

TREC 2019-A Incident Streams Track

Guidelines v1.1, 13th March 2019

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 tweet stream 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. Indeed, we showed that the amount of actionable information that could be useful for response officers on Twitter is significant (up-to 10% post-filtering), although this varies greatly with event type

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.

Envisaged System

For context, below is an example of a system that might provide social media information to emergency response officers.

EVENT TRACKERHelp Simulator Log In Sign Up



Current Events


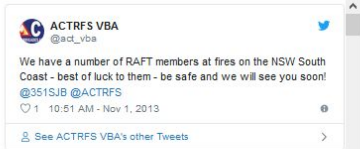
- 2012 Philippines Flood
- 2013 Alberta Flood
- 2013 Australia Bushfire
- 2013 Boston Bombings
- 2013 Manila Floods
- 2013 Queensland Flood
- 2013 Typhoon Yolanda

2013 Australia Bushfire

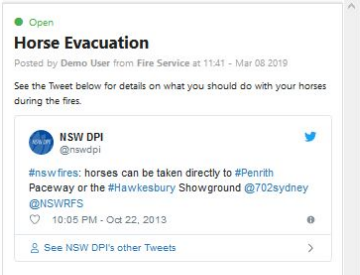
The 2013 New South Wales bushfires were a series of bushfires in Australia across the state of New South Wales primarily starting, or becoming notable, on the 13 October 2013. The user is a response officer in the command and control center for the New South Wales state in Australia.

[Wikipedia Page](#)

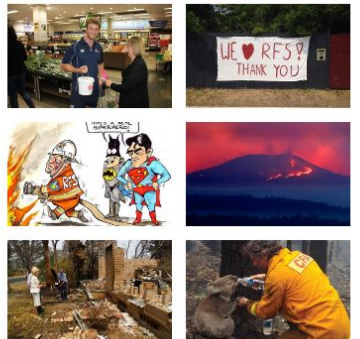
Twitter Feed




Event Tracker Feed



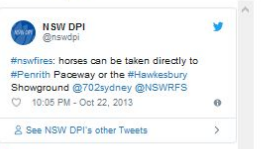
Media Feed





Advice



Move People



Donations



Goods and Services

No tweets found for selected event.

Event Tracker App: Developed by Charlie Thomas, University of Glasgow

Task: Classifying Tweets by Information Type (High-Level) v2

The TREC-IS 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*. Each tweet should be assigned as many categories as are appropriate.

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...).

Dataset

For this track we have select 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 in two ways. First, as a list of tweet identifiers for each incident. In this case, 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. Second, we host a server containing the tweets that you can use directly. In this case you will download a client program that will perform the download. More information about download methods can be found [here](#).

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**

Submitting

Participants submit the output of their system over a set of designated 'test' events, denoted 'TRECIS-CTIT-H 2019-A Test' (Classifying Tweets by Information Type High-Level 2019-A Test). A single participant can submit the output of multiple systems if desired, up to a maximum of four new systems (if you wish to submit more than this then contact the organizers). You may also submit the output of systems from previous TREC-IS editions, these do not count towards

the 4 system limit (if you participated in previous editions then please do submit the output of those older systems, so we can better track performance across editions). 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:

```
TRECIS-CTIT-H-Test-022 Q0 991459953742262272 1 0.8 ["Request-GoodsServices","Report-MultimediaShare"] myrun
TRECIS-CTIT-H-Test-022 Q0 991855886363541507 2 0.4 ["Report-MultimediaShare"] myrun
...
TRECIS-CTIT-H-Test-022 Q0 991855942093291520 863 0.1 ["Other-Discussion"] myrun
TRECIS-CTIT-H-Test-023 Q0 992010886465314816 1 0.6 ["Report-Factoid","Report-MultimediaShare"] 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 **tweet number in the stream**, sometimes referred to as the rank. Start at 1 for each event and count up.
5. The fifth field is a score, this should be how important you consider the information contained within the tweet to 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**. Values should be between 0 and 1.
6. The sixth field is the **information types** within the ontology. Only *high-level* types are valid categories for this task. This should be a comma-delimited list as illustrated above.
7. The seventh field is the **run tag**, this should be a unique identifier for your system. Please make this actually unique to your institution.

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, for the different information types. We will also provide weighted metrics that prioritize the categories that are more important (actionable).

Task 1 Training Examples

In addition to the new 'test' events, participants can also use the previous events from TREC-IS 2018 to evaluate (or train if using a machine learned approaches) their systems prior to running them on the new 'test' events. For each of the 2018 events 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.

Participants may use the previous 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/2019/ITR-H.types.v3.json>

Timeline

Guidelines released	11th March 2019
TRECIS-CTIT-H 2019-A Test release	6th May 2019
Run Assessing period	6th - 31th May 2019
Runs due	3rd June 2018
Scores returned to participants	17th June 2019

Summary of the 2018 Edition

TREC-IS also ran successfully in 2018, attracting participants from 11 international research groups. Full details can be found at <http://trecis.org> and we provide a short summary below for context.

2018 Task: There was a single task for the first year of the track: classifying tweets by information type (high-level). The goal of this task was for systems to categorize the tweets for a series of event/incident's streams into different information feeds that might be consumed by different public safety personnel or used for post-event analysis. In particular, we provided participants an [ontology of 24 information types](#). In effect, the task aim was to assign one of these ontology labels (information types) to each tweet within each event stream. In addition, participants were also to provide an priority score for each tweet, indicating how important the information within that tweet is, and hence ultimately whether the emergency response officer/PIO should be shown that tweet.

Each participant developed a system that processed the tweet stream for each event in time order, as if the event was occurring in real-time. As the system processes the stream, it emits

individual posts over time, categorizing those tweets into the information types from the ontology and assigning priority scores. Due to the time-critical nature of the task, decisions for each post were made immediately, i.e. a system had to choose to emit or discard a post immediately as it is processed.

2018 Dataset: We provided participants [6 training events](#) containing [pre-labeled tweets](#) and [15 test events](#). These events span six event types: wildfire, earthquake, flood, typhoon/hurricane, bombing and shooting. For each event type we also provided the participants a user profile document summarizing the aims/information needs of the emergency response officer/PIO for that event.

2018 Submission: Each participating group was allowed to submit 4 run files to TREC for evaluation. Unlike classical TREC tracks, participants return categories and scores for the entire tweet stream for each event. Runs follow a standard TREC format:

2018 Assessment: The test dataset comprises a total of ~23,000 tweets over the 15 test events. For this task we do not apply system pooling. Instead, all tweets were labelled by TREC Assessors. In particular, assessors were asked to assign one or more of the information types to each tweet and select an priority level (from Critical to Low).

2018 Evaluation: The aim of TREC-IS evaluation is to test the performance of a system in terms of its ability to identify relevant content for each of the ontology entries, as well as to evaluate whether systems would end up showing irrelevant or non-useful tweets the response officer/PIO. Primary system performance is measured using classical filtering metrics per selected ontology entry. More precisely, the target metrics are:

- **Precision:** The proportion of posts returned for the ontology entry are relevant for that entry.
- **Recall:** The proportion of all posts identified for the current ontology entry that were returned.
- **Accuracy:** The proportion of correctly classified posts (both true positives and true negatives) among the total number of posts pooled.

Meanwhile, the priority scores provided by systems are evaluated in terms of Root Mean Squared Error (RMSE) against the human-generated priority levels (when mapped to a continuous scale).

Additional information about the 2018 edition can be found at:

- http://dcs.gla.ac.uk/~richardm/TREC_IS/papers.html