TREC 2018 Incident Streams Track

Guidelines v1.0, 16th May 2018

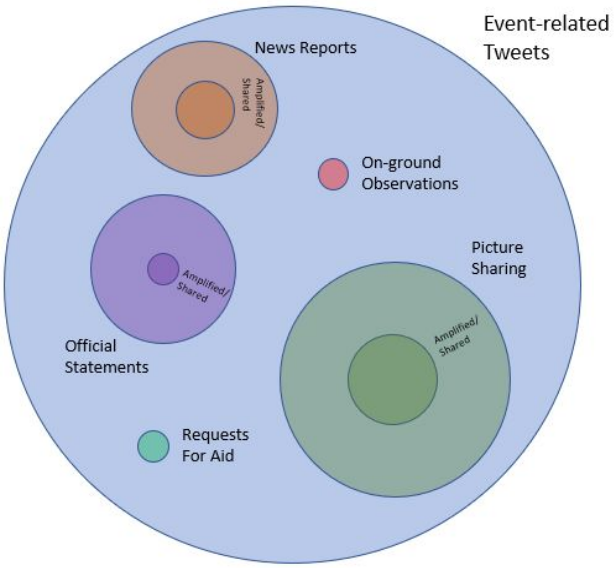
Coordinators:

Richard McCreadie, University of Glasgow
Cody Buntain, University of Maryland
Ian Soboroff, NIST

Motivation

People often turn to social media during emergencies as a source for information. Increasingly, we expect some information posted to social media to be important to emergency responders and public safety personnel. However, at this point in time, few technologies exist to help those users filter a social media stream down to actionable information or to route that information to the right safety sector for planning.

Given the notional tweetstream about an emergency like a wildfire in proximity to people's homes, we can imagine a range of information types that might be shared during the incident. The vast majority of tweets, might be expressions of sentiment, solidarity, and wishes to help from around the world. More valuable than those are reports from news services and government officials that contain useful information for people in the area of the incident. Meanwhile, the most relevant information is contained within the small number of tweets by people in the affected region who are reporting first-hand about conditions on the ground and immediate safety and health needs.



This track is sponsored in part by NIST efforts aimed at developing technology to support public safety, and hence we have a focus on local incidents rather than major disasters.

Dataset

For this track we have selected a number of events/incidents of different types, e.g. earthquakes, hurricanes, public or shootings. For each incident, we have a stream of tweets related to the incident, collected using hashtags and keyword monitoring. Each incident stream should be treated as an independent dataset for purposes of this track – systems can assume that an upstream system is providing basic filtering of the Twitter feed. The incidents and streams come from two sources. One is crisislex.org, and the other are collections curated by the organizers representing current events.

These datasets will be distributed as a list of tweet identifiers for each incident. Participants will need to fetch the actual JSON tweets using publicly available tools, such as twarc (https://github.com/DocNow/twarc), the TREC Microblog Track twitter-tools (https://github.com/lintool/twitter-tools), or any other tool for crawling twitter data.

Each incident/event is accompanied by a brief "topic statement" in the TREC style:

```
<top>
<num>Number: 001 </num>
<title>colorado wildfires</title>
<type>wildfire</type>
<url>https://en.wikipedia.org/wiki/2012_Colorado_wildfires</url>
<narr> The Colorado wildfires were an unusually devastating series of fires
in the US state of Colorado, which occurred throughout June, July, and
August 2012.
</narr>
</top>
```

Not all topics will have the 'url' field, and systems **should not** use the referenced pages in their systems; we are including those links as documentation for the incidents, but since they contain retrospective information that couldn't be available during the incident tweetstream, using it would be anachronistic.

Event Types

The incident streams task is focused on emergency/crisis-type events. The event types that you may need to process are:

- **wildfire**, **earthquake**, **flood**, **typhoon/hurricane**, **bombing**, **shooting**

Tasks

There is only a single task for the first year of the track (2018): *classifying tweets by information type (high-level)*.

Task: Classifying Tweets by Information Type (High-Level)

The goal of this task is for systems to categorize the tweets in each event/incident's stream into different information feeds that might be consumed by different public safety personnel or used for post-event analysis. In particular, we have developed a multi-layer ontology of information types (described later). The nodes in this ontology represent the different information types. In effect, the task aim is to assign ontology labels (information types) to each tweet within the event stream. A public safety officer can then 'subscribe' to the information types that are useful for fulfilling their role, e.g. shared images from the disaster area, or first-hand reports of unsafe conditions.

As noted above, the ontology has multiple layers, moving from generic information types to the very specific. For this reason, we denote information types as either `top-level intent', 'high-level' or 'low-level'. For example, a top-level intent might be 'Reporting' (the user is reporting some information). Within reporting, a high-level type might be 'Service Available' (the user is reporting that some service is being provided). Within service available, a low-level type might be 'Shelter Offered' (shelter is offered for affected citizens).

This task is Classifying Tweets by Information Type **(high-level)**. I.e. the goal is to categorize tweets into the information types listed as *high-level*. One category per tweet.

Participants can process the data as a single *batch*, or as a tweet-ordered *stream*. Your system can be either *fully automatic*, involving no human intervention once the data is exposed to the system, or *manual*, which includes any human intervention (like relevance feedback, manual query construction, online supervised learning…).

Task 1 Submissions

Participants submit the output of their system over a set of designated 'test' events, denoted 'TRECIS-CTIT-H 2018 Test' (Classifying Tweets by Information Type High-Level Test). A single participant can submit the output of multiple systems if desired, up to a maximum of four (if you wish to submit more than this then contact the organizers). We refer to a single submission as a **'run'**.

When submitting a run, it should be uploaded as a single gzip compressed text file. This file should contain one line for each tweet within the stream for the test events, in a slightly-modified TREC format, as shown below:

```
1 Q0 991459953742262272 1 37.5 Request-GoodsServices myrun
1 Q0 991855886363541507 2 33.2 Report-MultimediaShare myrun
...
1 Q0 991855942093291520 999 0.5 Other-Discussion myrun
2 Q0 992010886465314816 1 55.2 Report-Factoid myrun
...
```

There are seven fields, as follows:

1. The first field is the **incident identifier** (the contents of the "<num>" tags in the incident topic statement)
2. The second field is a literal "**Q0**" (this is kept because the evaluation script expects it)
3. The third field is the **tweet ID** of the tweet, an 18-19 digit number
4. The fourth field is the **rank** (per-event, please rank your tweets by the fifth field below)
5. The fifth field is a score, this should be how important you consider the information contained within the tweet ot be. Depending on your system, you might simply assign scores to each high level category, or use deeper analysis of the tweet text to generate an **importance score**.
6. The sixth field is the **information type** within the ontology. Only *high-level* types are valid categories for this task.
7. The seventh field is the **runtag**, this should be a unique identifier for your system.

For consistency please use **tab** characters between fields. Participants **categorize all tweets for each event** (this is important to enable future analysis of systems).

Task 1 Assessment

We will evaluate the performance of each submitted run at NIST. This is operationalized by having human assessors manually label a subset of the tweets returned within your run(s). It is expected that we will pool updates from each of your runs, prioritizing those with high importance scores, while also diversifying across information categories.

Task 1 Metrics

The primary metric for evaluation will be classification F1 score, micro averaged over the different information types (one-vs-all). Micro averaged precision, recall, and accuracy will also be reported.

Task 1 Training Examples

In addition to the ‘test’ events (‘TRECIS-CTIT-H 2018 Test’), participants will also be provided with a number of other events that they can use to evaluate (or train if using a machine learned approaches) their systems prior to running them on the ‘test’ events. We refer to these as training events, denoted as ‘TRECIS-CTIT-H Training’. For each training event we provide the tweet stream, as with the ‘test’ events. However, we also provide the following information for a subset of the tweets within those streams:

- **High-level Information Types**: These are human selected labels for a subset of the tweets for the training events.
- **Importance Scores**: These are derived from human selected importance labels for the tweets. The possible labels are: Critical, High, Medium, Low and Irrelevant. We map these to numerical scores as follows: Critical=1.0, High=0.75, Medium=0.5, Low=0.25 and Irrelevant=0.0.
- **Indicator Terms**: These are optional terms that the human annotators selected when choosing an information type to explain why they chose that type. For instance, for the high-level label ‘CallToAction-Evaluate’, indicator terms selected might include ‘leave’ or ‘NOW!’.

Participants may use the training events however they wish when developing or tuning their systems.

Ontology

As mentioned above, along with the event tweet stream, we also provide an ontology of information types that may be of interest to public safety personnel. These form the information types that you are to assign to each tweet. Rather than providing the entire ontology, we instead provide only the high-level types that you are to use as categories. These are provided in a JSON format file. For each information type we provide the following information:

```
{
      "id": "Request-GoodsServices",
      "desc": "The user is asking for a particular service or physical
               good.",
      "level": "High-level",
      "intentType": "Request",
      "exampleLowLevelTypes": [
            "PsychiatricNeed",
            "Equipment",
            "ShelterNeeded",
            "Vehicles"
```

```
    ]
}
```

The ontology can be accessed at:

➢ http://trecis.org/2018/ITR-H.types.json

Timeline

Guidelines released	21st May 2018
TRECIS-CTIT-H Training release	30th June 2018
TRECIS-CTIT-H 2018 Test release	30th July 2018
Runs due	30th August 2018
Run Assessing period	5th–29th September 2018
Scores returned to participants	15th October 2018
Notebook papers due	2nd November 2018
TREC	14th-16th November 2018