International Journal of Disaster Risk Reduction 78 (2022) 103107

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

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

When a disaster happens, we are ready: Location mention recognition from crisis tweets[☆]

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

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

ARTICLE INFO

Keywords:
Geolocation recognition
Social good
Twitter

ABSTRACT

Geolocation information is important for humanitarian organizations to gain situational awareness and deliver timely aid during disasters. Towards addressing the problem of recognizing locations, i.e., Location Mention Recognition (LMR), within social media posts during disasters, past studies mainly focused on proposing techniques that assume the availability of abundant training data at the disaster onset. In this work, we adopt the more realistic assumption that *no* (i.e., zero-shot setting) or *as little as a few hundred* examples (i.e., few-shot setting) from the just-occurred event is available for training. Specifically, we examine the effect of training a BERT-based LMR model on past events using different settings, datasets, languages, and geo-proximity. Extensive empirical analysis provides several insights for building an effective LMR model during disasters, including (i) Twitter crisis-related and location-specific data from geographically-nearby disaster events is more useful than all other combinations of training datasets in the zero-shot monolingual setting, (ii) using as few as 263–356 training tweets from the target language (i.e., few-shot setting) remarkably boosts the performance in the cross- and multilingual settings, and (iii) labeling about 500 target event's tweets leads to an acceptable LMR performance, higher than F_1 of 0.7, in the monolingual settings. Finally, we conduct an extensive error analysis and highlight issues related to the quality of the available datasets and weaknesses of the current model.

1. Introduction

Twitter has shown to be an effective medium for gaining situational awareness and performing an urgent need assessment of the affected population during sudden-onset disasters [2,3]. The microblogging platform often breaks news and is thus considered a

[☆] This manuscript describes a major extension to an earlier version published at COLING'20 [1]. Improvements over the previous work include fine-tuning the BERT-based LMR model for all setups over different hyper-parameters. This makes the results more reliable and the conclusions more solid. Additionally, we study the cross- and multilingual training data setup under the condition of absent target training data. For that, we conducted experiments on Italian and Turkish languages. This way we investigate various dimensions for data transfer learning including the type of documents (e.g., web documents, tweets), the type of entities (e.g., LOC, ORG, etc.), the type of the disaster events (e.g., earthquake, flood, etc.), the geographical proximity of disaster events (e.g., within same country or continent), and the language (e.g., English, Italian, etc.). We further provide a deeper analysis of the results. We also study the effect on the model's performance when incrementally labeling tweets from the target disaster event. This allows us to recommend the minimum number of tweets to label at the onset of disaster events to obtain an acceptable performing LMR model. Moreover, we conduct a failure analysis to understand the causes behind the model failure and to gain some insights on how to improve it further for future work.

[*] Corresponding author. Computer Science and Engineering Department, College of Engineering, Qatar University, Doha, Qatar.
E-mail addresses: rs081123@qu.edu.qa (R. Suwaileh), telsayed@qu.edu.qa (T. Elsayed), mimran@hbku.edu.qa (M. Imran), hsajjad@hbku.edu.qa (H. Sajjad).

https://doi.org/10.1016/j.ijdrr.2022.103107
Received 30 August 2021; Received in revised form 1 June 2022; Accepted 5 June 2022
Available online 15 June 2022

low-latency source for timely access to information. As disaster events happen and develop, many people rush to Twitter to report their updates, incidents around them, and more importantly, request assistance from government authorities and fellow citizens. These requests often contain location information, usually in the unstructured form [4]. This information can be invaluable for response authorities to be aware of the situation on the ground, and to effectively reach the affected people on the spot. Due to the limitations of using the traditional (e.g., hot-lines) and non-traditional (e.g., physical remote sensing) technologies, response authorities are increasingly incorporating *social sensing* in their operations [5].

For effective situational awareness via social sensing, responders depend on the availability and quality of geolocation information on Twitter [6]. First, they need to determine whether the reported incidents or requests are within their jurisdiction areas; otherwise, they redirect to the responsible authority [7]. Second, responders use geolocation information to locate events and create crisis maps for timely decision-making and response deployment and evaluation.[1] These include (1) spatial situational awareness maps, (2) hotspots maps of causalities, damages, and resources scarcity, (3) eyewitnesses and resources maps (e.g., food, shelters, etc.), (4) population mobility maps (e.g., evacuation behavior), or (5) disaster impact maps to plan for recovery, among others.

Response agencies' requirements to use Twitter for situational awareness or mapping tasks (as listed above) vary. In a participatory design workshop, police officers, firefighters, and paramedics, among other responders, provided example tweets that they look for on Twitter during disasters. Most examples contain fine-grained location mentions (LMs), e.g., streets, buildings, etc. [8]. On the other hand, a survey for emergency managers and disaster responders indicated their need for coarse-grained locations that they use for immediate relief activities and impact assessments [9–11].

Furthermore, the effectiveness of geolocation information has been shown during many real-world disasters. For instance, during the earthquake in Port-au-Prince in Haiti in 2010, the emergency personnel exploited Ushahidi platform[2] to map geotagged tweets.[3] Similarly, in the aftermath of the Christchurch Earthquake in New Zealand, geo-tagged e-mails and messages were used to map the earthquake incidents and consequences. Furthermore, the Red Cross in the United States used around 10k photos and videos of damages by the Sandy Hurricane in 2012. This data was shared over Instagram by eyewitnesses with geo-tagged information.[4]

However, a major obstacle toward using raw Twitter data for situational awareness is that only approximately 1–4% of tweets are tagged for places or GPS coordinates [12]. What is more worrisome is that Twitter announced on June 18, 2019 that it would removed the geotagging feature in tweets as people tend to use imprecise geolocation for their shared content.[5] Despite that, eyewitnesses of incidents still mention locations within the text of tweets, which indeed emphasizes the need for *automatic* location extraction and inference tools. While extensive research has been conducted on processing tweets for crisis management [13–21], a limited focus has been given to the *Location Mention Prediction* (LMP) problem [22].

The LMP systems consist of two main components; a *Location Mention Recognition* (LMR) model that extracts toponyms, i.e., places or location names from tweets, and a *Location Mention Disambiguation* (LMD) model that resolves the potential LMs to defined geographical points from a geo-positioning database (i.e., gazetteer). In Fig. 1, we illustrate our perception of the role of the LMR module in the disaster response pipeline. The upstream modules (filtering blocks) aim to keep informative posts of high quality for LMP and discard all other content (e.g., non-relevant, spam, fake or rumors, bot-generated, useless content). The arrangement of upstream modules can be changed according to the responders' needs. The downstream processing aims to geographically analyze and visualize the Twitter information. This processing can run manually (e.g., by stakeholders) or automatically (e.g., using machine learning models).

In this work, we only focus on the LMR task. Two main factors that influence the robustness of an LMR system are: (i) the learning model, and (ii) the dataset used to train the learning model. 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 general task of LMR by definition, which aims to extract the 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 slangs [23]. Moreover, entities mentioned in tweets may have inconsistent capitalization, which is one of the main features standard NER systems rely on [22]. The second common approach is employing gazetteers to maintain highly precise location mentions recognizers. The gazetteer-based models are restricted to the geographical coverage of their databases. Additionally, the noisy nature of the Twitter stream contradicts the nature of gazetteers causing the so-called mismatch problem. More 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.

To address the aforementioned challenges, we employ the BERT model as it achieved state-of-the-art results in many NLP tasks [24]. Moreover, deep learning models eliminate the cost of hand-crafting features, which allows us to overcome the limitations of gazetteer-based and traditional learning models that highly depend on feature engineering. While most of the 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 the efficiency of deploying the LMR models in emergencies. We

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

[2] www.ushahidi.com/.

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

[4] www.techradar.com/news/internet/dealing-with-disaster-how-social-media-is-helping-save-the-world-1203809 The Federal Emergency Management Agency (FEMA) and the City of New Orleans had also employed the eyewitness information to plan their response efforts during Hurricane Isaac 2012. The reported data contains flooding locations, road closures, and utility outages.

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

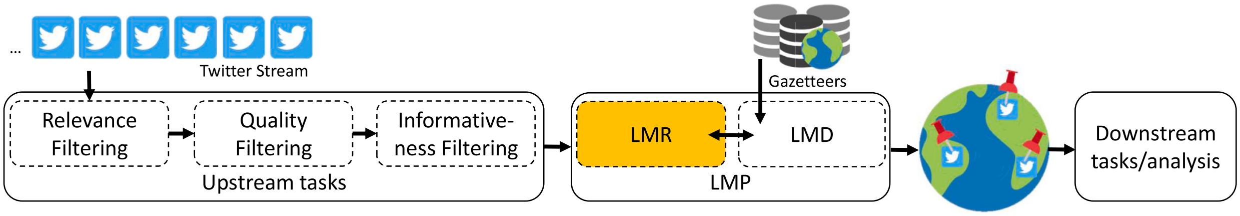

Fig. 1. Positioning the LMR task in the disaster response pipeline.

run our experiments in two settings related to training data augmentation strategies: (i) zero-shot setting, where there is no available training data at all, 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 training 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 the 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 in an effort to empirically provide answers to them. We fix our learning model to a state-of-the-art model (i.e., BERT-based 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 gives better performance than using 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 training on past disasters of the same type as the target disaster generally improves the performance. While labeled data from the target event yield the best performance, we note that using labeled data from disasters that happened nearby is helpful when the target labeled data is not available. Additionally, training on little labeled data, around 263–356 tweets, from the target language significantly boosts 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-crafted features by exploiting a BERT-based LMR model.
- We explore different data transfer setups with a variety of aspects 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 reasonabley-performing LMR model.
- We conduct failure analysis on the BERT-based model to gain insights for the future development of LMR models.

For reproducibility, we make the NER datasets (in BILOU format), the steps to acquire the licensed Twitter crisis-related datasets, the best performing models, and the steps to run the BERT-based LMR model publicly available.[6]

The rest of the article is organized as follows. We summarize related work in Section 2. We present an overview of the LMR problem and define it formally in Section 3. We discuss the experimental setup in Section 4. We thoroughly analyze the results, answer the research questions, and discuss the lessons we learned in Section 5. To gain insightful lessons for future development of LMR models, we conduct failure analysis in Section 6. We discuss the implications and limitation of our study in section 7. We finally conclude and

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

list some future directions in Section 8.

2. Related work

In this section, we discuss different uses of Twitter information for crisis management (Section 2.1). We then focus on research studies that perform geolocation information extraction from Twitter (Section 2.2). Lastly, we describe efforts toward automating the location extraction process from tweets during disaster events (Section 2.3).

2.1. Information extraction from Twitter for crisis management

Recently, Hiltz et al. [25] conducted a survey to prioritize the tasks (features in systems) 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. Invaluable efforts have been made by technologists to enable utilizing social media for preparedness, relief, and recovery of emergencies [13]. The proposed tasks involve but are not limited to, detecting disasters and incidents [14,15,26], filtering relevant tweets [17–19,27], summarizing them [16], identifying situational reports [20,28–30], identifying actionable information [3,7,21,31], and geolocation and other types of information extraction [22, 32–35].

Nevertheless, the current solutions are rarely deployed by relief organizations [36,37] due to several reasons. For instance, the unreliability of information, the inefficiency of solutions in disaster scenarios, the lack of readiness for customization for the different needs of different stakeholders, to name a few. Fortunately, there are some recent efforts to bridge the gap between the technologists and responders from the relief organizations by understanding their needs and the utility of existing solutions [7,8,21,25,31,38].

For example, Hughes & Shah [38] 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 et al. [28] was the first to define a list of 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) [31], availability and needs for different resources (generally or for a specific location), and activities of relief organizations [21]. Further efforts have been put to establish the definition of actionability and define the criteria for ranking actionable tweets [39]. Zade et al. [7] have then 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 et al. [8] investigated the characteristics of actionable tweets by interviewing administrators, telecommunicators, and first responders.

2.2. Geolocation extraction for crisis management

According to Hiltz et al. [25]; *grouping social media content on a map by their geographic locations* is the most demanded feature by responders. Another important feature was to *automatically geo-tag 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 [7,8]. Geolocation information allows responders to locate resources (e.g., food, shelters, etc.) and their status, activities (e.g., evacuation zones), casualties, and damages [8–11]. Thereby, since the early emergence of social media, relief organizations have utilized geo-tagged content in their activities. A good case in point is the exploitation of mashup technologies to map the disaster event and improve situational awareness. These applications follow a crowd-source-based model for collecting updates (e.g., comments, photos, or videos on incidents, etc.). There are different mashup platforms that can be used to map crisis data (real-time mapping of situational and actionable data), such as Ushahidi,[7] Geofeedia,[8] ESRI-ArcGIS,[9] Google Crisis Response,[10] and Factal.[11]

Many such services were developed and deployed during past disasters, including Port-au-Prince Earthquake in Haiti in 2020,[12] Typhoon Haiyan in the Philippines (Visov) in 2013, and Chennai Floods in India in 2016.[13] 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.[14]

Additionally, Roy et al. [40] created dynamic disruption maps from Twitter data to visualize types of distributions and their status. For that, they employed the NLTK NER model [41] to extract location mentions from tweets' text. Hong & Frias-Martinez [42]; on the other hand, used the geolocation information extracted from Twitter data to model the patterns of evacuation flows at different coarse-grained geographical levels such as country, state, and areas. This study is limited to users having the automatic geotagging feature of tweets enabled which allows tracking and analyzing users' location during the disaster event. However, this feature has been

[7] www.ushahidi.com/.

[8] en.wikipedia.org/wiki/Geofeedia.

[9] www.esri.com/en-us/disaster-response/overview.

[10] crisisresponse.google/.

[11] www.factal.com.

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

[13] www.ushahidi.com/support/examples-of-deployments.

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

removed by Twitter in 2019.[15] Following a similar line of analysis, Roy & Hasan [43] used the geolocation information of tweets to infer the individual evacuation behavior of people 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 et al. [44] 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 that is used by the mapping subsystem to pin the content on the map. The system started operating in early 2015 and improved further in 2017 [45]. Kosugi et al. [46] introduced a better and user-friendly web-based real-time system to support collaborative response with the same use cases including (1) reporting, (2) displaying reports (latest or nearby) and facilities' locations (evacuation places or disaster base medical centers), and (3) searching reports. Several other state-of-the-art applications of social media informatics in disasters are reviewed by Zhang et al. [47] and their challenges are highlighted. The authors also determined 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 on different geolocation tasks (e.g., user location prediction, tweet location prediction, and location mention extraction and disambiguation) and data domains (news articles, research articles, social media posts, etc.) [22]. The central task among these is the LMR task as it aids them by extracting evidence from the text for the user location, tweet location, incident location, etc. Location taggers have to address many challenges associated with the nature of the Twitter stream, such as tweet sparsity (short text of 280 character limit), processing noisy tweet stream (slang with lots of nicknames, abbreviations, shortcuts, etc.), processing rapidly changing stream with new emerging LMs [48]. Other technical challenges involve the time constraint of the solutions, the efficiency of deployment, and the coverage of a variety of areas instead of limiting the search space to pre-loaded gazetteers [48]. There are also task-specific challenges that have to be addressed, including toponymic polysemy, i.e., toponyms having multiple meanings. For example, "Washington" could mean the "Washington DC state" or the president "George Washington" [49]. Additionally, the lack of context when mentioning references of the general locations is a common challenge in crisis-related social media posts [6]. For example, during flooding, users could mention "the river" instead of mentioning the name of the flooded river because the geographic context is known among users discussing the event.

2.3. Location Mention Recognition for crisis management

Existing LMR studies exploit different techniques and features to extract Location Mentions (LMs) from text. Several of the proposed approaches are gazetteer-based in which public geo-positioning databases are employed, such as Geonames[16] [50,51], OpenStreetMap[17] [32,51,52], Official New Zealand gazetteer[18] [53], and Alexandria Digital Library Gazetteer[19] [54], among others.

One important processing step in gazetteer-based approaches is the matching phase. Different matching methods were proposed, such as language modeling [52], rule-based methods [50], lexical-based methods [51], and ensemble-based parsers (lexico-semantic and machine learning techniques) [53]. Although the gazetteer-based models achieve relatively high precision, as they verify the candidate LMs using gazetteers, their main drawback is the inability to detect toponyms that do not appear in the gazetteers. To evade acquiring annotated data at all, which is a bottleneck during emergencies, Al-Olimat et al. [32] proposed an unsupervised statistical tagger that identifies the LMs by traversing a tree of n-grams while matching them against region-specific gazetteers. While all those studies explore ways to improve the recognition methods, we explore, in this work, the choice of training data to overcome this challenge. In particular, we study the effectiveness of exploiting available in-, cross-, and out-of-domain training data to build LMR models that learn the patterns of location in tweets without relying on gazetteers.

Alternatively, there are supervised learning techniques that are exploited for LMR. In the fifth Australasian Language Technology Association (ALTA) shared task on identifying LMs in tweets [55], the participants explored several techniques such as feature engineering, ensemble classifiers, rule-based classification, knowledge infusion, CRFs sequence labelers, and semi-supervision techniques. They also explored different feature types, including geospatial, structural, and lexical features. Using noisy gazetteers and a variety of hand-engineered features showed improvements in many comparisons. However, there is still a big room for improvement to build robust LMR models with a minimal cost, which we explore through training BERT-based LMR model (with no expensive hand-engineered features or gazetteers) with different combinations of available data (with no data annotation, at least initially).

An intuitive solution for the LMR problem is to employ existing NER tools, since the LMR task is a subtask of the NER task. For instance, the StanfordNER model was employed for this purpose by several studies [55–57] after retraining it on Twitter datasets to effectively identify the location mentions in tweets. Nizzoli et al. [58] employed TAGME[20] to identify meaningful short phrases in the text and link them to articles on Wikipedia. Wang & Hu [59]: alternatively, adopted the three top systems from the "Toponym Resolution in Scientific Papers" SemEval 2019 task [60]. Although those studies used general-purpose taggers, their experiments did not investigate the gains and losses of considering all types of entities against using only the location entity. In our work, we extensively investigate this aspect to understand the effect of focusing the training on location entities on the performance of the LMR models.

The recent trend of Natural Language Processing (NLP) development is using deep learning approaches and BERT-like models.

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

[16] http://www.geonames.org/.

[17] http://www.openstreetmap.org/.

[18] http://www.linz.govt.nz/placenames/find-names/nzgazetteer-official-names.

[19] www.library.ucsb.edu/map-imagery-lab/alexandria-digital-library-gazetteer.

[20] tagme.d4science.org/tagme/.

However, up to our knowledge, there are only a couple of studies that exploit deep learning LMR models for the crisis domain [1,35]. Wang et al. [35] employed BiLSTM-CRF model [33]. Unlike our work, where we exclude the target data from training, these studies used the target events for training and testing. The assumption of the availability of annotated target data is quite impractical during the early phases of emergencies, due to the time constraints and cost of data annotation. In our work, we use a BERT-based model [1].

Among all the reviewed research studies, there is no single study that explores the setups we propose in this paper. Typically, the proposed approaches are trained and tested on the target events, assuming the training data is available at the onset of disaster events, which is often not true. Furthermore, none of the existing studies investigate the usefulness of labeled data from past events, as well as whether geospatial proximity is an essential aspect that has to be considered when choosing the past events for training. Additionally, cross- and multilingual transfer learning techniques are understudied for the location recognition task. The only work that studies the cross-lingual setup is by Gelernter & Zhang [61]; but the experiments are conducted on English and Spanish languages with translation tools, not the Turkish and Italian languages that we cover in our work. Although there is a lack of large-scale representative recognition datasets, at the time of this writing, there is no single study that explores the cost of collecting annotations to mitigate the effect of delaying the relief authorities' response during disaster events. In this work, we present an exploration for the cost of collecting annotated dataset at the onset of disaster events.

3. Problem overview

At the onset of a disaster event, response organizations and first-responders relying on Twitter need geolocation information of reports or tweets about the crisis event in general as well as those seeking immediate help. In this case, the goal is to detect the mentions of locations in the textual content of a tweet reporting an event or asking for help. This is different than determining the location information present in the Twitter user's profile, which is often used to find the user's home location.

Table 1 shows a few tweets with different types of location mentions posted at the time of real-world disaster events. Tweet #1 during the Chennai Floods in 2015 is requesting a boat to a very specific location (in this case a street name). Similarly, tweet #2 is an important situational awareness report about a bridge being collapsed. The author mentions the name of the bridge i.e., "Adayar Bridge Saidapet", which represents a very specific fine-grained location information. Tweet #3 reports the water level at Greens Bayou with respect to the National Weather Service standards during the Houston Floods in 2016. The full address is mentioned in the tweet as "Greens Bayou at Ley Rd, Houston, TX". On the other hand, tweets #4 & #5 share floods updates and warning during Houston and Louisiana Floods in 2016, respectively. In both tweets, fine-grained locations are mentioned (a river and streets). Differently, Tweet #6 reports a storm heading south from Evangeline to Allen areas during Louisiana Floods in 2016. Tweet #7 conveys very important information about kids safety at "Cashmere kindergarten". Similarly, tweets #8 & #9 report flooding on the roads of "Ocean city, New Jersey", and "East 8th street" and "Avenue C" streets caused by Hurricane Sandy in 2012.

We notice that the geolocation granularity of place names mentioned in these tweets vary from coarse-grained to fine-grained. While fine-grained locations are considered more actionable, in this work, we do not distinguish between fine and coarse-grained locations. We aim to recognize and extract *all* types of toponyms (i.e., places or location names) from tweets.

Accordingly, we define the **Location Mention Recognition (LMR)** task as *the automatic extraction of toponyms from text*. In this work, we limit the scope from two angles; we focus on *tweets* and more specifically *crisis-related* tweets that are shared *during emergencies and natural disasters*.

The problem 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 $L_t = \{l_i; i \in [1, n_t]\}$ in the tweet t, where l_i is the ith location mention and n_t is the total number of location mentions in t, if any. Each location mention may span one or more *tokens*. In this work, we follow the *BILOU* annotation scheme with 5 classes,[21] due to its superior performance over the commonly adopted *BIO* scheme [62–64]. In the *BIO* scheme, labels identify the position of every token in a location mention (LM), e.g., "B" denotes the beginning token of an LM, "I" denotes a token inside an LM, and "O" denotes a token outside of an LM. The *BILOU* scheme extends the *BIO* scheme with two more positional tags, i.e., "L" denotes the last token in an LM and "U" indicates that the LM has only one token such as "London". Therefore, we define the LMR as a *multi-class classification task on the token level.*

4. Experimental setup

In this section, we describe the details of our experimental setup. We present the datasets in Section 4.1 and the experimental configurations in Section 4.2. We then discuss the base LMR model in Section 4.3 followed by the evaluation measures in Section 4.4.

4.1. Datasets

To answer the research questions listed in Section 1, we mainly need three types of datasets: (i) *general-purpose NER dataset*, (ii) *Twitter NER dataset*, and (iii) *crisis-related multilingual Twitter LMR datasets*. Table 2 shows various statistics of all the datasets we used in our experiments, which are described below.

- **General-purpose NER dataset:** A well-known candidate for this category is the CoNLL-2003 NER dataset [65], which comprises newswire text from Reuters, tagged with four different entity types, namely PER, LOC, ORG, and MISC. Overall, the dataset

[21] BILOU classes are: Beginning, Inside, Last, Outside, and Unit.

Table 1
Tweets from real-world disaster events with location mentions (underlined). HRC, EQK, and FLD refer to Hurricanes, Earthquakes, and Floods, respectively.

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

Table 2
Statistics of the datasets used in our experiments. The numbers in parentheses shows the percentage of training data. HRC, EQK, and FLD refer to Hurricanes, Earthquakes, and Floods, respectively. 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 non-location token.

Dataset	Lang	Country	# Docs	# Locs	Annotations				
					B	I	L	U	O
CoNLL-2003	EN	Global	22,137	10,645	1,041 (69)	116 (70)	1,041 (690	6,099 (67)	250,660 (68)
BTC	EN	Global	9,383	2,852	665 (100)	293 (100)	665 (100)	2,187 (100)	168,721 (100)
Chennai FLD	EN	IND	1,500	2,226	840 (80)	275 (78)	840 (80)	1,386 (80)	22,194 (70)
Houston FLD	EN	US	1,500	1,701	508 (81)	155 (84)	508 (81)	1,193 (81)	22,114 (70)
Louisiana FLD	EN	US	1,500	1,396	227 (81)	77 (78)	227 (81)	1,169 (81)	24,620 (69)
HRC Sandy	EN	US	1,996	735	665 (79)	665 (79)	595 (81)	70 (76)	32,525 (70)
ChCh EQK	EN	NZ	1,999	291	220 (79)	220 (79)	544 (80)	71 (77)	27,633 (71)
Milan Blackout	IT	IT	391	705	114 (71)	27 (78)	114 (71)	591 (69)	6,083 (71)
Turkish EQK	TR	TR	2,000	442	28 (68)	0 (0)	28 (68)	414 (71)	16,081 (68)

contains 22,137 sentences and 35,089 entities. We use the standard training and development segments for training and tuning hyper-parameters.

- **Twitter NER dataset:** We use the Broad Twitter Corpus (BTC) as our Twitter NER dataset [66]. It consists of 9,515 tweets, which are tagged with three entity types, namely PER, LOC, and ORG. The dataset has a broad coverage of 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 "New Year". We randomly sampled 90% of the dataset for training and 10% for development.
- **Crisis-related Twitter LMR datasets:** As the main focus of this work is to guide the development of a robust LMR system for toponym extraction from crisis-related tweets, we use several Twitter datasets from real-world disasters to perform extensive experiments. In total, we use seven datasets in this category; three of them represent floods, two earthquakes, one hurricane, and one blackout. The floods datasets consist of 4,500 tweets from *Chennai Floods 2015*, *Louisiana Floods 2016*, and *Houston Floods 2016* [32]. The tweets in these datasets are tagged using several location-related tags. In this work, we only use *inLOC* and *outLOC*, which indicate if the location is within or outside the disaster affected areas, respectively. We further filter out all hashtags used to collect the datasets, thus limiting their effect towards biasing the models' training process. The remaining four datasets in this category are adopted from Middleton et al. [48]. 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 *Christchurch Earthquake* are English language datasets, *Milan Blackout* is in Italian language, and *Turkey Earthquake* is in Turkish language.

4.2. *Experimental configurations*

We used several training and testing configurations in our experiments. In this section, we first define the adopted terminology, then 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 of any topic type (e.g., general or event-oriented); however the target dataset is *always* a crisis-related Twitter dataset.

Furthermore, we use different terminology to articulate the *match* between the source and target datasets in our experiments. 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). 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 the case when we use the NER source dataset with only the LOC entity and discard all other entity types (e.g., PER, ORG, and MISC). By doing so, we convert the LOC entity into the BILOU scheme and the non-LOC entities are labeled as "O".
- *DIS.others* denotes the case 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 the earthquake datasets for training.
- *DIS_<source_type>.others* denotes the case when the target disaster is of 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), the 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 the case 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 datasets but test on Italian or Turkish datasets.
- *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 datasets combined with a few Italian or Turkish tweets, but test on Italian or Turkish datasets.
- *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 datasets.

4.3. LMR model

Pretrained models, such as BERT, have shown impressive performance in the sequence modeling tasks including the NER task [24]. In this work, we employ the *BERT-LARGE-CASED* model in all experiments, except for the Corss- and multi-lingual training, we use the *BERT-BASE-MULTILINGUAL-CASED* model. We added a linear classification layer on top of the BERT model and fine-tune it using the source datasets. For Twitter datasets, we preprocessed the tweets to remove 'RT' token (indicating retweets), user mentions, non-ASCII characters, and URLs. We also segmented the hashtags using the word segment library,[22] since some location mentions appear as subtokens of hashtags in the adopted datasets.

During training, we tuned the batch size, number of training epochs, and the learning rate hyper-parameters using the value ranges recommended by Devlin et al. [24] as: 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 the space of hyper-parameters 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 the reliability of the results, and eventually use the median F_1 score from the five runs. We finally select the best combination of hyper-parameters, and report its F_1 score on the test set. We found the best hyper-parameters are mostly different from the default settings across different training setups [1]. There is no one combination of hype-parameters that fits all experimental setups (refer to results in Table 6, Appendix.A).

4.4. 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,[23] which adopts the evaluation scripts used to evaluate the chunking tasks (e.g., named-entity recognition) in CoNLL-2000 NER shared task [67]. The package evaluates the model's output on entity-level rather than token-level.[24] We use the default micro-average metric to account for the class imbalance issue in our datasets (see class distributions in Table 2).

5. Results and analysis

In this section, we discuss the seven research questions in detail, we present the experiments that we carried out to answer each of

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

[23] https://pypi.org/project/seqeval/.

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

them, and we analyze their results. 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 5.1) or with location entity alone (Section 5.2). We further study the performance when training on *in-domain* and *cross-domain* data in Sections 5.3 and 5.4, respectively. We then study the performance of the LMR models when considering the geographic proximity of disaster events during training (Section 5.5). Moreover, we discuss the effectiveness of cross-lingual setting when training on data in different languages than the language of the tweets discussing the target event in Section 5.6. We finally explore the gain in performance when incrementally acquiring training data from the target disaster in Section 5.7.

5.1. General-purpose (out-of-domain) training with multiple entities (RQ1)

Due to the limited LMR labeled data, we study the effect of using general-purpose NER datasets to train our LMR model. We hypothesize that since the general-purpose NER data is larger in size and has *location* as one of the entity types, using 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. the 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 4.1). Fig. 2 presents the results (the second and third bars from left in all charts). In all datasets, except Hurricane Sandy, the *BTC.ner* model outperforms the *CoNLL.ner* model, suggesting that the general-purpose datasets that are built on documents written in formal language might not be suitable for disaster-related tweets. However, the performance of such models is comparable in the case of Hurricane Sandy.

To answer **RQ1**, we conclude that Twitter NER datasets are more effective than general-purpose NER datasets for training an LMR model for toponyms 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 indeed a valuable resource for training an LMR system when no other data is available e.g., at the onset of a disaster event. 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 Twitter data.

5.2. 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 Fig. 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. We noticed that the improvement is clearly evident in precision but not recall (refer to results in Table 6 in Appendix.A), suggesting that focusing the training on locations only significantly 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, different from other types of entities, location mentions are usually attached to their category (e.g., LOC street, LOC city, etc.) or surrounded by adpositions such as "near", "at", or "10 Km away from". Based on such results, we conclude that the location-specific datasets are better for training

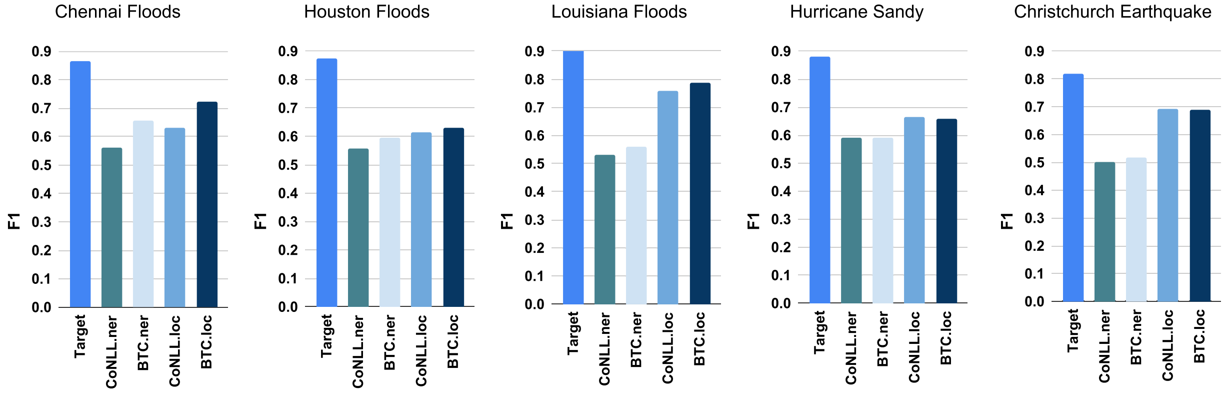

Fig. 2. The results of exploiting out-of-domain general purpose datasets for training an LMR model.

the LMR model compared to the general-purpose NER datasets, which answers **RQ2**.

Although the CoNLL-2003 dataset is 2.4X larger in size 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 datasets (i.e., Hurricane Sandy and Christchurch Earthquake) and CoNLL-2003 dataset, which justifies such an unexpected high performance. For example, one of the top frequent LMs in Hurricane Sandy dataset is "New York", which appears 289 times (208, 24, and 57 times in training, development, and test sets, respectively). 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.

5.3. *Crisis-related training (RQ3)*

Thus far, we confirmed our need for location-specific data to train the LMR system. However, the location mentions in the general stream, in contrast to disaster-specific streams, might appear in different patterns. To clarify, people tend to use more accurate and full addresses of locations when reporting incidents happening during emergencies, aiming to help responders make immediate actions (e.g., Tweet #1 in Table 1). To investigate further, we train our LMR model using a combination of the *BTC.loc* dataset (as using it achieved the best F_1 score earlier in most of the 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 with BTC NER dataset 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.

We show the results of these runs in Fig. 3. Generally, the results are not consistent across disasters, hence we cannot draw a clear conclusion on which setup is clearly the best. As references, we compare the results with the case when we train on the target dataset (denoted as *Target* in Fig. 3) and with *BTC.loc* (as using it mostly achieved the best F_1 score among the *non-target* setups). It is evident that using training data other than the target data shows significant degradation in performance with respect to the *Target* model. This finding emphasizes the importance of providing in-domain (i.e., *Target*) data to achieve 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, that is 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 automatically augment labeled data to improve the performance over time. This can be achieved by exploiting active learning, 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 have a bigger impact on the model than training on combined data.

To answer **RQ3**, we conclude that using disaster-related training data helped improve the LMR model by 5.3% on average for all

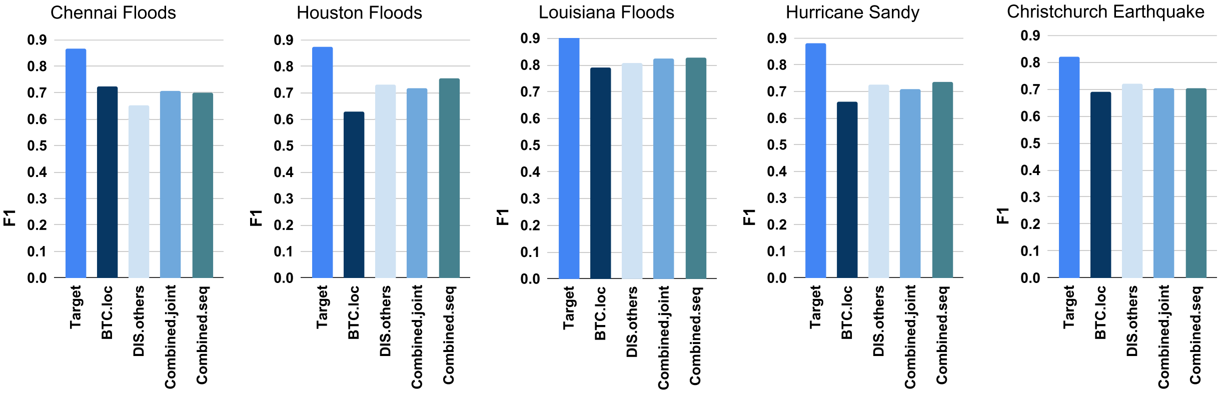

Fig. 3. The F_1 results of exploiting in- and out-of-domain data for training an LMR model.

datasets except Chennai Floods.

5.4. Cross-domain training (RQ4)

Although using disaster-related training data showed 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 of 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).

Fig. 4 shows the results. The missing bars in the case of Hurricane Sandy and Christchurch Earthquake are due to the fact that we only have one hurricane event and one earthquake event.

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 performance of training on FLD and HRC are slightly close for the Houston and Louisiana Floods. We suspect the reason to be the close geographical 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, Figs. 9–13 in Appendix.B). Specifically, 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 training on disasters of the same type generally achieves better performance. To further understand these results, next we explore how the geospatial proximity of source events to the target event affects the performance.

5.5. Geo proximity-based training (RQ5)

The location mentions from within the affected areas of a target disaster are expected to emerge in the tweets stream over time. However, such locations may not be seen by LMR models trained on past disaster events. We anticipate that 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, counties, etc.). 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, if the target disaster is Hurricane Sandy, we train on Houston and Louisiana Floods.
- *DIS_IN.FLD*: training on Chennai Floods happened in India.

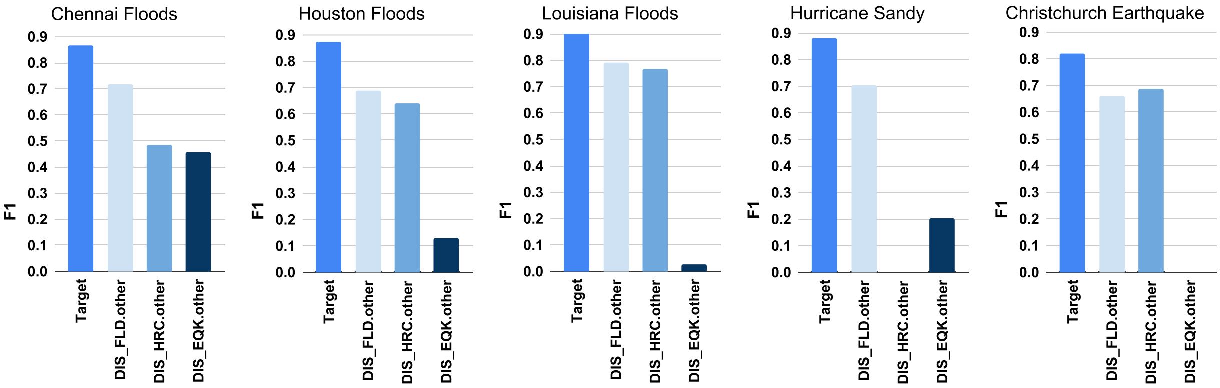

Fig. 4. The F_1 results of training on cross-domain data. Missing bars means there is no more than one disaster dataset of target type.

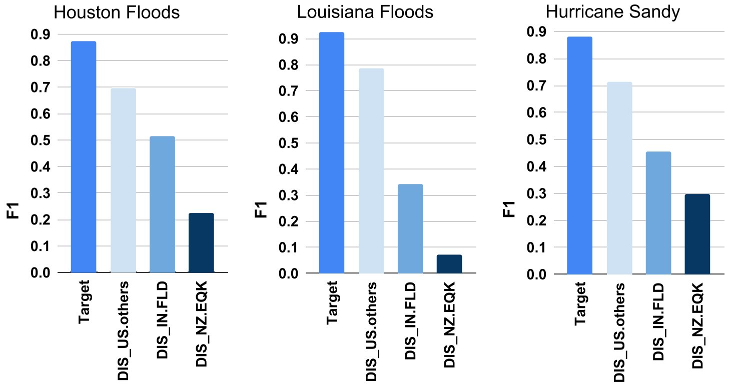

Fig. 5. The F_1 results of training on geo-proximity-based data.

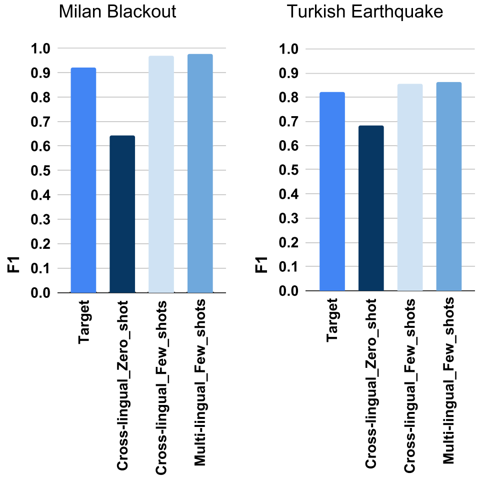

Fig. 6. The F_1 results of cross-lingual and multilingual training.

- *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 disasters that happened in the US; for other areas (NZ and IN), we do not have more than one disaster-specific dataset. Nonetheless, the results in Fig. 5 clearly indicate that training on source disasters that happened in close proximity areas (with respect to the target event) achieves the best performance regardless of the disaster event type. To answer **RQ5**, we suggest training on disasters that happened in close proximity areas to the location of the target event to achieve the best performance regardless of the type of the “source” and “target” disasters.

5.6. Cross-lingual training (RQ6)

Thus far, we 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) in which 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.

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.

Fig. 6 demonstrates that the performance of *Cross-lingual_Zero_shot* is acceptable but still far away from the improvement level achieved by the Target setup. However, adding little training data from the target (312 and 1400 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.

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.[25] We show the distribution of languages in Fig. 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.[26] 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 fair 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.

5.7. 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 model 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 have to be trained on target data, whenever available, to reach the highest possible performance. To simulate the process of acquiring target labeled data during disaster

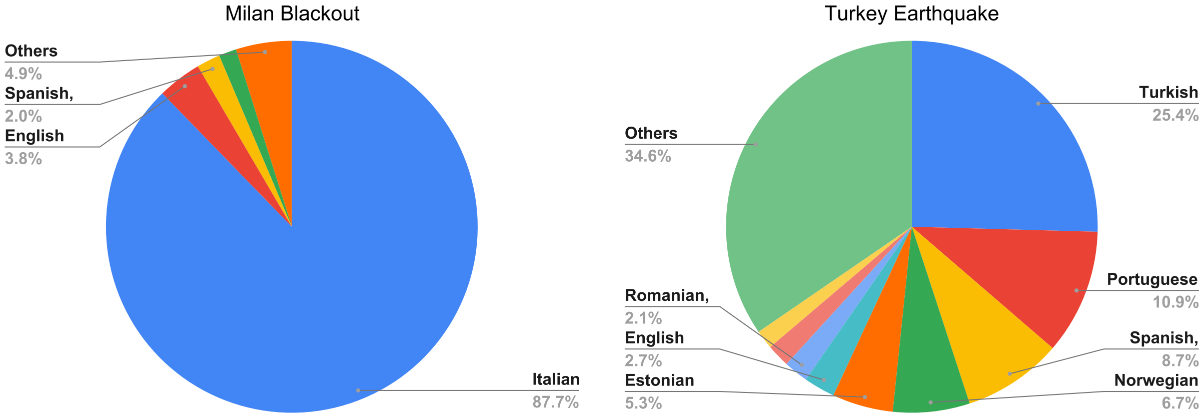

Fig. 7. The language distribution in Milan Blackout and Turkey Earthquake datasets.

[25] https://pypi.org/project/langdetect/.

[26] https://en.wikipedia.org/wiki/Languages_of_Turkey.

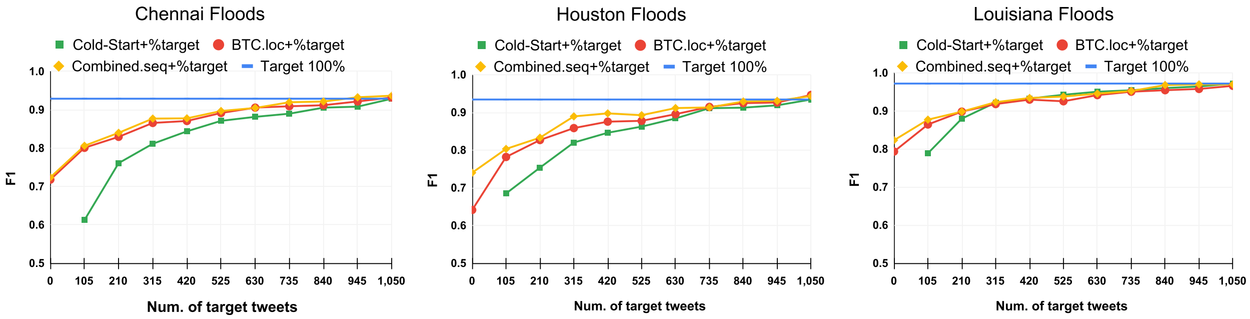

Fig. 8. The F_1 results of incrementally training on target data.

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 1050 tweets from every disaster. We first train our model with a base training dataset, then incrementally add 105 tweets in chronological order from the target event. This number constitutes 10% of the entire training data that can be labeled within our predefined budget. We use only the floods datasets as they contain the tweets' timestamps that we used 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 scenario when there is no data to pre-train the LMR model.
- *BTC.loc*+%: we use BTC dataset with only 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 5.2). This setup simulates the scenario when we have general-purpose and disaster-related tweets 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 the *Target* baseline, i.e., when training only on the entire 1050 training tweets.[27]

Results in Fig. 8 show that in all training setups, increasing the training data improves the performance continuously. 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 clearly better than the *Cold-Start* setup until we reach 40% in Chennai and Houston Floods and 20% in Louisiana Floods. Furthermore, *Combined.seq* is even more promising compared to *BTC.loc* as it contains cross- and in-domain data (i.e., disaster-specific tweets) that seem to slightly improve performance 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 until collecting 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, once we have labeled about 1000 tweets from the target event, there is almost no benefit of leveraging other external training data.

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 budget of annotation, we suggest labeling around 500 tweets to achieve reasonable performance (about 0.9 F_1 in the three disasters we experimented with), in addition to the cross- and out-or-domain training.

6. Error analysis

To better understand the different types of errors that 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.

[27] Note that the performance of the *Target* in these experiments is different than previous research questions due to sorting tweets chronologically.

6.1. 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 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 malformed LM. Table 3 shows the number of errors representing false positives, false negatives, partial matches, and malformed LMs (in the test set only). Table 4 shows example tweets along with their error types highlighted.

The false positives are mostly valid fine- and coarse-grained LMs that are 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 perhaps still needs more data to better recognize the LMs (tweets #3–6, 12–14). The partial matches are more common in Chennai and Houston Floods compared to 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, along, etc.) as part of the LM (e.g., tweet #7) which potentially confuses the model. Additionally, there are location types such as "mosque", "ATM", "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 recognize these LMs as part of precedence LMs, or as independent LMs.

Below we list our general observations based on the close examination of these errors.

a. Surprisingly, we observed that most of the false positives are indeed *valid* LMs. They are either fine- or coarse-grained locations that are 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" that is mentioned to describe the way grills are cooked 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 indeed emphasizes the need for an accurate preprocessing pipeline, including spell checking and hashtag segmentation, before the recognition phase.
c. There are location types that are not actual geographical points such as "area", "ATM", "mosque", etc. 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 not always annotated using the same sequence of BILOU tags. For example, the "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 the "HCL" is labeled as a single token LM (e.g., tweets #11 and #12). Additionally, the "5th street" in tweet #10 appears twice in the dataset, but it is labeled as a LM in the original tweet, but not in its retweet.
e. There are ambiguous locations that we could not resolve to existing locations such as "World", "Swayamsevaks", "Congress", etc. 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 another level of 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 Fig. 10).
g. Some ORG entities are correctly labeled by the LMR model as non-location, but they appear in the gold annotations as LMs. For example, "FEMA" in Tweet #15 refers to the "Federal Emergency Management Agency". Additionally, in Fig. 11, 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 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
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**

Table 4
Examples of errors of BERT-based 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.

Tweet #	**Error Type**	**Tweet text**
Tweet #1	False positive	`Here’s a look @ our downtown parking lot outside of HCSO headquarters downtown. #HouNews #TurnAroundDontDrown [url]`
Tweet #2	False positive	`Are the roads flooded in manassas??`
Tweet #3	False positive & False negative	`Check out Louisiana Grills at Rich & John’s "The Stove Shop" #summer #BBQ [url]`
Tweet #4	False negative: misspelling	`Another big quake, 6.3 has hit Christchurh during work hours. Not a pretty site. Buildings damaged, some collapsed.`
Tweet #5	False negative: concatenated	`– #apollohospitals– helplines –share – [user_mention] [user_mention] [user_mention] [user_mention] [user_mention] [url]`
Tweet #6	False negative: location type	`All the #mosque in chennai now open for food and stay. Thank you Allah !!`
Tweet #7	Partial match: adpositions	`.[user_mention] Navy rescue team deployed in Gandhi nagar area, beyond Adyar Bridge along Buckingham Canal`
Tweet #8	Partial match & Malformed	`Crescent College (B. S. Abdur Rahman University), #Vandalur is open for shelter. [...]`
Tweet #9	False positive & Malformed	`#SANDY ive never seen nyc look like this #HurricaneSandy the flooding is unreal....long beach`$_{L-LOC}$ `long island #unrecognizable`
Tweet #10	False positive: inconsistencies	`Looks like Malad :p [̈user_mention]]: Street flooding #NYC: 48th Ave between 5th and Center Blvd #Sandy [url]`
Tweet #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..`
Tweet #12	False negative: inconsistencies	`Need help to people who are in ELCOT branch HCL office, shollinganallur`
Tweet #13	False negative: ambiguous	`Swayamsevaks preparing & distributing food to around 1500 poor people in flood affected Lakshmipuram`
Tweet #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.`
Tweet #15	False Negative: ORG not LOC	`Louisiana Flooding, One Week Later: Author: J essica StapfUnprecedented. Historic. E ... disaster fema`

6.2. 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[28] and manual search (in case Google API does not return any result).

We show the number of LMs for each granularity level in Table 5. The "Other" type represents locations that were not resolved through Google Places API as well as through manual search. Overall, we notice that most of the errors are fine-grained and the majority of them originated from flood-related disasters. Moreover, the model seems to make more mistakes in detecting coarse-grained locations from Hurricane Sandy and the Christchurch Earthquake (tweets #9 and #4 in Table 4, respectively).

7. Discussion

Thus far, we discussed how geolocation information plays an important role in relief activities during emergencies. To tackle the LMR problem we studied the choice of training data that can influence the robustness of LMR models. In this section, we elaborate on the implications of our study from both the technical and the crisis management domain aspects (Section 7.1). We then discuss the limitations of our study (Section 7.2) and the limitations of using geographical information from Twitter (Section 7.3).

7.1. Implications of the study

This explorative study has a direct impact on the *efficiency* of crisis managers' response *at the lowest cost*, due to overcoming the challenge of collecting labeled data, which forms the bottleneck of running the downstream tasks that rely on geolocation information. Avoiding the response delay is the ultimate goal for responders during emergencies. Thus, deviating from the typical focus of LMR model development, we focused our exploration on the choice of existing training data. We proposed different factors that could influence the performance of the LMR models, such as *data domain*, *entity type*, *disaster domain*, *geo-proximity*, and *language*.

From the technical perspective, we explored two setups: (1) eliminating the labeled data acquisition phase (zero-shot setting), or (2) assuming we have a few, as hundreds, of training examples available (few-shot setting). For the zero-shot setup, we studied the effect of all the aforementioned factors. We studied the first four factors in monolingual setup and the last factor in cross- and multi-lingual setups.

For all factors, we observed that the performance of the LMR model improves as we focus the training on the factor of interest. For example, for the *data domain* factor, the performance improves as we use more domain-focused training data. While previous studies only suggest using Twitter data, our findings suggest prioritizing the datasets based on their domain and then their availability. Additionally, training on *LOC-only NER datasets* can improve the performance by approximately 12% in F_1 score. This indeed enables the rapid deployment of initial LMR models at the onset of disaster events by using the data at hand. While a relatively good model is deployed, more training data from the target disaster event can be labeled to improve the LMR model further. On the responders' side, the quick deployment enables rapid response, but with careful consideration to the degree of uncertainty of the model decisions.

For the *disaster domain* and *geo-proximity* factors, unsurprisingly, we found that focusing the training on the same type of disaster and geographically-nearby events as the target disaster event boosts the performance compared to employing events of different types or those happened in far-away areas. These two aspects had never been explored in existing studies, so our finding paves the way for future successful deployments by suggesting to consider the *disaster domain* and *geo-proximity* factors when training LMR models in the crisis domain.

The cross- and multi-lingual setups for LMR are understudied in literature [61], thus we employed the multi-lingual BERT model to investigate these setups. Our conclusion for the *language* factor suggests that even with the absence of training data from the target language, we can still build descent LMR model using data from other languages. For responders, regardless of the language used by people living in the impacted area by a disaster event, deploying initial acceptable performing models is possible which in turn supports their rapid response.

Furthermore, whenever feasible, we recommend labeling around 500 tweets and combining them with all other available data to reach a minimum F_1 performance of 85%. Although we noticed the limited annotation budgets by the small size of the available datasets [32,48,51,53,55,68] (ranging from 1000 to 6648 tweets), investigating the minimal annotations cost was not explored at all in previous studies. This conclusion gives some clues for the response authorities on the budget they may spend to improve initially-deployed models.

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

Table 5
Location types of miss predicted LMs.

	Chennai FLD			Houston FLD			Louisiana FLD			HRC Sandy			ChCh EQK			Total
	FP	FN	PM	FP	FN	PM	FP	FN	PM	FP	FN	PM	FP	FN	PM	
Fine-grained	0	11	23	19	11	7	3	13	2	6	6	6	2	6	6	121
Coarse-grained	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**

7.2. *Limitations of the study*

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, there are two factors related to the construction of the dataset that could affect our results and conclusions. First, the datasets that we adopt are scarce with limited coverage of disaster types and geographical areas. Over and above that, the datasets are randomly drawn from collected tweet datasets using hashtags. Relying on hashtags is a major limitation for capturing the relevant posts discussing the target disaster event, as there are some users not using hashtags while posting about the event [18,69]. Additionally, hashtag-based datasets might have different characteristics compared to other datasets collected using different ways such as geographical-based datasets, which could affect the drawn conclusions [27].
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 same 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 Fig. 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. Considering contextual translation of the data might lead to different conclusions.
e. We discussed some issues in the annotations of the used datasets in Section 6, that could affect the performance of our BERT-based 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 listed issues above motivate the need for a larger, domain and geographically diverse, and consistently annotated LMR datasets. We keep the exploration of this issue and the developing of standard annotation guidelines for future work.

7.3. *Limitations of social media data*

In addition to the limitations of our study, there are some limitations related to the information extracted from social media, such as incomplete information, false information, etc. Here, we discuss these limitations in the context of extracting geolocation information for the crisis management applications [70–73]. We discuss a few in this section:

- The geolocation information extracted from social media does not generally provide the full picture of the crisis [70], but alternatively complements other sources of data. A serious challenge caused this is the "digital divide" due to several reasons, such as losing internet connection or suffering from the limited connectivity (e.g., cable disruption or displacement) [71], having restricted or no access at all to the internet (e.g., low-income communities[29]), not using social media at all (e.g., elderly), or avoiding social media due to safety and privacy concerns (e.g., conflicts and violence events).

[29] https://data.worldbank.org/indicator/IT.NET.USER.ZS?locations=XO

- The datasets are randomly drawn from a bounded period while tracking the event over Twitter (i.e., using hashtags). The tracking time might not cover the entire pre-disaster and post-disaster periods in which the underlying causes and consequences of he disaster events can be studied [71].
- The quality of the user-generated text is not guaranteed. Hence, the response authorities have to pre-qualify the posts before inferring and extracting the geolocation information during emergencies. Our perception for the required pre-qualification (filtering) modules is depicted in Fig. 1 as upstream tasks. We elaborate on the filtering modules in the following:
  - Despite the benefits of using hashtags, there is no guarantee on the integrity of tweets posted on trending hashtags during emergencies. Hashtag riding is a common issue as some accounts use them to promote themselves (self-promotion), their ideas and believes (propaganda), their products or organizations (commercial), etc. For example, some new agencies might share dramatic media and titles about the event to increase their followers and readership [72], let alone the spammer and bot accounts that distract the public.
  - The information obtained from social media needs accurate validations for credibility and reliability. Since this issue is more severe during the crises, response authorities should account for this.
  - The response authorities would be interested in obtaining the geolocation information for solely the *informative* posts (e.g., situational updates or actionable posts) that help them in performing their activities. That requires the authorities to (automatically) filter the posts by informativeness.

8. Conclusion

This work contributes towards a crucial task, i.e., *Location Mention Recognition* in the crisis management domain. We formulated several research questions for which evidence-based answers were unknown. We designed an extensive and reliable experimental setup where several experiments investigate the effectiveness of training on general-purpose NER datasets from news articles and tweets. We demonstrate how the performance of a 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 the general-purpose Twitter NER data is preferred over the general-purpose web NER data; and crisis-related Twitter data is preferred over the general-purpose Twitter data. Furthermore, our results suggest that training on disaster events data from similar type or geographically-nearby events to the target event boosts the performance compared to training on different event types or distant events. We further show how training on previous disasters of a different language 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 combine them with all available data to obtain an LMR model that hits a performance level greater than 85% F_1 score. Overall, we remark that our findings shape the future directions in this line of research.

In addition to addressing the discussed limitations of our study, we plan, in the future, to work on the technical aspects of this work. This involves improving the learning model, using more advanced computational transfer learning methods to mitigate the negative effects of all our defined factors, modifying the classification layer built over BERT, employing other learning models, and studying different expansion ways to improve the tweets representation for the recognition task.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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

Appendix A. Full Results

Detailed results, including out-domain training, in/out-domain training, cross-domain training, and training based on geo-proximity of events (Table 6).

Table 6

Full results of different domain setups. Best F1 scores of non-target training setups are boldfaced. E refers to the number of training epochs, BS refers to the training batch size, and LR refers to the learning rate (Adam).

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

Appendix B. Location distribution

The location distribution of the English disaster-specific tweet datasets are depicted in Figs. 9–13).

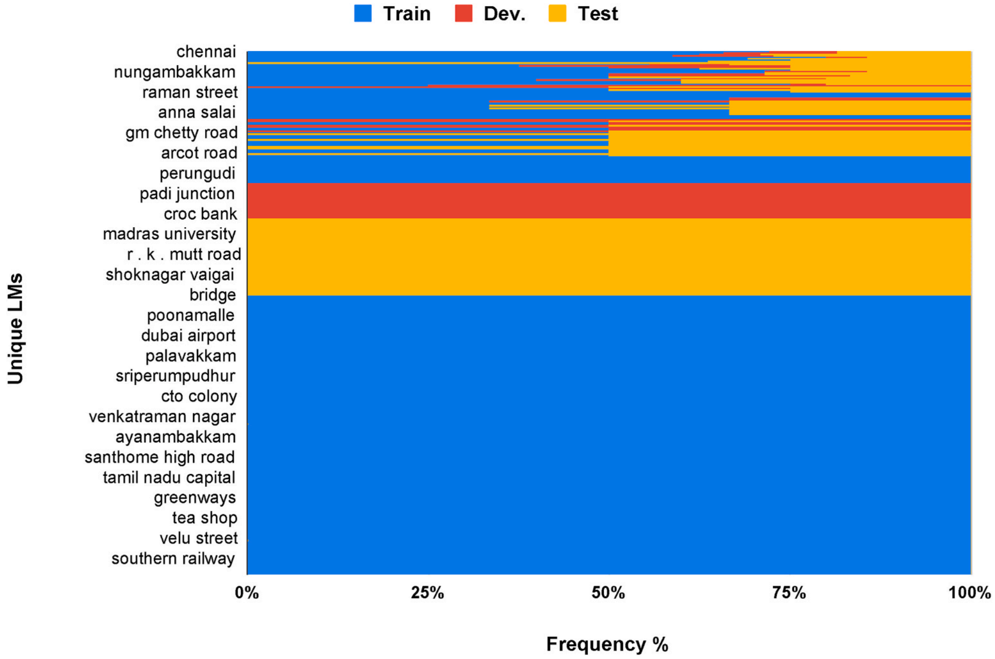

Fig. 9. The LMs distribution across training, development, and test data for Chennai Floods disaster dataset.

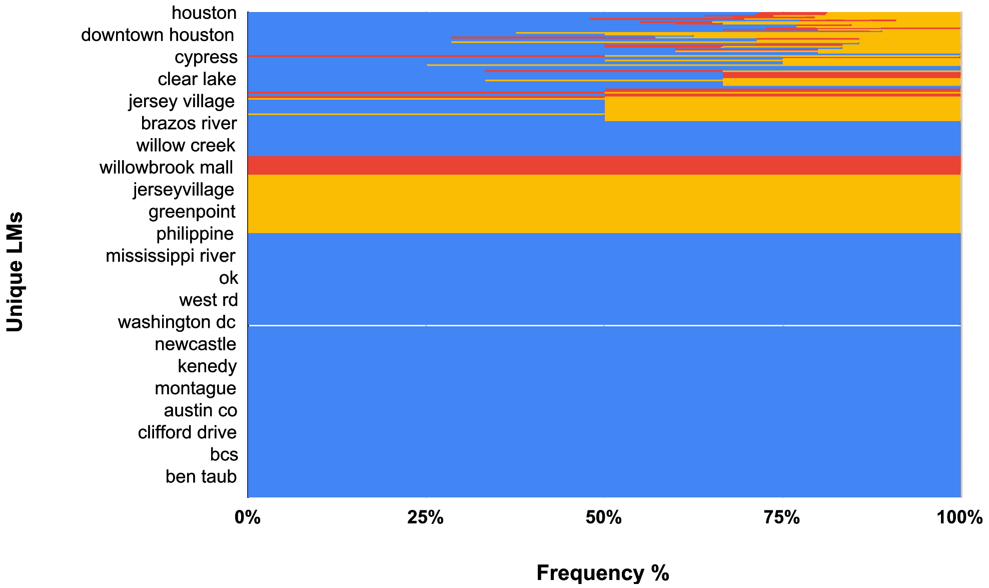

Fig. 10. The LMs distribution across training, development, and test data for Houston Floods disaster dataset.

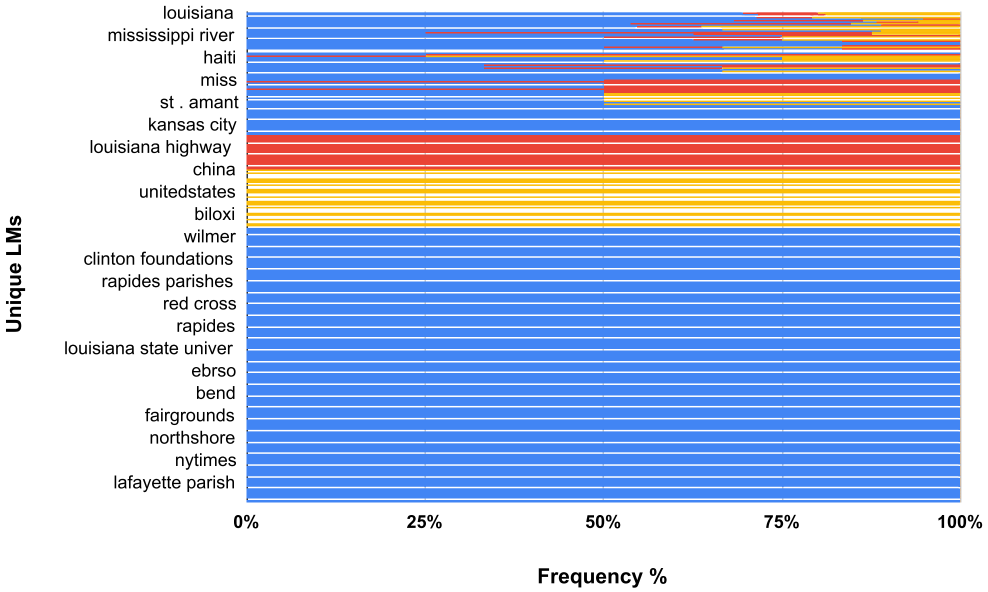

Fig. 11. The LMs distribution across training, development, and test data for Louisiana Floods disaster dataset.

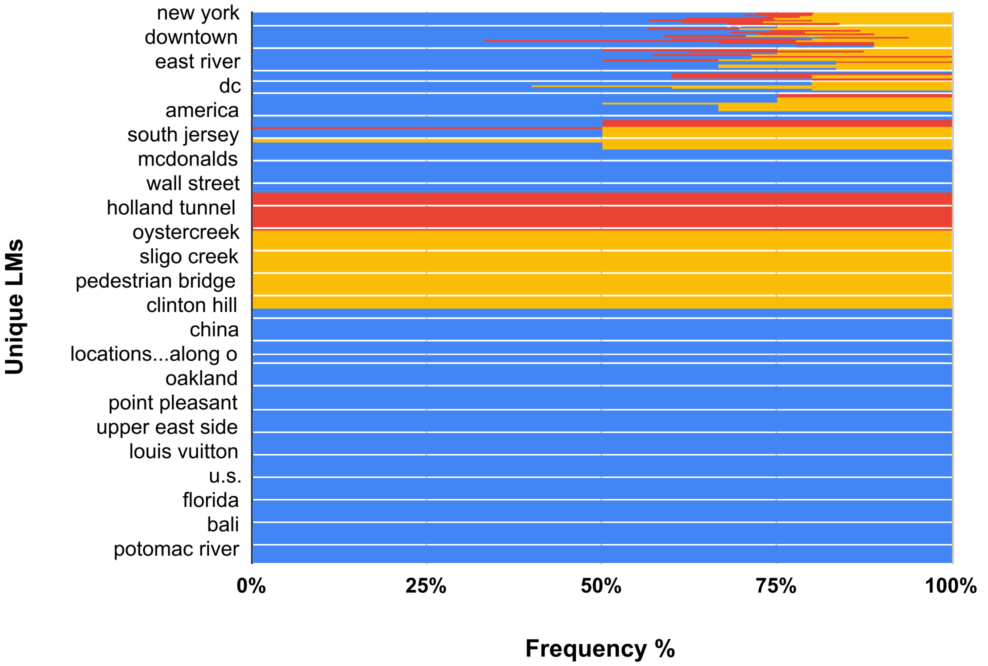

Fig. 12. The LMs distribution across training, development, and test data for Hurricane Sandy disaster dataset.

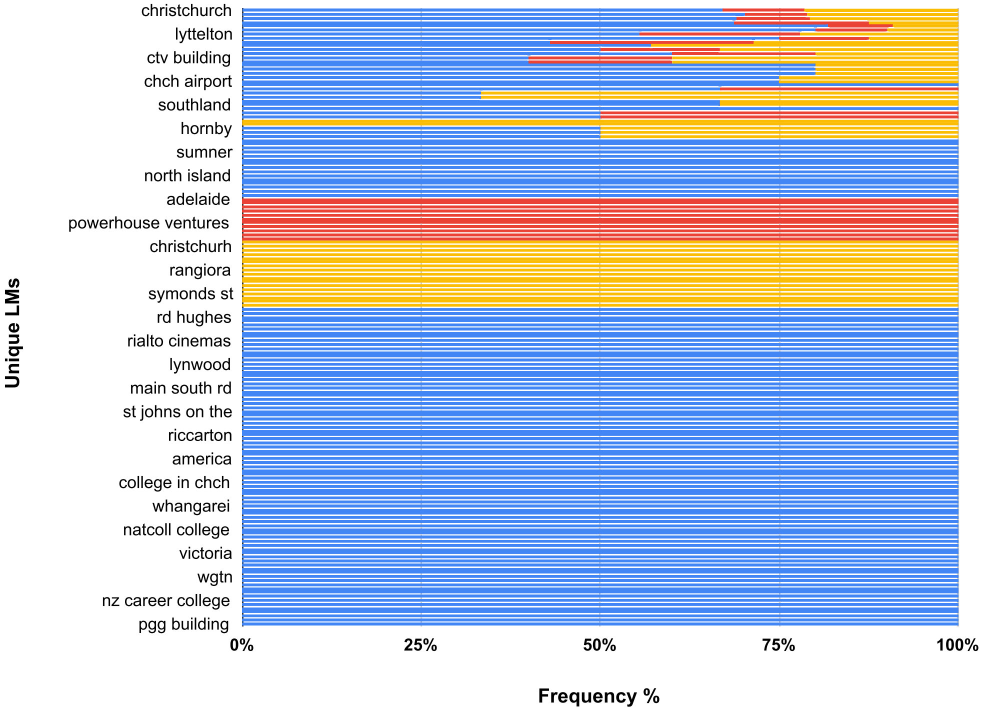

Fig. 13. The LMs distribution across training, development, and test data for Christchurch Earthquake disaster dataset.

References

[1] R. Suwaileh, M. Imran, T. Elsayed, 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, 2020, pp. 6252–6263.

[2] S.E. Vieweg, Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications, Ph.D. thesis, University of Colorado at Boulder, 2012.

[3] H. Purohit, C. Castillo, M. Imran, 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.

[4] Y. Hu, J. Wang, How do people describe locations during a natural disaster: an analysis of tweets from hurricane harvey, in: Leibniz International Proceedings in Informatics, LIPIcs, 2020, p. 177, doi:10.4230/LIPIcs.GIScience.2021.I.6.

[5] I. Weber, M. Imran, F. Ofli, F. Mrad, J. Colville, M. Fathallah, A. Chaker, W.S. Ahmed, Non-traditional data sources: providing insights into sustainable development, Commun. ACM 64 (2021) 88–95.

[6] R. Grace, Toponym usage in social media in emergencies, Int. J. Disaster Risk Reduc. 52 (2021), 101923, https://doi.org/10.1016/j.ijdrr.2020.101923.

[7] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, K. Starbird, From situational awareness to actionability: towards improving the utility of social media data for crisis response, Proc. ACM Hum.-Comput. Interact. 2 (2018), https://doi.org/10.1145/3274464.

[8] J. Kropczynski, R. Grace, J. Coche, S. Halse, E. Obeysekare, A. Montarnal, F. Benaben, A. Tapia, 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, 2018, pp. 428–438. Wellington, New Zealand, https://hal-mines-albi.archives-ouvertes.fr/hal-01987793.

[9] C. Reuter, T. Ludwig, M.-A. Kaufhold, T. Spielhofer, Emergency services' attitudes towards social media: a quantitative and qualitative survey across europe, Int. J. Hum. Comput. Stud. 95 (2016) 96–111, https://doi.org/10.1016/j.ijhcs.2016.03.005. https://www.sciencedirect.com/science/article/pii/S1071581916000379.

[10] S. McCormick, New tools for emergency managers: an assessment of obstacles to use and implementation, Disasters 40 (2016) 207–225, https://doi.org/10.1111/disa.12141.

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

[12] B. Huang, K.M. Carley, A large-scale empirical study of geotagging behavior on Twitter, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 365–373.

[13] M. Imran, C. Castillo, F. Diaz, S. Vieweg, Processing Social Media Messages in Mass Emergency: A Survey, ACM Comput. Surv., 2015, p. 47, https://doi.org/10.1145/2771588.

[14] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, H. Perez-Meana, J. Portillo-Portillo, V. Sanchez, L.J. García Villalba, Using twitter data to monitor natural disaster social dynamics: a recurrent neural network approach with word embeddings and kernel density estimation, Sensors 19 (2019), https://doi.org/10.3390/s19071746. https://www.mdpi.com/1424-8220/19/7/1746.

[15] M. Sreenivasulu, M. Sridevi, Comparative study of statistical features to detect the target event during disaster, Big Data Mining and Analytics 3 (2020) 121–130, https://doi.org/10.26599/BDMA.2019.9020021.

[16] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, 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.

[17] R. Mazloom, H. Li, D. Caragea, C. Caragea, M. Imran, A hybrid domain adaptation approach for identifying crisis-relevant tweets, Int. J. Inf. Syst. Crisis Response Manag. 11 (2019) 1–19.

[18] L.S. Snyder, Y.-S. Lin, M. Karimzadeh, D. Goldwasser, D.S. Ebert, Interactive learning for identifying relevant tweets to support real-time situational awareness, IEEE Trans. Visual. Comput. Graph. 26 (2019) 558–568.

[19] X. Ning, L. Yao, B. Benatallah, Y. Zhang, Q.Z. Sheng, S.S. Kanhere, Source-aware crisis-relevant tweet identification and key information summarization, ACM Trans. Internet Technol. 19 (2019) 1–20.

[20] M. Marbouti, 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.

[21] M. Basu, K. Ghosh, S. Ghosh, Information retrieval from microblogs during disasters: in the light of IRMiDis task, SN Computer Science 1 (2020) 61, https://doi.org/10.1007/s42979-020-0065-1.

[22] X. Zheng, J. Han, A. Sun, A survey of location prediction on Twitter, IEEE Trans. Knowl. Data Eng. 30 (2018) 1652–1671.

[23] B. Han, P. Cook, T. Baldwin, Lexical normalization for social media text, ACM Transactions on Intelligent Systems and Technology 4 (2013) 1–27.

[24] J. Devlin, M.-W. Chang, K. Lee, 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, 2019, pp. 4171–4186.

[25] S.R. Hiltz, A.L. Hughes, M. Imran, L. Plotnick, R. Power, M. Turoff, Exploring the usefulness and feasibility of software requirements for social media use in emergency management, Int. J. Disaster Risk Reduc. 42 (2020), 101367, https://doi.org/10.1016/j.ijdrr.2019.101367.

[26] B. Poblete, J. Guzmán, J. Maldonado, F. Tobar, Robust detection of extreme events using twitter: worldwide earthquake monitoring, IEEE Trans. Multimed. 20 (2018) 2551–2561, https://doi.org/10.1109/TMM.2018.2855107.

[27] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg, Crisislex: a lexicon for collecting and filtering microblogged communications in crises, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[28] S. Vieweg, A.L. Hughes, K. Starbird, 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.

[29] A.M. MacEachren, A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, J. Blanford, Senseplace2: geotwitter analytics support for situational awareness, in: 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2011, pp. 181–190.

[30] S. Vieweg, C. Castillo, 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.

[31] R. McCreadie, C. Buntain, 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.

[32] H. Al-Olimat, K. Thirunarayan, V. Shalin, 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, 2018, pp. 1986–1997.

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

[34] R.D. Das, R.S. Purves, Exploring the potential of Twitter to understand traffic events and their locations in Greater Mumbai, India, IEEE Trans. Intell. Transport. Syst. 21 (2020) 5213–5222.

[35] J. Wang, Y. Hu, K. Joseph, NeuroTPR: a neuro-net toponym recognition model for extracting locations from social media messages, Trans. GIS 24 (2020) 719–735.

[36] C. Reuter, Crisis 2.0: towards a systematization of social software use (ijiscram), in: Emergent Collaboration Infrastructures: Technology Design for Inter-organizational Crisis Management, Springer Fachmedien Wiesbaden, Wiesbaden, 2015, pp. 35–48, https://doi.org/10.1007/978-3-658-08586-5_4.

[37] C. Reuter, A.L. Hughes, M.-A. Kaufhold, Social media in crisis management: an evaluation and analysis of crisis informatics research, Int. J. Hum. Comput. Interact. 34 (2018) 280–294, https://doi.org/10.1080/10447318.2018.1427832.

[38] A.L. Hughes, 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 GROUP '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 399–408, https://doi.org/10.1145/2957276.2957307.

[39] H. Purohit, C. Castillo, M. Imran, 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.

[40] K.C. Roy, S. Hasan, P. Mozumder, A multilabel classification approach to identify hurricane-induced infrastructure disruptions using social media data, Comput. Aided Civ. Infrastruct. Eng. 35 (2020) 1387–1402.

[41] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, Inc.", 2009.

[42] L. Hong, V. Frias-Martinez, Modeling and predicting evacuation flows during hurricane irma, EPJ Data Science 9 (2020) 29.

[43] K.C. Roy, S. Hasan, Modeling the dynamics of hurricane evacuation decisions from twitter data: an input output hidden markov modeling approach, e102976, Transport. Res. C Emerg. Technol. 123 (2021), e102976, https://doi.org/10.1016/j.trc.2021.102976. doi:10.1016/j.trc.2021.102976.

[44] O. Uchida, M. Kosugi, G. Endo, T. Funayama, K. Utsu, S. Tajima, M. Tomita, Y. Kajita, Y. Yamamoto, A real-time information sharing system to support self-, mutual-, and public-help in the aftermath of a disaster utilizing twitter, IEICE Trans. Fund. Electron. Commun. Comput. Sci. E99.A (2016) 1551–1554, https://doi.org/10.1587/transfun.E99.A.1551.

[45] M. Kosugi, K. Utsu, S. Tajima, M. Tornita, Y. Kajita, Y. Yamamoto, O. Uchida, 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, https://doi.org/10.1109/ICT-DM.2017.8275693.

[46] M. Kosugi, K. Utsu, M. Tomita, S. Tajima, Y. Kajita, Y. Yamamoto, O. Uchida, A twitter-based disaster information sharing system, in: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, pp. 395–399, https://doi.org/10.1109/CCOMS.2019.8821719.

[47] C. Zhang, C. Fan, W. Yao, X. Hu, A. Mostafavi, Social media for intelligent public information and warning in disasters: an interdisciplinary review, Int. J. Inf. Manag. 49 (2019) 190–207, https://doi.org/10.1016/j.ijinfomgt.2019.04.004. https://www.sciencedirect.com/science/article/pii/S0268401218310995.

[48] S.E. Middleton, L. Middleton, S. Modafferi, Real-time crisis mapping of natural disasters using social media, IEEE Intell. Syst. 29 (2014) 9–17.

[49] M. Avvenuti, S. Cresci, F. Del Vigna, M. Tesconi, Mapping to prioritize, Computer 49 (2016) 28–37.

[50] S. Malmasi, M. Dras, Location mention detection in tweets and microblogs, in: Conference of the Pacific Association for Computational Linguistics, 2016, pp. 123–134.

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

[52] S.E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, Y. Kompatsiaris, Location extraction from social media: geoparsing, location disambiguation, and geotagging, ACM Trans. Inf. Syst. 36 (2018) 1–27.

[53] J. Gelernter, S. Balaji, An algorithm for local geoparsing of microtext, GeoInformatica 17 (2013) 635–667.

[54] F. Abdelkoui, M.-K. Kholladi, Extracting criminal-related events from Arabic tweets: a spatio-temporal approach, J. Inf. Technol. Res. 10 (2017) 34–47.

[55] D. Molla, 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.

[56] L. Ghahremanlou, W. Sherchan, J.A. Thom, Geotagging twitter messages in crisis management, Comput. J. 58 (2015) 1937–1954.

[57] J. Yin, S. Karimi, J. Lingad, Pinpointing locational focus in microblogs, in: Proceedings of the 2014 Australasian Document Computing Symposium, 2014, pp. 66–72.

[58] L. Nizzoli, M. Avvenuti, M. Tesconi, S. Cresci, Geo-semantic-parsing: AI-powered geoparsing by traversing semantic knowledge graphs, Decis. Support Syst. 136 (2020), 113346.

[59] J. Wang, 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.

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

[61] J. Gelernter, W. Zhang, Cross-lingual geo-parsing for non-structured data, in: Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR 2013, 2013, pp. 64–71, https://doi.org/10.1145/2533888.2533943.

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

[63] H.-J. Dai, P.-T. Lai, Y.-C. Chang, R.T.-H. Tsai, Enhancing of chemical compound and drug name recognition using representative rag scheme and fine-grained tokenization, J. Cheminf. 7 (2015) S14.

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

[65] E.F. Tjong Kim Sang, 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.

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

[67] E.F. Tjong Kim Sang, 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.

[68] J.O. Wallgrün, M. Karimzadeh, A.M. MacEachren, S. Pezanowski, Geocorpora: building a corpus to test and train microblog geoparsers, Int. J. Geogr. Inf. Sci. 32 (2018) 1–29.

[69] C. Reuter, T. Ludwig, C. Kotthaus, M.-A. Kaufhold, E. von Radziewski, V. Pipek, Big data in a crisis? creating social media datasets for crisis management research, com 15 (2016) 249–264.

[70] T. Shelton, A. Poorthuis, M. Graham, M. Zook, Mapping the data shadows of hurricane sandy: uncovering the sociospatial dimensions of 'big data, Geoforum 52 (2014) 167–179.

[71] K. Crawford, M. Finn, The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters, Geojournal 80 (2015) 491–502, https://doi.org/10.1007/s10708-014-9597-z.

[72] V. Farida, Twitter as a reporting tool for breaking news, Digital Journalism 1 (2013) 27–47, https://doi.org/10.1080/21670811.2012.741316.

[73] T. Gillespie, The Politics of "platforms", 12, New Media & Society, 2010, pp. 347–364, https://doi.org/10.1177/1461444809342738.