CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing

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

Abstract

Time-critical analysis of social media streams is important for humanitarian organizations for planing rapid response during disasters. The *crisis informatics* research community has developed several techniques and systems for processing and classifying big crisis-related data posted on social media. However, due to the dispersed nature of the datasets used in the literature (e.g., for training models), it is not possible to compare the results and measure the progress made towards building better models for crisis informatics tasks. In this work, we attempt to bridge this gap by combining various existing crisis-related datasets. We consolidate eight human-annotated datasets and provide 166.1k and 141.5k tweets for *informativeness* and *humanitarian* classification tasks, respectively. We believe that the consolidated dataset will help train more sophisticated models. Moreover, we provide benchmarks for both binary and multiclass classification tasks using several deep learning architecrures including, CNN, fastText, and transformers. We make the dataset and scripts 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 are 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). Most of these studies use one of the publicly available datasets (Olteanu et al. 2014; Imran, Mitra, and Castillo 2016; Alam et al. 2018, 2020) and either propose a new model or report higher performance of an existing model. Typical classification tasks in the community include (i) *informativeness* (i.e., informative vs. not-informative messages), (ii) *humanitarian information types* (e.g., affected individual reports, infrastructure damage reports), and (iii) *event types* (e.g., flood, earthquake, fire).

Despite the recent focus of the *crisis informatics*[1] research community to develop novel and more robust computational algorithms and techniques to process social media data, we observe several limitations in the current literature. *First*, few efforts have been invested to develop standard datasets (specifically, train/dev/test splits) and benchmarks for the community to compare their results, models, and techniques. *Secondly*, most of the published datasets are noisy, e.g., CrisisLex (Olteanu et al. 2014) contains duplicate and near-duplicate content, which produces misleading classification performance. Moreover, some datasets (e.g., CrisisLex) consist of tweets from several languages without any explicit language tag, to separate the data of a particular language of interest.

To address such limitations, in this paper, we aim to develop a standard social media dataset for disaster response that facilitates comparison between different modeling approaches and encourages the community to streamline their efforts towards a common goal. We consolidate eight publicly available datasets (see Section 3). The resulting dataset is larger in size, has better class distribution compared to the individual datasets, and enables building of robust models that performs better for various tasks (i.e., informativeness and humanitarian) and datasets.

The consolidation of datasets from different sources involves various standardization challenges. One of the challenges is the inconsistent class labels across various data sources. We map the class labels using their semantic meaning—a step performed by domain experts manually.

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

Another challenge 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. In summary, the contributions of this study are as follows:

- We consolidate eight publicly available disaster-related datasets by manually mapping semantically similar class labels, which leads to a larger dataset.
- We carefully cleaned various forms of duplicates, and assigned a language tag to each tweet.
- We provide benchmark results on English tweets set using state-of-the-art machine learning algorithms such as Convolutional Neural Networks (CNN), fastText (Joulin et al. 2017) and pre-trained transformer models (Devlin et al. 2019) for two classifications tasks, i.e., *Informativeness* (binary) and *Humanitarian type* (multi-class) classification.[2] The benchmarking encourages the community towards a 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 versions, *(iii)* a subset of the filtered data used for the classification experiments in this study.

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

2 Related Work

2.1 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. In addition, the issue of exact- and near-duplicate content in existing datasets can lead to misleading performance as mentioned earlier. This problem become more visible while consolidating existing datasets. Alam, Muhammad, and Ferda (2019); Kersten et al. (2019) and Wiegmann et al. (2020) have previously worked in the direction to consolidate social media disaster response data. A major limitation of the work by Alam, Muhammad, and Ferda (2019) is that the issue of duplicate and near-duplicate content have not been addressed when combining the different datasets. This issue resulted in an overlap between train and test sets. In terms of label mapping the work of Alam, Muhammad, and Ferda (2019) is similar to the current study. Kersten et al. (2019) focused only on informativeness[3] classification and combined five different datasets. This study has also not focused on exact- and near-duplicate content, which exist in different datasets. The study in Wiegmann et al. (2020) also compiled existing resources for *disaster event types* (e.g., Flood, Fire) classification, which consists of a total of 123,166 tweets from 46 disasters with 9 disaster types. This is different from our work as we address *informativeness* and *humanitarian* classification tasks. Addressing *disaster event types* classification is beyond the scope of our current study.

A fair comparison of the classification experiment is also difficult with previous studies as their train/dev/test splits are not public, except the dataset by Wiegmann et al. (2020). We address such limitations in this study, i.e., we consolidate the datasets, eliminate duplicates, and release standard dataset splits with benchmark results.

In terms of defining class labels (i.e., tagsets) for crisis informatics, most of the earlier efforts are discussed in (Imran et al. 2015; Temnikova, Castillo, and Vieweg 2015; Stowe et al. 2018; Wiegmann et al. 2020). Various recent studies (Olteanu et al. 2014; Imran, Mitra, and Castillo 2016; Alam et al. 2018; Stowe et al. 2018) use similar definitions for class labels. Unlike them, (Strassel, Bies, and Tracey 2017) defines more fine-grained 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 important for humanitarian aid for disaster response tasks, which are common across the publicly available resources. Some of the real-time applications that are currently using such labels include AIDR (Imran et al. 2014), CREES (Burel and Alani 2018), and TweetTracker (Kumar et al. 2011).

2.2 Classification Algorithms:

Despite the fact that a majority of studies in *crisis informatics* literature employ traditional machine learning algorithms, several recent works explore deep learning algorithms in disaster-related tweet classification tasks. The study of (Nguyen et al. 2017) and (Neppalli, Caragea, and Caragea 2018) performed comparative experiments between different classical and deep learning algorithms including Support Vector Machines, Logistic Regression, Random Forests, Recurrent Neural Networks, 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, Kersten, and Klan 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. 2019), and XLNet (Yang et al. 2019) for different NLP tasks. The

[2] We only focused on two tasks for this study and we aim to address *event types* task in a future study.

[3] Authors used *related* vs. *not-related* in their study.

study by (Jain, Ross, and Schoen-Phelan 2019) and (Wiegmann et al. 2020) investigates these embedding representations for disaster tweet classification tasks.

3 Data Curation

3.1 Data Consolidation

We consolidate eight datasets that were labeled for different disaster response classification tasks and whose labels can be mapped consistently for two tasks: *informativeness* and *humanitarian information type* classification. 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. Below we provide a brief overview of the datasets we used for consolidation.

1. **CrisisLex** is one of the largest publicly-available datasets, 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* binary classification task.
2. **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, Mitra, and Castillo 2016).
3. **SWDM2013** dataset consists of data from two events: *(i)* the Joplin collection contains tweets from the tornado that struck Joplin, Missouri on May 22, 2011; *(ii)* The Sandy collection contains tweets collected from Hurricane Sandy that hit Northeastern US on Oct 29, 2012 (Imran et al. 2013a).
4. **ISCRAM2013** dataset consists of tweets from two different events occurred in 2011 (Joplin 2011) and 2012 (Sandy 2012). Note that this set of tweets are different than SWDM2013 set even though they are collected from same events (Imran et al. 2013b).
5. **Disaster Response Data (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.[4]
6. **Disasters on Social Media (DSM)** dataset comprises 10K tweets collected and annotated with labels *related* vs. *not-related* to the disasters.[5]

[4]https://appen.com/datasets/combined-disaster-response-data/

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

Source	Total	Mapping		Filtering	
		Info	Hum	Info	Hum
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	6,869	6,116
Total	**206,411**	**195,991**	**167,878**	**166,098**	**141,533**

Table 1: Different datasets and their sizes (number of tweets) before and after label mapping and filtering steps. Info: Informativeness, Hum: Humanitarian

7. **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 for this dataset is targeted for three classification tasks: *(i)* informative *vs.* not-informative, *(ii)* humanitarian categories (eight classes) and *(iii)* damage severity assessment.
8. **AIDR** dataset is obtained from the *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.

First part of Table 1 summarizes original sizes of the datasets. The CrisisLex and CrisisNLP datasets are the largest and second-largest datasets, respectively, which are currently widely used in the literature. The SWDM2013 is the smallest set. However, it is one of the earliest datasets used by the crisis informatics community.

3.2 Class Label Mapping

The datasets come with different class labels. We create a set of common class labels by manually mapping semantically similar labels into one cluster. 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 aid*[6] 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 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. The reason is to reduce additional manual effort as it requires relabeling them for multiclass setting. Moreover, many tweets

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

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. We could not map them for the humanitarian task.

3.3 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 (O'Connor, Krieger, and Ahn 2010).[7] 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 (Olteanu et al. 2014) dataset that has been collected during Northern Italy Earthquakes, 2012.[8]

Then, we use a similarity-based approach to remove the near-duplicates. To do this, we first convert the tweets into vectors using bag-of-ngram approach as a vector representation. We use 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* (first tweet in Table 2) 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 checked 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 and manual checking, 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.

We analyzed the data to understand which events and datasets have more exact- and near-duplicates. Figure 2 provides counts for both exact- and near-duplicates for informativeness tweets. In the figure, we report total number (in parenthesis the number represents percentage of reduction) of duplicates (i.e., exact and near) for each dataset. The CrisisLex and CrisisNLP have higher number of duplicates comparatively, however, it is because those two are relatively larger in size. For each of these datasets, we analyzed different events where duplicates appear most. In CrisisLex, the majority of the exact duplicates appear in "Queensland floods (2013)"[9] 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 a similar trend.

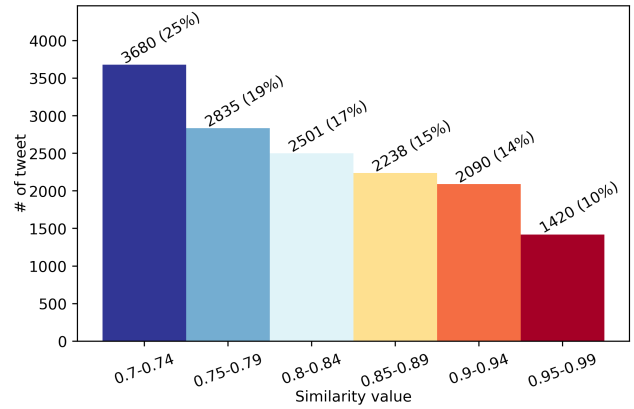

Figure 1: Number of near-duplicates in different bins obtained from consolidated informativeness tweets after label mapping. Tweets will lower similarity (< 0.7) bins are not reported here.

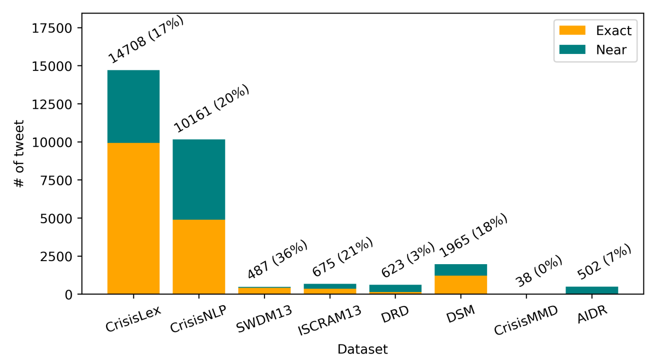

Figure 2: Exact- and near-duplicates in informativeness tweets. Number on top of each bar represents total number, and the number in the parenthesis represents percentage.

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 dupli-

[7] https://github.com/brendano/ark-tweet-nlp

[8] http://en.wikipedia.org/wiki/2012_Northern_Italy_earthquakes

[9] Event name refers to the event during which data has been collected by the respective data authors (see Section 3.1).

#	Tweet	Tokenized	Sim.	Dup.
1	Live coverage: Queensland flood crisis via @Y7News http://t.co/Knb407Fw	live coverage queensland flood crisis via url	0.788	✓
	Live coverage: Queensland flood crisis - Yahoo!7 http://t.co/U2hw0LWW via @Y7News	live coverage queensland flood crisis yahoo url via		
6	Halo tetangga. Sabar ya. RT @AJEnglish: Flood worsens in eastern Australia http://t.co/YfokqBmG	halo tetangga sabar ya rt flood worsens in eastern australia url	0.787	✓
	RT @AJEnglish: Flood worsens in eastern Australia http://t.co/kuGSMCiH	rt flood worsens in eastern australia url		
7	"@guardian: Queensland counts flood cost as New South Wales braces for river peaks http://t.co/MpQskYt1". Brisbane friends moved to refuge.	queensland counts flood cost as new south wales braces for river peaks url brisbane friends moved to refuge	0.778	✓
	Queensland counts flood cost as New South Wales braces for river peaks http://t.co/qb5UuYf9	queensland counts flood cost as new south wales braces for river peaks url		
8	RT @FoxNews: #BREAKING: Numerous injuries reported in large explosion at #Texas fertilizer plant http://t.co/oH93niFiAS". Brisbane friends moved to refuge.	rt breaking numerous injuries reported in large explosion at texas fertilizer plant url	0.744	✓
2	Obama to attend memorial service for victims of Texas explosion: The president will meet with victims of the d... http://t.co/VgGdVATn1b	obama to attend memorial service for victims of texas explosion the president will meet with victims of the d url	0.732	✗
	Obama to attend memorial service for victims of Texas explosion http://t.co/f6JXfzd7QZ	obama to attend memorial service for victims of texas explosion url		

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. The symbol (✓) indicates a duplicate, which we dropped and the symbol (✗) indicates a non-duplicate, which we kept in our dataset.

cates from the consolidated dataset would potentially lead to misleading performance results in the classification experiments.

3.4 Adding Language Tags

Some of the existing datasets contain tweets in various languages (i.e., Spanish and Italian) without explicit assignment of a language tag. In addition, many tweets have code-switched (i.e., multilingual) content. For example, the following tweet has both English and Spanish text: *"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.

We decided to provide a language tag for each tweet if it is not available with the respective dataset. We used the language detection API of Google Cloud Services[10] for this purpose. In Figure 3, we report language distribution for the top nineteen languages consists of more than 20 tweets. 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. Note that most of the non-English tweets appear in the CrisisLex dataset. We believe our language tags will enable future studies to perform multilingual analysis.

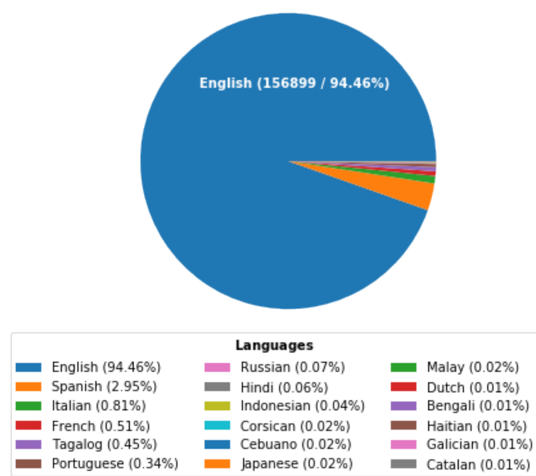

Figure 3: Distribution of top nineteen languages (>= 20 tweets) in the consolidated informativeness tweets.

3.5 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 SWDM2013 and ISCRAM2013 datasets. For the humanitarian task, some class labels are not present in different datasets. Only 17 tweets with the label "terrrorism 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

[10]`https://cloud.google.com/translate/docs/advanced/detecting-language-v3`. Note, it is a paid service, therefore, we have not used this service for the tweets for which language tags are available.

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 (***tweets containing multiple languages***) distribution of ***informativeness*** across different sources.

Class	CrisisLex	CrisisNLP	SWDM13	ISCRAM13	DRD	CrisisMMD	AIDR	Total
Affected individual	3,740	-	-	-	-	471	-	4,211
Caution and advice	1,774	1,137	117	412	-	-	161	3,601
Disease related	-	1,478	-	-	-	-	-	1,478
Displaced and evacuations	-	495	-	-	-	-	50	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
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**	**6,116**	**141,533**

Table 4: Data (***tweets containing multiple languages***) distribution of ***humanitarian*** categories across different datasets.

"not humanitarian" is relatively higher (37.40%) than other class labels. 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.

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 this and next Section 5 we report the class label distribution and results on the selected class labels with English tweets only.

4.1 Models and Architectures

In this section, we describe the details of our classification models. For the experiments, we use CNN, fastText, and pre-trained transformer models.

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).

fastText: For the fastText (Joulin et al. 2017), we used pre-trained embeddings trained on Common Crawl, which is released by fastText for English.

Transformer models: 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 the disaster domain. In this work, we choose the pre-trained models BERT (Devlin et al. 2019), DistilBERT (Sanh et al. 2019), and RoBERTa (Liu et al. 2019) for the classification tasks.

Model Settings: We train the CNN models using the Adam optimizer (Kingma and Ba 2014). The batch size is 128 and maximum number of epochs is set to 1000. We use a filter size of 300 with both window size and pooling length of 2, 3, and 4, and a dropout rate 0.02. We set *early stopping* criterion based on the accuracy of the development set with a patience of 200. For the experiments with fastText, we used default parameters except: *(i)* the dimension is set to 300, *(ii)* minimal number of word occurrences is set to 3, and *(iii)* number of epochs is 50. We train transformers models using the Transformer Toolkit (Wolf et al. 2019). For each model, we use an additional task-specific layer. We fine-tune the model using fine-tuning procedure as prescribed by (Devlin et al. 2019). Due to the instability of the pre-trained models as reported in (Devlin et al. 2019), we perform ten re-runs for each experiment using different seeds, and we select the model that performs best on the dev set. For transformer-based models, we used a learning rate of $2e-5$, and a batch size of 8. More details of the parameters setting can be found in the released scripts.

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**

Table 5: Data split and their distributions with the consolidated ***English*** tweets dataset.

Train	**Test**	**Acc**	**P**	**R**	**F1**
	Informativeness				
CrisisLex (2C)	1. CrisisLex	0.945	0.945	0.950	0.945
	2. CrisisNLP	0.689	0.688	0.690	0.689
	3. Consolidated	0.801	0.807	0.800	0.803
CrisisNLP (2C)	4. CrisisNLP	0.832	0.832	0.830	0.832
	5. CrisisLex	0.712	0.803	0.710	0.705
	6. Consolidated	0.725	0.768	0.730	0.727
Consolidated (2C)	7. CrisisLex	0.943	0.943	0.940	0.943
	8. CrisisNLP	0.829	0.828	0.830	0.828
	9. Consolidated	0.867	0.866	0.870	0.866
	Humanitarian				
CrisisLex (6C)	10. CrisisLex	0.921	0.920	0.920	0.920
	11. CrisisNLP	0.554	0.546	0.550	0.544
	12. Consolidated	0.694	0.601	0.690	0.633
CrisisNLP (10C)	13. CrisisNLP	0.780	0.757	0.780	0.762
	14. CrisisLex	0.769	0.726	0.770	0.714
	15. Consolidated	0.666	0.582	0.670	0.613
Consolidated (11C)	16. CrisisLex	0.908	0.916	0.910	0.912
	17. CrisisNLP	0.768	0.753	0.770	0.753
	18. Consolidated	0.835	0.827	0.840	0.829

Table 6: Classification results using **CNN** for the individual and consolidated datasets. 6C, 10C, and 11C refer to six, ten and eleven class labels respectively.

4.2 Preprocessing and Evaluation

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.

Evaluation: 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.

4.3 Experimental Settings

Individual vs. Consolidated Datasets: The motivation of these experiments is to investigate whether model trained with consolidated dataset generalizes well across different sets. For the individual dataset classification experiments, we selected CrisisLex and CrisisNLP as they are relatively larger 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.

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, Haddow, and Birch (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, Dabre, and Kurohashi 2017; Sajjad et al. 2017). Motivated by it, we propose an event-aware training mechanism. Given a set of m disaster event types $\mathbf{D} = \{d_1, d_2, ..., d_m\}$ where disaster event type d_i includes earthquake, flood, fire, and hurricane. For a disaster event type d_i, $\mathbf{T_i} = \{t_1, t_2, ..., t_n\}$ 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.

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. 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 `UNK` to 5% of the tweets of every disaster. Note that `UNK` is now distributed across all disaster event types and is a good representation of an unknown event.

5 Results and Discussions

5.1 Individual vs. Consolidated Datasets

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. The model trained with individual dataset shows that performance is higher on the corresponding set but low on other sets. For example, for informativeness task, the model trained with CrisisLex performs better on CrisisLex but not on CrisisNLP and Consolidated sets. We see similar pattern for CrisisNLP. However, the model trained with Consolidated data shows similar performance as individual sets (i.e., CrisisLex and CrisisNLP) but higher on consolidated set. A comparison is shown in highlighted rows in the Table 6. 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. This proves that model with consolidated dataset generalizes well, obtaining similar performance on individual sets and higher on the consolidated set.

Between CrisisLex and CrisisNLP, the performance is higher on CrisisLex dataset for both informativeness and humanitarian tasks (1st *vs.* 4th row in Table 6 for the informativeness, and 10th *vs.* 13th row for the humanitarian task in the same table.). This might be due to the CrisisLex dataset being larger than the CrisisNLP dataset. The cross dataset (i.e., train on CrisisLex and evaluate on CrisisNLP) results shows that there is a drop in performance. For example, there is 14.3% difference in F1 on CrisisNLP data using the CrisisLex model for the informativeness task. Similar behavior observed when evaluated the CrisisNLP model on the CrisisLex dataset. 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 of designing 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.

Comparing Models: In Table 7, we report the results using CNN, fastText and transformer based models. We report weighted F1 for all models and tasks. The transformer based models achieve higher performance compared to the CNN and fastText. We used three transformer based models, which varies in the parameter sizes. However, in terms of performance, they are quite similar.

Class-wise Results Analysis: In Table 8, we report class-wise performance of both CNN and BERT models for the humanitarian task. BERT performs better than or on par with CNN across all classes. More importantly, BERT performs substantially better than CNN in the case of minority classes as highlighted in the table. 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".

Train	Test	CNN	FT	BERT	D-BERT	RT
	Informativeness					
CrisisLex (2C)	1. CrisisLex	0.945	0.940	0.949	0.949	0.938
	2. CrisisNLP	0.689	0.687	0.698	0.681	0.694
	3. Consolidated	0.803	0.791	0.806	0.808	0.810
CrisisNLP (2C)	4. CrisisNLP	0.832	0.816	0.833	0.834	0.823
	5. CrisisLex	0.705	0.728	0.749	0.739	0.726
	6. Consolidated	0.727	0.733	0.753	0.755	0.744
Consolidated (2C)	7. CrisisLex	0.943	0.917	0.940	0.938	0.946
	8. CrisisNLP	0.828	0.811	0.825	0.828	0.829
	9. Consolidated	0.866	0.844	0.872	0.870	0.883
	Humanitarian					
CrisisLex (6C)	10. CrisisLex	0.920	0.911	0.934	0.935	0.937
	11. CrisisNLP	0.544	0.549	0.615	0.628	0.632
	12. Consolidated	0.633	0.605	0.766	0.770	0.784
CrisisNLP (10C)	13. CrisisNLP	0.762	0.759	0.791	0.788	0.789
	14. CrisisLex	0.714	0.719	0.842	0.845	0.850
	15. Consolidated	0.613	0.627	0.709	0.707	0.727
Consolidated (11C)	16. CrisisLex	0.912	0.903	0.923	0.921	0.931
	17. CrisisNLP	0.753	0.760	0.786	0.787	0.784
	18. Consolidated	0.829	0.824	0.860	0.856	0.872

Table 7: Classification results (weighted-F1) using CNN, fastText (FT) and transformer based models. D-BERT: DistilBERT, RT: RoBERTa.

	CNN			BERT		
Class	**P**	**R**	**F1**	**P**	**R**	**F1**
Affected individual	0.760	0.720	0.740	0.752	0.808	0.779
Caution and advice	0.630	0.630	0.630	0.675	0.707	0.691
Displaced and evacuations	0.490	0.180	0.260	0.491	0.535	0.512
Donation and volunteering	0.700	0.790	0.740	0.764	0.807	0.785
Infrastructure and utilities damage	0.650	0.660	0.660	0.696	0.717	0.706
Injured or dead people	0.760	0.780	0.770	0.812	0.845	0.828
Missing and found people	0.470	0.170	0.240	0.457	0.466	0.462
Not humanitarian	0.900	0.930	0.920	0.934	0.920	0.927
Requests or needs	0.850	0.840	0.850	0.909	0.901	0.905
Response efforts	0.330	0.070	0.120	0.349	0.308	0.327
Sympathy and support	0.760	0.640	0.690	0.751	0.725	0.738

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

5.2 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 6). 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 tags. For example,

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
fastText	0.824	0.824	0.824	0.824	0.795	0.794	0.795	0.794
BERT	0.872	0.872	0.872	**0.872**	0.865	0.866	0.865	0.865
DistilBERT	0.875	0.874	0.875	**0.874**	0.864	0.863	0.864	0.863
RoBERTa	0.879	0.879	0.879	**0.878**	0.870	0.871	0.870	**0.870**

Table 9: Results of event-aware experiments using the ***consolidated dataset***.

Model	Informativeness				Humanitarian			
	Acc	P	R	F1	Acc	P	R	F1
Monolingual model								
CNN	0.828	0.827	0.828	0.828	0.647	0.650	0.647	0.648
fastText	0.821	0.820	0.821	0.820	0.663	0.662	0.663	0.662
BERT	0.873	0.872	0.873	0.872	0.772	0.771	0.772	0.771
DistilBERT	0.872	0.871	0.872	0.871	0.771	0.770	0.771	0.770
RoBERTa	0.880	0.879	0.880	0.879	0.784	0.785	0.784	0.784
Multilingual model								
BERT-m	0.879	0.879	0.879	0.879	0.781	0.783	0.781	**0.781**
DistilBERT-m	0.873	0.872	0.873	0.872	0.772	0.771	0.772	0.771
XLM-RoBERTa	0.879	0.879	0.879	0.879	0.788	0.789	0.788	0.788

Table 10: Results of consolidated (***multilingual***) datasets (class label distributions are reported in Table 3 and 4) for both tasks and different mono and multilingual models. BERT-m: bert-base-multilingual-uncased, DistilBERT-m: distilbert-base-multilingual-cased

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.

5.3 Discussions

Social media data is noisy and it often poses a challenge for labeling and training classifiers. Our analysis on publicly available datasets reveals that one should 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.

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, Muhammad, and Ferda 2019; Alam et al. 2021). 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, Muhammad, and Ferda 2019), the reported F-measure for informativeness and humanitarian tasks are 0.93 and 0.78, respectively. In (Alam et al. 2021), the best reported F-measure for humanitarian task is 0.781. 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 cleaned from duplicate content, which might have led to misleading performance results otherwise.

Our initial consolidated datasets (i.e., Table 3 and 4) contains multilingual content with more class labels and different types of content (e.g., disease-related), therefore, an interesting future research could be to try different pretrained multilingual models to classify tweets in different languages. We have run a set of preliminary experiments using our initial consolidated datasets, and using monolingual model such as CNN, fastText, BERT, DistilBERT, RoBERTa, and multilingual versions of the mentioned transformer models. The results are reported in Table 10. We observe that performance dropped significantly for the humanitarian task compared to English-only dataset. For example, $\sim$8% drop using BERT model. Note that test set for English tweets (Table 5) is a subset of this set of tweets. From the results of multilingual versions of BERT (BERT-m), we see that there is an increase in performance compared to other mono-lingual models, however, the results are still far below. Such a finding shows an interesting avenue for further research. Another future research direction would be to use multilingual models for the zero-shot classification of tweets.

The competitive performance of transformer based models encourages us to try deeper models such as Google T5 (Raffel et al. 2020). For the transformer 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.

Our released dataset and benchmark results will help the research community to develop better models and compare results. The inclusion of language tags can help to conduct multilingual experiments in future research. The resulting 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, filtering exact- and near-duplicates, and providing benchmarks based on state-of-the-art CNN, FastText, and transformer-based models. Our experimental results and data splits can be useful for future research in order to conduct multilingual studies, developing new models and cross-domain experiments.

Ethics and Broader Impact

The developed consolidated labeled dataset is curated from different publicly available sources. The consolidated labeled dataset can be used to develop models for humanitarian response tasks and can be useful to fast responders. We release the dataset by maintaining the license of existing resources. We try to maintain FAIR principles,[11] and a datasheet (Gebru et al. 2018) will be available on the data package website.

References

Alam, F.; Muhammad, I.; and Ferda, O. 2019. CrisisDPS: Crisis Data Processing Services. In *ISCRAM*.

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

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.; Qazi, U.; Imran, M.; and Ofli, F. 2021. HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. In *ICWSM*.

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

Burel, G.; Saif, H.; Fernandez, M.; and Alani, H. 2017. On Semantics and Deep Learning for Event Detection in Crisis Situations. In *SemDeep at ESWC*.

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

Chu, C.; Dabre, R.; and Kurohashi, S. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *ACL*.

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.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv:1803.09010* .

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 *WWW*, 159–162.

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

Imran, M.; Elbassuoni, S. M.; Castillo, C.; Diaz, F.; and Meier, P. 2013b. Extracting information nuggets from disaster-related messages in social media. In *ISCRAM*.

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*.

[11] http://www.force11.org/group/fairgroup/fairprinciples

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* .

Kruspe, A.; Kersten, J.; and Klan, F. 2019. Detecting Event-Related Tweets by Example using Few-Shot Models. In *ISCRAM*.

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

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* .

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.

O'Connor, B.; Krieger, M.; and Ahn, D. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

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:1802.05365* .

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR* 21(140): 1–67.

Sajjad, H.; Durrani, N.; Dalvi, F.; Belinkov, Y.; and Vogel, S. 2017. Neural Machine Translation Training in a Multi-Domain Scenario. In *IWSLT*.

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

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. In *NAACL*, 35–40.

Stowe, K.; Palmer, M.; Anderson, J.; Kogan, M.; Palen, L.; Anderson, K. M.; Morss, R.; Demuth, J.; and Lazrus, H. 2018. Developing and evaluating annotation procedures for twitter data during hazard events. In *LAW-MWE-CxG*, 133–143.

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.

Temnikova, I. P.; Castillo, C.; and Vieweg, S. 2015. EMTerms 1.0: A Terminological Resource for Crisis Tweets. In *ISCRAM*.

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.

A Appendices

A.1 Events and class label mapping

In Table 11, 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 12, we report class label mapping for ISCRAM2013, SWDM2013, CrisisLex and CrisisNLP datasets. The first column in the table 12 shows the mapped class for both informative and humanitarian tasks. Note that all humanitarian class labels also mapped to informative, and not humanitarian labels mapped to not-informative in the data preparation step. In Table 13, 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 14. The CrisisMMD dataset has been annotated for informativeness and humanitarian task, therefore, very minor label mapping was needed as shown in Table in 15. 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 16.

A.2 Examples of Tweets with Similarity

In Table 17, we report example tweets with different similarity values to justify the selection of similarity threshold. We have chosen a value of > 0.75 to filter duplicate tweets.

B Experimental Parameters

In this section, we report parameters for CNN and BERT model. In addition, we discuss the computing infrastructures that we used for the experiments.

B.1 CNN Parameters

- Batch size: 128
- Filter size of 300
- Window size of 2, 3, and 4
- Pooling length of 2, 3, and 4
- Number of epochs: 3000
- Max seq length: 60
- Patience for early stopping: 200
- Learning rate (Adam): 1.0E-05
-

B.2 FastText Parameters

We mostly used default parameters as can be found int our released package.

- Embedding dimension: 300
- Minimal number of word occurrences: 3
- Number of epochs: 50

B.3 Transformer Models' Parameters

Below we list the hyperparameters that we used for training across all transformer based models. All experimental scripts will be publicly Hyper-parameters include:

- Batch size: 8
- Number of epochs: 10
- Max seq length: 128
- Learning rate (Adam): 2e-5

Number of parameters:

- **BERT** (bert-base-uncased): L=12, H=768, A=12, total parameters = 110M; where L is number of layers (i.e., Transformer blocks), H is the hidden size, and A is the number of self-attention heads.
- **DistilBERT** (distilbert-base-uncased): it is a distilled version of the BERT model consists of 6-layer, 768-hidden, 12-heads, 66M parameters.
- **RoBERTa** (roberta-large): it is using the BERT-large architecture consists of 24-layer, 1024-hidden, 16-heads, 355M parameters.

B.4 Detail Results

In Table 18 and 19, we provide detail results for different datasets (English Tweets) with different models.

Dataset	Year	Event name
ISCRAM2013		
ISCRAM2013	2011	Joplin
SWDM2013		
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	Philipinnes Floods
CrisisLexT26	2012	Philippines Typhoon Pablo
CrisisLexT26	2012	Venezuela Refinery Explosion
CrisisLexT26	2012	Guatemala Earthquake
CrisisLexT26	2012	Colorado Wildfires
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	Phillipines Typhoon Yolanda
CrisisLexT26	2013	Queensland Floods
CrisisLexT26	2013	Singapore haze
CrisisLexT26	2013	West-Texas explosion
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 11: 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.
✗	Animal management	CrisisNLP-volunteers	Pets and animals, living, missing, displaced, or injured/dead
Caution and advice	Caution and advice	CrisisLexT26	If a message conveys/reports information about some warning or a piece of advice about a possible hazard of an incident.
Disease related	Disease signs or symptoms	CrisisNLP-CF	Reports of symptoms such as fever, cough, diarrhea, and shortness of breath or questions related to these symptoms.
Disease related	Disease transmission	CrisisNLP-CF	Reports of disease transmission or questions related to disease transmission
Disease related	Disease Treatment	CrisisNLP-CF	Questions or suggestions regarding the treatments of the disease.
Disease related	Disease Prevention	CrisisNLP-CF	Questions or suggestions related to the prevention of disease or mention of a new prevention strategy.
Disease related	Disease Affected people	CrisisNLP-CF	Reports of affected people due to the disease
Displaced and evacuations	Displaced people	CrisisNLP-volunteers	People who have relocated due to the crisis, even for a short time (includes evacuations)
Displaced and evacuations	Displaced people and evacuations	CrisisNLP-CF	People who have relocated due to the crisis, even for a short time (includes evacuations)
Donation and volunteering	Donation needs or offers or volunteering services	CrisisNLP-CF	Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services
Donation and volunteering	Donations and volunteering	CrisisLexT26	Needs, requests, or offers of money, blood, shelter, supplies, and/or services by volunteers or professionals.
Donation and volunteering	Donations of money	CrisisNLP-volunteers	Donations of money
Donation and volunteering	Donations of money goods or services	SWDM2013/ISCRAM2013	If a message speaks about money raised, donation offers, goods/services offered or asked by the victims of an incident.
Donation and volunteering	Donations of supplies and or volunteer work	CrisisNLP-volunteers	Donations of supplies and/or volunteer work
Donation and volunteering	Money	CrisisNLP-volunteers	Money requested, donated or spent
Donation and volunteering	Shelter and supplies	CrisisNLP-volunteers	Needs or donations of shelter and/or supplies such as food, water, clothing, medical supplies or blood
Donation and volunteering	Volunteer or professional services	CrisisNLP-volunteers	Services needed or offered by volunteers or professionals
Informative	Informative	CrisisNLP-CF	2014 Iceland Volcano en, 2014 Malaysia Airline MH370 en
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 taking place.
Informative	Informative direct or 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.
Informative	Informative indirect	SWDM2013/ISCRAM2013	If the message is of interest to other people beyond the author's immediate circle, and seems to be seen/heard by the person on the radio, TV, newspaper, or other source. The message must specify the source.
Informative	related and informative	CrisisLexT26	Related to the crisis and informative: if it contains useful information that helps understand the crisis situation.
Infrastructure and utilities damage	Infrastructure damage	CrisisNLP-volunteers	Houses, buildings, roads damaged or utilities such as water, electricity, interrupted
Infrastructure and utilities damage	Infrastructure and utilities	CrisisNLP-volunteers	Buildings or roads damaged or operational; utilities/services interrupted or restored
Infrastructure and utilities damage	Infrastructure	CrisisNLP-volunteers	Infrastructure
Infrastructure and utilities damage	Infrastructure and utilities damage	CrisisNLP-CF	Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored.
Injured or dead people	Injured or dead people	CrisisNLP-CF	Reports of casualties and/or injured people due to the crisis.
Injured or dead people	Injured and dead	CrisisNLP-volunteers	Injured and dead
Injured or dead people	Deaths reports	CrisisNLP-CF	Injured and dead
Injured or dead people	Casualties and damage	SWDM2013/ISCRAM2013	If a message reports the information about casualties or damage done by an incident.
Missing and found people	Missing trapped or found people	CrisisNLP-volunteers	Missing, trapped, or found people—Questions and/or reports about missing or found people.
Missing and found people	People Missing or found	CrisisNLP-volunteers	People missing or found.
Missing and found people	People Missing found or seen	CrisisNLP-volunteers	If a message reports about the missing or found person effected by an incident or seen a celebrity visit on ground zero.
Not humanitarian	Not applicable	CrisisLexT26	Not applicable
Not humanitarian	Not related to crisis	CrisisNLP-volunteers	Not related to this crisis
Not humanitarian	Not informative	CrisisNLP-volunteers, CrisisLexT26	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 labeled	CrisisLexT26	Not labeled
Not humanitarian	Not related or irrelevant	CrisisNLP-CF, CrisisNLP-volunteers	1. Not related or irrelevant; 2. Unrelated to the situation or irrelevant
Not humanitarian	Not related to the crisis	CrisisNLP-volunteers	Not related to crisis
Not humanitarian	Not relevant	CrisisLexT26	Not relevant
Not humanitarian	Off-topic	CrisisLexT6;	Off-topic
Not humanitarian	Other	CrisisNLP-volunteers	if the message is not in English, or if it cannot be classified.
Not humanitarian	Not related	CrisisLexT26	Not related
Not humanitarian	Not physical landslide	CrisisNLP-volunteers	The item does not refer to a physical landslide
Not humanitarian	Terrorism not related	CrisisNLP-volunteers	If the tweet is not about terrorism related to the flight MH370
Other relevant information	Other relevant information	CrisisNLP-volunteers	1. Other useful information that helps understand the situation; 2. Informative for emergency/humanitarian response, but in none of the above categories, including weather/evacuations/etc.
Other relevant information	Other relevant	CrisisNLP-volunteers	1. Other useful information that helps understand the situation; 2. Informative for emergency/humanitarian response, but in none of the above categories, including weather/evacuations/etc.
Other relevant information	Other useful information	CrisisLexT26	1. Other useful information not covered by any of the following categories: affected individuals, infrastructure and utilities, donations and volunteering, caution and advice, sympathy and emotional support.
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; CrisisNLP	Relevant
Personal update	Personal	CrisisNLP-volunteers	If the tweet conveys some sort of personal opinion, which is not of interest of a general audience.
Personal update	Personal only	CrisisNLP-volunteers	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.
Personal update	Personal updates	CrisisNLP-volunteers	1. Status updates about individuals or loved ones.
Physical landslide	Physical landslide	CrisisNLP-volunteers	The item is related to a physical landslide
Requests or needs	Needs of those affected	CrisisNLP-volunteers	Needs of those affected
Requests or needs	Requests for help needs	CrisisNLP-volunteers	Something (e.g. food, water, shelter) or someone (e.g. volunteers, doctors) is needed.
Requests or needs	Urgent needs	CrisisNLP-volunteers	Something (e.g. food, water, shelter) or someone (e.g. volunteers, doctors) is needed.
Response efforts	Humanitarian aid provided	CrisisNLP-volunteers	Affected populations receiving food, water, shelter, medication, etc. from humanitarian/emergency response organizations.
Response efforts	Response efforts	CrisisNLP-volunteers	All info about responders. Affected populations receiving food, water, shelter, medication, etc. from humanitarian/emergency response organizations.
Sympathy and support	Sympathy and emotional support	CrisisNLP-volunteers	Sympathy and emotional support
Sympathy and support	Sympathy and support	CrisisLexT26	1. Thoughts, prayers, gratitude, sadness, etc.
Sympathy and support	Personal updates sympathy support	CrisisNLP-volunteers	Personal updates, sympathy, support.
Sympathy and support	Praying	CrisisNLP-volunteers	If author of the tweet prays for flight MH370 passengers.
Terrorism related information	Terrorism related	CrisisNLP-volunteers	If the tweet reports possible terrorism act involved.

Table 12: Class label mapping and grouping for CrisisLex, CrisisNLP, ISCRAM2013, and SWDM2013 datasets. The symbol (✗) indicates we do not map the tweets with that label for this study.

Original class	Class label mapping	
	Informative	**Humanitarian**
Related	Informative	✗
Aid related	Informative	Requests or needs
Request	Informative	Requests or needs
Offer	Informative	Donation and volunteering
Medical help	Informative	Requests or needs
Medical products	Informative	requests or needs
Search and rescue	Informative	displaced and evacuations
Security	✗	✗
Military	✗	✗
Water	Informative	Requests or needs
Food	Informative	Requests or needs
Shelter	Informative	Requests or needs
Clothing	Informative	Requests or needs
Money	Informative	Requests or needs
Missing people	Informative	Missing and found people
Refugees	Informative	Requests or needs
Death	Informative	Injured or dead people
Other aid	Informative	Requests or needs
Infrastructure related	Informative	Infrastructure and Utilities damage
Transport	Informative	Infrastructure and utilities damage
Buildings	Informative	Infrastructure and utilities damage
Electricity	Informative	Infrastructure and utilities damage
Hospitals	Informative	Infrastructure and utilities damage
Shops	Informative	Infrastructure and utilities damage
Aid centers	Informative	Infrastructure and utilities damage
Other infrastructure	Informative	Infrastructure and Utilities damage

Table 13: Class label mapping for Disaster Response Data (DRD). The symbol (✗) indicates we do not map the tweets with that label for this study.

Original class	Mapped class
Relevant	Informative
Not Relevant	Not informative

Table 14: Class label mapping for Disasters on Social Media (DSM) dataset.

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

Table 15: Class label mapping for CrisisMMD.

Original class	Class label mapping	
	Informative	**Humanitarian**
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 related 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 16: Class label mapping for AIDR system.

#	Tweet	Tokenized	Sim.	Dup.
1	RT @rosemaryCNN: As flood waters recede in Qld, #Australia, attention turns 2 relief & recovery. Police reportedly find a 5th victim ...	rt as flood waters recede in qld australia attention turns relief recovery police reportedly find a th victim	0.882	✓
	As flood waters recede in Qld, #Australia, attention turns 2 relief & recovery. Police reportedly find a 5th victim in a car #CNN	as flood waters recede in qld australia attention turns relief recovery police reportedly find a th victim in a car cnn		
2	Queensland counts flood cost as New South Wales braces for river peaks - The Guardian: The GuardianQueensland co... http://t.co/PyGhSzbG	queensland counts flood cost as new south wales braces for river peaks the guardian the guardianqueensland co url	0.856	✓
	Queensland counts flood cost as New South Wales braces for river peaks - The Guardian http://t.co/njADhrdc #News	queensland counts flood cost as new south wales braces for river peaks the guardian url news		
3	He's no Anna Bligh! @abcnews LIVE: Queensland Premier Campbell Newman is giving an update on Queensland flood crisis http://t.co/pXxoxLOe	he 's no anna bligh live queensland premier campbell newman is giving an update on queensland flood crisis url	0.808	✓
	AUSTRALIA: RT @abcnews: LIVE: Queensland Premier Campbell Newman is giving an update on Queensland flood crisis http://t.co/Jj9S057T	australia rt live queensland premier campbell newman is giving an update on queensland flood crisis url		
4	Australia lurches from fire to flood http://t.co/C6x8Uxnk	australia lurches from fire to flood url	0.807	✓
	Australia lurches from fire to flood #climatechange #globalwarming http://t.co/MZa6H3QC	australia lurches from fire to flood climatechange globalwarming url		
5	Live coverage: Queensland flood crisis via @Y7News http://t.co/Knb407Fw	live coverage queensland flood crisis via url	0.788	✓
	Live coverage: Queensland flood crisis - Yahoo!7 http://t.co/U2hw0LWW via @Y7News	live coverage queensland flood crisis yahoo url via		
6	Halo tetangga. Sabar ya. RT @AJEnglish: Flood worsens in eastern Australia http://t.co/YfokqBmG	halo tetangga sabar ya rt flood worsens in eastern australia url	0.787	✓
	RT @AJEnglish: Flood worsens in eastern Australia http://t.co/kuGSMCiH	rt flood worsens in eastern australia url		
7	"@guardian: Queensland counts flood cost as New South Wales braces for river peaks http://t.co/MpQskYt1". Brisbane friends moved to refuge.	queensland counts flood cost as new south wales braces for river peaks url brisbane friends moved to refuge	0.778	✓
	Queensland counts flood cost as New South Wales braces for river peaks http://t.co/qb5UuYf9	queensland counts flood cost as new south wales braces for river peaks url		
8	RT @FoxNews: #BREAKING: Numerous injuries reported in large explosion at #Texas fertilizer plant http://t.co/oH93niFiAS". Brisbane friends moved to refuge.	rt breaking numerous injuries reported in large explosion at texas fertilizer plant url	0.744	✗
	Numerous injuries reported in large explosion at Texas fertilizer plant: DEVELOPING: Emergency crews in Texas ... http://t.co/Th5Yzvdg5m	numerous injuries reported in large explosion at texas fertilizer plant developing emergency crews in texas url		
9	Obama to attend memorial service for victims of Texas explosion: The president will meet with victims of the d... http://t.co/VgGdVATn1b	obama to attend memorial service for victims of texas explosion the president will meet with victims of the d url	0.732	✗
	Obama to attend memorial service for victims of Texas explosion http://t.co/f6JXfzd7QZ	obama to attend memorial service for victims of texas explosion url		
10	RT @RobertTaylors: Shooting Reported at Los Angeles International Airport: There are reports of a shooting incident Friday mornin... http:/…	rt shooting reported at los angeles international airport there are reports of a shooting incident friday mornin http …	0.705	✗
	RT @BuzzFeed: There Are Reports Of A Shooting At Los Angeles International Airport http://t.co/9TgunRXajQ	rt there are reports of a shooting at los angeles international airport url		
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	✗
	Hurricane Sandy view from the top of the NYTimes building http://t.co/pLiXlaHI	hurricane sandy view from the top of the nytimes building url		

Table 17: 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.

Train	Test	Acc	P	R	F1	Train	Test	Acc	P	R	F1
		CNN						**FastText**			
CrisisLex	CrisisLex	0.945	0.945	0.950	0.945	CrisisLex	CrisisLex	0.940	0.940	0.940	0.940
CrisisLex	CrisisNLP	0.689	0.688	0.690	0.689	CrisisLex	CrisisNLP	0.685	0.694	0.685	0.687
CrisisLex	Consolidated	0.801	0.807	0.800	0.803	CrisisLex	Consolidated	0.789	0.803	0.789	0.791
CrisisNLP	CrisisNLP	0.832	0.832	0.830	0.832	CrisisNLP	CrisisNLP	0.817	0.816	0.817	0.816
CrisisNLP	CrisisLex	0.712	0.803	0.701	0.705	CrisisNLP	CrisisLex	0.729	0.781	0.729	0.728
CrisisNLP	Consolidated	0.725	0.768	0.730	0.727	CrisisNLP	Consolidated	0.731	0.754	0.731	0.733
Consolidated	CrisisLex	0.943	0.943	0.940	0.943	Consolidated	CrisisLex	0.918	0.918	0.918	0.917
Consolidated	CrisisNLP	0.829	0.828	0.830	0.828	Consolidated	CrisisNLP	0.812	0.811	0.812	0.811
Consolidated	Consolidated	0.867	0.866	0.870	0.866	Consolidated	Consolidated	0.845	0.844	0.845	0.844
		BERT						**RoBERTa**			
CrisisLex	CrisisLex	0.949	0.949	0.949	0.949	CrisisLex (6C)	CrisisLex	0.938	0.939	0.938	0.938
CrisisLex	CrisisNLP	0.704	0.703	0.704	0.698	CrisisLex	CrisisNLP	0.724	0.728	0.724	0.715
CrisisLex	Consolidated	0.806	0.808	0.806	0.806	CrisisLex	Consolidated	0.803	0.802	0.803	0.802
CrisisNLP	CrisisNLP	0.834	0.834	0.834	0.833	CrisisNLP (10C)	CrisisNLP	0.824	0.823	0.824	0.823
CrisisNLP	CrisisLex	0.752	0.819	0.752	0.749	CrisisNLP	CrisisLex	0.731	0.811	0.731	0.726
CrisisNLP	Consolidated	0.751	0.782	0.751	0.753	CrisisNLP	Consolidated	0.742	0.769	0.742	0.744
Consolidated	CrisisLex	0.940	0.940	0.940	0.940	Consolidated	CrisisLex	0.946	0.946	0.946	0.945
Consolidated	CrisisNLP	0.825	0.825	0.825	0.825	Consolidated	CrisisNLP	0.831	0.832	0.831	0.830
Consolidated	Consolidated	0.872	0.872	0.872	0.872	Consolidated (11C)	Consolidated	0.884	0.883	0.884	0.883
		DistilBERT									
CrisisLex	CrisisLex	0.949	0.950	0.949	0.949						
CrisisLex	CrisisNLP	0.691	0.693	0.691	0.681						
CrisisLex	Consolidated	0.808	0.808	0.808	0.808						
CrisisNLP	CrisisNLP	0.834	0.834	0.834	0.834						
CrisisNLP	CrisisLex	0.743	0.818	0.743	0.739						
CrisisNLP	Consolidated	0.752	0.783	0.752	0.755						
Consolidated	CrisisLex	0.938	0.938	0.938	0.938						
Consolidated	CrisisNLP	0.829	0.828	0.829	0.828						
Consolidated	Consolidated	0.871	0.870	0.871	0.87						

Table 18: Results of ***Informativeness*** task using different models on ***consolidated English Tweets***.

Train	Test	Acc	P	R	F1	Train	Test	Acc	P	R	F1
		CNN						**FastText**			
CrisisLex (6C)	CrisisLex	0.921	0.920	0.920	0.920	CrisisLex (6C)	CrisisLex	0.914	0.909	0.914	0.911
CrisisLex	CrisisNLP	0.554	0.546	0.550	0.544	CrisisLex	CrisisNLP	0.578	0.543	0.578	0.549
CrisisLex	Consolidated	0.694	0.601	0.690	0.633	CrisisLex	Consolidated	0.670	0.578	0.670	0.605
CrisisNLP (10C)	CrisisNLP	0.780	0.757	0.780	0.762	CrisisNLP (10C)	CrisisNLP	0.774	0.756	0.774	0.759
CrisisNLP	CrisisLex	0.769	0.726	0.770	0.714	CrisisNLP	CrisisLex	0.773	0.729	0.773	0.719
CrisisNLP	Consolidated	0.666	0.582	0.670	0.613	CrisisNLP	Consolidated	0.690	0.616	0.690	0.627
Consolidated	CrisisLex	0.908	0.916	0.910	0.912	Consolidated	CrisisLex	0.901	0.906	0.901	0.903
Consolidated	CrisisNLP	0.768	0.753	0.770	0.753	Consolidated	CrisisNLP	0.769	0.759	0.769	0.760
Consolidated (11C)	Consolidated	0.835	0.827	0.840	0.829	Consolidated (11C)	Consolidated	0.830	0.821	0.830	0.824
		BERT						**RoBERTa**			
CrisisLex (6C)	CrisisLex	0.934	0.935	0.934	0.934	CrisisLex (6C)	CrisisLex	0.936	0.938	0.936	0.937
CrisisLex	CrisisNLP	0.567	0.757	0.567	0.615	CrisisLex	CrisisNLP	0.588	0.753	0.588	0.632
CrisisLex	Consolidated	0.757	0.794	0.757	0.766	CrisisLex	Consolidated	0.775	0.810	0.775	0.784
CrisisNLP (10C)	CrisisNLP	0.793	0.79	0.793	0.791	CrisisNLP (10C)	CrisisNLP	0.789	0.790	0.789	0.789
CrisisNLP	CrisisLex	0.855	0.863	0.855	0.842	CrisisNLP	CrisisLex	0.856	0.876	0.856	0.850
CrisisNLP	Consolidated	0.738	0.717	0.738	0.709	CrisisNLP	Consolidated	0.746	0.748	0.746	0.727
Consolidated	CrisisLex	0.920	0.927	0.920	0.923	Consolidated	CrisisLex	0.927	0.934	0.927	0.931
Consolidated	CrisisNLP	0.785	0.787	0.785	0.786	Consolidated	CrisisNLP	0.784	0.785	0.784	0.784
Consolidated (11C)	Consolidated	0.859	0.861	0.859	0.860	Consolidated (11C)	Consolidated	0.871	0.873	0.871	0.872
		DistilBERT									
CrisisLex (6C)	CrisisLex	0.935	0.935	0.935	0.935						
CrisisLex	CrisisNLP	0.579	0.754	0.579	0.628						
CrisisLex	Consolidated	0.763	0.792	0.763	0.770						
CrisisNLP (10C)	CrisisNLP	0.791	0.786	0.791	0.788						
CrisisNLP	CrisisLex	0.860	0.868	0.860	0.845						
CrisisNLP	Consolidated	0.738	0.730	0.738	0.707						
Consolidated	CrisisLex	0.918	0.924	0.918	0.921						
Consolidated	CrisisNLP	0.791	0.786	0.791	0.787						
Consolidated (11C)	Consolidated	0.857	0.856	0.857	0.856						

Table 19: Results of ***Humanitarian*** task using different models on ***consolidated English Tweets***.