Reframing Crisis Information Extraction as a Sequence Tagging Objective with Augmented Twitter Data

Sam Showalter
UC, Irvine
showalte@uci.edu

Edgar Robles
UC, Irvine
roblesee@uci.edu

Preethi Seshadri
UC, Irvine
preethis@uci.edu

Abstract

With the advent of global communication, it has become standard to disseminate information about a crisis immediately on social media, particularly Twitter. This information is decentralized, uncurated, and disorganized, which can lead to confusion and panic. Organizing tweets into humanitarian categories is essential for ensuring a timely and thorough response to tragedy as many aid groups often align to specific humanitarian needs (e.g. injury, property damage, rescue, etc.). This crucial step in crisis response – extracting and organizing information to inform aid groups and gather resources – can be automated with Natural Language Processing (NLP). In comparison to previous work, we take a novel approach to extracting Twitter information by framing the objective as sequence tagging, where a model assigns each token its most relevant humanitarian label. In doing so, we confer two beneficial properties. Since tagging is done per token, information on a crisis can be organized at a more granular level. Moreover, this approach is not reliant on a particular lexicon or form, and can be generalized to new texts including news, blogs, and other sources. However, the only crisis data currently available tags entire sequences with a single label. We circumvent this issue by defining a novel data augmentation approach to convert our problem into a sequence tagging scheme, and quantitatively validate its efficacy on HumAID, the largest dataset of labeled crisis tweets ever collected. Our code can be found here: https://github.com/samshowalter/nltweetrelief

1 Introduction

Swift and comprehensive response to disaster situations is crucial for maintaining safety in society. However, it can be difficult to quickly understand the full extent of a disaster, as the severity may not be immediately known. Compounding this effect, the most current information on a crisis is usually posted to Twitter, often with little organization or structure. However, progress in deep learning and NLP has enabled scientists to categorize tweets to improve disaster response. In particular, connecting crisis information on social media to humanitarian causes is an ongoing body of research.

Nevertheless, no researchers have framed this problem as sequence tagging and instead generalize humanitarian semantics across an entire passage, sentence, or tweet. This is an incorrect assumption that can lead to information loss and confusion. Humanitarian topics are often mentioned together in crisis messages, and categorizing an entire phrase with one label does not adequately separate this information. For example, the phrase "*Our hearts go out to those affected by the fire that has injured 12 citizens; several people are still missing and we will begin a search.*" includes three humanitarian labels as defined by our dataset HumAID - sympathy, injury/death, and missing people. Most crisis systems today would consider the phrase as a whole and fail to identify and separate the diversity of information. This is due in large part to a lack of crisis datasets that model information extraction as a sequence tagging objective. This gap should be filled; as shown in the example, humanitarian topics are correlated in most crisis messages, making organization of information difficult and topics entangled. Fortunately, HumAID provides an excellent opportunity to fill this gap programmatically. The dataset consists of over 40,000 tweets and is carefully curated for consistency, since each tweet primarily aligns with a single humanitarian cause. With data augmentation, we can alter HumAID to provide a sequence tagging dataset. Our approach does not negate the need for a human-labeled sequence tagging dataset of crisis information, but rather illustrates a cheap method of generating a proxy. In turn, our contributions are as follows:

1. Formalize crisis information extraction as a sequence tagging objective and justify the need for this more granular approach
2. Define a novel data augmentation algorithm that converts single-label crisis tweets into a format conducive for sequence tagging
3. Verify SOTA NLP models effectively generalize from augmented datasets with bootstrapped uncertainty bounds to new crises
4. Qualitatively and quantitatively verify that crisis sequence taggers trained from Twitter can transfer effectively to other lexicons

2 Related Work

Multilabel classification has been previously explored to extract crisis information (Schulz et al., 2014), but not in the context of using modern language models or sequence tagging. Unsupervised information extraction from crisis tweets has also been applied to rank their importance (Interdonato et al., 2018), but not necessarily to categorize them by humanitarian topic. This rank was also applied to the entire tweet and did not identify the most salient tokens. Effective methods for assessing tweet relevance to a crisis have recently been developed as well (Kruspe et al., 2020), but are not useful in organizing information to facilitate a response. Mapping crises with Twitter's geolocation data (Middleton et al., 2014) takes a spatial approach to this problem, which has been included in holistic report creation from summarized text (Di Corso et al., 2017). Unfortunately, these reports attempt to encompass all information and do not separate it along humanitarian topics, which makes it more difficult for agencies to ensure a swift response. For example, a deluge of information on missing persons is not relevant for an agency that specializes in infrastructure restoration. Access to this information can become a hindrance if it obscures the most salient information an agency needs. Instead, effective organization of information by humanitarian topic may facilitate an accelerated and thorough response.

3 Experimental Design

3.1 Dataset

HumAID (Firoj Alam, 2021) is the largest collection of disaster related tweets, containing 19 major natural disaster events (e.g. Ecuador earthquake, Hurricane Florence, California wildfires, etc.) that occurred between 2016-2019. For each tweet, it also includes annotations of humanitarian categories (e.g. loss of life, injury, property damage, etc.), one per tweet. This dataset was recently released on April 8th, 2021 and represents the most comprehensive collection of crisis text.

3.2 Data Augmentation

It is common for news updates about a crisis to discuss multiple humanitarian categories together. Ideally, an information extraction system would be able to label the portions of text corresponding to each category at the lowest granularity possible. Instead of attempting to classify tweets in a unilabel or multilabel objective, we feel an optimal system should be designed as a sequence tagger.

We believe that the release of the HumAID dataset presents an opportunity to develop a sequence tagging NLP system that can achieve these goals. It is important to note each tweet in the dataset belongs to a specific disaster and a single humanitarian category. Since each tweet in our dataset is aligned with a single humanitarian label, we define a data augmentation approach that samples sets of tweets within a given crisis of varying size. With a varying number of randomly sampled tweets, we then create *passages* – collections of tweet sequences from a single crisis concatenated together – that serve as our augmented training samples. For each tweet, the label is broadcasted across the sequence. This makes the large assumption that each tweet is semantically homogeneous relative to its label. While HumAID was curated meticulously, we find that this is not always the case. However, when trained on a large, augmented dataset, our models are not greatly affected by this assumption.

After training this data on all disasters up to 2019, we evaluate the system on generated *passages* for disasters occurring from 2019 onward. There are several benefits to this data augmentation objective, including a training dataset of virtually unlimited size. Additionally, we feel that training models on a dataset of sequence-tagged passages will allow for more complex, real-world use cases.

In our approach, we only ensure each passage was drawn from the same crisis. Therefore, one limitation is that generated passages may sometimes lack natural flow and cohesiveness, since they are synthesized. However, crisis information often appears in this form anyway, due to the rushed nature of the release.

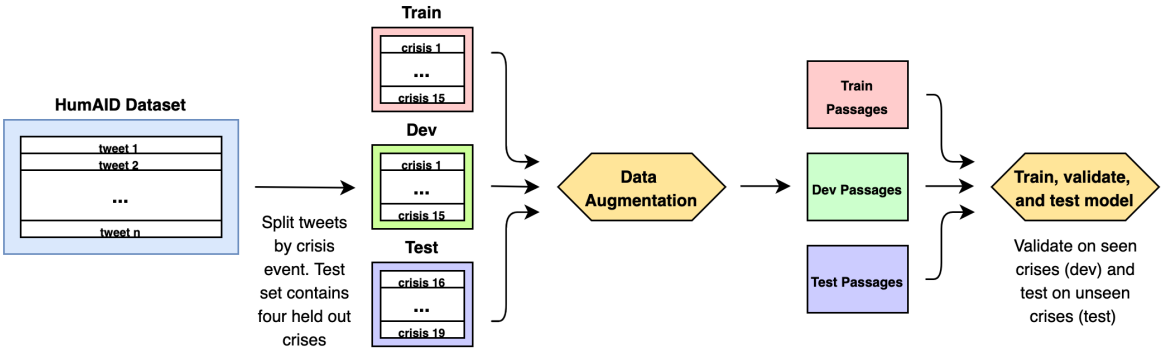

Figure 1: Conceptual overview of experimental pipeline, including data augmentation, training, and testing

Algorithm 1 Data Augmentation Algorithm

Init $\mathcal{T}_c$ = Tweet set for crisis $c, \forall c \in \mathcal{C}$
`tweet_num` $\sim U\{2, ..., k\}$
`crisis_event` $\sim U\{1, ..., |\mathcal{C}|\}$
Init `batch_size`, `num_batches`

for i in `num_batches` **do**
  for j in `batch_size` **do**
    Sample $t \sim$ `tweet_num`
    Sample $c \sim$ `crisis_event`
    Sample $\{w_{0c}, w_{1c}, ..., w_{tc}\} \sim \mathcal{T}_c$
    for w_{jc} $\forall j \in \{0, ..., t\}$ **do**
      $\mathbf{w_{jc}} = \texttt{tokenize}(w_{jc})$
      $\ell_{\mathbf{jc}} = \texttt{broadcast}(\ell_{jc} \times |\mathbf{w_{jc}}|)$
    end for
  end for
  $\mathbf{b_i} = \texttt{Cat}((\mathbf{w_{jc}}, \ell_{\mathbf{jc}}))$ $\forall j$
end for
$\mathbf{B} = \texttt{Cat}(\mathbf{b_i})$ $\forall i$
return $\mathbf{B}$

3.3 Processing Tweets

Compared to literary texts or news articles, tweets require nuanced tokenization. We leveraged NLTK's tweet tokenizer with a few additional processing steps to better handle Twitter-specific content (e.g. preserving hashtags, removing mentions, formatting numbers, encoding emojis, and removing unknown characters/symbols).

3.4 Evaluation Plan

One way to frame this problem on augmented data is through multilabel classification, where we predict the various humanitarian categories present in a *passage*. While multilabel classification helps with identifying the salient themes as a whole, it does not help with the task of explicitly separating crisis information by topic. Instead, sequence tagging is a more applicable and promising approach. For a given sequence of tokens, $x = x_1, ..., x_n$, sequence tagging predicts a sequence of labels of the same length, $y = y_1, ..., y_n$, where $y_i \in \{1...L\}$ are the labels of interest. In our case, each word in a passage is tagged with a single humanitarian label corresponding to the humanitarian category of the original tweet the word belongs to, and our goal is to predict these tags. This approach can be used to disentangle topics and by doing so should ideally recover the original, independent tweets present in a passage. While this might seem counterintuitive or circular, it effectively converts our problem into sequence tagging. Most importantly, with an effectively pretrained model like BERT we show that we can generalize from this contrived sequence data to learn intra-sequence semantic shifts in unseen messages, and effectively extract them.

Furthermore, we qualitatively analyze how well these models adapt to new lexicons when applied to longer articles. This evaluation is particularly important for two reasons. First, generalization on past crises does not imply effective future prediction. Therefore, it is essential that our model can generalize beyond its training crisis set, which we test with four held-out crises from 2019. It should also be able to generalize to unseen information from new lexicons beyond Twitter, which we evaluate using curated examples.

4 Results and Discussion

4.1 Model Training and Evaluation

Effective generalization is particularly difficult to achieve in sequence tagging due to the large variation present in the input. This difficulty is compounded as the number of potential tags grows. For

	Precision	Recall	Accuracy	F1
AlBERT	55.29 / 46.01 (0.7) / (0.95)	55.64 / 44.87 (0.67) / (0.8)	55.64 / 44.87 (0.67) / (0.8)	55.0 / 44.63 (0.71) / (0.83)
DistilBERT	92.7 / 65.11 (0.46) / (0.88)	92.69 / 64.72 (0.45) / (0.85)	92.69 / 64.72 (0.45) / (0.85)	92.67 / 64.79 (0.46) / (0.86)
DistilRoBERTa	**93.55 / 67.97** (0.38) / (0.79)	**93.49 / 66.7** (0.39) / (0.81)	**93.49 / 66.7** (0.39) / (0.81)	**93.5 / 67.1** (0.39) / (0.8)
SqueezeBERT	92.85 / 65.82 (0.44) / (0.86)	92.78 / 64.14 (0.44) / (0.78)	92.78 / 64.14 (0.44) / (0.78)	92.79 / 64.71 (0.44) / (0.81)
LSTM (*baseline*)	16.31 / 19.69 (0.62) / (0.75)	12.44 / 10.56 (0.3) / (0.27)	12.44 / 10.56 (0.3) / (0.27)	10.75 / 9.88 (0.27) / (0.24)

Table 1: Mean and (standard deviation (std.)) performance on dev and test data for a selection of SOTA models and an LSTM baseline. The average support size of each dataset is also provided. Results were extracted from 100 trials of validation sets of 30 batches, each containing 32 *passages*. Mean and (std.) support (# of tokens) for datasets was 6,881,956 and (104,283) respectively. Models built from pretrained HuggingFace Transformers.

`HumAID`, we tag our tweets with one of ten labels, shown in Table 3 below. As a baseline, we include a LSTM in our model set. In addition, we make use of pretrained DistilBERT and DistilRoBERTa models. BERT is considered the state of the art in many NLP tasks, including sequence tagging. DistilRoBERTa is a derivative built to handle particularly large sequences.

To assist these models in generalizing to new crises and lexical forms, we generate a completely unique augmented training set for every epoch. All models were trained for 500 epochs, and then evaluated on held-out validation data pulled from the same crises as the training set. Additionally, data augmentation enables us to bootstrap uncertainty bounds. Shown in Table 1, we display the mean and standard deviation of model performance over 100 generated validation datasets of 30 batches, each with 32 samples. The same process is conducted for test datasets, which generate *passages* from a completely unseen set of crises. Interestingly, the average tweet length of these held-out crises appears much larger than the validation dataset's, as indicated by the mean support size.

Impressively, both BERT models are able to achieve validation performances beyond 90%, far better than the LSTM's generalization capability. However, performance decreases substantially when applying the models to new datasets. Even so, model performance still remains near 70% and is predictive enough to be applied effectively in practice, as shown by our qualitative results.

Tag	Prec.	Rec.	F1	%
Rescue, Don. & Vol.	83.54 (1.01)	79.09 (1.19)	81.25 (0.87)	36.06 (0.97)
Inj. or Dead ppl.	77.51 (3.22)	83.44 (2.44)	80.32 (2.14)	4.37 (0.31)
Disp. ppl. & Evac.	77.23 (2.68)	83.2 (2.44)	80.07 (2.02)	6.1 (0.4)
Sympathy & Supp.	74.38 (2.03)	71.8 (1.83)	73.05 (1.43)	12.68 (0.66)
Missing People	70.24 (19.95)	79.41 (19.81)	71.37 (16.59)	0.03 (0.02)
Property Damage	65.96 (2.98)	59.98 (3.28)	62.76 (2.47)	7.11 (0.43)
Caut. & Advice	55.89 (3.35)	58.74 (2.83)	57.22 (2.54)	8.13 (0.57)
Other Rel. Info	38.91 (2.15)	44.18 (2.12)	41.35 (1.86)	15.06 (0.61)
Req. or Urg. Needs	52.26 (5.59)	32.47 (4.08)	39.96 (4.34)	2.41 (0.27)
Not Humanit.	33.83 (3.11)	42.99 (3.61)	37.8 (2.99)	8.05 (0.52)

Table 2: Mean and (standard deviation) performance of DistilRoBERTa on test crisis data separated by label as well as percent composition in the dataset, sorted by F1 Score in descending order, Results extracted from 100 trials of validation sets of 30 batches, each containing 32 *passages*. Bolded topics represent more abstract concepts, likely contributing to the relatively poor performance.

	Year	Prec.	Rec.	F1
Canada Wildfire	2016 2016	75.94 (1.51)	75.8 (1.40)	75.56 (1.40)
Mexico E-quake	2017 2017	81.31 (1.48)	79.37 (1.35)	80.0 (1.39)
Kerala Floods	2018 2018	67.6 (1.45)	62.56 (1.39)	64.18 (1.36)
H-cane Dorian	2019 2019	57.62 (1.59)	57.79 (1.52)	56.97 (1.58)
Mean	-	**70.62**	**68.88**	**69.18**
Crisis	-	(1.51)	(1.42)	(1.43)

Table 3: Mean and (standard deviation) performance of DistilRoBERTa on unseen test crises. Results were extracted from 100 trials of validation sets of 30 batches, each containing 32 *passages*.

Further evidence that validates our data augmentation protocol can be seen in the standard deviation metrics we track on model performance. With each *passage* essentially guaranteed to be unique, these models witness a tremendous amount of variation when evaluated on augmented datasets. Regardless, performance across all metrics never varies by more that 1% in standard deviation, verifying this protocol can train a robust sequence tagger by augmenting a unilabel dataset. This robustness appears even more pronounced when examining per-crisis metrics on the validation dataset.

However, when generalizing to new crises, performance of these models becomes far more variable. It appears that the models are able to understand semantic information about specific crises during training and apply this at test time, leading to a noticeable performance gap between seen and unseen crises at test time. We explore overfitting further by stratifying performance by humanitarian label for training and testing sets, shown in Table B.4 in the appendix. Ordered by test F1 score in descending order, we see that performance begins to drop dramatically as the humanitarian topic becomes more vague (topic bolded). For example, language about `injured or dead people` is fairly prescriptive in nature, and the model detected it with far higher performance than `other relevant information` or `not humanitarian`. Intuitively, this is somewhat expected. For many NLP applications, performance suffers as the task becomes more abstract.

4.2 Transitioning to New Lexicons

As mentioned in our introduction, a current gap in crisis NLP is a lack of granularity in extracting disaster information. To address this, let us return to our hypothetical example sentence. When fed through DistilRoBERTa, we achieve the following: "*Our hearts go out to those affected by the fire that has injured 12 citizens; several people are still missing and we will begin a search.*" for topics sympathy, injury /death, and missing people. With only a small mis-classification error on setup words ("that","has"), DistilRoBERTa is able to extract the relevant humanitarian details crucial for response. This is particularly impressive, since the training data it was exposed to simply concatenated tweets and never included compact topic switching. The fact DistilRoBERTa can identify humanitarian topic switches in only a few tokens speaks to its comprehensive pretraining as well the efficacy of our data augmentation system. When no sequence tagging data exists for crisis (as it does today), our framework appears to be a viable alternative.

Now let us move beyond this toy example. Consider the following examples using DistilRoBERTa on a Wikipedia article on Hurricane Dorian and a BBC release on Mexico's Puebla Earthquake. Colors denote the following tags: injury/death, evacuation, caution/advice, infra./util. damage, rescue/vol. effort, and other relevant information. No topic flips occur during redacted . . . sections.

Hurricane Dorian Wikipedia Article

> In preparation ... [many states] declared a state of emergency and [many] ... issued evacuation orders. [Dorian] made landfall in the Bahamas in Elbow Cay, ... and damage ... was catastrophic due to the intense storm conditions ... with thousands of homes destroyed and at least 77 direct deaths were recorded.

Mexico Puebla Earthquake BBC Article

> Elsewhere 15 people were killed when a church ... collapsed during Mass. Puebla governer was quoted ... saying [a] volcano [nearby] had a small eruption as a result of the tremor ... [and] schools would be closed and public transport would be free to allow people to get home. Emergency workers have been searching through the night ...

	Prec.	Rec.	F1	Partial Seq.	Total Seq.	Mean Flips	>1 Tag Seq. %	Tag Streak	Seq. Len.	# Toks.
Mex. E-quake	70.56	71.81	68.39	87.65	40.74	0.95	58.02	12.50	23.60	1912
Canada W-fire	80.75	79.14	79.43	91.55	50.70	0.66	46.48	22.83	40.85	2900
Kerala Floods	74.90	77.57	75.44	85.71	65.08	0.14	11.11	29.09	33.97	2140
H-cane Dorian	72.70	63.73	60.20	73.86	34.09	0.59	37.50	17.75	28.23	2484
Mean Crisis	**74.73**	**73.06**	**70.87**	**84.70**	**47.65**	**0.59**	**38.28**	**20.54**	**31.66**	**2359**

Table 4: Performance and descriptive statistics of 12 manually labeled news articles (3 for each crisis). In addition to traditional performance, partial and total sequence examine the percent of sequences the agent labeled partially and totally correct. In addition, for each crisis the mean number of topic flips per sequence, as well as the percentage of sequences with >1 tag, are recorded. Lastly, we note the mean tag streak, sequence length, and the support.

Wikipedia and news articles are typically used to provide a comprehensive, dense overview of a topic, making it a useful testing ground for our sequence tagger. Our qualitative findings, highlighted above, are compelling. It is clear that our tagger does not rely on contrived phrases (e.g. tweets or sentences) or structure. Instead, it extracts information semantically, flexibly switching topics as necessary. In order to further strengthen our analysis of new lexicons beyond qualitative review, we also manually collected and annotated 12 articles – three for each held-out test crisis. These articles were taken from a variety of sources, including major news outlets, blogs, and Wikipedia. In total, this totalled nearly 10,000 tokens of text for classification, as shown in Table 4.

Though a fair amount of variation still exists, the mean performance across our labeled news articles was very similar to the held-out performance on tweets. In fact, the overall performance was slightly better, indicating that perhaps formal articles are easier to parse than free-form tweets.

Additionally, we document salient features of news article structure. First, we note that the percentage of sequences in an average news article that contain multiple topics (and hence would require sequence tagging) is non-trivial at roughly 38%. The tag streak data further increases our level of granularity in this exploration and denotes the mean continuous streak of tokens for a topic before a flip occurs relative to average sequence length.

When compared to our held-out tweet dataset, these articles bear many structural differences. There is a wide variety in the homogeneity of sequences. Articles on Kerala Floods rarely saw topic changes (0.14 mean flip rate), while Mexico's Earthquake had an average of nearly one per sequence. Furthermore, the label composition of this dataset was far different than the tweet data, with `other relevant information` comprising over 37% of the dataset. This represented the most notable failure mode of the model. For long-form news, there is far more ancillary information provided – testimonials, quotes, opinions, etc. It was challenging for our model to determine what should be considered relevant information and what should not. With `other relevant information` and `not humanitarian` representing negation categories from the other tags, the most common failure was mistaking a more explicit category for other relevant information, and vice versa. Even so, performance on this new lexicon as a whole was both consistent with original test-set findings and robust. There were no discernable trends between performance and the type of crisis, the year, or the medium of communication.

5 Conclusion

Information extraction during a crisis is a challenging problem. Commonly, writers will frantically post messages to social media that contain information on a variety of humanitarian topics. In many cases, this leaves unilabel classification insufficient for information extraction. Alternatively, no sequence tagging dataset currently exists for crisis information. We demonstrate that in this absence our data augmentation protocol with a large dataset can simulate sequence tagging data and train a language model to generalize to new crises as well as lexicons. While there is still room for improvement in generalization performance on unseen tasks, we believe that this approach provides a solid foundation upon which further advances can be made.

References

E. Di Corso, F. Ventura, and T. Cerquitelli. 2017. All in a twitter: Self-tuning strategies for a deeper understanding of a crisis tweet collection. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3722–3726.

Muhammad Imran Ferda Ofli Firoj Alam, Umair Qazi. 2021. Humaid: Human-annotated disaster incidents data from twitter. In *15th International Conference on Web and Social Media (ICWSM)*.

R. Interdonato, A. Doucet, and J. Guillaume. 2018. Unsupervised crisis information extraction from twitter data. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 579–580.

A. Kruspe, J. Kersten, and F. Klan. 2020. Review article: Detection of informative tweets in crisis events. *Natural Hazards and Earth System Sciences Discussions*, 2020:1–18.

S. E. Middleton, L. Middleton, and S. Modafferi. 2014. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17.

A. Schulz, E.L. Mencía, T.T. Dang, and B. Schmidt. 2014. Evaluating multi-label classification of incident-related tweets. 1141:26–33.

A Selected Article Confusion Matrix

The figures below show the confusion matrix for the 12 selected and manually annotated articles of the 4 held-out crises. As a reference, the index-legend lookup table is provided below

	0	1	2	3	4	5	6	7	8	9
0	**813**	27	6	0	0	74	407	0	38	0
1	1	**514**	0	0	0	5	44	0	23	0
2	5	44	**891**	9	4	23	103	20	17	0
3	0	30	21	**537**	16	47	59	0	8	0
4	0	0	0	0	**0**	0	0	0	0	0
5	22	17	0	0	0	**473**	71	0	19	0
6	131	58	86	22	0	836	**2747**	14	58	2
7	0	0	0	0	0	0	0	**9**	8	0
8	0	48	4	0	0	0	26	0	**747**	0
9	0	0	0	13	0	12	47	0	0	**180**

Table 5: Confusion matrix for manually labeled articles for four held-out test set crises. Correct predictions are bolded along the trace of the matrix, and the index reference legend for the labels can be found below.

	Decoded Tag Name
0	Caution and Advice
1	Displaced People and Evacuations
2	Infrastructure and Utility Damage
3	Injured or Dead People
4	Missing or Found People
5	Not Humanitarian
6	Other Relevant Information
7	Requests or Urgent Needs
8	Rescue, Volunteering, or Donation Effort
9	Sympathy and Support

Table 6: Index and corresponding expanded name of all tags utilized in the HumAID dataset. For brevity, these are shortened in the article body, but the semantic meaning should be clear.

B Model Performance by Crises, Dev

On the following page, we provide the per-crisis performance on the validation datasets, broken out by validation crises. These crises, but not the specific tweets examined, were seen at training time. As shown, SOTA NLP models can easily generalize to language on seen crises. Performance drops substantially for unseen test crises, but still remains useful for real-world application.

	Year	Precision	Recall	Accuracy	F1
Hurricane Matthew	2016 2016	93.33 (1.51)	92.94 (1.61)	92.94 (1.61)	93.0 (1.58)
Italy Earthquake	2016 2016	95.98 (1.02)	95.91 (1.0)	95.91 (1.0)	95.85 (1.04)
Kaikoura Earthquake	2016 2016	91.93 (1.56)	91.4 (1.61)	91.4 (1.61)	91.52 (1.6)
Ecuador Earthquake	2016 2016	96.44 (0.94)	96.36 (0.94)	96.36 (0.94)	96.33 (0.95)
Hurricane Harvey	2017 2017	92.16 (1.31)	91.93 (1.39)	91.93 (1.39)	91.89 (1.4)
Hurricane Irma	2017 2017	91.43 (1.49)	91.18 (1.53)	91.18 (1.53)	91.19 (1.54)
Hurricane Maria	2017 2017	91.92 (1.51)	91.58 (1.56)	91.58 (1.56)	91.59 (1.55)
Srilanka Floods	2017 2017	95.26 (1.06)	94.92 (1.14)	94.92 (1.14)	94.91 (1.14)
California Wildfire	2018 2018	91.4 (1.39)	91.1 (1.45)	91.1 (1.45)	91.08 (1.45)
Hurricane Florence	2018 2018	91.15 (1.56)	90.8 (1.61)	90.8 (1.61)	90.83 (1.61)
Cyclone Idai	2019 2019	94.28 (1.32)	94.18 (1.32)	94.18 (1.32)	94.06 (1.37)
Midwest U.S. Floods	2019 2019	91.13 (1.78)	90.56 (1.92)	90.56 (1.92)	90.58 (1.93)
Pakistan Earthquake	2019 2019	94.45 (1.11)	94.06 (1.31)	94.06 (1.31)	93.95 (1.33)

Table 7: Mean and (standard deviation) performance of DistilRoBERTa on dev set crises. Results extracted from 100 trials of validation sets of 30 batches, each containing 32 *passages*.