IDRISI-RA: The First Arabic Location Mention Recognition Dataset of Disaster Tweets

Reem Suwaileh
Computer Science and Engineering Department
Qatar University
Doha, Qatar
rs081123@qu.edu.qa

Muhammad Imran
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
mimran@hbku.edu.qa

Tamer Elsayed
Computer Science and Engineering Department
Qatar University
Doha, Qatar
telsayed@qu.edu.qa

Abstract

Extracting geolocation information from social media data enables effective disaster management, as it helps response authorities; for example, in locating incidents for planning rescue activities, and affected people for evacuation. Nevertheless, geolocation extraction is greatly understudied for the *low resource languages* such as *Arabic*. To fill this gap, we introduce IDRISI-RA, the first publicly-available Arabic Location Mention Recognition (LMR) dataset that provides human- and automatically-labeled versions in order of thousands and millions of tweets, respectively. It contains both location mentions and their types (e.g., district, city). Our extensive analysis shows the decent geographical, domain, location granularity, temporal, and dialectical coverage of IDRISI-RA. Furthermore, we establish baselines using the standard Arabic NER models and build two simple, yet effective, LMR models. Our rigorous experiments confirm the need for developing specific models for Arabic LMR in the disaster domain. Moreover, experiments show the promising domain and geographical generalizability of IDRISI-RA under zero-shot learning.

1 Introduction

Worldwide, and in the Arab world, Twitter has played a critical operational role in crisis management. The Beirut explosion in 2020 is an excellent case in point, where on-site individuals started intuitively responding to each other using geotagged tweets. What makes tweets invaluable is the presence of location mentions at different granularity (Grace et al., 2018; McCormick, 2016; Reuter et al., 2016; Kropczynski et al., 2018). Response authorities exploit this geographical information to effectively manage emergencies using *Crisis Maps*. Although the geographical dimension adds situational and operational values to Twitter data, on 18 June 2019 Twitter discontinued the geotagging feature in tweets.[1] This necessitates the need to develop automatic geolocation tools. Nevertheless, the main obstacle for the Arabic language, which is a low-resource language, the LMR task is severely understudied due to several factors, including the absence of a unified evaluation framework constituting annotated datasets, a representative set of baselines, and fair evaluation metrics.

To address these barriers, we focus on the Location Mention Recognition (LMR) task and introduce IDRISI-RA,[2] the first human-labeled dataset comprising Arabic tweets from 7 disaster events (*gold* annotations). We also introduce the first large-scale automatically-labeled tweets (*silver* annotations) from 22 disaster events that cover all Arab world. IDRISI-RA covers the most occurring disaster types that happened in the Arab world. More importantly, it is labeled for two annotation types that are location mentions (i.e., toponym textual spans) and their types (e.g., city, street, POIs, etc.). Hence, it supports *type-less* LMR, where the model detects toponyms at any granularity, and *type-based* LMR, where the model distinguishes LMs types while detecting them (Suwaileh et al., 2023).

Although adapting Named Entity Recognition (NER) models and English datasets goes a long way towards tackling the LMR task, researchers have empirically shown that specialized LMR models yield better performance and are more effective for emergency management tasks (Suwaileh et al., 2022). For the Arabic language, only a few datasets exist yet suffer from their limited geographical, domain, and dialectal coverage. What exacerbates the low resources issue for the Arabic language is the unavailability of LMR datasets, except a few task-specific ones, including traffic surveillance

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

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

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16298–16317

and event detection (Al Emadi et al., 2017; Alkouz and Al Aghbari, 2018; Alkouz and Al Aghbari, 2020; Bahnasy et al., 2020). Therefore, in an effort to expedite the development of Arabic LMR models and shape future directions, we perform extensive experiments to empirically answer the following research questions:

RQ1: Are standard Arabic NER models sufficient for effective LMR over disaster tweets?

RQ2: Can LMR models trained on IDRISI-RA generate generalizable LMR models that reasonably perform on unseen disaster events?

RQ3: Can LMR models trained on IDRISI-RA generate domain generalizable LMR models that reasonably perform on unseen disaster events of the same or different types?

RQ4: Can LMR models trained on IDRISI-RA generate geographically generalizable LMR models that reasonably perform on unseen disaster events that happened in different countries?

Our rigorous analyses and experiments necessitate the development of specialized LMR models for the disaster domain. Additionally, the experiments demonstrate promising domain and geographic generalizability of IDRISI-RA under *zero-shot learning*.

The contributions of this paper are as follows:

- We present IDRISI-RA,[3] the first public human-labeled Arabic LMR dataset (gold version) of about 4.6k tweets. The dataset covers diverse disaster types and countries.
- We release the largest automatically-labeled Arabic LMR dataset (silver version), constituting about 1.2M tweets.
- We annotate the location mentions into coarse- and fine-grained location types to enable hierarchical LM recognition, disambiguation, and evaluation.
- We benchmark IDRISI-RA using the standard Arabic NER models and our own simple yet competitive LMR models to establish a set of baselines for the community.
- We empirically demonstrate that IDRISI-RA is a reasonably generalizable dataset.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents an overview of the LMR problem. Section 4 describes the design objective, dataset creation, and annotation. Section 5 analyzes the reliability, coverage, and diversity of IDRISI-RA. Sections 6-7 present benchmarks and generalizability study on IDRISI-RA. Section 8 presents the silver version of IDRISI-RA. Sections 9-10 discuss a few use cases for utilizing IDRISI-RA and the ethical considerations. We finally conclude and list a few future directions in Section 11.

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

2 Related Work

In this section, we review the Arabic Twitter NER and LMR datasets. We present their characteristics and issues and discuss how IDRISI-RA dataset overcomes these limitations. In Table 1, we summarize the existing NER and LMR datasets.

2.1 Twitter NER Datasets

An intuitive solution to tackle the LMR task is to use general-purpose NER models for LM detection or train their LMR models using Twitter NER datasets. For example, Chen et al. (2019) built a multilingual LMR system that employs a translation module at its core to process Arabic documents. The study focuses on detecting the location mentions in newswire, broadcast news and conversations, and conversational telephone speeches.

Although the Arabic NER datasets could be sufficient for training acceptable LMR models at the onset of disaster events, several challenges are associated with this line of research. Despite the fact that the NER models trained on web data perform poorly on tweets (Darwish and Gao, 2014), the Arabic NER studies have a limited focus on Twitter. While this requires creating domain-specific datasets, existing public Arabic Twitter NER datasets suffer from the limited size, domain, and geographic coverage (Darwish, 2013; Aguilar et al., 2018; Jarrar et al., 2022). Darwish and Gao (2014) introduced the first Arabic Twitter NER dataset that contains 5,069 tweets and 1,300 LMs, 299 of which are unique. Aguilar et al. (2018) created a dataset that constitutes 12,334 tweets and 5,306 LMs. However, only the training and development sets are public to the research community. As of Jan 2023, we managed to crawl 11,155 tweets containing 1,412 LMs, 440 of which are unique. Recently, Jarrar et al. (2022) created the first Arabic Multi-domain nested NER dataset. It contains a subset of 5,653 tweets containing 435 entities of types LOC and GPE, 175 of which are unique.

A major challenge in processing Arabic documents is handling the dialectical (colloquial) text

Dataset	Task	# Tweets	# LM (uniq)	Annotation	Public
Darwish (2013)	NER	5,069	1,300 (299)	In-house	Yes
Aguilar et al. (2018)	NER	11,155	1,412 (440)	In-house	Yes
Jarrar et al. (2022)	NER	5,653	435 (175)	In-house	Yes
Al Emadi et al. (2017)	LMR	-	-	In-house	No
Alkouz and Al Aghbari (2018)	LMR	100	-	In-house	No
Alkouz and Al Aghbari (2020)	LMR	-	-	In-house	No
Bahnasy et al. (2020) *	LMR	297,150	-	Automatic	No
IDRISI-RA_gold *	LMR	4,593	5,236 (918)	In-house	Yes
IDRISI-RA_silver *	LMR	1,205,373	884,217 (18,609)	Automatic	Yes

Table 1: Summary of the existing NER and LMR datasets. * indicates the disaster-related datasets, entirely or partially. – indicates the information that we could not obtain to the best of our effort. The dialectical distribution is unknown for all datasets, except for Aguilar et al. (2018).

commonly used over Twitter. Although the MSA-EGY (Aguilar et al., 2018) dataset is of reasonable size, it is limited to only MSA and Egyptian dialects, suffering from its limited geographical coverage. Other datasets do not report their dialectical distributions.

Furthermore, the Arabic NER datasets are randomly filtered using the sampling Twitter API; they are not disaster-specific datasets. These datasets could serve at the onset of disaster events for deploying acceptable LMR models but should be augmented with disaster-specific Twitter data for developing robust LMR models (Suwaileh et al., 2022).

IDRISI-RA addresses these limitations by being an event-centric dataset that geographically covers all Arab countries and reasonably represents their dialects.

2.2 Twitter LMR Twitter

Alkouz and Al Aghbari (2018) adopted an English LMR system (Malmasi and Dras, 2015) to extract LMs from English and Arabic traffic-related tweets (filtered using traffic keywords such as "traffic" and "jam"). The system issues the n-grams extracted from the tweet text against Google Place API and assigns the latitude and longitude coordinates to n-grams that obtained results from the API. The resultant data, however, is geographically limited to United Arab Emirates (UAE). There are a couple of other cross-lingual traffic monitoring systems for English and Arabic languages (Al Emadi et al., 2017; Alkouz and Al Aghbari, 2020). However, these datasets are not public and limited in size as they contain around 500-600 tweets. Bahnasy et al. (2020) employed LM extraction to aid event detection over Arabic tweets. Although the dataset is large in size, its geographical coverage is limited to Egypt, dialectical coverage is limited to Egyptian dialect, and its disaster domain is limited to fire, flood, and pandemic disaster types only. In contrast, IDRISI-RA dataset contains other disaster events that represent the most happening disaster types in Arabic-speaking countries. These events happened in 22 different countries. IDRISI-RA also captures a good coverage of Arabic dialects.

3 Problem Overview

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

There are two types of LMR tasks. The first type recognizes toponyms (e.g., Paris, New York) without their types (e.g., city, state) and is denoted as "type-less LMR". The second type recognizes toponyms and also distinguishes between location types (e.g., country, city, and street) and is denoted as "type-based LMR" (Suwaileh et al., 2023). The latter better serves the development and evaluation

of geolocation processing systems in light of the responders' needs. It enables a variety of downstream tasks (e.g., crisis maps) at different location granularity, in addition to being crucial for accurately disambiguating the toponyms.

The LMR problem is formally defined as follows: given a tweet t, 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 i^{th} 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 (`bilou_tag`) with 5 classes: "B" denotes the beginning token of an LM, "I" denotes a token inside an LM, "L" denotes the last token in an LM, "O" denotes a token outside of an LM, and "U" indicates that the LM has only one token such as "Doha". Therefore, we define the LMR as a *multi-class classification task on the token level*.

A *type-less* LMR model predicts whether a token is part of a location mention (`<bilou_tag>-LOC`) or not ("O"), while a *type-based* LMR model predicts whether a token is part of LM of specific type (`<bilou_tag>-<location_type>`) or not ("O"). Where `<location_type>` is one of the 9 types presented in Section 4.2.

4 Dataset Construction

This section discusses design objectives, data selection, and data annotation details.

4.1 Design Objectives and Data Selection

During the development of our Arabic LMR dataset, we established five key design objectives: (1) to expand geographical coverage by incorporating as many Arab countries as possible, (2) to incorporate a variety of recurrent disaster types specific to the Arab world, (3) to ensure a balance of location granularity (i.e., cities, districts), (4) to ensure broad temporal coverage, and (5) to increase relevance to disaster response and management tasks by discerning informative content, as social media event streams are often noisy. We analyzed the existing disaster-related Twitter datasets in Arabic and selected Kawarith (Alharbi and Lee, 2021), as it contains tweets from 22 disaster events from the Arab world. We selected tweets (ids) from seven events (listed in Table 2) labeled as relevant for humanitarian purposes. The selected tweet ids (6,182) are used to download full tweet content using the Twitter API, which resulted in 4,593 tweets.

4.2 Dataset Annotation

We perform two types of annotation on the selected data. The first involves human annotators identifying toponyms, such as geographical names of places, within the tweet text. In the second, the annotators assign location types to the identified toponyms. These location types include country, province, city/town, district, neighborhood, road/street, natural points of interest such as rivers and seas, and human-made points of interest such as schools and hospitals. Toponyms that do not belong to the defined location types (e.g., islands, villages, camps) are assigned the "other location" label.

Seven graduate-level students were trained[4] to carry out the annotation task, voluntarily without any monetary or course credit benefits, using the WebAnno NLP annotation tool[5].

We selected the WebAnno tool as it supports Unicode right-to-left languages (e.g., Arabic). To ensure the quality of annotations, we selected the annotators to be either a citizen or having a good familiarity with the country of the disaster event. All annotators had to pass a quiz of 20 tweets before being eligible to start the annotation task.

Disagreements between annotators were examined by an additional meta-annotator and resolved. In Table 2, columns "# LMs (unique)", we show the total number of annotated LMs and the unique number of LMs after de-duplicating them per event in parentheses. The unique number of LMs varies according to the granularity of the affected area. On average, 26% of the LMs are unique. We further report the "Hapax" per data split.

5 Dataset Description and Quality

In this section, we evaluate IDRISI-RA datasets for reliability, diversity, and coverage.

5.1 Reliability

To evaluate the quality of the dataset, we compute the Inter-annotator Agreement (IAA) that quantifies the reliability of annotations. We compute Cohen's Kappa (Cohen, 1960) for both annotation tasks separately and jointly. Results in Figure 1, show the average reliability achieved is 83% (almost perfect), 67% (substantial), 70% (substantial), for LOC (i.e., toponym identification task), TYPE

[4]We share the annotation guidelines publicly in the GitHub repository: https://github.com/rsuwaileh/IDRISI/

[5]https://webanno.github.io/

Event	# T	# $T_{\|LM\|=0}$	# LMs (unique)	$Hapax_{tr}$	$Hapax_{dv}$	$Hapax_{ts}$
Jordan FLD 2018	527	107	897 (108)	2	5	0
Kuwait FLD 2018	1,269	503	1,137 (140)	25	1	11
Cairo BMB 2019	268	1	623 (024)	5	1	1
Hafr FLD 2019	514	46	752 (112)	27	6	5
Dragon STR 2020	305	122	338 (160)	45	4	7
Beirut BMB 2020	349	63	550 (061)	9	0	9
CoVID-19	1,361	777	939 (313)	70	15	24
Total	4,593	1,619	5,236 (918)	183	32	57

Table 2: Statistics of tweets (referred as T) in IDRISI-RA dataset. "Hapax" refers to the number of LMs that appear once in training (tr), development (dv), and test (ts) sets.

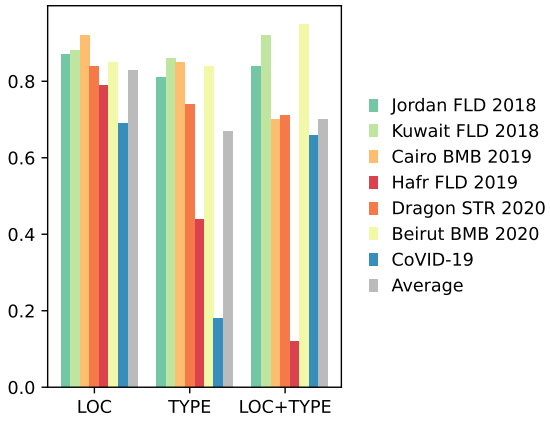


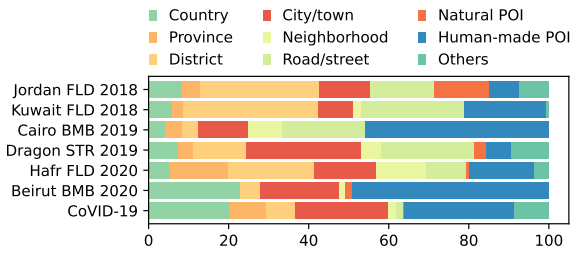
Figure 1: The Inter Annotator Agreement using Cohen's Kappa for IDRISI-RA per event. 0.2, 0.4, 0.6, 0.8 indicate the degree of reliability as slight, fair, moderate, and substantial, respectively.

(i.e., location type assignment), and LOC+TYPE, respectively. All events show high-quality annotations, except the "Hafr Floods 2019" event with 12% agreement for the TYPE task (slight reliability) and 44% for LOC+TYPE (moderate reliability). Upon investigation, we found that "Hafr Albatten" is the most frequent LM in the dataset; one annotator assigns "city" type for all occurrences, and the other assigns "province". While both annotators are correct (as in the Arab world, both types are used interchangeably), we anticipate the agreement level to increase when accepting both types. Furthermore, the COVID-19 event shows slight agreement for the TYPE task due to similar reasons across Arab countries.

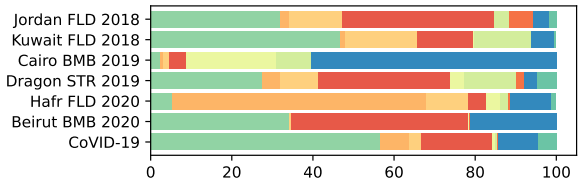
5.2 Coverage and Diversity

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

Geographical Coverage: IDRISI-RA covers 5 distinct Arab countries, namely Jordan, Kuwait, Egypt, Saudi Arabia, Lebanon. Additionally, the whole Arab world is represented by COVID-19 pandemic event. Figure 2a shows the distribution of distinct LMs per location type.



(a) Distinct Location Mentions.



(b) All Location Mentions.

Figure 2: Distribution of location types in IDRISI-RA.

Domain Coverage: IDRISI-RA represents the most happening disaster types in the Arab world that are discussed over Twitter (Alabbas et al., 2017; Alharbi and Lee, 2019; Ameen et al., 2020; Alharbi and Lee, 2021), including 3 floods, 2 explosions, 1 storm, and the global COVID-19 pandemic.

Location Types Coverage: Figure 2b shows the distribution of the location types. It is evident that the coarse-grained (e.g., Country, State, and City) LMs dominate the dataset due to Kawarith collection strategy that depends on tracking relevant keywords, mostly hashtags which are the names of the coarse affected areas by the disaster event.

Temporal Coverage: IDRISI-RA covers recent disaster events happened between 2018-2020. The time span for events is approximately 8.8 days, on average (refer to Table 6 in Appendix A). In Figure 5 in Appendix A, we depict the number of tweets during two events showing the coverage over important developments.

Dialectical Distribution: To analyze the distribution of dialects vs. MSA in IDRISI-RA, we employed the *ASAD* dialectical classifier (Hassan et al., 2021). We found around 86.8% of the tweets in the dataset are MSA. Table 3 shows the dialectical distribution of around 13% tweets. The largest portion goes to the Egyptian dialect as Cairo BMB 2019 and Dragon STR 2020 happened in Egypt. The next dialect is Kuwaiti as Kuwait FLD 2018 event contains the second top number of tweets. Qatari and Saudi dialects are very close to the Kuwaiti dialect which explains their prevalence.

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

Table 3: The dialects distribution in IDRISI-RA. The 18 countries are represented by their 2-letter ISO codes.[6]

6 Benchmarking Experiments

In this section, we discuss IDRISI-RA benchmarking experiments that we conducted to provide baselines for the research community.

6.1 Experimental Setup

We benchmark IDRISI-RA datasets under different tasks, data, and domain setups. The LMR task setups are *type-less* and *type-based*. The data setups are *(i) random*, where we ignore tweets posting timestamp, and *(ii) time-based*, where tweets are chronologically ordered. Tweets are randomly shuffled and split into 70% training, 10% development, and 20% test sets, per event. We report the detailed stats in Table 6 in Appendix A.

6.2 Learning Models

We employ one deep learning-based and three traditional machine learning models for benchmarking, as described below.

- ***CAMeLBERT-Mix*** (CML) (Inoue et al., 2021): An NER model that is trained on ANERcorp dataset, including MSA, Dialectal Arabic (DA), and Classical Arabic (CA) data.
- ***Farasa*** (FRS) (Abdelali et al., 2016): A commonly used NER model in Farasa Arabic tool.
- ***Conditional Random Fields*** (CRF) (Lafferty et al., 2001): We employ CRF due to its competitive performance in the NER task. We used word syntactic features, including the suffix, POS tag, and the context (adjacent words and their syntactical features).
- ***BERT-based*** (BRT) (Abdul-Mageed et al., 2020): We selected MARBERT model for its superiority in Arabic Twitter NER when used for embeddings (Benali et al., 2022).

All these scientific artifacts are used according to their terms and conditions for research purposes.

6.3 Hyperparameter Tuning

During training, we tuned the hyperparameters of the BERT-based model including the sequence length, the batch size, the number of training epochs, and the learning rate in light of the recommended values by (Devlin et al., 2019). For the CRF-based models, we experimented with five available training algorithms and their hyperparameters. We report the full details on hyperparameters tuning in Appendix B.

6.4 Evaluation Measures

To evaluate the LMR models, we compute the harmonic mean (F_1 score) of Precision (P) and Recall (R). We extended the *seqeval (v1.2.2)* package [7] to evaluate the models per tweet and report the average performance, and reward the models when correctly predicting no LMs for a tweet.

6.5 Results

Type-less LMR: Table 4 presents the F_1 results of all type-less models. The detailed results including precision and recall are in Appendix B. MARBERT-based model (BRT) achieves the best performance for both *random* and *time-based* scenarios. Next in order are the CRF, FARASA (FRS), and CAMeLBERT-Mix (CML). Although the CAMeLBERT-Mix is considered a BERT-based model, it shows poor performance compared to MARBERT-based model, as it was fine-tuned on

[6] en.wikipedia.org/wiki/ISO_3166-1_alpha-2

[7] https://github.com/chakki-works/seqeval

news wire documents for the NER (entities include LOC, ORG, PER, and MISC) task.

Type-based LMR: Results in Table 4 show that the MARBERT-based LMR is evidently the best model for the *random* data setup. We anticipate the reason behind the lower performance of the CRF model to be the limited features used to train the Arabic version (refer to Section 6.2). The CRF model exhibits comparable F_1 scores to the MARBERT-based model in the *time-based* data setup. To answer **RQ1**, we confirm the need for specific-LMR datasets and models that can perform effectively over disaster tweets.

7 Dataset Generalizability

In this section, we empirically study the generalizability of IDRISI-RA dataset. For that, we employ the best LMR model, MARBERT-based (refer to Section 6); hereafter, we refer to it as "the model". We study three dimensions: *(i)* **generalizability to unseen events** regardless of their type and geolocation, *(ii)* generalizability to unseen events of the same or different disaster types (**domain generalizability**), and *(iii)* generalizability to unseen events that happened in the same or different countries (**geographical generalizability**).

7.1 Experimental Setups

We run our experiments under both *type-less* and *type-based* task setups for only *random* data setup. We tune the hyperparameters of the model for every setup (refer to Section 6.3). We define the *source dataset* as the dataset (or the combination of datasets) used to *train* the model, and the *target dataset* as the dataset used to *test* it.

Domain generalizability: We examine the model's performance under cross- and in-domain transfer setups (Suwaileh et al., 2020). The "domain" in our experiments refers to the type of disaster event. IDRISI-RA dataset covers the four most occurring disaster types in the Arab world: flood, bombing, storm, and pandemic. A transfer data setup is composed of source-target pair, resulting in 16 runs.

Geographical generalizability: We examine the model's performance over events that took place in different countries than the source dataset. IDRISI-RA covers five countries (refer to Table 6 in Appendix A), besides the global COVID-19. A transfer data setup is composed of source-target pair, resulting in 42 runs after excluding the *target* runs.

7.2 Results

Generalizability to unseen events: Table 5 shows the results of the model both with and without (i.e., zero-shot) the target event. The results for the *type-less* LMR demonstrate the potential of IDRISI-RA dataset under the zero-shot setting. The difference against the target runs is mostly negligible. Due to the difficulty of the *type-based* LMR, the performance under zero-shot learning is significantly lower than the target runs. However, the zero-shot results are still within a reasonable range (i.e., average F_1 0.88), which demonstrates the effectiveness of models trained on IDRISI-RA. To answer **RQ2**, we confirm that training on IDRISI-RA generates generalizable Arabic LMR models that achieve, on average, around 0.75 and 0.88 F_1 scores for the *type-less* and *type-based* LMR, respectively.

Domain generalizability: Figure 3 illustrates the F_1 scores of the models over the target sets.

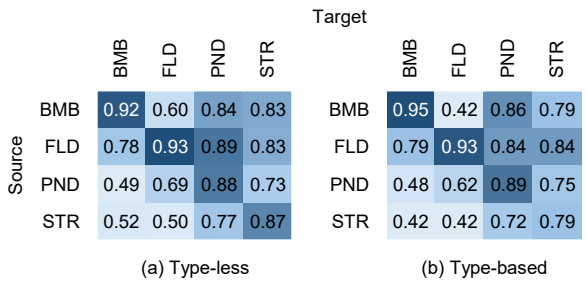


Figure 3: The F_1 results for the domain generalizability within IDRISI-RA under *random* data setup.

In-Domain: Ideally, the best results should lay on the diagonal which depicts the in-domain setup. This assumption holds for all runs, except the STR-to-STR runs in the *type-based* LMR (Figure 3.b). Training on "bombing" data in the BMB-to-STR setup achieves comparable results to training on "storm" data in the STR-to-STR, because both source and target data share the same or close affected areas (Egypt and Lebanon), which could imply the overlap of toponyms' occurrences and patterns. The "bombing" (BMB) includes data from the *Cairo Bombing 2019* in Egypt and the *Beirut Explosion 2020* in Lebanon. The "storm" (STR) test data contains Dragon storms 2020 that affected Egypt and Jordan, among a few LMs from Levantine Arabic. Moreover, the FLD-to-STR run achieves 6.3% better performance compared to the STR-to-STR run, as the FLD source data is approximately 7.5 times larger in size than the "storm" STR source. The effect of training dataset size

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

Table 4: The F_1 results for the LMR models on IDRISI-RA.

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

Table 5: The F_1 results for the MARBERT-based LMR model under *zero-* and *target* training setups.

on these results could be confirmed by the relatively low F_1 scores when the model trained on the "storm" data that has the smallest training data.

Cross-Domain: Generally, the right upper part above the diagonal shows better results than the counterpart, except for the BMB-to-FLD where the size of training data influences the results. We also note here that the model is tuned for every source-to-target transfer setup over the development splits, hence, the poor results on the test splits could indicate overfitting that prevents generalizability. This motivates the use of more advanced transfer learning techniques. To answer **RQ3**, we confirm that IDRISI-RA can generate acceptable domain generalizable models for the most disaster types. It also provides challenging examples for the LMR models.

Geographical generalizability: Figure 4 shows the F_1 scores of the models over the target countries that are the same or different than the affected area of the source data. The model achieves approximately 0.61 and 0.84 F_1 scores, on average, for *type-less* and *type-based* LMR, respectively. The top performance is achieved over "GL" target data that refers to the COVID-19 event due to its geographical coverage. To answer **RQ4**, we found that IDRISI-RA can generate reasonable geographically generalizable models.

8 Silver IDRISI-RA

Given the dearth of Arabic LMR datasets, we expand the size of IDRISI-RA dataset by employing the best-performing LMR model to infer labels for 1.2 million Arabic tweets posted during 22 disaster events. We call the resulting data as the *silver* version to indicate the reliability of annotations. The inference model is trained using the entire *type-based* and *type-less* train sets from all events. All development partitions are used for tuning the hyperparameters of the MARBERT-based model. The model applied to the tweets resulted in 884,217 LMs (18,609 distinct). We note that this large dataset can potentially be useful for a variety of LMR approaches, e.g., domain adaptation, transfer learning, and semi-supervised learning.

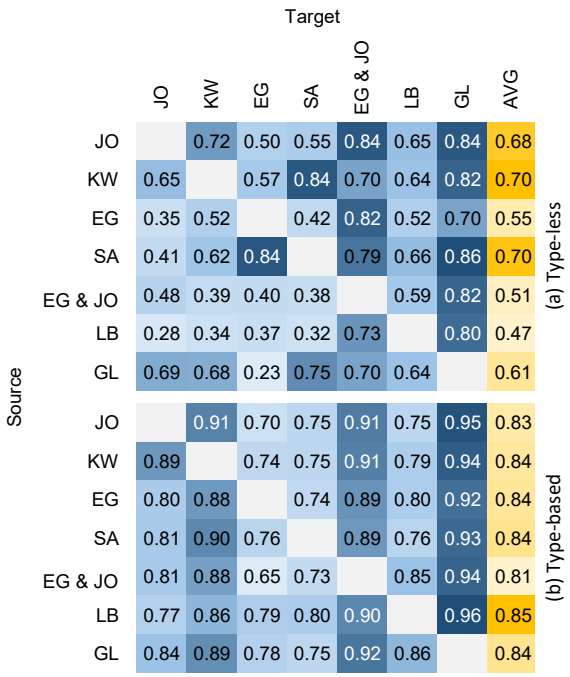


Figure 4: The F_1 results for the geographical generalizability within IDRISI-RA under *random* data setup.

9 Research Use Cases

Releasing IDRISI-RA empowers research on different applications, other than the geolocation. In this sections, we discuss a few use cases.

Event/incident detection: People tend to mention where events/incidents take place when they report them (Hu and Wang, 2020). Following (Sankaranarayanan et al., 2009) and (Watanabe et al., 2011), IDRISI-RA can be employed to detect event/incident on Twitter by extracting LMs.

Relevance filtering: Prior studies show that the geographical references in social media messages could indicate their relevance and informativeness for a disaster event (Vieweg et al., 2010; De Albuquerque et al., 2015). Therefore, IDRISI-RA can be utilized to train relevance filtering models.

Displacement monitoring: Extracting LMs from tweets shared by displaced people and refugees can help predicting their future paths. In fact, monitor the population movement during emergencies is very useful for responders. Thus, IDRISI-RA can be exploited in modeling the population movement.

Geographical retrieval: IDRISI-RA could serve the GIR retrieval techniques that rely on detecting locations and spatial references in queries and documents (García-Cumbreras et al., 2009).

10 Ethical Considerations

While Twitter allows users to self-manage their privacy by disabling the geo-tagging functionalities, "even well informed and rational individuals cannot appropriately self-manage their privacy" (Solove, 2012) due to lack of awareness on how such data can be collected or commercially used. However, geolocation extraction could be justified in the context of social good during *natural* disaster events where protecting geographical privacy could be of least importance as affected people need to rescue their lives or get basic necessities of life (Crawford and Finn, 2015). The privacy of affected people during *human-made* disaster events is more critical because revealing their locations during conflicts and wars could risk their lives. Therefore, we limit our dataset to *natural* disaster events and COVID-19 pandemic and apply a couple of de-identification steps to protect users' privacy (refer to Section A). We further limit the dataset usage to research purposes only by releasing it under the Creative Commons Attribution 4.0 International License.[8] Additionally, we emphasize that systems developed for LMR using IDRISI-RA dataset should implement proper mechanism for preserving user privacy.

11 Conclusion

We introduced IDRISI-RA, the first Arabic LMR Twitter dataset. It contains 22 disaster events of different types that happened in the Arab region. We manually- (gold) and automatically annotated (silver) about 4.6K and 1.2M tweets. Both versions are annotated for location mentions and location types which form the value and uniqueness of IDRISI-RA. Our analysis showed that IDRISI-RA is second to none in empowering research for Arabic LMR. Additionally, the extensive experiments emphasize the need for developing LMR-specific models for the disaster domain. The developed LMR baselines are simple yet competitive ones. The results also demonstrated the decent generalizability of IDRISI-RA. For future work, we plan to extend the annotations for the Location Mention Disambiguation (LMD) task. We further plan to explore different transfer learning, domain adaptation, and active learning techniques to tackle the LMR task.

[8] https://creativecommons.org/licenses/by/4.0/legalcode

Limitations

There are a few shortcomings that we discuss below:

Underrepresented fine-grained LMs: Although we had chosen a careful sampling method focused of event-centric informative dataset aiming to increase the likelihood of fine-grained LMs' occurrence (Kitamoto and Sagara, 2012), we think the low frequency of fine-grained LMs in IDRISI-RA is a major limitation as it contains solely 25.5% fine-gained LMs.

Human errors: There are some human errors made during annotation due to the difficulty of the task.

- Annotators sometimes fail in distinguishing between *Location* and *Organization* entities (e.g., "Red Cross").
- Different location types could be used interchangeably for the same locations which forms a difficulty for annotators (refer to Section 5.1).
- Annotators highlight the locations that are mentioned as descriptions within the context of the tweet.

We plan to overcome these errors as part of Location Mention Disambiguation (LMD) annotation that aim to remove ambiguity of geo/geo entities (as a sequel to the geo/non-geo LMR annotations).

Temporary locations: Temporary facilities (i.e., medical camps, shelters, etc.) are constructed during emergencies to provide resources and support for the affected people. However, these facilities could be disassembled (e.q., quarantine centers) once the emergency event is over. Additionally, the names of some locations could change during emergencies. For example, allocating a specific school as a shelter and giving it a new expressive name (e.g., "main shelter"). Once the disaster event is over, the school will return to providing its original services. The difficulty of these temporary locations lies in their need for context when resolved. Although they are important for the affected people and response authorities, not all of them are labeled in IDRISI-RA.

Generalizability: Due to the absence of *public* LMR datasets, we could not compare the generalizability of IDRISI-RA against existing LMR datasets. Hence, we study the generalizability within IDRISI-RA for domain and geographical aspects.

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. We would like to thank the in-house annotators for their valuable help including Noura Abdullah, Aisha Suwaileh, Rasha Hamdoon, Najlaa Alfuhaida, Hana Shamayleh, Lamiaa Basyoni, Sara Alrasbi, and Nada Abo Eita.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *NAACL*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Noora Al Emadi, Sofiane Abbar, Javier Borge-Holthoefer, Francisco Guzman, and Fabrizio Sebastiani. 2017. Qt2s: A system for monitoring road traffic via fine grounding of tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Waleed Alabbas, Haider M. al Khateeb, Ali Mansour, Gregory Epiphaniou, and Ingo Frommholz. 2017. Classification of colloquial arabic tweets in real-time to detect high-risk floods. In *2017 International Conference On Social Media, Wearable And Web Analytics (Social Media)*, pages 1–8.

Alaa Alharbi and Mark Lee. 2019. Crisis detection from arabic tweets. In *Proceedings of the 3rd workshop on arabic corpus linguistics*, pages 72–79.

Alaa Alharbi and Mark Lee. 2021. Kawarith: an Arabic Twitter corpus for crisis events. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Balsam Alkouz and Zaher Al Aghbari. 2018. Leveraging cross-lingual tweets in location recognition. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0084–0089. IEEE.

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

Yasmeen Ali Ameen, Khaled Bahnasy, and Adel E. Elmahdy. 2020. Classification of arabic tweets for damage event detection. *International Journal of Scientific & Engineering Research*, 11.

Khaled Bahnasy, Adel El-Mahdy, et al. 2020. Twitter analysis based on damage detection and geoparsing for event mapping management. *Future Computing and Informatics Journal*, 5(1):1.

Brahim Ait Benali, Soukaina Mihi, Nabil Laachfoubi, and Addi Ait Mlouk. 2022. Arabic named entity recognition in arabic tweets using bert-based models. *Procedia Computer Science*, 203:733–738.

Xu Chen, Judith Gelernter, Han Zhang, and Jin Liu. 2019. Multi-lingual geoparsing based on machine translation. *Future Generation Computer Systems*, 96:667–677.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Kate Crawford and Megan Finn. 2015. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, 80:491–502.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567.

Kareem Darwish and Wei Gao. 2014. Simple effective microblog named entity recognition: Arabic as an example. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2513–2517.

Joao Porto De Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International journal of geographical information science*, 29(4):667–689.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Miguel Á García-Cumbreras, José M Perea-Ortega, Manuel García-Vega, and L Alfonso Ureña-López. 2009. Information retrieval with geographical references. relevant documents filtering vs. query expansion. *Information processing & management*, 45(5):605–614.

Rob Grace, Jess Kropczynski, and Andrea Tapia. 2018. Community coordination: Aligning social media use in community emergency management. In *Proceedings of the 15th ISCRAM Conference*.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118.

Yingjie Hu and Jimin Wang. 2020. How Do People Describe Locations during a Natural Disaster: An Analysis of Tweets from Hurricane Harvey. *Leibniz International Proceedings in Informatics, LIPIcs*, 177.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, June*.

Asanobu Kitamoto and Takeshi Sagara. 2012. Toponym-based geotagging for observing precipitation from social and scientific data streams. In *Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '12, page 23–26, New York, NY, USA. Association for Computing Machinery.

Jessica Kropczynski, Rob Grace, Julien Coche, Shane Halse, Eric Obeysekare, Aurelie Montarnal, Frederick Benaben, and Andrea Tapia. 2018. Identifying Actionable Information on Social Media for Emergency Dispatch. In *ISCRAM Asia Pacific 2018: Innovating for Resilience – 1st International Conference on Information Systems for Crisis Response and Management Asia Pacific.*, pages p.428–438, Wellington, New Zealand.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289.

Shervin Malmasi and Mark Dras. 2015. Location mention detection in tweets and microblogs. In *Conference of the Pacific Association for Computational Linguistics*, pages 123–134. Springer.

Sabrina McCormick. 2016. New tools for emergency managers: an assessment of obstacles to use and implementation. *Disasters*, 40(2):207–225.

Christian Reuter, Thomas Ludwig, Marc-André Kaufhold, and Thomas Spielhofer. 2016. Emergency services' attitudes towards social media: A quantitative and qualitative survey across europe. *International Journal of Human-Computer Studies*, 95:96–111.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51.

Daniel J Solove. 2012. Introduction: Privacy self-management and the consent dilemma. *Harv. L. Rev.*, 126:1880.

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

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

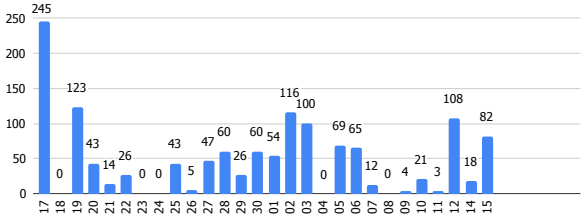
Reem Suwaileh, Muhammad Imran, Tamer Elsayed, and Hassan Sajjad. 2020. Are we ready for this disaster? towards location mention recognition from crisis tweets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6252–6263.

Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088.

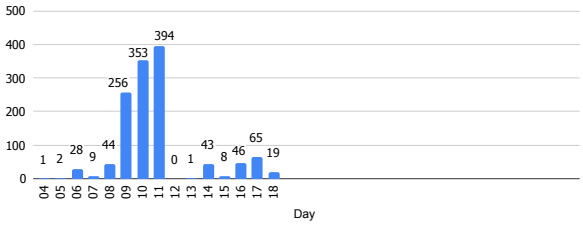
Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. 2011. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2541–2544.

A Data Release

The IDRISI-RA dataset is released[9] data setups that are *random* and *Time-based*. The location mention and location type annotations are made available for the community to enable development of *type-less* and *type-based* LMR models. The data is released in **JSONL** format where every lines



(a) COVID-19 (17 May - 15 Jun)



(b) Kuwait FLD 2018 (04 Nov - 18 Nov)

Figure 5: The temporal coverage of tweets in IDRISI-RA.

corresponds to one tweet with the following properties: "text", "created_at", "info_class" adopted from Kwaraith dataset, and "location_mentions".

We processed the data to de-identify it as follows:

- We do not release the the user identifiers, i.e., "user_id".
- We replace the user mentions (i.e.,"@") in the tweet text by "@0" of the same length as the mention length. For example, if the mention is "@someuser", we replace it with "@00000000".
- We keep the tweet ids, i.e., "id", to allow re-crawling tweets for extracting more information, e.g., meta data, social network properties, etc. This allows developing LMR models that utilize different features beyond the textual content.

Table 6 shows the detailed statistics of IDRISI-RA dataset for the random and time-based setups per event. Figure 6 presents example Arabic tweets translated into English and highlights different types of LMs.

In Figure 5, we depict the temporal coverage of COVID-19 and Kuwait FLD 2018 events.

B Hyperparameter Tuning and Results

For the BERT-based models, we tuned the sequence length, the batch size, the number of training epochs, and the learning rate in light of the recommended values by (Devlin et al., 2019) as: batch size of 8, 16 or 32, number of epochs of 2, 3,

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

Event	Arabic tweet	English Translation (Google)
Jordan FLD 2018	إخلاء عائلة إلى مخيم الكرامة أماكن أخرى الكشفي بالشونة الجنوبية منطقة #الاردن #سيول_الاردن [رابط]	Evacuation of a family to Al-Karama Scout Camp$_{\text{Other location}}$ in Southern Shuna$_{\text{District}}$ #jordan #jordan_floods [URL]
Kuwait FLD 2018	هيئة الطرق: تم سحب مياه نفق المنقف طريق/شارع بالكامل وسيتم افتتاحه لمرتادي الطرق بالساعات القادمة #الكويت دولة #امطار_الكويت	Roads Authority: The Mangaf tunnel$_{\text{Road/street}}$ has been completely withdrawn, and it will be opened to road users in the coming hours #kuwait$_{\text{Country}}$ #kuwait_floods
Cairo BMB 2019	عاجل| #الصحة: 17 مصابا في الحصر المبدئي لانفجار المعهد القومى للأورام مَعْلَم من صنع الإنسان [رابط] #معهد_الأورام مَعْلَم من صنع الإنسان	Urgent | #health_authority: 17 injured in the initial count of the National Cancer Institute$_{\text{Human-made POI}}$ explosion [URL] #cancer_institue$_{\text{Human-made POI}}$
Hafr FLD 2019	حي الخالدية حي واللي قريب منه الله يستر عليهم المفروض يهاجرون موقعهم خطر مرة والسيول جتهم من كل جهة 🚤 🚤 #حفرالباطن_الان	Al-Khalidiyah neighborhood$_{\text{Neighborhood}}$ and the one nearby May God protect them They are supposed to migrate Their site is dangerous And torrents came from every direction 🚤 🚤 #hafar_albatin_now
Beirut BMB 2020	#انفجار_مرفا_بيروت #لبنان | حريق كبير في جبل مشغرة مَعْلَم طبيعي طال المنازل #بيروت_عم_تبكي	#beirut_port_explusion #lebanon | A big fire in Mashghara Mountain$_{\text{Natural_POI}}$ affected homes #beirut_crying

Figure 6: Tweet Example from IDRISI-RA. Highlighted text and subscripts refer to LMs and their location types.

or 4, and learning rate of 5E-5, 3E-5, or 2E-5. We also experimented with a sequence length of 128 or 256.

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

We ran our hyperparameter tuning experiments for about 3 days on a cluster of 46 GPUs of different NVIDIA models including p100, v100, v100-NVLINK, and T4.

Tables 7 and 8 show the best hyper-parameters and detailed results for the CRF and BERT-based LMR models for IDRISI-RA, respectively. Tables 9 shows the best hyper-parameters of the models under *disaster domain transfer* setting.

Event	Country	Time Period	Tweets					LMs				
			TRN	DEV	TST	All	LM_0	TRN	DEV	TST	All	Uniq
Random												
Jordan FLD 2018	JO	10/25 - 11/18	371	53	103	527	107	614	89	194	897	108
Kuwait FLD 2018	KW	11/04 - 11/18	889	127	253	1,269	503	820	93	224	1,137	140
Cairo BMB 2019	EG	08/04 - 08/04	189	27	52	268	1	417	66	140	623	24
Hafr FLD 2019	SA	10/25 - 10/29	364	52	98	514	46	520	83	149	752	112
Dragon STR 2020	EG & JO [a]	03/11 - 03/15	217	31	57	305	122	252	27	59	338	160
Beirut BMB 2020	LB	08/04 - 08/07	245	35	69	349	63	378	49	123	550	61
CoVID-19	Global	05/17 - –/– [b]	959	137	265	1,361	777	637	96	206	939	313
Time-based												
Jordan FLD 2018	JO	10/25 - 11/18	371	53	103	527	107	631	79	187	897	108
Kuwait FLD 2018	KW	04/11 - 04/18	889	127	253	1269	503	772	112	253	1,137	140
Cairo BMB 2019	EG	08/04 - 08/04	189	27	52	268	1	458	53	112	623	24
Hafr FLD 2019	SA	10/25 - 10/29	364	52	98	514	46	526	79	147	752	112
Dragon STR 2020	EG & JO [a]	03/11 - 03/15	217	31	57	305	122	248	39	51	338	160
Beirut BMB 2020	LB	08/04 - 08/07	245	35	69	349	63	400	55	95	550	61
CoVID-19	Global	05/17 - –/– [b]	959	137	265	1361	777	599	109	231	939	313
All	-	-	3,234	462	897	4,593	1,619	3,634	526	1,076	5,236	918

Table 6: Detailed information and statistics of IDRISI-RA datasets, both *random* and *time-based* setups. "TRN", "DEV", "TST" refer to the training, development, and test splits, respectively. LM_0 refers to the number of tweets with no LMs.

[a] While Dragon storms have affected several Arab countries and LMs are from different countries, Kawarith creators had intentionally focused on the Egyptian tweets. We found LMs from Jordan as well.

[b] The last tweet in the chronologically sorted tweets was published in 2020/06/15 but the pandemic was ongoing.

Event	Algo.	HP1	HP2	P	R	F1
Random data setup Type-less LMR						
Jordan Floods	arow	*variance*=0.1	*gamma*=0.16	0.841	0.845	0.843
Kuwait Floods	arow	*variance*=1	*gamma*=0.125	0.865	0.603	0.711
Cairo Bombing	l2sgd	*c2*=0.2	*ce*=1e-2	0.971	0.964	0.968
Hafr Floods	lbfgs	*c1*=0.05	*c2*=0.05	0.881	0.799	0.838
Dragon Storms	pa	*c*=1	*error_sensitive*=TRUE	0.787	0.627	0.698
Beirut Explosion	arow	*variance*=0.1	*gamma*=0.5	0.943	0.813	0.873
CoVID-19	arow	*variance*=1	*gamma*=1	0.787	0.539	0.640
Type-based LMR under random data setup						
Jordan Floods	lbfgs	*c1*=0.25	*c2*=0.25	0.766	0.786	0.776
Kuwait Floods	arow	*variance*=0.5	*gamma*=0.1	0.822	0.530	0.644
Cairo Bombing	l2sgd	*c2*=0.9	*ce*=1e-4	0.929	0.938	0.933
Hafr Floods	l2sgd	*c2*=0.5	*ce*=1e-4	0.891	0.776	0.829
Dragon Storms	ap	*epsilon*=1e-5	-	0.659	0.569	0.611
Beirut Explosion	l2sgd	*c2=0.15*	*ce=1e-2*	0.692	0.874	0.772
CoVID-19	arow	*variance=0.1*	*gamma=0.125*	0.676	0.597	0.634
Type-less LMR under time-based data setup						
Jordan Floods	pa	*c*=0	*error_sensitive*=TRUE	0.837	0.837	0.837
Kuwait Floods	pa	*c*=0	*error_sensitive*=TRUE	0.904	0.904	0.904
Cairo Bombing	pa	*c*=2	*error_sensitive*=TRUE	0.714	0.708	0.708
Hafr Floods	l2sgd	*c2*=0.75	*ce*=1e-6	0.861	0.861	0.859
Dragon Storms	pa	*c*=0	*error_sensitive*=TRUE	0.872	0.872	0.872
Beirut Explosion	l2sgd	*c2*=0.3	*ce*=1e-3	0.701	0.703	0.701
CoVID-19	pa	*c*=0	*error_sensitive*=TRUE	0.928	0.928	0.928
Time-based data setup Type-based LMR						
Jordan Floods	arow	*variance*=0.25	*gamma*=0.1	0.776	0.778	0.775
Kuwait Floods	pa	*c*=0	*error_sensitive*=TRUE	0.891	0.891	0.891
Cairo Bombing	l2sgd	*c2*=0.05	*ce*=1e-6	0.740	0.741	0.737
Hafr Floods	l2sgd	*c2*=0.05	*ce*=1e-6	0.882	0.883	0.882
Dragon Storms	pa	*c*=0	*error_sensitive*=TRUE	0.880	0.880	0.880
Beirut Explosion	arow	*variance*=0.5	*gamma*=0.16	0.617	0.643	0.621
CoVID-19	pa	*c*=0	*error_sensitive*=TRUE	0.901	0.901	0.901

Table 7: The best hyper-parameters and results for CRF model over IDRISI-RA. The column “Algo." refers to the training algorithm of CRF. The “HP1" and “HP2" refer to the tuned hyper-parameters with respect to the algorithm.

Event	Random							Time-based						
	e	bs	lr	sl	P	R	F1	e	bs	lr	sl	P	R	F1
Type-less														
Jordan Floods	3	8	3e-5	256	0.954	0.957	0.953	3	8	3e-5	128	0.911	0.916	0.903
Kuwait Floods	4	16	3e-5	256	0.935	0.925	0.928	3	16	3e-5	128	0.895	0.905	0.893
Cairo Bombing	3	8	3e-5	256	0.995	0.986	0.989	2	8	3e-5	128	0.934	0.939	0.936
Hafr Floods	4	8	3e-5	128	0.883	0.883	0.879	4	8	3e-5	256	0.878	0.897	0.878
Dragon Storms	3	8	3e-5	128	0.878	0.873	0.870	4	8	4e-5	128	0.882	0.868	0.869
Beirut Explosion	4	8	3e-5	128	0.885	0.851	0.855	4	8	3e-5	256	0.601	0.611	0.582
CoVID-19	3	8	3e-5	128	0.889	0.884	0.881	3	8	3e-5	256	0.896	0.914	0.897
Type-based														
Jordan Floods	4	8	3e-5	128	0.916	0.907	0.908	4	8	3e-5	256	0.880	0.872	0.862
Kuwait Floods	4	8	3e-5	128	0.933	0.925	0.925	3	8	3e-5	256	0.874	0.892	0.879
Cairo Bombing	4	8	3e-5	128	0.984	0.970	0.975	4	8	3e-5	256	0.930	0.935	0.931
Hafr Floods	4	8	3e-5	256	0.870	0.857	0.856	3	8	3e-5	256	0.841	0.854	0.838
Dragon Storms	4	8	3e-5	256	0.798	0.789	0.787	4	8	3e-5	256	0.726	0.722	0.714
Beirut Explosion	4	8	4e-5	256	0.854	0.821	0.813	4	8	5e-5	128	0.616	0.635	0.596
CoVID-19	4	8	3e-5	256	0.895	0.898	0.893	4	8	3e-5	128	0.888	0.898	0.886

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

Source-Target	e	bs	lr	sl	P	R	F1
Type-less							
BMB-BMB	4	8	3e-5	128	0.935	0.917	0.918
BMB-FLD	4	8	3e-5	128	0.584	0.664	0.596
BMB-PND	4	8	3e-5	128	0.854	0.840	0.839
BMB-STR	4	8	3e-5	128	0.839	0.833	0.831
FLD-BMB	3	8	3e-5	256	0.843	0.762	0.779
FLD-FLD	3	8	3e-5	256	0.940	0.928	0.930
FLD-PND	3	8	3e-5	256	0.898	0.887	0.887
FLD-STR	3	8	3e-5	256	0.839	0.819	0.826
PND-BMB	3	8	3e-5	128	0.526	0.524	0.488
PND-FLD	3	8	3e-5	128	0.686	0.744	0.687
PND-PND	3	8	3e-5	128	0.889	0.884	0.881
PND-STR	3	8	3e-5	128	0.749	0.716	0.728
STR-BMB	3	8	3e-5	128	0.574	0.535	0.518
STR-FLD	3	8	3e-5	128	0.491	0.544	0.501
STR-PND	3	8	3e-5	128	0.792	0.768	0.773
STR-STR	3	8	3e-5	128	0.878	0.873	0.870
Type-based							
BMB-BMB	4	8	3e-5	256	0.972	0.934	0.945
BMB-FLD	4	8	3e-5	256	0.396	0.505	0.422
BMB-PND	4	8	3e-5	256	0.876	0.859	0.858
BMB-STR	4	8	3e-5	256	0.798	0.785	0.786
FLD-BMB	3	8	3e-5	256	0.850	0.763	0.786
FLD-FLD	3	8	3e-5	256	0.937	0.935	0.933
FLD-PND	3	8	3e-5	256	0.854	0.846	0.842
FLD-STR	3	8	3e-5	256	0.855	0.838	0.842
PND-BMB	4	8	3e-5	256	0.513	0.507	0.481
PND-FLD	4	8	3e-5	256	0.603	0.703	0.622
PND-PND	4	8	3e-5	256	0.895	0.898	0.893
PND-STR	4	8	3e-5	256	0.781	0.743	0.752
STR-BMB	4	8	3e-5	256	0.469	0.428	0.418
STR-FLD	4	8	3e-5	256	0.406	0.466	0.421
STR-PND	4	8	3e-5	256	0.743	0.715	0.72
STR-STR	4	8	3e-5	256	0.798	0.789	0.787

Table 9: The best hyper-parameters and results for the BERT-based model under *disaster domain transfer* setting e, bs, lr, and sl refer to the hyper-parameters, number of epochs, batch size, learning rate, and sequence length, respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Discussed two main risks: - Noisiness of SM data in Section 3.1 Design Objectives and Data Selection - Revealing individuals' (potentially vulnerable ones) information in Appendix A Data Release

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3 Dataset Construction 5.2 Learning Models

☑ B1. Did you cite the creators of artifacts you used?
5.2 Learning Models

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
5.2 Learning Models Appendix A Data Release

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
5.2 Learning Models

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix A Data Release

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4 Dataset Description and Quality

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3 Dataset Construction Appendix A Data Release

C ☑ Did you run computational experiments?

5 Benchmarking Experiments 6 Dataset Generalizability

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.3 Hyperparameter Tuning Appendix B Hyperparameters Tuning and Results

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.1 Experimental Setup 5.3 Hyperparameter Tuning 6.1 Experimental Setups Appendix B Hyperparameters Tuning and Results

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.5 Results 6.2 Results Appendix B Hyperparameters Tuning and Results We note that, as our aim in this paper is to provide baselines for the community and not to develop models that advance the SOTA, we extensively tune hyperparameters without repeating experiments with different seeds to report an anggregated (e.g., avg) results.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.2 Learning Models 5.4 Evaluation Measures We use default models and parameters

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3.2 Dataset Annotation

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3.2 Dataset Annotation

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3.2 Dataset Annotation

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The annotators are volunteering students and were informed that their annotations will be used for research purposes in the humanitarian domain.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We collect, annotate, and share the data under Developer Agreement and Policy. No approval from other ethics review board.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.2 Dataset Annotation