Description of Sources

name of source	URL	quality of source (high/low)	number of (annotated) units	explanation of quality of source
Sources weren't downloaded explicity into a file, urllib.request and BeautifulSoup were used to fetch and parse html sources	http://www.disaster-report.com/	Low (blogs, people's comments, or twitter feeds)	98	This was a blog on Disaster Report has been created by Ashish Khanal
	http://www.worldvision.org/disaster-response-news-stories/worst-natural-disasters-2015			Report by Chris Huber
	https://www.theguardian.com/world/2016/jul/27/flooding-in-india-affects-16m-people-and-submerges-national-park	High (News articles)	11	Unlike other news articles this one had more information relevant to domain of interest
	http://www.cnn.com/2016/04/04/asia/pakistan-floods/index.html		<5	Very few sentence units was able to be tagged as other news were irrelevant to domain of interest
	http://www.rollingstone.com/culture/pictures/5-recent-underreported-environmental-disasters-20140407/elk-river-chemical-spill-0660454		<5	Very few sentence units was able to be tagged as other news were irrelevant to domain of interest
	http://www.usnews.com/news/national/articles/2007/10/29/a-timeline-of-natural-disasters-in-california		16	Unlike other news articles this one had more information relevant to domain of interest

Feature Description

name of feature	a representative example of it	range of values
Parts of Speech tag leveraging python wrapper for Standford CoreNLP https://github.com/hhsecond/corenlp_pywrap	JJ/VBN/NNP/CD	Penn Treebank tag set http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html
Named Entity Recognition leveraging python wrapper for Standford CoreNLP https://github.com/hhsecond/corenlp_pywrap	PERSON/ DATE/ LOCATION / ORGANIZATION	7 class: Location, Person, Organization, Money, Percent, Date, Time http://nlp.stanford.edu/software/CRF-NER.shtml
Capitalization	0/1	0 or 1 indicating if the token's first character is capitalized or not
Numeral	0/1	0 or 1 indicating if the token is a number or not
Sequence_Code	B-X/I-X/O	B-X/I-X/O to capture sequence of tokens B-X if beginning a sequence, I-X if within a sequence else O

The following labels were used:
       d : date, m : month,   y : year , p : place, n : noAffected(no of people affected) ,
       i : irrelevant
       c : cause(EarthQuake/OilSpill/FireAccident/typhoon) (ie. cause of event)
       e : effect(Kills/Injures/Devastates/Destroy)(i.e effects of event)

Each word/token of a sentence was tagged with aforementioned features in order :
Token, POS_TAG, NER_TAG, Capitalization, Numeral, Sequence_Code, Label

As observed in training and testing files, the same format was used for classification using crf++ :
https://taku910.github.io/crfpp/

Test Results

Report Scores can be generated using :
        python generate_report.py combined_quality_test_output.txt
        python generate_report.py high_quality_test_output.txt
        python generate_report.py low_quality_test_output.txt

precision	recall	F1	category	dataset
0.71	0.59	0.65	"cause" (EarthQuake/OilSpill/FireAccident/typhoon) (ie. cause of event)	High and Low Quality Sources
0.83	0.56	0.67		High Quality Source
0.62	0.62	0.62		Low Quality Source
0.50	0.10	0.17	"effect" (Kills/Injures/Devastates/Destroy)	High and Low Quality Sources
1.00	0.20	0.33		High Quality Source
0.00	0.00	0.00		Low Quality Source
0.92	0.99	0.95	"irrelevant"	High and Low Quality Sources
0.94	0.99	0.99		High Quality Source
0.89	0.98	0.93		Low Quality Source
1.00	1.00	1.00	"month"	High and Low Quality Sources
1.00	1.00	1.00		High Quality Source
1.00	1.00	1.00		Low Quality Source
0.88	0.58	0.70	"noAffected"	High and Low Quality Sources
1.00	0.83	0.91		High Quality Source
0.67	0.33	0.44		Low Quality Source

1.00	0.94	0.97	"place"	High and Low Quality Sources
1.00	1.00	1.00		High Quality Source
1.00	0.83	0.91		Low Quality Source
1.00	0.86	0.92	"year"	High and Low Quality Sources
1.00	0.67	0.80		High Quality Source
1.00	1.00	1.00		Low Quality Source

The following are weighted scores for different test sets
    ('weighted_avg of f1-scores for all classes: ', 0.9) : combined high and low quality tests
    ('weighted_avg of f1-scores for all classes: ', 0.93) : high_quality_test
    ('weighted_avg of f1-scores for all classes: ', 0.85) : low_quality_test

We observe that month, year, place gets classified with above 0.9 f1-scores across all the dataset

It is interesting to note that a sequence label like noAffected(which tries to capture no of people affected) ex : "17 people died….", "...killing more than 8,800 people", were classified with 0.91 f1-scores in high-quality sources where-as low as 0.41, drop of 47% in low-quality sources.

"cause" label representing cause of disaster approximates 0.6 f1 scores across all datasets. However "effect" label representing consequence of disaster had 0 f1-scores for low-quality dataset.

For well-known label categories such as month,year,place quality of test-sets didn't matter much, but we have significant impact of quality of test-sets for labels such as noAffected, cause, effect