TREC 2021-A Incident Streams Track

Guidelines v4.0.beta, 8 March 2021

Coordinators:

Richard McCreadie, University of Glasgow
Cody Buntain, New Jersey Institute of Technology
Ian Soboroff, NIST

Changelog:

- V4.0.alpha - Alpha 2021-A Track Updates
    - Merging Tasks 1 and 3 based on results of information-type distributions across COVID-19 and non-COVID events
    - Dropping Task 2 as results suggest systems trained on the full set of information types outperform those trained on restricted information-type sets
    - Updated submission section to reflect new JSON runfile format
    - Added a COVID-19-specific evaluation for assessing all systems against a subset of pandemic events in addition to the full TREC-IS event set
- V3 - 2020-B edition updates
- V2 - 2020-A edition updates

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. Despite this expectation, few technologies exist to filter a social media stream down to actionable information or to route that information to the appropriate stakeholder (e.g., public health officials, emergency response officers, etc.).

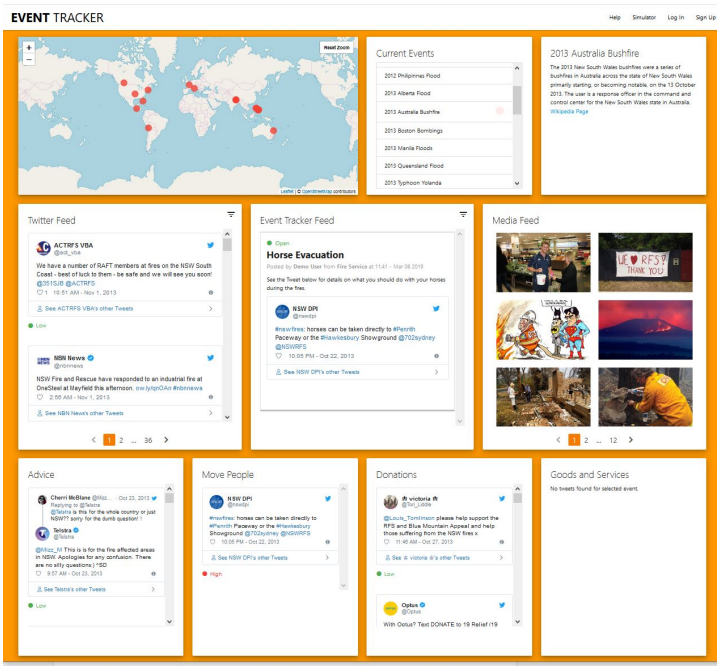
Given the notional tweet stream about an emergency like a wildfire in proximity to people's homes, we can imagine a range of information types that might be shared during the incident. Much of this content might be expressions of sentiment, solidarity, and wishes to help from around the world, but more 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 might be 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 health and safety needs (e.g., requests for rescue). In previous editions of TREC-IS, we have shown 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.

Hence, this track is motivated by the need for technology to support emergency response officers and other stakeholders *and* for technology assessment tools to instill trust in this technology.

This track is sponsored in part by NIST, and is aimed at developing technology to support public safety, and hence we have a focus on local incidents rather than major disasters. An overview of the previous TREC-IS 2019 editions can be found here.

Envisaged System

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



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

Overview of Tasks

In the 2021-A edition of TREC-IS, we are returning to a single task and merging COVID-19 events into this task:

- Task 1. All High-Level Information Type Classification, v2.1

Task 1. All High-Level Information Type Classification, v2.1

Systems participating in this task will be given tweet streams from a collection of crisis events and should classify each tweet as having one or more of the 25 high-level information types described in the ontology section below. Critically, each tweet should be assigned as many categories as are appropriate.

The nodes in this ontology represent various information types that might be needed by emergency response officers across a range of disasters. 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, first-hand reports of unsafe conditions, or volunteer coordination efforts.

While the ontology has multiple layers (moving from generic information types to the very specific), 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 targets the "high-level" labels, though participants are welcome to build multi-layered systems that first classify the "top-level intent" before tagging the high-level information type, which constitutes the primary output for this task. The high-level types are listed, alphabetically, in the Ontology section below.

For input, 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, etc.).

Datasets

In keeping with past TREC-IS editions, we have selected a number of emergency events covering several different types:

- Wildfires,
- Structural fires,
- Earthquakes,
- Floods,
- Tropical storms (e.g., hurricanes, typhoons),
- General storms (e.g., tornadoes, mudslides),
- Mass violence (e.g., shootings, bombings, or hostage situations),
- Industrial accidents (e.g., explosions),

- Public health emergencies (e.g., COVID-19, Zika, epidemics).

As with 2020-B, we will release larger datasets for 2021-A that have not been previously assessed by human annotators. Participant systems should assign categories and priorities to every message in these datasets, and TREC-IS coordinators will evaluate participant systems on a pooled set of content from these larger datasets.

For each incident, we have a stream of related tweets, collected using hashtags, keyword, user, and geolocation monitoring. Each incident stream should be treated as an independent dataset, and systems can assume that an upstream system is providing basic filtering and de-duplication of the Twitter feed (i.e., each event dataset has already been marginally filtered for relevance prior to arrival at your system). These streams have been collected from previous crisis informatics datasets (e.g., http://crisislex.org/ or http://aidr.qcri.org/) with more recent events having been curated by the TREC-IS organizers.

These datasets will be distributed via a host server 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>
```

NOTE: 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.

Submitting

Participants submit the output of their system over a set of designated ‘test’ events, denoted ‘TRECIS-CTIT-H 2021-A Test’ (Classifying Tweets by Information Type High-Level 2021-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, and 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. Participants should **categorize all tweets for each event** (this is important to enable future analysis of systems).

In prior editions of TRECIS, the track used a standard qrel format, but in the 2021 editions, the track is moving to a new **JSON-based** format. This format should be newline-delimited, such that each line is a standalone JSON object. A pretty-printed version of an entry follows:

```
{
    "topic": "TRECIS-CTIT-H-Test-022",
    "runtag": "myrun",
    "tweet_id": "991855886363541507",
    "priority": 0.67,
    "info_type_scores": [0.2,0.31,0.1,0.7,0.0,...],
    "info_type_labels": [0,0,0,1,0,...]
}
```

This format contains six fields, as follows:

1. A **"topic"** field referencing the **incident identifier** (the contents of the "<num>" tags in the incident topic statement)
2. A **"runtag"** field identifying your particular run (exclude your team name, as that information is recorded during run submission.
3. A **"tweet ID"** field containing a string version of the tweet's unique identifier.
4. A **"priority"** field that shows how important you consider the information contained within the tweet to be for a response officer, and should be a decimal value between 0 and 1, 0 indicating lowest priority and 1 indicating highest.
5. An **"info_type_scores"** field that reflects probabilities for the **information types** within the ontology, ordered alphabetically (use the order listed in the Ontology section). Each *high-level* type in the task must have an associated probability. This should be a comma-delimited list as illustrated above.
    a. TREC-IS coordinators will use these scores for selecting which tweets will be pooled for evaluation.

6. An **“info_type_labels”** field that specifies which labels this system associates with this tweet. This field should contain a binary array, with 1 indicating that information type is associated with this tweet. Use the same order as in the **“info_type_scores”** field.
    a. While you should be able to generate **info_type_labels** from **info_type_scores**, different groups may use different thresholds to select labels from scores. We ask systems to provide these labels to avoid TREC-IS coordinators from imposing a thresholding mechanism across all participants.

Task Assessment

Each submitted run and its performance will be evaluated at NIST via human assessors who manually label a subset of the tweets returned within your run(s). As in TREC-IS 2020-B, 2021-A will rely solely on pooling for evaluation. 2021-A will release a large volume of unlabeled tweets from a number of different crisis events, and participant systems are expected to generate information type and priority labels for every tweet in these datasets. For evaluation, TREC-IS coordinators will pool results from all participant systems and sample according to information-type scores provided by the participants. NIST assessors will then evaluate a subset of these pooled messages, and participant systems will be assessed against these manually assessed subsets.

Task Metrics

To evaluate the performance of participant systems, we currently report two groups of metrics, namely: *Information Feed* and *Prioritization*.

For *Information Feed*, each run will be evaluated by 1) its overall classification accuracy, micro-averaged across events and macro-averaged across information types, 2) its overall F1 score, macro-averaged across all information types and mirco-averaged across events; and 3) its F1 score among six actionable information types.

For *Prioritization*, we report two metrics: 1) its overall prioritization error, micro-averaged across events and macro-averaged across all information types; and 2) a normalized, discounted cumulative gain evaluated across the top 100 tweets, micro-averaged across all test events.

We explain the metrics and reasoning in more detail here.

COVID-Specific Evaluation

While the track has folded the COVID-19 task into the main task, TREC-IS will include a separate evaluation using only COVID-specific events. The metrics in this evaluation will be the same as above, but the events on which these metrics are generated will include only COVID-19-related events.

While all systems will be evaluated on this subset, participants wishing to optimize their systems for COVID-19 may be particularly interested in this evaluation.

Training Examples

Participants can use assessor data and events from any prior TREC-IS edition to evaluate (or train if using machine learned approaches) their systems. For each of the previous 2018, 2019, and 2020 events, we provide tweet streams and 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.

Baselines

This edition will include several baseline systems for evaluation, to include: random baselines for information categorization and prioritization, a dictionary-based baseline for information categorization, and a prioritization baseline based on ordering tweets by the average information type priority (consult the 2019 Overview paper here for a visualization of and more information on average information type priority). We expect this last baseline to be particularly strong and suggest participants make use of information type when calculating priority.

Ontology

Along with the event tweet stream, we 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.

The 25 high-level information types, in alphabetical order, are:

- CallToAction-Donations
- CallToAction-MovePeople
- CallToAction-Volunteer
- Other-Advice
- Other-ContextualInformation
- Other-Discussion
- Other-Irrelevant
- Other-Sentiment
- Report-CleanUp
- Report-EmergingThreats
- Report-Factoid
- Report-FirstPartyObservation
- Report-Hashtags

- Report-Location
- Report-MultimediaShare
- Report-News
- Report-NewSubEvent
- Report-Official
- Report-OriginalEvent
- Report-ServiceAvailable
- Report-ThirdPartyObservation
- Report-Weather
- Request-GoodsServices
- Request-InformationWanted
- Request-SearchAndRescue

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.v4.json

Timeline

Guidelines released	8 March 2021
TRECIS-CTIT-H 2021-A Test release	12 March 2021

TREC-IS 2020-B Participant Workshop	26 March 2021
Runs due	3 May 2021
Scores returned to participants	1 June 2021