The data is collected from Crisislex. The data set is ***CrisislexT26*** and is available at <https://github.com/sajao/CrisisLex/blob/master/releases/CrisisLexT26-v1.0.zip?raw=true>.

The original data collection contains tweets that were collected during the course of 26 large crisis events between 2012 and 2013. Around 1000 tweets per crisis have been labelled for their informativeness, information type, and source.

In this sub collection we have chosen 9 events from 26 that are predominantly in English. The following events have been chosen: Australia Bushfire, Colorado Wildfire, Colorado Floods, Queensland Floods, Savar Building collapse, West Texas Explosion, Boston Bombing, Los Angeles Airport Shooting, and Singapore Haze.

The informativeness is classified under 4 categories: Related and informative, Related but not informative, Not related, and not applicable. For performing a binary classification the data from those events (Events containing majority of documents in English were considered), documents labelled as *Related and informative* have been taken as *POSITIVE* labelled documents, while documents labelled as *Not Related* and *Not Applicable* have been labelled as *NEGATIVE* labelled documents.

Two instances of such documents are shown below (in the context of Colorado wildfire):

|  |  |
| --- | --- |
| Text | Score |
| #Evacauation center Cache La Poudre Middle School 3515 West County Road 54G in Laporte. #Colorado #Wildfire | 1 |
| Denver Post: #Colorado governor signs bill creating rules for public access to ballots http://t.co/Hmp1wQ8P #opengov | 0 |

There are 3 folders : BabelNet Annotation, CRISISLEXT26\_9events\_2\_labels and CRISISLEXT26\_9events\_Original\_files\_tvs

The 2 labels folder contains CSV files (tab separated) that contain the tweet IDs, the corresponding tweet and a label (0/1) representing if this is a crisis related tweet or not.

The original file folder contains the original CSV (tab separated) files provided by the CrisisLex.

The BabelNet annotation folder contains the annotations of the records for each event, contained in the CSV files in the folder CRISISLEXT26\_9events\_2\_labels.

The csv files (comma separated) contain Babel Synset Annotation for each term occurring in each tweet. The unique ID provided in the first column is the Tweet serial number reference in the csv file for corresponding event in *2 labels* folder. Babel Synsets were generated by using Babelfy annotation API provided by BabelNet.

e.g.

"1","Total","bn:03683702n","http://babelnet.org/rdf/s03683702n","http://dbpedia.org/resource/Total\_S.A.","BABELFY","1"

Tweet serial number, Word Annotated, Babel Synset ID (annotation ID in BabelNet), BabelNet URI, DBpedia URI, Source, Label\_ID