Standardizing and Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing

Firoj Alam        Hassan Sajjad        Muhammad Imran        Ferda Ofli
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
{fialam, hsajjad, mimran, fofli}@hbku.edu.qa

Abstract

Time-critical analysis of social media streams is important for humanitarian organizations to plan rapid response during disasters. The crisis informatics research community has developed several techniques and systems to process and classify big crisis related data posted on social media. However, due to the dispersed nature of the datasets used in the literature, it is not possible to compare the results and measure the progress made towards better models for crisis informatics. In this work, we attempt to bridge this gap by standardizing various existing crisis-related datasets. We consolidate labels of eight annotated data sources and provide 166.1k and 141.5k tweets for informativeness and humanitarian classification tasks, respectively. The consolidation results in a larger dataset that affords the ability to train more sophisticated models. To that end, we provide baseline results using CNN and BERT models. We make the dataset available at
https://crisisnlp.qcri.org/crisis_datasets_benchmarks.html

1 Introduction

At the onset of a disaster event, information pertinent to situational awareness such as reports of injured, trapped, or deceased people, urgent needs of victims, and infrastructure damage reports is most needed by formal humanitarian organizations to plan and launch relief operations. Acquiring such information in real-time is ideal to understand the situation as it unfolds. However, it is challenging as traditional methods such as field assessments and surveys are time-consuming. Microblogging platforms such as Twitter have been widely used to disseminate situational and actionable information by the affected population. Although social media sources are useful in this time-critical setting, it is, however, challenging to parse and extract actionable information from big crisis data available on social media (Castillo, 2016).

The past couple of years have witnessed a surge in the research works that focus on analyzing the usefulness of social media data and developing computational models to extract actionable information. Among others, proposed computational techniques include information classification, information extraction, and summarization (Imran et al., 2015; Rudra et al., 2018). Most of these studies use one of the publicly available datasets, reported in [Olteanu et al., 2014] [Imran et al., 2016] [Alam et al., 2018], either proposing a new model or reporting higher performance of an existing model. Typical classification tasks in the community include (i) informativeness (i.e., informative reports vs. not-informative reports), (ii) humanitarian information type classification (e.g., affected individual reports, infrastructure damage reports), and (iii) event type classification (e.g., flood, earthquake, fire).

Despite the recent focus of the crisis informatics research community to develop novel and more robust computational algorithms and techniques to process social media data, very limited efforts have been invested to develop standard datasets and benchmarks for others to compare their results, models, and techniques. In this paper, we develop a standard social media dataset for disaster response to facilitate comparison between different modeling approaches and to encourage the community to streamline their efforts towards a common goal. We can create such a standard benchmark dataset thanks to the publicly available datasets. The consolidated data is also larger in size and has better class distribution compared to the individual datasets, which are two important data features for building better models.

https://en.wikipedia.org/wiki/Disaster_informatics
We consolidate eight annotated datasets, namely, CrisisLex (Olteanu et al., 2014; Olteanu et al., 2015), CrisisNLP (Imran et al., 2016), SWDM2013 (Imran et al., 2013a), ISCRAM13 (Imran et al., 2013b), Disaster Response Data (DRD), Disasters on Social Media (DSM), CrisisMMD (Alam et al., 2018), and data collected by AIDR system (Imran et al., 2014). One of the challenges while consolidating the datasets is the inconsistent class labels across the datasets. One of the earlier efforts of defining the class labels and terminologies is the work of Temnikova et al. (2015). The CrisisLex, CrisisNLP and CrisisMMD datasets used similar definitions discussed in (Temnikova et al., 2015). Across several studies, a commonality exists at the semantic level of class labels used in different datasets. In this study, we map the class labels across datasets using their semantic meaning—a step performed by domain experts manually.

Another challenge while consolidating different social media datasets is to tackle the duplicate content that is present within or across datasets. There are three types of duplicates: (i) tweet-id based duplicates (i.e., same tweet appears in different datasets), (ii) content-based duplicates (i.e., tweets with different ids have same content), which usually happens when users copy-paste tweets, and (iii) near-duplicate content (i.e., tweets with similar content), which happens due to retweets or partial copy of tweets from other users. We use cosine similarity between tweets to filter out various types of duplicates. The contributions of this work are as follows.

- We consolidate all publicly available disaster-related datasets by manually mapping semantically similar class labels. We filter exact and near-duplicate tweets to clean the data and avoid any experimental biases.
- We provide benchmark results using state-of-the-art learning algorithms such as Convolutional Neural Networks (CNN) and pre-trained BERT models (Devlin et al., 2018) for two classifications tasks, i.e., Informativeness (binary) and Humanitarian type (multi-class) classification. The benchmarking encourages the community towards comparable and reproducible research.
- For the research community, we aim to release the dataset in multiple forms as, (i) a consolidated class label mapped version, (ii) exact- and near-duplicate filtered version obtained from previous version, (iii) a subset of the filtered data used for the classification experiments in this study.
- Our released dataset also includes a language tag, which enables the use of multilingual information in classification and is a promising future research direction.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the existing work. Section 3 describes our data consolidation procedures, and Section 4 describes the experiments and Section 5 presents and discusses the results. Finally, Section 6 concludes the paper.

## 2 Related Work

**Dataset Consolidation:** In crisis informatics research on social media, there has been an effort to develop datasets for the research community. An extensive literature review can be found in (Imran et al., 2015). Although there are several publicly available datasets that are used by the researchers, their results are not exactly comparable due to the differences in class labels and train/dev/test splits. Alam et al. (2019) and Kersten et al. (2019) have previously worked in this direction to consolidate social media disaster response data. However, both of these studies have limitations because Alam et al. (2019) did not consider the issue of duplicate and near-duplicate content when combining different datasets while Kersten et al. (2019) focused only on informativeness classification. A fair comparison of the classification experiment is also difficult with these two studies as their train/dev/test splits are not public. We address such limitations in this study, i.e., we consolidate the datasets, eliminate duplicates, and release train/dev/test splits publicly with benchmark results.

In terms of defining class labels (i.e., ontologies) for crisis informatics, most of the earlier efforts are discussed in Imran et al. (2015) and Temnikova et al. (2015). Various recent studies (Olteanu et al., 2014; Imran et al., 2016; Alam et al., 2018) use a similar definitions.

[https://www.figure-eight.com/dataset/combined-disaster-response-data/](https://www.figure-eight.com/dataset/combined-disaster-response-data/)
[https://data.world/crowdflower/disasters-on-social-media](https://data.world/crowdflower/disasters-on-social-media)

Note that in this study informativeness classification is also referred to as related vs. not-related.
Table 1: Different datasets and their sizes before and after label mapping and filtering steps.

| Source         | Total | Mapping | Informativeness | Humanitarian | Informativeness | Humanitarian |
|----------------|-------|---------|-----------------|--------------|-----------------|--------------|
| CrisisLex      | 88,015| 84,407  | 84,407          | 69,699       | 69,699          |              |
| CrisisNLP      | 52,656| 51,271  | 50,824          | 40,401       | 40,074          |              |
| SWDM13         | 1,543 | 1,344   | 802             | 857          | 699             |              |
| ISCRAM13       | 3,617 | 3,196   | 1,702           | 2,521        | 1,506           |              |
| DRD            | 26,235| 21,519  | 7,505           | 20,896       | 7,419           |              |
| DSM            | 10,876| 10,800  | 0               | 8,835        | 0               |              |
| CrisisMMD      | 16,058| 16,058  | 16,058          | 16,020       | 16,020          |              |
| AIDR           | 7,411 | 7,396   | 6,580           | 6,116        |                 |              |
| **Total**      | **206,411** | **195,991** | **167,878**   | **166,098**  | **141,533**    |              |

Different from them, Strassel et al. (2017) defines categories based on need types (e.g., evacuation, food supply) and issue type (e.g., civil unrest). In this study, we use the class labels that are highly important for humanitarian aid for disaster response task, which also has a commonality across the publicly available resources.

Classification Algorithms: Despite a majority of studies in crisis informatics literature employ traditional machine learning algorithms for automatic event detection, event type classification, and fine-grained humanitarian information type classification, several recent works explore deep learning algorithms in disaster-related tweet classification tasks. The study of Nguyen et al. (2017) and Neppalli et al. (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) and few-shot learning (Kruspe et al., 2019). Very recently different types of embedding representations have been proposed in literature such as Embeddings from Language Models (ELMo) (Peters et al., 2018), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), and XLNet (Yang et al., 2019) for different NLP tasks. For disaster-related classification, Jain et al. (2019) investigates these embedding representations and achieves similar results.

3 Data Consolidation

We consolidate some of the most prominent, publicly-available social media datasets that were labeled for different disaster response classification tasks. In doing so, we deal with two major challenges: (i) discrepancies in the class labels used across different datasets, and (ii) exact- and near-duplicate content that exists within as well as across different datasets.

In this study, we focus on eight datasets that have annotations and can be mapped consistently for two tasks: informativeness classification and humanitarian information type categorization. These datasets include CrisisLex (CrisisLexT6 (Olteanu et al., 2014), CrisisLexT26 (Olteanu et al., 2015)), CrisisNLP (Imran et al., 2016), SWDM2013 (Imran et al., 2013a), ISCRAM13 (Imran et al., 2013b), Disaster Response Data (DRD), Disasters on Social Media (DSM), CrisisMMD (Alam et al., 2018), and data collected by AIDR system (Imran et al., 2014). Second column of Table 1 summarizes original sizes of the datasets. From the table, we observe that CrisisLex and CrisisNLP are the largest and second-largest datasets, respectively, which are currently widely used in the literature. The SWDM2013 is the smallest set, which is one of the earliest datasets for the crisis informatics community. Below we elaborate on the details of the data consolidation process of these datasets.

[Note that the AIDR system data has been annotated by domain experts and is available upon request.](https://www.figure-eight.com/dataset/combined-disaster-response-data/)
[https://data.world/crowdflower/disasters-on-social-media]
3.1 Class Label Mapping

To combine these datasets, we create a set of common class labels by manually mapping class labels that come from different datasets but have the same or similar semantic meanings. For example, the label “building damaged,” originally used in the AIDR system, is mapped to “infrastructure and utilities damage” in our final dataset. Some of the class labels in these datasets are not annotated for humanitarian purposes, therefore, we have not included them in the consolidated dataset. For example, we do not select tweets labeled as “animal management” or “not labeled” that appear in CrisisNLP and CrisisLex26. This causes a drop in the number of tweets for both informativeness and humanitarian tasks as can be seen in Table 1 (Mapping column). The large drop in the CrisisLex dataset for the informativeness task is due to the 3,103 unlabeled tweets (i.e., labeled as “not labeled”). The other significant drop in the number of training examples for the informativeness task is in the DRD dataset. This is because many tweets were annotated with multiple labels, which we have not included in our consolidated dataset.

Many tweets in these datasets were labeled for informativeness only. For example, the DSM dataset is only labeled for informativeness, and a large portion of the DRD dataset is labeled for informativeness only. Therefore, we were not able to map them for the humanitarian task. More details of this mapping for different datasets are reported in the supplementary material.

3.2 Exact- and Near-Duplicate Filtering

To develop a machine learning model, it is important to design non-overlapping train/dev/test splits. A common practice is to randomly split the dataset into train/dev/test sets.

This approach does not work with social media data as it generally contains duplicates and near duplicates. Such duplicate content, if present in both train and test sets, often leads to overestimated test results during classification. Filtering the near-and-exact duplicate content is one of the major steps we have taken into consideration while consolidating the datasets.

We first tokenize the text before applying any filtering. For tokenization, we used a modified version of the Tweet NLP tokenizer [9]. Our modification includes lowercasing the text and removing URL, punctuation, and user id mentioned in the text. We then filter tweets having only one token. Next, we apply exact string matching to remove exact duplicates. An example of an exact duplicate tweet is: “RT Reuters: BREAKING NEWS: 6.3 magnitude earthquake strikes northwest of Bologna, Italy: USGS”, which appear three times with exact match in CrisisLex26 [10] dataset that has been collected during Northern Italy Earthquakes, 2012.

Then, we use a similarity-based approach to remove the near-duplicates. To do this, we first convert the tweets into vectors of uni- and bi-grams with their frequency-based representations. We then use cosine similarity to compute a similarity score between two tweets and flag them as duplicate if their similarity score is greater than the threshold value of 0.75. In the similarity-based approach, threshold selection is an important aspect. Choosing a lower value would remove many distant tweets while choosing a higher value would leave several near-duplicate tweets in the dataset. To determine a plausible threshold value, we manually looked into the tweets in different threshold bins (i.e., 0.70 to 1.0 with 0.05 interval) as shown in Figure 1 which we selected from consolidated informativeness dataset. By investigating the distribution, we concluded that a threshold value of 0.75 is a reasonable choice. From the figure we can clearly see that choosing a lower threshold (e.g., < 0.75) removes larger number of tweets. Note that rest of the tweets have similarity lower than what we have reported in the figure. In Table 2, we provide a few examples for the sake of clarity.

To understand, which events and in which dataset has more exact- and near-duplicate we attempted to analyze them. In Figure 2 we provide such duplicates counts for both exact- and near-duplicates for informativeness tweets. In the figure, we report total number (in parenthesis the number represent percentage of reduction) of duplicates (i.e., exact and near) for each dataset. The figure shows that CrisisLex and CrisisNLP have higher duplicates comparatively, however, it is because those two are larger.
datasets comparatively. For each of these datasets, we wanted to see which event’s duplicates appear most. In CrisisLex, the majority of the exact duplicates appear in “Queensland floods (2013)” consisting of 2270 exact duplicates. The second majority is “West Texas explosion 2013)” event, which consists of 1301 duplicates. Compared to CrisisLex, the exact duplicates are low in CrisisNLP, and the majority of such duplicates appear in the “Philippines Typhoon Hagupit (2014)” event with 1084 tweets. For the humanitarian tweets, we observed similar characteristics of Figure 2.

As indicated in Table 1, there is a drop after filtering, e.g., ~25% for informativeness and ~20% for humanitarian tasks. It is important to note that failing to eradicate duplicates from the consolidated dataset would potentially lead to misleading performance results in the classification experiments.

Figure 1: Number of near-duplicates in different bins obtained from consolidated informativeness tweets after label mapping. Tweets will lower similarity (< 0.7) bins are not reported here.

Figure 2: Exact- and near-duplicates in informativeness tweets. Number on top of each bar represent total number and the number in the parenthesis represent percentage of consolidated exact and near duplicates from the respective dataset.

Note that the event name that we are referring here is the events during which data has been collected by the respective data authors. We provided such information as a part of supplementary material.
| # | Tweet | Tokenized | Sim. | Dup. |
|---|---|---|---|---|
| 1 | RT @rosamaryCNN: As flood waters recede in Qld, #Australia, attention turns 2 relief & recovery. Police reportedly found a 5th victim... | rt as flood waters recede in qld australia attention turns relief recovery police reportedly found a th victim | 0.882 | × |
| 2 | Queensland counts flood cost as New South Wales braces for river peaks - The Guardian: The Guardian-Queensland co... | queensland counts flood cost as new south wales braces for river peaks the guardian-queensland co url | 0.856 | × |
| 3 | He’s no Anna Bligh! @abcnews LIVE: Queensland Premier Campbell Newman is giving an update on Queensland flood crisis http://t.co/pyxX0LoE | he’s no anna bligh live queensland premier campbell newman is giving an update on queensland flood crisis url | 0.808 | × |
| 4 | Australia lurches from fire to flood http://t.co/C6sx8Uxnk | australia lurches from fire to flood url | 0.807 | × |
| 5 | Live coverage: Queensland flood crisis via @Y7News http://t.co/Kkhbf07Fw | live coverage queensland flood crisis via url | 0.788 | × |
| 6 | Halo tetangga. Sabar ya. RT @AJEnglish: Flood worsens in eastern Australia http://t.co/IfokMQmB | halo tetangga sabar ya rt flood worsens in eastern australia url | 0.787 | × |
| 7 | Queensland counts flood cost as New South Wales braces for river peaks http://t.co/MPQskYt1 | queensland counts flood cost as new south wales braces for river peaks url brisbane friends moved to refuge. | 0.778 | × |
| 8 | RT @FoxNews: #BREAKING: Numerous injuries reported in large explosion at #Texas fertilizer plant http://t.co/H93riJAS7 | rt breaking numerous injuries reported in large explosion at texas fertilizer plant url | 0.744 | ✓ |
| 9 | Obama to attend memorial service for victims of Texas explosion: The president will meet with vic... | obama to attend memorial service for victims of texas explosion the president will meet with victims of the d url | 0.732 | ✓ |
| 10 | RT @RobertTaylors: Shooting Reported at Los Angeles International Airport: There are reports of a... | rt shooting reported at los angeles international airport there are reports of a shooting incident friday morning http ... | 0.705 | ✓ |
| 11 | @BuzzFeed: Watch Hurricane Sandy roll in from the top of the @nytimes building http://t.co/dL2g3sAH | watch hurricane sandy roll in from the top of the building url | 0.709 | ✓ |

Table 2: Examples of near-duplicates with similarity scores selected from informativeness tweets. Duplicates are highlighted. Sim. refers to similarity value. Dup. refers to whether we consider them as duplicate and filtered. The symbol (×) indicates a duplicate, which we dropped and the symbol (✓) indicates a not duplicate, which we have included in our dataset.
3.3 Adding Language Tags

While combining the datasets, we realized that some of them contain tweets in different languages (i.e., Spanish and Italian) other than English. In addition, many tweets have code-switched (i.e., multilingual) content. For example, the following tweet has both English and Spanish: “It’s #Saturday, #treat yourself to our #Pastel tres leches y compota de mora azul. https://t.co/WMpmu27P9X”. Note that Twitter tagged this tweet as English whereas the Google language detector service tagged it as Spanish with a confidence score of 0.379. After we realized this multilingual issue in the datasets, we decided to provide a language tag for each tweet if the language tag is not available with the respective dataset. For example, the tweets annotated by volunteers in the CrisisNLP dataset have language tags provided by Twitter whereas no language tag is provided with the CrisisLex dataset. For these tweets, we used the language detection API of Google Cloud Services[12,13]. We provided a language tag and confidence score obtained from the language detection API. Hence, with the consolidated dataset, we include a language tag for all tweets. In Figure 3, we report the distribution of languages with more than 20 tweets in the datasets. Among different languages of informativeness tweets, English tweets appear to be highest in the distribution compared to any other language, which is 94.46% of 156,899, as shown in Figure 3. Note that most of the non-English tweets appear in the CrisisLex dataset.

![Figure 3: Distribution of top nineteen languages (>= 20 tweets) in the consolidated informativeness tweets.](https://cloud.google.com/translate/docs/advanced/detecting-language-v3)

3.4 Data Statistics

Distribution of class labels is an important factor for developing the classification model. In Table 3 and 4, we report individual datasets along with the class label distribution for informativeness and humanitarian tasks, respectively. It is clear that there is an imbalance in class distributions in different datasets and some class labels are not present. For example, the distribution of “not informative” class is very low in SWDM13 and ISCRAM13 datasets. For the humanitarian task, some class labels are not present in different datasets. Only 17 tweets with the label “terrorism related” are present in CrisisNLP. Similarly, the class “disease related” only appears in CrisisNLP. The scarcity of the class labels poses a great challenge to design the classification model using individual datasets. Even after combining the datasets,
the imbalance in class distribution seems to persist (last column in Table 4). For example, the distribution of “Not humanitarian” is relatively higher (37.40%) than other class labels, which might have to be under-sampled for training the classification model. In Table 4, we highlighted some class labels, which we dropped in the rest of the classification experiments conducted in this study, however, tweets with those class labels will be available in the released datasets. The reason for not including them in the experiments is that we aim to develop classifiers for the disaster response tasks only.

| Class                  | CrisisLex | CrisisNLP | SWDM13 | ISCRAM13 | DRD | DSM | CrisisMMD | AIDR | Total  |
|------------------------|-----------|-----------|--------|----------|-----|-----|-----------|------|--------|
| Informative            | 42,140    | 23,694    | 716    | 2,443    | 14,849 | 3,461 | 11,488    | 2,968 | 101,759 |
| Not informative        | 27,559    | 16,707    | 141    | 78       | 6,047  | 5,374 | 4,532     | 3,901 | 64,339 |
| Total                  | 69,699    | 40,401    | 857    | 2,521    | 20,896 | 8,835 | 16,020    | 6,869 | 166,098 |

Table 3: Data distribution of informativeness across different sources.

| Class                  | CrisisLex | CrisisNLP | SWDM13 | ISCRAM13 | DRD | DSM | CrisisMMD | AIDR | Total |
|------------------------|-----------|-----------|--------|----------|-----|-----|-----------|------|-------|
| Affected individual    | 3,740     | -         | -      | -        | -   | -   | -         | -    | 4,211 |
| Caution and advice     | 1,774     | 1,137     | 117    | 412      | -   | -   | -         | -    | 3,601 |
| Disease related        | -         | 1,478     | -      | -        | -   | -   | -         | -    | 1,478 |
| Displaced and evacuations | -      | 495       | -      | -        | -   | -   | -         | -    | 545   |
| Donation and volunteering | 1,932   | 2,882     | 27     | 189      | 10  | 3,286 | 24        | 8,350 |
| Infrastructure and utilities damage | 1,353   | 1,721     | -      | -        | 877 | 1,262 | 283       | 5,496 |
| Injured or dead people | -         | 2,151     | 139    | 125      | -   | 486  | 267       | 3,168 |
| Missing and found people | -        | 443       | -      | 43       | -   | 40   | 46        | 572   |
| Not humanitarian       | 27,559    | 16,708    | 142    | 81       | -   | 4,538| 3,911     | 52,939|
| Other relevant information | 29,562 | 8,188     | -      | -        | -   | 5,937| 939       | 44,626|
| Personal update        | -         | 116       | 274    | 656      | -   | -   | -         | -    | 1,046 |
| Physical landslide     | -         | 538       | -      | -        | -   | -   | -         | -    | 538   |
| Requests or needs      | -         | 215       | -      | -        | 6,532| -   | 257       | 7,004 |
| Response efforts       | -         | 1,114     | -      | -        | -   | -   | 1,114     | -    | 7,004 |
| Sympathy and support  | 3,779     | 2,872     | -      | -        | -   | -   | -         | 178  | 6,829 |
| Terrorism related      | -         | 16        | -      | -        | -   | -   | -         | -    | 16    |
| Total                  | 69,699    | 40,074    | 699    | 1,506    | 7,419| 16,020| 6,869     | 141,533|

Table 4: Data distribution of humanitarian categories across different datasets.

4 Experiments

Although our consolidated dataset contains multilingual tweets, we only use tweets in English language in our experiments. We split data into train, dev, and test sets with a proportion of 70%, 10%, and 20%, respectively, also reported in Table 5. As mentioned earlier we have not selected the tweets with highlighted class labels in Table 4 for the classification experiments. Therefore, in the rest of the paper, we report the class label distribution and results on the selected class labels with English tweets only.

4.1 Experimental Settings

Individual vs. Consolidated Datasets: The motivation of these experiments is to investigate whether consolidated dataset helps in improving the classification performance. For the individual dataset classification experiments, we selected CrisisLex and CrisisNLP as they are reasonably large in size and have a reasonable number of class labels, i.e., six and eleven class labels, respectively. Note that these are subsets of the consolidated dataset reported in Table 5. We selected them from train, dev and test splits of the consolidated dataset to be consistent across different classification experiments. To understand the effectiveness of the smaller datasets, we run experiments by training the model using smaller datasets and evaluating using the consolidated test set.

CNN vs. BERT using Consolidated Dataset: The recent development of the pre-trained BERT model has shown success in different downstream NLP tasks. In this study, we wanted to compare the performance of the widely used CNN model with BERT model. We chose to use the consolidated dataset for the experiments.
### Table 5: Data split and their distributions with the consolidated English tweets dataset.

| Informativeness          | Train | Dev  | Test  | Total  |
|--------------------------|-------|------|-------|--------|
| Informative              | 65826 | 9594 | 18626 | 94046  |
| Not informative          | 43970 | 6414 | 12469 | 62853  |
| **Total**                | 109796| 16008| 31095 | 156899 |

| Humanitarian             |       |      |       |        |
|--------------------------|-------|------|-------|--------|
| Affected individual      | 2454  | 367  | 693   | 3514   |
| Caution and advice       | 2101  | 309  | 583   | 2993   |
| Displaced and evacuations| 359   | 53   | 99    | 511    |
| Donation and volunteering| 5184  | 763  | 1453  | 7400   |
| Infrastructure and utilities damage | 3541 | 511  | 1004  | 5056   |
| Injured or dead people   | 1945  | 271  | 561   | 2777   |
| Missing and found people | 373   | 55   | 103   | 531    |
| Not humanitarian         | 36109 | 5270 | 10256 | 51635  |
| Requests or needs        | 4840  | 705  | 1372  | 6917   |
| Response efforts         | 780   | 113  | 221   | 1114   |
| Sympathy and support     | 3549  | 540  | 1020  | 5109   |
| **Total**                | 61235 | 8957 | 17365 | 87557  |

**Event-aware Training**  
The availability of annotated data for a disaster event is usually scarce. One of the advantages of our compiled data is to have identical classes across several disaster events. This enables us to combine the annotated data from all previous disasters for the classification. Though this increases the size of the training data substantially, the classifier may result in sub-optimal performance due to the inclusion of heterogeneous data (i.e., a variety of disaster types and occurs in a different part of the world).

Sennrich et al. (2016) proposed a tag-based strategy where they add a tag to machine translation training data to force a specific type of translation. The method has later been adopted to do domain adaptation and multilingual machine translation (Chu et al., 2017). Motivated by it, we propose an event-aware training mechanism. Given a set of disaster event types \( D = \{ d_1, d_2, ..., d_m \} \) where disaster event type \( d_i \) includes earthquake, flood, fire, hurricane. For a disaster event type \( d_i \), \( T_i = \{ t_1, t_2, ..., t_n \} \) are the annotated tweets. We append a disaster event type as a token to each annotated tweet \( t_i \). More concretely, say tweet \( t_i \) consists of \( k \) words \( \{ w_1, w_2, ..., w_k \} \). We append a disaster event type tag \( d_i \) to each tweet so that \( t_i \) would become \( \{ d_i, w_1, w_2, ..., w_k \} \). We repeat this step for all disaster event types present in our dataset. We concatenate the modified data of all disasters and use it for the classification. Different from concatenating the original data, we are essentially preserving the domain information present in the data while making use of all of the data for the classification.

The event-aware training requires the knowledge of the disaster event type at the time of the test. If we do not provide a disaster event type, the classification performance will be suboptimal due to a mismatch between train and test. In order to apply the model to an unknown disaster event type, we modify the training procedure. Instead of appending the disaster event type to all tweets of a disaster, we randomly append disaster event type \( \text{UNK} \) to 5% of the tweets of every disaster. Note that \( \text{UNK} \) is now distributed across all disaster event types and is a good representation of an unknown event.

### 4.2 Models and Architectures

In this section, we describe the details of our classification models. For the experiments, we use CNN and pre-trained BERT model for experimentation.

**Classification using CNN**  
The current state-of-the-art disaster classification model is based on the CNN architecture. We used similar architecture as proposed by Nguyen et al. (2017).

**Classification using BERT**  
Pre-trained models have achieved state-of-the-art performance on natural language processing tasks and have been adopted as feature extractors for solving down-stream tasks such as question answering, and sentiment analysis. Though the pre-trained models are mainly trained on non-Twitter text, we hypothesize that their rich contextualized embeddings would be beneficial for
### Table 6: Classification results for the individual and consolidated train sets using the CNN model. 2C, 6C, 10C, and 11C refer to two, six, ten and eleven class labels, respectively.

|                  | Train         | Test         | Acc | P   | R   | F1  |
|------------------|---------------|--------------|-----|-----|-----|-----|
| **Informativeness** |               |              |     |     |     |     |
| CrisisLex (2C)   | Consolidated  | 0.801        | 0.807| 0.800| 0.803|
| CrisisNLP (2C)   | Consolidated  | 0.725        | 0.768| 0.730| 0.727|
| Consolidated (2C)| Consolidated  | 0.867        | 0.866| 0.870| 0.866|
| **Humanitarian** |               |              |     |     |     |     |
| CrisisLex (6C)   | Consolidated  | 0.694        | 0.601| 0.690| 0.633|
| CrisisNLP (10C)  | Consolidated  | 0.666        | 0.582| 0.670| 0.613|
| Consolidated (11C)| Consolidated | 0.835        | 0.827| 0.840| 0.829|

The disaster domain. In this work, we choose the pre-trained model BERT (Devlin et al., 2018) for the classification task. We follow the standard fine-tuning procedure with a task-specific layer added on top of the BERT architecture.

**Model Settings** We train the CNN models using the Adam optimizer (Kingma and Ba, 2014). The maximum number of epochs is set to 1000. We set early stopping criterion based on the accuracy of the development set with a patience of 200. We use a filter size of 300 filters with both window size and pooling length of 2, 3, and 4. We use BERT-base model (Devlin et al., 2018) using the Transformer Toolkit (Wolf et al., 2019). The model consists of 12 layers plus an additional task-specific layer. We fine-tune the model using the default settings for three epochs as prescribed by Devlin et al. (2018).

### 4.3 Preporcessing and Evaluation

**Data Preprocessing** Prior to the classification experiment, we preprocess tweets to remove symbols, emoticons, invisible and non-ASCII characters, punctuations (replaced with whitespace), numbers, URLs, and hashtag signs. We also remove stop words.

**Evaluation Settings** To measure the performance of each classifier, we use weighted average precision (P), recall (R), and F1-measure (F1). The rationale behind choosing the weighted metric is that it takes into account the class imbalance problem.

### 5 Results and Discussions

#### 5.1 Individual vs. Consolidated Dataset

In Table 6 we report the classification results for individual vs. consolidated datasets for both informativeness and humanitarian tasks using the CNN model. As mentioned earlier, we selected CrisisLex and CrisisNLP to conduct experiments for the individual datasets. Between CrisisLex and CrisisNLP, the performance is higher with CrisisLex dataset for both informativeness and humanitarian tasks. This might be due to the CrisisLex dataset being larger than the CrisisNLP dataset. The model trained using the consolidated dataset achieves 0.866 (F1) for informativeness and 0.829 for humanitarian, which is better than the models trained using individual datasets. In the humanitarian task, for different datasets in Table 6, we have different number of class labels. We report the results of those classes only for which the model is able to classify. For example, the model trained using the CrisisLex data can classify tweets using one of the six class labels (see Table 4 for excluded labels with highlights). The experiments with smaller datasets for both informativeness and humanitarian tasks show the importance to design a classifier using a larger dataset. Note that humanitarian task is a multi-class classification problem which makes it a much more difficult task than the binary informativeness classification.

#### 5.2 CNN vs. BERT

Table 7 compares the performance of CNN and BERT on the consolidated datasets. For the informativeness task, there is no difference in performance between CNN and BERT. However, on the humanitarian task, the BERT model outperforms CNN model by an absolute margin of 3.5 points in F1. The BERT model
Informativeness Humanitarian

| Model       | Acc    | P     | R     | F1    | Acc    | P     | R     | F1    |
|-------------|--------|-------|-------|-------|--------|-------|-------|-------|
| CNN         | 0.867  | 0.866 | 0.870 | **0.866** | 0.835  | 0.827 | 0.840 | **0.829** |
| BERT        | 0.866  | 0.866 | 0.866 | **0.865** | 0.866  | 0.865 | 0.866 | **0.864** |

Table 7: Classification results of the consolidated dataset using CNN and BERT.

| Class                              | CNN P | CNN R | CNN F1 | BERT P | BERT R | BERT F1 |
|------------------------------------|-------|-------|--------|--------|--------|---------|
| Affected individual                | 0.760 | 0.720 | 0.740  | 0.771  | 0.835  | 0.802   |
| Caution and advice                 | 0.630 | 0.630 | 0.630  | 0.684  | 0.720  | 0.702   |
| Displaced and evacuations          | 0.490 | 0.180 | 0.260  | 0.523  | 0.586  | 0.552   |
| Donation and volunteering          | 0.700 | 0.790 | 0.740  | 0.746  | 0.825  | 0.783   |
| Infrastructure and utilities damage| 0.650 | 0.660 | 0.660  | 0.727  | 0.704  | 0.716   |
| Injured or dead people             | 0.760 | 0.780 | 0.770  | 0.822  | 0.866  | 0.844   |
| Missing and found people           | 0.470 | 0.170 | 0.240  | 0.512  | 0.427  | 0.466   |
| Not humanitarian                   | 0.900 | 0.930 | 0.920  | 0.933  | 0.926  | 0.929   |
| Requests or needs                  | 0.850 | 0.840 | 0.850  | 0.916  | 0.907  | 0.912   |
| Response efforts                   | 0.330 | 0.070 | 0.120  | 0.424  | 0.226  | 0.295   |
| Sympathy and support               | 0.760 | 0.640 | 0.690  | 0.770  | 0.732  | 0.751   |

Table 8: Class-wise classification results of the consolidated dataset using CNN and BERT.

We further investigate the classification results of the CNN models for the minority class labels. We observe that the class “response efforts” is mostly confused with “donation and volunteering” and “not humanitarian”. For example, the following tweet with “response efforts” label, “I am supporting Rebuild Sankhu @crowdfunderuk #crowdfunder http://t.co/WBsKGZHHSj”, is classified as “donation and volunteering”. We also observe similar phenomena in minority class labels. The class “displaced and evacuations” is confused with “donation and volunteering” and “caution and advice”. It is interesting that the class “missing and found people” is confused with “donation and volunteering” and “not humanitarian”. The following “missing and found people” tweet, “RT @Fahdhusain: 11 kids recovered alive from under earthquake rubble in Awaran. Shukar Allah!!”, is classified as “donation and volunteering”.

5.3 Event-aware

In Table 9, we report the results of the event-aware training using both CNN and BERT. The event-aware training improves the classification performance by 1.3 points (F1) using CNN for the humanitarian task compared to the results without using event information (see Table 7). However, no improvement has been observed for the informativeness task. The training using event information enables the system to use data of all disasters while preserving the disaster-specific distribution.

Event-aware training is also effective in the advent of a new disaster event. Based on the type of a new disaster, one may use appropriate tags to optimize the classification performance. The event-aware training can be extended to use more than one tag. For example, in addition to preserving the event information, one can also append a tag for the disaster region. In this way, one can optimize the model for more fine-grained domain information.

The event-aware training with BERT does not provide better results in any of the tasks, which requires further investigation and we leave it as a future study.

5.4 Discussions

Social media data is noisy and it often poses a challenge for labeling and training classifiers. While investigating the publicly available datasets, we realized that it is important to follow a number of steps before preparing and labeling any social media dataset, not just the dataset for crisis computing. Such
steps include (i) tokenization to help in the subsequent phase, (ii) remove exact- and near-duplicates, (iii) check for existing data where the same tweet might be annotated for the same task, and then (iv) labeling. For designing the classifier, we postulate checking the overlap between training and test splits to avoid any misleading performance results.

The classification performance that we report is considered as benchmark results, which can be used to compare in any future study. The current state-of-art for informativeness and humanitarian tasks can be found in (Burel et al., 2017) [Alam et al., 2019]. The F-measure for informativeness and humanitarian tasks are reported as 0.838 and 0.613, respectively, on the CrisisLex26 dataset in (Burel et al., 2017). Whereas in (Alam et al., 2019), the reported F-measure for informativeness and humanitarian tasks are 0.93 and 0.78, respectively. It is important to emphasize the fact that the results reported in this study are reliable as they are obtained on a dataset that has been cleansed from duplicate content, which might have led to misleading performance results otherwise.

The competitive performance of BERT encourages us to try deeper models such as BERT-large [Devlin et al., 2018] and Google T5 [Raffel et al., 2019] models. Another interesting angle is to use pre-trained multilingual models to classify tweets in different languages. A future research direction is to use multilingual models for the zero-shot classification of tweets. For the BERT-based model, it is important to invest the effort to try different regularization methods to obtain better results, which we foresee as a future study. From the event-aware experiments, we see that it helps to improve the classification performance, which could also be a future research avenue.

As we aim to release different versions of the dataset with benchmark results for the research community, we believe that it will help the community to develop better models and compare results. Our consolidated dataset also includes a language tag, which will help to conduct multilingual experiments in future research. The resulting consolidated dataset covers a time-span starting from 2010 to 2017, which can be used to study temporal aspects in crisis scenarios.

6 Conclusions

The information available on social media has been widely used by humanitarian organizations at times of a disaster. Many techniques and systems have been developed to process social media data. However, the research community lacks a standard dataset and benchmarks to compare the performance of their systems. We tried to bridge this gap by consolidating existing datasets and providing benchmarks based on state-of-the-art CNN and BERT models.

Table 9: Classification results of the event-aware experiments using the consolidated dataset.

| Model | Informativeness | Humanitarian |
|-------|-----------------|--------------|
|       | Acc  | P    | R    | F1    | Acc  | P    | R    | F1    |
| CNN   | 0.868 | 0.868 | 0.870 | 0.867 | 0.847 | 0.841 | 0.850 | 0.842 |
| BERT  | 0.860 | 0.861 | 0.860 | 0.860 | 0.834 | 0.837 | 0.834 | 0.835 |

References

Firoj Alam, Ofli Ferda, and Imran Muhammad. 2018. Crisismmmd: Multimodal twitter datasets from natural disasters. In Proc. of the 12th ICWSM, 2018, pages 465–473. AAAI press, 1.

Firoj Alam, Imran Muhammad, and Ofli Ferda. 2019. Crisisdps: Crisis data processing services. In Proc. of 16th ISCRAM.

Grégoire Burel and Harith Alani. 2018. Crisis event extraction service (crees)-automatic detection and classification of crisis-related content on social media. In Proc. of the 15th ISCRAM, 2018.

Grégoire Burel, Hassan Saif, Miriam Fernandez, and Harith Alani. 2017. On semantics and deep learning for event detection in crisis situations. In Workshop on Semantic Deep Learning (SemDeep), at ESWC 2017, 5.

Carlos Castillo. 2016. Big Crisis Data. Cambridge University Press.
Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013a. Practical extraction of disaster-relevant information from social media. In Proc. of the 22nd WWW, pages 1021–1024. ACM.

Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013b. Extracting information nuggets from disaster-related messages in social media. In Proc. of the 12th ISCRAM.

Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In Proc. of the ACM Conference on WWW, pages 159–162. ACM.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. ACM Computing Surveys, 47(4):1.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In Proc. of the LREC, 2016, Paris, France, 5. ELRA.

Pallavi Jain, Robert Ross, and Bianca Schoen-Phelan. 2019. Estimating distributed representation performance in disaster-related social media classification. In 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE.

Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. 2019. Robust filtering of crisis-related tweets. In ISCRAM.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Anna Kruspe, Jens Kersten, and Friederike Klan. 2019. Detecting event-related tweets by example using few-shot models. In Proc. of the 16th ISCRAM.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. 2018. Deep neural networks versus naïve bayes classifiers for identifying informative tweets during disasters. In Proc. of the 15th ISCRAM, 2018.

Dat Tien Nguyen, Kamla Al-Mannai, Shafiq R Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In Proc. of the 11th ICWSM, 2017, pages 632–635. AAAI press.

Brendan O’Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In Fourth International AAAI Conference on Weblogs and Social Media.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In Proc. of the 8th ICWSM, 2014. AAAI press.

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In Proc. of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pages 994–1009. ACM.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Koustav Rudra, Pawan Goyal, Niloy Ganguly, Prasenjit Mitra, and Muhammad Imran. 2018. Identifying sub-events and summarizing disaster-related information from microblogs. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 265–274. ACM.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40, San Diego, California, June. Association for Computational Linguistics.
Stephanie M Strassel, Ann Bies, and Jennifer Tracey. 2017. Situational awareness for low resource languages: the lorelei situation frame annotation task. In *SMERP@ ECIR*, pages 32–41.

Irina P Temnikova, Carlos Castillo, and Sarah Vieweg. 2015. Entterms 1.0: A terminological resource for crisis tweets. In *ISCRAM*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

7 Supplementary Material

7.1 Dataset Details

**CrisisLex** is one of the largest publicly-available datasets, *CrisisLex*, which consists of two subsets, i.e., CrisisLexT26 and CrisisLexT6 (Olteanu et al., 2014). CrisisLexT26 comprises data from 26 different crisis events that took place in 2012 and 2013 with annotations for informative vs. not-informative as well as humanitarian categories (six classes) classification tasks among others. CrisisLexT6, on the other hand, contains data from six crisis events that occurred between October 2012 and July 2013 with annotations for related vs. not-related classification task.

**CrisisNLP** is another large-scale dataset collected during 19 different disaster events that happened between 2013 and 2015 and annotated according to different schemes including classes from humanitarian disaster response and some classes related to health emergencies (Imran et al., 2016).

**SWDM2013** dataset consists of data from two events. The Joplin collection contains tweets from the tornado that struck Joplin, Missouri on May 22, 2011. The Sandy collection contains tweets collected from Hurricane Sandy that hit Northeastern US on Oct 29, 2012 (Imran et al., 2013a).

**ISCRAM2013** dataset consists of tweets from two different events occurred in 2011 (Joplin 2011) and 2012 (Sandy 2012). The Joplin 2011 data consists of 4,400 labeled tweets collected during the tornado that struck Joplin, Missouri (USA) on May 22, 2011, whereas Sandy 2012 data consists of 2,000 labeled tweets collected during the Hurricane Sandy, that hit Northeastern US on Oct 29, 2012.

**DRD** consists of tweets collected during various crisis events that took place in 2010 and 2012. This dataset is annotated using 36 classes that include informativeness as well as humanitarian categories.

**DSM** dataset comprises 10K tweets collected and annotated with labels related vs. not-related to the disaster.

**CrisisMMD** is a multimodal dataset consisting of tweets and associated images collected during seven disaster events that happened in 2017 (Alam et al., 2018). The annotations available and relevant to this study include two classification tasks: informative vs. not-informative and humanitarian categories (eight classes).

**AIDR** is the labeled dataset obtained from *AIDR system* (Imran et al., 2014) that has been annotated by domain experts for different events and made available upon requests. We only retained labeled data that are relevant to this study.

7.2 Events and class label mapping

In Table 10 we report the events associated with the respective datasets such as ISCRAM2013, SWDM2013 CrisisLex and CrisisNLP. The time-period is from 2011 to 2015, which is a good representative of temporal aspects. In Table 11 we report class label mapping for ISCRAM2013, SWDM2013 CrisisLex and CrisisNLP datasets. The first column in the table shows the mapped class for both informative and humanitarian tasks. Note that all humanitarian class labels also mapped to informative and

---

14 [https://data.world/crowdflower/disasters-on-social-media](https://data.world/crowdflower/disasters-on-social-media)
not humanitarian labels mapped to not-informative in the data preparation step. In Table 12, we report the class label mapping for informativeness and humanitarian tasks for DRD dataset. The DSM dataset only contains tweets labeled as relevant vs not-relevant which we mapped for informativeness task as shown in Table 13. The CrisisMMD dataset has been annotated for informativeness and humanitarian task, therefore, very minor label mapping was needed as shown in Table in 14. The AIDR data has been labeled by domain experts using AIDR system and has been labeled during different events. The label names we mapped for informativeness and humanitarian tasks are shown in Table 15.
| Dataset       | Year     | Event name                      |
|--------------|----------|---------------------------------|
| ISCRAM2013   | 2011     | Joplin                          |
| SWDM2013     | 2012     | Sandy                           |
| **CrisisLex**|          |                                 |
| CrisisLexT6  | 2012     | US Sandy Hurricane              |
| CrisisLexT6  | 2013     | Alberta Floods                  |
| CrisisLexT6  | 2013     | Boston Bombings                 |
| CrisisLexT6  | 2013     | Oklahoma Tornado                |
| CrisisLexT6  | 2013     | Queensland Floods               |
| CrisisLexT6  | 2013     | West Texas Explosion            |
| CrisisLexT26 | 2012     | Costa Rica Earthquake           |
| CrisisLexT26 | 2012     | Italy Earthquakes               |
| CrisisLexT26 | 2012     | Philippines Floods              |
| CrisisLexT26 | 2012     | Philippines Typhoon Pablo       |
| CrisisLexT26 | 2012     | Venezuela Refinery Explosion    |
| CrisisLexT26 | 2013     | Alberta Floods                  |
| CrisisLexT26 | 2013     | Australia Bushfire              |
| CrisisLexT26 | 2013     | Bangladesh Savar building collapse |
| CrisisLexT26 | 2013     | Bohol Earthquake                |
| CrisisLexT26 | 2013     | Boston Bombings                 |
| CrisisLexT26 | 2013     | Brazil Nightclub Fire           |
| CrisisLexT26 | 2013     | Canada Lac Megantic Train Crash |
| CrisisLexT26 | 2013     | Colorado Floods                 |
| CrisisLexT26 | 2013     | Glasgow Helicopter Crash        |
| CrisisLexT26 | 2013     | Italy Sardinia Floods           |
| CrisisLexT26 | 2013     | LA Airport Shootings            |
| CrisisLexT26 | 2013     | Manila Floods                   |
| CrisisLexT26 | 2013     | NY Train Crash                  |
| CrisisLexT26 | 2013     | Philippines Typhoon Yolanda     |
| CrisisLexT26 | 2013     | Queensland Floods               |
| CrisisLexT26 | 2013     | Singapore haze                  |
| CrisisLexT26 | 2013     | West-Texas explosion            |
| CrisisLexT26 | 2012     | Guatemala Earthquake            |
| **CrisisNLP**|          |                                 |
| CrisisNLP-CF | 2013     | Pakistan Earthquake             |
| CrisisNLP-CF | 2014     | California Earthquake           |
| CrisisNLP-CF | 2014     | Chile Earthquake                |
| CrisisNLP-CF | 2014     | India Floods                    |
| CrisisNLP-CF | 2014     | Mexico Hurricane Odile          |
| CrisisNLP-CF | 2014     | Middle-East Respiratory Syndrome|
| CrisisNLP-CF | 2014     | Pakistan Floods                 |
| CrisisNLP-CF | 2014     | Philippines Typhoon Hagupit     |
| CrisisNLP-CF | 2014     | Worldwide Ebola                 |
| CrisisNLP-CF | 2015     | Nepal Earthquake                |
| CrisisNLP-CF | 2015     | Vanuatu Cyclone Pam             |
| CrisisNLP-volunteers | 2014-2015 | Worldwide Landslides         |
| CrisisNLP-volunteers | 2014   | California Earthquake           |
| CrisisNLP-volunteers | 2014 | Chile Earthquake                |
| CrisisNLP-volunteers | 2014 | Iceland Volcano                 |
| CrisisNLP-volunteers | 2014 | Malaysia Airline MH370          |
| CrisisNLP-volunteers | 2014 | Mexico Hurricane Odile          |
| CrisisNLP-volunteers | 2014 | Middle-East Respiratory Syndrome|
| CrisisNLP-volunteers | 2014 | Philippines Typhoon Hagupit     |
| CrisisNLP-volunteers | 2015 | Nepal Earthquake                |
| CrisisNLP-volunteers | 2015 | Vanuatu Cyclone Pam             |

Table 10: Events in CrisisLex, CrisisNLP, ISCRAM2013 and SWDM2013 datasets.
| Mapped class                  | Original class                              | Source                        | Annotation Description                                                                                                                                                                                                 |
|------------------------------|---------------------------------------------|-------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Affected individual          | Affected individuals                        | CrisisLexT26                  | Deaths, injuries, missing, found, or displaced people, and/or personal updates.                                                                                                                                                                                                 |
| Caution and advice           | Caution and advice                          | CrisisLexT26                  | It’s a message conveying/reporting information about some warning or a piece of advice about a possible hazard of an incident.                                                                                                                                                    |
| Disease related              | Disease signs or symptoms                   | CrisisNLPCF                  | Reports of symptoms such as fever, cough, diarrhea, and symptoms of breath or questions related to these symptoms.                                                                                                                                                       |
| Disease related              | Disease transmission                        | CrisisNLPCF                  | Reports of disease transmission or questions related to disease transmission.                                                                                                                                                                                                  |
| Disease related              | Disease Treatment                            | CrisisNLPCF                  | Questions or suggestions regarding the treatments of the disease.                                                                                                                                                                                                            |
| Disease related              | Disease Prevention                           | CrisisNLPCF                  | Questions or suggestions related to the prevention of disease or mention of a new prevention strategy.                                                                                                                                                                    |
| Disease related              | Disease Affected people                     | CrisisNLPCF                  | Reports of affected people due to the disease.                                                                                                                                                                                                                              |
| Displaced and evacuations    | Displaced people                             | CrisisNLPCF                  | People who have relocated due to the crisis, even for a short time (includes evacuations).                                                                                                                                                                                     |
| Donation and volunteering    | Donations and volunteering                  | CrisisNLPCF                  | Needs, requests, or offers of money, blood, shelter, supplies, and/or services by volunteers or professionals.                                                                                                                                                          |
| Donation and volunteering    | Donations of money                           | SWDM2013/ISCRAM2013           | Donations of money                                                                                                                                                                                                                                                          |
| Donation and volunteering    | Donations of money goods or services        | SWDM2013/ISCRAM2013           | Donations of money goods or services.                                                                                                                                                                                                                                        |
| Donation and volunteering    | Donations of supplies and/or volunteer work | SWDM2013/ISCRAM2013           | Donations of supplies and/or volunteer work                                                                                                                                                                                                                                 |
| Donation and volunteering    | Money                                       | CrisisNLPCF                  | Money requested, donated or spent                                                                                                                                                                                                                                         |
| Donation and volunteering    | Shelter and supplies                         | CrisisNLPCF                  | Needs or donations of shelter and/or supplies such as food, water, clothing, medical supplies or blood.                                                                                                                                                                    |
| Donation and volunteering    | Volunteers or professional services         | SWDM2013/ISCRAM2013           | Services needed or offered by volunteers or professionals.                                                                                                                                                                                                                  |
| Informative                  | Informative direct                          | SWDM2013/ISCRAM2013           | If the message is of interest to other people beyond the author’s immediate circle, and seems to be written by a person who is a direct eyewitness of what is happening.                                                                                           |
| Informative                  | Informative indirect                        | SWDM2013/ISCRAM2013           | If the message is of interest to other people beyond the author’s immediate circle, but there is not enough information to tell if it is a direct report or a repetition of something from another source.                                |
| Not humanitarian             | Not related to crisis                        | CrisisNLPCF                  | Not related to this crisis. 1. Refers to the crisis, but does not contain useful information that helps you understand the situation; 2. Not related to the Typhoon, or not relevant for emergency/humanitarian response; 3. Related to the crisis, but not informative: if it refers to the crisis, but does not contain useful information that helps understand the situation. |
| Not humanitarian             | Not related to the crisis                   | CrisisNLPCF                  | Not related to crisis.                                                                                                                                                                                                                                                         |
| Not humanitarian             | Not relevant                                | CrisisNLPCF                  | Not relevant                                                                                                                                                                                                                                                             |
| Not humanitarian             | Oil topic                                   | CrisisNLPCF                  | Not applicable                                                                                                                                                                                                                                                             |
| Not humanitarian             | Other                                       | CrisisNLPCF                  | Not relevant                                                                                                                                                                                                                                                               |
| Not humanitarian             | Not related to the physical landslide       | CrisisNLPCF                  | If the item does not reference a physical landslide.                                                                                                                                                                                                                       |
| Not humanitarian             | Terrorism not related                       | CrisisNLPCF                  | If the tweet is not about terrorism related to the flight MH370. 1. Other useful information that helps understand the situation; 2. Informativo para emergencia/humanitaria response, but in none of the above categories, including weather/evacuations; 3. Other useful information that helps understand the situation, but in none of the above categories, including weather/evacuations; 4. Other useful information not covered by any of the following categories: actors, infrastructure and utilities, donations and volunteering, caution and advice, sympathy and emotional support. |
| Other relevant information   | Other useful information                     | CrisisNLPCF                  | Related to the crisis, but not informative: if it refers to the crisis, but does not contain useful information that helps understand the situation.                                                                                                                                 |
| Other relevant information   | Related but not informative                 | CrisisLexT26                  | Related to the crisis, but not informative: if it refers to the crisis, but does not contain useful information that helps understand the situation.                                                                                                                             |
| Other relevant information   | Relevant                                     | CrisisLexT26                  | Relevant                                                                                                                                                                                                                                                                   |
| Personal update              | Personal                                     | CrisisNLPCF                  | If the tweet conveys some sort of personal opinion, which is not of interest of a general audience. 1. Personal and only useful to a small circle of family/friends of the author; 2. If a message is only of interest to its author and her immediate circle of family/friends and does not convey any useful information to other people who do not know the author. 3. Status updates about individuals or loved ones. |
| Personal update              | Personal only                                | CrisisNLPCF                  | If the tweet conveys some sort of personal opinion, which is not of interest of a general audience. 1. Personal and only useful to a small circle of family/friends of the author; 2. If a message is only of interest to its author and her immediate circle of family/friends and does not convey any useful information to other people who do not know the author. 3. Status updates about individuals or loved ones. |
| Physical landslide           | Physical landslide                           | CrisisNLPCF                  | The item is related to a physical landslide                                                                                                                                                                                                                                 |
| Requests or needs            | Needs of those affected                      | CrisisNLPCF                  | Needs of those affected                                                                                                                                                                                                                                                         |
| Requests or needs            | Requests for help needs                     | CrisisNLPCF                  | Something (e.g. food, water, shelter) or someone (e.g. volunteers, doctors) is needed.                                                                                                                                                                                        |
| Requests or needs            | Urgent needs                                 | CrisisNLPCF                  | Something (e.g. food, water, shelter) or someone (e.g. volunteers, doctors) is needed.                                                                                                                                                                                        |
| Response efforts             | Humanitarian aid provided                    | CrisisNLPCF                  | Affected populations receiving food, water, shelter, medication, etc. from humanitarian/emergency response organizations. All info about responders. Affected populations receiving food, water, shelter, medication, etc. from humanitarian/emergency response organizations.                                                      |
| Response efforts             | Response efforts                             | CrisisNLPCF                  | Response efforts                                                                                                                                                                                                                                                             |
| Sympathy and support         | Sympathy and emotional support               | CrisisNLPCF                  | Sympathy and emotional support 1. Thoughts, prayers, gratitude, sadness, etc.                                                                                                                                                                                                  |
| Sympathy and support         | Sympathy and support                         | CrisisNLPCF                  | Sympathy and emotional support 1. Thoughts, prayers, gratitude, sadness, etc.                                                                                                                                                                                                  |
| Sympathy and support         | Personal updates sympathy support            | CrisisNLPCF                  | Personal updates, sympathy, support                                                                                                                                                                                                                                           |
| Sympathy and support         | Praying                                      | CrisisNLPCF                  | If the author of the tweet prays for the flight MH370 passengers.                                                                                                                                                                                                              |
| Terrorism related information| Terrorism related                            | CrisisNLPCF                  | If the tweet reports possible terrorism act involved.                                                                                                                                                                                                                         |

Table 11: Class label mapping and grouping for CrisisLex, CrisisNLP, ISCRAM2013, and SWDM2013 datasets.
| Original class              | Class label mapping                   |
|----------------------------|---------------------------------------|
| Related                    | Informative                           |
| Aid related                | Requests or needs                     |
| Request                    | Requests or needs                     |
| Offer                      | Donation and volunteering             |
| Medical help               | Requests or needs                     |
| Medical products           | Requests or needs                     |
| Search and rescue          | displaced and evacuations              |
| Security                   |                                       |
| Military                   |                                       |
| Water                      | Requests or needs                     |
| Food                       | Requests or needs                     |
| Shelter                    | Requests or needs                     |
| Clothing                   | Requests or needs                     |
| Money                      | Requests or needs                     |
| Missing people             | Missing and found people              |
| Refugees                   | Requests or needs                     |
| Death                      | Injured or dead people                |
| Other aid                  | Requests or needs                     |
| Infrastructure related     | Infrastructure and Utilities damage    |
| Transport                  | Infrastructure and utilities damage    |
| Buildings                  | Infrastructure and utilities damage    |
| Electricity                | Infrastructure and utilities damage    |
| Hospitals                  | Infrastructure and utilities damage    |
| Shops                      | Infrastructure and utilities damage    |
| Aid centers                | Infrastructure and utilities damage    |
| Other infrastructure       | Infrastructure and Utilities damage    |

Table 12: Class label mapping for Disaster Response Data (DRD).

| Original class | Mapped class |
|----------------|--------------|
| Relevant       | Informative  |
| Not Relevant   | Not informative |

Table 13: Class label mapping for Disasters on Social Media (DSM) dataset.

| Original class          | Class label mapping                  |
|-------------------------|--------------------------------------|
| Affected individuals    | Informative                          |
| Infrastructure and utility damage | Affected individual               |
| Injured or dead people  | Informative                          |
| Missing or found people | Informative                          |
| Not relevant or cant judge | Not informative                      |
| Other relevant information | Other relevant information        |
| Rescue volunteering or donation effort | Donation and volunteering    |
| Vehicle damage          | Informative                          |
|                         | Infrastructure and utilities damage  |

Table 14: Class label mapping for CrisisMMD.
| Original class                                      | Class label mapping                      |
|----------------------------------------------------|------------------------------------------|
| Blocked roads                                      | Informative Infrastructure and utilities damage |
| Blood or other medical supplies needed             | Informative Requests or needs            |
| Building damaged                                   | Informative Infrastructure and utilities damage |
| Camp shelter                                       | Informative Requests or needs            |
| Casualties and damage                              | Informative Infrastructure and utilities damage |
| Caution and advice                                 | Informative Caution and advice           |
| Clothing needed                                    | Informative Requests or needs            |
| Damage                                             | Informative Infrastructure and utilities damage |
| Displaced people                                   | Informative Displaced and evacuations    |
| Donations                                          | Informative Donation and volunteering    |
| Food and or water needed                           | Informative Requests or needs            |
| Food water                                         | Informative Requests or needs            |
| Humanitarian aid provided                          | Informative Response efforts             |
| Informative                                        | Informative Informative                  |
| Infrastructure and utilities                       | Informative Infrastructure and utilities damage |
| Infrastructure damage                              | Informative Infrastructure and utilities damage |
| Injured dead                                       | Informative Injured or dead people       |
| Injured or dead people                             | Informative Injured or dead people       |
| Loss of electricity                                | Informative Infrastructure and utilities damage |
| Loss of internet                                   | Informative Infrastructure and utilities damage |
| Missing trapped or found people                    | Informative Missing and found people     |
| Money                                              | Informative Requests or needs            |
| Money needed                                       | Informative Requests or needs            |
| Needs and requests for help                        | Informative Requests or needs            |
| Non emergency but relevant                         | Informative \*                         |
| None of the above                                  | Not informative Not humanitarian         |
| Not informative                                    | Not informative Not humanitarian         |
| Not relevant or irrelevant                         | Not informative Not humanitarian         |
| Not relevant                                       | Not informative Not humanitarian         |
| Not relevent                                       | Not informative Not humanitarian         |
| Other relevant                                     | Informative Other relevant information   |
| Other relevant information                         | Informative Other relevant information   |
| Other useful for response                          | Informative Other relevant information   |
| Relief and response efforts                         | Informative Requests or needs            |
| Requests for help needs                            | Informative Requests or needs            |
| Response efforts                                    | Informative Requests or needs            |
| Shelter                                            | Informative Requests or needs            |
| Shelter and supplies                               | Informative Requests or needs            |
| Shelter needed                                     | Informative Requests or needs            |
| Shelter or supplies needed                         | Informative Requests or needs            |
| Sympathy and emotional support                     | Informative Sympathy and support         |
| Urgent needs                                       | Informative Requests or needs            |

Table 15: Class label mapping for AIDR system.