HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks

Firoj Alam, Umair Qazi, Muhammad Imran, Ferda Ofli
Qatar Computing Research Institute, HBKU, Qatar
{fialam,uqazi,mimran,fofli}@hbku.edu.qa

Abstract

Social networks are widely used for information consumption and dissemination, especially during time-critical events such as natural disasters. Despite its significantly large volume, social media content is often too noisy for direct use in any application. Therefore, it is important to filter, categorize, and concisely summarize the available content to facilitate effective consumption and decision-making. To address such issues automatic classification systems have been developed using supervised modeling approaches, thanks to the earlier efforts on creating labeled datasets. However, existing datasets are limited in different aspects (e.g., size, contains duplicates) and less suitable to support more advanced and data-hungry deep learning models. In this paper, we present a new large-scale dataset with ∼77K human-labeled tweets, sampled from a pool of ∼24 million tweets across 19 disaster events that happened between 2016 and 2019. Moreover, we propose a data collection and sampling pipeline, which is important for social media data sampling for human annotation. We report multiclass classification results using classic and deep learning (fastText and transformer) based models to set the ground for future studies. The dataset and associated resources are publicly available at `https://crisisnlp.qcri.org/humaid_dataset.html`.

1 Introduction

Recent studies highlight the importance of analyzing social media data during disaster events (Imran et al. 2015; Alam et al. 2020) as it helps decision-makers to plan relief operations. However, most of the actionable information on social media is available in the early hours of a disaster when information from other traditional data sources is not available. However, utilizing this information requires time-critical analysis of social media streams for aiding humanitarian organizations, government agencies, and public administration authorities to make timely decisions and to launch relief efforts during emergency situations (Starbird et al. 2010; Vieweg et al. 2010; Alam et al. 2021a). Among various social media platforms, Twitter has been widely used, on one hand, to disseminate information, and on the other, to collect, filter, and summarize information (Alam, Ofli, and Imran 2019). As the volume of information on social media is extremely high (Castillo 2016), automated data processing is necessary to filter redundant and irrelevant content and categorize useful content. There are many challenges to dealing with such large data streams and extracting useful information. Those include parsing unstructured and brief content, filtering out irrelevant and noisy content, handling information overload, among others.

Typical approaches tackling this problem rely on supervised machine learning techniques, i.e., classify each incoming tweet into one or more of a pre-defined set of classes. In the past, several datasets for disaster-related tweets classification were published (Olteanu et al. 2014; Imran, Mitra, and Castillo 2016; Alam, Ofli, and Imran 2018). These resources have supported NLP community to advance research and development in the *crisis informatics*[1] domain in many ways (Purohit and Sheth 2013; Burel and Alani 2018; Imran et al. 2014; Kumar et al. 2011; Okolloh 2009; Alam, Imran, and Ofli 2019). Deep neural networks have shown SOTA performance in many NLP tasks and application areas. However, deep learning algorithms are usually data-hungry, whereas the existing datasets in the crisis informatics domain are limited in different respects, which restricts the development of more sophisticated deep learning models.

We have so far investigated the existing datasets to understand their limitations for future research. These limitations can be summarized as follows. The existing datasets cover small-scale events. They contain exact-or-near duplicate tweets (e.g., CrisisLexT26 (Olteanu et al. 2014)),[2] which affects robustness of the trained models. They are usually dominated by tweets that come from outside of disaster hit areas and are usually about prayers and thoughts. We have also examined the existing literature to identify which categories are important for humanitarian organizations to extract actionable information and facilitate response efforts (Nemeskey and Kornai 2018; Kropczynski et al. 2018; Strassel, Bies, and Tracey 2017). Such an analysis and understanding has motivated us to develop a new, large-scale dataset that can take crisis informatics research

[1] `https://en.wikipedia.org/wiki/Disaster_informatics`

[2] Existing datasets contains such duplicates for different reasons: retweeted tweet, and same tweet collected in different data collection.

to the next level by affording the ability to develop more sophisticated models.

Hence, in this paper, we present the largest publicly available human annotated Twitter dataset, called **HumAID**: **Hum**an-**A**nnotated Disaster **I**ncidents **D**ata, for crisis informatics research. It has the following characteristics. *(i)* The dataset contains over 77,000 labeled tweets, which were sampled from 24 million tweets collected during 19 major real-world disasters that took place between 2016 and 2019, including hurricanes, earthquakes, wildfires, and floods. *(ii)* HumAID encompasses different disaster types across different time frames and locations. *(iii)* The dataset is more balanced in terms of disaster types and more consistent in terms of label agreement with regards to the existing datasets. *(iv)* Thanks to our carefully designed data filtering and sampling pipeline, HumAID consists of tweets that are more likely to be from the disaster-hit areas, and hence, contain more useful information coming from eyewitnesses or affected individuals. *(v)* Our annotation scheme consists of 11 categories representing critical information needs of a number of humanitarian organizations, including United Nations OCHA's needs reported in the MIRA framework[3] and previous studies (Vieweg, Castillo, and Imran 2014; Olteanu et al. 2014; Imran, Mitra, and Castillo 2016; Alam, Ofli, and Imran 2018; Mccreadie, Buntain, and Soboroff 2019). *(vi)* Finally, HumAID is the largest dataset in comparison to the existing datasets in the crisis informatics domain.

Our focus was developing a large-scale human-labeled English tweets dataset covering several categories useful for humanitarian organizations during *natural disasters*. To obtain such a large-scale dataset we used Amazon Mechanical Turk[4] for the annotation. Furthermore, we used the labeled tweets to obtain benchmark results with both classical (i.e., SVM and RF) and deep learning algorithms (i.e., fastText and transformer based models). Our extensive experiments show that deep learning models outperform traditional supervised learning algorithms. Last but not least, we share dataset, and data splits with the research community for both reproducibility and further enhancements.

The rest of the paper is organized as follows. Section 2 provides a brief overview of previous work. Section 3 describes our data collection approaches, and Section 4 provides annotation procedures. We report experimental results in Section 5 and discuss future research directions in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

Over the past few years, there has been a major research effort on analyzing social media content (mainly Twitter and Facebook) for humanitarian aid. Key challenges addressed in these studies include data filtering, classification, information extraction, and summarization to enhance situational awareness and mine actionable information (Sakaki, Okazaki, and Matsuo 2010; Imran et al. 2014; Saravanou et al. 2015; Tsou et al. 2017; Martinez-Rojas, del Carmen Pardo-Ferreira, and Rubio-Romero 2018). Most of these studies have been possible thanks to the publicly available datasets.

2.1 Existing Datasets

Below we provide a brief overview of the existing datasets.

CrisisLex comprises two datasets, i.e., CrisisLexT26 and CrisisLexT6 (Olteanu et al. 2014). The CrisisLexT26 dataset consists of ∼28,000 labeled tweets from 26 different disaster events that took place in 2012 and 2013. It includes disaster type and sub-type, and coarse- and fine-grained humanitarian class labels. CrisisLexT6 contains ∼60,000 labeled tweets from six disaster events that occurred between October 2012 and July 2013. Annotation of CrisisLexT6 includes *related* vs. *not-related*.

CrisisMMD is a multimodal and multitask dataset consisting of ∼18,000 labeled tweets and associated images (Alam, Ofli, and Imran 2018; Ofli, Alam, and Imran 2020). Tweets have been collected from seven natural disaster events that took place in 2017. The annotations include three tasks: (i) *informative* vs. *not-informative*, (ii) humanitarian categories (eight classes), and (iii) damage severity levels (three classes). The third annotation task, i.e., damage severity (mild, severe and none), was applied only on images.

CrisisNLP consists of ∼50,000 human-labeled tweets collected from 19 different disasters that happened between 2013 and 2015, and annotated according to different schemes including classes from humanitarian disaster response and some classes pertaining to health emergencies (Imran, Mitra, and Castillo 2016).

Disaster Response Data contains 30,000 tweets with 36 different categories, collected during disasters such as an earthquake in Haiti in 2010, an earthquake in Chile in 2010, floods in Pakistan in 2010, Hurricane Sandy in USA in 2012, and news articles.[5]

Disasters on Social Media dataset comprises 10,000 tweets annotated with labels *related* vs. *not-related* to the disasters.[6]

SWDM2013 consists of two data collections. The Joplin collection contains 4,400 labeled tweets collected during the tornado that struck Joplin, Missouri on May 22, 2011. The Sandy collection contains 2,000 labeled tweets collected during Hurricane Sandy, that hit Northeastern US on Oct 29, 2012 (Imran et al. 2013).

Eyewitness Messages dataset contains ∼14,000 tweets with labels (i) direct-eyewitness, (ii) indirect-eyewitness, (iii) non-eyewitness, and (iv) don't know, for different event types such as flood, earthquake, fire, and hurricane (Zahra, Imran, and Ostermann 2020).

Arabic Tweet Corpus consists of tweets collected during four flood events that took place in different areas of the Arab world (i.e., Jordan, Kuwait, northern Saudi Arabia, and

[3] https://www.humanitarianresponse.info/en/programme-cycle/space/document/mira-framework

[4] https://www.mturk.com/

[5] https://www.figure-eight.com/dataset/combined-disaster-response-data/

[6] https://data.world/crowdflower/disasters-on-social-media

western Saudi Arabia) in 2018 (Alharbi and Lee 2019). The dataset contains 4,037 labeled tweets with their relevance and information type.

TREC Incident Streams dataset has been developed as part of the TREC-IS 2018 evaluation challenge and consists of 19,784 tweets labeled for actionable information identification and assessing the criticality of the information (Mccreadie, Buntain, and Soboroff 2019). This dataset is developed based on CrisisLex, CrisisNLP and the data collected using Gnip services.

Disaster Tweet Corpus 2020 is a compilation of existing datasets for *disaster event types* classification (Wiegmann et al. 2020).

2.2 Modeling Approaches

For disaster response, typical social media content classification task include *(i)* informative vs non-informative, (also referred as related *vs.* not-related, or on-topic *vs.* off-topic), *(ii)* fine-grained humanitarian information categories, *(iii)* disaster event types, *(iv)* damage severity assessment. To address such tasks classical algorithms have been widely used in developing classifiers in the past (Imran et al. 2015). However, deep learning algorithms have recently started receiving more attention due to their successful applications in various natural language processing (NLP) and computer vision tasks. For instance, (Nguyen et al. 2017) and (Neppalli, Caragea, and Caragea 2018) perform comparative experiments between different classical and deep learning algorithms including Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). Their experimental results suggest that CNN outperforms other algorithms. Though in another study, (Burel and Alani 2018) reports that SVM and CNN can provide very competitive results in some cases. CNNs have also been explored in event type-specific filtering model (Kersten et al. 2019). Recent successful embedding representations such as Embeddings from Language Models (ELMo) (Peters et al. 2018), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019), and XLNet (Yang et al. 2019) have also been explored for disaster related tweet classification tasks (Jain, Ross, and Schoen-Phelan 2019; Wiegmann et al. 2020; Alam et al. 2021b). From a modeling perspective, our work is different than previous work in a way that we have used both classical and deep learning algorithms with different transformer-based models, which can serve as a strong baseline for future study.

3 Data Collection, Filtering and Sampling

We used the AIDR system (Imran et al. 2014) to collect data from Twitter during 19 disaster events occurred between 2016 and 2019. AIDR is a publicly available system, which uses the Twitter streaming API for data collection.[7] The data collection was performed using event-specific keywords and hashtags. In Table 1, we list details of the data collection period for each event. In total, 24,859,155 tweets were collected from all the events. Data annotation is a costly procedure, therefore, we investigated how to filter and sample data that can maximize the quality of the labeled data.

3.1 Data Filtering

To prepare data for manual annotation, we perform the following filtering steps:

1. **Date-based filtering:** For some events, that data collection period extends beyond the actual event dates. For instance, for Hurricane Florence, our data collection period is from Sep 11, 2018 to Nov 17, 2018 although the hurricane actually dissipated on Sep 18. For this purpose, we restrict the data sampling period to actual event days as reported on their Wikipedia page.
2. **Location-based filtering:** Since our data collection was based on event-specific keywords (i.e., we did not restrict the data collection to any geographical area), it is likely that a large portion of the collected data come from outside the disaster-hit areas. However, the most useful information for humanitarian organizations is the information that originates from eyewitnesses or people from the disaster-hit areas. Therefore, we discarded all tweets outside the disaster-hit areas by using a geographic filter. The geographic filter uses one of the three fields (i.e., *geo*, *place*, or *user location*) from a tweet. We prioritize the *geo* field, as it comes from the user device as GPS coordinates. If the *geo* field is not available, which is the case for 98% of the tweets, we use the *place* field. The *place* field comes with a geographical bounding box, which we use to determine whether a tweet is inside or outside an area. As our last option, we use the *user location* field, which is a free-form text provided by the user. Next, we use the Nominatim[8] service (an OpenStreetMap[9] database) to resolve the provided location text into city, state, and country information. The resolved geo-information is then used to filter out tweets which do not belong to a list of locations that we manually curate for each event. A list of target locations for each event is provided in the dataset bundle.
3. **Language-based filtering:** We choose to only annotate English tweets due to budget limitations. Therefore, we discard all non-English tweets using the Twitter provided language metadata for a given tweet. It would be interesting to annotate tweets in other languages in future studies.
4. **Classifier-based filtering:** After applying the filters mentioned above, the remaining data is still in the order of millions (i.e., $\sim$7 million according to Table 1), a large proportion of which might still be irrelevant. To that end, we trained a Random Forest classifier[10] using a set of humanitarian categories labeled data reported in (Alam, Imran, and Ofli 2019), which consists the classes similar to our annotation task described in the next section. We

[7] http://aidr.qcri.org/

[8] http://nominatim.org

[9] https://www.openstreetmap.org

[10] The choice of this algorithm was based on previous studies and its use in practical application (Imran et al. 2014). Also because it is computationally simple, which enabled us to classify a large number of tweets in a short amount of time.

Event name	Total	Date range	Date	Location	Language	Classifier	WC	De-dup.	Sampled
2016 Ecuador Earthquake	1,756,267	04/17 – 04/18	884,783	172,067	19,988	11,269	11,251	2,007	2,000
2016 Canada Wildfires	312,263	05/06 – 05/27	312,263	66,169	66,169	5,812	5,796	2,906	2,906
2016 Italy Earthquake	224,853	08/24 – 08/29	224,853	15,440	15,440	6,624	6,606	1,458	1,458
2016 Kaikoura Earthquake	318,256	09/01 – 11/22	318,256	44,791	44,791	11,854	11,823	3,180	3,180
2016 Hurricane Matthew	1,939,251	10/04 – 10/10	82,643	36,140	36,140	10,116	10,099	2,111	2,100
2017 Sri Lanka Floods	40,967	05/31 – 07/03	40,967	4,267	4,267	2,594	2,594	760	760
2017 Hurricane Harvey	6,384,625	08/25 – 09/01	2,919,679	1,481,939	1,462,934	638,611	632,814	97,034	13,229
2017 Hurricane Irma	1,641,844	09/06 – 09/17	1,266,245	563,899	552,575	113,757	113,115	29,100	13,400
2017 Hurricane Maria	2,521,810	09/16 – 10/02	1,089,333	541,051	511,745	202,225	200,987	17,085	10,600
2017 Mexico Earthquake	361,040	09/20 – 09/23	181,977	17,717	17,331	11,662	11,649	2,563	2,563
2018 Maryland Floods	42,088	05/28 – 06/07	42,088	20,483	20,483	7,787	7,759	1,155	1,140
2018 Greece Wildfires	180,179	07/24 – 08/18	180,179	9,278	9,278	4,896	4,888	1,815	1,815
2018 Kerala Floods	850,962	08/17 – 08/31	757,035	401,950	401,950	225,023	224,876	29,451	11,920
2018 Hurricane Florence	659,840	09/11 – 09/18	483,254	318,841	318,841	38,935	38,854	13,001	9,680
2018 California Wildfires	4,858,128	11/10 – 12/07	4,858,128	2,239,419	2,239,419	942,685	936,199	103,711	10,225
2019 Cyclone Idai	620,365	03/15 – 04/16	620,365	47,175	44,107	26,489	26,469	5,236	5,236
2019 Midwestern US Floods	174,167	03/25 – 04/03	174,167	96,416	96,416	19,072	19,037	3,427	3,420
2019 Hurricane Dorian	1,849,311	08/30 – 09/02	1,849,311	993,982	993,982	137,700	136,954	18,580	11,480
2019 Pakistan Earthquake	122,939	09/24 – 09/26	122,939	34,200	34,200	16,180	16,104	2,502	2,500

Table 1: Event-wise data distribution, filtering and sampling. WC: Word Count, De-dup.: De-duplication. Date range format is MM/DD and the year is specified in the Event name column.

follow widely used train/dev/test (70/10/20) splits to train and evaluate the model. We prepossessed the tweets before training the classifier, which include removing stop words, URLs, user mentions, and non-ASCII characters. The trained classifier achieved an F1=76.9%, which we used to classify and eliminate all the irrelevant tweets, i.e., tweets classified as not-humanitarian.

5. **Word-count-based filtering:** We retain tweets that contain at least three words or hashtags. The rationale behind such a choice is that tweets with more tokens tend to provide more information and likely to have additional contextual information useful for responders. URLs and numbers are usually discarded while training a classifier, thus we ignore them while counting the number of tokens for a given tweet.
6. **Near-duplicate filtering:** Finally, we apply de-duplication to remove exact and near-duplicate tweets using their textual content. This consists of three steps: *(i)* tokenize the tweets to remove URL, user-mentions, and other non-ASCII characters; *(ii)* convert the tweets into vectors of uni- and bi-grams with their frequency-based representations, *(iii)* compute the cosine similarity between tweets and flag the one as duplicate that exceed a threshold. Since threshold identification is a complex procedure, therefore, we follow the findings in (Alam et al. 2021b), where a threshold of 0.75 is used to flag duplicates.

3.2 Sampling

Although the filtering steps help reduce the total number of tweets significantly while maximizing the information theoretic value of the retained subset, there are still more tweets than our annotation budget. Therefore, in the sampling step, we select n random tweets from each class while also maintaining a fair distribution across classes. In Table 1, we summarize the details of the data filtering and sampling including total number of tweets initially collected as well as the total number of tweets retained after each filtering and sampling step for each event. In particular, the last column of the table indicates the total number of tweets sampled for annotation for each disaster event.

4 Manual Annotations

Since the main purpose of this work is to create a large-scale dataset that can be used to train models that understand the type of humanitarian aid-related information posted in a tweet during disasters, we first define what "humanitarian aid" means. For the annotation we opted and redefined the annotation guidelines discussed in (Alam, Ofli, and Imran 2018).

Humanitarian aid:[11] In response to humanitarian crises including natural and human-induced disasters, humanitarian aid involves assisting people who need help. The primary purpose of humanitarian aid is to save lives, reduce suffering, and rebuild affected communities. Among the people in need belong homeless, refugees, and victims of natural disasters, wars, and conflicts who need necessities like food, water, shelter, medical assistance, and damage-free critical infrastructure and utilities such as roads, bridges, power-lines, and communication poles.

Based on the *Humanitarian aid* definition above, we define each humanitarian information category below.[12] The annotation task was to assign one of the below labels to a tweet. Though multiple labels can be assigned to a tweet, however, we limited it to one category to reduce the annotation efforts.

L1: Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster;

[11] https://en.wikipedia.org/wiki/Humanitarian_aid

[12] Note that we also supplemented these definitions by showing example tweets in the instructions.

Event name	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	Total
2016 Ecuador Earthquake	30	3	70	555	10	23	18	81	91	394	319	1,594
2016 Canada Wildfires	106	380	251	4	-	79	13	311	20	934	161	2,259
2016 Italy Earthquake	10	3	54	174	7	9	10	52	30	312	579	1,240
2016 Kaikoura Earthquake	493	87	312	105	3	224	19	311	24	207	432	2,217
2016 Hurricane Matthew	36	38	178	224	-	76	5	328	53	326	395	1,659
2017 Sri Lanka Floods	28	9	17	46	4	20	2	56	34	319	40	575
2017 Hurricane Harvey	541	688	1,217	698	10	410	42	1,767	333	2,823	635	9,164
2017 Hurricane Irma	613	755	1,881	894	8	615	60	2,358	126	1,590	567	9,467
2017 Hurricane Maria	220	131	1,427	302	11	270	39	1,568	711	1,977	672	7,328
2017 Mexico Earthquake	35	4	167	254	14	38	3	109	61	984	367	2,036
2018 Maryland Floods	70	3	79	56	140	77	1	137	1	73	110	747
2018 Greece Wildfires	26	7	38	495	20	74	4	159	25	356	322	1,526
2018 Kerala Floods	139	56	296	363	7	456	65	955	590	4,294	835	8,056
2018 Hurricane Florence	1,310	637	320	297	-	1,060	95	636	54	1,478	472	6,359
2018 California Wildfires	139	368	422	1,946	179	1,318	68	1,038	79	1,415	472	7,444
2019 Cyclone Idai	89	57	354	433	19	80	11	407	143	1,869	482	3,944
2019 Midwestern U.S. Floods	79	8	140	14	1	389	27	273	46	788	165	1,930
2019 Hurricane Dorian	1,369	802	815	60	1	874	46	1,444	179	987	1,083	7,660
2019 Pakistan Earthquake	71	-	125	401	1	213	32	154	19	152	823	1,991
Total	**5,404**	**4,036**	**8,163**	**7,321**	**435**	**6,305**	**560**	**12,144**	**2,619**	**21,278**	**8,931**	**77,196**

Table 2: Distribution of annotations across events and class labels.

- **L2: Displaced people and evacuations:** People who have relocated due to the crisis, even for a short time (includes evacuations);
- **L3: Infrastructure and utility damage:** Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
- **L4: Injured or dead people:** Reports of injured or dead people due to the disaster;
- **L5: Missing or found people:** Reports of missing or found people due to the disaster event;
- **L6: Not humanitarian:** If the tweet does not convey humanitarian aid-related information;
- **L7: Don't know or can't judge:** If the tweet is irrelevant or cannot be judged due to non-English content.
- **L8: Other relevant information:** If the tweet does not belong to any of the above categories, but it still contains important information useful for humanitarian aid, belong to this category;
- **L9: Requests or urgent needs:** Reports of urgent needs or supplies such as food, water, clothing, money, medical supplies or blood;
- **L10: Rescue, volunteering, or donation effort:** Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, people in shelter facilities, donation of money, or services, etc.;
- **L11: Sympathy and support:** Tweets with prayers, thoughts, and emotional support;

For the manual annotation, we opted to use Amazon Mechanical Turk (AMT) platform. In crowdsourcing, one of the challenges is to find a large number of qualified workers while filtering out low-quality workers or spammers (Chowdhury et al. 2014). To tackle this problem, a typical approach is to use qualification tests followed by a gold standard evaluation (Chowdhury et al. 2015). We created a qualification test consisting of 10 tweets. To participate in the task, each annotator first needs to pass the qualification test. In order to pass the test, the annotator needs to correctly answer at least 6 out of 10 tweets. The gold standard evaluation is performed at the HIT (i.e., Human Intelligence Task) level. A HIT consists of 30 tweets and in each HIT there are 10 gold standard tweets (i.e., tweets with known labels) and 20 tweets with unknown labels. These 10 gold standard tweets are selected from a pool of tweets labeled by domain experts. Note that developing a gold standard dataset is another costly procedure in terms of time and money. Therefore, we first randomly selected the tweets from different events by focusing on disaster types such as hurricane, flood, fire and earthquake, and then domain experts manually labeled them.

The annotator who participates in the HIT needs to read each tweet and assign one of the above labels to complete the task. The participation and completion of a HIT by an annotator are referred to as an assignment. We set the assignment approval criterion to 70%, which means an assignment of the HIT will be automatically approved if the annotator correctly labels at least 7 out of 10 gold standard tweets.

For each HIT and the associated tweets, we wanted to have three judgments. As our HIT design consists of 20 tweets with unknown labels and we wanted to automatically approve the HIT, we set the HIT approval criterion to 66%. That is, a HIT is approved if the annotators agree on a label for at least 14 out of 20 tweets. In order to approve the label for a tweet, we also set a threshold of 66%, which means out of three annotators two of them have to agree on the same label. Since the social media content is highly noisy and categories can be subjective, we choose to use a minimum threshold of 66% for the agreement of the label for each tweet.

4.1 Crowdsourcing Results

In Table 1, the last column represents the number of tweets sampled for the annotation (i.e., 109,612 in total). Since in AMT our minimum threshold to accept a HIT was 66%, we can expect to acquire agreed labels for a minimum of 66% of the tweets. In Table 2, we present the annotated dataset, which consists of class label distribution for each event along with the total number of annotated tweets. In summary, we have ∼70% tweets with agreed labels, which results in 77,196 tweets. To compute the annotation agreement we considered the following evaluation measures.

1. Fleiss kappa: It is a reliability measure that is applicable for any fixed number of annotators annotating categorical labels to a fixed number of items, which can handle two or more categories and annotators (Fleiss, Levin, and Paik 2013). However, it can not handle missing labels, except for excluding them from the computation.
2. Average observed agreement: It is an average observed agreement over all pairs of annotators (Fleiss, Levin, and Paik 2013).
3. Majority agreement: We compute the majority at the tweet level and take the average. The reason behind this is that for many tweets the number of annotators vary between three and five, and hence, it is plausible to evaluate the agreement at the tweet level.
4. Krippendorff's alpha: It is a measure of agreement that allows two or more annotators and categories (Krippendorff 1970). Additionally, it handles missing labels.

For the first two methods, we selected three annotations whereas for the last two methods we considered all annotations (i.e., three to five). In Table 3, we present the annotation agreement for all events with different approaches mentioned above. The average agreement score varies 55% to 83%. Note that in Kappa measurement value of 0.41-0.60, 0.61-0.80, and 0.81-1 refers to the moderate, substantial, and perfect agreement, respectively (Landis and Koch 1977). Based on such measurements we conclude that our annotation agreement score leads to moderate to the substantial agreement. We also investigated the quality of the annotated labels and it suggests that tweet texts clearly demonstrate the labels, as can be seen in Table 4.

4.2 Lexical Analysis and Statistics

To understand the lexical content, we check the number of tokens for each tweet in each event. This information help in understanding the characteristics of the dataset. For example, the maximum number of tokens can help define max sequence length in deep learning-based architectures such as CNN. In Table 5, we provide results for each event. The minimum number of tokens is 3 for all the events, therefore, we have not reported that number in the table. From the table, we can observe that for some events, (e.g., Hurricane Maria, Maryland Floods), the max token limits are higher. This is because Twitter extended its character limit to 280 from September 2017. In Figure 1, we provide statistics of the tweet lengths in the overall dataset in different bins. The majority of the tweets is appearing with a range of length 10-20 tokens, second bin in the figure.

Event name	Fleiss (κ)	K-α	A/O	M/A
2016 Ecuador Earthquake	0.65	0.64	0.73	0.86
2016 Canada Wildfires	0.54	0.61	0.63	0.85
2016 Italy Earthquake	0.64	0.69	0.74	0.89
2016 Kaikoura Earthquake	0.57	0.57	0.63	0.81
2016 Hurricane Matthew	0.60	0.57	0.66	0.82
2017 Sri Lanka Floods	0.47	0.51	0.61	0.83
2017 Hurricane Harvey	0.55	0.57	0.62	0.82
2017 Hurricane Irma	0.55	0.53	0.63	0.80
2017 Hurricane Maria	0.54	0.53	0.62	0.80
2017 Mexico Earthquake	0.57	0.62	0.67	0.86
2018 Maryland Floods	0.52	0.56	0.58	0.81
2018 Greece Wildfires	0.63	0.65	0.70	0.86
2018 Kerala Floods	0.50	0.50	0.63	0.82
2018 Hurricane Florence	0.50	0.54	0.57	0.80
2018 California Wildfires	0.55	0.59	0.61	0.83
2019 Cyclone Idai	0.51	0.52	0.61	0.82
2019 Midwestern U.S. Floods	0.48	0.50	0.58	0.80
2019 Hurricane Dorian	0.59	0.55	0.65	0.81
2019 Pakistan Earthquake	0.55	0.61	0.65	0.85
Average	0.55	0.57	0.64	0.83

Table 3: Annotation agreement scores for different events. Metrics: Fleiss κ, Krippendorff alpha (K-α), Average observed agreement (A/O), Average majority agreement (M/A).

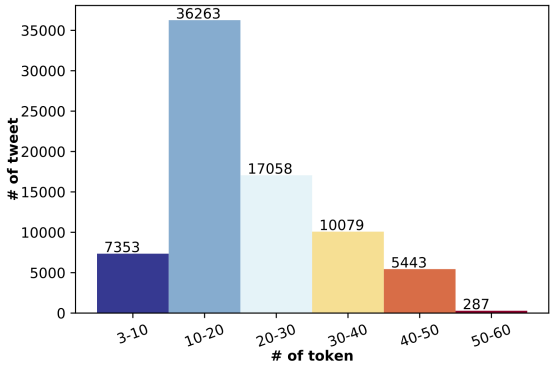

Figure 1: Number of tweet with different lengths in overall dataset.

5 Experiments and Results

In this section, we describe the details of our classification experiments and results. To run the experiments, we split data into training, development, and test sets with a proportion of 70%, 10%, and 20%, respectively. We removed low prevalent classes (i.e., number of tweets with a class label less than 15, e.g., in event *2016 Ecuador Earthquake*, there were only 3 tweets with the class label *Displaced people and evacuations*) in some events from the classification experiments. This approach reduced from 77196 to 76484 tweets for the experiments. Event-wise data split and class label distribution are reported in Table 6.

We ran the classification experiments at three levels: *(i)* event level, *(ii)* event-type level, and *(iii)* all data combined. The purpose of event and event type level experiments is to provide a baseline, which can be used to compare cross event experiments in future studies.

Tweet	Label
VLEs extending helping hands to provide relief material to affected people in Keralas Chenganoor, Alapuzha and Idukki districts. They distributed food, water, clothing, medicine etc to flood people as humanitarian work is still on. #KeralaFloodRelief #KeralaFloods2018 #OpMadad	L10
@narendramodi Sir, Chenganoor area is in very dangerous condition..We need more army assistance there..Please Please help.@PMOIndia #KeralaFloodRelief	L9
In this difficult period. My prayers for all flood affected people of Kerala. We know Kerala is most beautiful state of india and people of Kerala part of UAE success. Let us extend our hands in support in their difficulties. #KeralaFloods #Kerala_in_our_hearts_	L11

Table 4: Examples of annotated tweets.

Event name	Std.	Mean	Max
2016 Ecuador Earthquake	4.24	13.84	27
2016 Canada Wildfires	4.10	14.27	28
2016 Italy Earthquake	4.44	14.11	25
2016 Kaikoura Earthquake	5.01	15.39	28
2016 Hurricane Matthew	4.51	15.76	29
2017 Sri Lanka Floods	4.15	15.93	24
2017 Hurricane Harvey	4.70	15.45	31
2017 Hurricane Irma	4.89	15.47	29
2017 Hurricane Maria	4.95	15.96	51
2017 Mexico Earthquake	4.64	15.49	37
2018 Maryland Floods	11.06	22.75	51
2018 Greece Wildfires	11.73	23.06	54
2018 Kerala Floods	11.43	26.38	54
2018 Hurricane Florence	10.98	25.57	55
2018 California Wildfires	12.02	24.72	57
2019 Cyclone Idai	11.12	28.46	53
2019 Midwestern U.S. Floods	11.62	27.50	54
2019 Hurricane Dorian	12.14	25.73	57
2019 Pakistan Earthquake	11.71	22.80	54

Table 5: Descriptive statistics (i.e., std., max and mean number of token) for each event

For the data splits we first create splits for each event separately. Then, for the event-type experiments, we combine the training, development, and test sets of all the events that belong to the same event type. For example, we combine all training sets of specific earthquake collections into the general earthquake-type training set. Since combining data from multiple events can result in near-duplicate tweets[13] across different data splits (i.e., training, development, and test), we applied the same near-duplicate removal approach discussed earlier to eliminate such cases. With this approach, we removed only nine tweets, which leads to having a total of 76,475 tweets in event-type experiments. Similarly, the same duplicate removal approach was applied when combining data from all the events, which also reduced another 9 tweets, resulting 76,466 tweets.

[13]This happens when more than one crisis events occur at the same time and same tweets are collected for different events.

To measure the performance of each classifier, we use weighted average precision (P), recall (R), and F1-measure (F1). The choice of the weighted metric is to factor in the class imbalance problem.

5.1 Preprocessing

Tweet text consists of many symbols, emoticons, and invisible characters. Therefore, we preprocess them before using in model training and classification experiments. The preprocessing part includes removal of stop words, non-ASCII characters, punctuations (replaced with whitespace), numbers, URLs, and hashtag signs.

5.2 Models

For this study, we focus on multiclass classification experiments using both classical and deep learning algorithms discussed below. As for the classical models, we used the two most popular algorithms *(i)* Random Forest (RF) (Breiman 2001), and *(ii)* Support Vector Machines (SVM) (Platt 1998).

As deep learning algorithms, we used FastText (Joulin et al. 2017) and transformer-based models such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), XLM-RoBERTa (Conneau et al. 2019) and DistilBERT (Sanh et al. 2019). The reason to choose XLM-RoBERTa is that some tweets can have mix-language (e.g., English tweets with some French words) and we wanted to see how model performs given that it is a multilingual model.

5.3 Classification Experiments

To train the classifiers using the aforementioned *classical algorithms*, we converted the preprocessed tweets into bag-of-n-gram vectors weighted with logarithmic term frequencies (tf) multiplied with inverse document frequencies (idf). Since contextual information, such as n-grams, are useful for classification, we extracted unigram, bigram, and trigram features. For both SVM and RF we use grid search to optimize the parameters.

For FastText, we used pre-trained embeddings trained on Common Crawl[14] and default hyperparameter settings.

For transformer-based models, we use the Transformer Toolkit (Wolf et al. 2019) and fine-tune each model using the settings as described in (Devlin et al. 2019) with a task-specific layer on top of the transformer model. Due to the instability of the pre-trained models as reported by (Devlin et al. 2019), we do 10 runs of each experiment using different random seeds and choose the model that performs the best on the development set. For training the BERT model for each event, event-type and all combined dataset, we use a batch size of 32, learning rate of 2e-5, maximum sequence length 128, and fine tune 10 epochs with the 'categorical cross-entropy' as the loss function.

5.4 Results

In Table 7, we report the classification results (weighted F1) for each event, event type and combined dataset with all

[14]https://fasttext.cc/docs/en/crawl-vectors.html

Ev.	L1	L2	L3	L4	L5	L6	L8	L9	L10	L11
1	21/3/6	-	49/7/14	388/57/110	-	16/2/5	57/8/16	64/9/18	276/40/78	223/33/63
2	74/11/21	266/39/75	176/25/50	-	-	55/8/16	218/32/61	14/2/4	653/95/186	113/16/32
3	-	-	38/5/11	122/18/34	-	-	36/5/11	21/3/6	218/32/62	405/59/115
4	345/50/98	61/9/17	218/32/62	73/11/21	-	157/23/44	218/32/61	17/2/5	145/21/41	302/44/86
5	25/7/4	27/7/4	125/35/18	157/44/23	-	53/15/8	229/66/33	37/11/5	228/65/33	276/79/40
6	20/3/5	-	12/2/3	32/5/9	-	14/2/4	39/6/11	24/3/7	223/32/64	28/4/8
7	379/55/107	482/70/136	852/124/241	488/71/139	-	287/42/81	1237/180/350	233/34/66	1976/288/559	444/65/126
8	429/62/122	528/77/150	1317/192/372	626/91/177	-	430/63/122	1651/240/467	88/13/25	1113/162/315	397/58/112
9	154/22/44	92/13/26	999/145/283	211/31/60	-	189/28/53	1097/160/311	498/72/141	1384/202/391	470/69/133
10	24/4/7	-	117/17/33	178/26/50	-	27/4/7	76/11/22	43/6/12	688/100/196	257/37/73
11	49/7/14	-	55/8/16	39/6/11	98/14/28	54/8/15	96/14/27	-	51/7/15	77/11/22
12	18/3/5	-	27/4/7	346/50/99	14/2/4	52/8/14	111/16/32	18/2/5	249/36/71	225/33/64
13	97/14/28	39/6/11	207/30/59	254/37/72	-	319/47/90	669/97/189	413/60/117	3005/438/851	585/85/165
14	917/134/259	446/65/126	224/33/63	208/30/59	-	742/108/210	445/65/126	38/5/11	1034/151/293	330/48/94
15	97/14/28	258/38/72	295/43/84	1362/199/385	125/18/36	923/134/261	727/106/205	55/8/16	991/144/280	330/48/94
16	62/9/18	40/6/11	248/36/70	303/44/86	13/2/4	56/8/16	285/41/81	100/15/28	1308/191/370	338/49/95
17	55/8/16	-	98/14/28	-	-	272/40/77	191/28/54	32/5/9	552/80/156	116/16/33
18	958/140/271	561/82/159	571/83/161	42/6/12	-	612/89/173	1011/147/286	125/18/36	691/101/195	758/110/215
19	50/7/14	-	87/13/25	281/41/79	-	149/22/42	108/15/31	13/2/4	106/15/31	576/84/163

Table 6: Event-wise data split and distribution of class labels. First column (i.e., Ev) enumerates all 19 events in the same ordered as in Table 1. The numbers in each cell represent train/dev/test for each class label and event.

models. The column *# Cls* reports number of class labels available after removing low prevalent class labels. Between classical algorithms, overall, the performance of SVM is better than RF.

The comparison between SVM and FastText, the performances are quite close for many events, and event types. The transformer based models are outperforming across events, event types, and combined dataset.

The comparison of different transformer-based models entails that DistilBERT shows similar performance as opposed to BERT model, therefore, the use of DistilBERT in real applications might be a reasonable choice. RoBERTa and XLM-RoBERTa are outperforming BERT and DistilBERT, which comes with the cost of their large number of parameters. In terms of comparing monolingual (RoBERTa) vs. multilingual (XLM-RoBERTa) version of RoBERTa the gain with monolingual training is higher. In terms of comparing event type and all data, for the earthquake event type, we attain higher performance as the classifier is trained and evaluated on nine class labels as opposed to ten class labels for other event types and the combined dataset.

6 Discussions

There has been significant progress in crisis informatics research in the last several years due to the growing interest in the community and publicly available resources. One of the major interests is analyzing social media data and finding actionable information to facilitate humanitarian organizations. Models have been developed using publicly available datasets to streamline this process. Currently publicly available datasets are limited in different respects such as duplicates, and no fixed test set for the evaluation. These issues make it difficult to understand whether one approach outperforms another.

While closely inspecting the available datasets, we observed that there are duplicates and near-duplicates, which results in misleading performance figures. Another major limitation is that the reported results are not comparable because there is no fixed test set for the evaluation. That is, each set of results has been reported on its own data split, which makes it difficult to understand whether one approach outperforms another. To address such limitations and advance the crisis informatics research, in this study, we report our efforts to develop the largest-to-date Twitter dataset (i.e., $\sim$77,000 tweets) focusing on humanitarian response tasks. While developing the dataset we carefully designed a unique *data filtering and sampling pipeline*, which ensured the following characteristics: *(i)* four major disaster types, *ii)* disasters have occurred in different parts of the world at different times, *(iii)* selected samples are from disasters that created impact on land and caused major damage, *(iv)* content language is English, *(v)* data sampling is based on an in-house classifier to eliminate irrelevant content, *(vi)* there are at least three words in a tweet, *(vii)* exact and near-duplicates are removed (even across events for the combined dataset), and *(viii)* moderate to substantial inter-annotator agreement.

While using AMT for annotation, we ensured higher quality annotations by putting in place a qualification test, including gold-standard evaluation inside the tasks, and requiring an agreement score of minimum 66%.

To allow for accurate performance comparison and reporting across future studies, we make the data splits publicly available. We have conducted experiments using different classification algorithms on these data splits to provide baseline results for future research. In total, our experimental setup consists of more than 1000 experiments. Event-wise classification results can be useful to realize within- and across-event experimental comparisons. Whereas, the event-type results can be helpful to develop a generalized event-type-based model and compare it with new approaches. Similarly, a more generalized classifier can be developed using the combined dataset and our provided results can be helpful in future experimental comparison.

Data	# Cls	RF	SVM	FT	BERT	D-B	RT	X-R
2016 Ecuador Earthquake	8	0.784	0.738	0.752	0.861	**0.872**	**0.872**	0.866
2016 Canada Wildfires	8	0.726	0.738	0.726	**0.792**	0.781	0.791	0.768
2016 Italy Earthquake	6	0.799	0.822	0.821	0.871	0.878	**0.885**	0.877
2016 Kaikoura Earthquake	9	0.660	0.693	0.658	**0.768**	0.743	0.765	0.760
2016 Hurricane Matthew	9	0.742	0.700	0.704	0.786	0.780	**0.815**	0.784
2017 Sri Lanka Floods	8	0.613	0.611	0.575	0.703	0.763	0.727	**0.798**
2017 Hurricane Harvey	9	0.719	0.713	0.718	0.759	0.743	**0.763**	0.761
2017 Hurricane Irma	9	0.693	0.695	0.694	0.722	0.723	**0.730**	0.717
2017 Hurricane Maria	9	0.682	0.682	0.688	0.715	0.722	**0.727**	0.723
2017 Mexico Earthquake	8	0.800	0.789	0.797	0.845	0.854	**0.863**	0.847
2018 Maryland Floods	8	0.554	0.620	0.621	0.697	0.734	0.760	**0.798**
2018 Greece Wildfires	9	0.678	0.694	0.667	**0.788**	0.739	0.783	0.783
2018 Kerala Floods	9	0.670	0.694	0.714	0.732	0.732	0.745	**0.746**
2018 Hurricane Florence	9	0.731	0.717	0.735	0.768	0.773	**0.780**	0.765
2018 California Wildfires	10	0.676	0.696	0.713	0.760	**0.767**	0.764	0.757
2019 Cyclone Idai	10	0.680	0.730	0.707	0.790	0.779	**0.796**	0.793
2019 Midwestern U.S. Floods	7	0.643	0.632	0.624	0.702	0.706	**0.764**	0.726
2019 Hurricane Dorian	9	0.688	0.663	**0.693**	0.691	0.691	0.686	0.691
2019 Pakistan Earthquake	8	0.753	0.766	0.787	0.820	0.822	**0.834**	0.827
Earthquake	9	0.766	0.783	0.789	0.833	**0.839**	0.836	0.837
Fire	10	0.685	0.717	0.727	0.771	0.771	**0.787**	0.779
Flood	10	0.653	0.693	0.704	0.749	0.734	**0.758**	0.755
Hurricane	10	0.702	0.716	0.730	0.740	**0.742**	0.741	0.739
All	10	0.707	0.731	0.744	0.758	0.758	**0.760**	0.758
Average		**0.700**	**0.710**	**0.712**	**0.768**	**0.769**	**0.781**	**0.777**

Table 7: Classification results (weighted F1) for events, event-type and combined (All) dataset. Cls: Number of class labels, FT: FastText, X-R: XLM-RoBERTa, D-B: DistilBERT, RT: RoBERTa, Best results are highlighted with bold form.

7 Conclusions

The information available on social media has been widely used by humanitarian organizations at times of a disaster, which has been posted during an ongoing crisis event. However, most of these posts are not useful or relevant and need to be filtered out to have a concise summary. Besides, fine-grained analysis and understanding are also necessary to take actionable decisions. Such fine-grained analysis could be a report of "infrastructure or utility damage," "urgent needs," and so on. This requires having a classifier that can categorize such information. Our study focused on creating a dataset, which can be used to train a classifier and to categorize such information useful for actionable decisions. To this end, HumAID is the largest dataset, which will be publicly available for the research community. We also provide classification benchmark results, which can used to compare in future studies.

Broader Impact

We collected tweets from Twitter using Twitter streaming API by following its terms of service. The annotated dataset can be used to develop a model for humanitarian response tasks. We release the dataset by maintaining Twitter data redistribution policy.

References

Alam, F.; Alam, T.; Ofli, F.; and Imran, M. 2021a. Social Media Images Classification Models for Real-time Disaster Response. *arXiv:2104.04184* .

Alam, F.; Imran, M.; and Ofli, F. 2019. CrisisDPS: Crisis Data Processing Services. In *ISCRAM*.

Alam, F.; Ofli, F.; and Imran, M. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *ICWSM*, 465–473.

Alam, F.; Ofli, F.; and Imran, M. 2019. Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria. *Behaviour & Information Technology* 1–31.

Alam, F.; Ofli, F.; Imran, M.; Alam, T.; and Qazi, U. 2020. Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response. In *ASONAM*, 151–158.

Alam, F.; Sajjad, H.; Imran, M.; and Ofli, F. 2021b. CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. In *ICWSM*.

Alharbi, A.; and Lee, M. 2019. Crisis Detection from Arabic Tweets. In *Workshop on Arabic Corpus Ling.*, 72–79.

Breiman, L. 2001. Random forests. *Machine learning* 45(1): 5–32.

Burel, G.; and Alani, H. 2018. Crisis Event Extraction Service (CREES)-Automatic Detection and Classification of Crisis-related Content on Social Media. In *ISCRAM*.

Castillo, C. 2016. *Big Crisis Data*. Cambridge University Press.

Chowdhury, S. A.; Calvo, M.; Ghosh, A.; Stepanov, E. A.; Bayer, A. O.; Riccardi, G.; García, F.; and Sanchis, E. 2015. Selection and aggregation techniques for crowdsourced semantic annotation task. In *Proc. of 16th ISCA*.

Chowdhury, S. A.; Ghosh, A.; Stepanov, E. A.; Bayer, A. O.; Riccardi, G.; and Klasinas, I. 2014. Cross-language transfer of semantic annotation via targeted crowdsourcing. In *ISCA*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov,

V. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Fleiss, J. L.; Levin, B.; and Paik, M. C. 2013. *Statistical methods for rates and proportions*.

Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys* 47(4): 67.

Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; and Vieweg, S. 2014. AIDR: Artificial intelligence for disaster response. In *Proc. of the WWW*, 159–162.

Imran, M.; Elbassuoni, S.; Castillo, C.; Diaz, F.; and Meier, P. 2013. Practical extraction of disaster-relevant information from social media. In *WWW*, 1021–1024.

Imran, M.; Mitra, P.; and Castillo, C. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *LREC*.

Jain, P.; Ross, R.; and Schoen-Phelan, B. 2019. Estimating Distributed Representation Performance in Disaster-Related Social Media Classification. In *ASONAM*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*, 427–431.

Kersten, J.; Kruspe, A.; Wiegmann, M.; and Klan, F. 2019. Robust Filtering of Crisis-related Tweets. In *ISCRAM*.

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

Kropczynski, J.; Grace, R.; Coche, J.; Jalse, S.; Obeysekare, E.; Montarnal, A.; Benaben, F.; and Tapia, A. 2018. Identifying Actionable Information on Social Media for Emergency Dispatch. *ISCRAM* .

Kumar, S.; Barbier, G.; Abbasi, M. A.; and Liu, H. 2011. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In *ICWSM*.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* 159–174.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692* .

Martinez-Rojas, M.; del Carmen Pardo-Ferreira, M.; and Rubio-Romero, J. C. 2018. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management* 43: 196–208.

Mccreadie, R.; Buntain, C.; and Soboroff, I. 2019. TREC Incident Streams: Finding Actionable Information on Social Media. In *ISCRAM*.

Nemeskey, D. M.; and Kornai, A. 2018. Emergency vocabulary. *Information Systems Frontiers* 20(5): 909–923.

Neppalli, V. K.; Caragea, C.; and Caragea, D. 2018. Deep Neural Networks versus Naïve Bayes Classifiers for Identifying Informative Tweets during Disasters. In *ISCRAM*.

Nguyen, D. T.; Al-Mannai, K.; Joty, S. R.; Sajjad, H.; Imran, M.; and Mitra, P. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks. In *ICWSM*, 632–635.

Ofli, F.; Alam, F.; and Imran, M. 2020. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. *arXiv:2004.11838* .

Okolloh, O. 2009. Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory learning and action* 59(1): 65–70.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *ICWSM*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* .

Platt, J. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft.

Purohit, H.; and Sheth, A. P. 2013. Twitris v3: From Citizen Sensing to Analysis, Coordination and Action. In *ICWSM*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 851–860.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108* .

Saravanou, A.; Valkanas, G.; Gunopulos, D.; and Andrienko, G. 2015. Twitter floods when it rains: a case study of the UK floods in early 2014. In *WWW*, 1233–1238.

Starbird, K.; Palen, L.; Hughes, A. L.; and Vieweg, S. 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *CSCW*, 241–250.

Strassel, S. M.; Bies, A.; and Tracey, J. 2017. Situational Awareness for Low Resource Languages: the LORELEI Situation Frame Annotation Task. In *SMERP@ ECIR*, 32–41.

Tsou, M.-H.; Jung, C.-T.; Allen, C.; Yang, J.-A.; Han, S. Y.; Spitzberg, B. H.; and Dozier, J. 2017. Building a Real-Time Geo-Targeted Event Observation (Geo) Viewer for Disaster Management and Situation Awareness. In *ICC*.

Vieweg, S.; Castillo, C.; and Imran, M. 2014. Integrating social media communications into the rapid assessment of sudden onset disasters. In *SocInfo*, 444–461.

Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *SIGCHI*, 1079–1088.

Wiegmann, M.; Kersten, J.; Klan, F.; Potthast, M.; and Stein, B. 2020. Analysis of Detection Models for Disaster-Related Tweets. In *ISCRAM*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*, volume 32.

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