Description of the dataset

==========================

The HumAID Twitter dataset consists of several thousands of manually annotated tweets that have been collected during nineteen major natural disaster events including earthquakes, hurricanes, wildfires, and floods, which happened during 2016 to 2019 across different parts of the World. It is the largest social media dataset (~77K) for crisis informatics so far. The annotations consist of following humanitarian categories.

Humanitarian categories

\* Caution and advice

\* Displaced people and evacuations

\* Dont know cant judge

\* Infrastructure and utility damage

\* Injured or dead people

\* Missing or found people

\* Not humanitarian

\* Other relevant information

\* Requests or urgent needs

\* Rescue volunteering or donation effort

\* Sympathy and support

Data format and directories

===========================

The data directory contains a sample dataset:

\* events/ This directory contains sub-directories for each event. In which each event directory contains tab-separated (i.e., TSV) three files, i.e., train, dev and test. Each TSV file stores ground-truth annotations for the aforementioned humanitarian categories. The data format of these files is described in detail below.

Format of the TSV files

---------------------------------------------------------

Each TSV file contains the following columns, separated by a tab:

\* tweet\_id: corresponds to the actual tweet id from Twitter.

\* tweet\_text: corresponds to the tweet text.

\* class\_label: corresponds to a label assigned to a given tweet text.

Citation

========

Terms of Use

============

Please see Licensing.txt