Kawārith: an Arabic Twitter Corpus for Crisis Events

Alaa Alharbi[1,2] **Mark Lee**[1]
[1]School of Computer Science, University of Birmingham, Birmingham, UK
[2]College of Computer Science and Engineering, Taibah University, Medina, KSA
{*axa1314,m.g.lee*}*@cs.bham.ac.uk*
alaharbi@taibahu.edu.sa

Abstract

Social media (SM) platforms such as Twitter provide large quantities of real-time data that can be leveraged during mass emergencies. Developing tools to support crisis-affected communities requires available datasets, which often do not exist for low resource languages. This paper introduces Kawārith[1] a multi-dialect Arabic Twitter corpus for crisis events, comprising more than a million Arabic tweets collected during 22 crises that occurred between 2018 and 2020 and involved several types of hazard. Exploration of this content revealed the most discussed topics and information types, and the paper presents a labelled dataset from seven emergency events that serves as a gold standard for several tasks in crisis informatics research. Using annotated data from the same event, a BERT model is fine-tuned to classify tweets into different categories in the multi-label setting. Results show that BERT-based models yield good performance on this task even with small amounts of task-specific training data.

1 Introduction

Recent studies have revealed the significant role of Twitter during emergency events (Imran et al., 2013a; Olteanu et al., 2015; Imran et al., 2016). In particular, Twitter messages have been shown to contain valuable and timely crisis-related information that serves to enhance situational awareness and supports humanitarian response efforts. Accessible datasets enable crisis informatics researchers to leverage this valuable user-generated data.

The present study focuses on crisis-related Arabic Twitter content. Alshaabi et al. (2020) reported that Arabic is one of the most used languages on Twitter. Nevertheless, this content has attracted relatively little research attention. Arabic dialects, which differ in their morphology, phonology and syntax, have limited resources. The lack of sufficient resources makes developing natural language processing (NLP) tools utilising Arabic SM messages more challenging. To the best of our knowledge, there is no accessible Arabic Twitter corpus of multi-crisis events. For that reason, we assembled a large-scale crisis-related corpus[2] of more than a million tweets from 22 Middle East crises involving different hazard types. Using NLP techniques, our detailed analysis of the corpus identified the main information types shared on Twitter, and we used these to label a subset of the data. The dataset can facilitate information extraction from SM to support situational awareness and enable decision-makers to respond effectively during mass emergencies. The paper makes a number of contributions to the existing literature.

- Creation and publication of a large-scale crisis-related Arabic Twitter corpus.
- Analysis of the corpus content, including the main information categories of conversations posted during a range of crisis events.
- Compilation of 405 domain-independent multi-dialect Arabic stop words[3].
- Manually annotated Arabic Twitter dataset of more than 12k messages from seven different crises.

The remainder of the paper is organised as follows. Section 2 outlines relevant related work. Section 3 describes how data were collected from the Twitter platform to create Kawārith. Section 4 provides a detailed description and analysis of the corpus. Section 5 discusses the annotation scheme, and section 6 presents the results of fine-tuning a BERT language model to automatically classify the la-

[1]This is the Romanised form of the Arabic word كوارث, meaning crises.

[2]The corpus is available at https://github.com/alaa-a-a/kawarith.

[3]The stop word list is available at https://github.com/alaa-a-a/multi-dialect-arabic-stop-words.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 42–52
Kyiv, Ukraine (Virtual), April 19, 2021.

belled dataset. The paper concludes in Section 7.

2 Related Work

In the publicly available Twitter crisis-related corpora in the crisis informatics literature, supervised approaches have been used to extract messages of interest from human-labelled datasets, including actionable information that enhances situational awareness. In the present context, accessible unlabelled corpora have been used for a range of purposes, including key topic extraction, social analytics and public sentiment assessment during emergency events.

Imran et al. (2013a,b) shared two annotated datasets labelled for two tasks: identifying informative messages that contribute to situational awareness and assigning these to information categories such as *cautions* and *donations*. The first set, ISCRAM2013, comprises 3617 tweets about the Joplin tornado, and the second includes 2987 tweets collected during Hurricane Sandy in 2012. Cobo et al. (2015) collected and published $\sim 2K$ Spanish Tweets from the Chilean earthquake in 2010, which were human-labelled as *relevant* or *irrelevant* to the crisis. Cresci et al. (2015) created the SoSItalyT4 dataset, comprising $\sim 5.6K$ Italian tweets collected during four natural disasters in Italy between 2009 and 2014, which were annotated for the purposes of damage assessment as *damage*, *no damage* or *irrelevant*.

One of the largest labelled and publicly accessible crisis datasets is CrisisLex, which incorporates two collections: CrisisLexT6 (Olteanu et al., 2014) and CrisisLexT26 (Olteanu et al., 2015). CrisisLexT6 includes $60K$ English tweets from six crisis events during 2012 and 2013. The messages were annotated by relatedness to the event in question (*related* vs. *not related*). CrisisLexT26 contains tweets collected during 26 crises that also occurred in 2012 and 2013. About $28K$ posts were annotated in terms of informativeness (*informative* vs. *uninformative*), information type and information source, and most of the subsequent research followed the proposed taxonomies. Imran et al. (2016) released CrisisNLP, a corpus of $\sim 52K$ annotated tweets collected during 19 crisis events between 2013 and 2015. The tweets were manually labelled by information type by paid workers and volunteers. For the purpose of coordinating humanitarian response efforts, different annotation schemes (labels) were used for different event types. Most messages in CrisisLex and CrisisNLP were written in English but they also include posts written in other languages such as Italian, Spanish and French, and both datasets have been widely used in the Twitter crisis detection literature.

Alam et al. (2018a) created CrisisMMD, a multimodal dataset of $\sim 16K$ English tweets with attached images collected in 2017 from seven natural disasters. Entries were labelled according to three dimensions: informativeness, information categories and damage severity. TREC-IS[4] (McCreadie et al., 2020) provided crisis-related Twitter datasets from 48 past emergency events, which were manually labelled by information types and priority levels. Alharbi and Lee (2019) shared a dataset of $4K$ Arabic tweets that refer to four high-risk floods in three Middle Eastern countries during October and November 2018, annotated by relatedness and information type. Hamoui et al. (2020) presented FloDusTA data, comprises $\sim 9k$ tweets from floods, dust storms, and traffic accident events that occurred in Saudi Arabia. The messages were labeled by event type and time of occurrence (*historical*, *immediate*, *future* or *irrelevant*). To identify different types of eyewitness report during crises, Zahra et al. (2020) published a labelled dataset of $\sim 14K$ English tweets gathered during four natural disasters. Kozlowski et al. (2020) built a dataset of $\sim 13K$ French tweets collected from different ecological crises, labelled for *relatedness*, *urgency* and *intention to act* as information types that reflect categories in other studies.

Several large published Twitter crisis corpora that provide no human labels can only be reused by reassembling the data from tweet IDs. Unlike small datasets, Twitter's Developer Policy did not allow the distribution of tweets for large-scale datasets (Zubiaga, 2018). Examples include $6M$ geo-tagged tweet IDs from Hurricane Sandy (Wang et al., 2015), $\sim 7M$ English tweets from Hurricane Harvey (Phillips, 2017) and $35M$ tweet IDs related to Hurricanes Irma and Harvey. Alam et al. (2018b) also published a Twitter corpus of more than $8M$ message IDs collected in 2017 from Hurricanes Irma, Harvey and Maria. More recently, research interest has focused increasingly on the COVID-19 pandemic, and several crisis informatics studies have released large-scale Twitter datasets. While some of these are limited to a single language like English (Lamsal, 2020; Gupta et al., 2020) or Ara-

[4] http://dcs.gla.ac.uk/~richardm/TREC_IS/

bic (Alqurashi et al., 2020; Haouari et al., 2020; Addawood, 2020), others have created multi-lingual datasets (Chen et al., 2020; Banda et al., 2020; Qazi et al., 2020; Singh et al., 2020; Alshaabi et al., 2021). Liu et al. (2020) created EPIC, an epidemic corpus consisting of $\sim 20M$ tweets related to various diseases.

To contribute to this body of research, the present study introduces a large-scale Arabic Twitter corpus and a subset of tweets from seven crises, manually labelled in terms of their content. In contrast to previous presented work, we consider a multi-label annotation scheme, which is described in detail below.

3 Crisis Events and Data Collection

The Kawārith corpus comprises Arabic tweets from 22 crisis events that occurred between October 2018 and September 2020. Kawarith focuses on high- to medium-risk events that are most likely to trigger substantial Twitter activity and encompasses a wide range of hazard types, including floods, shootings, bombing, wildfires, pandemics, sandstorms and explosions. Table 1 lists these crises by date; flood events occurring in the same area are referenced by country and year of occurrence. As these crises occurred in diverse Arabic-speaking regions, the corpus should include tweets written in different dialects. Previous studies have revealed that Arabic dialects are strongly present in SM, although many messages—especially those sent from news and organisation accounts—are written in Modern Standard Arabic (MSA).

Using the Twitter search API [5], these data were collected iteratively. During each crisis, we began by using trending crisis-related hashtags or keywords as query terms. If no relevant trends were found during this initial phase of data collection, we used the API to search Twitter using a logical AND combination of the terms *hazard type* and *crisis location*. Additionally, as an alternative search term, we linked the two terms in hashtag form, as we observed that people tended to use crisis-related hashtags like #سيول_الكويت "#Kuwait_floods" and #جائحة_كورونا "#corona_pandemic" for the Kuwait floods and COVID-19 crises, respectively. This first step led us to crawl an initial set of tweets, which was manually inspected to identify any new hashtags that related strongly to the event. The dataset was then expanded by tracking these hashtags, and this step was repeated until no new relevant hashtags could be found. Finally, we updated our query to include all manually selected keywords linked by logical OR to extract crisis messages in the next timeframe. Concurrently, we updated the query with any new relevant keywords emerging as trends on Twitter. Keywords could be in the form of observed hashtags, phrases or single words. We adopted a cautious approach to keyword selection, often using event-specific (discriminative) terms rather than hazard descriptors such as أمطار غزيرة "heavy rain" or يغرق "drowns" to reduce false positives—especially for flood events, which usually occurred simultaneously. Terms such as country name hashtags were generally disregarded, especially if the event had little impact on that country. The decision to use such terms as search queries was generally based on recently retrieved tweets; a candidate term was added to the query if it retrieved event-relevant messages. Importantly, as we favoured precision rather than recall, many relevant tweets were likely to be missed. However, we are satisfied that our data captured the key aspects of the crises. In the case of COVID-19, we tracked only nine keywords referring to the crisis by name because the event has triggered many other topics (such as conspiracy theories) that were not immediately relevant to our purposes. As our study focused on building an Arabic dataset, data collection was confined to tweets that Twitter tagged as Arabic, and this language parameter necessarily excluded tweets in other languages. In Lebanon, for instance, people also tweeted in Arabizi (Romanised Arabic), English and other languages, which may account for the relatively small volume of Arabic data crawled for those events despite their severity and relevance.

Data gathering continued from the first day of a crisis until the end, which we chose to define as the point at which it no longer triggered conversations on Twitter and related keywords no longer appeared in the Twitter trending list for that geographical area. We treated long-term crises like the COVID-19 as exceptions to this rule. In the case of COVID-19, data collection was delayed until near the peak of the epidemic in the Middle East. In other words, the goal was to obtain representative rather than comprehensive samples. Lists of the query terms and collection dates have been included in the published data. In total, we

[5] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

year	crisis name	country	start date	# tweets	# unique authors	# tweets by verified users
2018	Jordan floods	Jordan	25/10/18	8493	5376	452
2018	Kuwait floods-18	Kuwait	04/11/18	34315	20285	637
2018	Qurayyat floods	Saudi Arabia	10/11/18	9731	6781	176
2018	Hafr Albatin floods-18	Saudi Arabia	14/11/18	6069	4218	105
2018	Leeth floods	Saudi Arabia	23/11/18	9596	6170	99
2019	Khartoum massacre	Sudan	03/06/19	12305	6811	50
2019	Cairo bombing	Egypt	04/08/19	2018	1320	182
2019	Lebanon wildfires	Lebanon	13/10/19	8585	5907	100
2019	Egypt floods	Egypt	21/10/19	10938	4138	51
2019	Hafr Albatin floods-19	Saudi Arabia	25/10/19	14546	8398	120
2019	Karbala massacre	Iraq	28/10/19	11961	6593	328
2019	Dubai floods	United Emirates	10/11/19	2480	1983	75
2019	Coronavirus disease	Worldwide	01/12/19	775169	345381	16295
2019	Lebanon floods	Lebanon	09/12/19	8415	5272	148
2019	Kuwait floods-19	Kuwait	15/12/19	25491	15566	312
2020	Dragon storms	Egypt	12/03/20	92014	49037	1479
2020	Aden floods	Yemen	21/04/20	37019	10638	147
2020	Oman floods	Oman	30/05/20	80673	25240	755
2020	Ta'if floods	Saudi Arabia	24/07/20	25424	13524	69
2020	Beirut explosion	Lebanon	04/08/20	307795	158427	7584
2020	Syria wildfires	Syria	03/09/20	22632	15162	167
2020	Sudan floods	Sudan	04/09/20	153126	96257	815
Total				1658795	812484	30146

Table 1: List of crises sorted by date, with tweets and users statistics

collected $1,658,795$ unique tweets from 22 emergency events. Apart from COVID-19, which was global, the crises occurred in eleven different countries[6] (see Table 1).

4 Data Description

4.1 Tweet-related & User-related Statistics

The Twitter search API supports search of tweets published in the previous seven days. As tweets matching different queries within the same timeframe might be captured on multiple occasions during our iterative collection process, we removed repetitive messages (ID-based duplicates) from the corpus and retained only tweets with unique IDs. Table 1 shows the number of unique tweets and unique authors for each crisis, along with the number of tweets published by verified accounts. In total, only $\sim 1.8\%$ of messages were sent by such accounts, indicating that few crisis-related tweets were generated by public interest accounts (e.g. media, government) which are typically verified. We found a strong Pearson correlation of 0.76 between tweets posted by verified users and those that included URLs. The Cairo bombing returned the largest percentages of both (9.02% and 24.43%, respectively). This suggests that many of the tweets related to this event were published by authentic news accounts rather than by the general public, who might have little to share about an instantaneous and focalised event of this kind. On average, only 7.9% of the corpus tweets include URLs. Overall, the corpus includes $40,175$ unique links, excluding links pointing to other posts in quote tweets.

4.2 Tweet Content Analysis

4.2.1 Content Redundancy

To explore the amount of duplicated content in the corpus, two tweets were considered duplicates if they exhibited a matching sequence of tokens (words or emojis). To identify duplicate content, we first cleaned the tweet text by removing RT, URL, user name, punctuation and special characters. This pre-processing also removed diacritics and elongation. This process revealed that more than half of the tweets in the corpus were duplicates, most of which were retweets. Other identical messages included shared news, emergency updates and instructions. We anticipated that this content was received and copied from different sources. We also expected that posts expressing emotional support would include similar common prayers and condolence phrases. In addition, we noticed that many nearly identical tweets were spam that included similar text (tokens), with shared shortened links referring to the same URL or to URLs with similar content. Spammers habitually exploit

[6]Dragon storms have affected several countries, but we focused our collection on the Egyptian Twitter content.

crisis name	# new messages	# retweets	# messages with unique text	% duplicates
Jordan floods	2379	6114	2383	71.94%
Kuwait floods-18	5504	28811	6139	82.11%
Qurayyat floods	903	8828	885	90.91%
Hafr Albatin floods-18	734	5335	786	87.05%
Leeth floods	1898	7698	1945	79.73%
Khartoum massacre	974	11331	1296	89.47%
Cairo bombing	747	1271	711	64.77%
Lebanon wildfires	1353	7232	3122	63.63%
Egypt floods	2207	8731	2384	78.20%
Hafr Albatin floods-19	1475	13071	2023	86.09%
Karbala massacre	2147	9814	1880	84.28%
Dubai floods	416	2064	383	84.56%
Coronavirus disease	189697	585472	250980	67.62%
Lebanon floods	2899	5516	3275	61.08%
Kuwait floods-19	5947	19544	6984	72.60%
Dragon storms	23125	68889	21815	76.29%
Aden floods	6640	30379	6274	83.05%
Oman floods	15843	64830	18224	77.41%
Ta'if floods	3910	21514	4612	81.86%
Beirut explosion	54956	252839	63408	79.40%
Syria wildfires	6459	16173	6160	72.78%
Sudan floods	45702	107424	23577	84.60%

Table 2: Kawārith content redundancy statistics

trending hashtags to advertise and spread malicious content. A duplicate could be a new message. Non-duplicates are messages with unique text, whether new or retweeted. Table 2 shows the percentage of duplicates in Kawārith by crisis, along with the number of new messages and retweets.

4.2.2 Prevalent Topics

To inspect keywords and prevalent topics in the corpus at event level, we employed word cloud and probabilistic topic models. Prior to topic identification, we performed two main steps. First, we removed noise by eliminating URLs, user names, punctuation, emojis and stop words. We also omitted hashtags from the vocabulary, as these were used as query terms and therefore occurred with greater frequency. The second step involved four types of letter normalisation: different forms of alef أ، إ، آ were normalised to ا, alef maqsora ى to ya ي, wāw mahmoza ؤ to wāw و and ta marbouta ة to ha ه. Stop words were removed from the vocabulary because of their high frequency of occurrence without adding meaningful content to the domain in question. For this purpose, we employed Arabic stop words from NLTK toolkit (Bird, 2006) and Alrefaie's repository[7], which contain 243 and 750 such words from MSA and classical Arabic, respectively. We found that many of the dialectal stop words in our corpus are not used in MSA, as Arabic-speaking people also tweet in their own dialects, and to the best of our knowledge there is no available domain-independent multi-dialect Arabic stop word list. To identify such words, samples of tweets were collected from the countries in our list and the dialectal stop words were manually identified from the most frequent words in each sample.

Using the Mo3jam dictionary[8], we added synonyms in other dialects, taking account of spelling variations. For example, the word لِسّع (lsʕ[9]) "not yet", which blends لساعته "to this moment" or حتى هذه الساعة "until this moment" (Aldrsoni, 2012), also takes the form لسّاته (lsAth). Arabic speakers tend to adopt a phonological system of spelling when writing non-MSA words, and the former could also be written as لسّه (lsh), لسّا (lsA) or لسّى (lsý). We also included common misspellings of frequently occurring words such as أصلن (ÂSln) "ever" for the word أصلاً (ÂSlA). It is important to note that as we disregarded diacritics, homographic stop words that share spellings with commonly used non-stop words were discarded. For instance, to avoid filtering out the word دُوَل which translates as 'countries', we removed the word دُوْل (dwl) which means 'these/those' in Egyptian and Higazi dialects. Our final list contained 405 multi-dialect Arabic stop words. Adding these words to the NLTK and Alrefaie's

[7] https://github.com/mohataher/arabic-stop-words

[8] https://en.mo3jam.com/

[9] We used Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007)

lists, 1177 (1098 after letter normalisation) words were identified as stop words to be filtered out before identifying keywords and underlying topics.

Most frequent words

To gain insights of the most frequent unigrams and bigrams, we used word clouds to visualise the text corpus for each crisis. Figure 1 shows word clouds of the top 200 words associated with 6 crises. In general, the diagrams show that the most frequently occurring terms are location names. For many events, terms related to emotional support and prayers show a high rate of occurrence. A closer look reveals that most of the top 200 terms are hazard descriptors; for example, common terms for flood events include الدفاع المدني "civil defense", الأمطار "rains", البنية التحتية "infrastructure", as well as victim names and many weather-related terms. In the case of COVID-19, prevalent terms include وزارة الصحة "Ministry of Health", الإجراءات الاحترازية "prevention measures", إصابة جديدة "new case" and الصحة العالمية "WHO". Human-induced crises share many common event-independent terms such as دماء "bloods" ,مستشفيات "hospitals" and مصابين "injured people", along with crisis-specific terms. Unlike other crises, events like the Khartoum massacre generated words related to internet blockage. Following duplicate filtration, visualisation revealed further hazard-related words, confirming that a single topic may dominate the dataset. Retweeted messages or content duplicates are not necessarily relevant to the crisis, as spam messages associated with crisis-related hashtags sometimes attract a large number of shares, and word clouds may include irrelevant terms (e.g. advertisements). For instance, phrases about invoking blessings upon the prophet Mohammed populate the Dragon storm diagram because the event occurred on Friday. This confirms the importance of removing duplicates and irrelevant posts from crisis data in pursuit of meaningful insights. Figure 2 shows words frequently associated with the Khartoum massacre and Dragon storms before and after duplicate removal. Observation suggests that one crisis can be discussed using data from another; for example #لبنان_ينتفض (which relates to the Lebanese protests) is the second most frequent hashtag in the Lebanese floods data. For that reason, it is useful to identify messages in terms of crisis type following data collection.

Figure 1: Word clouds showing the top 200 words from Kawārith for 6 selected crises

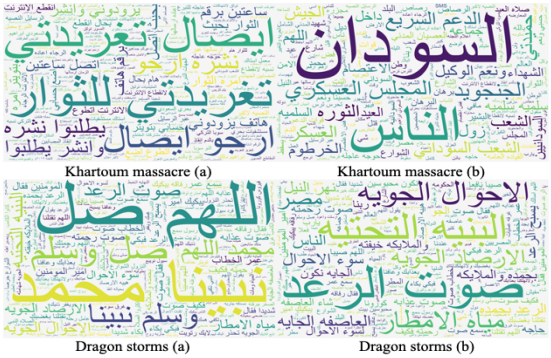

Figure 2: Word clouds for two crises: (a) for all data and (b) after duplicate removal

LDA Topics and Content Categorisation

We investigated the main topics discussed during crises using the Latent Dirichlet Allocation (LDA) modelling technique (Blei, 2012). LDA is a probabilistic topic model commonly used to disclose hidden topics in unstructured data. In this study, we used c_v coherence measure as an indicator to identify the optimum number of topics per event. To assign interpretations to topics, we manually analysed examples of tweets containing the top topic words to explore their meanings in context. Use of the selected dialectal stop words improved c_v scores by 37% on average. Analysis of example tweets yielded the following findings.

- Most of the broad information types used to coordinate response efforts as reported in previous work (Olteanu et al., 2015; Imran et al., 2016) appear in our corpus, varying in frequency across events. For covid19, in-domain fine-grained

topics were identified, including disease spread, COVID-19 symptoms, treatment, volunteering, prevention measures, cautions, other relevant events, prayers, opinions and personal messages.
- Many of the relevant tweets represent public opinion. Other on-topic tweets describe relevant events and consequences such as authority resignations.
- A tweet can communicate different information types; for example, the message below includes warning updates about affected individuals and weather conditions:

> #الاردن #هاجل. وفاة طفلة بسبب السيول الجارفة اليوم الجمعة وفقدان عدد من الاشخاص في عدة مناطق. وتشهد الاردن ومناطق واسعة من #السعودية موجة من الأمطار الغزيرة وتشكلا للسيول. ويتوقع اشتداد الحالة الجوية هذا المساء وغدا السبت. #سيول_الاردن #وسم #غدق #طقس
>
> "#Jordan #Breaking. A child has died in flash flooding this Friday and several people are missing in many areas. Jordan and wide regions of Saudi are witnessing heavy rainfall leading to floods. Severe weather conditions are expected this evening and tomorrow. #Jordan_flood #wasm #Ghadag #weather"

In this study, we employed a multi-label annotation scheme to categorise messages as different information types based on manual interpretation of prevalent topics and in light of earlier taxonomies. Unlike previous work taxonomies, we did not consider 'other useful information' class since usefulness is a subjective concept depending on the person who receives the information. Instead, we introduced 'other crisis updates' as a *"catchall"* category for other information that vary across crises such as flood level and wind. We observed that such updates were usually mentioned as *caution*. Hence, we merged these two categories together. As a few messages related to donation and volunteering in most cases, we merged this category with affected individuals, as these were usually mentioned together. We also tagged opinions, supplications and prayers. These may not be of use to humanitarian responders or contribute to situational awareness but can be used for other purposes, including opinion mining and measurement of event impact. Table 3 shows some example tweets from each category. In the case of COVID-19, a tweet was classified as either relevant or irrelevant to the event[10]. The following section describes the manual labelling process.

[10]To comply with Twitter's policies, we explicitly avoided coding data about users' health.

5 Manual Annotation and Inter-rater Reliability

Parts of our corpus were manually labelled to facilitate automatic identification of information categories by machine learning algorithms and to assess the frequency of the different message types posted on Twitter during emergencies. Seven crises were selected for annotation: the Jordan floods, Kuwait floods-18, Hafr Albatin floods-19, the Cairo bombing, the Dragon storms, the Beirut explosion and COVID-19. We focused on flood events as frequent occurrences in the Middle East.

To construct the labelled dataset, we considered only tweets with unique texts. We excluded duplicate messages as identified in section 4.2.1 to avoid labelling messages with the same content. We did not consider propagating labels to duplicate tweets after labelling the unique messages to avoid experimental bias in classification. Including duplicates in the dataset results in an overestimated performance if there is an overlap between test and training data (Alam et al., 2020). We also removed tweets containing less than 4 tokens as these are too short to convey any meaningful message. When calculating a tweet's length, we split the hashtags. As noted earlier, user mentions and URLs were not considered proper tokens. Each hyperlink was replaced with the Arabic word رابط to inform coders of a link referring to a website, image or video. Annotators were not required to visit the hyperlinks, as tweets were judged only on their text content. We sampled a different number of tweets for annotation from each event, ensuring that samples were taken from different timeframes. About 70–85% of flood events data were considered, along with all unique examples from the Cairo bombing (which contains only 711 distinct tweets). In the case of Dragon storms and Beirut explosion, about 1050 posts were sampled from each crisis. Regarding COVID-19, we considered 2005 tweets.

The data was annotated by volunteers. All annotators were native Arabic speakers. Each example was judged by two annotators, who were provided with annotation instructions and a news or Wikipedia article summarising the crisis. To begin, coders were trained using a short quiz with examples from each category and explanations of the correct answers. To further ensure reliability, annotators were tested on 30 examples from one event, and only those scoring 70% were allowed to proceed. The judgments were provided by 21

Label	Tweet
Affected individuals & help	عشرة عناصر من اطفاء بيروت مفقودين. "Ten members of Beirut firefighters are missing." قسم الطوارئ في مستشفى اوتيل ديو يستغيث ويطلب للتبرع بالدم. "The Emergency Department at the Hôtel-Dieu hospital calls for help and appeals for blood donations. "
Infrastructure & utilities damage	قطع المياه عن محافظة القاهرة بالكامل لسوء الأحوال الجوية - بوابة الشروق. "Water is cut off in Cairo governorate due to bad weather - Al Shorouk Gate."
Caution, preparations & other crisis updates	الارصاد تحذر امطار وسيول وانخفاض في درجات الحرارة ورمال واتربة. وتعطيل للدراسة الخميس بسبب الاحوال الجوية. "The Meteorological Department warns of rains, floods, temperature drop, sand and dust. And schools are closing on Thursday due to weather conditions."
Emotional support, prayers & supplications	اللهم احفظ مصر وأهلها. "May Allah save Egypt and its people."
Opinions & criticism	الاستقالات لا تكفي، وانما المطلوب محاسبة ومحاكمة كل مسئول مهمل. "Resignations are not enough, what is required is accountability and trial for every negligent official. "
Irrelevant	في أعماقنا رعد وبرق وعواصف وأمطار لا تشير إليها الأرصاد الجوية. "Deep inside us are thunder, lightning, storms and rain that have never been detected by weather forecast."

Table 3: Labels with example tweets

Label	Jordan floods	Kuwait floods-18	Hafr Albatin floods-19	Cairo bombing	Dragon storms	Beirut explosion
Affected individuals & help	331	414	83	138	70	186
Infrastructure & utilities damage	39	271	100	17	105	64
Caution, preparations & other crisis updates	268	980	475	214	252	170
Emotional support, prayers & supplications	709	816	202	222	120	277
Opinions & criticism	604	1355	189	181	221	198
Irrelevant	118	399	637	6	309	177
Total number of labelled examples	2000	4100	1615	706	1010	1010

Table 4: Distribution of labels: Flood crises, Cairo bombing, Dragon storms and Beirut explosion

Label	COVID-19
Relevant	1782
Irrelevant	223
Total number of labelled examples	2005

Table 5: Distribution of labels: COVID-19

trusted coders. (The annotation instructions have been included with the published datasets.) Average inter-rater agreement for the seven events as measured by Krippendorff's alpha was about 0.7, indicating substantial agreement and clear instructions. If two annotators disagreed, the message was judged by a third person; if coder 3 did not agree with coder 1 or 2, majority voting was applied to select the label agreed by at least two coders. A tweet was discarded if all three annotators completely disagree with each other. In total, we obtained $12,446$ labelled examples. Tables 4 and 5 show the distribution of categories and the total number of labelled tweets for the seven events. The dataset is imbalanced, and the distribution of information types varies across events. On average, only 4.4% of dataset instances have more than one label. Most relevant messages conveyed emotional support, opinions, cautions and crisis updates. Among COVID-19 tweets, we observed that the largest category of relevant messages relates to disease spread. The non-negligible percentage of irrelevant tweets (15% of the dataset) highlights the need for a classification step following data collection to filter out irrelevant posts.

6 Tweets Classification

To benchmark the dataset, we fine-tuned the Arabic Bidirectional Encoder Representation from Transformer (AraBERT) base model (Antoun et al., 2020). For reproducibility, we split the data into stratified train and test sets (80% and 20%, respectively). We fine-tuned a BERT model for each event using its training data by adding a linear classification layer on top of the BERT model, preceded by a dropout layer of a probability (0.2) to prevent the model from over-fitting. The learning rate of Adam optimizer was set to $5e$-5 and the loss function to binary cross-entropy. We set the maximum sequence length to 60 tokens as the longest tweet in Kawārith has 60 words. All models were trained for 5 epochs, which was empirically chosen. We experimented with batch sizes of 8 and 32. Each experiment was repeated three times with random seeds of $\{a, b, c\}$ because different seeds can yield considerably different results (Dodge et al., 2020).

Event	Acc.		Mac. f1		Mic. f1		HL	
	batch=8	batch=32	batch=8	batch=32	batch=8	batch=32	batch=8	batch=32
Jordan floods	82.8	85.8	72	75.6	83.65	86.59	0.055	0.045
Kuwait floods-18	81.38	81.99	79.51	80.3	82.3	83	0.06	0.057
Hafr Albatin floods-19	82.7	82.93	79.52	79.55	83.38	83.92	0.057	0.054
Cairo bombing	85.3	83.84	70.42	57.33	86.74	84.98	0.048	0.053
Dragon storms	78.7	78.81	77.33	78.73	79.84	81.01	0.071	0.066
Beirut explosion	76.37	76.17	72.91	73.45	76.77	78.06	0.079	0.073
COVID-19	94.1	93.5	82.95	81.9	-	-	-	-

Table 6: The accuracy, f1 scores and Hamming loss of the AraBERT model on the test set for each crisis

Text was pre-processed by removing noise as described in section 4.2.2. Evaluation was based on accuracy, known as Hamming score for the multi-label setting (Godbole and Sarawagi, 2004), macro f1, micro f1 and Hamming loss (HL) measures. The HL captures the fraction of labels that are incorrectly predicted. The binary COVID-19 data was evaluated using the accuracy and macro f1. The average scores of the three runs for each experiment are displayed in Table 6.

Results show that BERT achieved good accuracy and micro-f1 scores despite the relatively small training data used to fine-tune the models. Obtaining low HL scores indicates that a few labels were incorrectly predicted. While achieving a high macro-f1 is challenging with a skewed class distribution, BERT achieved macro-f1 scores higher than 70% in almost all cases without handling the data imbalance problem. Investigating different data augmentation techniques to improve the models' performance is left for future work. Generally, we found that fine-tuning in batches of 32 yields small gains in performance. In the case of the Cairo bombing, batch size of 8 achieved better results in all random runs, which we relate to the small number of training data (only 565 examples). When repeating the experiments for the other events using less than (randomly selected) 600 training examples, we found that it is more effective to fine-tune the models in batches of 8 instead of 32, which resulted in improvements in most cases.

Looking at the classification errors of the best models, we found that the models' mispredictions are not associated with a specific information category, and error types vary across events. We observed that many of the irrelevant tweets (especially from flood crises) were mistakenly classified as opinions. Most of these misclassified irrelevant tweets are ambiguous messages that express negative opinions about some topics that could be related to the crisis in some way. It is important to note that the two main annotators disagreed on most of such ambiguous posts. For many events, the models confused some messages from 'opinions & criticism' class with other classes. Classifying COVID-19 data obtained high accuracy and f1 scores. Performing such binary classification to filter out spam and other irrelevant posts prior to message categorisation could alleviate errors of classifying them as opinions. For future work, we will explore the BERT's performance when classifying cross-event data.

7 Conclusion

This paper introduced Kawārith, an Arabic Twitter corpus for 22 crises. We also reported a preliminary analysis of tweet content and provided a gold-standard multi-label dataset comprising $12k$ unique tweets. We believe this corpus can be leveraged for several tasks, including crisis detection and crisis type classification. Assigning messages to categories in order to identify informative posts can enhance situational awareness and assist emergency responders in organising effective relief efforts. The labelled dataset can also be utilised to gauge public opinion and sentiment during crises.

References

Aseel Addawood. 2020. Coronavirus: Public arabic twitter dataset.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018a. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*.

Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. 2018b. A twitter tale of three hurricanes: Harvey, irma, and maria. *arXiv preprint arXiv:1805.05144*.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2020. Standardizing and benchmarking crisis-related social media datasets for humanitarian information processing. *arXiv preprint arXiv:2004.06774*.

Sulaiman Aldrsoni. 2012. معجم اللهجات المحكية *[Dictionary of Spoken Dialects]*. King Fahd National Library, Riyadh, KSA.

Alaa Alharbi and Mark Lee. 2019. Crisis detection from Arabic tweets. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 72–79, Cardiff, United Kingdom. Association for Computational Linguistics.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Thayer Alshaabi, Michael V Arnold, Joshua R Minot, Jane Lydia Adams, David Rushing Dewhurst, Andrew J Reagan, Roby Muhamad, Christopher M Danforth, and Peter Sheridan Dodds. 2021. How the world's collective attention is being paid to a pandemic: Covid-19 related n-gram time series for 24 languages on twitter. *Plos one*, 16(1):e0244476.

Thayer Alshaabi, David R Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. 2020. The growing echo chamber of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *arXiv preprint arXiv:2003.03667*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688*.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

Alfredo Cobo, Denis Parra, and Jaime Navón. 2015. Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1189–1194.

Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell'Orletta. 2015. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1195–1200.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 22–30. Springer.

Raj Kumar Gupta, Ajay Vishwanath, and Yinping Yang. 2020. Covid-19 twitter dataset with latent topics, sentiments and emotions attributes. *arXiv preprint arXiv:2007.06954*.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On arabic transliteration. In *Arabic computational morphology*, pages 15–22. Springer.

Btool Hamoui, Mourad Mars, and Khaled Almotairi. 2020. FloDusTA: Saudi tweets dataset for flood, dust storm, and traffic accident events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1391–1396, Marseille, France. European Language Resources Association.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. *arXiv preprint arXiv:2004.05861*.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013a. Extracting information nuggets from disaster-related messages in social media. In *Iscram*.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013b. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. 2020. A three-level classification of french tweets in ecological crises. *Information Processing & Management*, 57(5):102284.

Rabindra Lamsal. 2020. Corona virus (covid-19) tweets dataset. ieee dataport.

Junhua Liu, Trisha Singhal, Lucienne Blessing, Kristin L Wood, and Kwan Hui Lim. 2020. Epic: An epidemics corpus of over 20 million relevant tweets. *arXiv preprint arXiv:2006.08369*.

Richard McCreadie, Cody Buntain, and Ian Soboroff. 2020. Incident streams 2019: Actionable insights and how to find them. In *Proceedings of the International ISCRAM Conference*.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Eighth international AAAI conference on weblogs and social media*.

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

Mark Edward Phillips. 2017. Hurricane harvey twitter dataset.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15.

Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.

Haoyu Wang, Eduard Hovy, and Mark Dredze. 2015. The hurricane sandy twitter corpus. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Kiran Zahra, Muhammad Imran, and Frank O Ostermann. 2020. Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, 57(1):102107.

Arkaitz Zubiaga. 2018. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984.