

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

Firoj Alam, Hassan Sajjad, Muhammad Imran, Ferda Ofli
Qatar Computing Research Institute (QCRI), HBKU, Qatar



Highlights

- We consolidate eight human-annotated datasets and provide 166.1k and 141.5k tweets for informativeness and humanitarian classification tasks
- We carefully cleaned various forms of duplicates, and assigned a language tag to each tweet
- We provide benchmark results on English tweets using state-of-the-art machine learning algorithms such as CNN, fastText and pre-trained transformer models
- We release the dataset in multiple forms that can help the community to perform different types of classification experiments
- Publicly available at https://crisisnlp.qcri.org/crisis_datasets_benchmarks.html

