



BIG DATA CHALLENGE

1. BACKGROUND

The frequency and severity of natural disasters are growing day by day. These natural disasters (e.g. hurricanes) cause damages or disruptions to the infrastructures that significantly affect human lives and wellbeing. Social media monitoring gives an opportunity to collect the disruption information during an ongoing disasters. But the volume of social media data and lack of enough manpower make it impossible to process these data during natural disasters.

2. OBJECTIVE

Imagine you are hired as a Data Scientist for a disaster management organization and you are requested to create an algorithm which classifies different types of disruption information gathered from social media platforms. The purpose of this algorithm will be to make disruption identification easier. This will reduce the human efforts and enable disaster management agencies to make a proactive decision for disruption response and recovery.

3. ELIGIBILITY

Each university may have one (1) team of up to five (5) members.

4. PROCEDURE

Each team will need to bring a laptop to the conference event.

For this challenge, at the beginning of the conference 1000 posts collected from Twitter during hurricane Irma will be given to teams. Each tweet is geo-tagged. You have to label each tweet as follows:

- Whether a given tweet is hurricane related or not (**e.g., 1- hurricane related, 0- not hurricane related**)
- If a tweet is describing an actual disruption or not (**e.g., 1- actual disruption, 0- not disruption**)
- What is the type of disruption among the 6 options: **power (1), water (2), communication (3), wastewater (4), transportation (5), and other (6)**. For multiple disruptions, one label should be sufficient.

Prior to the competition, 40 sample tweets will be provided with the actual labels to give the format of the tweets as well as ability to test your tagging scheme.





5. GUIDELINES

1. Your output will be evaluated against the data labeled by a human.
2. The total number of submissions allowed by a team is during the conference is 3. Each attempt will be given a score (to allow you to improve your tagging); however, only the final submission will be used for competition scoring. The submission procedure will be released prior to the competition.
3. You can develop a computer algorithm or use any existing tool for labeling the tweets. There is no restriction on the tools that you can use.

6. SCORING

For a tweet, 1 point will be given for correctly labeling each aspect. The maximum possible score for each tweet is 3. For any incorrect labeling 0.5 points will be deducted. For example:

If a tweet is labeled as: 1 1 2 – hurricane related/actual disruption/water infrastructure.

If you predict the label as 1 1 4 then it will receive 1.5 points (1 + 1 - 0.5).

A few scoring examples are given below:

Actual Label			Predicted Label			Score
1 (hurricane related)	1 (actual disruption)	2 (water)	1	1	3	1.5
0 (not hurricane related)	0 (no disruption)	6 (other)	1	1	1	-3
1 (hurricane related)	1 (actual disruption)	3 (communication)	1	1	3	3

7. JUDGING

Teams will be ranked based on the resulting score from their final submission.

8. QUESTIONS

Direct any questions to [conference organizers](#). Answers will be posted on the Q&A page.

