, we briefly summarize our proposed solutions and findings in Sections 1.3 and 1.4. We then thoroughly elaborate on the contributions of this dissertation in Section 1.5.

1.1. Research Problem: Location Mention Prediction

To articulate the role of an LMP module in the emergency management domain, we depict a high-level computational response pipeline in Figure 1.1. While tweets are noisy and arrive at a very high rate, the responders demand high-quality situational reports and actionable tweets for reliable decision-making and relief deployment. Hence, pre-filters have to precede the LMP components; these pre-filters are depicted as upstream tasks in Figure 1.1. Among them are (i) relevance filters: to discard all irrelevant content to the target disaster event, (ii) qualification filters: to discard spam, rumors, and bot-generated content, among other low-quality content, (iii) informativeness filters: to filter out sympathy, opinions, and criticism, among other less informative content. The

pre-qualified tweets continue to the LMP module that constitutes two main components: (i) *Location Mention Recognition* (LMR) to extract toponym spans from the text of tweets, and (ii) *Location Mention Disambiguation* (LMD) to link the potential extracted LMs to existing toponyms in a geo-positioning database (i.e., gazetteer). Finally, the output of the LMP components can be directly used by the disaster response authorities or fed into other downstream tasks such as mapping services (refer to Section 6.2.2). Figure 1.2 illustrates a high-level overview of the LMP problem and its tasks.

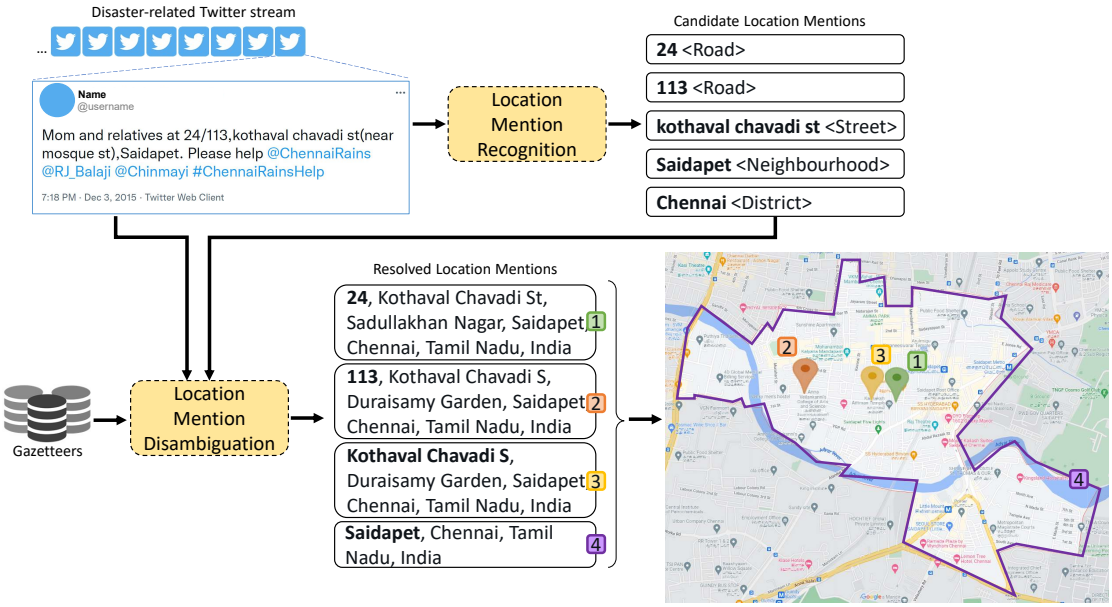


Figure 1.2. High-level overview of the LMP tasks.

The LMR and LMD tasks have different names in the literature. For example, some studies refer to LMR as location extraction or geoparsing. The LMD has been alternatively named resolution, linking (looking up a geo-positioning database to find matches), or geocoding (assigning geo-coordinates to LMs regardless of the sources used). In rare cases, geoparsing could jointly refer to both LMR and LMD tasks. We use LMR and LMD task names throughout this dissertation for expository clarity.

Table 1.1 shows a few example tweets shared during different real-world disaster

events including *Chennai floods 2015*, *Houston floods 2016*, *Louisiana floods 2016*, *Christchurch earthquake 2012*, and *Hurricane Sandy 2012* [14], [15]. These tweets are important for the relief organizations and first-responders, who exploit Twitter to extract (i) situational reports such as incidents and casualties statistics, or (ii) actionable information, including the calls for rescue, requests for resources, need for volunteers, to name a few.

Locations appear in different patterns and types in tweets, such as:

- *Full address*: Examples are the location where rescue boats are needed in tweet #1, and the location of reported water levels in tweet #3.

- *Administrative divisions*: Different administrative levels are commonly mentioned during disasters. For instance, states are mentioned for high level updates, such as "TX" and "LA" in tweets #3 and #5, respectively. Cities are also reported, such as "Valley Mills" and "Ocean City" in tweets #4 and #8, respectively. Neighborhoods are also mentioned, such as "Greens Bayou" in tweet #3.

- *Points-of-interest (POI)*: Fine-grained locations appear frequently on Twitter during disaster events. For instance, the "Bosque River" in tweet #4 presents a natural POI where a flood warning is issued. The human-made POIs are also frequent such as institutions (e.g., "Cashmere kindergarten" in tweet #7), or worship places (e.g., "Tangipahoa Parish" in tweet #3).

- *Streets*: Similar to the POIs, streets are important to study the impact of disaster on the population and responders mobility during disaster events. This includes bridges, e.g., "Adayar Bridge Saidape" and "TVK bridge" in tweets #1 and #2, respectively, and flooded streets, e.g., "East 8th" and "Avenue C" in tweet #9.

Table 1.1. Tweets from real-world disaster events with location mentions (gray-shaded). HRC, EQK and FLD refer to Hurricane, Earthquake, and Floods respectively.

T#	Dataset	Tweet text
#1	Chennai FLD	[user_mention] Dear Friends, Pl help by sending boat to 54 and 58, Vivekananda Nagar Street, Nesapakkm, Chennai [...]
#2		[user_mention] Fear bridge being washed away. Adayar Bridge Saidapet. Hope TVK bridge is holding up fine at Malhar [url]
#3	Houston FLD	#USGS08076700 - Greens Bayou at Ley Rd, Houston, TX is above NWS flood stage (30ft) [URL]
#4		FWD cancels Flood Warning for North Bosque River at Valley Mills [TX] [url] #ntxwx
#5	Louisiana FLD	Flash Flood Warning for Livingston, St. Helena, and Tangipahoa Parish in LA until 7:45am Saturday.
#6		This line of storms in Evangeline is moving to the southwest towards Allen, which will bring heavy rainfall #LAwx [url]
#7	ChCh EQK	RT [user_mention]: all kids safe at Cashmere kindergarten. #eqnz
#8	HRC Sandy	All roads into and out of Ocean City, New Jersey are closed due to flooding that has cut off the popular Jersey... [url]
#9		Flooding at East 8th and Avenue C before the blackout (GIF) [url]

1.2. Challenges

Learning to detect and disambiguate location mentions in tweets is a non-trivial task. As a result, LMP systems have to address many challenges related to the nature of Twitter stream or the difficulty of LMR and LMD tasks. This section elaborates on the challenges we address while tackling the LMP problem.

1.2.1. Twitter Stream Challenges

Data domain-wise, processing Twitter data requires tackling various challenges stemming from its stream's nature. We discuss a few in the following:

Tweet sparsity: Tweets are limited to only 280 characters causing the *lack of context* challenge for the learning models, which requires enriching the context of tweets using different techniques, e.g., text expansion.

Hashtag riding: Spammers typically use viral hashtags for advertisements, self-promotion, and propaganda, to list a few. A potential way to alleviate this challenge is pre-qualifying tweets before applying LMP. This enables the delivery of high-quality tweets to responders and reduce the processing time.

Mismatch between tweets and gazetteers: Compared to gazetteers, Twitter stream is noisy; tweets contain informal language, misspellings, grammar mistakes, shortened words, and slang, causing the so-called mismatch challenge [16]. Employing semantic models alongside the lexical models for text processing could alleviate this challenge. We list different types of issues in the following with examples in Table 1.2:

- **Nicknames**: Some places have common nicknames used by locals. For example, in Tweet #1, *Chennai* is nicknamed "The Detroit of India". The nicknames often do not exist in the gazetteers.

- **Abbreviations**: Short names of places are prevalent on Twitter due to the character limit of tweets. For example, "T. Nagar" and "GM Chetty Road" are abbreviations of "Theagaraya Nagar" and "Gopathi Narayanaswami Chetty", respectively, in Tweet #3.

- **Misspellings**: Misspellings and grammar mistakes are common over Twitter. For

Table 1.2. Example tweets from Chennai floods to illustrate the challenges of processing Twitter stream for LMP task. LMs are gray-shaded in text.

T#	Challenge	Tweet text
#1	Nickname	#ChennaiFloods sad to see the state of city. Detroit of India is suffering. Hv personal experienced.
#2	Capitalization	Accommodation in t nagar to 30-50 people in Rameswaram Road, T. Nagar. Contact 9843111199 #ChennaiRainsHelp #ChennaiFloods chennai micro
#3	Abbreviations	Anyone around T. Nagar, needing shelter or food, can approach the Gurudwara on GM Chetty Road #Chennai
#4	Misspelling	Medical students of shri ramchandra medical college in chennai stranded without supplies. Need help.
#5	Shortcuts	sm 1 help providing water 50 children @Lawrence Charitable Trust.safe.2/4,1st cross st,3rd avenue,AshokNagar-LakshmanSruti #ChennaiFloods

instance,“**shri** ramchandra medical college" in Tweet #4 should be written as “**sri** ramchandra medical college".

- **Shortcuts**: Users tend to use shortened words due to the character limit of tweets. For example, using “st” instead of “road”, in Tweet #5. Also, using “@” symbol instead of the literal “at” prepositions in the same tweet.

- **Capitalization**: Users tend to ignore capitalization when writing tweets (e.g., “chennai" instead of “Chennai" in Tweet #4.

1.2.2. LMP Task-Specific Challenges

Tackling the LMP task imposes addressing several challenges. We discuss some of them below and elaborate further on LMR and LMD task-specific challenges in Sections 1.2.3 and 1.2.4.

Scarcity of labeled data: Supervised learning algorithms need a large representative dataset to perform effectively. However, acquiring training datasets is critical in disaster management where systems must be trained and deployed promptly *in real-time*. We extensively review the existing public LMR and LMD datasets and discuss their limitations in Sections 2.4.2.1 and 2.5.2.1.

Time-criticality of solutions: Empirically effective solutions are not necessarily ready for efficient deployment at the onset of disaster events. Therefore, the developed systems must be trained and evaluated to run in real-time to enable effective crisis management.

1.2.3. LMR-Specific Challenges

Different challenges arise when tackling the LMR task, including:

Emerging locations: New toponyms emerge while disaster events develop over the Twitter stream. Recognizing frequent toponyms is easier than unseen ones as learnable models often memorize the vocabulary of toponyms rather than learning their syntactic and semantic patterns, preventing generalizing to unseen data.

Toponymic polysemy: Location mentions might have different meanings referring to different entity types other than locations. For example, "Sabah Al-Ahmad" could mean the "Sabah Al-Ahmad city, in Kuwait" or the former Emir of Kuwait "Sabah Al-Ahmad Al-Jaber Al-Sabah".

Incompleteness of gazetteers: The gazetteer-based LMR approaches verify LMs using gazetteers before making the final predictions. Nevertheless, many locations, especially the fine-grained locations, might not appear in the gazetteers causing failure in detection.

Temporary locations: Temporary facilities (i.e., medical camps and shelters) are constructed during emergencies to provide resources and support the affected people. How-

ever, these facilities are disassembled (e.q., quarantine centers) once the emergency is over. Additionally, the names of some locations could change during emergencies, such as converting schools into shelters and giving them new expressive names (e.g., “main shelter”). Once the disaster event is over, schools will return to providing their original services. The difficulty of detecting and disambiguating these temporary locations is due to the need to comprehend their context.

1.2.4. LMD-Specific Challenges

Tackling the LMD task requires alleviating different challenges, including:

Toponymic homonymy: The same location name might refer to different locations. For instance, "Doha" might refer to the "capital city of Qatar state" or the "Doha city in Kuwait," "Kuwait" could refer to the "State of Kuwait" or the capital city "Kuwait" (different granularity), and "Ooredoo" may refer to any branch of the telecommunications service provider inside, or outside Qatar.

Incompleteness of gazetteers: LMD systems will not be able to resolve missing LMs in gazetteers when detected by the LMR systems. Therefore, consolidating multiple gazetteers is a pivotal solution to cover as many locations as possible with different properties into a unified gazetteer. Nevertheless, the augmentation and deduplication processes are nontrivial.

Dynamism: The toponyms in gazetteers are dynamic and do change over time due to: (i) changing names (or properties) of locations (e.g., street names), (ii) deleting locations (e.g., permanently closing a restaurant), or (iii) building and opening new facilities (e.g., malls, airports, and parks) which require maintaining up-to-date gazetteers.

1.3. Overview of Proposed Solutions

Two main factors that influence the robustness of an LMP system are: (i) the learning model, and (ii) the dataset used to train the model. As for the learning model for the LMR task, there are two well-established approaches. The first is adopting existing general-purpose Named Entity Recognition (NER) taggers. NER task aims to extract entity mentions (e.g., location and organization) in a given text. However, the general-purpose NER systems fail to generalize to Twitter data due to the noisiness of the Twitter stream (refer to the "mismatch" challenge in Section 1.2.1). The second common approach is matching potential LMs against gazetteers. However, the gazetteer-based approaches fail to generalize to toponyms that do not exist in the employed gazetteers (refer to the "incompleteness" issue in Section 1.2.4). Recent studies have proposed deep learning approaches to alleviate the mismatch and incompleteness challenges. Nevertheless, deep learning models are data hungry and demand longer training time, which introduces a limitation when deploying them at the onset of disaster events (refer to the "scarcity of labeled data" and "time-criticality of solutions" challenges in Section 1.2.2). To address the challenges above, we fine-tune a BERT pre-trained model [17] for the LMR task for two reasons. First, it eliminates the cost of hand-crafting features. Second, it does not require huge training data to perform reasonably.

As for the training data, existing studies assume sufficient training data is available. However, we explore how the choice of training and evaluating the LMR models influences their performance in the disaster management domain. Hence, our empirical exploration contributes to the effectiveness and efficiency of deploying the LMR models in emergencies. We investigate the effect of multiple factors on the LMR model during training, including the *data domain*, *entity types*, *disaster domain*, *geo-proximity*, and

language.

The dataset exploration we performed has revealed many limitations to address, such as the limited size, the confined domain and geographical coverage, the absence of location type annotations, and the inconsistency of annotations, among others. Unfortunately, there is no *public* Arabic LMR dataset up to the time of this writing. In fact, the absence of large and generalizable LMR datasets impedes the comparison of existing LMR models. Thus, we build IDRISI-R datasets that consist of the largest to date available English LMR dataset (IDRISI-RE) and the first public Arabic LMR dataset (IDRISI-RA), which contributes to enriching the low resources languages. For that, we extend the English HumAID and Arabic Kawarith disaster datasets that are labeled for humanitarian categories to combat the low-quality content (refer to the "hashtag riding" challenge in Section 1.2.1). Additionally, we conduct extensive analysis on the reliability (high inter-annotator agreement), coverage (geographical, domain, temporal, dialectical for Arabic, and location type granularity), and generalizability (domain, geographical, and over unseen events) of the datasets showing that IDRISI-R is second to none.

We further extend the IDRISI-R datasets for the LMD task and introduce IDRISI-D datasets (IDRISI-DE English dataset and IDRISI-DA Arabic dataset). IDRISI-DE is the largest English LMD dataset (toponym-wise) to date. IDRISI-DA is the first Arabic LMD dataset that enriches the low resources languages. We annotate both datasets by (i) linking LMs to toponyms in OpenStreetMap (OSM) gazetteer, and (ii) judging the usefulness of different features (e.g., URLs, replies, entities) for disambiguating LMs. The goal of exploring the usefulness of features is to tackle the tweet sparsity challenge (refer to Section 1.2.1) by context expansion.

As for the learning models, we have also trained our two models, namely CRF_{LMR}

and BERT$_{LMR}$ models. We extensively compare our models against a representative set of English LMR models and Arabic NER models over IDRISI-R datasets. Our models showed state-of-the-art performance under different data and task setups. As for the LMD task, different from existing studies, we perceive the LMD task as a ranking problem allowing a lenient hierarchical evaluation of solutions. We further propose employing a pre-trained model (BERT$_{LMD}$) to account for the efficient deployment of models in the disaster domain with competitive effectiveness.

1.4. Findings Overview

Below, we elaborate on the essential findings that pave the way for better future research on LMP in the crisis domain.

Our empirical exploration of the best practices of using the existing datasets and tools for training and evaluation at the onset of disaster events suggests that disaster-specific Twitter datasets are the best when compared to the general-purpose web and Twitter datasets [18], [19]. Limiting the training data to location entities, in contrast to using all types of entities (e.g., person or organization), makes a notable difference in performance. We also found that considering the disaster domain and geographical proximity are critical factors in improving the LMR performance. Additionally, as little as 263-356 tweets from the target language could improve the performance when combined with all available multilingual data. One fundamental rule we confirm in our study is training on all available data from all domains to minimize the labeling cost at the onset of disaster events; this means technically developing and testing the models under zero-shot learning. When little budget is available to annotate tweets from the target event, labeling only 500 tweets would be sufficient to obtain reasonable

LMR models; this means technically developing and testing the models under few-shot learning.

Over and above, our exploration on existing datasets highlighted the need for larger and more generalizable datasets for English and Arabic. We target English for having several datasets that suffer from many shortcomings. We target Arabic for being a low resources language with no public Twitter LMR datasets. Our rigorous empirical analysis on the generalizability of our IDRISI-R datasets 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 geographical coverage and data size are the top influencers on the generalizability of the LMR datasets. Our experiments confirm the reasonable generalizability of IDRISI-R datasets. Both IDRISI-RE and IDRISI-RA datasets show decent reliability, reasonable geographical, domain, temporal coverage, and location type annotations. Moreover, the benchmarking experiments testify BERT_{LMR} as the state-of-the-art LMR model over both IDRISI-RE and IDRISI-RA datasets.

Furthermore, both IDRISI-RE and IDRISI-RA datasets are labeled for features' usefulness. The analysis of the manual annotations showed that the event context, hashtags, and other location mentions appearing within the same tweet are helpful for accurate disambiguation. Our experiments confirm that the BERT_{LMD} model is competitive over IDRISI-DE dataset and provides a state-of-the-art performance over IDRISI-RA dataset.

Throughout the dissertation, we elaborate on these findings and their respective research questions.

1.5. Contributions

The contribution of this dissertation is multifold, covering both learning models and datasets. In Table 1.3, we link these contributions to the respective chapters.

a. We exhaustively and comparatively review the disaster-specific LMP solutions and evaluation tools over crisis tweets. The review describes the current status, the associated challenges, and the future directions. It also attempts to link stakeholders' requirements with existing geolocation computational tasks.

b. We empirically study the best practices of exploiting the existing resources and tools for effective LMR at the onset of disaster events:

   - We tackle the bottleneck of lack of annotated data, drawbacks of gazetteer-based solutions, and the high cost of hand-engineered features by exploiting the vanilla pre-trained BERT model for LMR (BERT$_{LMR}$).
   - We study the effect of different factors, including data domain, entity types, disaster domain, geo-proximity, and language, under zero-shot learning. Out of this exploration, we recommend the best practices for deploying reasonable LMR models at the onset of disaster events.
   - We investigate the cost of incrementally acquiring target labeled data at the onset of disaster events for training reasonably performing LMR model (i.e., few-shot learning).
   - We conduct a failure analysis on our BERT$_{LMR}$ model to gain insights for the future development of LMR models.

c. We release six public LMR and LMD datasets[2] including:

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

- IDRISI-RE: The largest to date *manually*-labeled English LMR dataset (gold version). It contains around 20.5K tweets and 21.9K LMs.
- IDRISI-RA: The first *manually*-labeled Arabic LMR dataset (gold version). It contains around 4.6K tweets and 5.2K LMs.
- IDRISI-RE$_{silver}$: The largest *automatically*-labeled English LMR dataset (silver version) constituting around 57K tweets and 43.4K LMs.
- IDRISI-RA$_{silver}$: The largest *automatically*-labeled Arabic LMR dataset (silver version) constituting around 1.2M tweets and 884K LMs.
- IDRISI-DE: The largest to date *manually*-labeled English LMD dataset of around 9.6K tweets.
- IDRISI-DA: The first *manually*-labeled Arabic LMD dataset of around 4.7K tweets.

d. The value of the IDRISI datasets is due to the types of annotations that we collected, including:

- The LMR annotations include location mentions and their coarse- and fine-grained location types.
- The LMD annotations comprise linking location mentions to toponyms in OpenStreetMap (OSM) and the human assessment of the usefulness of different features for disambiguation, including the event context, URLs, hashtags, entities, and other locations.

e. A key advantage of IDRISI datasets is their *domain* and *geographical* generalizability. We empirically demonstrate that

- IDRISI-RE dataset is the best generalizable dataset compared to the existing

English datasets.

- IDRISI-RA dataset is a reasonably generalizable dataset.

f. We establish a set of baselines for the research community:

   - We benchmark the IDRISI-RE dataset using diverse and representative English LMR models
   - We benchmark the IDRISI-RA dataset using standard Arabic NER models.

g. We develop and present the state-of-the-art English and Arabic BERT$_{LMR}$ models for the disaster domain.

h. We develop and present the competitive English and state-of-the-art Arabic BERT$_{LMD}$ models for the disaster domain.

We motivate and elaborate on each of these contributions in their respective chapters. We also list the research outcomes in Section 6.3.

Table 1.3. Links between contributions and respective chapters in the dissertation.

		Chapter #			
		2	3	4	5
Contribution #	a	✓			
	b		✓		
	c			✓	
	d			✓	
	e			✓	
	f			✓	
	g			✓	
	h				✓

The remainder of this dissertation is organized as follows. Chapter 2 discusses the usefulness of exploiting Twitter in the crisis management domain (Section 2.1) and

presents the Twitter geolocation studies (Section 2.2). It also briefly reviews the non-disaster LMP studies (Section 2.3). It then thoroughly discusses both LMR (Section 2.4) and LMD (Section 2.5) approaches in the disaster management domain. Chapter 3 discusses our empirical exploration of the best practices of utilizing the existing resources and tools for the *Location Mention Recognition* task at the onset of disaster events. It presents the problem definition (Section 3.1), methodology (Section 3.2), experimental setups (Section 3.3), results (Section 3.4), and a failure analysis (Section 3.5). We conclude the chapter with a list of limitations of the study in Section 3.6. Chapter 4 presents our efforts in creating IDRISI datasets and benchmarks. It introduces IDRISI-R English (IDRISI-RE) and Arabic (IDRISI-RA) LMR datasets and benchmarks in Sections 4.2 and 4.3, respectively. This includes construction (Sections 4.2.1 and 4.3.1), description and quality (Sections 4.2.2 and 4.3.2), benchmarking experiments (Sections 4.2.3 and 4.3.3), and empirical analysis on generalizability (Sections 4.2.4 and 4.3.4). It then presents the silver LMR datasets in Section 4.4. The chapter also introduces IDRISI-D English (IDRISI-DE) and Arabic (IDRISI-DA) LMD datasets in Section 4.5. It presents the construction efforts of the dataset (Section 4.5.1), and description and quality analysis (Section 4.5.2). The chapter concludes with a discussion on datasets' limitations in Section 4.6. Chapter 5 discusses the *Location Mention Disambiguation* task in detail including the problem formulation (Section 5.1), methodology (Section 5.2), experimental setups (Section 5.3), and results (Section 5.4). Chapter 6 sums up the entire dissertation in Section 6.1. It also discusses this dissertation's theoretical, practical, and research implications in Section 6.2. It then lists the research outcomes in Section 6.3 and elaborates on potential future directions in Section 6.4.

CHAPTER 2: RELATED WORK

In this section, we thoroughly review the literature. We first discuss the information extraction research from Twitter for crisis management in Section 2.1. Next, we discuss the geolocation information extraction over Twitter for crisis management in Section 2.2. We then provide an overview of LMP studies in general domains in Section 2.3. Finally, we exhaustively review the LMR and LMD tasks in Sections 2.4 and 2.5, respectively.

2.1. Information Extraction from Twitter for Crisis Management

Recently, Hiltz, Hughes, Imran, *et al.* [20] conducted a survey to prioritize the computational tasks developed by technologists based on the guidance of experts from the crisis management domain in different countries. The study showed that Twitter is the most preferred social media platform by experts during emergencies, alongside Facebook. Technologists have made invaluable efforts to utilize social media for preparedness, relief, and recovery of emergency situations [21]. The proposed tasks involve but are not limited to, detecting disasters and incidents [22]–[24], summarizing them [25], filtering relevant tweets [26]–[29], identifying situational reports [30]–[33], identifying actionable information [3], [34]–[36], geolocation inference and other types of information extraction [14], [37]–[40].

Nevertheless, current solutions are rarely deployed by relief organizations [41], [42] due to several reasons. For instance, the unreliability of information, the inefficiency of solutions in disaster scenarios, and the lack of customized solutions for the different needs of different stakeholders. Fortunately, there are some recent efforts to bridge the gap between technologists and responders from relief organizations by understanding

their needs [20] and the utility of existing solutions [3].

For example, Hughes and Shah [43] proposed a monitoring and analytical application that helps Public Information Officers (PIO) to document and report information about emergencies from social media. The development of the application was in light of observation of the PIO activities. Furthermore, Vieweg, Hughes, Starbird, *et al.* [30] was the first to define different types of situational updates grounded on a manual exploration of disaster datasets. The list of types has evolved to include finer classes and actionable classes, such as calls for actions (evacuation, volunteers, donations) [35], availability and needs for different resources (generally or for a specific location), and activities of relief organizations [36]. Further efforts have been put into establishing the definition of actionability and defining the criteria for ranking actionable tweets to prioritize response [44]. Furthermore, Zade, Shah, Rangarajan, *et al.* [3] moved the attention beyond extracting decision-support reports to aiming at supporting mission-specific responders. Through surveys and interviews with response authorities, they explored the varying definitions of actionability for different responders. Similarly, Kropczynski, Grace, Coche, *et al.* [8] investigated the characteristics of actionable tweets by interviewing administrators, telecommunicators, and first responders. Researchers have also spent efforts on extracting geolocation information. Next, we elaborate further on this and show the vital role of geolocation information in the disaster management domain.

2.2. Geolocation Information Extraction over Twitter for Crisis Management

According to Hiltz, Hughes, Imran, *et al.* [20], *grouping social media content on a map by their geographic locations* is the most demanded feature by responders. Another important feature was to *automatically geotag posts*. In addition to providing

a spatial view of the situation during emergencies, these features facilitate efficient management of relief activities by making actions or routing them to the responsible authorities [3], [8]. Geolocation information allows responders to locate resources (e.g., food, shelters, etc.) and their status, activities (e.g., evacuation zones), causalities, and damages [8]–[11]. Since the early emergence of social media, relief organizations have utilized geotagged content. A good case in point is exploiting mashup technologies to map disaster events and improve situational awareness. Those applications follow a crowd-source-based model for collecting updates (e.g., comments, photos, or videos of incidents). Different mashup platforms have been used to map crisis data (real-time mapping of situational and actionable data), such as Ushahidi,[1] Geofeedia,[2] ESRI-ArcGIS,[3] Google Crisis Response,[4] and Factal.[5]

Many such services were deployed during past disasters, including the 2020 earthquake in Port-au-Prince, Haiti,[6] Typhoon Haiyan (Visov) in 2013, and Chennai Floods 2016, in India.[7] Furthermore, Fairfax County in Virginia, US, explored the usefulness of Geofeedia that provides location-based analytical modules to monitor and aggregate various social media data, including Twitter, Instagram, and YouTube. The county also took part in releasing the "National Capital Region News and Information" portal that uses a web-based system to exploit social media and geotagging capabilities for crisis communication and management.[8]

Additionally, Roy, Hasan, and Mozumder [45] created dynamic disruption maps from Twitter data to visualize types of disruption and their status. They employed

[1] www.ushahidi.com/
[2] en.wikipedia.org/wiki/Geofeedia
[3] www.esri.com/en-us/disaster-response/overview
[4] crisisresponse.google/
[5] www.factal.com
[6] www.hsdl.org/?abstract&did=805223
[7] www.ushahidi.com/support/examples-of-deployments
[8] www.hsdl.org/?abstract&did=805223

the NLTK NER model [46] to extract location mentions from tweets' text. Hong and Frias-Martinez [47], on the other hand, used the geolocation information extracted from Twitter data to model evacuation flow patterns at different coarse-grained geographical levels such as country, state, and areas. This study is limited to users with the automatic geotagging feature of tweets enabled, which allows tracking and analyzing users' location during the disaster. However, Twitter removed this feature in 2019.[9] Following a similar line of analysis, Roy and Hasan [48] used the geolocation information of tweets to infer the evacuation behavior of individuals during hurricanes, including whether people evacuated or not, when did they evacuate, and what are their destination points. The study aimed to extract the effect of evacuation behavior on highway traffic.

Moreover, Uchida, Kosugi, Endo, *et al.* [49] implemented a real-time system to support the collaborative response through reporting and retrieving Twitter disaster-related content. The system combined two web-based subsystems for sharing and mapping information. The information-sharing subsystem attaches the user location to the tweet used by the mapping subsystem to pin the content on the map. The system started operating in early 2015 and improved further in 2017 [50]. Kosugi, Utsu, Tomita, *et al.* [51] introduced a better and more user-friendly web-based real-time system to support collaborative response with the same types of use cases, including reporting, displaying reports (latest or nearby) and facilities' locations (evacuation places or disaster base medical centers), and searching reports. Zhang, Fan, Yao, *et al.* [52] reviewed several other state-of-the-art applications of social media informatics in disaster events and highlighted their challenges. The authors also proposed some research frontiers for social media informatics in the disaster management domain.

At the other end of the spectrum, a large body of the technical literature focuses

[9] twitter.com/TwitterSupport/status/1141039841993355264

on different geolocation computational tasks (e.g., user and tweet location prediction, and location mention prediction) and diverse data domains (news articles, research articles, and social media posts) [37]. The central task among these is the LMR task, which aids all others by extracting evidence from the text for the user, tweet, and incident locations. Therefore, we exhaustively review the LMP studies in the following sections.

2.3. Location Mention Prediction over Twitter for General Domains

The first exploration of the LMR task is dated back to the Message Understanding Conference (MUC) [53] in 1996 as part of the Named Entity Recognition (NER) task. The LMD task, on the other hand, has its root, as part of the Entity Linking (EL) task in Natural Language Processing (NLP) as cross-document coreference resolution and in databases as record linkage [54]. The practical role of LMR and LMD tasks is evident through the different available spatial address geocoding systems and APIs such as commercial Google Maps platform,[10] the open-source QGIS,[11] ArcGIS,[12] and TomTom,[13] among others. Moreover, the LMR and LMD tasks are essential for different domains such as crisis management domain [14], [15], [18], [55]–[72], traffic monitoring [39], [73]–[75], POIs recommendation [76], [77], geographical text analysis and retrieval [78], [79], to name a few. In this section, we briefly review the non-disaster-specific LMR and LMD solutions. We elaborate thoroughly on the disaster-specific LMR and LMD solutions in Sections 2.4 and 2.5, respectively.

[10] https://developers.google.com/maps
[11] https://qgis.org/en/site/
[12] https://geocode.arcgis.com/
[13] https://developer.tomtom.com/

2.3.1. Location Mention Recognition

Departing from the NER general task, Hoang and Mothe [80] analyzed the trade-off between recall and precision of NER tools alongside a filtering step using DBpedia[14] over tweets. The NER tools are StanfordNER, Gate NLP framework (Gate) [81], and Ritter tool (originally trained StarfordNER on Twitter data) [82]. They reported increased recall levels when using more than one NER tool. On the other hand, the precision improves when filtering tweets to determine whether they contain toponyms or not. TwitterStand [83] was among the first studies to exploit geotagging tweets' content for automatic extraction and mapping of breaking news. The tool uses a training-free method to extract candidate LMs to identify key phrases in tweets using term frequency. To resolve the candidate LMs, it searches the GeoNames gazetteer using pre-defined heuristics. Differently, Malmasi and Dras [84] extract noun phrases (NPs) from tweets using a recursive rule-based tree parser. To link extracted locations to Geonames' toponyms, they apply fuzzy matching. The major weakness of the gazetteer-based methods is the mismatch between the noisy Twitter stream and the often clean gazetteers. To alleviate this issue, Sultanik and Fink [85] proposed an Information Retrieval (IR) approach to identify the location mentions in tweets. They indexed the locations of gazetteers by their phonetic encodings using a K-D data structure to enable efficient matching using fuzzy matching to mitigate misspellings in the input data (e.g., Qatr versus Qatar.). For that, their matching component hashes the gazetteer's toponyms based on a phonetic encoding algorithm dubbed "Double Metaphone". More interestingly, a couple of studies [56], [66] adopted an ensemble-based LMR model to achieve high coverage of recognized locations. Zhang and Gelernter [66] combined the output of four LMR models

[14] http://dbpedia.org/snorql/

with different techniques, including lexico-semantic based, rules-based building, rules-based street, and machine learning-based parsers. Differently, Kinsella, Murdock, and O'Hare [86] proposed learning a language model of varying location granularity using geotagged tweets. Tweets are collected using geo-coordinates. Furthermore, in 2014, the topic of the fifth Australasian Language Technology Association (ALTA) shared task was on LMR in tweets [87]. Participants explored techniques such as feature engineering, ensemble classifiers, rule-based classification, knowledge infusion, CRFs sequence labelers, and semi-supervision. As for features, they used different features, including geospatial, structural, and lexical features. Participants used a retrained StanfordNER on tweet datasets as well.

2.3.2. Location Mention Disambiguation

A couple of studies [88], [89] had tackled the LMR and LMD tasks jointly to allow passing feedback between the LMR and LMD models for a robust LMP. Guo, Chang, and Kiciman [88] trained a structural SVM model to perform recognition and disambiguation tasks. Ji, Sun, Cong, *et al.* [89] used beam search to find the best combination of recognition and disambiguation labels. Nevertheless, this approach leads to error propagation between LMR and LMD models. To elaborate, inaccurately detected LMs would negatively affect the performance of the LMD model as it would fail to resolve them.

Contrarily, Li, Hu, Feng, *et al.* [90] conducted a coherence linking at the user level since location mentions in tweets posted by users usually fall within their home city. After detecting the user's home city using LMs appearing in the user timeline, they disambiguate the target LMs according to their relation to the home city. Similarly,

Ji, Sun, Cong, *et al.* [89] used a coherence measure that relies on the average distance between potentially relevant toponyms to the LMs.

2.4. Disaster-Specific Location Mention Recognition

In this section, we discuss the LMR studies from two angles, their technical solutions (Section 2.4.1) and evaluation tools (Section 2.4.2).

2.4.1. Solutions

Existing studies exploit different techniques and features to extract location mentions (LMs) from text [37]. However, it is worth mentioning that comparing the performance across approaches is unattainable due to the absence of a unified evaluation framework. Hence, this section's will solely focus on methodology.

2.4.1.1. NER-based Approaches with Domain Transfer

An intuitive LMR solution is to exploit off-the-shelf NER models. Existing studies did not only employ the existing NER models directly but also used the NER datasets to train their own LMR models. For instance, Lingad, Karimi, and Yin [55] explored the effectiveness of four NER models on toponyms extraction over disaster-related tweets. The models are StanfordNER [91], OpenNLP,[15] Yahoo! PlaceMaker,[16] and TwitterNLP [92]. The results showed that StanfordNER is the top-performing tool when retrained on Twitter data; otherwise, it poorly performs. Gelernter and Balaji [56] developed GeoLocator that uses OpenCalais NER tool [93] to extract toponyms and facilities (i.e., buildings). To combat the gazetteer incompleteness issue, they augmented

[15] `https://opennlp.apache.org/`
[16] No longer available.

OpenCalais with a list of building types to improve its recall when detecting buildings. Recently, GazPNE2 [94] exploited Stanza NER model to accelerate recognition and detect hard LMs.

Later, several LMR studies exploited StanfordNER tool due to its superiority. Among the four participating teams in the LMR shared task in the ALTA Workshop 2014 [87], a couple used the StanfordNER tool for toponym recognition. One team used it in a data transfer mode as a pre-trained model within an ensemble LMR system with rule-based modules that identify abbreviations and location specifiers in text [57]. Alternatively, Liu, Rahimi, Salehi, *et al.* [58] retrained the StanfordNER model over the ALTA training data. Ghahremanlou, Sherchan, and Thom [59] and Yin, Karimi, and Lingad [60] had also retrained StandfordNER using tweet datasets to improve its effectiveness. Furthermore, Mao, Thakur, Sparks, *et al.* [61] compared three NER models, including the original StanfordNER model, a retrained version on tweets, and the Bi-LSTM model to map places of power outages using Twitter data.

Following the same line of research, Nizzoli, Avvenuti, Tesconi, *et al.* [65] have recently used the NER dataset from the Named Entity rEcognition and Linking (NEEL) challenge [95] to train their LMR model. They then employed the TAGME tool [96] to capture meaningful short phrases in the text and match them against Wikipedia articles to detect LMs. Wang and Hu [68], alternatively, retrained the three top systems from the Toponym Resolution in Scientific Papers task, at SemEval 2019 [97], on CoNLL 2003 NER Web dataset [98]. The systems are *DM_NLP* [69], *UniMelb* [70], and *UArizona* [71]. The LMR models are at their core Bidirectional Long Short Term Memory (BiLSTM) network (Discussed further in Section 2.4.1.4). We discuss the disambiguation components of these systems in Section 2.5 in detail. Furthermore, to

create dynamic disruption maps during disaster events, Roy, Hasan, and Mozumder [45] extracted LMs from tweets using the NLTK-NER model [46].

To sum up, a key motivation for adopting *domain transfer* techniques is to mitigate the response latency of relief authorities at the onset of disaster events. The latency could occur due to the time-costly annotation of target disaster data. Directly applying the NER models trained on web data does not lead to effective performance in the disaster domain. Although this group of studies had used general-purpose taggers, their experiments did not investigate the gains and losses of considering all types of entities against the location entity during training. In this dissertation, we empirically investigate how limiting the training on location entities affects the performance of LMR models. Additionally, existing studies did not consider evaluation under the zero-shot setup. Hence their performance cannot be anticipated for future disaster events, i.e., when deployed during real-world emergencies. Differently, we study the performance of LMR models under zero- and few-shot learning to examine their potential to generalize.

2.4.1.2. Gazetteer-based Approaches

The subsequent research direction was to develop disaster-specific LMR models. The intuitive approach is to verify the potential LMs against a geolocation database (i.e., gazetteers) while detecting them. Many of the existing LMR models are gazetteer-based models with two consecutive components [14], [15], [56], [62], [63]: (i) *extraction*: aims to detect potential LMs from text, and (ii) *retrieval*: aims to link the candidate LMs with toponyms in gazetteers. Hence, these approaches could be categorized under joint approches of recognition and disambiguation, as the resulting LMs correspond to existing toponyms in gazetteers.

Several existing gazetteers were employed in the gazetteer-based approaches, including Geonames [62], [63], OpenStreetMap [14], [15], [63], and National Geospatial Intelligence Agency gazetteer of New Zealand [56], to name a few.

The gazetteer-based approaches extract LMs from the textual content; however, they only qualify LMs after verifying them against a gazetteer. For example, after extracting candidate LMs using OpenCalais NER model, the GeoLocator [56] matches (exact matching) the potential LMs against a gazetteer after correcting misspellings. To alleviate the mismatch issue between the user-generated text (tweets) and gazetteers, GeoLocator expands tweets with abbreviations and acronyms using a C4.5 decision tree classifier. Middleton, Kordopatis-Zilos, Papadopoulos, *et al.* [15] proposed the *map-database* approach that relies on direct matching with the gazetteer. An index of locational phrases is constructed by augmenting different variations of LMs in the OpenStreetMap gazetteer using a set of heuristics. The collected variations are represented as n-grams before being indexed and searched. All combinations of n-grams for the input tweet are issued against the index of the locational phrases.

Another direction for matching gazetteers is utilizing language models. For instance, Middleton, Kordopatis-Zilos, Papadopoulos, *et al.* [15] adopted a language modeling approach, namely the *lm-tags-gazetteer* that was proposed by the top team in the MediaEval 2016 Placing Task [99] after extending it for location entities. The extended system uses gazetteers and a large geo-tagged social media dataset (Flicker posts) to build a language model. Then, the language model is computed regionally to account for the location indicative terms.

Similarly, Al-Olimat, Thirunarayan, Shalin, *et al.* [14] proposed an unsupervised statistical approach to construct regional language models. The tagger identifies the LMs

by simultaneously traversing a tree of n-grams and matching them against a pre-built region-specific gazetteer. Alternatively, Dutt, Hiware, Ghosh, *et al.* [63] applied syntactical heuristics to identify candidate LMs before matching them against the gazetteer. After identifying the nouns in a text using a POS tagger, they consider the candidate nouns followed by common suffixes as LMs. Then, a suffix list is pre-compiled using different naming conventions of locations (e.g., streets and cities). To tokenize text, they also constructed a prefix list of prepositions (e.g., at) and directions (e.g., north).

Acquiring labeled data is a key bottleneck during emergencies. The elegance of gazetteer-based solutions lies in their essence being unsupervised approaches that enable them to evade the need for acquiring annotated data at all. Additionally, the gazetteer-based approaches do achieve high precision levels. However, albeit being training-free models and highly accurate, they have two main drawbacks. First, the noisiness of Twitter streams causes a mismatch between the textual content of tweets and gazetteers, which introduces two challenges: (i) the need for careful text preprocessing (e.g., spell checking), and (ii) the need for augmenting all variations of the LMs including their abbreviations and acronyms into gazetteers for more effective matching. As an alternative, these steps could be replaced by semantic text representation models. Second, the incompleteness of gazetteers affects the performance of LMR models when detecting correct LMs that do not exist in the used gazetteer. To alleviate these challenges, we propose using the pre-trained BERT_{LMR} model which learns semantic and contextual text features without relying on gazetteers.

2.4.1.3. Learning-based Approaches

The machine learning-based (ML-based) LMR models alleviate the limitations of the gazetteer-based LMR solutions. The virtue of ML-based approaches lies in their promising ability to generalize beyond the seen data if supported with good features. Before discussing the ML-based models, we first discuss the different features used to train these models. Note that most are token-level features (except one type of geographical features, and temporal features), as the LMR task is defined here as a sequential token-level tagging/classification task.

Textual features: While words are the essential component of tweets, they constitute the basic feature in all the ML-based approaches. Tweets are typically represented as a bag of words. *N-grams* are employed at both word- [100]–[102] and character- [72] levels.

Lexical features: Following the NER tools, which are trained on formal documents and heavily rely on *capitalization*, different features of capitalization were used for the LMR models [101], [102]. For instance, binary features for whether all characters are uppercased, all characters are lowercased, only the first character is uppercased, and mixed capitalization. Also, the prior probabilities for (i) having the first character capitalized, and (ii) having all characters capitalized. While capitalization shows a strong signal for entities, Twitter users do typically ignore capitalization. Thus, applying letter case correction before recognition is essential to use the NER models. Furthermore, Li and Sun [101] used an indicator feature for whether a token is numeric or alphanumeric.

Contextual features: The bi-directional context of tokens is usually considered in the Conditional Random Fields (CRFs) classifiers to capture the boundaries of locations by adding the adjacent words within a window of a maximum size of 2 [100], [102] or

5 [101]. Furthermore, *word embeddings* have been used for the LMR task [18], [64], [72] to contextually represent tokens/tweets using pre-trained models such as GloVe [103] and BERT [17].

Syntactic features: Assigning the *Part-Of-Speech tags* (POS), e.g., noun, verb, adjective, among other types, to words (or tokens) showed to be effective when combined with other features [100]–[102].

Geographical features: To trade-off between precision and recall, a few studies [72], [101], [102] pre-labeled the tweet tokens using a toponym inventory. This approach is proven to be influential on the LMR performance. To build the toponym inventory, Li and Sun [101] labeled the common names of POIs mentioned in tweets with the associated Foursquare check-ins. This method generates a Twitter-like noisy gazetteer. Alternatively, Han, Yepes, MacKinlay, *et al.* [102] combined the GeoNames gazetteer with a manually-crafted list of location abbreviations and codes to account for the incompleteness and mismatch challenges. They leveraged the ConceptMapper [104] to link locations extracted from gazetteers and use them for tweet tokens representation. Xu, Pei, Li, *et al.* [72] assigned the BIO-like LMR pre-labels predicted using a CRFs model to represent the tweet tokens alongside their distributed word and character representations. Crafting this type of feature is similar to the phase of *extracting* LMs candidates from gazetteers in the gazetteer-based approaches (refer to Section 2.4.1.2).

Entity features: This type of feature is more general than the *Geographical features* in which the LMR models are fed with the NER tags extracted by NER models and their confidence scores [65]. In addition to that, some Entity features are extracted from knowledge bases such as the DBpedia ontology class of the entity, the number of classes and superclasses of the entity, the node degree of the entity, the length of the

corresponding Wikipedia article in characters, among other features [65].

Temporal features: These features aim to capture common words based on chronology. For example, Li and Sun [101] manually compiled a time-trend list of 36 common English verbs, auxiliary verbs, adjectives, or adverbs with scores of 1, 0, and -1 representing the future-, present-, and past-trends, respectively. These scores are used to compute the time-trend score per tweet by averaging the scores of the tweet tokens that appear in the time-trend list.

To this end, we discuss the proposed ML-based approaches that exploit the different categories of features. We note that the Stanford NER tool employs a linear chain CRF model in its core [100]; hence all studies that leverage it are considered ML-based as long as a gazetteer verification does not interrogate their output. The first employment of LM-based approaches for the LMR task is dated back to 2014 when the PETAR system was introduced [101], [105]. PETAR uses a linear-chain Conditional Random Fields (CRF) model and is trained over features from all feature categories above. To overcome the mismatch issue in informal abbreviations and misspellings, Li and Sun [101] applied the Brown clustering technique [106] that groups tokens appearing in similar contexts. Concomitantly, Han, Yepes, MacKinlay, *et al.* [102] leveraged various lexical, semantic, syntactic, and geographical features on top of a CRF classifier [87].

Although CRF models accompanied by noisy gazetteers and a variety of hand-engineered features have achieved competitive performance in different studies [87], [101], [102], [105], their main limitation is the expensive feature engineering phase. There is still a big room for improvement to build robust LMR models with a minimal cost, which we explore through training BERT$_{LMR}$ model (to reduce the training time) with different combinations of available data (to eliminate the data annotation time).

2.4.1.4. Riding the Wave of Deep Learning Models

More recently, a few studies proposed Deep Learning (DL) approaches. The first exploitation of DL approaches for LMR in the disaster domain was proposed in 2019 by Kumar and Singh [64]. They trained a Convolutional Neural Network (CNN) model to learn tweets representation and perform the recognition. Xu, Pei, Li, *et al.* [72] proposed DLocRL, a deep learning sequential pipeline for location recognition and disambiguation in tweets. For recognition, they employed the BiLSTM-CRF model with contextualized word- and character-level features concatenated with their geographical CRF pre-labels. Following the same approach, Wang et al. [40] proposed the NeuroTPR model that employs BiLSTM-CRF NER model [107] for recognition.

Furthermore, Wang and Hu [68] employed Toponym Resolution systems that exploit neural-based BiLSTM recognition modules. The *DM_NLP* [69] model learns character- and word-level text features to represent documents. The learned representations go through a CRF layer added to the BiLSTM model to generate the final LMR token-level labels. The model uses other syntactic features such as POS, StanfordNER, and chunking labels. It also uses contextual features that improve the performance of the model. Differently, *UniMelb* [70] is a BiLSTM-based model that learns word-level text representation and employs a self-attention mechanism with a binary softmax layer on top of it. The *UArizona* [71] goes to the extreme and learns from concatenated word, character, and affix-level representations of text data. Akin to *DM_NLP*, a CRF layer is added on top of the BiLSTM model. nLORE [108] is a deep learning-based model that exploits LORE's [109] rule-based features for recognition.

Diverging from others, Hu, Al-Olimat, Kersten, *et al.* [110] introduced the GazPNE recognizer, which is an unsupervised model that fuses CNN and Bi-LSTM

models to learn from around 4.6M positive training examples extracted from gazetteers and 220M negative synthesized examples. The advantage of GazPNE is that it does not require labeled examples for training. GazPNE2 [94] is an enhanced version of GazPNE that employs an LMD module to improve the LMR accuracy. In addition, it uses synthesized training data extracted from gazetteers to train a CLSTM model.

Recently, Khanal, Traskowsky, and Caragea [111] and Khanal and Caragea [112] exploited the transform-based pre-trained models for the LMR task. Khanal, Traskowsky, and Caragea [111] further pre-trained LUKE [113] model on their data to learn contextualized entity embeddings that allow optimizing a self-attention mechanism for recognizing LMs. Khanal and Caragea [112] investigated multi-task learning for different crisis management computational tasks, including LMR, key-phrase identification, eyewitness identification, and humanitarian categories classification. They empirically confirmed the positive impact of multi-task learning on LMR performance.

The main disadvantage of the DL approaches is being data-hungry. To alleviate this issue, we followed three directions. First, we report the first results for employing the pre-trained BERT to reduce the amount of training data; we introduce BERT$_{LMR}$ [18]. Second, we explore different transfer learning setups that account for the difference between the sources and target disaster events for different factors, including the data domain, entity type, disaster domain, geo-proximity, and language. We attempt to find the best non-target training data at the onset of disaster events. Third, we build the largest-scale manually- and automatically-labeled LMR datasets for both English and Arabic languages. We anticipate these invaluable resources to empower research on LMR.

2.4.2. Evaluation

Unfortunately, the lack of a unified evaluation framework prevented comparing the methods discussed in Section 2.4 under fair conditions. Long-delayed, the first effort to provide a unified framework for the LMP task was the EUPEG framework [114] in 2019. EUPEG provides access to 5 non-disaster-specific LMP models and eight general-purpose datasets. Unfortunately, only one is a Twitter dataset, GeoCorpora [115]. Surprisingly, solely Wang, Hu, and Joseph [40] used the framework.

A fair evaluation framework for the LMR task in the disaster domain has to provide diverse evaluation datasets, evaluation measures, and a set of representative baselines. This section compares the available evaluation datasets and the commonly-used LMR evaluation measures and setups.

2.4.2.1. Datasets

In this section, we review the Twitter NER datasets, the general LMR datasets, and the disaster-specific LMR datasets. First, we present their characteristics and issues and discuss how our LMR datasets, IDRISI-R, overcome these limitations. Then, in Tables 2.1 and 2.2, we summarize the existing NER and LMR English and Arabic datasets, respectively.

Twitter NER Datasets: As LMR is a subtask of NER by definition, different studies explored the effectiveness of the general off-the-shelf NER tools for LMR or retrained their LMR models using NER datasets [19], [40], [55], [80]. On the other hand, the Arabic NER studies have a limited focus on Twitter. In this section, we review the English and Arabic NER datasets.

- ***English Datasets***: Out of the six Twitter NER datasets presented in Table 2.1,

Table 2.1. Summary of the NER and LMP English datasets. "Type" and "Pblc" columns indicate whether the dataset contains location types annotations and whether it is public, respectively. "*" indicates the disaster-related datasets, entirely or partially. "$^{+D}$" indicates LMD datasets.

Dataset	# Twt	# LM (unique)	Annotation	Type	Pblc
Twitter NER datasets					
Ritter et al. [92]	2,400	276 (193)	In-house	×	✓
Liu et al. [116]	12,245	-	In-house	×	×
Li et al.[117]	7,750	-	In-house	×	×
Gelernter et al. [118] *	4,488	2,866 (-)	Translation	×	×
WNUT2017 [119] *	2,296	773 (559)	In-house	×	✓
BTC [120] *	9,551	3,114 (1,295)	In- & Crowd	×	✓
General Twitter LMP datasets					
Sultanik et al. [85]	500	99 (-)	In-house	-	×
Zhang et al. [66]	956	1,393 (779)	In-house	-	×
Inkpen et al. [121]	6,000	4,369 (-)	In-house	-	×
Ji et al. [89] $^{+D}$	3,611	1,542 (-)	In-house	-	×
Li et al.[101], [105]	3,570	2,056 (906)	Automatic	-	×
Kumar et al. [64]	5,107	3,230 (-)	In-house	-	×
Chen et al. [122]	6,571	2,604 (-)	In-house	✓	✓
Disaster-specific Twitter LMP datasets					
GEL [56] *	3,987	-	In-house	✓	×
MID [123] *	3,996	2,030 (451)	In-house	×	✓
ALTA [87] *	3,003	4,854 (1,704)	Crowd	×	✓
OLM [14] *	4,500	5,323 (1,619)	In-house	×	✓
DUT [63] *	1,000	∼100 (-)	In-house	×	×
GeoCorpora [115] * $^{+D}$	6,648	3,100 (1,119)	Crowd	×	✓
HU1 [124] *	1,000	2,139 (989)	In-house	✓	✓
HU3 [125] *	3,000	3,530 (1,351)	In-house	×	✓
FGLOCTweet [126] *	9,435	5,958 (3,457)	Automatic	×	✓
KHAN [111] *	9,339	9,655 (1,639)	Crowd	✓	✓
Bahnasy et al. [127] *	297,150	-	Automatic	×	×
IDRISI					
IDRISI-RE$_{gold}$ *	20,514	21,879 (3,830)	Crowd	✓	✓
IDRISI-RE$_{silver}$ *	56,682	43,404 (2,675)	Automatic	✓	✓

only three are made public [82], [119], [120]. We note here that there is a burgeoning literature on NER and available datasets, but we solely list the ones used in the LMR research. The main drawback of the Ritter, Clark, Etzioni, *et al.* [92] and WNUT2017 [119] datasets is the modest number of *Location*

Table 2.2. Summary of the NER and LMR Arabic datasets. "Type" and "Pblc" columns indicate whether the dataset contains location types annotations and whether it is public, respectively. "*" indicates the disaster-related datasets, entirely or partially.

Dataset	# Twt	# LM (unique)	Annotation	Type	Pblc
Twitter NER datasets					
Darwish [128]	5,069	1,300 (299)	In-house	×	✓
Aguilar et al. [129]	11,155	1,412 (440)	In-house	×	✓
Jarrar et al. [130]	5,653	435 (175)	In-house	×	✓
Twitter LMR datasets					
Al Emadi et al. [73]	-	-	In-house	×	×
Alkouz et al. [131]	100	-	In-house	×	×
Alkouz et al. [74]	-	-	In-house	×	×
Bahnasy et al. [127] *	297,150	-	Automatic	×	×
IDRISI					
IDRISI-RA$_{gold}$ *	4,593	5,236 (918)	In-house	✓	✓
IDRISI-RA$_{silver}$ *	1,205,373	884,217 (18,609)	Automatic	✓	✓

entities. The Broad Twitter Corpus (BTC) [120] that constitutes the largest public Twitter NER dataset offers roughly 2,852 *Location* entities. Nevertheless, our experiments demonstrated that the disaster-specific datasets are preferable over the general-purpose data, e.g., BTC, for training LMR models in the disaster domain [18].

- ***Arabic Datasets***: Table 2.2 presents the Arabic NER dataset (refer to the first group). Although the Arabic NER datasets could be sufficient for training acceptable LMR models at the onset of disaster events, several challenges are associated with this line of research. Even though the NER models trained on web data perform poorly on tweets [132], the Arabic NER studies have a limited focus on Twitter. While this requires creating data domain-specific datasets, a few public Arabic Twitter NER datasets suffer from the limited size and inadequate coverage [128]–[130]. Darwish and Gao [132] introduced the first Arabic Twitter NER dataset containing 5,069 tweets and 1,300 LMs, 299 of which are unique. Aguilar,

AlGhamdi, Soto, *et al.* [129] created a dataset comprising 12,334 tweets and 5,306 LMs. However, only the training and development sets are public for the research community. As of Jan 2023, we managed to crawl 11,155 tweets containing 1,412 LMs, 440 of which are unique. Recently, Jarrar, Khalilia, and Ghanem [130] created the first Arabic Multi-domain nested NER dataset. It contains a subset of 5,653 tweets containing 435 entities of LOC and GPE, only 175 of which are unique. Additionally, a significant challenge in processing Arabic documents is handling the common dialectical (colloquial) text over Twitter. Although the MSA-EGY [129] dataset is of reasonable size, it is limited to only MSA and Egyptian dialect, which could make it geographically confined. Other datasets do not report the dialectical distributions. Furthermore, the Arabic NER datasets are randomly filtered using the sampling Twitter API; they are not disaster-specific. These datasets could serve at the onset of disaster events for deploying acceptable LMR models but should be augmented with disaster-specific Twitter data to improve accuracy [19]. We aim to address these limitations while constructing IDRISI-RA by being an event-centric dataset that geographically covers all Arab countries and reasonably represents their dialects.

General LMR Twitter Datasets:

- ***English Datasets***: The LMR problem is of interest to many domains such as emergency management (refer to Section 2.4.1), traffic monitoring [74], [133], [134], POIs recommendation [76], [77], geographical text analysis and retrieval [78], [79], to name a few. There are a few existing general English LMR tweet datasets (refer to the second group of datasets in Table 2.1). These datasets are not event-centric; however, they are useful to evaluate the generalizability of the LMR

models to domains other than the disaster domain. Although these datasets are of appropriate sizes and mostly labeled by experts, only one is publicly-available (i.e., [122]) due to the issue of third-party copyrights.

- ***Arabic Datasets***: Table 2.2 presents the LMR Arabic dataset (refer to the second group). Alkouz and Al Aghbari [131] adopted an English LMR system [84] to extract LMs from English and Arabic traffic-related tweets (filtered using traffic keywords such as "traffic" and "jam"). The system issues the n-grams extracted from the tweet text against Google Place API and assigns the latitude and longitude coordinates to n-grams that obtained results from the API. The resultant data, however, is geographically limited to United Arab Emirates (UAE). There are a couple of other cross-lingual traffic monitoring systems (for English and Arabic languages) [73], [74], however, the datasets are nonpublic and very small. They contain around 500-600 tweets, and the percentage of Arabic data is undetermined.

Disaster-Specific LMR Twitter Datasets: Despite the abundance of English and Arabic disaster datasets that are made available by academic researchers, e.g., [135]–[139], a few of them are labeled for the LMR task. In this section, we review both disaster-specific Twitter Recognition and Disambiguation (LMD) datasets as the latter could be used to develop and evaluate LMR models. We further compare our IDRISI-R dataset to the existing ones.

- ***English Datasets***: Generally, the existing datasets (refer to the third group in Table 2.1) are limited in size, with the largest being the KHAN dataset which constitutes 9,339 tweets and 9,655 LMs [115]. While the size of the datasets forms a challenge for the supervised models, these 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), three hurricane events happened in the US (ALTA and KHAN), two earthquake events happened in New Zealand (MID) and Nepal (KHAN), one bombing event happened in Sri Lanka (KHAN), and one COVID-19 dataset (KHAN). The GeoCorpora dataset is collected using general disaster keywords such as "earthquake", "flood", "fires", among others, with the most coverage of toponyms for the United States (42%), the United Kingdom (12%), among other countries. Albeit the good diversity, the small number of data per disaster type forms the main barrier to exploiting these datasets.

As for the location types, not all datasets contain location types for LMs [56], [111], [122], [123]. Chen's dataset [122] contains very generic types (point and area) or particular location types (road and river). GEL [56] contains 4,000 tweets collected during the 2011 Christchurch Earthquake in New Zealand. The four categories of LMs are street, building, toponym, and abbreviation. However, the GEL dataset is not available to the research community. MID dataset [123] contains three types of locations including "admin", "building", and "transport". Similarly, KHAN dataset [111] has categories of locations that further requires finer annotations to split types out. For example, all fine-grained locations (e.g., buildings, landmarks) are labeled into one category called "lan". We augment annotations for different coarse-grained and fine-grained location types in IDRISI-RE to overcome this shortcoming.

Additionally, new LMs within the affected areas emerge in the Twitter stream during disasters, demanding extended data coverage during the entire disaster

period. Nevertheless, the OLM dataset is the only dataset that we could analyze its temporal coverage as other datasets either do not release the IDs (MID), or the event notion is ignored when they are collected (GeoCorpora is a keyword-based dataset) or released (all events are merged in ALTA and KHAN datasets). Using the tweets we managed to crawl at the time of this writing, we found that the OLM dataset misses long critical periods, especially during the Chennai Floods 2015. To articulate, while the floods happened between 8 Nov - 14 Dec 2015,[17] the tweets only cover the period between 2-4 Dec 2015.

Moreover, while an LMR dataset could cover all relevant topics discussed during the disaster, it has to contain informative and actionable tweets useful for the responders. Unfortunately, only the ALTA dataset is labeled for relevance, and the GEL nonpublic dataset is labeled for informativeness. We aim to select events that are already filtered for relevance and contain informative tweets when constructing IDRISI-R datasets. The main imperfection of the LMR datasets is the inconsistency of the "location mentions" definition between and within datasets. Indeed, the guidelines used to train annotators are rarely discussed [66], [115]. Therefore, we release our annotation task instructions that articulate our "location mention" definition, and we further elaborate on them in Section 4.2.2.1.

- ***Arabic Datasets***: Table 2.2 presents the only disaster-specific Arabic LMR dataset. Bahnasy, El-Mahdy, *et al.* [127] employed LM extraction to aid event detection over Arabic tweets. Although the dataset is large, it is confined from different angles; its geographical coverage is limited to Egypt; its dialectical coverage is limited to Egyptian dialect; and its disaster domain is limited to fire, flood, and

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

pandemic disaster types only. In contrast, IDRISI-R dataset contains diverse disaster types that form the most happening types in Arabic-speaking countries. These events happened in 22 different countries. IDRISI-RA also captures a fair coverage of Arabic dialects.

2.4.2.2. Evaluation Measures

The standard evaluation measures for LMR effectiveness are Accuracy (Acc), Precision (P), Recall (R), and F_β score (the harmonic mean of Precision and Recall), per entity. Nevertheless, researchers compute these measures in different ways based on three main factors: (1) handling the *partial matches*, i.e., whether to reward the model when detecting part of the LM span, (2) evaluating *per tweet or event*, i.e., whether to report token-level macro or micro average performance, and (3) handling the *true negatives*, i.e., whether to reward the models when they correctly predict no LMs for a single tweet. The partial matches are typically ignored from evaluation except in a few studies, e.g., [14], [87], [110]. Molla et al. [87] report that they account for partial matches but do not elaborate on their strategy nor make their evaluation script public. Al-Olimat et al. [14] and Hu et al. [110] penalize models by adding 0.5 to both false positives and false negatives counts before computing the Precision and Recall. Our evaluation script accounts for factors (2) and (3) but not factor (1) as it requires further investigation that we keep for the future.

Given the scarcity of publicly-available large-scale representative LMR datasets and the critical real-time nature of the deployment of LMR models at the onset of disaster events, LMR systems should be extensively evaluated in different possible scenarios. Typically, the existing models are evaluated under an unrealistic assumption that labeled

target data is available during the early stages of disaster events. As acquiring labeled data is costly during disasters, researchers should report the performance of their models under the zero-shot setup, where the model had never been introduced to training data from the target disaster. Therefore, studying the models' generalizability under zero-shot learning is important while considering different factors such as disaster domain, geographical proximity, and language. In this dissertation, we investigate LMR model performance under these different scenarios, we explored the effect of data domain, entity type, disaster domain, geographical proximity, and language under the zero-shot setup using five disaster-specific datasets [14][15]. Our rigorous experiments suggest that "target" evaluation setups show misleadingly high performance compared to cross- and out-of-domain scenarios during emergencies.

2.5. Disaster-Specific Location Mention Disambiguation

In this section, we discuss the LMD studies and discuss their technical solutions (Section 2.5.1) and evaluation tools (Section 2.5.2).

2.5.1. Solutions

This section reviews the current solutions from the methodology perspective due to the absence of a unified evaluation framework that allows performance comparison across proposed solutions. A few studies tackle the LMD task using limited machine learning and deep learning approaches. Thus, we present the hand-crafted and automatically-computed features employed before discussing the existing approaches. Akin to the LMR systems, existing approaches exploit contextual features at character and word levels (e.g., learned by CNNs), syntactic (e.g., POS tags), and geographical

features (extracted from gazetteer attributes) for tackling the LMD task. In addition to a few other features we discuss below:

Contextual features: In addition to the features discussed in Section 2.4, the gazetteer entries lack context to learn rich representations. Thus, external resources, such as Wikipedia articles, are useful for expanding their representations. Articles could be used as a whole or only extracted segments for efficiency concerns.

Similarity features: These features are computed by the similarity of the candidate LMs extracted from tweets against the toponym names in gazetteers. The similarity is encoded using exact matching, substring matching (i.e., candidate LM partially matches a gazetteer toponym or the opposite), prefix matching (LM matching the beginning of a gazetteer toponym or the opposite), Jaccard similarity, or Levenshtein similarity.

Gazetteer features: The properties of toponyms in gazetteers are employed to capture relevance signals and prioritize gazetteer candidates. These features change according to the employed gazetteer. In addition, different attributes are helpful, including the popularity of the toponym, the number of ancestor LMs, and the administrative division level, among others.

Mention neighbors features: The LMs often co-occur with their child (fine-grained) or parent (coarse-grained) toponyms as illustrated in Figure 2.1. i.e., addresses. upon this observation, the mention neighbors features improved the LMD performance. Using all LMs in the tweet besides the target LM, we can encode the relationship between the multi-level mention lists by examining whether the gazetteer (i) toponym, (ii) the ancestor, or (iii) alternate names exist in the mention lists of the LM. The collective disambiguation considers the co-occurring LMs as features (refer to Section 2.5.1.3).

There are three commonly used English digital gazetteers, namely:

Figure 2.1. Example user profile during floods in Chennai.

- Geonames: Geographical database covers all countries and contains over 11M unique places (with 25 million different geographical names) available for download. The locations are categorized into nine main types (e.g., country, parks, village, and road) and subcategorized into 645 feature codes.
- OpenStreetMap: An international street-level gazetteer constructed by a community of mappers. They continuously add and maintain data about streets, trails, and POIs worldwide.
- Foursquare: A database with more than 105M placenames collected using a collective crowdsourcing method. Specifically, the platform logs the users' check-ins from social media platforms such as Twitter and Instagram.

2.5.1.1. Learning-based Models

Middleton et al. [15] trained an SVM model on gazetteer-based features, including location type, population, and alternative names. Following the same line, the disambiguation models of the toponym resolution system employed by Wang and Hu [68] are essentially machine learning models: (i) *DM_NLP* [69] is a Light Gradient Boosting Machine (LightGBM) model trained on similarity scores, contextual representations, gazetteer attributes, and mention list features. (ii) *UniMelb* [70] is a Support Vector Machine (SVM) that uses different feature types such as the history results in the training dataset, population, gazetteer attributes, similarity, and mention neighbors features. (iii) *UArizona* [71] is a heuristic-based system that selects the toponym with the highest population in gazetteers.

2.5.1.2. Deep Learning Models

Xu et al. [72] proposed a novel attention-based two-pairs of bi-LSTMs for matching location mentions against the Foursquare gazetteer. The Foursquare gazetteer constitutes a collection of location profiles. Each location profile (*lp*) contains several attributes, including title, category, address, tips, tips' counts, visitors count, visits count, and rating. To process the *lps*, the researchers represent the category attribute in a one-hot vector, the word-based attributes are represented by averaging their TF-IDF vector representations, and the numeric-based attributes are normalized using the global gazetteer counts. Finally, all these representations are concatenated before being fed to the disambiguation model. On the other hand, tweet-LM pairs are represented using their text contextual representation; text contextual representation attended to the *lp* representation and geographical distance. The two-pair networks learn the left (starting from the first token in the tweet and ending at the end of the LM) and right (starting from the LM and ending at the last token of the tweet) contexts of the LM. The geographical distance is measured using the Manhattan distance between the geo-coordinates of the user location, if available, and every *lp*. Both representations then go through a fully-connected layer to learn disambiguation.

2.5.1.3. Collective Disambiguation

To disambiguate location mentions in the same tweet, such as "Kuwait" and "Ooredoo," one might expect that "Ooredoo" is a branch in "Kuwait" not in "Qatar" (the original headquarter). The general approaches differ in whether to resolve entities in isolation, using a pair-wise strategy, or collectively. Inspired by the pair-wise methods, Zhang et al. [66] consider the hierarchy of the location mentions in tweets when resolving

them. Recently, Xu et al. [72] collectively disambiguated all LMs in a tweet by adopting the Pair Linking algorithm [67] that improved the disambiguation performance.

2.5.2. Evaluation

Evaluating LMD systems requires ground truth data of tweets containing the LMs and their corresponding toponyms from gazetteers. A non-gazetteer-specific dataset would contain the full addresses of the LMs allowing the LMD systems to freely chose their geo-positioning database. Next, we discuss the LMD datasets and evaluation measures in the literature.

2.5.2.1. Datasets

There needs to be more Twitter disaster-specific LMD datasets. Table 2.1 presents only two LMD datasets (marked by "$^{+D}$") and their statistics. Only GeoCorpora [115] is available for the research community. Wang and Hu [68] evaluated it using eight different datasets available through EUPEG framework [114], solely one of which is a tweet dataset that is GeoCorpora. Xu et al. [72] used the dataset released by Ji et al. [89], but it is not public for the research community.

2.5.2.2. Evaluation Measures

Similar to the LMR task, researchers evaluate the LMD systems using Precision (P), Recall (R), and the F_β score. Karimzadeh [140] proposed more evaluation measures that overcome the shortcoming of P, R, and F_β scores. Moreover, distance-based methods are also used in non-disaster-specific studies to evaluate LMD systems in which the distance between the GPS coordinates of the gold and predicted LMs is computed using great circle distance. The Median and Mean Distance Errors then measure the system's

overall performance. Other discrete measures such as Acc, P, R, and F_β can evaluate the predictions within a distance d that is commonly set to 161 KM (100 miles). To articulate, Acc@d is the fraction of correctly predicted LMs within d.

CHAPTER 3: LOCATION MENTION RECOGNITION

The LMR task is generally defined as *the automatic extraction of toponyms from text*. The scope of this chapter is limited from two angles; the extraction is focused on the textual content of *tweets*, and more specifically *disaster-related* tweets that are posted *during disaster events*.

Two main factors that influence the robustness of an LMR system are: (i) the learning model, and (ii) the dataset used to train the classifier. As for the learning model, there are two well-established approaches. The first is adopting existing general-purpose Named Entity Recognition (NER) taggers. NER is the generalized task of LMR, by definition, which aims to extract entity mentions in a given text. However, the general-purpose NER systems do not effectively extract toponyms from Twitter messages because tweets often contain informal language, misspellings, grammar mistakes, shortened words, and slang [16]. Moreover, entities mentioned in tweets may have inconsistent capitalization, which is one of the main features that standard NER systems rely on [37]. The second common approach is employing gazetteers to maintain highly-precise location mentions recognizers. However, the gazetteer-based models are restricted to the geographical coverage of their databases. Additionally, the noisiness of Twitter stream contracts the gazetteers' formal nature, causing the so-called mismatch problem. Recently, several deep learning approaches were proposed. However, the general practice for these solutions is to train the models using data from the target disaster event, which is usually scarce or hard to obtain at the onset of disaster events.

We employ the BERT model to address the challenges mentioned above, which achieved state-of-the-art results in many NLP tasks with little data [17]. Moreover, it eliminates the cost of hand-crafting features, allowing us to overcome the limita-

tions of gazetteer-based and traditional learning models that highly depend on feature engineering.

While most existing studies focus on learning algorithms and assume sufficient training data is available, we explore how the choice of a training dataset influences the performance of an LMR system in the domain of humanitarian crises where the cost and time of acquiring training data should be minimized. This exploration, thus, contributes to the effectiveness and efficiency of deploying the LMR models in emergencies. We run our experiments in two settings related to training data augmentation strategies: (i) zero-shot setting where there is no available training data, and (ii) few-shot setting where limited training examples, in order of hundreds, are available. In the zero-shot setting, we investigate the effect of multiple factors on the LMR model during training, including the *data domain*, *entity types*, *disaster domain*, *geo-proximity*, and *language*. In the few-shot setting, we investigate the performance of multilingual models when trained with limited labeled data from the target language. We also seek to determine the performance gains of our best LMR model when incrementally adding labeled data from the target event in a monolingual setting. This enables us to learn the minimal cost of acquiring data from just-occurred disasters.

Considering all these diverse settings, we formulate our research questions as follows:

- **RQ1**: How effective is the LMR system when trained on the *web-based general-purpose* NER datasets with all types of entities, including location (LOC), organization (ORG), person (PER), and miscellaneous (MISC), versus *Twitter general-purpose* datasets?
- **RQ2**: How effective are the general-purpose web-based datasets compared with

general-purpose Twitter datasets when using *only LOC entities* (i.e., without ORG and PER)?

- **RQ3**: Does training on *crisis-related Twitter* datasets improve the performance of the LMR system compared to the *general-purpose Twitter* datasets?
- **RQ4**: Does training on combined data from different types of crisis events yield better performance than training on data from the same type of events?
- **RQ5**: Does the geospatial proximity of training events to the target event affect the performance?
- **RQ6**: Can a model trained on one language be used to recognize location mentions from another language?
- **RQ7**: How many target event tweets are required to train a reasonably performing (e.g., F_1 >= 0.70) LMR model?

The research on the LMR task is currently lacking answers to these questions. In this work, we perform extensive experiments to provide answers to them empirically. We fix our learning model to a pre-trained model (i.e., BERT$_{LMR}$) and use a variety of datasets, i.e., web-based general-purpose, Twitter general-purpose, and Twitter crisis-specific. Our results suggest that the general-purpose datasets are not the best for LMR in crisis tweets. Moreover, the types of entities (e.g., person or organization) used to train a model make a difference. Specifically, training using only location entities performs better than training on all entity types. Furthermore, while Twitter datasets are preferred over general-purpose datasets, we observe that Twitter crisis-related datasets help achieve better performance. More interestingly, we found that training on past disasters of the same type as the target disaster generally improves performance. While labeled data from the target event yield the best performance, we note that labeling data

from nearby disasters is helpful when the target labeled data is unavailable. Additionally, training on little labeled data, around 263-356 tweets, from the target language notably improves the performance when combined with all available multilingual data. Finally, we suggest training on all available data from all domains to minimize the labeling cost at the onset of disaster events. Labeling around 500 tweets would generally be sufficient to obtain an acceptable LMR model.

The contributions of this paper are as follows:

- We tackle the bottleneck of lack of annotated data, drawbacks of gazetteer-based solutions, and the expense of hand-engineered features by exploiting a BERT-based LMR model, BERT$_{LMR}$.
- We explore different data transfer setups, including data domain, types of entities, disaster domain, geo-proximity, and language. We further suggest the best option for each aspect at the onset of disaster events.
- We study the cost of incrementally acquiring labeled data at the onset of disaster events for training reasonably performing LMR model.
- We conduct failure analysis on BERT$_{LMR}$ to gain insights for the future development of LMR models.

For reproducibility, we make the NER datasets (in BILOU format), steps to acquire the licensed datasets, the best performing models, and the steps to run BERT$_{LMR}$ model publicly available.[1]

The remainder of this chapter is organized as follows. First, we discuss the LMR problem definition in Section 3.1. Then, we present the LMR learnable model in Section 3.2. Next, we explain the experimental setups in Section 3.3. We then discuss

[1] https://github.com/rsuwaileh/TLLMR4CM.git

the results and conduct failure analysis in Sections 3.4 and 3.5, respectively. We finally discuss the limitation of our empirical study in Section 3.6.

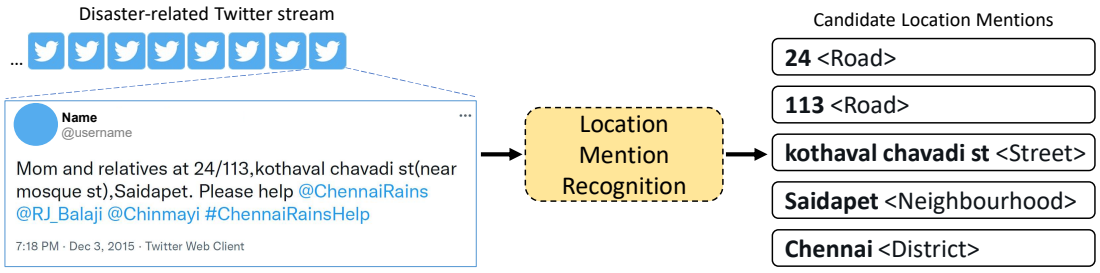


Figure 3.1. High-level overview of the LMR task

3.1. Problem Definition

The LMR task is formally defined as follows: Given a tweet t that is related to a disaster event e, the LMR system aims to identify all location mentions (LMs): $LM_t = \{lm_i; i \in [1, n_t]\}$ in the tweet t, where lm_i is the i^{th} location mention and n_t is the total number of location mentions in t, if any. The location mention may constitute one or more *tokens* in the tweet text.

To distinguish the LMR from other tasks, we emphasize that LMR aims at removing the geo/non-geo ambiguity of tokens in text. It is also known as *location extraction* or *geoparsing* in the literature.

There are two task setups for LMR. The first recognizes toponyms without their types, denoted as "type-less" LMR and the second distinguishes between types of LMs (e.g., country, city, and street), denoted as "type-based" LMD. The latter better serves developing and evaluating geolocation processing systems in light of the responders' needs. Furthermore, 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. Table 1.1 shows a few example tweets shared during different real-world disaster events.

3.2. Transfer Learning for LMR using BERT-based model

Pretrained models, such as BERT, have shown impressive performance in the sequence modeling tasks, including the NER task [17]. In this work, we employ BERT-LARGE-CASED model in all experiments, except for the cross- and multilingual training, we use BERT-BASE-MULTILINGUAL-CASED model. We added a linear classification layer on top of the BERT model and fine-tuned it using the source dataset. We preprocessed the tweets to remove 'RT,' user mentions, non-ASCII characters, and URLs. We also segmented the hashtags using the word segment library,[2] since some location mentions appear as subtokens of hashtags in the ground truth of datasets we use for evaluation. From hereafter, we refer to it as "BERT$_{LMR}$".

We perceive the LMR task as a multi-class sequence tagging task and use the *BILOU* scheme, which we adopt from NER studies, due to its superior performance over the *BIO* scheme [141]–[143]. In the *BILOU* scheme, tokens are assigned positional tags; "B" denotes the beginning token of an LM; "I" denotes a token inside an LM; "L" denotes the last token in an LM and "U" indicates that the LM has only one token, such as "Qatar"; and "O" denotes a token outside any LM.

3.3. Experimental Setup

In this section, we describe the details of our experimental setup. First, we present the datasets in Section 3.3.1 and the experimental configurations in Section 3.3.2. We

[2] http://www.grantjenks.com/docs/wordsegment/

then discuss the hyper-parameters fine-tuning in Section 3.3.3 followed by the evaluation measures in Section 3.3.4.

3.3.1. Datasets

To answer **RQ1.1**-**RQ1.7**, we mainly need three types of datasets: (i) *general-purpose NER dataset*, (ii) *Twitter NER dataset*, and (iii) *crisis-related multilingual Twitter LOC dataset*. Tables 3.1 and 3.2 show various statistics of all the datasets we used in our experiments, described below.

- **General-purpose NER dataset:** A well-known candidate for this category is the CoNLL-2003 NER dataset [98], which comprises newswire text from Reuters, tagged with four different entity types, namely PER, LOC, ORG, and MISC. Overall, the dataset contains 22,137 sentences and 35,089 entities. In addition, we use the standard training development segments for training and tuning hyper-parameters.
- **Twitter NER dataset:** We use the Broad Twitter Corpus (BTC) as our Twitter NER dataset [144]. It consists of 9,515 tweets tagged with three entity types: PER, LOC, and ORG. The dataset broadly covers spatial, temporal, and social aspects. Various segments in the dataset represent different types of data collection and annotation methodologies. For instance, *Segment A* comprises random samples of UK tweets about the “New Year". We randomly sampled 90% of the dataset for training and 10% for development.
- **Crisis-related Twitter LMR dataset:** As this work mainly focuses on developing a robust LMR system for toponym extraction from crisis-related tweets, we use several Twitter datasets from real-world disasters to perform extensive experiments.

Table 3.1. Statistics of the datasets used in our experiments. HRC, EQK, and FLD refer to hurricanes, earthquakes, and floods.

Dataset	Lang	Country	# Docs	# Locs
CoNLL-2003	EN	Global	22,137	10,645
BTC	EN	Global	9,383	2,852
Chennai FLD	EN	IND	1,500	2,226
Houston FLD	EN	US	1,500	1,701
Louisiana FLD	EN	US	1,500	1,396
HRC Sandy	EN	US	1,996	735
ChCh EQK	EN	NZ	1,999	291
Milan Blackout	IT	IT	391	705
Turkish EQK	TR	TR	2,000	442

We use seven datasets in this category; three represent floods, two earthquakes, one hurricane, and one blackout. The floods dataset consists of 4,500 tweets from *Chennai floods 2015*, *Louisiana floods 2016*, and *Houston floods 2016* [14]. The tweets in these datasets are tagged using several location-related tags. In this work, we only use inLOC and outLOC, indicating whether the location is within or outside the disaster-affected areas. We filter out all hashtags used to collect the datasets, thus limiting their effect towards biasing the model's training process. The remaining four datasets in this category are adopted from [145]. This source contains 6,386 multilingual tweets in total. It contains English, Italian, and Turkish tweets from four disasters, namely *Hurricane Sandy 2012*, *Christchurch earthquake 2012*, *Milan blackout 2013*, and *Turkey earthquake 2013*. Hurricane Sandy and the Christchurch earthquake are English datasets, the Milan blackout is in Italian, and the Turkey earthquake is in Turkish.

Table 3.2. BILOU tokens' statistics of the datasets used in our experiments. The numbers in parentheses show the percentage of training data. HRC, EQK, and FLD refer to hurricanes, earthquakes, and floods. For the annotations, "U" denotes a single-token (unit) LM. "B", "I", and "L" denote the beginning, inside, and last tokens of an LM, respectively. "O" denotes a non-location token.

Dataset	B	I	L	U	O
CoNLL-2003	1,041 (69)	116 (70)	1,041 (690)	6,099 (67)	250,660 (68)
BTC	665 (100)	293 (100)	665 (100)	2,187 (100)	168,721 (100)
Chennai FLD	840 (80)	275 (78)	840 (80)	1,386 (80)	22,194 (70)
Houston FLD	508 (81)	155 (84)	508 (81)	1,193 (81)	22,114 (70)
Louisiana FLD	227 (81)	77 (78)	227 (81)	1,169 (81)	24,620 (69)
HRC Sandy	665 (79)	665 (79)	595 (81)	70 (76)	32,525 (70)
ChCh EQK	220 (79)	220 (79)	544 (80)	71 (77)	27,633 (71)
Milan Blackout	114 (71)	27 (78)	114 (71)	591 (69)	6,083 (71)
Turkish EQK	28 (68)	0 (0)	28 (68)	414 (71)	16,081 (68)

3.3.2. Experimental Configurations

We used several training and testing configurations in our experiments. In this section, we define the adopted terminology and discuss the different generic experimental configurations.

We define the "**source dataset**" as the dataset (or the combination of datasets) that we use to *train* our LMR model and the "**target dataset**" as the dataset on which we *test* our LMR model. The source dataset can be of any document type (e.g., web articles or tweets) and any topic type (e.g., general or event-oriented); however, the target dataset is *always* a crisis-related Twitter dataset.

Furthermore, our experiments use different terminology to articulate the *match* between the source and target datasets. First, we use "**domain**" to refer to the domain of the target dataset, which is always of a specific disaster type. We use "**in-domain**" to denote the case when the source and target datasets are of the same disaster type, e.g., a hurricane. We use "**cross-domain**" to denote the case when the source and target

datasets are both disasters *but* of different types (e.g., earthquake vs. flood). Finally, we use "**out-of-domain**" to denote the case when the source dataset is not a disaster dataset (e.g., general tweets or web articles).

Using the above terminology, we define different configurations based on the source and target datasets as follows:

- <*source dataset*>.`ner` denotes the case when we use the NER source dataset with all entity types (e.g., LOC, PER, ORG, and MISC) in the BILOU scheme.
- <*source dataset*>.`loc` denotes when we use the NER source dataset with only the LOC entity and discard all other entity types (e.g., PER, ORG, and MISC). Doing so converts the LOC entity into the BILOU scheme, and the non-LOC entities are labeled as "O".
- `DIS`.`others` denotes when the source dataset includes all English disaster datasets, regardless of the type, except the target dataset. For example, if the target event is *Chennai floods*, then we use the other two flood events (i.e., *Louisiana floods* and *Houston floods*) in addition to the hurricane and earthquake datasets for training.
- `DIS_`<*source_type*>.`others` denotes the case when the target disaster is of a different type than the *source_type*, which (in our experiments) can be either Floods (FLD), Hurricane (HRC), or Earthquake (EQK).
- `DIS_`<*source_area*>.`others` denotes the case when the target disaster happens in a different geographical area than the *source_area*, which (in our experiments) can be either India (IN), United States (US), or New Zealand (NZ).
- "Combined" denotes the case when we use different document types (i.e., web and tweets) in our source dataset. In this case, we use "joint" ("seq") to denote the case when we feed the different types together in one stage (sequentially in two

stages) while training our model.

- <*source dataset*>_%Target denotes when we use a percentage of the target data for training.
- `Cross-lingual_Zero_shot` denotes the case when we train on a disaster dataset in a source language and test on a different target language. For example, we train on English tweets but test on Italian or Turkish tweets.
- `Cross-lingual_Few_shots` denotes the case when we train on a disaster dataset in a source language with a few examples from a different target language. We test on the target language. For example, we train on English tweets combined with a few Italian or Turkish tweets but test on Italian or Turkish tweets.
- `Multilingual_Few_shots` denotes the case when we train on disaster datasets in multiple languages *including* the target language and test on the target language. For example, we train on all available languages, and test on the Italian or Turkish tweets.

3.3.3. Hyper-parameters Tuning

During training, we tuned the hyper-parameters such as batch size, number of training epochs, and the learning rate using configuration values recommended in [17] as the batch size of 16 or 32, number of epochs of 2, 3, or 4, and learning rate of 5E-5, 3E-5, or 2E-5. For every experimental configuration, we search hyper-parameters space using the grid search method on the development set. We further experiment with five different seed initialization values for every combination of hyper-parameters, seeking reliability of results, and eventually use the median F_1 score from the five runs. We finally select the best hyper-parameter combination and report its F_1 score on the test

set. We found that the best hyper-parameters primarily differ from the default settings across different training setups [18]. No one combination of hype-parameters fits all experimental setups (refer to results in Table A.1, Appendix A).

3.3.4. Evaluation Measures

To measure the effectiveness of the LMR model over different setups, we compute Precision (P), Recall (R), and their harmonic mean (F_1 score) for each entity (i.e., location mention) using the *seqeval (v1.2.2)* package,[3] which adopts the evaluation scripts used to evaluate the chunking tasks (e.g., named-entity recognition) in CoNLL-2000 NER shared task [146]. The package evaluates the model's output on the entity-level rather than the token-level.[4] We use the default micro-average metric to account for the class imbalance issue in our datasets (see class distributions in table 3.2).

3.4. Results and Analysis

In this section, we discuss the seven research questions in detail, present the experiments we carried out to answer each of them and analyze their results. First, we explore the usefulness of exploiting "out-of-domain" training data with either multiple entity types such as person and organization alongside the location (Section 3.4.1) or with location entity alone (Section 3.4.2). We further study the performance when training on "in-domain" and "cross-domain" data in Sections 3.4.3 and 3.4.4, respectively. We then study the performance of the LMR models when considering the geographic proximity of disaster events during training (Section 3.4.5). Moreover, we discuss the effectiveness of cross-lingual setup when training on data in a different language than the language

[3] `https://pypi.org/project/seqeval/`

[4] The seqeval package uses all predicted and all gold LMs to compute precision and recall, respectively. Malformed tag sequences are discarded from evaluation.

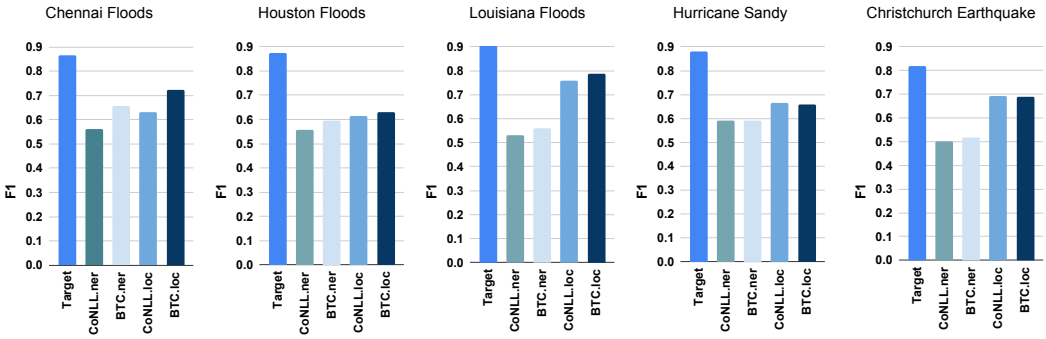


Figure 3.2. The results of exploiting out-of-domain general-purpose datasets for training an LMR model.

of the tweets discussing the target event in Section 3.4.6. We finally explore the gain in performance when incrementally acquiring training data from the target disaster in Section 3.4.7.

3.4.1. General-Purpose (Out-of-Domain) Training with Multiple Entities (RQ1)

Due to the limited location-labeled data, we study the effect of using general-purpose NER datasets to train our LMR model. Since general-purpose NER data is more prominent in size and has location as an entity type, it might be sufficient to train models that effectively recognize toponyms in tweets posted during disasters. This is useful in emergencies when time is critical and acquiring new training data is time-consuming and expensive. The delay in response may negatively affect relief actions.

To this end, we explore the usefulness of the general-purpose NER dataset vs. Twitter NER dataset for the LMR task. We use the following training settings:

- *CoNLL*.`ner`: Using the CoNLL-2003 dataset with all entity types (LOC, PER, ORG, and MISC) for training.
- *BTC*.`ner`: Using the BTC dataset with all entities (LOC, PER, and ORG) for training.

We test our LMR model on each crisis-related Twitter dataset (refer to Section 3.3.1). Figure 3.2 presents the results (the second and third bars from left in all charts). Except for Hurricane Sandy, the *BTC*.`ner` model outperforms the *CoNLL*.`ner` model in all datasets, suggesting that the general-purpose datasets built on documents written in formal language might not be suitable for disaster-related tweets. However, such models' performance is comparable to Hurricane Sandy's case.

To answer **RQ1**, we conclude that Twitter NER datasets are more effective than general-purpose NER datasets for training an LMR model for toponym recognition in disaster-related tweets. While the general-purpose NER models did not outperform the Twitter-based models in any of the setups above, the general-purpose NER datasets are a valuable resource for training an LMR system when no other data is available, e.g., at the onset of a disaster event. Furthermore, they exhibit an acceptable performance ranging between 0.5 and 0.6 F_1 (with even better performance if trained only on LOC entities, see **RQ2** below) given the unavailability of domain data.

3.4.2. General-Purpose (Out-of-Domain) Training with Location Entities (RQ2)

Similar to RQ1, we aim to determine the effectiveness of an LMR model trained on general-purpose (out-of-domain) datasets, but this time *excluding* non-location entities such as PER, ORG, etc.

To this end, we adopt the following training settings:

- *CoNLL*.`loc`: Using the CoNLL-2003 dataset with only the LOC entity.
- *BTC*.`loc`: Using the BTC dataset with only the LOC entity.

According to the results in Figure 3.2 (considering the fourth and fifth bars from left in all charts), training the LMR model using only LOC entity improves the

performance by 5.6-22.7% and 2.6-22.6% across the different disasters for *CoNLL* and *BTC* respectively. However, we noticed that the improvement is evident in precision but not recall (refer to results in Table A.1 in Appendix A), suggesting that focusing the training on locations only improves the precision of recognizing locations with little or no degradation in recall (except for Hurricane Sandy's where degradation reached about 12%). We anticipate the reason to be the distinct patterns of LMs compared to other entities in the data. For instance, as opposed to other types of entities, location mentions are usually attached to their category or surrounded by adpositions such as "near", "at", or "10Km away from". Based on such results, we conclude that the location-specific datasets are better for training the LMR model than the general-purpose NER datasets, which answer **RQ2**.

Although the CoNLL-2003 dataset is 2.4% larger and it contains 3.7 times the number of LMs compared to the BTC dataset, we notice that *BTC*.`loc` model is better than *CoNLL*.`loc` on the three *flood* datasets, but not on Hurricane Sandy and Christchurch Earthquake datasets. Upon investigation, we interestingly found a noticeable overlap between some of the top frequent LMs in those two disaster-related datasets and the CoNLL-2003 dataset, which justifies such an unexpected high performance. For example, one of the top frequent LMs in Hurricane Sandy is "New York", which appears 289 times (208, 24, and 57 times in training, development, and test sets, respectively). On the other hand, the same LM appears 123 times in CoNLL-2003 dataset (100, 23, and 27 in training, development, and test data, respectively). Similarly, the second top LM in Christchurch Earthquake dataset, which is "New Zealand" with frequency = 57 (40, 5, and 12 times in training, development, and test data, respectively), is mentioned 50 times in CoNLL-2003 dataset (41, 9, and 10 in training, development, and test data

respectively). On the contrary, the top 3 most-frequent LMs in Chennai, Houston, and Louisiana floods appear 0, 16 (12, 0, and 4 in training, development, and test subsets, respectively), and 0 times, respectively, in CoNLL-2003 dataset.

3.4.3. Crisis-Related Training (RQ3)

Thus far, we confirmed our need for location-specific data to train the LMR system. However, in contrast to disaster-specific streams, the location mentions in the general streams might appear in different patterns. To clarify, people might use more accurate and complete addresses of locations when reporting incidents happening during emergencies, aiming to help responders take immediate actions (e.g., Tweet #1 in Table 1.2). To investigate further, we train our LMR model using a combination of *BTC*.`loc` dataset (as using it achieved the best F_1 score earlier in most datasets) and the available crisis-related datasets. By this, we aim to address **RQ3**: *Does training on crisis-related Twitter datasets improve the performance of the LMR system compared to the general-purpose Twitter datasets?*

An interesting aspect to explore in this context is the effect of combining the in-, cross-, and out-of-domain data. To address this, we train an LMR model using crisis-related datasets as follows.

- `DIS.others`: Combining all disaster datasets except the target disaster for training.
- *Combined*.`joint`: Combining in-, cross-, and out-of-domain datasets for training. Specifically, we use *BTC*.`loc` and all `DIS.others` for training. All the datasets are merged before training.
- *Combined*.`seq`: Using *BTC*.`loc` and `DIS.others` for training; however, we first train a model using the former and then fine-tune it using the latter.

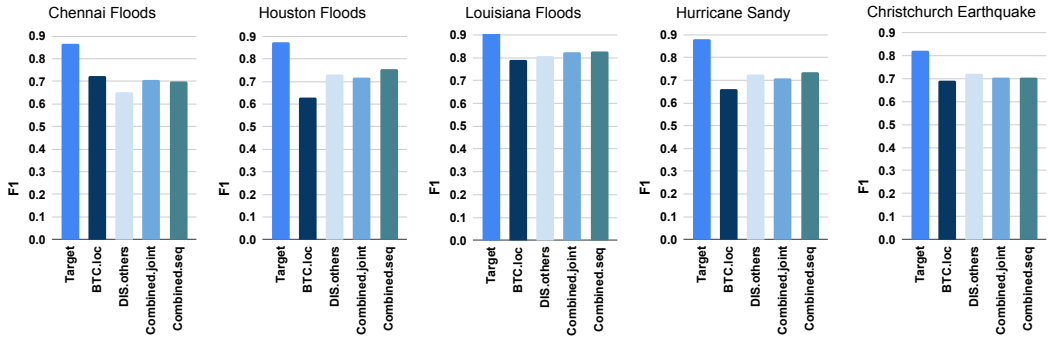


Figure 3.3. The F_1 results of exploiting in- and out-of-domain data for training an LMR model

We show the results of these runs in Figure 3.3. Generally, the results are inconsistent across disasters; hence we cannot draw a clear conclusion on which setup is the best. As references, we compare the results with the case when we train on the target dataset (denoted as Target in Figure 3.3) and with *BTC*.`loc` (as using it mostly achieved the best F_1 score among the non-target setups). Using training data other than the target data shows significant degradation in performance concerning the Target model. This finding emphasizes the importance of providing in-domain (i.e., "target") data for better effectiveness. Additionally, employing only in- and cross-domain data (i.e., `DIS.others`) shows improvement against *BTC*.`loc`, except for the Chennai floods. These results confirm the potential of using in- and cross-domain data for better performance.

Moreover, combining in-, cross-, and out-of-domain training data provide reasonable performance comparable to `DIS.others` for early location extraction when a sudden disaster happens. In the worst scenario, such a reasonable model can be employed to augment labeled data to improve performance over time automatically. This can be achieved by exploiting active learning, and automatic labeling, among other known data augmentation techniques.

Furthermore, the *Combined*.`seq` setup is slightly better than the *Combined*.`joint`

setup by approximately 1.6% on average across all datasets, except for the Chennai floods. This is intuitive since the fine-tuning, exclusively on in- and cross-domain disaster data, should impact the model more than training on combined/mixed data.

To answer RQ3, we conclude that disaster-related training data helped improve the LMR model by 5.3% on average for all datasets except Chennai floods.

3.4.4. Cross-Domain Training (RQ4)

Although using disaster-related training data shows slight gains in most cases, the improvement is still far from the “Target" performance. We anticipate the problem to be the difference in disaster types that we employed for training. Consequently, we study the effect of training on “cross-domain" data, i.e., training on data from previous disasters but a different type than the target, compared to the case when both the source and target disasters are of the same type. In this section, we address **RQ4**: *Is training on combined data from different types of crisis events (cross-domain) better than training on data from the same type of events (in-domain)?*

To this end, we use the following training setups:

- `DIS_FLD.others`: Using data from all flood events for training and testing on other disasters (in this case, other disasters are of type FLD, HRC, and EQK).
- `DIS_HRC.others`: Using data from the hurricane event for training and testing on other disasters (in this case, other disasters are of type FLD and EQK).
- `DIS_EQK.others`: Using data from the earthquake event for training and testing on other disasters (in this case, other disasters are of type FLD and HRC).

Figure 3.4 shows the results. The missing bars in the case of Hurricane Sandy and Christchurch earthquake are because we only have one hurricane and one earthquake

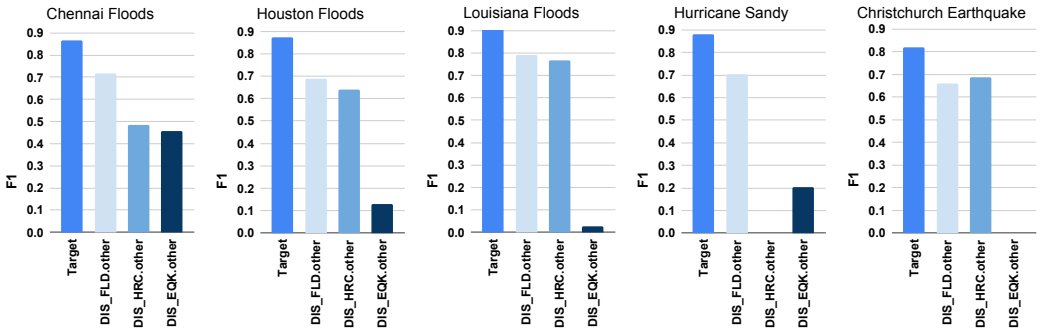


Figure 3.4. The F_1 results of training on cross-domain data. Missing bars indicate no more than one disaster dataset of the target type.

events.

Looking at the results when the target type is floods (the first three sub-figures), training on disasters of the same type as the target (FLD) consistently achieves better performance compared to training on HRC and EQK data. Interestingly, the training performance on FLD and HRC is slightly close for the Houston and Louisiana floods. We suspect the reason is the proximity between the affected areas of Hurricane Sandy, Houston floods, and Louisiana floods, which enhances the model's ability to detect more LMs.

We also notice that training on EQK data is consistently the worst across all disasters. Upon investigation, we found that the location distribution in Christchurch Earthquake is highly skewed (refer to location distributions, Figures E.1-E.5 in Appendix E). Precisely, the location mention "Christchurch" constitutes 262 out of 527 locations (49.7%) and 84 out of 156 locations (53.8%) in the training and test data, respectively. Moreover, 68% of the tweets constituting this dataset have no locations. For this reason, we believe that this dataset is inadequate for training compared to other datasets.

To answer **RQ4**, we found that training on disasters of the same type generally

achieves better performance. To further understand these results, we explore how the geospatial proximity of source events to the target event affects the performance in the following.

3.4.5. Geo Proximity-based Training (RQ5)

The location mentions within the affected areas of a target disaster are expected to emerge in the tweets stream over time. However, LMR models trained on past disaster events data have not seen such locations. Employing an LMR model trained on the closer geographical area as the target disaster (within the same country in our experiments) can alleviate this issue. A concrete example of this is the case of Louisiana floods when trained on Hurricane Sandy data (refer to the previous section). To elaborate, not all countries exhibit the same naming formats (e.g., using street numbers in contrast to names) and administrative levels (e.g., states and counties). In this section, we address **RQ5**: *How does the geospatial proximity of source events to the target event affect the performance?*

To address this question, we use the following training settings:

- `DIS_US.other`: Combining all events from the USA except the target for training. For example, we train on Houston and Louisiana floods if the target disaster is Hurricane Sandy.
- `DIS_IN.FLD`: Training on Chennai Floods happened in India.
- `DIS_NZ.EQK`: Training on Christchurch Earthquake happened in New Zealand.

Due to the lack of diverse disaster-specific labeled data for the LMR task, we could conduct experiments only on target datasets of disaster events in the US; for other areas (NZ and IN), we do not have more than one disaster-specific dataset. Nonetheless,

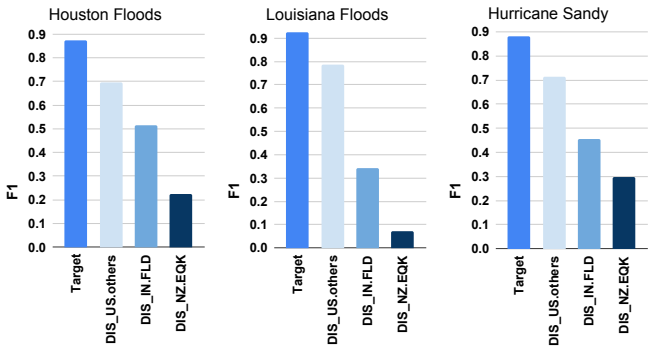


Figure 3.5. The F_1 results of training on geo-proximity-based data.

the results in Figure 3.5 indicate that training on source disasters nearby areas (with respect to the target event) achieves the best performance regardless of the type of disaster.

To answer **RQ5**, we suggest training on disaster events that happened in close areas to the target event to achieve the best performance regardless of the type of disaster.

3.4.6. Cross-lingual Training (RQ6)

Thus far, we have studied the performance of the LMR model from different aspects (i.e., entity type, disaster type, geographical proximity) in a monolingual setup. However, disasters may occur in areas of low-resource languages (e.g., Italian and Turkish) where little or no training data is available. This motivates us to study the performance of LMR models in cross- and multilingual setups. In this section, we address **RQ6**: *Can a model trained on one language effectively recognize location mentions in another language?*

To address this research question, we select three languages, namely English, Italian, and Turkish, based on the availability of labeled data. The source language can be monolingual (English only), bilingual (English and Italian, or English and Turkish), or multilingual (English, Italian, and Turkish). The target is either Italian or Turkish.

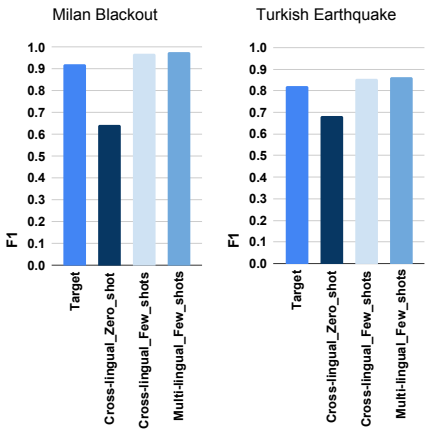


Figure 3.6. The F_1 results of cross-lingual and multilingual training.

We use the following setups:

- `Cross-lingual_Zero_shot`: We fine-tune multilingual BERT on the monolingual source language (English) using the `Combined.joint` and test on the target language (Italian or Turkish).
- `Cross-lingual_Few_shots`: We fine-tune multilingual BERT on the monolingual source language (English) and a *little* data from the target language (bilingual). We then test on the target language.
- `Multilingual_Few_shots`: We fine-tune multilingual BERT on the training data of all available languages, including the target language (multilingual). We test on the target language.

Figure 3.6 demonstrates that the performance of `Cross-lingual_Zero_shot` is acceptable but still far away from the performance level achieved by the *Target* setup. However, adding little training data from the target (312 and 1,400 from the Milan Blackout and Turkey Earthquake disasters, respectively) in `Cross-lingual_Few_shots` and `Multilingual_Few_shots` setups significantly increases the F_1 score of the LMR model to beat the “Target" setup by 4.6% and 3.9%, respectively.

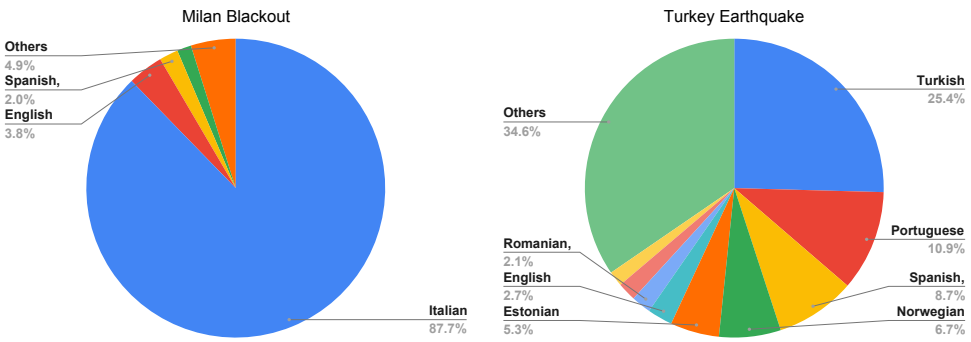


Figure 3.7. The language distribution in Milan Blackout and Turkey Earthquake datasets

Interestingly, `Multilingual_Few_shots` is slightly better than the `Cross-lingual_Few_shots` for both Milan Blackout and Turkey Earthquake. We anticipate the reason to be the popularity of Italian and Turkish languages in both countries. To investigate, we analyzed the language distribution of both datasets using the *langdetect* tool.[5] We show the distribution of languages in Figure 3.7. Surprisingly, the Italian and Turkish tweets constitute only 87.7% and 25.4% of the Milan Blackout and Turkey Earthquake datasets, respectively. The Turkish dataset is much noisier than the Italian dataset due to the popularity of other languages in the country.[6] Additionally, the Turkey Earthquake dataset contains 2.2% Italian tweets, which might explain its usefulness in training when the target is the Italian language. To answer **RQ6**, we conclude that training the multilingual LMR model with no target data achieves adequate performance, but using as little as 263-356 training examples in the target language, which constitutes 87.7% and 25.4% of the Milan Blackout and Turkey Earthquake training data respectively, notably improves the performance.

[5] `https://pypi.org/project/langdetect/`
[6] `https://en.wikipedia.org/wiki/Languages_of_Turkey`

3.4.7. Incremental Training with Target (RQ7)

To this end, we confirmed the need for using reasonably large disaster-specific training data (non-target data) to build an acceptable performing LMR model at the onset of the disaster events. Nevertheless, training robust (highly accurate) LMR models is crucial during emergencies as relief responders are expected to use the geolocation information from these models to make critical decisions. According to our findings in addressing the previous research questions, the LMR models must be trained on target data whenever possible to reach the highest possible performance. To simulate acquiring target-labeled data during disaster events, we study the effect of *gradually* feeding the LMR model with increasing amounts of the target data. Precisely, we aim to determine the minimum number of tweets to annotate from the target data to improve the model performance before reaching a stable performance.

In this section, we address **RQ7**: *How many target event tweets are required to train a reasonably performing (i.e., F_1 >= 0.70) LMR model?* To answer this research question, we explore two aspects: (1) the number of tweets to annotate before reaching a high stable performance, and (2) the best base training data to start with besides the target data. Furthermore, we assume our annotation budget (i.e., cost and time) is sufficient to label 1,050 tweets from every disaster. Therefore, we first train our model with a base training dataset, then incrementally add 105 tweets from the target event chronologically. This number constitutes 10% of the entire training data that can be labeled within our predefined budget. We use only the flood datasets as they contain the tweets' timestamps for the chronological sorting of the training data. We experimented with three different setups of base training datasets:

- *Cold-Start+%*: We do not use any *base* training data. This setup simulates the

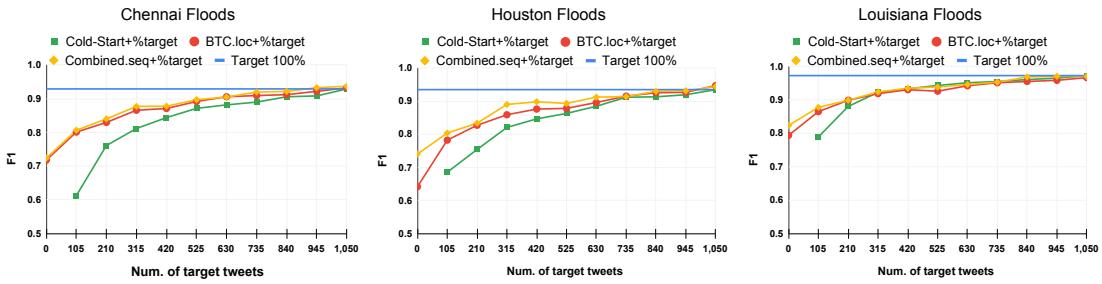


Figure 3.8. The F_1 results of incremental training on target data.

scenario when there is no data to pre-train the LMR model.

- *BTC*.`loc`+%: We use the BTC dataset with only the LOC entity type as the base training data. This setup simulates the scenario when we do not have disaster-related tweets to use for training.
- *Combined*.`seq`+%: We use the best-performing data setup as the base training data (refer to Section 3.4.2). This setup simulates the scenario when we have general-purpose and disaster-related tweet data for training.

In all setups, we increment the training with 10% from the target data and report performance at each increment. As a reference, we also show `Target` baseline, i.e., when training only on the entire 1,050 training tweets.[7]

Results in Figure 3.8 show that increasing the training data improves performance in all training setups. We also notice performance stability when reaching 70% of the target data (training on 735 tweets) and afterward across all base setups for all datasets. Additionally, using “Combined.seq" and “BTC.loc" base training is better than the *Cold-Start* setup until we reach 40% in Chennai and Houston floods and 20% in Louisiana. Furthermore, “Combined.seq" is even more promising than “BTC.loc" as it contains cross- and in-domain data (i.e., disaster-specific tweets) that seem to

[7]Note that the performance of the “Target" in these experiments is different from previous research questions due to sorting tweets chronologically.

improve performance slightly over time. Indeed, these observations show the advantage of exploiting external training data (i.e., out-of-domain or cross-domain) at the onset of disasters to allow some time to collect target data for training.

Although the *cold-start* setup of base training is the worst compared to `BTC.loc` and `Combined.joint`, it exceeds F_1 of 0.8 when using approximately 30% of the target training in Chennai and Houston and 20% in Louisiana. This again emphasizes the need for target data for training. The results also indicate that leveraging other external training data is almost no benefit once we have labeled about 1k tweets from the target event.

Therefore, for **RQ7**, we suggest training on all available training data, regardless of their domain, at the onset of a disaster event to allow some time for annotating target tweets. As for the annotation budget, we suggest labeling around 500 tweets to achieve good performance (about 0.9 F_1 in the three disasters we experimented with), in addition to the cross- and out-or-domain training.

3.5. Error Analysis

To better understand the different types of errors our best model makes, in this section, we closely examine the results of the model. Specifically, we investigate the results obtained from the “target" setup for the five English disaster datasets.

3.5.1. Error Types

We examine four types of errors, i.e., false positives, false negatives, partial matches, and malformed on the entity level. Given the fact that LMs can be composed of more than one token (e.g., “New York"), we look at cases where the predicted LM

Table 3.3. The types of errors in “Target" runs in English disaster datasets. “Partial match +" and “Partial match -" indicate when the predicted LM contains more or less tokens than the gold LM, respectively.

Error Type	Chennai FLD	Houston FLD	Louisiana FLD	HRC Sandy	ChCh EQK
False positive	0	27	11	30	22
False negative	19	21	17	16	19
Partial match +	10	5	0	0	0
Partial match -	18	9	5	7	7
Malformed	28	12	14	10	6
Total	75	74	47	63	54

tokens either partially match the corresponding gold LM (i.e., “partial match -”) or contains extra tokens compared to its gold LM (i.e., “partial match +”). We denote these cases as partial matches. Additionally, any predicted multi-token LM must start with B-LOC tag, end with L-LOC, and I-LOC used for in-between tokens; otherwise, it is considered a malformed LM. Table 3.3 shows the number of errors representing false positives, false negatives, partial matches, and malformed LMs (in the test set only). Tables 3.4 and 3.5 show example tweets along with their error types highlighted.

The false positives are mostly valid fine- and coarse-grained LMs not annotated in the datasets. For example, “HCSO" in tweet#1 and “manassas" in tweet #2 refer to fine- and coarse-grained LMs, respectively, that were not labeled as such. The false negatives are common in all datasets, which implies that the model may still need more data to recognize the LMs better (tweets #3-6, 12-14). The partial matches are more common in Chennai and Houston floods than in other disasters. The partial errors and malformed sequences of *BILOU* tags are more common in Chennai floods because the gold annotations in this dataset *inconsistently sometimes* include the location type (e.g., area) and prepositions (e.g., beyond and along) as part of the LM (e.g., tweet #7) which potentially confuses the model. Additionally, there are location types such as “mosque",

“ATM," and “area," to name a few cases, that are annotated as unit locations (U-LOC), e.g., in tweet #6. These annotation decisions confused the model as to whether it should recognize them as part of precedence LMs or independent LMs.

Below we list our general observations based on closely examining these errors.

a. Surprisingly, we observed that most of the false positives are indeed *valid* LMs. They are either fine- or coarse-grained locations not annotated in the datasets. This highlights a potential issue with the existing datasets. We note here that some of the LMs are not actual location names but were used within a general context, such as “Louisiana" which is mentioned to describe the way grills are made in tweet #3. However, other LMs like "manassas" in tweet #2 should be detected by the LMR model.

b. There are LMs that are misspelled such as “Christchurh" in tweet #4 and concatenated such as “apollohospitals" instead of “apollo hospitals" in tweet #5. This emphasizes the need for an accurate preprocessing pipeline before the recognition phase, including spell-checking and hashtag segmentation.

c. There are location types that are not actual geographical points, such as “area", “ATM," and “mosque". An example of this issue is tweet #6. We suggest filtering these sources of errors before using the datasets for training.

d. There are inconsistent annotations of LMs. LMs that appear multiple times in the dataset are only sometimes annotated using the same sequence of BILOU tags. For example, “South Texas" appears “south$_O$ TX$_U$", “south$_O$ Texas$_O$", and “Southeast$_B$ Texas$_L$" in the same dataset. Also, “HCL office" appears twice in tweets, but it is labeled once as a multi-token LM, and another time only “HCL"

is labeled as a single token LM (e.g., tweets #11 and #12). Additionally, “5th street" in tweet #10, which appears twice in the dataset, is labeled as an LM in the original tweet but not annotated in its retweet.

e. There are ambiguous locations that we could not resolve to exist, such as “World", “Swayamsevaks," and “Congress." An example of this issue is tweet #13.

f. There are abbreviated LMs such as “global hosp" instead of “global hospital" in tweet #14, which adds difficulty to recognizing LMs. The issue becomes more challenging when the LM abbreviations are common English words, such as “ok," that appear in the training data of “Houston Floods" dataset to denote “Oklahoma" state (Refer to Figure E.2).

g. The LMR model correctly ignores some ORG entities, but they appear in the gold annotations as LMs. For example, “FEMA" in Tweet #15 refers to the “Federal Emergency Management Agency". Additionally, in Figure E.3, the ’Clinton Foundations’ and ’nytimes’ are ORG entities rather than LOC entities. This issue and its reverse exist across the datasets we adopted in our experiments. The reverse of it can be illustrated by ORG entities in the “Houston Floods" event, where “NWS" and “TXDOT" local organizations are correctly labeled as LMs by the LMR model, but they are not gold LMs. Discovering such annotations in the LMR datasets shows the difficulty of the annotation task conducted by human annotators, as they are confused by whether such entities are mentioned within the context of the tweet as organizations or locations of the offices of these organizations. Such confusion does negatively affect the LMR model performance.

Table 3.4. Examples of errors of BERT$_{LMR}$ model. Underlined text is the gold LM. The double-underlined text refers to gold LMs in two duplicates of the same tweet. Highlighted text is the predicted LM.

T#	Error Type	Tweet text
#1	False positive	Here's a look @ our downtown parking lot outside of HCSO headquarters **downtown**. #**Hou**News #TurnAroundDontDrown [url]
#2	False positive	Are the roads flooded in manassas??
#3	False positive & False negative	Check out Louisiana Grills at **Rich & John's** "The Stove Shop" #summer #BBQ [url]
#4	False negative: misspelling	Another big quake, 6.3 has hit **Christchurh** during work hours. Not a pretty site. Buildings damaged, some collapsed.
#5	False negative: concatenated	#**apollohospitals**- helplines -share - [user_mention] [user_mention] ... [url]
#6	False negative: location type	All the #**mosque** in **chennai** now open for food and stay. Thank you Allah !!
#7	Partial match: adpositions	[user_mention] Navy rescue team deployed in **Gandhi nagar area**, **beyond Adyar Bridge along Buckingham Canal**
#8	Partial match & Malformed	Crescent College (**B. S. Abdur Rahman University**), #Vandalur is open for shelter. [...]
#9	False positive & Malformed	#SANDY ive never seen nyc look like this #HurricaneSandy the flooding is unreal....long beach$_{L-LOC}$ **long island** #unrecognizable
#10	False positive: inconsistencies	Looks like Malad :p [̈user_mention]]: Street flooding #**NYC**: **48th Ave** between **5th** and **Center Blvd** #Sandy [url]
#11	Partial match: inconsistencies	anyone to help ppl stuck in **HCL office** for 4 days at **Navalur**. Boats r reqd 2 transport ppl home. Phones unreachable..

Table 3.5. Examples of errors of BERT$_{LMR}$ model. Underlined text is the gold LM. The double-underlined text refers to gold LMs in two duplicates of the same tweet. Highlighted text is the predicted LM.

#	Error Type	Tweet text
#12	False negative: inconsistencies	Need help to people who are in **ELCOT** branch **HCL** office, **shollinganallur**
#13	False negative: ambiguous	**Swayamsevaks** preparing & distributing food to around 1500 poor people in flood affected **Lakshmipuram**
#14	False negative: abbreviation	[user_mention] [user_mention] No1 is allowed even 1 km ahead of **global hosp**. Poliz army local helpers available on spot as of 6pm today.
#15	False Negative: ORG not LOC	**Louisiana** Flooding, One Week Later: Author: J essica StapfUnprecedented. Historic. E ... disaster **fema**

3.5.2. Location Types

We further analyzed the errors based on their granularity level, i.e., *fine-grained* location mentions such as point-of-interest (POI), road/street name, and neighborhood, and *coarse-grained* locations such as country, county, and state. We first labeled the LMs, which represent one of the error types mentioned above, with their granularity level using Google Places API[8] and manual search (in case Google API does not return any result).

We show the number of LMs for each granularity level in Table 3.6. For example, the “Other" type represents locations that were not resolved through Google Places API as well as through manual search. Overall, we notice that most errors are fine-grained, originating from flood-related disasters. Moreover, the model makes more mistakes in detecting coarse-grained locations from Hurricane Sandy and the Christchurch earthquake (tweets #9 and 4 in Table 3.4, respectively).

[8] https://developers.google.com/maps/documentation/places/web-service/overview

Table 3.6. Location types of miss predicted LMs. “FG” and “CG” refer to fine-grained and coarse-grained locations. “FP”, “FN”, and “PM” refer to false positives, false negatives, and partial matches.

	Chennai FLD			Houston FLD			Louisiana FLD			HRC Sandy			ChCh EQK			
	FP	FN	PM	FP	FN	PM	FP	FN	PM	FP	FN	PM	FP	FN	PM	Total
FG	0	11	23	19	11	7	3	13	2	6	6	6	2	6	6	121
CG	0	5	4	4	9	7	3	4	3	21	10	1	20	13	1	105
Other	0	3	1	4	1	0	5	0	0	3	0	0	0	0	0	17
Total	0	19	28	27	21	14	11	17	5	30	16	7	22	19	7	243

3.6. Limitations

The methodological limitations of our study are associated with the technical aspect of the LMR model and the experimental evaluation (setups and datasets). We list the main ones in the following:

a. Generally, two factors related to the construction of the dataset could affect our results and conclusions. First, the datasets we adopt are scarce, with limited disaster types and geographical area coverage. Over and above that, the datasets are randomly drawn from collected tweet datasets using hashtags. Relying on hashtags is a significant limitation for capturing the relevant posts discussing the target disaster event, as some users do not use hashtags while posting about the event [28], [147]. Additionally, hashtag-based datasets might have different characteristics than other datasets collected in different ways, such as geographical-based ones, affecting the drawn conclusions [26].

b. We explored the effect of the *data domain*, *entity type*, *disaster domain*, and *geo-proximity* factors in monolingual setup for English language in the crisis management domain. Our conclusions might not translate to other languages and domains.

c. The available datasets for experiments have limited disaster domain and geographical coverage. Thus, while studying the effect of associated factors, the conclusions might not translate to other datasets with better coverage and exact size for different disaster types and geographical areas.

d. For the *language* factor, we studied the cross- and multi-lingual setups for English, Italian, and Turkish languages. Hence, we note that:

   - The datasets we used, as shown in Figure 3.7, are not pure for their respective languages. Further filtering by language is required for solid conclusions. However, due to multiple reasons, including (1) the small size of the data, (2) the employment of the multilingual BERT model, and (3) the concern of making results comparable to future studies, we opted to use the data as is.
   - Generalizing our conclusions to other languages requires further investigation.
   - Our study relied on the power of the multilingual BERT model. Contextual translation of the data might lead to different conclusions.

e. We discussed some issues in the annotations of the used datasets in Section 3.5, that could affect the performance of our BERT$_{LMR}$ model. Unfortunately, there are no standard guidelines for annotating LMR datasets of a higher standard yet. In fact, the annotation guidelines used to train the annotators for constructing the LMR datasets used in our study were *not* made publicly available to the community.

All the issues listed above motivate the need for a larger, domain- and geographically diverse, consistently annotated LMR datasets. Therefore, we construct our datasets

in the following chapter.

There is no LMP unified evaluation framework with all essential components, including annotated datasets, diverse open-source (or public black box) baselines, and fair evaluation metrics for the disaster management domain over Twitter. In fact, the absence of large and generalizable datasets, specifically, makes the comparison difficult between the existing LMP systems. Therefore, we exhaustively review the efforts made to provide LMR and LMD datasets in Sections 2.4.2.1 and 2.5.2.1, respectively. In a nutshell, the existing English LMP tweet datasets are either nonpublic non-disaster-specific [39], [64], [66], [73], [85], [89], [101], [105], [121], or disaster-specific but suffer from several limitations [14], [56], [63], [87], [111], [115], [123]–[126] such as the limited size, the confined domain and geographical coverage, the absence of location type annotations, among others.

Table 4.1. Datasets and Benchmarks chapter outline.

Topic	IDRISI-RE	IDRISI-RA	IDRISI-DE	IDRISI-DA
Introduction	4.2	4.3	4.5	4.5
Constructions	4.2.1	4.3.1	4.5.1	4.5.1
Analysis	4.2.2	4.3.2	4.5.2	4.5.2
Benchmarking	4.2.3	4.3.3	5.4	5.4
Generalizability	4.2.4	4.3.4	-	-

In this chapter, we discuss our effort to fill this gap in the disaster management domain. Specifically, we introduce IDRISI[1] largest to date and generalizable datasets and establish a set of competitive baselines for the LMP task. We start the section by setting our objectives for the LMP datasets in Section 4.1. We then thoroughly discuss the LMR and LMD datasets' creation, analysis, and benchmarking. Table 4.1 outlines the chapter for all datasets. The introduction discusses the motivation, research questions,

[1]Named after Muhammad Al-Idrisi, who is one of the pioneers and founders of advanced geography: `https://en.wikipedia.org/wiki/Muhammad_al-Idrisi`.

and contributions. For datasets construction, we discuss the selection, sampling, and annotation efforts. We then analyze the reliability and coverage of the IDRISI-R datasets. We also analyze the reliability and usefulness of features in IDRISI-D datasets. For benchmarking, we employ a representative set of models under different data and task setups and discuss their results. Furthermore, we study the domain and geographical generalizability of IDRISI-R datasets. Finally, we discuss the limitations of IDRISI datasets in Section 4.6.

4.1. Objectives

Creating LMP datasets for practical, event-centric, and fine-grained evaluation requires identifying characteristics that guide the dataset construction efforts. Grounded on our review of past efforts (refer to Section 2.4.2.1 in Chapter 2), we introduce a set of characteristics that, we anticipate, can form optimal LMP datasets in the following:

O1. ***Geographical coverage***: The naming conventions of places vary from one country to another, which decisively affects the performance of LMP models. The wider the geographical coverage of an LMP dataset, the more naming conventions it captures. While constructing IDRISI, we aim to capture various naming conventions by annotating disaster events covering many English-speaking countries.

O2. ***Domain coverage***: At the onset of disaster events, acquiring training data is impractical and expensive. Alternatively, an acceptable performing LMP model could be trained using previous disasters of the same type (i.e., in-domain data) [19]. However, as such an approach is infeasible due to the limited domain coverage of existing datasets, we aim to cover various disaster types with a greater number of tweets when constructing IDRISI.

O3. ***Location type annotations***: The location types (e.g., cities, streets, and POIs) allow customizing the downstream applications to meet the responders' needs, such as generating crisis maps at different location granularity. Additionally, the evaluation per location type shows the weaknesses and strengths of LMP models based on the responders' preferences. Therefore, when constructing IDRISI, we aim to annotate the LMs into location types to remedy such deficiency.

O4. ***Large-scale***: Learning models, in particular deep neural networks, are data hungry. Thus, models trained on many training examples tend to yield higher performance and generalize to unseen data. However, most of the existing LMP datasets are limited in size (refer to Table 2.1). We aim to overcome this shortcoming while creating IDRISI by annotating more LMs than 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 proper temporal coverage of disaster events, we aim to overcome this issue while creating IDRISI.

O6. ***Relevance and informativeness***: An LMP dataset has to contain informative and actionable tweets to support effective disaster management. Unlike existing datasets, in constructing IDRISI, we extend a dataset already labeled for informativeness. This simulates the expected input to the LMP models in real-world information processing systems for disaster management.

O7. ***Dialectical Arabic coverage***: Dialects are widespread in Arabic tweets; therefore, a diverse set of them should be represented in Arabic datasets, besides Modern Standard Arabic (MSA). Therefore, we aim to cover as many dialectical tweets as

possible in IDRISI.

We emphasize here that these objectives constitute a generalizable LMP dataset and should be collectively achieved to eliminate any barrier against establishing an effective and fair evaluation framework for LMP tasks. Throughout this chapter, we elaborate on how IDRISI achieves these objectives.

4.2. English LMR Datasets and Benchmarks

In this section, we introduce IDRISI-RE dataset,[2] the largest to date manually-labeled (*gold* version) and automatically-labeled (*silver* version) tweet datasets for LMR comprising 19 English events, whose tweets are labeled to identify both toponyms and their geographical types. IDRISI-RE covers 19 disaster incidents 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? (**RQ8**)
- Unseen events of *different* disaster types? (**RQ9**)
- Unseen events that happen in the *same geographical* areas? (**RQ10**)
- Unseen events that happen in *different geographical* areas? (**RQ11**)

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.

[2]The "R" and "E" letters refer to the **R**ecognition task and the **E**nglish language, respectively.

- The geographical coverage and the data size are the top influencers on the generalizability of the English LMR datasets.
- IDRISI-RE shows a decent reliability level and reasonable geographical, domain, temporal, and location granularity coverage.
- A thorough experimental evaluation of a representative set of LMR models shows 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.[3] 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 LMP models and to allow finer evaluation and comparison between 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.

4.2.1. Construction

Two main factors guided the choice of our underlying dataset. First, while responders look for *informative posts on Twitter*, the tweets become more invaluable with

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

the geographical context [20]. Second, *the likelihood of LMs occurrence* increases during events [148]. Consequently, we selected an *event-centric* dataset 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 [149] due to its geographically broad coverage and disaster domain diversity.

We carried out two annotation versions: (i) the *gold* annotations using crowdsourced human labels, and (i) the *silver* annotations using an automatically-learned model.

4.2.1.1. Gold Dataset Sampling

We focus our sampling on the most informative tweets to label them by human annotators. In the following, we describe the sampling methods we used for IDRISI-RE.

Before creating our pool of tweets for manual annotation (i.e., gold), we dropped the less informative classes (to relief authorities) in 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 an overall cost budget of $4,300, 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 sample a maximum of 1,300 tweets per event randomly. 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.

4.2.1.2. Gold Annotations

We gather two types of annotation on the selected data in two tasks. The first task involves human annotators identifying toponyms within the tweet text, such as geographical names of places. In the second task, they assign location types to the identified toponyms. These location types include country, province/state, city/town, district, neighborhood, road/street, natural points of interest like river and sea, and human-made points of interest such as schools and hospitals. Toponyms not belonging to the defined location types are assigned the “other location" label. We provided detailed annotation guidelines for annotators with examples to clearly articulate our definition of location mentions.[4]

We used *Appen* crowdsourcing platform[5] 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.[6]

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

[4] https://github.com/rsuwaileh/IDRISI/tree/main/LMR/annotation_guidelines
[5] http://success.appen.com
[6] https://spacy.io/

level is below a minimum confidence of 80%,[7] we allowed dynamically collecting of up to five more annotations by different annotators, 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 the case of ties, we selected the longest span. Finally, to ensure the quality of labels, we deleted all annotated spans of length equal 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 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 while disambiguating them.

Table 4.10 shows the final number of tweets (column "Tweets"), the number of tweets with no LMs (column "Tweets$_{|LM|=0}$"), and the total number of annotated LMs with the unique LMs in parentheses (column "LMs (uniq)").

[7]The confidence level is computed by adding up the confidence scores of the contributed workers.

4.2.2. Description and Quality

In this section, we thoroughly evaluate the IDRISI-RE dataset in terms of reliability, consistency, coverage, and diversity.

4.2.2.1. 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-α) [150] to measure the reliability of annotations. Unlike Fleiss Kappa, k-α 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-α 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-α for *TYPE* annotation (which implicitly encodes the *LOC* annotation). Figure 4.1 shows the k-α 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 acceptable reliability. Overall, the IAA shows that the annotations are highly reliable for three events, acceptable for twelve events, and low quality for four events.

LM definition across English datasets: Table 4.2 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

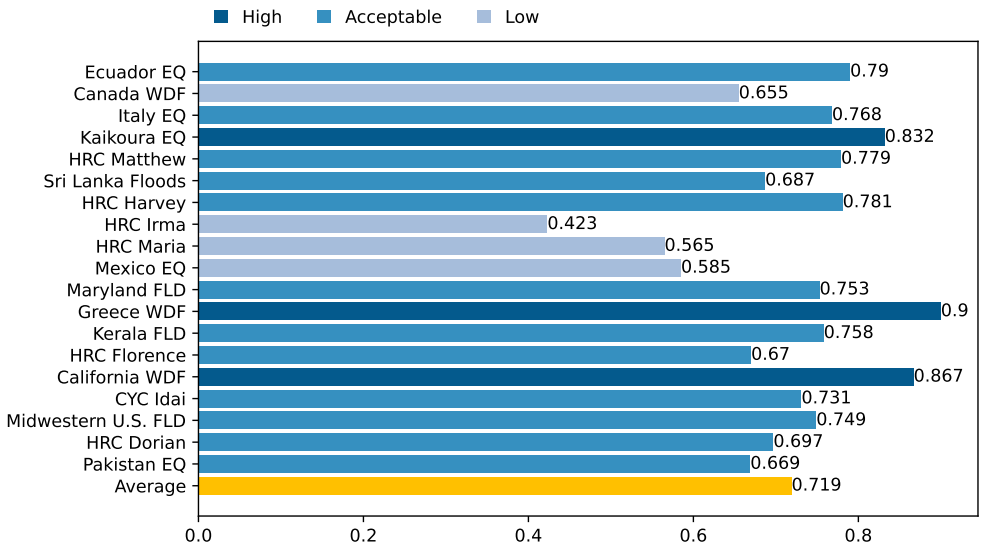


Figure 4.1. k-α for IDRISI-RE per disaster event.

considered LMs or not in the corresponding datasets. Table 4.3 presents example tweets from IDRISI-RE to articulate our LM definition and distinguish it from other datasets. In particular, 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 4.3). The locations within user mentions (tokens start with "@") are considered LMs in the ALTA and KHAN datasets while ignored in all other datasets. Although user mentions could indicate the location of incidents discussed in the tweet, we do not consider them as LMs in IDRISI-RE because they typically refer to organizations or people, 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 1km away from Jatlan". We follow the same intuition for addresses and routes. For instance, in Tweet #4, the consecutive LMs have to be labeled independently.[8]

Table 4.2. 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 [123]	✓	✓	×	×
ALTA [87]	✓	✓	✓	✓
OLM [14]	✓	×	×	✓
GeoCorpora [115]	✓	×	×	×
HU1 [124]	✓	×	×	×
HU3 [125]	✓	×	×	×
KHAN [111]	✓	✓	×	✓
FGLOCTweet [126]	✓	✓	×	✓
IDRISI-RE	✓	×	×	×

Table 4.3. Example tweets from IDRISI-RE dataset. In our annotation guidelines, the bold and gray-shaded LMs represent the undesired and desired LMs, respectively.

T#	Tweet
#1	To all my followers please RT: Where to #Donate to #Mexico #Earthquake Victims - @nytimes #PrayFor**Mexico**
#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** ...
#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

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

4.2.2.2. Coverage and Diversity

In this section, we discuss how IDRISI-RE satisfies the properties presented in Section 4.1.

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 4.1). 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 4.1), 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 [26], [135], [136], [151]–[153]. 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 4.1). Figure 4.2 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 that the key factor that explains the dominance of coarse-grain LMs is the HumAID

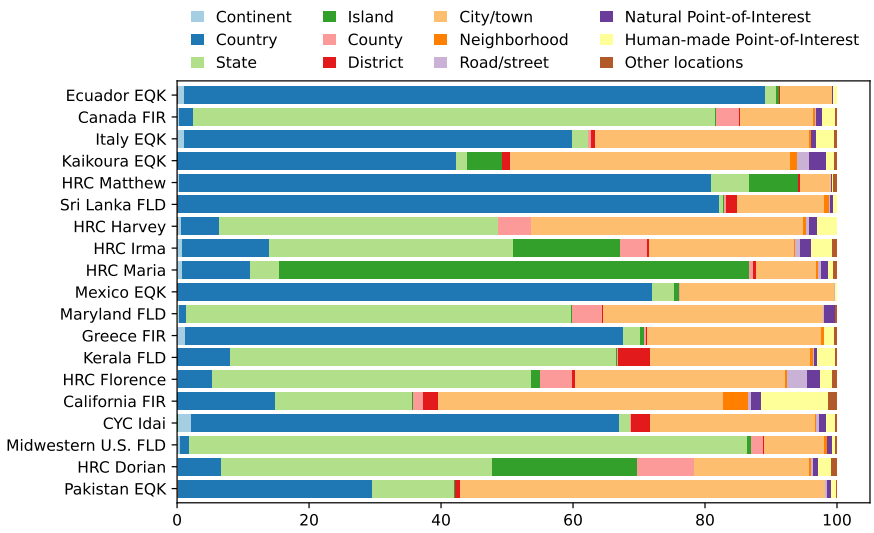


Figure 4.2. Distribution of location types in IDRISI-RE. HRC, EQK, FLD, CYC, and FIR refer to Hurricanes, Earthquakes, Floods, Cyclones, and Wildfires, respectively.

dataset creation method. HumAID was collected by tracking relevant keywords to the disaster events, usually the coarse-grained impacted areas' names. Indeed, these coarse-grained LMs are less challenging to detect by annotators. Consequently, we could not prevent annotators from detecting them, or 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 likely 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 2.1, showing that IDRISI-RE contains the maximum number of unique LMs (3,830 LMs). Figures E.6-E.9 show the distribution of the top 15 LMs per disaster event in Appendix E.

Temporal Coverage: Ideally, the event-centric datasets should span over the entire period of a disaster event to allow the response authorities to operate efficiently during all phases of the disaster events (refer to **O5** in Section 4.1). The events in IDRISI-RE

were crawled two days before and two days after their peak incidents [149]. In Figure B.1 in Appendix B, we depict the number of tweets during two events showing the coverage of important developments.

4.2.3. 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* LMR (refer to Section 3.1). We also use two data setups: *(i) random* and *(ii) time-based*. We ignore tweets' timestamps in the random setup and randomly select train, development, and test examples. The data is chronologically ordered in the time-based setup before splitting it into training, development, and test sets. Tweets are randomly shuffled and split into 70% training, 10% development, and 20% test sets per event. We report the detailed stats in Tables B.1 and B.2 in Appendix B.

4.2.3.1. Learning Models

We have developed our own LMR models:

- CRF$_{LMR}$ [154]: The Conditional Random Fields (CRFs) model is a competitive probabilistic tagging algorithm, which can be used as a standalone tagger [101], [105] or integrated into an LMR system [38]–[40], [155]. We used the *crfsuite* library[9] to train a CRF$_{LMR}$ 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$_{LMR}$ [19]: This models is a fine-tuned pre-trained BERT model for the LMR task. We add a linear layer on top of the vanilla BERT model (refer to Section 3.2).

[9] https://sklearn-crfsuite.readthedocs.io/

We have also select a representative set of LMR models as described below:

- GPNE [110]: An *unsupervised* LMR model that specifically shows better detection performance for location mentions within the disaster-hit areas.
- GPNE_2 [94]: An enhanced version for GazPNE that employs an LMD module to improve the LMR accuracy. It exploits the Stanza NER model to accelerate recognition and detect hard LMs. This system uses synthesized training data extracted from gazetteers. Hence, we cannot retrain/fine-tune it using our data. Instead, we use its public CLSTM-trained model.
- NTPR_O [40]: A neural-based toponym recognition tool trained on recurrent neural networks. We run the **o**riginal trained model 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 hyper-parameters and adopt the values used by the authors [40].
- 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 [109]: An untrainable rule-based recognition model [109]. We run the original application that is made public by the authors.
- NLORE [108]: 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.[10]

4.2.3.2. 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.

[10]The authors promise to make it open source in the future.

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_{LMR} model, 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 $\{2^{-i} | i \in [0, 3]\}$.

4.2.3.3. 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 the entity level rather than the token level. Our evaluation differs from *seqeval*[11] in three aspects: (1) it evaluates per tweet and reports the average performance, (2) it rewards the models when they

[11] `https://pypi.org/project/seqeval/`

correctly predict no LMs for a single tweet, and (3) it accepts BILOU-like or JSON formats.

Table 4.6. 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$_{LMR}$	BERT$_{LMR}$	CRF$_{LMR}$	BERT$_{LMR}$
Ecuador EQK	0.932	**0.939**	0.910	**0.926**
Canada FIR	**0.853**	0.733	**0.865**	0.771
Italy EQK	**0.906**	0.890	**0.881**	**0.881**
Kaikoura EQK	0.879	**0.909**	0.875	**0.899**
HRC Matthew	0.901	**0.919**	0.899	**0.952**
Sri Lanka FLD	0.910	**0.925**	0.897	**0.912**
HRC Harvey	0.906	**0.909**	**0.914**	0.895
HRC Irma	**0.906**	0.833	**0.893**	0.823
HRC Maria	0.882	**0.924**	0.890	**0.897**
Mexico EQK	0.838	**0.913**	0.880	**0.911**
Maryland FLD	0.751	**0.892**	**0.873**	0.805
Greece FIR	0.896	**0.925**	0.886	**0.887**
Kerala FLD	**0.880**	**0.880**	0.857	**0.919**
HRC Florence	**0.879**	0.772	**0.889**	0.778
California FIR	0.907	**0.909**	**0.902**	**0.902**
Cyclone Idai	0.877	**0.900**	0.852	**0.895**
Midwestern U.S. FLD	0.917	**0.936**	0.920	**0.944**
HRC Dorian	**0.875**	0.858	**0.865**	0.852
Pakistan EQK	0.820	**0.894**	0.780	**0.828**
Average	0.880	**0.883**	**0.880**	0.878

4.2.3.4. Benchmarking Results

Type-less LMR: In this setup, the LMR models recognize LMs, regardless of their types. Tables 4.4 and 4.5 present the F_1 results of all LMR models over IDRISI-RE events. We also report the detailed results, including precision and recall, with the best hyper-parameters for the BERT$_{LMR}$ and CRF$_{LMR}$ models in Appendix C. 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 show the second-best performance followed by the CRF$_{LMR}$ 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_{LMR} model. In contrast to the *random* setup, the NTPR_O performance is better than the CRF_{LMR} model under the *time-based* data setup. Specifically, the CRF_{LMR} 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 others. GPNE shows poorer performance than the other models. However, its new release (GPNE_2) outperforms LORE and NLORE under the *random* and *time-based* data setups and the CRF_{LMR} model in the *time-based* data setup.

Type-based LMR: In this setup, the LMR models recognize the LMs and predict their types simultaneously. Table 4.6 showed the results of all models over IDRISI-RE dataset. The CRF_{LMR} 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.

4.2.4. 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 required to obtain a model's generalizability is its training dataset. However, most existing datasets lack essential characteristics to achieve better generalizability. To overcome these issues, IDRISI-RE dataset is designed to cover data events that span broader geographical locations and multiple disaster types/domains (e.g., floods, earthquakes). 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 2.4.2.1). 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 4.2.3); 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 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, after combining events, we divide the training set 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 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 solid conclusions regarding generalizability.

4.2.4.1. 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).*

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 a *zero-shot* learning setup for specific events. We note that *all* of the 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 D.1 in Appendix D 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 made public by the authors. We used only the Hurricane Michael 2018 event since the other events have very few tweets in the order of tens, which is inadequate for training the model [19]. 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). We used the tracking keywords used in crawling the dataset to extract matching tweets for each domain [115], [126]. We excluded the hurricane tweets from the FGLOCTweet dataset due to the small number of relevant tweets (only 13). We then partition each domain's tweets into 70% training, 10% development, and 20% test. We split the GEO dataset because no standard splits are public for the research community. We also split the 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.

Results and Discussion: We made several observations on the model's performance and analyzed the results to answer the *domain* generalizability research questions: can an LMR model that is trained on IDRISI-RE generalize to:

- Unseen events of the same disaster type? (**RQ8**)
- Unseen events of different disaster types? (**RQ9**)

In-domain: To address **RQ8**, 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 Figure 4.3 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:

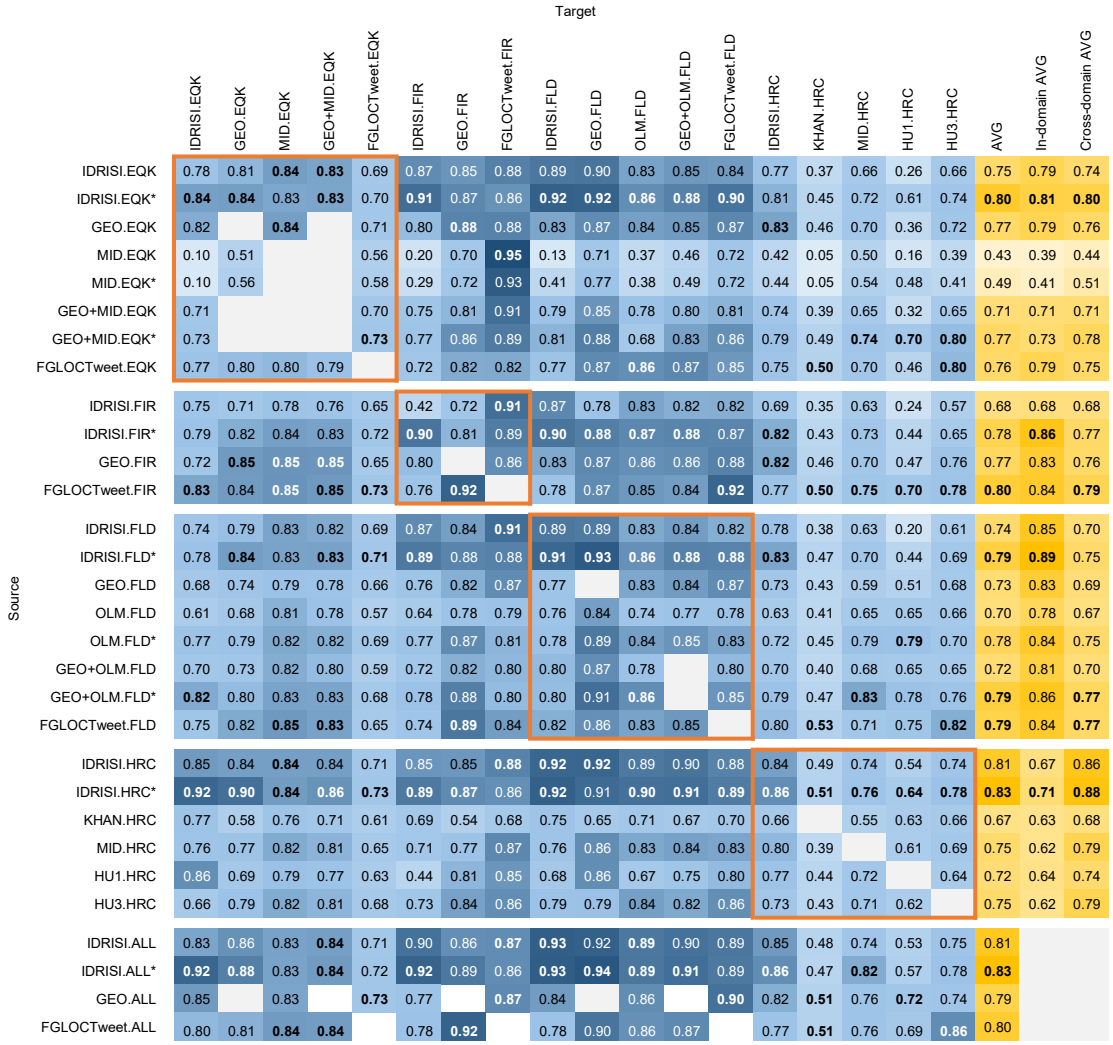


Figure 4.3. The F_1 results of the domain generalizability experiments of IDRISI-RE against existing datasets. The best results per column are **boldfaced** column-wise, per disaster domain. EQK, FIR, FLD, and HRC refer to Earthquake, Wildfire, Flood, and Hurricane, respectively.

- *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 domain. In particular, over all domains, the *IDRISI.<domain>** sources sets consistently generate better models compared to *IDRISI.<domain>* sources sets. These results emphasize the importance of acquiring 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 in improving performance. For instance, we found that 40% of the LMs in *GEO.EQK* source set is 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 shows that training on *GEO.EQK* generates a superior model than training on *IDRISI.EQK*.

To answer **RQ8**, we show that IDRISI-RE dataset generates the best domain generalizable models per domain, compared to the other LMR datasets.

Cross-domain: To address **RQ9**, we study the domain generalizability of IDRISI-RE within different disaster types for source and target sets. Figure 4.3 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 "Christchurch" LM 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, 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 constitutes 20-22% for each 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*: We do not apply size normalization for these models. 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 slightly lower performance by approximately 2.5%).

To answer **RQ9**, 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 distinguishing it from the existing datasets, training on *IDRISI.ALL** generates LMR models that surpass the ones trained on *GEO.ALL*, *FGLOCTweet.ALL*, and *IDRISI.ALL*, on average.

4.2.4.2. 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).*

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 the reported experiments are under zero-shot learning (experiments where the training and test sets include the same event are hidden/greyed in the figures).

Results and Discussion:

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? (**RQ10**)
- Unseen events that happen in different geographical areas? (**RQ11**)

Geographical Generalizability within the same country: Figure 4.4 presents the F_1 scores for the geographical generalizability experiments for the events in the United States. We limit our experiments to events 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** outper-

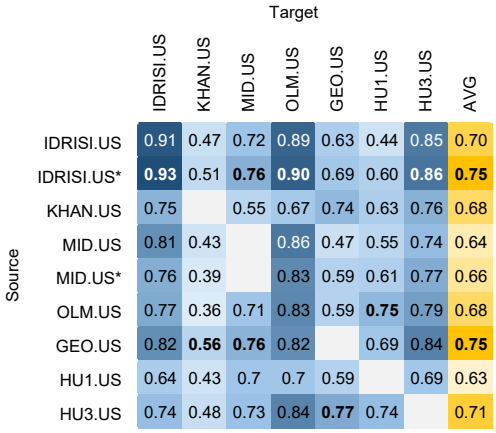


Figure 4.4. The geographical inter-generalizability F_1 results for IDRISI-RE for the *geographical few-shot learning*. The blue color scale is global for the entire matrix. The best results per column are **boldfaced**.

forms 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 **RQ10**, we conclude that the models trained on IDRISI-RE exhibit an acceptable F_1 average score of 0.75. Furthermore, 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: Figure 4.5 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 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*

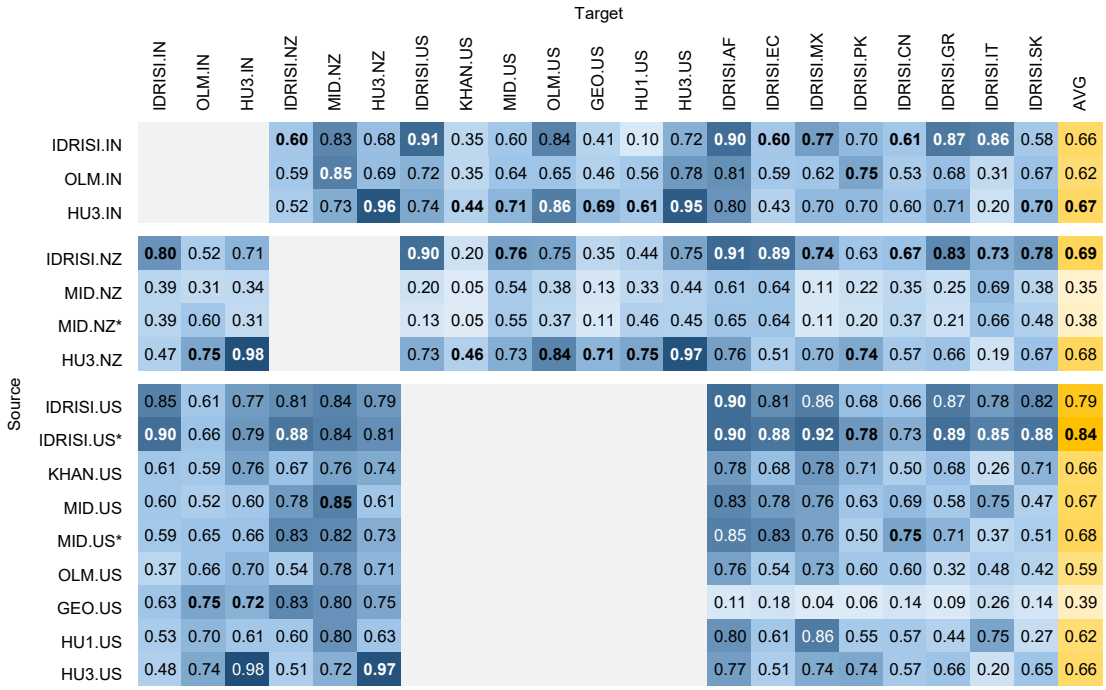


Figure 4.5. The geographical inter-generalizability F_1 results for IDRISI-RE for the *geographical zero-shot learning*. IN, NZ, and the US refer to India, New Zealand, and the United States, respectively. The blue color scale is global for the entire matrix. The best results per geographical area per column are boldfaced.

geographical areas. The high performance of *HU3.IN* on *HU3.NZ* and *HU3.US* leads to the best average score, yet comparable to *IDRISI.IN*. Similarly, the high performance of *HU3.NZ* on *HU3.IN* and *HU3.US* leads to a comparable average against *IDRISI.NZ*.

To answer **RQ11**, it is pretty evident that training on IDRISI-RE generates the best-performing LMR models that can reasonably generalize to events that happened in different geographical areas. On the other hand, the performance of models generated by other datasets is usually relatively poor.

4.2.4.3. Domain Transfer within IDRISI-RE

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 4.2.3.2) for each transfer setup over the development sets (same events as the training/source sets). IDRISI-RE covers four disaster types: hurricane, earthquake, flood, and wildfire. A transfer data setup comprises a source-target pair, resulting in 16 setups.

Results and Discussion: Figure 4.6 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.

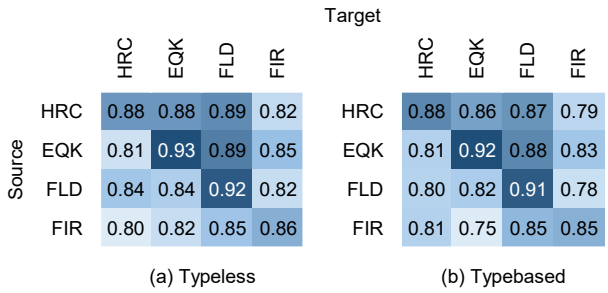


Figure 4.6. The F_1 results for the domain transfer experiments within IDRISI-RE. HRC, EQK, FLD, and FIR refer to HRCs, EQKs, FLD, and FIR, respectively.

4.3. Arabic LMR Datasets and Benchmarks

Worldwide and in the Arab region, Twitter has played a critical operational role in crisis management. The Beirut explosion in 2020 is an excellent case in point, where on-site individuals started intuitively responding to each other. What makes tweets invaluable is location mentions at different granularity [8]–[11]. Response authorities exploit this geographical information to effectively manage emergencies using *Crisis Maps*. Although the geographical dimension adds situational and operational values to Twitter data, Twitter announced on 18 June 2019 that it removed the geotagging feature in tweets.[12] This necessitates developing automatic geolocation tools. Nevertheless, the main obstacle for the Arabic language is being a low-resource language where the geolocation tasks are severely understudied due to the absence of a unified evaluation framework constituting annotated datasets, a representative set of baselines, and fair evaluation metrics.

To address these barriers, we introduce IDRISI-RA,[13] the first human-labeled dataset comprising Arabic tweets from 7 disaster events (*gold* annotations). IDRISI-RA offers the first large-scale automatically-labeled tweets (*silver* annotations) from 22

[12] https://twitter.com/TwitterSupport/status/1141039841993355264
[13] The "R" and "A" letters refer to the **R**ecognition task and the **A**rabic language, respectively.

disaster events that cover the Arab world. It also covers the most occurring disaster types in the Arab world. More importantly, it is labeled for two annotation types that are location mentions (i.e., toponym textual spans) and their types (e.g., city, street, and POIs). Hence, it supports *type-less* and *type-based* LMR.

Although adapting Named Entity Recognition (NER) models and datasets goes a long way towards tackling the LMR task, we have empirically shown that training English-specialized LMR models is compulsory for highly performing models in the emergency management domain (refer to Chapter 3) [19]. Translating this to Arabic LMR requires Twitter NER datasets; however, a few datasets exist yet suffer from the limited size and the confined domain, geographical, and dialectical coverage (refer to Section 2.4.2.1) [128]–[130].

What exacerbates the low resources issue for the Arabic LMR is the unavailability of the few LMR datasets created for traffic surveillance or event detection tasks [73], [74], [127], [131]. Therefore, to expedite the development of Arabic LMR models and shape the future directions, we perform extensive experiments to answer the following research questions empirically:

- **RQ12**: Are standard Arabic NER models sufficient for effective LMR over disaster tweets?
- Can LMR models trained on IDRISI-RA generate generalizable LMR models that reasonably perform on:
  - **RQ13**: Unseen disaster events?
  - **RQ14**: Unseen disaster events of the same or different types (domain generalizability)?
  - **RQ15**: Unseen disaster events in different countries (geographical general-

izability)?

Our rigorous analyses and experiments necessitate the development of specific LMR datasets and models that performs accurately in the disaster domain. Additionally, the experiments empirically confirm the promising generalizability of IDRISI-RA under *zero-shot learning*, and the reasonable domain and geographical generalizability.

The contribution of this paper is fourfold:

- We present IDRISI-RA,[14] the first public human-labeled Arabic LMR dataset (*gold* version) of about 4.6k tweets. The dataset covers diverse disaster types and countries.
- We release the largest automatically-labeled Arabic LMR dataset (*silver* version), constituting about 1.2M tweets.
- We annotate the location mentions into coarse- and fine-grained location types to enable hierarchical LM recognition, disambiguation, and evaluation.
- We benchmark IDRISI-RA using the standard Arabic NER models and our own simple yet competitive LMR models to establish a set of baselines for the research community.
- We empirically demonstrate that IDRISI-RA is a reasonably generalizable dataset.

4.3.1. Construction

Similar to IDRISI-RE, we selected an *event-centric* dataset already labeled for humanitarian categories. We analyzed multiple existing disaster-related tweet datasets and selected Kawarith [138], as it contains tweets from 22 disaster events from the Arab region, which makes it geographically wide and domain (disaster domain) diverse.

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

4.3.1.1. Gold Dataset Sampling

To sample the most informative tweets for human annotation, we selected tweets from seven events (listed in Table 4.10) labeled as relevant for humanitarian purposes. Selected tweets (6,182) are used to download full tweet content using the Twitter API, which resulted in 4,593 tweets.

4.3.1.2. Gold Annotations

We perform two annotation types on the selected data: (1) location mentions identification, such as geographical names of places, within the tweet text, and (2) location type selection for the identified toponyms. These location types include country, province/state, city/town, district, neighborhood, road/street, natural points of interest like rivers and seas, and human-made points of interest such as schools and hospitals. Toponyms not belonging to the defined location types are assigned the “other location" label. Detailed annotation guidelines are available in the GitHub repository.[15]

Seven graduate-level students were trained to carry out the annotation task using the WebAnno NLP annotation tool.[16] We selected the WebAnno tool as it supports Unicode right-to-left languages (e.g., Arabic). Furthermore, to ensure the quality of annotations, we selected the annotators to be either citizen or have a good familiarity with the country of the disaster event. Finally, all annotators had to pass a quiz of 20 tweets before being eligible to start the annotation task.

Disagreements between annotators were examined by an additional meta-annotator and resolved. In Table 4.10 column “LMs (uniq)”, we show the total number of annotated LMs and unique LMs in parentheses. The unique number of LMs changes

[15] https://github.com/rsuwaileh/IDRISI/tree/main/LMR/annotation_guidelines
[16] https://webanno.github.io/

depending on the granularity of the affected area. On average, 26% LMs are unique.

4.3.2. Description and Quality

In this section, we present a thorough evaluation of IDRISI-RA dataset in terms of reliability, consistency, coverage, and diversity.

4.3.2.1. Reliability

To evaluate the reliability of annotations, we computed Cohen's Kappa [156] for both annotation tasks separately and jointly. Results in Figure 4.7, show the average reliability achieved is 83% (almost perfect), 67% (substantial), 70% (substantial), for LOC (i.e., toponym identification task), TYPE (i.e., location type assignment), and LOC+TYPE, respectively. All events show high-quality annotations, except the "Hafr Floods 2019" event with 12% agreement for the TYPE task (slight reliability) and 44% for LOC+TYPE (moderate reliability). Upon investigation, we found that "Hafr Albatten" is the most frequent LM in the dataset; one annotator assigns "city" type for all occurrences and the other assigns "province." While both annotators are correct (as in the Arab world, both types are used interchangeably), we anticipate an increase in the agreement level when accepting both types. Furthermore, the COVID-19 event shows slight agreement for the TYPE task for similar reasons across Arab countries.

4.3.2.2. Coverage and Diversity

In this section, we discuss how IDRISI-RA satisfies the properties presented in Section 4.1.

Geographical Coverage: To ensure that IDRISI-RA can train geographical generalizable models, it has to cover a wide geographical Arab area (refer to **O1** in Section 4.1).

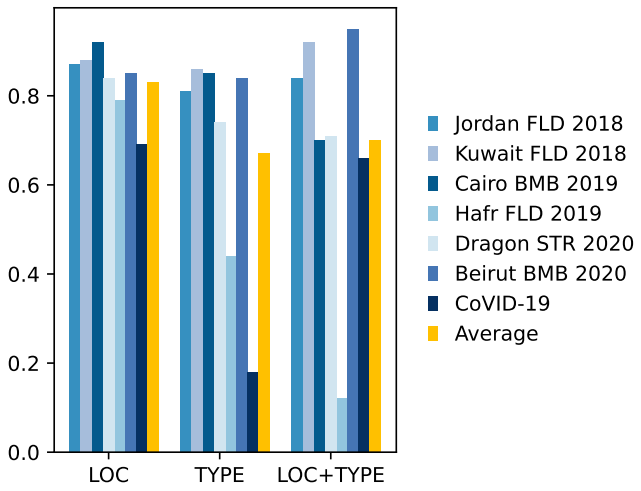


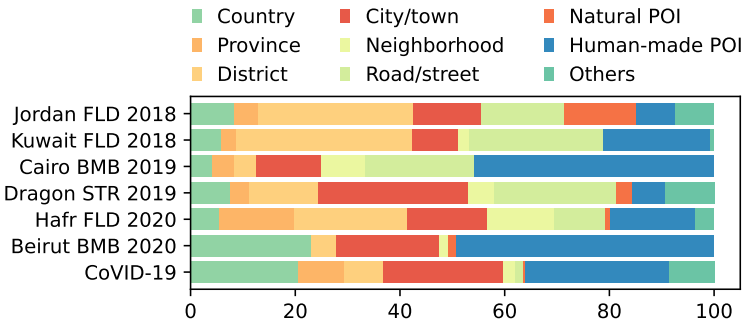
Figure 4.7. The Inter Annotator Agreement using Cohen's Kappa for IDRISI-RA per event. 0.2, 0.4, 0.6, and 0.8 indicate the degree of reliability as slight, fair, moderate, and substantial, respectively.

IDRISI-RA covers five different Arab countries, namely Jordan, Kuwait, Egypt, Saudi Arabia, and Lebanon. Additionally, the whole Arab region is represented by the COVID-19 pandemic event. Figure 4.8a shows the distribution of distinct LMs per location type.

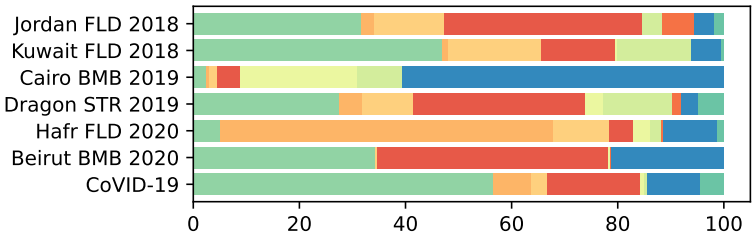
Domain Coverage: To remedy the lack of diversity in disaster types (refer to **O2** in Section 4.1), IDRISI-RA represents the most happening disaster types in the Arab region discussed over Twitter [138], [157]–[159], including three floods, two explosions, one storm, and the global COVID-19 pandemic.

Location Types Coverage: We labeled the dataset for coarse- and fine-grained location types to support advanced development and finer evaluation of LMR models. Figure 4.8b shows the distribution of the location types in IDRISI-RA. The coarse-grained (e.g., Country, State, and City) LMs dominate the dataset due to Kawarith collection strategy that depends on tracking relevant keywords, mostly hashtags which are the names of the coarse-grained affected areas by the disaster event.

Temporal Coverage: Covering long critical periods of disaster events allows the response authorities to operate efficiently (refer to **O5** in Section 4.1). For example, IDRISI-RA



(a) Distinct Location Mentions.



(b) All Location Mentions.

Figure 4.8. Distribution of location types in IDRISI-RA.

covers recent disaster events between 2018-2020. The period for events is approximately 8.8 days, on average (refer to Table B.3 in Appendix B). In Figure B.2 in Appendix B, we depict the number of tweets during two events showing the coverage of important developments.

Dialectical Distribution: To analyze the distribution of dialects vs. MSA in IDRISI-RA, we employed the *ASAD* dialectical classifier [160]. We found that around 86.8% of the tweets in the dataset are MSA. Table 4.7 shows the dialectical distribution of around 13% tweets. The largest portion goes to the Egyptian dialect, as Cairo BMB 2019 and Dragon STR 2020 happened in Egypt. The next dialect is Kuwaiti because Kuwait FLD 2018 event contains the second top number of tweets. Qatari and Saudi dialects are very close to the Kuwaiti dialect, explaining their prevalence.

Table 4.7. The dialects distribution in IDRISI-RA. The 18 countries are represented by their 2-letter ISO codes.

EG	KW	QA	SA	MA	LB
27.2%	26.6%	8.7%	7.4%	4.0%	3.8%
PS	BH	JO	LY	AE	SD
3.6%	3.4%	3.2%	2.3%	2.1%	2.1%
TN	DZ	OM	YE	IQ	SY
1.7%	1.1%	1.1%	0.8%	0.6%	0.6%

4.3.3. Benchmarking Experiments

To provide baselines for the LMR task, we benchmark IDRISI-RA dataset for different task, data, and disaster domain setups (to avoid redundancy, refer to Section 4.2.3 for details). We report the detailed statistics in Table B.3 in Appendix B.

4.3.3.1. Learning Models

We have developed our own LMR models:

- CRF$_{LMR}$ [154]: The Conditional Random Fields (CRF$_{LMR}$) is a competitive probabilistic tagging algorithm. We used word syntactic features, including the suffix, POS tag, and context (adjacent words and their syntactical features).
- BERT$_{LMR}$: We selected MARBERT model [161] for its superiority in Arabic Twitter NER when used for embeddings [162].

We tune the hyper-parameters of these two models (refer to Section 4.2.3.2).

We further employ two standard Arabic NER models for benchmarking, as described below.

- CAMELBERT-MIX (CML) [163]: An NER model trained on ANERcorp dataset, including MSA, Dialectal Arabic (DA), and Classical Arabic (CA) data.
- FARASA (FRS) [164]: A commonly used NER model in Farasa Arabic tool.

4.3.3.2. Results and Discussion

Table 4.8 presents the F_1 results of all setups over IDRISI-RA (refer to Section 4.2.3.3 for more details about the evaluation measures) of the adopted NER and our LMR models below. Detailed results are presented in Appendix C.

Type-less LMR: In this setup, the LMR models are only required to recognize LMs, regardless of their types. BERT$_{LMR}$ model achieves the best performance for both *random* and *time-based* scenarios. Next in order are the CRF$_{LMR}$, FARASA (FRS), and CAMeLBERT-Mix (CML). Although the CAMeLBERT-Mix is considered a BERT-based model, it shows poor performance compared to BERT$_{LMR}$, as it was fine-tuned on news wire documents for the NER (entities include LOC, ORG, PER, and MISC) task.

Type-based LMR: BERT$_{LMR}$ is evidently the best model for the *random* data setup. We anticipate the reason behind the lower performance of the CRF$_{LMR}$ model to be the limited features used to train the Arabic version (refer to Section 4.3.3.1). The CRF$_{LMR}$ model exhibits comparable F_1 scores to the BERT$_{LMR}$ model in the *time-based* data setup. To answer **RQ12**, we confirm the need for specific-LMR datasets and models that can perform effectively over disaster tweets.

4.3.4. Generalizability

In this section, we empirically study the generalizability of IDRISI-RA dataset. For that, we employ the best LMR model, BERT$_{LMR}$ (refer to Section 4.3.3);from hereafter, we refer to it as "the model". We study three dimensions: *(i)* **generalizability to unseen events** regardless of their type and location, *(ii)* generalizability to unseen events of the same or different disaster types (**domain generalizability**), and *(iii)* generalizability to unseen events that happened in the same or different countries (**geographical**

generalizability).

4.3.4.1. Experimental Setups

We run our experiments under both *type-less* and *type-based* task setups for only *random* data setup. We tune the model's hyper-parameters for every setup (refer to Section 4.2.3.2). 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.

Domain generalizability: We examine the model's performance under cross- and in-domain transfer setups [18]. The "domain" in our experiments refers to the type of disaster event. IDRISI-RA dataset covers the four most occurring disaster types in the Arab region: flood, bombing, storm, and pandemic. A transfer data setup comprises a source-target pair, resulting in 16 runs.

Geographical generalizability: We examine the model's performance over events in different countries than the source dataset. IDRISI-RA covers five countries (refer to Table B.3 in Appendix B), besides the global COVID-19. A transfer data setup comprises a source-target pair, resulting in 42 runs after excluding the *target* runs.

4.3.4.2. Results and Discussion

Generalizability to unseen events: Table 4.9 shows the model's results. The results for the *type-less* LMR demonstrate the potential of IDRISI-RA dataset under the zero-shot setting. The difference against the target runs is mostly negligible. Due to the difficulty of the *type-based* LMR, the performance under zero-shot learning is significantly lower than the target runs. However, the zero-shot results are within a reasonable range (i.e., average F_1 0.88), demonstrating the effectiveness of models trained on IDRISI-RA. To answer **RQ13**, we confirm that training on IDRISI-RA generates generalizable Arabic

LMR models that achieve, on average, around 0.75 and 0.88 F_1 scores for the *type-less* and *type-based* LMR, respectively.

Table 4.9. The F_1 results for the MARBERT$_{LMR}$ model under *zero-* and *target* training setups.

Task setup	Type-less				Type-based			
Data setup	Random		Time-based		Random		Time-based	
Training setup	Zero	Target	Zero	Target	Zero	Target	Zero	Target
Jordan FLD 2018	**0.768**	0.765	**0.759**	0.751	0.900	**0.967**	0.907	**0.957**
Kuwait FLD 2018	**0.853**	0.848	**0.830**	0.829	0.956	**0.982**	0.955	**0.972**
Cairo BMB 2019	**0.642**	0.632	**0.651**	0.626	0.835	**0.992**	0.805	**0.991**
Hafr FLD 2019	0.761	**0.762**	**0.754**	0.747	0.786	**0.971**	0.763	**0.965**
Dragon STR 2020	0.809	**0.814**	**0.829**	0.825	0.941	**0.946**	0.947	**0.950**
Beirut BMB 2020	0.616	**0.633**	0.594	**0.603**	0.914	**0.936**	0.841	**0.851**
COVID-19	0.879	**0.883**	0.842	**0.853**	0.961	**0.972**	0.960	**0.964**
Average	0.761	**0.762**	**0.751**	0.748	0.899	**0.967**	0.883	**0.950**

Domain generalizability: Figure 4.9 illustrates the F_1 scores of the models over the target sets.

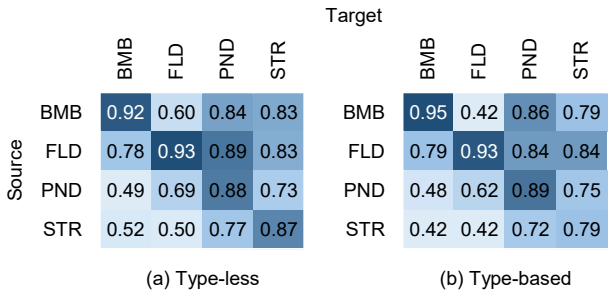


Figure 4.9. The F_1 results for the domain generalizability within IDRISI-RA under *random* data setup.

In-domain: Ideally, the best results should lay on the diagonal, which depicts the in-domain setup. This assumption holds for all runs, except the STR-to-STR runs in the *Type-based* LMR (Figure 4.9.b). Training on “bombing" data in the BMB-to-STR setup achieves comparable results to training on “storm" data in the STR-to-STR because both source and target data share the same or close affected areas (Egypt and Lebanon), which

could imply the overlap of toponyms' occurrences and patterns. The "bombing" (BMB) includes data from the *Cairo Bombing 2019* in Egypt and the *Beirut Explosion 2020* in Lebanon. The "storm" (STR) test data contains Dragon Storms 2020 that affected Egypt and Jordan, among a few LMs from Levantine Arabic. Moreover, the FLD-to-STR run achieves 6.3% better performance compared to the STR-to-STR run, as the FLD source data is approximately 7.5 times larger in size than the "storm" STR source. The effect of training dataset size on these results could be confirmed by the relatively low F_1 scores when the model trained on the "storm" data that has the smallest training data.

Cross-domain: Generally, the right upper part above the diagonal shows better results than the counterpart, except for the BMB-to-FLD, where the size of training data influences the results. We also note here that the model is tuned for every source-to-target transfer setup over the development splits; hence, the poor results on the test splits could indicate overfitting that prevents generalizability. This motivates the use of more advanced transfer learning techniques. Finally, to answer **RQ14**, we confirm that IDRISI-RA can generate acceptable domain generalizable models for most disaster types. It also provides challenging examples for the LMR models.

Geographical generalizability: Figure 4.10 shows the F_1 scores of the models over the target countries that are the same or different than the affected area of the source data. On average, the model achieves approximately 0.61 and 0.84 F_1 scores for *type-less* and *type-based* LMR, respectively. Due to its geographical coverage, the model achieves the top performance over "GL" target data (i.e., the COVID-19 event). To answer **RQ15**, we found that IDRISI-RA can generate reasonable geographically generalizable models.

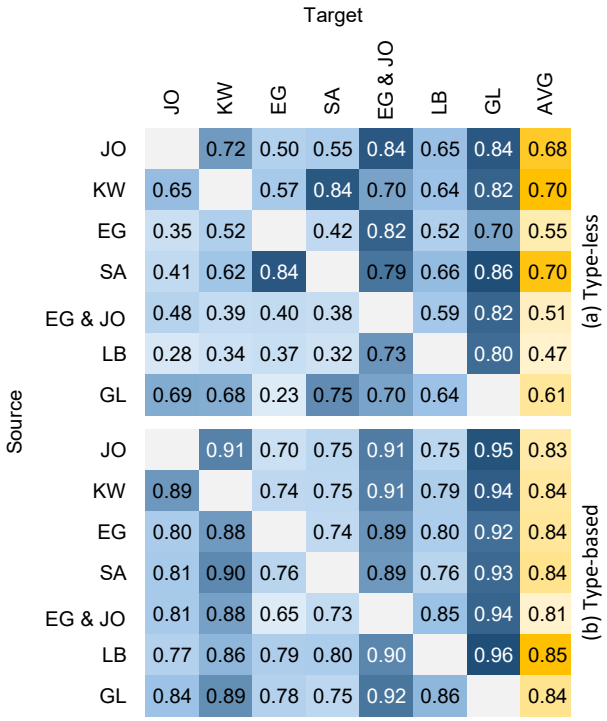


Figure 4.10. The F_1 results for the domain generalizability within IDRISI-RA under *random* data setup.

4.4. Silver Annotations

Thus far, we have discussed acquiring gold annotations using human workers. However, to increase the size of the dataset beyond our limited budget, we automatically amplify the size of IDRISI-R by using an automatic labeler, which is the best performing LMR models on the *gold* annotations (refer to Sections 4.2.3 and 4.3.3). More specifically, we trained BERT$_{LMR}$ models using the entire *gold* annotations (all events combined) of IDRISI-R and IDRISI-RA, separately. We then ran the resulting model on the tweets not sampled for the *gold* annotations from all disaster events, including the tweets that belong to the low informative classes. Out of this process, we constructed the largest automatically-labeled English and Arabic LMR datasets, comprising 56,682 and 1,205,373 tweets, respectively. We denote this version as *silver* to imply its level of

reliability and report its statistics in Table 4.10 (rows 2 and 4). We anticipate it to enable developing robust LMR models and support research on advanced learning techniques (e.g., transfer learning and domain adaptation).

Table 4.10. Tweet and Location Mention statistics of IDRISI-RE dataset.

Version	Tweets	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)
IDRISI-RA$_{gold}$	4,593	1,619	5,236 (918)
IDRISI-RA$_{silver}$	1,205,373	639,178	884,217 (18,609)

4.5. English and Arabic LMD Datasets and Benchmarks

In this section, we discuss the effort we made to extend the IDRISI-R datasets for the LMD task. First, we present the datasets' sampling and annotation in detail in Section 4.5.1. We then analyze the annotations in Section 4.5.2. We benchmark the datasets in the next chapter (Chapter 5, Section 5.4).

The research community lacks access to Twitter disaster-specific public LMD datasets, consequently preventing comparing existing studies. The only public English dataset is GeoCorpora [115]; however, it is a keyword-based dataset that misses the event context, which is important for disambiguating LMs. Furthermore, the dataset is limited to tweets containing the tracking keywords that may only appear in some informative tweets, which causes an information loss. Additionally, a fundamental limitation of GeoCorpora is the dominance of LMs from the United States (42%) and the United Kingdom (12%). More and above, up to the time of this writing, there are no Arabic LMD datasets.

In this section, we address these drawbacks and build IDRISI-D datasets[17] for

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

English (IDRISI-DE) and Arabic (IDRISI-DA) languages. IDRISI-DE is the largest-scale human-labeled tweet English dataset constituting 5,591 tweets and 9,685 LMs, 1,395 of which are unique. IDRISI-DA is the first public human-labeled Arabic dataset constituting 3,294 tweets and 6,445 LMs, 1,226 of which are unique. IDRISI-D datasets encompass all properties of IDRISI-R such geographical, domain, location granularity, temporal, informative, and dialectical (for Arabic) coverage.

Over and above, in an effort to alleviate the tweet sparsity issue (refer to Section1.2.1), we collect usefulness annotations for different features (e.g., hashtags, entities) and information sources (e.g., URLs) from the human-annotators.

To this end, we analytically answer the following research questions:

- **RQ16**: What features within tweets' textual content (replies, named entities, and other LMs) would enrich LM context for effective LMD?
- **RQ17**: What auxiliary information sources of information (hashtags, event context, and URLs) would enrich LM context for effective LMD?

The contributions of this section are as follows:

- We present IDRISI-DE, the largest *manually-labeled* public English LMD dataset of about 5,461 tweets.
- We present IDRISI-DA, the first Arabic LMD dataset containing about 2,909 tweets.
- We manually analyze the usefulness of different tweet features (hashtags, replies, named entities, and other LMs) and information sources (event context and URLs) for enriching the LM context for effective LMD.

4.5.1. Datasets Construction

In this section, we discuss the process of constructing IDRISI-D datasets. We start by describing the sampling in Section 4.5.1.1. We then discuss the annotation process in Section 4.5.1.2.

4.5.1.1. Dataset Sampling

Constrained by not overwhelming the *volunteered* annotators, we sampled a set of tweets from every disaster event while maintaining the distributions of LM types. This set of tweets went through an annotation pipeline of 3 phases. As fine-grained LMs are underrepresented in IDRISI-R (refer to Section 4.6) and we want to include all of them in IDRISI-D datasets, we carried out another annotation process for IDRISI-RE to include all fine-grained LMs those which were not sampled in the first annotation process. Due to a lack in budget, only one expert annotator conducted the latter annotation process. On the other hand, IDRISI-RA was entirely sampled for LMD annotations.

4.5.1.2. Dataset Annotation

The LMD annotation removes the ambiguity of geo/geo entities (as a sequel to the geo/non-geo LMR annotations). We collected the LMD annotations in 3 phases to increase reliability with a minimum load on the expert annotators:

PH 1 We selected two in-house volunteered annotators per event. The annotators are alumni who have a good familiarity with the country of the disaster event. When one of the in-house annotators' confidence level is low for a specific LM, or both annotators disagree, a meta annotator labels such cases in *Phase 2*.

PH 2 A meta annotator resolves the disagreement from *Phase 1* and labels the low confident examples. For that, she verifies the annotation by carefully searching OSM and Google. When she fails to disambiguate an LM, it goes to experts in *Phase 3*.

PH 3 Expert annotators disambiguate the hard unresolved LMs from *Phases 1* and *2*. Experts are residents of the countries where the disaster events took place.

In all phases, we asked annotators to (1) disambiguate the LMs, (2) assign a confidence score for their annotation, and (3) judge the usefulness of features and sources for disambiguation. For disambiguation (1), annotators search OpenStreetMap (OSM) gazetteer[18] after reading the tweet online and checking all relevant content, including the replies and the linked web pages. For the confidence level (2), annotators assign a score between 1-3 to show their confidence level. For the usefulness of features and information sources (3), we asked annotators whether the following features and information sources help understand the context of the LMs and resolve them:

- Event: The corresponding disaster event. Some NER tools consider the "Event" entity type (e.g., SpaCy tool). However, we limit the event notion to the corresponding event that the tweet discusses.
- Hashtags: The tweets posted on the same hashtags that appear within the tweet's text.
- Replies: The tweet thread or responses from the community.
- Other LMs: Other location mentions appear within the same tweet text.
- URLs: The linked web pages or media within the tweet text.
- Entities: Named entities that appear within the tweet text.

[18] `www.openstreetmap.org`

For every feature/source, annotators assign "Yes" if it is useful, "No" if it is useless, or "None" when it does not exist. We release the usefulness annotations and confidence scores within IDRISI-D datasets.

To avoid propagating human errors (refer to Section 4.6 for examples) from IDRISI-R datasets to IDRISI-D, we asked the annotators to modify LMs, add new LMs, or drop LMs in certain cases. In Table 4.13, we show example tweets and elaborate on them in the following:

- Adding new LMs: Crowd LMR annotators missed a few LMs. Hence, we allow the LMD annotators to add them if they are resolvable. For example, the "Pontagea Health Centre" in Tweet #1. We have added 27 LMs to IDRISI-DE while no LMs added to IDRISI-DA.
- Modifying LMs: Several cases require modification, such as decomposing addresses into address components (Tweet #2), separate multiple LMs (Tweet #3), fixing LM boundaries (Tweets #4-#6), merging LMs (Tweet #6), and fixing LM type (Tweet #8). We have modified 154 and 15 LMs in both IDRISI-DE and IDRISI-DA, respectively. The IDRISI-RA is cleaner than IDRISI-RE as it is in-house labeled.
- Dropping LMs: We asked annotators to drop LMs when they violate the LMR annotations guidelines. Cases include organization or person entities, nationalities, locational descriptions, among others. In Table 4.11, we show the types of wrong LMs dropped from IDRISI-D dataset and their statistics.
- Adding LMs to OSM: Annotators have added 27 and 171 LMs to OSM when they do not exist for both IDRISI-DE and IDRISI-DA, respectively.

Table 4.11. Error types of LMR annotations that were cleaned out in IDRISI-D

Error Type	IDRISI-DE	IDRISI-DA	Description
DESC	1044	0	Description of something related to the LOC
PER	124	0	Name or description of an individual person
ORG	337	10	Organization or campaign or group of people
NATION	145	0	Nationality or citizenship
AMBIG	110	18	Description of officially undefined location
MISSING	126	170	Valid LM that does not exist on OSM
ERROR	81	4	Conflicts IDRISI-R annotation guidelines
MULTIPLE	19	1	Referes to multiple locations (branches)
ALL	1,986	203	

We ran the task for ten weeks and obtained the final IDRISI-DE and IDRISI-DA datasets. We present the statistics of the resultant datasets in Table 4.12. Detailed statistics are presented Tables B.4-B.6 in Appendix B. "Coarse-grained" LMs include country, state, province, district, county, and city/town. "Fine-grained" LMs include neighborhood, road/street, and point-of-interest. "Others" refer to LMs that do not fall under coarse-grained and fine-grained types (e.g., villages).

Table 4.12. Tweet and Location Mention statistics of IDRISI-D dataset.

Dataset	Tweets	LMs (uniq)	Coarse-grained	Fine-grained	Others
IDRISI-DE	5,591	9,586 (1,601)	6,633	714	487
IDRISI-DA	2,869	3,893 (763)	2,326	1,506	54

Table 4.13. Examples of the annotations cases. Bold LMs are the wrong annotations in IDRISI-R. Gray-shaded LMs are the corrected version of LMs in IDRISI-D.

T#	Change	Tweet text
#1	Add	Pontagea Health Centre in **Beira**, #Mozambique, was partially destroyed by #CycloneIdai, with many services such as paediatrics and full maternity no longer available. Many medical supplies were lost or damaged.
#2	Modify: decompose	**High Springs Memorial Park, 17380 N.W. US Hwy 441.** Sandbags donations needed due to Santa Fe River flooding.
#3	Modify: separate	Please join us for Hurricane Maria relief this Saturday on Melrose St btwn **Buchwick & Broadway**. Every bit counts! #hurricanerelief #unidos
#4	Modify: offsets (multiple)	[user_mention] [user_mention] But if the victim is to be a non-agressor, then that still moralizes the deaths of all victims of terrorism, **9/11,California**,Sandy hook those people just were, ...
#5	Modify: offsets	Extremely heavy rains in lower **Shire River districts** of Chikwawa Nsanje in #Malawi's far south has been compounded by further rains from last week's #CycloneIdai. ...
#6	Modify: offsets	The University of **Nebraska Omaha** Love Your Melon Crew sure knows how to make kids happy - with potato chips and fruit (don't worry, other food was served)! Thank you for your continued support of #MealsThatHeal
#7	Modify: undefined	8AM #Maria update: Tropical Storm Warning in effect for the **Outer Banks of NC**. Good news, though, as central winds down to 75 mph.
#8	Modify: type	Amidst applause, Canadas rescue team arrives in **Mexico City Airport**$_{City \rightarrow POI}$ on Saturday #earthquake #CASDDA via [user_mention]
#9	Drop	**Rosen Hotels & Resorts** in Orlando announces availability of 30 guestrooms at [user_mention] for #HurricaneIrma evacuees. Call 407-996-9840.
#10	Drop	Fast-moving wildfires near #Athens have killed at least 76. #Europe has sweltered through an unusually hot and dry summer, breaking temperature records and fueling significant fires in several countries, including #**Sweden** and #Britain. ...

4.5.2. Description and Quality

4.5.2.1. Reliability

To evaluate the reliability of annotations, we measure the IAA for Phase 1 as it has more than two annotators using Cohen's Kappa [156]. We measure the IAA for the ability to resolve LMs. In other words, our classes are about whether an LM is

resolvable or not. We also compute the percentage of agreement for assigned toponyms from gazetteers by annotators. Tables 4.14 and 4.15 show the Cohen's Kappa and agreement percentages for IDRISI-DE and IDRISI-DA, respectively.

For IDRISI-DE, Table 4.14 shows Cohen's Kappa and agreement percentages. The average IAA is 0.83 (almost perfect). In detail, 15 events show almost perfect IAA (above 0.8), and the remaining events show substantial agreement (above 0.6). The agreement percentages for most of the events are above 85%. Exceptions are Mexico EQK and Pakistan EQK, which show approximately 68% and 76% agreement (substantial).

For IDRISI-DA, the results are presented in Table 4.15 for all events showing almost perfect IAA, except for Dragon STR 2020, which shows substantial IAA. On average, the dataset achieves around 90% IAA. Moreover, the agreement percentages for all events are above 90% (98% on average), which also confirms the quality of annotations.

These results statistically demonstrate the high quality of annotations for both IDRISI-DE and IDRISI-DA datasets. Furthermore, the disagreements were resolved in Phases 2 and 3 of annotation by meta-annotator and expert annotation, respectively, to ensure the quality of annotations with minimum cost.

Table 4.14. Inter-Annotator for Phase 1 annotation for IDRISI-DE per event. For Cohen's k, 0.2, 0.4, 0.6, and 0.8 indicate the degree of reliability as slight, fair, moderate, and substantial, respectively.

Event	Cohen's kappa	% Agreement
Ecuador EQK	0.91	97.41%
Canada FIR	0.87	96.71%
Italy EQK	0.89	98.24%
Kaikoura EQK	0.80	93.07%
HRC Matthew	0.80	98.21%
Sri Lanka FLD	0.69	89.47%
HRC Harvey	0.96	99.87%
HRC Irma	0.66	87.44%
HRC Maria	0.86	99.59%
Mexico EQK	0.92	67.97%
Maryland FLD	0.80	95.39%
Greece FIR	0.98	99.85%
Kerala FLD	0.87	98.91%
HRC Florence	0.89	91.36%
California FIR	0.83	97.28%
Cyclone Idai	0.82	98.27%
Midwestern U.S. FLD	0.80	99.68%
HRC Dorian	0.75	91.82%
Pakistan EQK	0.72	75.91%
Average	0.83	93.50%

Table 4.15. Inter-Annotator for Phase 1 annotation for IDRISI-DA per event. For Cohen's k, 0.2, 0.4, 0.6, 0.8 indicate the degree of reliability as slight, fair, moderate, and substantial, respectively

Event	Cohen's kappa	% Agreement
Jordan FLD 2018	0.95	99.01%
Kuwait FLD 2018	0.97	98.95%
Cairo BMB 2019	0.87	98.15%
Hafr FLD 2019	0.98	99.47%
Dragon STR 2020	0.75	93.69%
Beirut BMB 2020	0.81	97.68%
CoVID-19	0.95	98.93%
Average	0.90	97.98%

4.5.2.2. Usefulness Features

Tables 4.17 and 4.18 show the percentages of features' presence in IDRISI-DE and IDRISI-DA datasets, respectively. They also show the percentages of how useful are they. Apparently, the "Event", "Other LMs", and "Hashtags" are the most useful features for LMD. These features are more advantageous for fine-grained LMs.

Looking carefully at the usefulness annotations, we make different observations through examples in Table 4.16.

- Event: Usually, events are characterized by the geographical dimension (i.e., the affected area), which makes it recurrently contribute to narrowing the search space. This helps annotators in mitigating the "Toponymic homonymy" challenge. An exception to this are the LMs located outside the geographically affected area; however, these LMs are usually not of interest to the responders. In tweet #1, all results for "Corniche El Nile Street" are not within "Cairo" where the "Cairo BMB 2019" event occurred. Thus, searching toponyms within the affected area generates accurate annotations.

- Other LMs: This feature's usefulness is due to the geo-vicinity between co-occurring LMs with the same tweet. The geographical property between such LMs is usually inclusion and containment. Coarse-grained LMs are useful to disambiguate the finer-grained LMs that typically require tremendous effort. For instance, in tweet #2, "Nebraska" (State) was useful to disambiguate "Elkhorn River" (Human Point-of-Interest) to distinguish it from another part located in "West Virginia" (State). The usefulness percentages are lower than expected for this feature due to different reasons, e.g., the same LM is repeated in the same

tweet (Tweet #3).

- Hashtags: As most hashtags indicate the disaster event (e.g., #HurricaneHarvey), hashtags are equally important to the "Event" feature.

- Replies: We found a small number of tweets got interaction from the community. Hence, replies are only sometimes useful for LMD.

- URLs: They are sometimes useful when the linked web page elaborates on the geographical context of the reported information in the tweet. "Lake Butler" in tweet #4 is challenging LM. The annotator found an address in the linked Facebook post: "Lake Butler, FL, United States". Using this address, there are a matching administrative area called "Lake Butler" and a water feature called "Lake Butler". Using a comment on the Facebook post mentioning "Keystones Heights," which is closer to the administrative area "Lake Butler," the annotator successfully resolved this LM. The importance percentage of URLs is low as many of them are broken. Also, some of the linked articles require access subscription (Tweet #2).

- Entities: Are the nSamed entities that appear within the tweet text.

Table 4.16. Example tweets showing the usefulness of different features for the LMD annotation. Bold text indicates the LMs. The gray-shaded text indicates the features.

#	Useful features	Tweet text
#1	Event, Other LMs, Hashtag.	News websites quoted a security source as saying that when one of the speeding cars was driving in the opposite direction by mistake on **Corniche El Nile Street** in front of the #Cancer_Institute, it collided with 3 cars, which led to an explosion as a result of the collision.
#2	Other LMs	Human remains discovered along **Elkhorn River** after flooding, sheriff says https://buff.ly/2CEShla #Nebraska
#3	URL	In the wake of Hurricane Irma, we've planned a food distribution event in **Lake Butler** to help anyone affected by... fb.me/2fbe0b4YE
#4	None	Labatt to help those affected by Fort McMurray wildfire [...] **#FortMcMurray** #LCBO

Table 4.17. Statistics of the LMD features in IDRISI-DE dataset.

	Event	Hashtags	URLs	Replies	Other LMs	Entities
All LMs						
Exist	100.0%	63.9%	37.0%	0.4%	67.3%	31.2%
Doesn't exist	0.0%	36.1%	63.0%	99.6%	32.7%	68.8%
Fine-grained LMs						
Exist	100.0%	64.0%	34.3%	2.7%	65.5%	31.9%
Doesn't exist	0.0%	36.0%	65.7%	97.3%	34.5%	68.1%
Coarse-grained LMs						
Exist	100.0%	63.9%	37.2%	0.3%	67.7%	31.2%
Doesn't exist	0.0%	36.1%	62.8%	99.7%	32.3%	68.8%
All LMs						
Useful	98.4%	32.7%	3.9%	5.0%	38.3%	5.6%
Useless	1.6%	67.3%	96.1%	95.0%	61.7%	94.4%
Fine-grained LMs						
Useful	94.0%	54.7%	28.2%	0.0%	66.9%	12.3%
Useless	6.0%	45.3%	71.8%	100.0%	33.1%	87.7%
Coarse-grained LMs						
Useful	98.8%	30.9%	2.1%	32.1%	36.0%	5.1%
Useless	1.2%	69.1%	97.9%	67.9%	64.0%	94.9%

Table 4.18. Statistics of the LMD features in IDRISI-DA dataset.

	Event	Hashtags	URLs	Replies	Other LMs	Entities
All LMs						
Exist	100.0%	56.6%	41.9%	27.7%	42.7%	34.8%
Doesn't exist	0.0%	43.4%	58.1%	72.3%	57.3%	65.2%
Fine-grained LMs						
Exist	100.0%	77.5%	53.5%	59.8%	74.6%	63.8%
Doesn't exist	0.0%	22.5%	46.5%	40.2%	25.4%	36.2%
Coarse-grained LMs						
Exist	100.0%	50.6%	38.4%	17.8%	32.7%	25.8%
Doesn't exist	0.0%	49.4%	61.6%	82.2%	67.3%	74.2%
All LMs						
Useful	63.2%	22.2%	2.6%	0.9%	23.1%	2.0%
Useless	36.8%	77.8%	97.4%	99.1%	76.9%	98.0%
Fine-grained LMs						
Useful	89.8%	21.2%	3.6%	0.6%	19.8%	1.0%
Useless	10.2%	78.8%	96.4%	99.4%	80.2%	99.0%
Coarse-grained LMs						
Useful	54.4%	22.4%	2.0%	1.2%	24.8%	2.5%
Useless	45.6%	77.6%	98.0%	98.8%	75.2%	97.5%

To answer **RQ16**, our manual annotations confirm the utility of *other LMs* for effective LMD. The replies and named entities rarely appear in disaster tweets.

To answer **RQ17**, we confirm that the *event* and *hashtags* are the most useful sources for enriching the context of LMs for effective LMD. Indeed, the *hashtags* are typically used to refer to the disaster event (e.g., #HurricaneIrma).

The *other LMs*, *event*, and *hashtags* are more useful for disambiguating fine-grained LMs.

4.6. Limitations

Our thorough analysis shows shortcomings in the annotations of IDRISI that we discuss here.

- ***Underrepresented Fine-grained LMs***: Although we had chosen a careful sampling method focused on an event-centric informative dataset aiming to increase the likelihood of fine-grained LMs' occurrence [148], we think the low frequency of fine-grained LMs in IDRISI-RE and IDRISI-RA is a major limitation as they contain solely 9.77% and 25.5% fine-grained LMs, respectively.

- ***Human Errors***: Some human errors are made during LMR annotation due to the task's difficulty.

  - Annotators sometimes fail in distinguishing between *Location* and *Organization* entities (e.g., "Red Cross").
  - Different location types could be used interchangeably for the exact locations, which poses a difficulty for annotators (refer to Sections 4.2.2.1 and 4.3.2.1).
  - Annotators highlight the LMs when they appear as descriptions within the tweet's context.

  We fixed the majority of these errors as part of the Location Mention Disambiguation (LMD) annotation (refer to Section 4.5.1.2).

- ***Unlabeled Temporary Locations***: Although the temporary locations (refer to Section 1.2.3) are essential for the affected people and response authorities, not all of them are labeled in IDRISI.

- ***Unstudied Generalizability due to Absence of Arabic LMR datasets***: Due to the absence of *public* Arabic LMR datasets, we could not compare the generalizability of IDRISI-RA to any other datasets. Hence, we study the generalizability within IDRISI-RA for domain and geographical aspects.

- ***Ungeneralizable Conclusions for the LMD Usefulness Features***: We note that the conclusions we make on the usefulness of features and external information sources might not translate to other datasets and languages. Therefore, further empirical investigation is required to study the performance gains when employing the most useful ones for context expansion.

Table 4.4. 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$_{LMR}$	BERT$_{LMR}$	GPNE	GPNE$_2$	NTPR$_O$	NTPR$_R$	NTPR$_F$	LORE	NLORE
Ecuador EQK	0.866	**0.953**	0.242	0.741	0.840	0.920	0.921	0.653	0.632
Canada FIR	**0.732**	**0.732**	0.435	0.683	0.718	0.708	0.727	0.619	0.647
Italy EQK	0.558	**0.880**	0.730	0.214	0.828	0.851	0.863	0.200	0.167
Kaikoura EQK	0.878	**0.912**	0.594	0.730	0.787	0.906	0.896	0.711	0.756
HRC Matthew	0.890	**0.941**	0.141	0.923	0.862	0.915	0.929	0.857	0.882
Sri Lanka FLD	0.856	**0.917**	0.421	0.692	0.654	0.908	0.894	0.735	0.548
HRC Harvey	0.810	**0.906**	0.397	0.738	0.788	0.891	0.898	0.672	0.798
HRC Irma	0.773	**0.835**	0.369	0.713	0.704	0.814	0.801	0.651	0.735
HRC Maria	0.864	**0.925**	0.479	0.779	0.708	0.881	0.865	0.712	0.815
Mexico EQK	0.860	**0.929**	0.783	0.759	0.885	0.886	0.902	0.715	0.727
Maryland FLD	0.809	**0.890**	0.754	0.817	0.794	0.869	0.879	0.487	0.737
Greece FIR	0.839	0.927	0.792	0.730	0.807	**0.935**	0.929	0.694	0.686
Kerala FLD	0.725	**0.887**	0.664	0.480	0.718	0.863	0.873	0.430	0.441
HRC Florence	0.667	**0.755**	0.466	0.535	0.553	0.742	0.738	0.572	0.531
California FIR	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. FLD	0.904	**0.944**	0.680	0.785	0.772	0.929	0.920	0.706	0.716
HRC Dorian	0.820	**0.878**	0.589	0.757	0.760	0.870	0.858	0.616	0.722
Pakistan EQK	**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 4.5. 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$_{LMR}$	BERT$_{LMR}$	GPNE	GPNE$_2$	NTPR$_O$	NTPR$_R$	NTPR$_F$	LORE	NLORE
Ecuador EQK	0.716	0.916	0.164	0.703	0.854	**0.920**	0.874	0.640	0.563
Canada FIR	0.644	**0.767**	0.094	0.696	0.719	0.726	0.721	0.608	0.612
Italy EQK	0.504	**0.842**	0.357	0.276	0.777	0.770	0.768	0.234	0.232
Kaikoura EQK	0.755	0.896	0.169	0.693	0.769	**0.911**	0.879	0.723	0.731
HRC Matthew	0.790	0.944	0.045	0.862	0.866	0.936	**0.945**	0.872	0.892
Sri Lanka FLD	0.740	0.904	0.215	0.679	0.753	**0.919**	0.903	0.756	0.599
HRC Harvey	0.599	**0.894**	0.111	0.739	0.820	0.885	0.857	0.659	0.800
HRC Irma	0.538	**0.825**	0.111	0.722	0.683	0.805	0.813	0.668	0.732
HRC Maria	0.768	**0.904**	0.195	0.733	0.707	0.894	0.861	0.723	0.789
Mexico EQK	0.798	**0.911**	0.339	0.734	0.815	0.865	0.884	0.694	0.722
Maryland FLD	0.648	0.845	0.428	0.792	0.794	0.833	**0.892**	0.483	0.663
Greece FIR	0.778	**0.883**	0.389	0.767	0.777	**0.883**	0.842	0.706	0.684
Kerala FLD	0.638	**0.923**	0.273	0.575	0.786	0.909	0.888	0.530	0.553
HRC Florence	0.465	**0.784**	0.130	0.499	0.562	0.721	0.734	0.614	0.526
California FIR	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. FLD	0.792	**0.949**	0.440	0.789	0.819	0.927	0.930	0.746	0.754
HRC Dorian	0.470	0.862	0.137	0.767	0.791	0.833	**0.864**	0.548	0.639
Pakistan EQK	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

Table 4.8. The F_1 results for the LMR models on IDRISI-RA.

LMR setup	Type-less								Type-based			
Data setup	Random				Time-based				Random		Time-based	
Event	CML	FRS	CRF$_{LMR}$	BERT$_{LMR}$	CML	FRS	CRF$_{LMR}$	BERT$_{LMR}$	CRF$_{LMR}$	BERT$_{LMR}$	CRF$_{LMR}$	BERT$_{LMR}$
Jordan FLD 2018	0.517	0.650	0.843	**0.953**	0.491	0.641	0.776	**0.903**	0.837	**0.908**	0.775	**0.862**
Kuwait FLD 2018	0.320	0.688	0.711	**0.928**	0.294	0.625	0.644	**0.893**	0.904	**0.925**	**0.891**	0.879
Cairo BMB 2019	0.237	0.058	0.968	**0.989**	0.250	0.083	0.933	**0.936**	0.708	**0.975**	0.737	**0.931**
Hafr FLD 2019	0.303	0.286	0.838	**0.879**	0.319	0.276	0.829	**0.878**	**0.859**	0.856	**0.882**	0.838
Dragon STR 2020	0.579	0.737	0.698	**0.870**	0.615	0.702	0.611	**0.869**	**0.872**	0.787	**0.880**	0.714
Beirut BMB 2020	0.539	0.493	**0.873**	0.855	0.520	0.710	**0.772**	0.582	0.701	**0.813**	**0.621**	0.596
CoVID-19	0.238	0.845	0.640	**0.881**	0.266	0.800	0.634	**0.897**	**0.928**	0.893	**0.901**	0.886
Average	0.390	0.537	0.796	**0.908**	0.394	0.548	0.743	**0.851**	0.830	**0.880**	0.812	**0.815**

CHAPTER 5: LOCATION MENTION DISAMBIGUATION

The LMR task is generally defined as *the automatic linking of candidate location mentions in text to toponyms in gazetteers*. The scope of this chapter is limited from two angles; the disambiguation features are limited to the textual content of tweets, and more specifically disaster-related tweets posted during disaster events.

This chapter starts with formulating the LMD task in Section 5.1. Next, we discuss the proposed solution in Section 5.2. We then present the experimental evaluation in Section 5.3. We finally thoroughly discuss the results in Section 5.4.

Once the LMR system identifies the candidate location mentions, the next step is to resolve them into actual locations in a geo-positioning database (i.e., gazetteer); Location Mention Disambiguation (LMD). In other words, the system has to pin LMs on the map using a geographical representation such as the standard *geographic coordinate system* (GCS) or the *geocode system*. The GCS is a spherical coordinate system [165] that represents points on the earth using *longitude* and *latitude* angles measured with respect to the earth's center. The geocoding system represents geographical entities (points, lines, or polygons) using unique human-readable codes (or hashes) generated by dividing the geographic surface of the earth into grid cells at multi-level hierarchy. Figure 5.1 illustrates a high-level overview of the LMD task accompanied by an LMR component.

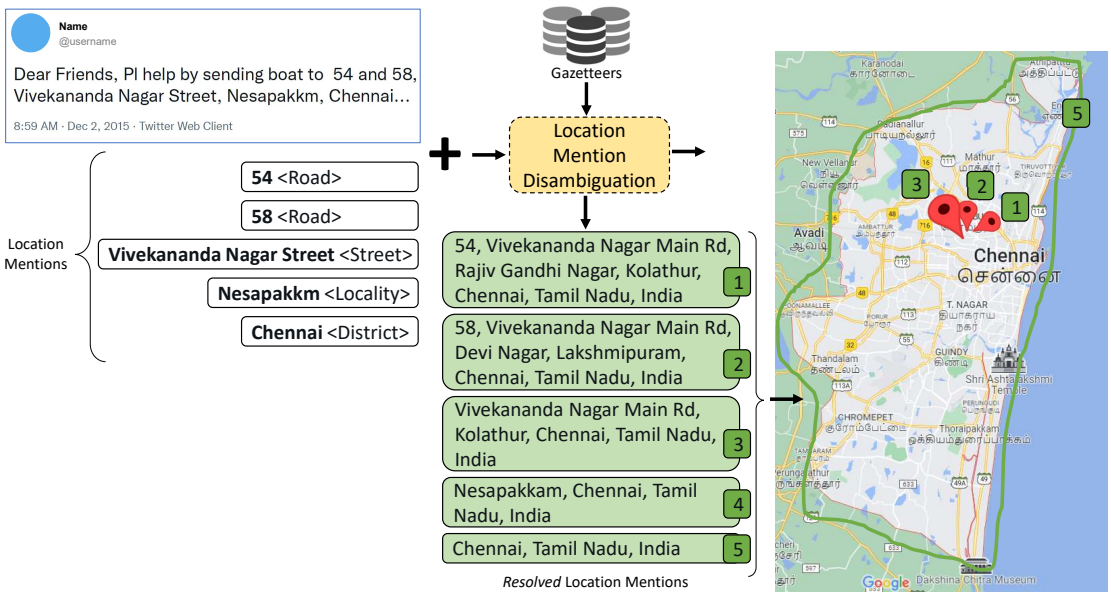


Figure 5.1. High-level overview of the LMD task

5.1. Problem Definition

To define the LMD task formally, we consider the following list of inputs:

- A tweet t that is related to a disaster event e
- A set of location mentions (LMs): $LM_t = \{lm_i; i \in [1, n_t]\}$ in tweet t, where lm_i is the i^{th} location mention and n_t is the total number of location mentions in t, if any.
- A geo-positioning database G (i.e., gazetteer) that consists of toponyms: $P = \{p_i; i \in [1, k]\}$, where p_i is the i^{th} toponym, and k is the number of location profiles existing in G. Different gazetteers may contain different properties for the same toponym, such as the name (in different languages), alternative names (e.g., exonyms), geo-coordinates (latitude and longitude), hierarchical address, location type (e.g., city, street, and POI), and other type-specific properties for different location types (e.g., “population" property for type “city").

Unlike the LMR task, LMD aims to resolve geo/geo ambiguity between candidate LMs extracted by the 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. We use a generic disambiguation terminology to cover all tasks' objectives.

The LMD system aims to match every location mention lm_i in the tweet t to one of the toponyms p_i in G that accurately represents lm_i, if exists. Otherwise, the system must abstain from prediction and declare the lm_i as irresolvable (or unlinkable). The irresolvable LMs are usually due to the incompleteness of existing crowdsourced digital gazetteers.

5.2. Disambiguation using BERT

Human annotators have prior accumulative knowledge that they exploit during the disambiguation labeling task. Such knowledge is unavailable for the LMD models that typically suffer from the cold start problem. Therefore, we employ the pre-trained model for disambiguation, BERT_{LMD}, in an attempt to account for efficiency and accelerate the model optimization for deployment in the disaster domain. Figure 5.2 depicts a high-level overview of our BERT_{LMD} model. Typically, LMD datasets contain the correct candidate toponyms extracted from gazetteers. To augment negative examples, we issue every gold LM against OpenStreetMap (OSM) online gazetteer and randomly pick a toponym that does not match it. We limit the negative examples to only one to balance the training data.

There are several toponym features in OSM, such as multilingual names, alternative names (e.g., exonyms), geo-coordinates (latitude and longitude), location type

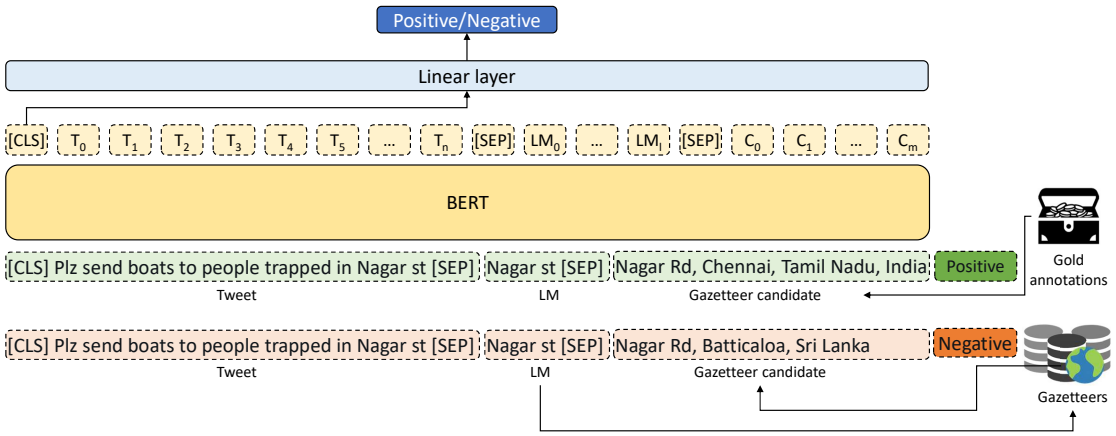


Figure 5.2. High-level overview of BERT$_{LMD}$ model (Training phase).

(e.g., city, street), address, population, place rank, and importance. In this approach, we limit the features of the candidate toponyms to textual features. Moreover, we only employ the full address of toponyms (dubbed "display_name" in OSM).

In this section, we answer the following research question for English (**RQ18**) and Arabic (**RQ19**) LMD: Do pre-trained models (BERT$_{LMD}$) perform more effectively for LMD task than the gazetteer retrieval-based (matching) approaches?

5.3. Evaluation Setup

In this section, we discuss the evaluation setup used to evaluate the proposed solution. In detail, we discuss the BERT$_{LMD}$ models in Section 5.3.1, the training and inference dataset in Section 5.3.2, the baselines in Section 5.3.3, and the evaluation measures and strategy in Section 5.3.4.

5.3.1. Hyper-parameter Tuning

For the BERT$_{LMD}$ model, we employ BERT-LARGE-CASED [17] and MARBERT [161] models for English and Arabic LMD models, respectively. Following the recommended values in [17], we tune the number of training epochs as 2, 3, or 4, and the learning rate

(Adam) as 5E-5, 3E-5, or 2E-5. We only remove diacritics from Arabic tweets and do not apply any further preprocessing.

5.3.2. Dataset

We use IDRISI-D datasets for evaluation (refer to Section 4.5 for details). Due to the imbalanced distribution of location types across the train, development, and test splits of IDRISI-R datasets, we used stratified sampling to repartitioned IDRISI-D datasets per event. As a result, we obtained fair distribution of location types across splits. To provide the learnable models with sufficient training data, we merged all events in each dataset (IDRISI-DE and IDRISI-DA).

5.3.3. Baselines

We have used retrieval-based and heuristic-based off-the-shelf LMD baselines.

- NOMINATIM [166]: A tool to search OSM data by name and address.
- GEOLOCATOR2 [66]: Off-the-shelf LMP system, which observed that the geographical coherence is effective in the location disambiguation task. It considers the hierarchy of location mentions in tweets when resolving them.
- GEOLOCATOR3 [66]: An improved version of CMU-geolocator that uses the population to post-filter retrieved results from Nominatim.
- GEOPARSEPY [15]: A trained SVM model on gazetteer-based features, including location type, population, and alternative names.

We use baselines for IDRISI-DE as they are originally developed for the English language. We use only NOMINATIM and GEOLOCATOR for IDRISI-DA as GEOPARSEPY cannot handle Arabic text. These LMD models are not disaster-specific, except GEOP-

ARSEPY. We could not employ the disaster-specific LMD models (i.e., DLocRL) as they are not public or open source, except GEOPARSEPY. Furthermore, re-implementation is not handy due to lacking technical details about these systems and nonpublic evaluation datasets.

5.3.4. Evaluation Measures and Strategy

To evaluate the effectiveness of the LMD models, we compute the Mean Reciprocal Rank ($MRR@r$) measure at different cutoff (ranks). We currently set $r = 1$,[1] but we can evaluate the LMD systems with different ranks when we perceive the task as a ranking problem. We exclude the distance-based methods (refer to Section 2.5.2.2) as tuning the distance threshold d needs further empirical investigation for different location types. We keep this for future work. Alternatively, we can leniently evaluate systems using hierarchical evaluation. Inspired by the evaluation of Twitter user geolocation [167], we evaluate the systems at different granularity, including country, state, county, city, district, neighborhood, street, and point-of-interest. Table 5.1 shows the address components involved in every evaluation level. We note that we do not evaluate at DISTRICT and NEIGHBORHOOD levels as corresponding location components are rarely filled in OSM or have a variety of type names (districts are sometimes classified as counties).

[1]The $MRR@1$ is equivalent to the accuracy measure for classification since for every LM, we have only one correct toponym.

Table 5.1. Evaluation levels and their corresponding location address components.

		Address components					
		Country	State	County	City	Street	Poi
Evaluation Levels	Country	✓					
	State/Province	✓	✓				
	County/District	✓	✓	✓			
	City/Town	✓	✓	✓	✓		
	Street	✓	✓	✓	✓	✓	
	Poi	✓	✓	✓	✓	✓	✓

5.4. Results and Discussion

In the section, we answer the research questions: Do pre-trained models (Bert$_{LMD}$) perform effectively for LMD task compared to the retrieval-based (matching) and heuristic-based approaches, for English (**RQ18**) and Arabic (**RQ19**) LMD?

5.4.1. English LMD

Table 5.2 shows the results of the Bert$_{LMD}$ model over IDRISI-DE and against the baselines. It is worth mentioning that Geolocator and Geoparsepy baselines rely on searching gazetteers and applying post-filters. It is evident that these post-filters are not effective for all evaluation levels, except for the Country level, and the raw results from Nominatim are more accurate. Geolocator systems show the best results for the Country level, but their performance decreases against the Bert$_{LMD}$ model at finer evaluation levels including State, City, Street and Poi, but not County. Nominatim is the top model at almost all evaluation levels. The Bert$_{LMD}$ model managed to compete with it at only the Poi evaluation level, which counts for the Bert$_{LMD}$ as the fine-grained LMs are of interest to the response authorities in the disaster domain [8]. The results confirm that disambiguating fine-grained LMs is more challenging than

coarse-grained LMs.

To answer **RQ18**, we confirm that the pre-trained BERT$_{LMD}$ model outperforms baselines at fine-grained evaluation levels. However, NOMINATIM is highly competitive for English LMD models.

Table 5.2. The results for the LMD models on IDRISI-DE dataset.

System	COUNTRY	STATE	COUNTY	CITY	STREET	POI
GEOLOCATOR2	**0.851**	0.601	0.316	0.244	0.022	0.015
GEOLOCATOR3	0.825	0.608	0.309	0.235	0.022	0.015
GEOPARSEPY	0.642	0.316	0.141	0.090	0.000	0.000
NOMINATIM	0.809	**0.663**	**0.379**	**0.355**	**0.244**	**0.073**
BERT$_{LMD}$	0.734	0.612	0.294	0.275	0.144	**0.073**

Table 5.3. The results for the LMD models on IDRISI-DA dataset.

System	COUNTRY	STATE	COUNTY	CITY	STREET	POI
GEOLOCATOR2	**0.454**	0.079	0.000	0.027	0.000	0.006
GEOLOCATOR3	0.443	0.073	0.000	0.021	0.000	0.006
NOMINATIM	0.430	0.220	0.029	0.165	0.130	0.106
BERT$_{LMD}$	**0.454**	**0.492**	**0.100**	**0.338**	**0.423**	**0.274**

5.4.2. Arabic LMD

Table 5.3 shows the results of BERT$_{LMD}$ models over IDRISI-DA and against the baselines. Similar to the English results, the GEOLOCATOR systems show high performance at COUNTRY level. However, their performance is comparable to the BERT$_{LMD}$ model. GEOLOCATOR systems fail at the fine-grained evaluation levels as they employ the GeoNames gazetteer that does not support Arabic for fine-grained locations. The NOMINATIM baseline is showing the best results among baselines, but it fails to outperform the BERT$_{LMD}$ at all evaluation levels.

To answer **RQ19**, we confirm that the pre-trained Arabic BERT$_{LMD}$ model wins by a large margin against all existing LMD systems, except at the COUNTRY evaluation

level where GEOLOCATOR2 shows comparable performance. This win counts for the Arabic language being a low-resource language; there is a lack of digital gazetteers that widely and granularly cover the Arab world.

The pre-trained BERT$_{LMD}$ models are promising, but they still require further improvements. The future directions for the LMD are two-fold: (i) enhancing the representation of LMs and toponyms, and (ii) employing advanced learning algorithms.

As for enhancing the representation of LMs, the manual annotations illustrated the usefulness of event context, hashtags, and other LMs for LMD. Therefore, LMD models could utilize these features to expand the LM context for effective disambiguation. As for enhancing the representation of toponyms, while we limit our study to the textual representation of toponyms, different features can be employed from gazetteers such as geo-coordinates (latitude and longitude), location type (e.g., city, street), population, among others.

On the other hand, employing other learning models would allow exploiting other types of features than the textual features. For instance, advanced algorithms, such as reinforcement learning, are worth exploring.

CHAPTER 6: CONCLUSION

In this chapter, we conclude with a summary of this dissertation (Section 6.1). We then thoroughly discuss this dissertation's theoretical, practical, and research implications in Section 6.2. We finally list our research outcomes in Section 6.3 and future directions in Section 6.4.

6.1. Conclusion

This dissertation contributes towards a crucial task, i.e., *Location Mention Prediction* in the crisis management domain. To sum up, we explored two main factors that influence the robustness of an LMP system: (i) the dataset used to train the model, and (ii) the learning model. As for the learning models, we introduce the state-of-the-art LMR models over English and Arabic disaster tweets. We further introduce competitive and state-of-the-art LMD English and Arabic models, respectively.

As for the training dataset, we formulated several research questions for which evidence-based answers were unknown. We designed an extensive and reliable experimental setup where several experiments investigate training effectiveness on general-purpose NER datasets from news articles and tweets. We demonstrate how the performance of an LMR model varies when trained on formal language (new articles) compared to informal language (tweets) as well as when trained on past disasters while considering the type, geoproximity, and language of the source and target disasters. Our findings suggest that Twitter-based NER labeled data is preferred over general-purpose data, and crisis-related labeled data is preferred over general-purpose Twitter data. Furthermore, our results suggest that training on disaster events data from similar types or geographically-nearby events to the target event improves performance compared to

training on different types or distant events. We further show how training on previous disasters of different languages than the target provides reasonable performing models that can be improved with little training from the target. Moreover, out of our investigation on the minimum number of tweets to label form the target event, we recommend labeling around 500 tweets and combining them with all available data to obtain an LMR model that achieves greater than 85% F_1 score. Overall, our findings shape the future directions in this line of research.

We introduced IDRISI-R datasets. IDRISI-RE is the large-scale *Location Mention Recognition* Twitter dataset comprising around 20k human-labeled and 57k machine-labeled tweets from 19 disaster events. The annotations include spans of location mentions in tweets' content and their geographical types, such as country, state, city, and street. The dataset is domain diverse and covers several countries across continents. Additionally, we benchmark IDRISI-RE using 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.

IDRISI-RA is the first Arabic LMR Twitter dataset. It contains 22 disaster events of different types that happened in the Arab region. We manually- (gold) and automatically (silver) annotated about 4.6K and 1.2M tweets. Both versions contain location mentions annotations and location types annotations. Our analysis showed that IDRISI-RA is second to none in empowering research for Arabic LMR. We confirm

the need for developing LMR-specific models for the disaster domain through extensive experiments using NER Arabic models. The developed LMR baselines are simple yet competitive ones. The results also demonstrated the decent generalizability of IDRISI-RA.

We extended further extended IDRISI-R datasets with LMD annotations and introduced IDRISI-D datasets. Both IDRISI-RE English and IDRISI-RA Arabic datasets are labeled for feature and information source usefulness. The analysis of the manual annotations showed that the event context, hashtags, and other location mentions appearing within the same tweet are helpful for accurate disambiguation.

As for the learning models, we adopted the pre-trained BERT model for both LMR and LMD tasks to compact all challenges for LMR and LMD. Our extensive LMR experiments under different task and data setups testify BERT_{LMR} as the state-of-the-art LMR model over both IDRISI-RE and IDRISI-RA datasets. Moreover, our experiments confirm that the BERT_{LMD} model is competitive over IDRISI-DE dataset and provides a state-of-the-art performance over IDRISI-RA dataset.

We provide all the resources and tools for the community to empower the development of LMP systems in the disaster management domain.

6.2. Implications

Compared to the public datasets, IDRISI datasets provide the largest, domain diverse, geographically representative, temporally representative, and informative LMR and LMD datasets. They also contain annotations for different coarse- (e.g., country, city) and fine-grained (e.g., street, POI) location types. In addition to that, the extensive empirical generalizability analysis showed that IDRISI are the best *domain* and

geographical generalizable LMP datasets. All these advantages of IDRISI cultivate the basis for empowering research on LMP in the disaster response and management domain. Indeed, the resources and models presented in this dissertation are useful for other contexts and domains.

In this section, we describe the theoretical (Section 6.2.1), practical (Section 6.2.2), and research (Section 6.2.3) implications of releasing IDRISI.

6.2.1. Theoretical Implications

While responders need to obtain all useful information that supports managing emergencies effectively and efficiently, the geographical context enables a better understanding of the development of disaster events and the affected people's behavior at the events' onset. For example, the geographical information is beneficial in creating diverse crisis maps (refer to Section 6.2.3). Making IDRISI public enables developing and evaluating generalizable LMP 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 datasets are designed to meet seven objectives, whose value we elaborate on in the following:

1. ***Geographical coverage***: Deploying geographically generalizable LMP models at the onset of disaster events require data that cover broad geographical areas. While IDRISI cover 22 English-speaking and the Arab world, they support 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 impacted people's behavior at different location granularity.

2. ***Domain coverage***: Similarly, building domain generalizable LMP models that are ready for deployment at the onset of the disaster events of any type (e.g., flood, earthquake) requires training them on data that is collected during diverse disaster events. The domain diversity of IDRISI datasets enable 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 dataset offer not only LM annotations but also location type annotations, they enable the development and evaluation of robust LMP models that aid in drawing situational-awareness, assessing the disaster impact, managing resources, and deploying relief plans, *at different location granularity*.
4. ***Large-scale***: The trainable LMP models, especially the deep learning-based models, require large training datasets to perform accurately. Thus, IDRISI, as being the largest to date and most generalizable datasets, support the responders to understand better the disaster events and the behavior of the impacted people.
5. ***Temporal coverage***: As IDRISI datasets cover the critical periods of the disaster events, they help different response authorities to better understand the disaster events and the behavior of affected people during different disaster phases (pre-disaster, during disaster events, and post-disaster).
6. ***Relevance and informativeness***: Providing the geographical context to only informative content, after discarding noise, is a high priority to aid the response authorities in understanding the updates of the disaster events on the ground. As IDRISI datasets solely contain informative tweets, they provide more realistic

data for training LMP models ready for direct integration in real-world information processing systems for disaster management.

7. ***Dialectical coverage***: As IDRISI Arabic datasets contain various Arabic dialects, they are suitable for training robust LMP models that generalize to unseen events happening in the Arab world.

Moreover, while all these design factors are important, our conclusions emphasize the influence of geographical coverage and data size for creating generalizable LMP datasets (refer to Sections 4.2.4 and 4.3.4).

6.2.2. Practical Implications

Using IDRISI datasets enable the deployment of different surveillance and decision-support systems during disaster events used by different response authorities. These systems employ the underlying applications discussed in Section 6.2.3 and generate reports at different location granularity for different phases of the disaster. These reports could be in the 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: These maps visualize and identify the most impactful incidents, such as infrastructure damage, power outage, and facilities closure, among others. They also help response authorities to manage relief activities and plan recovery.

Eyewitnesses maps: These maps locate eyewitnesses and first responders helping 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 near the incidents' locations.

Resources maps: These maps locate 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 on the need.

Population mobility maps: These maps aid in evacuating the vulnerable people away from the affected areas as they help in monitoring their movement in real-time, which in turn facilitates studying 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 can be utilized to build automatic domain and geographically generalizable LMP models that perform at acceptable accuracy levels.

6.2.3. Research Implications

IDRISI datasets enable research in different computational tasks, such as event/incident detection, relevance filtering, and geolocation tasks, to name a few. In addition to that, as they cover different types of disaster events, we anticipate them 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 [168]. Fortunately, people tend to mention where events/incidents take place when reporting them [124]. Harnessing the relation between the occurrence (e.g., peaks) of LMs in tweets and the likelihood of events and inci-

dents happening can aid early prediction and detection. For example, [169] and [170] proposed content analysis of tweets by extracting locations for event/incident detection.

Relevance filtering: A pivotal barrier to exploiting social media for crisis management is the noisiness of data which necessitates the need for automatic relevance filtering methods [171]. Prior studies show that the geographical references in social media messages could indicate their relevance and informativeness [30], [172]. Kaufhold, Bayer, and Reuter [173] achieved the best performance when incorporating location-related features in their rapid classification model. Thus, we anticipate IDRISI datasets 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 [174], (3) inferring user location [175], and (4) modeling user movement [176]. While all these tasks are crucial for crisis management, the LMP tasks, in particular, play an essential role in tackling all of them using text-based techniques [37]. For instance, combining extracted entities (e.g., LMs) from tweets and their relations inferred from a Knowledge-base leads to a noticeable improvement in the *user location prediction* model [177].

Displacement monitoring: Internal and cross-border displacement is a terrible consequence of crises. By early May 2019, displaced people reached about 41.3 million due to conflicts and violence.[1] 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 datasets support 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 [79]. IDRISI datasets serve the GIR retrieval techniques that rely

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

on detecting locations and spatial references in queries and documents [178]. Thus, the large size of IDRISI datasets provide a promising resource for augmenting spatial information of tweets for geographical indexing and retrieval over the Twitter streams. Additionally, as IDRISI datasets are characterized by their wide geographical coverage, we anticipate them to be a representative resource for Geographical retrieval.

6.3. Outcomes

This dissertation resulted in five major publications, one computational challenge, and one tutorial:

- Journal articles:
  - **Reem Suwaileh**, Tamer Elsayed, and Muhammad Imran. IDRISI-RE: A Generalizable Dataset with Benchmarks for Location Mention Recognition on Disaster Tweets. Information Processing and Management. 2023.
  - **Reem Suwaileh**, Tamer Elsayed, Muhammad Imran, and Hassan Sajjad. When a Disaster Happens, We Are Ready: Location Mention Recognition from Crisis Tweets. International Journal of Disaster Risk Reduction (IJDRR). 2022.
- Conference full/long papers:
  - **Reem Suwaileh**, Muhammad Imran, and Tamer Elsayed. IDRISI-RA: The First Arabic Dataset with Benchmarks for Location Mention Recognition on Disaster Tweets. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23), Toronto, Canada, July 9-14, 2023.
  - **Reem Suwaileh**, Muhammad Imran, Tamer Elsayed, and Hassan Sajjad. Are We Ready for this Disaster? Towards Location Mention Recognition

from Crisis Tweets. Proceedings of the 28th International Conference on Computational Linguistics (COLING'20), pp. 6252–6263, Barcelona, Spain (Online), December 8-13, 2020.

- Book chapter: **Reem Suwaileh**, Tamer Elsayed, and Muhammad Imran. Role of Geolocation Prediction in Disaster Management. International Handbook of Disaster Research (IHDR). Springer, 2023.
- Hosted the first version of a task on *Location Mention Recognition from Social Media Crisis-related Text* in the GeoAI Challenge Launched by the International Telecommunication Union (ITU) with Muhammad Imran, Ehsan Ullah, Lokendra Chauhan, Ferda Ofli, and Tamer Elsayed, 2022.
- Tutorial on *Geo-tagging text documents* with Umair Qazi, Ferda Ofli, and Imran Muhammad (QCRI), in the Artificial Intelligence for Collective Intelligence (AI4CI) hosted by Qatar Computing Research Institute (QCRI) and the United Nations Development Programme (UNDP), 2022.

Other publications that are not directly contributing to the topic of this dissertation:

- Workshop papers:
  - Fatima Haouari, Maram Hasanain, **Reem Suwaileh** and Tamer Elsayed. ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP) at EACL 2021. pp. 72-81, Kyiv, Ukraine (Virtual), April 2021.
  - Fatima Haouari, Maram Hasanain, **Reem Suwaileh** and Tamer Elsayed. ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation

Networks. Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP) at EACL 2021. pp. 82-91, Kyiv, Ukraine (Virtual), April 2021.

- Conference papers:
  - Maram Hasanain, Yassmine Barkallah, **Reem Suwaileh**, Mucahid Kutlu and Tamer Elsayed. ArTest: The First Test Collection for Arabic Web Search with Relevance Rationales. Proceedings of the 43rd annual international ACM conference on Research and development in information retrieval: SIGIR '20, pp. 2017-2020, Virtual Event, China, July 2020.
  - Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, **Reem Suwaileh**, and Fatima Haouari. CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media. Proceedings of the 42nd European Conference on Information Retrieval (ECIR), Lisbon, Portugal, Lecture Notes in Computer Science, vol. 12035, pp. 499–508, Springer, Cham, April 2020.
  - Shahad Alshalan, Raghad Alshalan, Hend Al-Khalifa, **Reem Suwaileh**, Tamer Elsayed. Improving Arabic Microblog Retrieval with Distributed Representations. Proceedings of the 15 th Asia Information Retrieval Societies Conference (AIRS 2019), Hong Kong, China, November 2019.
  - Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, **Reem Suwaileh**, Giovanni Da San Martino, Pepa Atanasova. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. In: Crestani F. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019. Lecture Notes in Computer

Science, vol. 11696, pp. 301-321. Springer, Cham, 2019.

- Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, **Reem Suwaileh**, Giovanni Da San Martino, Pepa Atanasova. CheckThat! at CLEF 2019: Automatic Identification and Verification of Claims. Proceedings of the 41st European Conference on Information Retrieval (ECIR), Cologne, Germany, Lecture Notes in Computer Science, vol. 11438, pp. 309–315, Springer, Cham, April 2019.
- Maram Hasanain, **Reem Suwaileh**, Tamer Elsayed, Alberto Barrón-Cedeño, Preslav Nakov. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. Proceedings of CheckThat! Lab at the 10th International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF'19), Lugano, Switzerland, Sep 2019

- Book chapters:
  - Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, **Reem Suwaileh**, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In: Arampatzis A. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2020. Lecture Notes in Computer Science, vol. 12260. Springer, Cham, 2020.
  - Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, **Reem Suwaileh**, Giovanni Da San Martino, Pepa Atanasova. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification

of Claims. In: Crestani F. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019. Lecture Notes in Computer Science, vol. 11696, pp. 301-321. Springer, Cham, 2019. 3.

6.4. Future Directions

There are several directions for future work. We elaborate on some of them in the following:

LMP problem modeling: We have studied the LMR and LMD tasks independently. However, optimizing the LMR and LMD models jointly rather than in a pipeline architecture would allow passing feedback between models to optimize accordingly. This direction was explored for the LMP tasks outside the disaster domain [89].

Learning features: Our study is limited to utilizing textual features for both LMR and LMD. However, other features like metadata and social networks (e.g., followers and interactions) are worth investigating.

Unified and up-to-date geo-positioning databases: The key bottleneck of gazetteer-based LMR solutions and LMD systems is the choice of geo-positioning databases. There is much room for contribution regarding augmentation, aggregation, and maintaining up-to-date geo-positioning databases.

Evaluation: For LMR evaluation, in almost all existing studies, the evaluation is limited to exact matches with gold annotations. The studies that focus on partial matches are heuristics-based. We plan to further study the partial matches for LMR. For LMD evaluation, we plan to investigate proper ways to tune the distance threshold d in the distance-based evaluation measures.

Efficiency analysis The literature and our work focus on the effectiveness of

models. While the disaster domain is time-critical, we plan to profile models and further analyze their efficiency when integrated into information processing systems for disaster management.

Deployment: We also plan to deploy both LMR and LMD models into online information processing systems for disaster management.

REFERENCES

[1] S. E. Vieweg, "Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications," Ph.D. dissertation, University of Colorado at Boulder, 2012, ISBN: 978-1-2673-3596-8.

[2] *Five essentials for the first 72 hours of disaster response*, 2017. [Online]. Available: https://www.unocha.org/story/five-essentials-first-72-hours-disaster-response.

[3] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird, "From situational awareness to actionability: Towards improving the utility of social media data for crisis response," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, 2018. DOI: 10.1145/3274464. [Online]. Available: https://doi.org/10.1145/3274464.

[4] J. Ziemke, "Crisis mapping: The construction of a new interdisciplinary field?" *Journal of Map & Geography Libraries*, vol. 8, no. 2, pp. 101–117, 2012.

[5] I. Weber, M. Imran, F. Ofli, *et al.*, "Non-traditional data sources: Providing insights into sustainable development," *Communications of the ACM*, vol. 64, no. 4, pp. 88–95, 2021.

[6] A. L. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergency events," *International journal of emergency management*, vol. 6, no. 3, pp. 248–260, 2009.

[7] R. Grace, "Toponym usage in social media in emergencies," *International Journal of Disaster Risk Reduction*, vol. 52, no. July 2020, p. 101 923, 2021, ISSN:

22124209. DOI: 10.1016/j.ijdrr.2020.101923. [Online]. Available: https://doi.org/10.1016/j.ijdrr.2020.101923.

[8] J. Kropczynski, R. Grace, J. Coche, *et al.*, "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.*, Wellington, New Zealand, Nov. 2018, p.428–438. [Online]. Available: https://hal-mines-albi.archives-ouvertes.fr/hal-01987793.

[9] R. Grace, J. Kropczynski, and A. Tapia, "Community coordination: Aligning social media use in community emergency management," in *Proceedings of the 15th ISCRAM Conference*, 2018.

[10] S. McCormick, "New tools for emergency managers: An assessment of obstacles to use and implementation," *Disasters*, vol. 40, no. 2, pp. 207–225, 2016. DOI: https://doi.org/10.1111/disa.12141. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/disa.12141. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/disa.12141.

[11] C. Reuter, T. Ludwig, M.-A. Kaufhold, and T. Spielhofer, "Emergency services' attitudes towards social media: A quantitative and qualitative survey across europe," *International Journal of Human-Computer Studies*, vol. 95, pp. 96–111, 2016, ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2016.03.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581916000379.

[12] *The ushahidi platform*. [Online]. Available: www.ushahidi.com/.

[13] *Innovative uses of social media in emergency management*, application/pdf, [Online; accessed 30 March 2022]. [Online]. Available: www.hsdl.org/?abstract&did=805223.

[14] H. Al-Olimat, K. Thirunarayan, V. Shalin, and A. Sheth, "Location name extraction from targeted text streams using gazetteer-based statistical language models," in *Proceedings of the 27th International Conference on Computational Linguistics*, Aug. 2018, pp. 1986–1997.

[15] S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, "Location extraction from social media: Geoparsing, location disambiguation, and geotagging," *ACM Transactions on Information Systems*, vol. 36, no. 4, pp. 1–27, Jun. 2018, ISSN: 1046-8188.

[16] B. Han, P. Cook, and T. Baldwin, "Lexical normalization for social media text," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 1, pp. 1–27, Feb. 2013, ISSN: 2157-6904.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*, Jun. 2019, pp. 4171–4186.

[18] R. Suwaileh, M. Imran, T. Elsayed, and H. Sajjad, "Are we ready for this disaster? towards location mention recognition from crisis tweets," in *Proceedings of the 28th International Conference on Computational Linguistics*, Dec. 2020, pp. 6252–6263.

[19] R. Suwaileh, T. Elsayed, M. Imran, and H. Sajjad, "When a disaster happens, we are ready: Location mention recognition from crisis tweets," *International Journal of Disaster Risk Reduction*, p. 103 107, 2022.

[20] S. R. Hiltz, A. L. Hughes, M. Imran, L. Plotnick, R. Power, and M. Turoff, "Exploring the usefulness and feasibility of software requirements for social media use in emergency management," *International Journal of Disaster Risk Reduction*, vol. 42, p. 101 367, 2020, ISSN: 2212-4209. DOI: https://doi.org/10.1016/j.ijdrr.2019.101367. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212420919311203.

[21] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Comput. Surv.*, vol. 47, no. 4, 2015, ISSN: 0360-0300. DOI: 10.1145/2771588. [Online]. Available: https://doi.org/10.1145/2771588.

[22] B. Poblete, J. Guzmán, J. Maldonado, and F. Tobar, "Robust detection of extreme events using twitter: Worldwide earthquake monitoring," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2551–2561, 2018. DOI: 10.1109/TMM.2018.2855107.

[23] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, *et al.*, "Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation," *Sensors*, vol. 19, no. 7, 2019, ISSN: 1424-8220. DOI: 10.3390/s19071746. [Online]. Available: https://www.mdpi.com/1424-8220/19/7/1746.

[24] M. Sreenivasulu and M. Sridevi, "Comparative study of statistical features to detect the target event during disaster," *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 121–130, 2020. DOI: 10.26599/BDMA.2019.9020021.

[25] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran, "Identifying sub-events and summarizing disaster-related information from microblogs," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 265–274.

[26] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[27] R. Mazloom, H. Li, D. Caragea, C. Caragea, and M. Imran, "A hybrid domain adaptation approach for identifying crisis-relevant tweets," *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, vol. 11, no. 2, pp. 1–19, 2019.

[28] L. S. Snyder, Y.-S. Lin, M. Karimzadeh, D. Goldwasser, and D. S. Ebert, "Interactive learning for identifying relevant tweets to support real-time situational awareness," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 558–568, 2019.

[29] X. Ning, L. Yao, B. Benatallah, Y. Zhang, Q. Z. Sheng, and S. S. Kanhere, "Source-aware crisis-relevant tweet identification and key information summarization," *ACM Transactions on Internet Technology (TOIT)*, vol. 19, no. 3, pp. 1–20, 2019.

[30] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "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*, 2010, pp. 1079–1088.

[31] A. M. MacEachren, A. Jaiswal, A. C. Robinson, *et al.*, "Senseplace2: Geotwitter analytics support for situational awareness," in *2011 IEEE conference on visual analytics science and technology (VAST)*, IEEE, 2011, pp. 181–190.

[32] S. Vieweg, C. Castillo, and M. Imran, "Integrating social media communications into the rapid assessment of sudden onset disasters," in *International Conference on Social Informatics*, Springer, 2014, pp. 444–461.

[33] M. Marbouti and F. Maurer, "Social media use during emergency response–insights from emergency professionals," in *Conference on e-Business, e-Services and e-Society*, Springer, 2016, pp. 557–566.

[34] H. Purohit, C. Castillo, M. Imran, and R. Pandev, "Social-EOC: Serviceability model to rank social media requests for emergency operation centers," in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2018, pp. 119–126, ISBN: 9781538660515.

[35] R. McCreadie, C. Buntain, and I. Soboroff, "Trec incident streams: Finding actionable information on social media," in *International Conference on Information Systems for Crisis Response and Management*, 2019, pp. 691–705, ISBN: 9788409104987.

[36] M. Basu, K. Ghosh, and S. Ghosh, "Information Retrieval from Microblogs During Disasters: In the Light of IRMiDis Task," *SN Computer Science*, vol. 1, no. 1, p. 61, 2020. DOI: 10.1007/s42979-020-0065-1.

[37] X. Zheng, J. Han, and A. Sun, "A Survey of Location Prediction on Twitter," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1652–1671, 2018, ISSN: 15582191. DOI: 10.1109/TKDE.2018.2807840. arXiv: 1705.03172v2. [Online]. Available: https://ieeexplore.ieee.org/document/8295255/%20https://doi.org/10.1109/TKDE.2018.2807840.

[38] C. Xu, J. Pei, J. Li, C. Li, X. Luo, and D. Ji, "DLocRL: A deep learning pipeline for fine-grained location recognition and linking in tweets," in *Proceedings of the World Wide Web Conference*, May 2019, pp. 3391–3397.

[39] R. D. Das and R. S. Purves, "Exploring the potential of Twitter to understand traffic events and their locations in Greater Mumbai, India," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 12, pp. 5213–5222, 2020, ISSN: 15580016.

[40] J. Wang, Y. Hu, and K. Joseph, "NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages," *Transactions in GIS*, vol. 24, no. 3, pp. 719–735, 2020, ISSN: 14679671.

[41] C. Reuter, "Crisis 2.0: Towards a systematization of social software use (ijiscram)," in *Emergent Collaboration Infrastructures: Technology Design for Inter-Organizational Crisis Management*. Wiesbaden: Springer Fachmedien Wiesbaden, 2015, pp. 35–48, ISBN: 978-3-658-08586-5. DOI: 10.1007/978-3-

658-08586-5_4. [Online]. Available: https://doi.org/10.1007/978-3-658-08586-5_4.

[42] C. Reuter, A. L. Hughes, and M.-A. Kaufhold, "Social media in crisis management: An evaluation and analysis of crisis informatics research," *International Journal of Human–Computer Interaction*, vol. 34, no. 4, pp. 280–294, 2018. DOI: 10.1080/10447318.2018.1427832. eprint: https://doi.org/10.1080/10447318.2018.1427832. [Online]. Available: https://doi.org/10.1080/10447318.2018.1427832.

[43] A. L. Hughes and R. Shah, "Designing an application for social media needs in emergency public information work," in *Proceedings of the 19th International Conference on Supporting Group Work*, ser. GROUP '16, Sanibel Island, Florida, USA: Association for Computing Machinery, 2016, pp. 399–408, ISBN: 9781450342766. DOI: 10.1145/2957276.2957307. [Online]. Available: https://doi.org/10.1145/2957276.2957307.

[44] H. Purohit, C. Castillo, M. Imran, and R. Pandey, "Social-eoc: Serviceability model to rank social media requests for emergency operation centers," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2018, pp. 119–126.

[45] K. C. Roy, S. Hasan, and P. Mozumder, "A multilabel classification approach to identify hurricane-induced infrastructure disruptions using social media data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 35, no. 12, pp. 1387–1402, 2020.

[46] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[47] L. Hong and V. Frias-Martinez, "Modeling and predicting evacuation flows during hurricane irma," *EPJ Data Science*, vol. 9, no. 1, p. 29, 2020.

[48] K. C. Roy and S. Hasan, "Modeling the dynamics of hurricane evacuation decisions from twitter data: An input output hidden markov modeling approach," en, *Transportation research part C: emerging technologies*, vol. 123, e102976–e102976, 2021, ISSN: 0968-090X. DOI: 10.1016/j.trc.2021.102976. [Online]. Available: http://dx.doi.org/10.1016/j.trc.2021.102976.

[49] O. Uchida, M. Kosugi, G. Endo, *et al.*, "A real-time information sharing system to support self-, mutual-, and public-help in the aftermath of a disaster utilizing twitter," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E99.A, no. 8, pp. 1551–1554, 2016. DOI: 10.1587/transfun.E99.A.1551.

[50] M. Kosugi, K. Utsu, S. Tajima, *et al.*, "Improvement of twitter-based disaster-related information sharing system," in *2017 4th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 2017, pp. 1–7. DOI: 10.1109/ICT-DM.2017.8275693.

[51] M. Kosugi, K. Utsu, M. Tomita, *et al.*, "A twitter-based disaster information sharing system," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019, pp. 395–399. DOI: 10.1109/CCOMS.2019.8821719.

[52] C. Zhang, C. Fan, W. Yao, X. Hu, and A. Mostafavi, "Social media for intelligent public information and warning in disasters: An interdisciplinary review," *International Journal of Information Management*, vol. 49, pp. 190–207, 2019, ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.

04.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0268401218310995.

[53] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[54] R. Reinanda, E. Meij, and M. de Rijke, "Knowledge Graphs: An Information Retrieval Perspective," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 4, pp. 1–158, 2020, ISSN: 1554-0669. DOI: 10.1561/1500000063. [Online]. Available: http://www.nowpublishers.com/article/Details/INR-063.

[55] J. Lingad, S. Karimi, and J. Yin, *Location extraction from disaster-related microblogs*. New York, New York, USA: Association for Computing Machinery, 2013, pp. 1017–1020, ISBN: 9781450320382. DOI: 10.1145/2487788.2488108. [Online]. Available: http://opennlp.apache.org%20http://dl.acm.org/citation.cfm?doid=2487788.2488108.

[56] J. Gelernter and S. Balaji, "An algorithm for local geoparsing of microtext," *Geoinformatica*, vol. 17, no. 4, pp. 635–667, Oct. 2013, ISSN: 1384-6175.

[57] P. Nand, R. Perera, A. Sreekumar, and L. He, "A multi-strategy approach for location mining in tweets: AUT NLP group entry for ALTA-2014 shared task," in *Proceedings of the Australasian Language Technology Association Workshop 2014*, Melbourne, Australia, Nov. 2014, pp. 163–170. [Online]. Available: https://aclanthology.org/U14-1024.

[58] F. Liu, A. Rahimi, B. Salehi, M. Choi, P. Tan, and L. Duong, "Automatic identification of expressions of locations in tweet messages using conditional random fields," in *Proceedings of the Australasian Language Technology Association Workshop 2014*, Melbourne, Australia, Nov. 2014, pp. 171–176. [Online]. Available: https://aclanthology.org/U14-1025.

[59] L. Ghahremanlou, W. Sherchan, and J. A. Thom, "Geotagging twitter messages in crisis management," *The Computer Journal*, vol. 58, no. 9, pp. 1937–1954, 2015, ISSN: 14602067.

[60] J. Yin, S. Karimi, and J. Lingad, "Pinpointing locational focus in microblogs," in *Proceedings of the 2014 Australasian document computing symposium*, ACM, 2014, p. 66.

[61] H. Mao, G. Thakur, K. Sparks, J. Sanyal, and B. Bhaduri, "Mapping near-real-time power outages from social media," *International Journal of Digital Earth*, vol. 12, no. 11, pp. 1285–1299, 2019, ISSN: 17538955. DOI: 10.1080/17538947.2018.1535000. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/17538947.2018.1535000.

[62] S. Malmasi and M. Dras, "Location mention detection in tweets and microblogs," in *Computational Linguistics*, K. Hasida and A. Purwarianti, Eds., Singapore: Springer Singapore, 2016, pp. 123–134, ISBN: 978-981-10-0515-2.

[63] R. Dutt, K. Hiware, A. Ghosh, and R. Bhaskaran, "SAVITR: A system for real-time location extraction from microblogs during emergencies," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1643–1649, ISBN: 9781450356404.

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

[65] L. Nizzoli, M. Avvenuti, M. Tesconi, and S. Cresci, "Geo-semantic-parsing: AI-powered geoparsing by traversing semantic knowledge graphs," *Decision Support Systems*, vol. 136, p. 113 346, 2020.

[66] W. Zhang and J. Gelernter, "Geocoding location expressions in Twitter messages: A preference learning method," *Journal of Spatial Information Science*, vol. 2014, no. 9, pp. 37–70, 2014.

[67] M. C. Phan, A. Sun, Y. Tay, J. Han, and C. Li, "Neupl: Attention-based semantic matching and pair-linking for entity disambiguation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1667–1676.

[68] J. Wang and Y. Hu, "Are we there yet? evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 2019, pp. 1–6, ISBN: 9781450369602.

[69] X. Wang, C. Ma, H. Zheng, *et al.*, "DM_NLP at SemEval-2018 task 12: A pipeline system for toponym resolution," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 917–923. DOI: 10.18653/v1/S19-2156. [Online]. Available: https://aclanthology.org/S19-2156.

[70] H. Li, M. Wang, T. Baldwin, M. Tomko, and M. Vasardani, "UniMelb at SemEval-2019 task 12: Multi-model combination for toponym resolution," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 1313–1318. DOI: 10.18653/v1/S19-2231. [Online]. Available: https://aclanthology.org/S19-2231.

[71] V. Yadav, E. Laparra, T.-T. Wang, M. Surdeanu, and S. Bethard, "University of Arizona at SemEval-2019 task 12: Deep-affix named entity recognition of geolocation entities," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 1319–1323. DOI: 10.18653/v1/S19-2232. [Online]. Available: https://aclanthology.org/S19-2232.

[72] C. Xu, J. Pei, J. Li, C. Li, X. Luo, and D. Ji, "DLocRL: A deep learning pipeline for fine-grained location recognition and linking in tweets," in *Proceedings of the World Wide Web Conference*, May 2019, pp. 3391–3397.

[73] N. Al Emadi, S. Abbar, J. Borge-Holthoefer, F. Guzman, and F. Sebastiani, "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, 2017.

[74] B. Alkouz and Z. Al Aghbari, "SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks," *Information Processing and Management*, vol. 57, no. 1, p. 102 139, 2020, ISSN: 03064573. DOI: 10.1016/j.ipm.2019.102139.

[75] Y. Zhang, X. Dong, D. Zhang, and D. Wang, "A syntax-based learning approach to geo-locating abnormal traffic events using social sensing," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2019, pp. 663–670.

[76] C. Zhang, K. Zhang, Q. Yuan, *et al.*, "Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 361–370.

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

[78] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 769–778.

[79] R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, and V. Murdock, *Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*. Now Foundations and Trends, 2018.

[80] T. B. N. Hoang and J. Mothe, "Location extraction from tweets," *Information Processing & Management*, vol. 54, no. 2, pp. 129–144, 2018, ISSN: 03064573.

[81] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani, "Twitie: An open-source information extraction pipeline for microblog text," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2013, pp. 83–90.

[82] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Jul. 2011, pp. 1524–1534.

[83] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: News in tweets," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2009, pp. 42–51, ISBN: 9781605586496.

[84] S. Malmasi and M. Dras, "Location mention detection in tweets and microblogs," in *Conference of the Pacific Association for Computational Linguistics*, Springer, 2015, pp. 123–134.

[85] E. A. Sultanik and C. Fink, "Rapid geotagging and disambiguation of social media text via an indexed gazetteer," *Proceedings of ISCRAM*, vol. 12, pp. 1–10, 2012.

[86] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in glasgow: Modeling locations with tweets," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, ACM, 2011, pp. 61–68.

[87] D. Molla and S. Karimi, "Overview of the 2014 alta shared task: Identifying expressions of locations in tweets," in *Proceedings of the Australasian Language Technology Association Workshop 2014*, 2014, pp. 151–156.

[88] S. Guo, M.-W. Chang, and E. Kiciman, "To link or not to link? a study on end-to-end tweet entity linking," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 1020–1030.

[89] Z. Ji, A. Sun, G. Cong, and J. Han, "Joint recognition and linking of fine-grained locations from tweets," in *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 1271–1281, ISBN: 9781450341431.

[90] G. Li, J. Hu, J. Feng, and K.-l. Tan, "Effective location identification from microblogs," in *2014 IEEE 30th International Conference on Data Engineering*, IEEE, 2014, pp. 880–891.

[91] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 363–370. DOI: 10.3115/1219840.1219885. [Online]. Available: https://aclanthology.org/P05-1045.

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

[93] OpenCalais, *Opencalais*, OpenCalais, Ed., [Online; edited 31 March 2022], 2022. [Online]. Available: https://github.com/ElusiveMind/opencalais.

[94] X. Hu, Z. Zhou, Y. Sun, *et al.*, "GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models," *IEEE Internet of Things Journal*, pp. 16 259–16 271, 2022.

[95] G. Rizzo, A. E. C. Basave, B. Pereira, *et al.*, "Making sense of microposts (# microposts2015) named entity recognition and linking (neel) challenge.," in *# MSM*, 2015, pp. 44–53.

[96] P. Ferragina and U. Scaiella, *TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities)*, 2010. DOI: 10.1145/1871437.1871689. [Online]. Available: http://dl.acm.org/citation.cfm?id=1871437.1871689 (visited on 06/01/2021).

[97] D. Weissenbacher, A. Magge, K. O'Connor, M. Scotch, and G. Gonzalez-Hernandez, "SemEval-2019 task 12: Toponym resolution in scientific papers," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Jun. 2019, pp. 907–916.

[98] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, 2003, pp. 142–147.

[99] G. Kordopatis-Zilos, A. Popescu, S. Papadopoulos, and Y. Kompatsiaris, "Placing images with refined language models and similarity search with pca-reduced vgg features.," in *MediaEval*, 2016.

[100] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 363–370. DOI: 10.3115/1219840.1219885. [Online]. Available: https://aclanthology.org/P05-1045.

[101] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, 2014, pp. 43–52.

[102] B. Han, A. J. Yepes, A. MacKinlay, and Q. Chen, "Identifying Twitter location mentions," in *Proceedings of the Australasian Language Technology Association Workshop 2014*, Melbourne, Australia, Nov. 2014, pp. 157–162. [Online]. Available: https://aclanthology.org/U14-1023.

[103] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[104] M. Tanenblatt, A. Coden, and I. Sominsky, "The conceptmapper approach to named entity recognition," in *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*, 2010.

[105] C. Li and A. Sun, "Extracting fine-grained location with temporal awareness in tweets: A two-stage approach," *Journal of the Association for Information Science and Technology*, vol. 68, no. 7, pp. 1652–1670, 2017.

[106] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based *n*-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–480, 1992. [Online]. Available: https://aclanthology.org/J92-4003.

[107] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 2016, pp. 260–270.

[108] N. J. Fernández and C. Periñán-Pascual, "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*, IOS Press, vol. 29, 2021, p. 243.

[109] N. J. F. Martínez and C. Periñán-Pascual, "Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets," *RAEL: revista electrónica de lingüística aplicada*, vol. 19, no. 1, pp. 136–163, 2020.

[110] X. Hu, H. Al-Olimat, J. Kersten, *et al.*, "GazPNE annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules," *International Journal of Geographical Information Science*, 2021.

[111] S. Khanal, M. Traskowsky, and D. Caragea, "Identification of fine-grained location mentions in crisis tweets," in *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2022, pp. 7164–7173. [Online]. Available: `https://aclanthology.org/2022.lrec-1.776`.

[112] S. Khanal and D. Caragea, "Multi-task learning to enable location mention identification in the early hours of a crisis event," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4051–4056.

[113] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *Pro-*

ceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online: Association for Computational Linguistics, Nov. 2020, pp. 6442–6454. DOI: 10.18653/v1/2020.emnlp-main.523. [Online]. Available: https://aclanthology.org/2020.emnlp-main.523.

[114] J. Wang and Y. Hu, "Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers," *Transactions in GIS*, vol. 23, no. 6, pp. 1393–1419, 2019.

[115] J. O. Wallgrün, M. Karimzadeh, A. M. MacEachren, and S. Pezanowski, "Geocorpora: Building a corpus to test and train microblog geoparsers," *International Journal of Geographical Information Science*, vol. 32, no. 1, pp. 1–29, 2018.

[116] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, 2011, pp. 359–367.

[117] C. Li, J. Weng, Q. He, *et al.*, "Twiner: Named entity recognition in targeted twitter stream," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2012, pp. 721–730.

[118] J. Gelernter and W. Zhang, "Cross-lingual geo-parsing for non-structured data," in *Proceedings of the 7th Workshop on Geographic Information Retrieval*, ACM, 2013, pp. 64–71.

[119] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham, "Results of the WNUT2017 shared task on novel and emerging entity recognition," in *Pro-*

ceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 140–147. DOI: 10.18653/v1/W17-4418. [Online]. Available: https://aclanthology.org/W17-4418.

[120] L. Derczynski, K. Bontcheva, and I. Roberts, "Broad Twitter corpus: A diverse named entity recognition resource," in *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, Dec. 2016, pp. 1169–1179.

[121] D. Inkpen, J. Liu, A. Farzindar, F. Kazemi, and D. Ghazi, "Detecting and disambiguating locations mentioned in twitter messages," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9042, Springer Verlag, 2015, pp. 321–332, ISBN: 9783319181165. DOI: 10.1007/978-3-319-18117-2_24. [Online]. Available: https://dev.twitter.com.

[122] P. Chen, H. Xu, C. Zhang, and R. Huang, "Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3329–3339. DOI: 10.18653/v1/2022.naacl-main.243. [Online]. Available: https://aclanthology.org/2022.naacl-main.243.

[123] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 9–17, 2014.

[124] Y. Hu and J. Wang, "How Do People Describe Locations during a Natural Disaster: An Analysis of Tweets from Hurricane Harvey," *Leibniz International Proceedings in Informatics, LIPIcs*, vol. 177, 2020, ISSN: 18688969. DOI: 10.4230/LIPIcs.GIScience.2021.I.6. eprint: 2009.12914. [Online]. Available: http://www.acsu.buffalo.edu/$%5Csim$yhu42/https://geoai.geog.buffalo.edu/people/.

[125] X. Hu, Z. Zhou, H. Li, *et al.*, "Location reference recognition from texts: A survey and comparison," *arXiv preprint arXiv:2207.01683*, 2022.

[126] N. J. Fernández-Martínez, "The FGLOCTweet corpus: An english tweet-based corpus for fine-grained location-detection tasks," *Research in Corpus Linguistics*, vol. 10, no. 1, pp. 117–133, 2022.

[127] K. Bahnasy, A. El-Mahdy, *et al.*, "Twitter analysis based on damage detection and geoparsing for event mapping management," *Future Computing and Informatics Journal*, vol. 5, no. 1, p. 1, 2020.

[128] K. Darwish, "Named entity recognition using cross-lingual resources: Arabic as an example," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 1558–1567.

[129] G. Aguilar, F. AlGhamdi, V. Soto, M. Diab, J. Hirschberg, and T. Solorio, "Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 138–147. DOI: 10.18653/v1/W18-3219. [Online]. Available: https://aclanthology.org/W18-3219.

[130] M. Jarrar, M. Khalilia, and S. Ghanem, "Wojood: Nested arabic named entity corpus and recognition using bert," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, June*, 2022.

[131] B. Alkouz and Z. Al Aghbari, "Leveraging cross-lingual tweets in location recognition," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*, IEEE, 2018, pp. 0084–0089.

[132] K. Darwish and W. Gao, "Simple effective microblog named entity recognition: Arabic as an example," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 2513–2517.

[133] J. D. G. Paule, Y. Sun, and Y. Moshfeghi, "On fine-grained geolocalisation of tweets and real-time traffic incident detection," *Information Processing & Management*, vol. 56, no. 3, pp. 1119–1132, 2019.

[134] L. Shang, Y. Zhang, C. Youn, and D. Wang, "Sat-geo: A social sensing based content-only approach to geolocating abnormal traffic events using syntax-based probabilistic learning," *Information Processing & Management*, vol. 59, no. 2, p. 102 807, 2022.

[135] A. Olteanu, S. Vieweg, and C. Castillo, "What to expect when the unexpected happens: Social media communications across crises," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, ACM, 2015, pp. 994–1009.

[136] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Practical extraction of disaster-relevant information from social media," in *Proceedings of the 22nd*

international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 1021–1024.

[137] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages," *arXiv preprint arXiv:1605.05894*, 2016.

[138] A. Alharbi and M. Lee, "Kawarith: An Arabic Twitter corpus for crisis events," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 42–52. [Online]. Available: https://aclanthology.org/2021.wanlp-1.5.

[139] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 82–91. [Online]. Available: https://aclanthology.org/2021.wanlp-1.9.

[140] M. Karimzadeh, "Performance evaluation measures for toponym resolution," in *Proceedings of the 10th Workshop on Geographic Information Retrieval*, ser. GIR '16, Burlingame, California: Association for Computing Machinery, 2016, ISBN: 9781450345880. DOI: 10.1145/3003464.3003472. [Online]. Available: https://doi.org/10.1145/3003464.3003472.

[141] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Jun. 2009, pp. 147–155.

[142] H.-J. Dai, P.-T. Lai, Y.-C. Chang, and R. T.-H. Tsai, "Enhancing of chemical compound and drug name recognition using representative rag scheme and fine-grained tokenization," *Journal of cheminformatics*, vol. 7, no. S1, S14, Jan. 2015.

[143] J. Yang, S. Liang, and Y. Zhang, "Design challenges and misconceptions in neural sequence labeling," in *Proceedings of the 27th International Conference on Computational Linguistics*, Aug. 2018, pp. 3879–3889.

[144] L. Derczynski, K. Bontcheva, and I. Roberts, "Broad Twitter corpus: A diverse named entity recognition resource," in *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, Dec. 2016, pp. 1169–1179.

[145] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 9–17, 2014.

[146] E. F. Tjong Kim Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task: Chunking," in *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*, 2000, pp. 127–132.

[147] C. Reuter, T. Ludwig, C. Kotthaus, M.-A. Kaufhold, E. von Radziewski, and V. Pipek, "Big data in a crisis? creating social media datasets for crisis management research," *i-com*, vol. 15, no. 3, pp. 249–264, 2016.

[148] A. Kitamoto and T. Sagara, "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, ser. GeoMM '12, Nara, Japan: Association for Computing Machinery, 2012, pp. 23–26, ISBN: 9781450315906. DOI: 10.1145/2390790.2390799. [Online]. Available: https://doi.org/10.1145/2390790.2390799.

[149] F. Alam, U. Qazi, M. Imran, and F. Ofli, "Humaid: Human-annotated disaster incidents data from twitter," in *15th International Conference on Web and Social Media (ICWSM)*, 2021.

[150] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970. DOI: 10.1177/001316447003000105. eprint: https://doi.org/10.1177/001316447003000105. [Online]. Available: https://doi.org/10.1177/001316447003000105.

[151] M. Imran, P. Mitra, and C. Castillo, "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 (LREC 2016)*, Portoroz, Slovenia: European Language Resources Association (ELRA), 2016, ISBN: 978-2-9517408-9-1.

[152] F. Alam, S. Joty, and M. Imran, "Domain adaptation with adversarial training and graph embeddings," 2018.

[153] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "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*, ACM, 2017, pp. 569–576.

[154] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[155] Z. Ji, A. Sun, G. Cong, and J. Han, "Joint recognition and linking of fine-grained locations from tweets," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 1271–1281, ISBN: 9781450341431.

[156] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[157] W. Alabbas, H. M. al-Khateeb, A. Mansour, G. Epiphaniou, and I. Frommholz, "Classification of colloquial arabic tweets in real-time to detect high-risk floods," in *2017 International Conference On Social Media, Wearable And Web Analytics (Social Media)*, 2017, pp. 1–8. DOI: 10.1109/SOCIALMEDIA.2017.8057358.

[158] A. Alharbi and M. Lee, "Crisis detection from arabic tweets," in *Proceedings of the 3rd workshop on arabic corpus linguistics*, 2019, pp. 72–79.

[159] Y. A. Ameen, K. Bahnasy, and A. E. Elmahdy, "Classification of arabic tweets for damage event detection," 2020.

[160] S. Hassan, H. Mubarak, A. Abdelali, and K. Darwish, "Asad: Arabic social media analytics and understanding," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021, pp. 113–118.

[161] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "Arbert & marbert: Deep bidirectional transformers for arabic," *arXiv preprint arXiv:2101.01785*, 2020.

[162] B. A. Benali, S. Mihi, N. Laachfoubi, and A. A. Mlouk, "Arabic named entity recognition in arabic tweets using bert-based models," *Procedia Computer Science*, vol. 203, pp. 733–738, 2022.

[163] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online): Association for Computational Linguistics, Apr. 2021.

[164] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *NAACL*, 2016.

[165] K.-T. Chang, *Introduction to geographic information systems*. McGraw-Hill Boston, 2008, vol. 4.

[166] *Nominatim api*. [Online]. Available: `https://nominatim.org/release-docs/develop/`.

[167] A. Mourad, F. Scholer, W. Magdy, and M. Sanderson, "A practical guide for the effective evaluation of twitter user geolocation," *ACM Transactions on Social Computing*, vol. 2, no. 3, pp. 1–23, 2019.

[168] G. Pettet, H. Baxter, S. M. Vazirizade, *et al.*, "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*, 2022.

[169] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: News in tweets," in *Proceedings of the 17th acm sigspatial*

international conference on advances in geographic information systems, 2009, pp. 42–51.

[170] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "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*, 2011, pp. 2541–2544.

[171] V. Lorini, C. Castillo, S. Peterson, *et al.*, "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*, 2021, pp. 772–777.

[172] J. P. De Albuquerque, B. Herfort, A. Brenning, and A. Zipf, "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management," *International journal of geographical information science*, vol. 29, no. 4, pp. 667–689, 2015.

[173] M.-A. Kaufhold, M. Bayer, and C. Reuter, "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning," *Information Processing & Management*, vol. 57, no. 1, p. 102 132, 2020.

[174] O. Ozdikis, H. Ramampiaro, and K. Nørvåg, "Locality-adapted kernel densities of term co-occurrences for location prediction of tweets," *Information Processing & Management*, vol. 56, no. 4, pp. 1280–1299, 2019.

[175] X. Luo, Y. Qiao, C. Li, J. Ma, and Y. Liu, "An overview of microblog user geolocation methods," *Information processing & management*, vol. 57, no. 6, p. 102 375, 2020.

[176] J. Wu, R. Hu, D. Li, L. Ren, W. Hu, and Y. Xiao, "Where have you been: Dual spatiotemporal-aware user mobility modeling for missing check-in poi identification," *Information Processing & Management*, vol. 59, no. 5, p. 103 030, 2022.

[177] T. Miyazaki, A. Rahimi, T. Cohn, and T. Baldwin, "Twitter geolocation using knowledge-based methods," in *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 7–16. DOI: 10.18653/v1/W18-6102. [Online]. Available: https://aclanthology.org/W18-6102.

[178] M. Á. García-Cumbreras, J. M. Perea-Ortega, M. García-Vega, and L. A. Ureña-López, "Information retrieval with geographical references. relevant documents filtering vs. query expansion," *Information processing & management*, vol. 45, no. 5, pp. 605–614, 2009.

APPENDIX A: DETAILED TRANSFER LMR RESULTS

Detailed results, including out-domain training, in/out-domain training, cross-domain training, and training based on geo-proximity of events (Table A.1).

Table A.1. Full results of different domain setups. Best F1 scores of non-target training setups are boldfaced. E, BS and LR refer to the number of training epochs, the training batch size, and the learning rate (Adam), respectively. For LR, 3, 4, and 5 represent values 5e-3, 5e-4, and 5e-5, respectively.

	Chennai FLD						Houston FLD						Louisiana FLD						HRC Sandy						ChCh EQK					
Source Data	E	BS	LR	P	R	F1	E	BS	LR	P	R	F1	E	BS	LR	P	R	F1	E	BS	LR	P	R	F1	E	BS	LR	P	R	F1
Target	3	16	5	0.86	0.86	0.87	3	16	5	0.87	0.88	0.87	4	16	3	0.94	0.91	0.92	4	16	5	0.88	0.90	0.88	4	32	5	0.82	0.83	0.82
Out-domain general purpose training results.																														
CoNLL.ner	4	16	3	0.57	0.56	0.56	4	16	3	0.59	0.53	0.56	4	16	3	0.42	0.72	0.53	4	16	3	0.49	0.73	0.59	4	16	3	0.38	0.74	0.50
BTC.ner	3	32	5	0.66	0.67	0.66	3	32	5	0.64	0.55	0.60	3	32	5	0.46	0.74	0.56	3	32	5	0.54	0.65	0.59	3	32	5	0.40	0.75	0.52
CoNLL.loc	3	16	3	0.78	0.53	0.63	3	16	3	0.84	0.48	0.61	3	16	3	0.89	0.66	0.76	3	16	3	0.69	0.61	0.66	3	16	3	0.68	0.71	0.69
BTC.loc	3	16	5	0.83	0.66	0.72	3	16	5	0.78	0.53	0.63	3	16	5	0.86	0.73	0.79	3	16	5	0.70	0.63	0.66	3	16	5	0.65	0.74	0.69
In & out-domain training results.																														
DIS.others	3	16	5	0.86	0.53	0.65	4	16	5	0.83	0.65	0.731	4	16	5	0.87	0.74	0.81	4	16	5	0.72	0.74	0.72	4	16	5	0.65	0.81	0.72
Combined.joint	3	16	5	0.85	0.61	0.71	4	16	3	0.83	0.63	0.72	4	16	3	0.89	0.77	0.82	2	16	5	0.70	0.71	0.71	3	16	3	0.64	0.78	0.70
Combined.seq	2	16	5	0.87	0.59	0.70	3	16	5	0.85	0.67	0.75	4	32	5	0.89	0.78	0.83	3	16	5	0.73	0.73	0.74	4	16	5	0.62	0.82	0.70
Cross-domain training results.																														
DIS.FLD.others	4	16	5	0.82	0.64	0.72	3	32	5	0.82	0.59	0.69	4	16	5	0.80	0.78	0.79	4	16	3	0.68	0.73	0.70	4	16	3	0.58	0.76	0.66
DIS.HRC.others	4	16	5	0.81	0.35	0.49	4	16	5	0.79	0.54	0.64	4	16	5	0.90	0.66	0.77							4	16	5	0.68	0.70	0.69
DIS.EQK.others	4	32	5	0.80	0.33	0.46	4	32	5	0.66	0.07	0.13	4	32	5	0.57	0.02	0.03	4	32	5	0.77	0.12	0.21						
Geo-proximity-based training results.																														
DIS_US.others							3	16	5	0.80	0.62	0.70	4	32	5	0.86	0.71	0.79	4	16	5	0.71	0.72	0.71						
DIS_IN.FLD							4	32	5	0.85	0.38	0.52	4	32	5	0.74	0.22	0.34	4	32	5	0.56	0.42	0.45						
DIS_NZ.EQK							4	16	5	0.71	0.13	0.22	4	16	5	0.63	0.04	0.07	4	16	5	0.78	0.19	0.30						

APPENDIX B: IDRISI DATA RELEASE

The IDRISI-RA dataset is released[1] data setups that are *random* and *Time-based*. The location mention and location type annotations are made available for the community to enable development of *type-less* and *type-based* LMR models. The data is released in **JSONL** format where every lines corresponds to one tweet with the following properties: "text", "created_at", "info_class" adopted from Kwaraith dataset, and "location_mentions".

Tables B.1, B.2, and B.3 show the detailed statistics of IDRISI-RE and IDRISI-RA datasets for the random and time-based setups per event.

In Figure B.1, we depict the temporal coverage of Cyclone Idai 2019 and Kerala FLD 2018 events from IDRISI-RE dataset. In Figure B.2, we depict the temporal coverage of COVID-19 and Kuwait FLD 2018 events from IDRISI-RA dataset.

Tables B.4-B.6 show detailed statistics of IDRISI-D datasets.

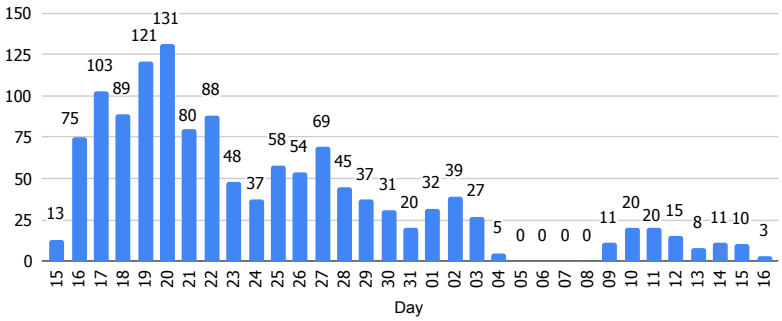
[1]This dataset is licensed under a Creative Commons Attribution 4.0 International License: `https://creativecommons.org/licenses/by/4.0/legalcode`

Table B.1. Detailed information and statistics of IDRISI-RE dataset for the *random* setup. "TRN", "DEV", "TST" refer to the training, development, and test splits, respectively. LM_0 refers to the number of tweets with no LMs.

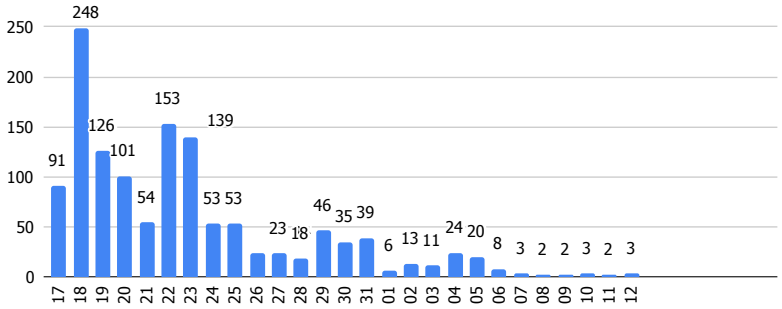
			Tweets					LMs				
Event	Country	Time Period	TRN	DEV	TST	All	LM_0	TRN	DEV	TST	All	Uniq
Ecuador EQK	EC	04/17 - 04/18	812	116	225	1,153	205	846	110	222	1,178	116
Canada FIR	CA	05/06 - 05/27	910	130	260	1,300	455	756	98	221	1,075	181
Italy EQK	IT	08/24 - 08/29	413	59	118	590	360	203	30	51	284	66
Kaikoura EQK	NZ	09/01 - 11/22	868	124	239	1,231	375	804	114	215	1,133	227
HRC Matthew	US	10/04 - 10/10	602	86	167	855	62	794	121	217	1,132	119
Sri Lanka FLD	SK	05/31 - 07/03	322	46	89	457	88	366	57	98	521	105
HRC Harvey	US	08/25 - 09/01	910	130	259	1,299	719	542	68	143	753	182
HRC Irma	US	09/06 - 09/17	910	130	258	1,298	563	658	83	210	951	354
HRC Maria	US	09/16 - 10/02	910	130	259	1,299	357	814	114	237	1,165	217
Mexico EQK	MX	09/20 - 09/23	910	130	260	1,300	132	972	132	292	1,396	116
Maryland FLD	US	05/28 - 06/07	301	43	78	422	28	473	97	123	693	89
Greece FIR	GR	07/24 - 08/18	679	97	191	967	123	984	143	238	1,365	156
Kerala FLD	IN	08/17 - 08/31	910	130	260	1,300	331	1,175	151	330	1,656	367
HRC Florence	US, CA	09/11 - 09/18	910	130	260	1,300	559	869	119	257	1,245	403
California FIR	US	11/10 - 12/07	910	130	260	1,300	277	939	125	269	1,333	165
Cyclone Idai	MZ, ZW, MW, MG	03/15 - 04/16	910	130	260	1,300	376	1,201	160	383	1,744	268
Midwest. US FLD	US	03/25 - 04/03	756	108	212	1,076	92	1,115	166	311	1,592	189
HRC Dorian	US	08/30 - 09/02	910	130	260	1,300	492	869	156	235	1,260	325
Pakistan EQK	PK	09/24 - 09/26	539	77	151	767	129	983	150	270	1,403	185
Total			14,392	2,056	4,066	20,515	4,723	15,363	2,194	4,322	21,879	3,830

Table B.2. Detailed information and statistics of IDRISI-RE dataset for the *time-based* setup. "TRN", "DEV", "TST" refer to the training, development, and test splits, respectively. LM_0 refers to the number of tweets with no LMs.

			Tweets					LMs				
Event	Country	Time Period	TRN	DEV	TST	All	LM_0	TRN	DEV	TST	All	Uniq
Ecuador EQK	EC	04/17 - 04/18	812	116	225	1,153	205	850	113	215	1,178	116
Canada FIR	CA	05/06 - 05/27	910	130	260	1,300	455	753	106	216	1,075	181
Italy EQK	IT	08/24 - 08/29	413	59	118	590	360	191	28	65	284	66
Kaikoura EQK	NZ	09/01 - 11/22	868	124	239	1,231	375	812	93	228	1,133	227
HRC Matthew	US	10/04 - 10/10	602	86	167	855	62	815	114	203	1,132	119
Sri Lanka FLD	SK	05/31 - 07/03	322	46	89	457	88	370	56	95	521	105
HRC Harvey	US	08/25 - 09/01	910	130	259	1,299	719	566	66	121	753	182
HRC Irma	US	09/06 - 09/17	910	130	258	1,298	563	701	93	157	951	354
HRC Maria	US	09/16 - 10/02	910	130	259	1,299	357	840	97	228	1,165	217
Mexico EQK	MX	09/20 - 09/23	910	130	260	1,300	132	984	130	282	1,396	116
Maryland FLD	US	05/28 - 06/07	301	43	78	422	28	475	68	150	693	89
Greece FIR	GR	07/24 - 08/18	679	97	191	967	123	934	132	299	1,365	156
Kerala FLD	IN	08/17 - 08/31	910	130	260	1,300	331	1,152	163	341	1,656	367
HRC Florence	US, CA	09/11 - 09/18	910	130	260	1,300	559	873	158	214	1,245	403
California FIR	US	11/10 - 12/07	910	130	260	1,300	277	945	133	255	1,333	165
Cyclone Idai	MZ, ZW, MW, MG	03/15 - 04/16	910	130	260	1,300	376	1,233	158	353	1,744	268
Midwest. US FLD	US	03/25 - 04/03	756	108	212	1,076	92	1,129	152	311	1,592	189
HRC Dorian	US	08/30 - 09/02	910	130	260	1,300	492	806	142	312	1,260	325
Pakistan EQK	PK	09/24 - 09/26	539	77	151	767	129	984	144	275	1,403	185
Total			14,392	2,056	4,066	20,515	4,723	15,412	3,146	4,320	21,879	3,830

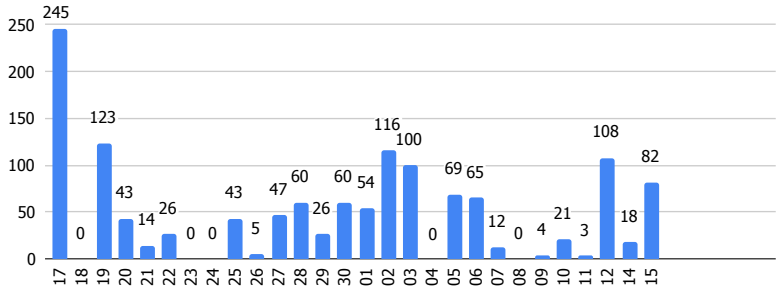


(a) Cyclone Idai 2019 (15 Mar - 16 Apr)

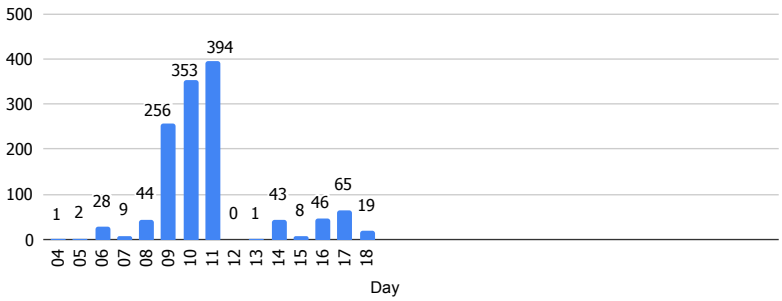


(b) Kerala FLD 2018 (17 Aug - 31 Aug)

Figure B.1. The temporal coverage of tweets in IDRISI-RA.



(a) COVID-19 (17 May - 15 Jun)



(b) Kuwait FLD 2018 (04 Nov - 18 Nov)

Figure B.2. The temporal coverage of tweets in IDRISI-RA.

Table B.3. Detailed information and statistics of IDRISI-RA datasets, both *random* and *time-based* setups. "TRN", "DEV", "TST" refer to the training, development, and test splits, respectively. LM_0 refers to the number of tweets with no LMs.

			Tweets					LMs				
Event	Country	Time Period	TRN	DEV	TST	All	LM_0	TRN	DEV	TST	All	Uniq
Random												
Jordan FLD 2018	JO	10/25 - 11/18	371	53	103	527	107	614	89	194	897	108
Kuwait FLD 2018	KW	11/04 - 11/18	889	127	253	1,269	503	820	93	224	1,137	140
Cairo BMB 2019	EG	08/04 - 08/04	189	27	52	268	1	417	66	140	623	24
Hafr FLD 2019	SA	10/25 - 10/29	364	52	98	514	46	520	83	149	752	112
Dragon STR 2020	EG & JO [a]	03/11 - 03/15	217	31	57	305	122	252	27	59	338	160
Beirut BMB 2020	LB	08/04 - 08/07	245	35	69	349	63	378	49	123	550	61
CoVID-19	Global	05/17 - –/–[b]	959	137	265	1,361	777	637	96	206	939	313
Time-based												
Jordan FLD 2018	JO	10/25 - 11/18	371	53	103	527	107	631	79	187	897	108
Kuwait FLD 2018	KW	04/11 - 04/18	889	127	253	1269	503	772	112	253	1,137	140
Cairo BMB 2019	EG	08/04 - 08/04	189	27	52	268	1	458	53	112	623	24
Hafr FLD 2019	SA	10/25 - 10/29	364	52	98	514	46	526	79	147	752	112
Dragon STR 2020	EG & JO [a]	03/11 - 03/15	217	31	57	305	122	248	39	51	338	160
Beirut BMB 2020	LB	08/04 - 08/07	245	35	69	349	63	400	55	95	550	61
CoVID-19	Global	05/17 - –/–[b]	959	137	265	1361	777	599	109	231	939	313
All			3,234	462	897	4,593	1,619	3,634	526	1,076	5,236	918

[a] While Dragon storms have affected several Arab countries and LMs are from different countries, Kawarith creators had intentionally focused on the Egyptian tweets. We found LMs from Jordan as well.

[b] The last tweet in the chronologically sorted tweets was published in 2020/06/15 but the pandemic was ongoing.

Table B.4. Detailed statistics of IDRISI-DE dataset. “TRN”, “DEV”, “TST” refer to the training, development, and test splits, respectively.

Event	Tweets	LMs	Uniq LMs	TRN	DEV	TST
Ecuador EQK	132	163	58	106 (65.0%)	21 (12.9%)	36 (22.1%)
Canada FIR	151	220	66	143 (65.0%)	26 (11.8%)	51 (23.2%)
Italy EQK	186	236	163	153 (64.8%)	40 (16.9%)	43 (18.2%)
Kaikoura EQK	209	319	28	208 (65.2%)	38 (11.9%)	73 (22.9%)
HRC Matthew	173	259	91	166 (64.1%)	41 (15.8%)	52 (20.1%)
Sri Lanka FLD	69	101	101	63 (62.4%)	17 (16.8%)	21 (20.8%)
HRC Harvey	536	748	183	495 (66.2%)	83 (11.1%)	170 (22.7%)
HRC Irma	179	254	132	164 (64.6%)	40 (15.7%)	50 (19.7%)
HRC Maria	195	273	122	178 (65.2%)	43 (15.8%)	52 (19.0%)
Mexico EQK	169	226	64	149 (65.9%)	31 (13.7%)	46 (20.4%)
Maryland FLD	101	206	25	134 (65.0%)	30 (14.6%)	42 (20.4%)
Greece FIR	743	1,218	38	785 (64.4%)	158 (13.0%)	275 (22.6%)
Kerala FLD	838	1,454	89	942 (64.8%)	213 (14.6%)	299 (20.6%)
HRC Florence	202	450	246	284 (63.1%)	45 (10.0%)	121 (26.9%)
California FIR	131	201	25	136 (67.7%)	24 (11.9%)	41 (20.4%)
Cyclone Idai	1,006	2,138	76	1,285 (60.1%)	331 (15.5%)	522 (24.4%)
Midwest. US FLD	187	371	38	219 (59.0%)	72 (19.4%)	80 (21.6%)
HRC Dorian	198	379	28	233 (61.5%)	49 (12.9%)	97 (25.6%)
Pakistan EQK	186	469	28	279 (59.5%)	51 (10.9%)	139 (29.6%)
All	5,591	9,685	1,601	6,122 (63.2%)	1,353 (14.0%)	2,210 (22.8%)

Table B.5. Detailed statistics of LMs in IDRISI-DE dataset per location type.

Event	Country	State	County	City	District	Neighborhood	Street	Point-of-Interest	Other locations
Ecuador EQK	133	3	1	1	20	0	0	0	2
Canada FIR	24	24	3	21	80	7	5	15	33
Italy EQK	129	6	0	0	95	0	1	1	3
Kaikoura EQK	72	1	0	13	140	27	2	29	11
HRC Matthew	209	9	0	0	12	0	0	1	0
Sri Lanka FLD	72	0	0	6	17	0	2	1	3
HRC Harvey	57	313	38	3	305	3	0	7	20
HRC Irma	32	68	7	2	47	9	0	14	20
HRC Maria	18	9	3	4	34	9	1	12	6
Mexico EQK	120	18	0	2	79	1	0	0	4
Maryland FLD	1	117	8	0	64	2	0	13	0
Greece FIR	791	29	2	3	369	0	2	0	8
Kerala FLD	106	746	2	6	435	3	91	9	50
HRC Florence	22	162	29	4	99	73	4	22	29
California FIR	3	94	9	4	49	3	1	13	21
Cyclone Idai	1,429	29	6	3	538	7	47	14	26
Midwest. US FLD	4	259	21	17	38	2	2	14	7
HRC Dorian	36	95	43	3	54	3	0	12	25
Pakistan EQK	100	36	0	0	203	2	2	8	13
All	3,358	2,018	172	92	2,678	151	160	185	281

Table B.6. Detailed statistics of IDRISI-DA dataset. “TRN”, “DEV”, “TST” refer to the training, development, and test splits, respectively.

Event	Tweets	LMs	Uniq LMs	TRN	DEV	TST
Jordan FLD 2018	508	1,222	71	407 (64.7%)	50 (7.9%)	172 (27.3%)
Kuwait FLD 2018	786	1,186	38	525 (60.2%)	132 (15.1%)	215 (24.7%)
Cairo BMB 2019	332	872	415	724 (60.1%)	214 (17.8%)	266 (22.1%)
Hafr FLD 2019	504	903	231	256 (59.7%)	53 (12.4%)	120 (28.0%)
Dragon STR 2020	200	429	130	559 (61.9%)	129 (14.3%)	215 (23.8%)
Beirut BMB 2020	307	629	162	749 (61.3%)	118 (9.7%)	355 (29.1%)
CoVID-19	657	1,204	179	777 (65.5%)	122 (10.3%)	287 (24.2%)
All	3,294	6,445	1,226	3,997 (62.0%)	818 (12.7%)	1,630 (25.3%)

Table B.7. Detailed statistics of LMs in IDRISI-DA dataset per location type.

Event	Country	Province	District	City	Neighborhood	Street	Point-of-Interest	Other locations
Jordan FLD 2018	373	22	159	472	0	56	119	21
Kuwait FLD 2018	620	15	213	183	5	74	71	5
Cairo BMB 2019	16	5	12	32	204	81	522	0
Hafr FLD 2019	56	547	126	56	7	7	91	13
Dragon STR 2020	106	19	41	138	14	56	25	30
Beirut BMB 2020	233	0	2	284	2	0	108	0
CoVID-19	717	116	23	205	8	9	97	29
All	2,121	724	576	1,370	240	283	1,033	98

APPENDIX C: IDRISI-R DETAILED FINE-TUNING RESULTS AND BEST HYPER-PARAMETERS

Tables C.1 and C.2 show the best hyper-parameters and detailed results of the BERT$_{LMR}$ model for both type-less and type-based LMR over IDRISI-RE. Tables C.3 and C.4 show the best hyper-parameters and detailed results of the CRF LMR models for both type-less and type-based recognition, respectively.

Tables C.5 and C.6 show the best hyper-parameters and detailed results for the CRF$_{LMR}$ and BERT$_{LMR}$ models for IDRISI-RA, respectively.

Table C.1. The best hyper-parameters and results of the BERT$_{LMR}$ model over IDRISI-RE under the *random* data setup. e, bs, lr, and sl refer to the hyper-parameters, number of epochs, batch size, learning rate, and sequence length, respectively.

Event	e	bs	lr	P	R	F1
Type-less						
Ecuador Earthquake	4	32	4e-5	0.960	0.958	0.953
Canada Wildfires	4	8	4e-5	0.733	0.749	0.732
Italy Earthquake	3	8	3e-5	0.881	0.886	0.880
Kaikoura Earthquake	3	8	3e-5	0.914	0.919	0.912
Hurricane Matthew	4	8	5	0.948	0.945	0.941
Sri Lanka Floods	3	16	4e-5	0.921	0.929	0.917
Hurricane Harvey	4	8	5	0.919	0.902	0.906
Hurricane Irma	4	8	3e-5	0.843	0.839	0.835
Hurricane Maria	2	8	4e-5	0.932	0.926	0.925
Mexico Earthquake	4	8	3e-5	0.932	0.932	0.929
Maryland Floods	3	16	5	0.895	0.901	0.890
Greece Wildfires	3	8	5	0.935	0.934	0.927
Kerala Floods	4	32	5	0.897	0.893	0.887
Hurricane Florence	4	8	4e-5	0.773	0.755	0.755
California Wildfires	3	16	3e-5	0.923	0.930	0.920
Cyclone Idai	3	8	4e-5	0.932	0.927	0.925
Midwestern U.S. Floods	4	8	5	0.948	0.957	0.944
Hurricane Dorian	4	8	5	0.874	0.893	0.878
Pakistan Earthquake	3	32	4e-5	0.876	0.902	0.877
Type-based						
Ecuador Earthquake	2	8	3e-5	0.951	0.940	0.939
Canada Wildfires	3	8	4e-5	0.733	0.749	0.733
Italy Earthquake	3	8	4e-5	0.894	0.894	0.890
Kaikoura Earthquake	4	16	5	0.914	0.916	0.909
Hurricane Matthew	4	32	5	0.931	0.923	0.919
Sri Lanka Floods	4	8	5	0.929	0.933	0.925
Hurricane Harvey	4	16	4e-5	0.921	0.905	0.909
Hurricane Irma	2	8	5	0.847	0.831	0.833
Hurricane Maria	2	8	5	0.936	0.924	0.924
Mexico Earthquake	2	16	4e-5	0.921	0.914	0.913
Maryland Floods	3	8	4e-5	0.906	0.894	0.892
Greece Wildfires	3	16	3e-5	0.927	0.940	0.925
Kerala Floods	4	8	5	0.891	0.885	0.880
Hurricane Florence	3	16	4e-5	0.795	0.774	0.772
California Wildfires	4	32	4e-5	0.913	0.919	0.909
Cyclone Idai	3	32	4e-5	0.906	0.906	0.900
Midwestern U.S. Floods	4	8	5	0.944	0.948	0.936
Hurricane Dorian	4	16	5	0.857	0.871	0.858
Pakistan Earthquake	4	8	5	0.899	0.908	0.894

Table C.2. The best hyper-parameters and results of the BERT$_{LMR}$ model over IDRISI-RE under the *time-based* data setup.. e, bs, lr, and sl refer to the hyper-parameters, number of epochs, batch size, learning rate, and sequence length, respectively.

Event	e	bs	lr	P	R	F1
Type-less						
Ecuador Earthquake	4	32	4e-5	0.923	0.921	0.916
Canada Wildfires	4	8	4e-5	0.768	0.779	0.767
Italy Earthquake	8	3e-5	0.840	0.849	0.842	
Kaikoura Earthquake	3	8	3e-5	0.912	0.893	0.896
Hurricane Matthew	4	8	5	0.949	0.956	0.944
Sri Lanka Floods	3	16	4e-5	0.904	0.918	0.904
Hurricane Harvey	4	8	5	0.900	0.893	0.894
Hurricane Irma	4	8	3e-5	0.829	0.833	0.825
Hurricane Maria	2	8	4e-5	0.913	0.909	0.904
Mexico Earthquake	4	8	3e-5	0.919	0.913	0.911
Maryland Floods	3	16	5	0.900	0.838	0.845
Greece Wildfires	3	8	5	0.897	0.895	0.883
Kerala Floods	4	32	5	0.927	0.934	0.923
Hurricane Florence	4	8	4e-5	0.801	0.785	0.784
California Wildfires	3	16	3e-5	0.914	0.906	0.906
Cyclone Idai	3	8	4e-5	0.911	0.900	0.898
Midwestern U.S. Floods	4	8	5	0.946	0.961	0.949
Hurricane Dorian	8	5	0.865	0.872	0.862	
Pakistan Earthquake	3	32	4e-5	0.830	0.878	0.836
Type-based						
Ecuador Earthquake	4	32	4e-5	0.941	0.922	0.926
Canada Wildfires	4	8	4e-5	0.772	0.780	0.771
Italy Earthquake	3	8	3e-5	0.879	0.888	0.881
Kaikoura Earthquake	3	8	3e-5	0.918	0.895	0.899
Hurricane Matthew	4	8	5	0.955	0.963	0.952
Sri Lanka Floods	3	16	4e-5	0.911	0.925	0.912
Hurricane Harvey	4	8	5	0.898	0.896	0.895
Hurricane Irma	4	8	3e-5	0.827	0.828	0.823
Hurricane Maria	2	8	4e-5	0.910	0.895	0.897
Mexico Earthquake	4	8	3e-5	0.918	0.914	0.911
Maryland Floods	3	16	5	0.851	0.795	0.805
Greece Wildfires	3	8	5	0.899	0.899	0.887
Kerala Floods	4	32	5	0.926	0.927	0.919
Hurricane Florence	4	8	4e-5	0.792	0.781	0.778
California Wildfires	3	16	3e-5	0.918	0.900	0.902
Cyclone Idai	3	8	4e-5	0.905	0.900	0.895
Midwestern U.S. Floods	4	8	5	0.944	0.957	0.944
Hurricane Dorian	4	8	5	0.864	0.860	0.852
Pakistan Earthquake	3	32	4e-5	0.819	0.868	0.828

Table C.3. 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. "var", "eps", and "g" refer to "var" "epsilon", and "g", respectively.

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	eps=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	eps=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	var=0.1	g=0.5	0.838	0.884	0.860
Maryland Floods	arow	var=0.1	g=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	eps=1e-5		0.705	0.745	0.725
Hurricane Florence	ap	eps=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	var=0.25	g=0.25	0.933	0.933	0.932
Canada Wildfires	ap	eps=0.01		0.853	0.853	0.853
Italy Earthquake	arow	var=1	g=0.125	0.906	0.906	0.906
Kaikoura Earthquake	arow	var=0.5	g=0.25	0.880	0.880	0.879
Hurricane Matthew	arow	var=1	g=0.1	0.902	0.905	0.901
Sri Lanka Floods	arow	var=1	g=0.1	0.911	0.911	0.910
Hurricane Harvey	arow	var=0.1	g=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	var=0.16	g=0.5	0.883	0.883	0.882
Mexico Earthquake	arow	var=0.25	g=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	var=0.5	g=0.1	0.895	0.901	0.896
Kerala Floods	arow	var=0.125	g=0.5	0.880	0.881	0.880
Hurricane Florence	arow	var=1	g=0.16	0.879	0.879	0.879
California Wildfires	arow	var=1	g=0.125	0.908	0.908	0.907
Cyclone Idai	arow	var=0.125	g=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	var=0.16	g=0.5	0.875	0.875	0.875
Pakistan Earthquake	arow	var=1	g=0.125	0.821	0.822	0.820

Table C.4. 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	eps=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	var=0.1	g=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	var=0.16	g=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	var=0.1	g=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	var=0.5	g=0.16	0.881	0.881	0.881
Kaikoura Earthquake	arow	var=0.16	g=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	var=1	g=0.25	0.900	0.896	0.897
Hurricane Harvey	ap	eps=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	var=1	g=0.1	0.890	0.890	0.890
Mexico Earthquake	arow	var=0.1	g=1	0.881	0.882	0.880
Maryland Floods	arow	var=1	g=0.16	0.875	0.878	0.873
Greece Wildfires	arow	var=0.25	g=0.5	0.886	0.890	0.886
Kerala Floods	arow	var=1	g=0.1	0.857	0.859	0.857
Hurricane Florence	arow	var=0.1	g=0.5	0.889	0.889	0.889
California Wildfires	arow	var=1	g=0.125	0.903	0.903	0.902
Cyclone Idai	arow	var=1	g=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	var=0.5	g=0.1	0.865	0.866	0.865
Pakistan Earthquake	arow	var=0.16	g=0.16	0.781	0.782	0.780

Table C.5. The best hyper-parameters and results for CRF model over IDRISI-RA. 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 Type-less LMR						
Jordan Floods	arow	*var*=0.1	*g*=0.16	0.841	0.845	0.843
Kuwait Floods	arow	*var*=1	*g*=0.125	0.865	0.603	0.711
Cairo Bombing	l2sgd	*c2*=0.2	*ce*=1e-2	0.971	0.964	0.968
Hafr Floods	lbfgs	*c1*=0.05	*c2*=0.05	0.881	0.799	0.838
Dragon Storms	pa	*c*=1	*e_sensitive*=TRUE	0.787	0.627	0.698
Beirut Explosion	arow	*var*=0.1	*g*=0.5	0.943	0.813	0.873
CoVID-19	arow	*var*=1	*g*=1	0.787	0.539	0.640
Random data setup Type-based LMR						
Jordan Floods	lbfgs	*c1*=0.25	*c2*=0.25	0.766	0.786	0.776
Kuwait Floods	arow	*var*=0.5	*g*=0.1	0.822	0.530	0.644
Cairo Bombing	l2sgd	*c2*=0.9	*ce*=1e-4	0.929	0.938	0.933
Hafr Floods	l2sgd	*c2*=0.5	*ce*=1e-4	0.891	0.776	0.829
Dragon Storms	ap	*eps*=1e-5	-	0.659	0.569	0.611
Beirut Explosion	l2sgd	*c2=0.15*	*ce=1e-2*	0.692	0.874	0.772
CoVID-19	arow	*var=0.1*	*g=0.125*	0.676	0.597	0.634
Time-based data setup Type-less LMR						
Jordan Floods	pa	*c*=0	*e_sensitive*=TRUE	0.837	0.837	0.837
Kuwait Floods	pa	*c*=0	*e_sensitive*=TRUE	0.904	0.904	0.904
Cairo Bombing	pa	*c*=2	*e_sensitive*=TRUE	0.714	0.708	0.708
Hafr Floods	l2sgd	*c2*=0.75	*ce*=1e-6	0.861	0.861	0.859
Dragon Storms	pa	*c*=0	*e_sensitive*=TRUE	0.872	0.872	0.872
Beirut Explosion	l2sgd	*c2*=0.3	*ce*=1e-3	0.701	0.703	0.701
CoVID-19	pa	*c*=0	*e_sensitive*=TRUE	0.928	0.928	0.928
Time-based data setup Type-based LMR						
Jordan Floods	arow	*var*=0.25	*g*=0.1	0.776	0.778	0.775
Kuwait Floods	pa	*c*=0	*e_sensitive*=TRUE	0.891	0.891	0.891
Cairo Bombing	l2sgd	*c2*=0.05	*ce*=1e-6	0.740	0.741	0.737
Hafr Floods	l2sgd	*c2*=0.05	*ce*=1e-6	0.882	0.883	0.882
Dragon Storms	pa	*c*=0	*e_sensitive*=TRUE	0.880	0.880	0.880
Beirut Explosion	arow	*var*=0.5	*g*=0.16	0.617	0.643	0.621
CoVID-19	pa	*c*=0	*e_sensitive*=TRUE	0.901	0.901	0.901

Table C.6. The best hyper-parameters and results of the BERT$_{LMR}$ model over IDRISI-RA 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.

	Random							Time-based						
Event	e	bs	lr	sl	P	R	F1	e	bs	lr	sl	P	R	F1
Type-less														
Jordan Floods	3	8	3e-5	256	0.954	0.957	0.953	3	8	3e-5	128	0.911	0.916	0.903
Kuwait Floods	4	16	3e-5	256	0.935	0.925	0.928	3	16	3e-5	128	0.895	0.905	0.893
Cairo Bombing	3	8	3e-5	256	0.995	0.986	0.989	2	8	3e-5	128	0.934	0.939	0.936
Hafr Floods	4	8	3e-5	128	0.883	0.883	0.879	4	8	3e-5	256	0.878	0.897	0.878
Dragon Storms	3	8	3e-5	128	0.878	0.873	0.870	4	8	4e-5	128	0.882	0.868	0.869
Beirut Explosion	4	8	3e-5	128	0.885	0.851	0.855	4	8	3e-5	256	0.601	0.611	0.582
CoVID-19	3	8	3e-5	128	0.889	0.884	0.881	3	8	3e-5	256	0.896	0.914	0.897
Type-based														
Jordan Floods	4	8	3e-5	128	0.916	0.907	0.908	4	8	3e-5	256	0.880	0.872	0.862
Kuwait Floods	4	8	3e-5	128	0.933	0.925	0.925	3	8	3e-5	256	0.874	0.892	0.879
Cairo Bombing	4	8	3e-5	128	0.984	0.970	0.975	4	8	3e-5	256	0.930	0.935	0.931
Hafr Floods	4	8	3e-5	256	0.870	0.857	0.856	3	8	3e-5	256	0.841	0.854	0.838
Dragon Storms	4	8	3e-5	256	0.798	0.789	0.787	4	8	3e-5	256	0.726	0.722	0.714
Beirut Explosion	4	8	4e-5	256	0.854	0.821	0.813	4	8	5	128	0.616	0.635	0.596
CoVID-19	4	8	3e-5	256	0.895	0.898	0.893	4	8	3e-5	128	0.888	0.898	0.886

Table C.7. The best hyper-parameters and results of the BERT$_{LMR}$ model under *disaster domain transfer* setting
e, bs, lr, and sl refer to the hyper-parameters, number of epochs, batch size, learning rate, and sequence length, respectively.

Source-Target	e	bs	lr	sl	P	R	F1
Type-less							
BMB-BMB	4	8	3e-5	128	0.935	0.917	0.918
BMB-FLD	4	8	3e-5	128	0.584	0.664	0.596
BMB-PND	4	8	3e-5	128	0.854	0.840	0.839
BMB-STR	4	8	3e-5	128	0.839	0.833	0.831
FLD-BMB	3	8	3e-5	256	0.843	0.762	0.779
FLD-FLD	3	8	3e-5	256	0.940	0.928	0.930
FLD-PND	3	8	3e-5	256	0.898	0.887	0.887
FLD-STR	3	8	3e-5	256	0.839	0.819	0.826
PND-BMB	3	8	3e-5	128	0.526	0.524	0.488
PND-FLD	3	8	3e-5	128	0.686	0.744	0.687
PND-PND	3	8	3e-5	128	0.889	0.884	0.881
PND-STR	3	8	3e-5	128	0.749	0.716	0.728
STR-BMB	3	8	3e-5	128	0.574	0.535	0.518
STR-FLD	3	8	3e-5	128	0.491	0.544	0.501
STR-PND	3	8	3e-5	128	0.792	0.768	0.773
STR-STR	3	8	3e-5	128	0.878	0.873	0.870
Type-based							
BMB-BMB	4	8	3e-5	256	0.972	0.934	0.945
BMB-FLD	4	8	3e-5	256	0.396	0.505	0.422
BMB-PND	4	8	3e-5	256	0.876	0.859	0.858
BMB-STR	4	8	3e-5	256	0.798	0.785	0.786
FLD-BMB	3	8	3e-5	256	0.850	0.763	0.786
FLD-FLD	3	8	3e-5	256	0.937	0.935	0.933
FLD-PND	3	8	3e-5	256	0.854	0.846	0.842
FLD-STR	3	8	3e-5	256	0.855	0.838	0.842
PND-BMB	4	8	3e-5	256	0.513	0.507	0.481
PND-FLD	4	8	3e-5	256	0.603	0.703	0.622
PND-PND	4	8	3e-5	256	0.895	0.898	0.893
PND-STR	4	8	3e-5	256	0.781	0.743	0.752
STR-BMB	4	8	3e-5	256	0.469	0.428	0.418
STR-FLD	4	8	3e-5	256	0.406	0.466	0.421
STR-PND	4	8	3e-5	256	0.743	0.715	0.72
STR-STR	4	8	3e-5	256	0.798	0.789	0.787

APPENDIX D: DETAILED DATA SETUPS FOR GENERALIZABILITY EXPERIMENTS

Tables D.1 and D.2 show the detailed data setups for the *domain* and *geographical* generalizability experiments, respectively.

Table D.1. 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 D.2. 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		✓				

APPENDIX E: LOCATION MENTION DISTRIBUTION

The location distribution of the English disaster-specific tweet datasets are shown in Figures E.1-E.5).

Figures E.6-E.9 show the distribution of top 15 frequent location mentions in IDRISI-RE dataset per disaster event.

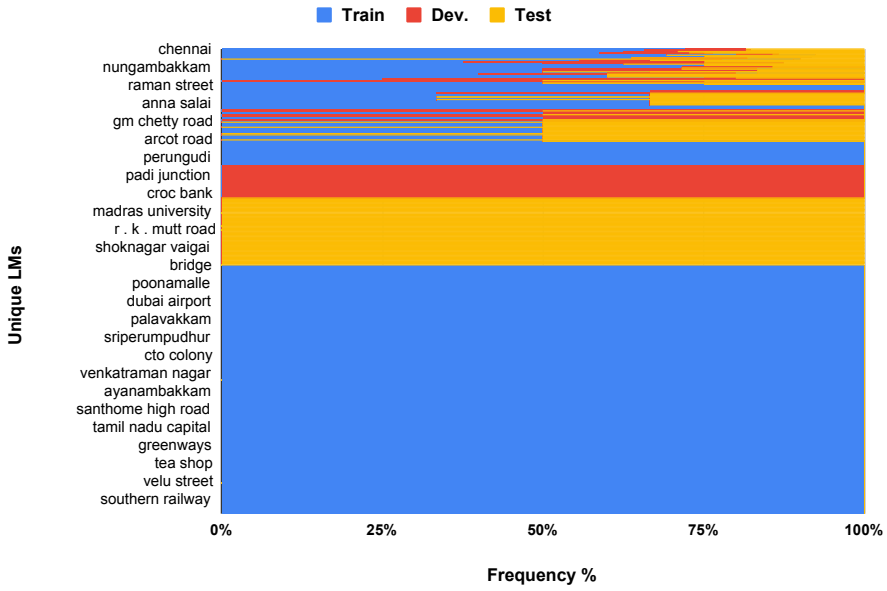


Figure E.1. The LMs distribution across training, development, and test data for Chennai Floods disaster dataset.

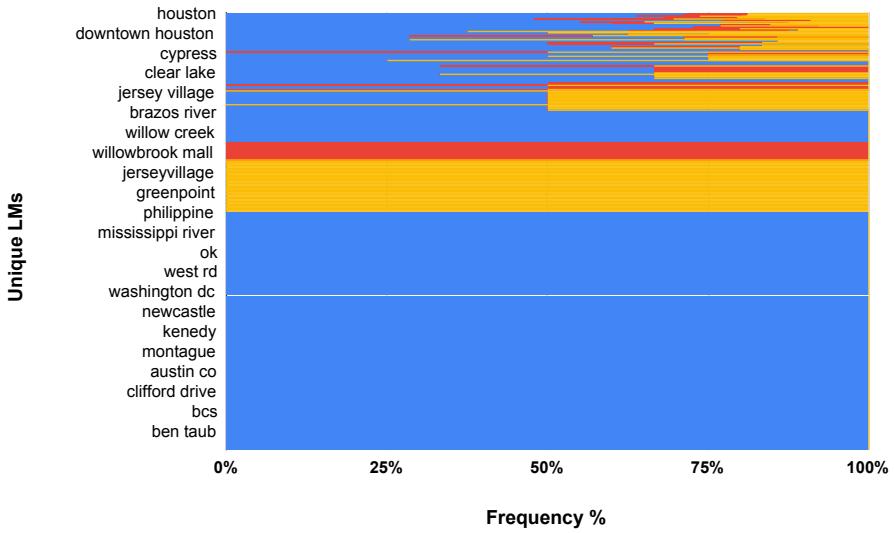


Figure E.2. The LMs distribution across training, development, and test data for Houston Floods disaster dataset.

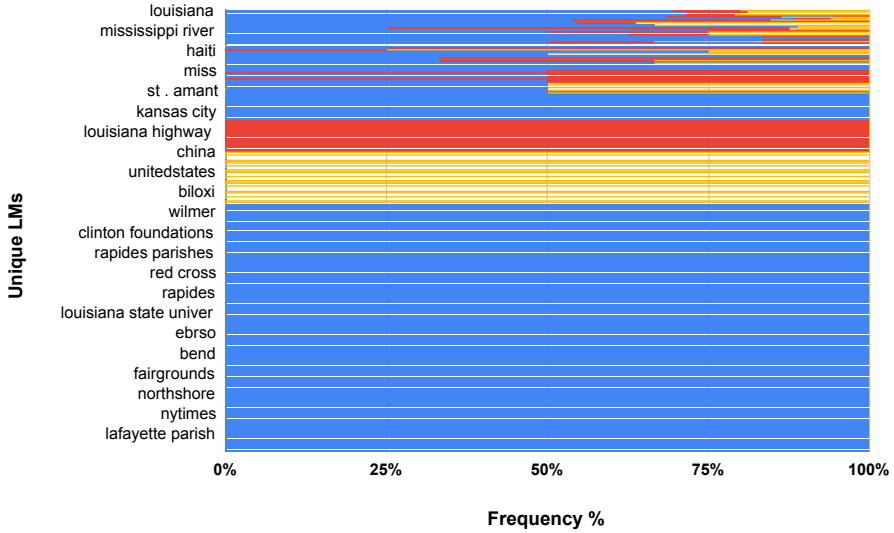


Figure E.3. The LMs distribution across training, development, and test data for Louisiana Floods disaster dataset.

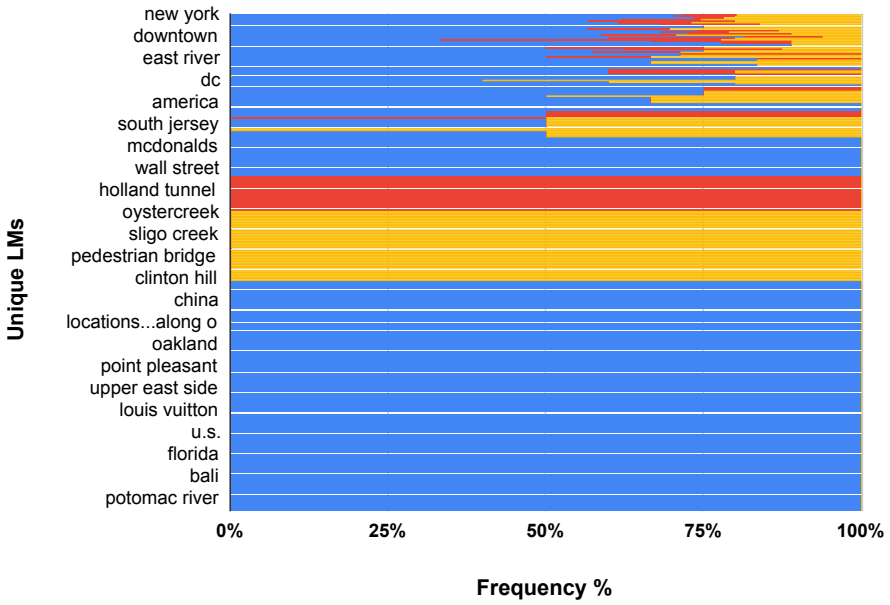


Figure E.4. The LMs distribution across training, development, and test data for Hurricane Sandy disaster dataset.

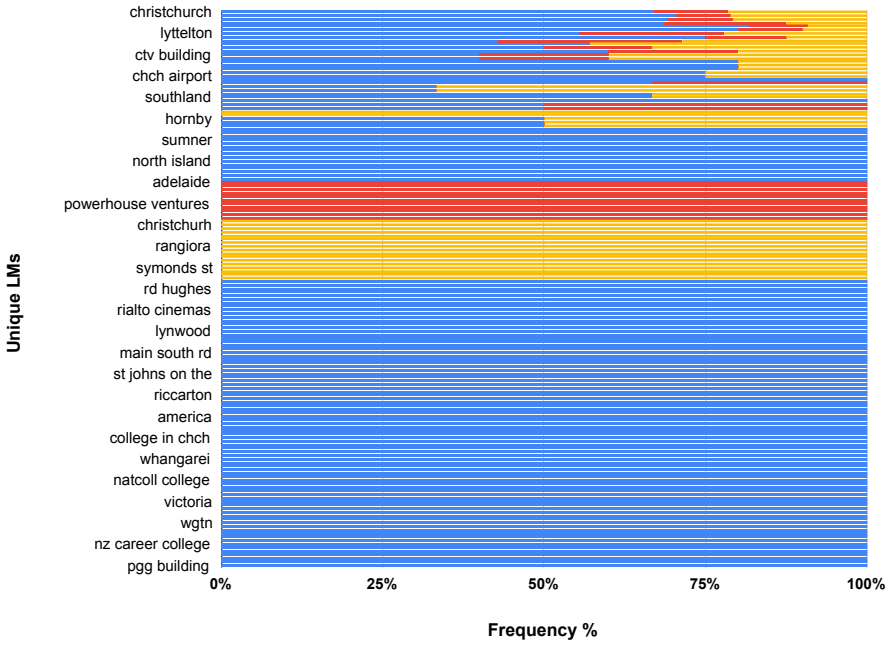


Figure E.5. The LMs distribution across training, development, and test data for Christchurch Earthquake disaster dataset.

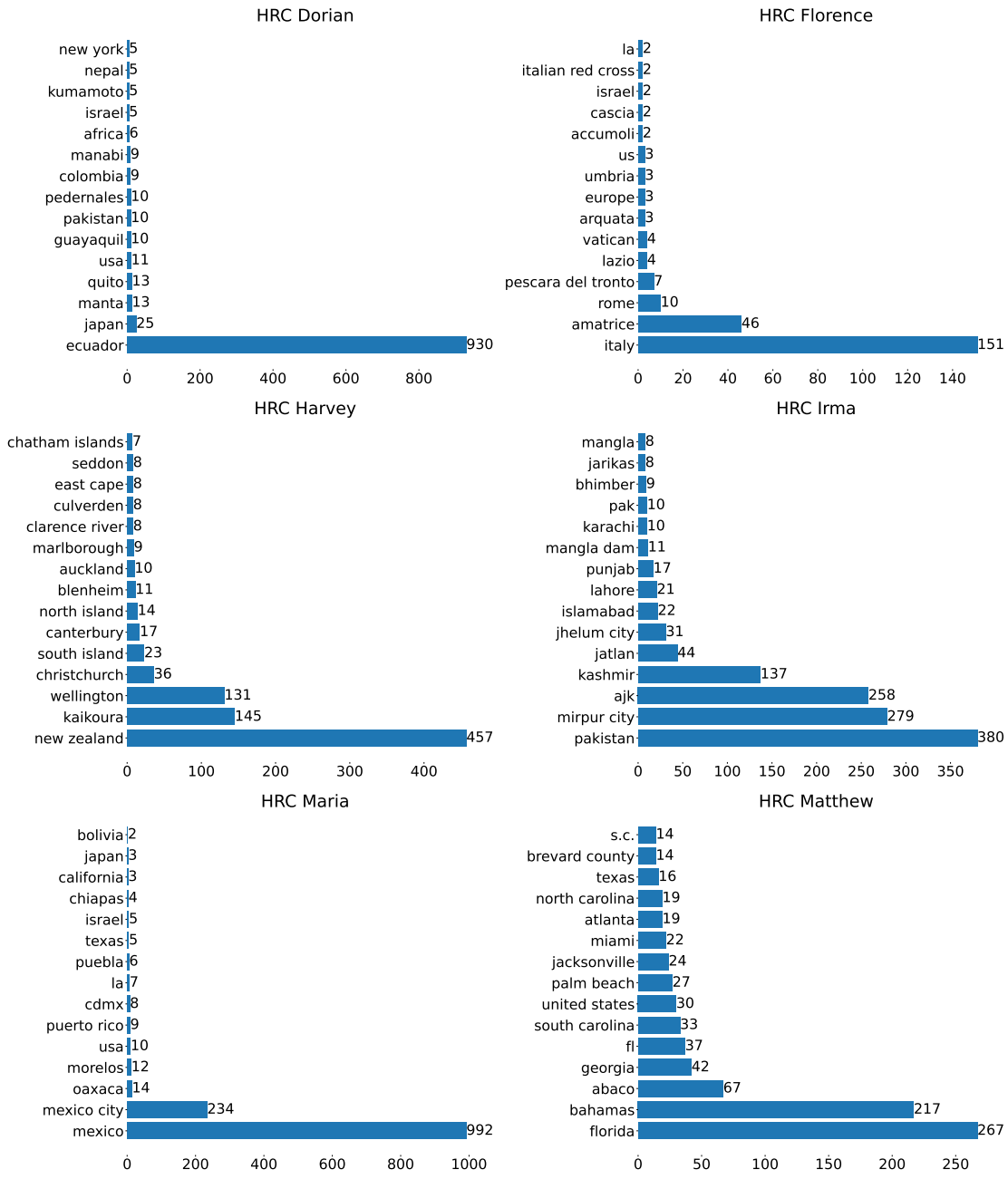


Figure E.6. The distribution of top 15 location mentions in IDRISI-RE per hurricane event.

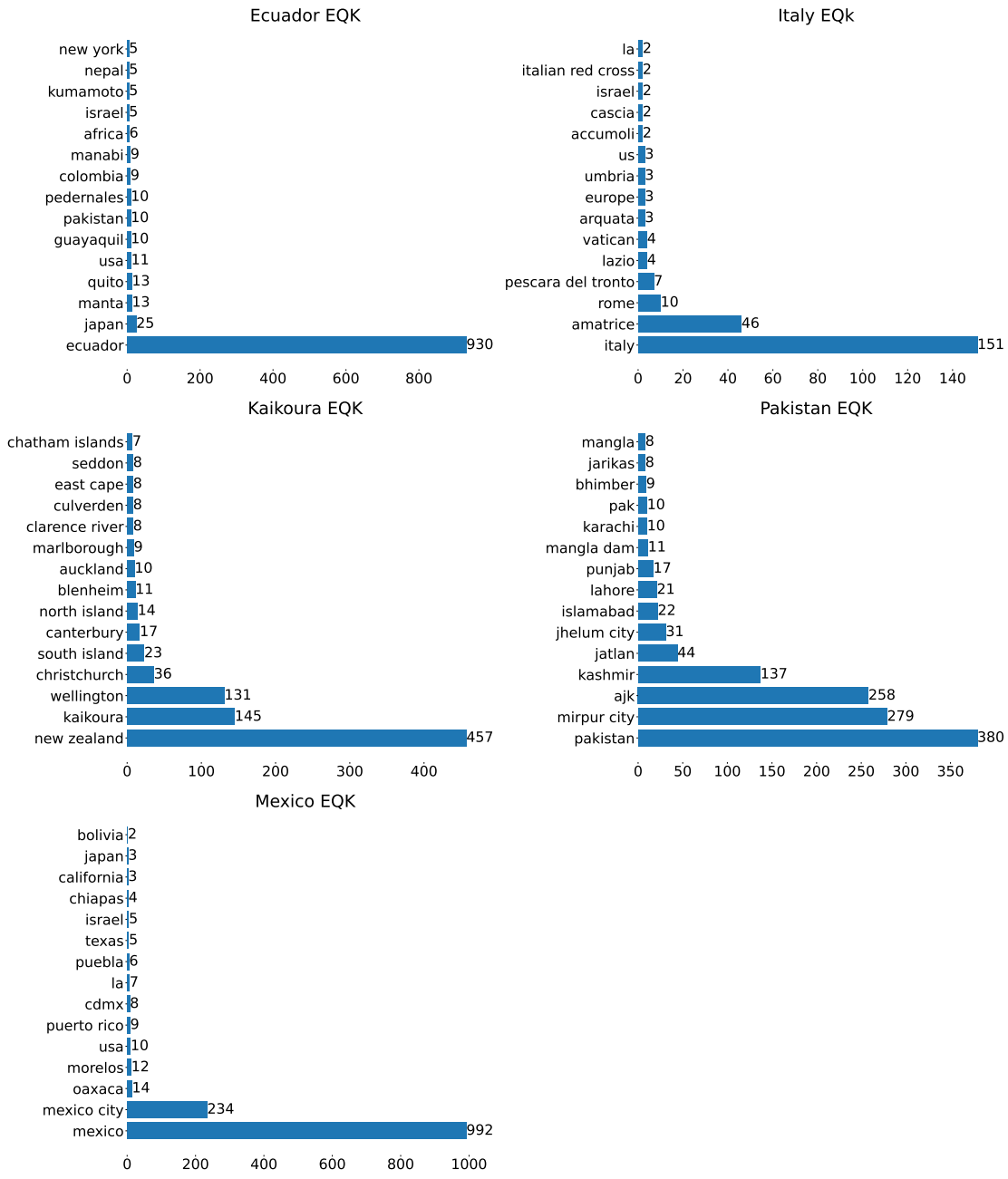


Figure E.7. The distribution of top 15 location mentions in IDRISI-RE per earthquake event.

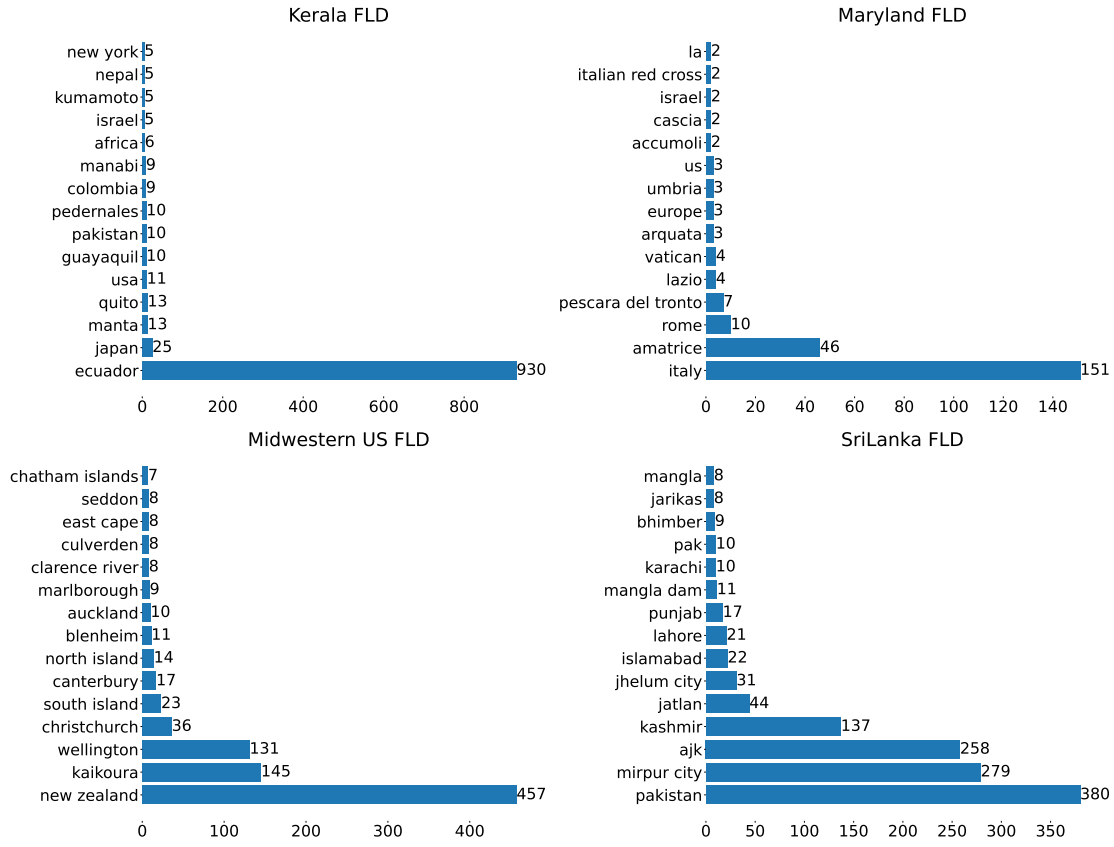


Figure E.8. The distribution of top 15 location mentions in IDRISI-RE per flood event.

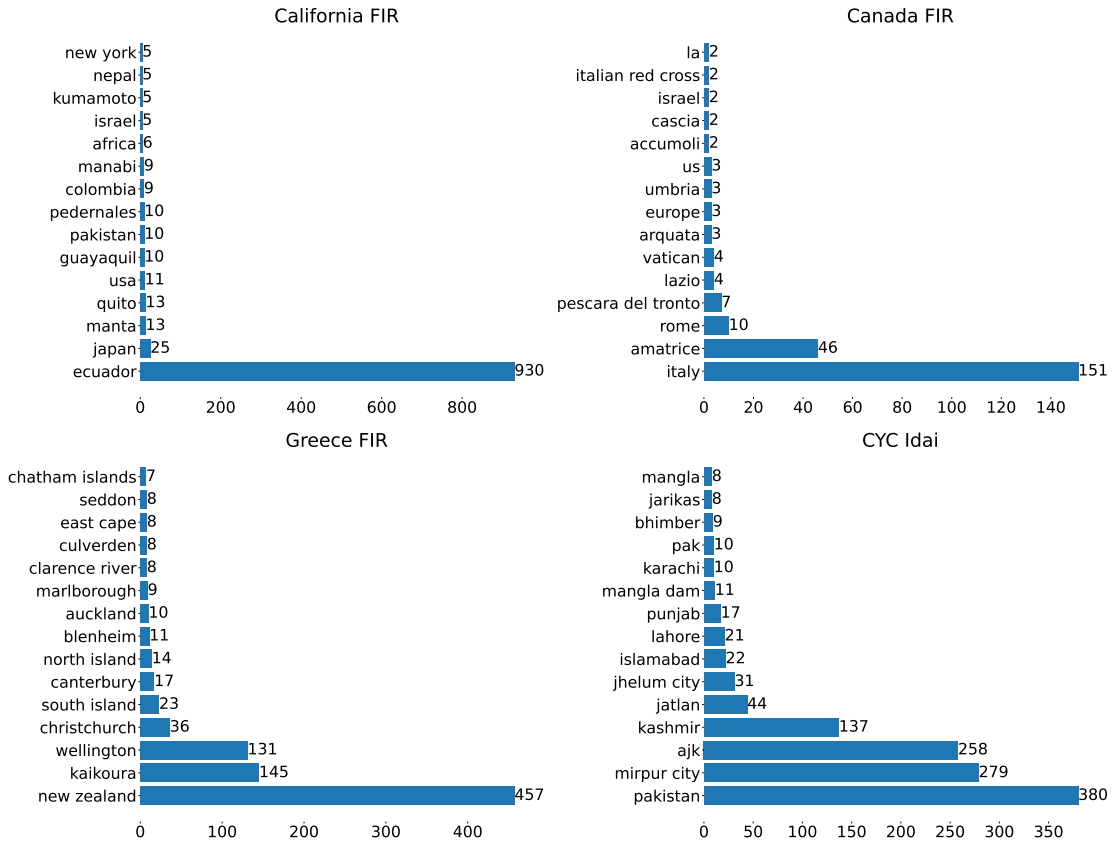


Figure E.9. The distribution of top 15 location mentions in IDRISI-RE per wildfire/cyclone event.