TREC 2020-B Incident Streams Track

Guidelines v3.0, 26 August 2020

Coordinators:

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

Changelog:

- V3.1 - Updated event types and reverted the priority label to a score rather than a label
- V3.0 - Updated 2020-A guidelines for 2020-B, which will use pooling for evaluation, changes the run submission format, and updates the tasks
    - The regions for Task 3, on COVID-19 data, has been modified
    - Task 3 information types have been updated
    - Changes to *Dataset*
        - To reflect pooling
    - Changes to *Submission* section:
        - 2020-B test event names are updated
        - The run submission format has been modified, replacing the label string with a vector of label scores, which we need for pooling
    - Changes to *Task Assessment*
        - 2020-B will transition fully to a pooling approach for participant evaluation
    - Changes to *Task 3 Training Data*
        - COVID-19-specific training data is now available
    - Changes to *Timeline*
        - Dates have been updated
        - Added a TREC-IS workshop to the schedule
- V2.1 - 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.

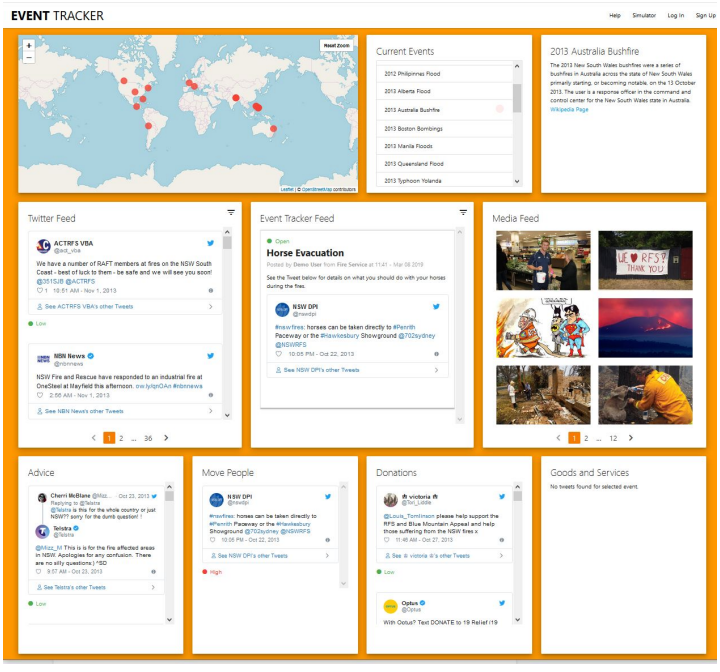
Much of this content might be expressions of sentiment, solidarity, and wishes to help from around the world, but more valuable than those are reports from news services and government officials that contain useful information for people in the area of the incident. Meanwhile, the most relevant information 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 2020-B edition of TREC-IS, we are maintain the task set we introduced in 2020-A (the main task from TREC-IS 2019, plus two additional tasks):

- Task 1. All High-Level Information Type Classification, v2.1
- Task 2. Selected High-Level Information Type Classification, v1.0
- Task 3. COVID-19-Specific Information-Type Classification, v2.0

Regardless of the task(s) in which a research group participates, submissions to TREC-IS will have the same form, as described in the Submission section.

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.

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.).

Task 2. Selected High-Level Information Type Classification, v1.0

A common concern in Task 1 is that the number of high-level information types makes it difficult to dive deeply into the labels. Obtaining a deeper understanding of these labels appears key to a high-performing system, however, as systems with strong feature engineering have performed highly in previous TREC-IS editions. To address this issue, TREC-IS 2020-B will continue 2020-A's Task 2, a restricted version of Task 1 that focuses only on 11 of the high-level information types. These 11 include the top six information types labeled as "actionable" in previous editions (i.e., the types that have, on average, the highest priority) as well as five other types selected from the full set used in Task 1. These types, along with a default "other" type, are as follows:

- Request-SearchAndRescue
- CallToAction-MovePeople
- Report-NewSubEvent
- Report-EmergingThreats
- Report-ServicesAvailable
- Request-GoodsServices
- Request-InformationWanted
- CallToAction-Volunteer
- Report-Location
- Report-FirstPartyObservation
- Report-MultimediaShare
- Other

This task will use the same datasets as Task 1. As in Task 1, each tweet should be assigned as many categories from this restricted set 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, etc.).

Task 3. COVID-19-Specific Information-Type Classification, v2.0

The above tasks continue the prior TREC-IS work on classifying information for wildfires, earthquakes, floods, typhoons/hurricanes, bombings, and shootings. With the current and

massively impactful global health crisis around COVID-19, however, it the track is also trying to provide public health officials and emergency response officers with additional tooling and evaluation data for future public health emergencies or resurgence of COVID-19. To this end, TREC-IS 2020-B continues the COVID-19 task specific to social media data about the pandemic. In 2020-A, this task restricted the set of information types as the track lacked training data for public health events. For 2020-B, we are removing the category restriction and making a collection of COVID-19 labels available, collected and manually assessed from three regions: Seattle, WA, Washington, DC, and New York City, NY.

For 2020-B, we will share a larger dataset for COVID-19, again focusing on geographically constrained areas:

- Miami, Florida
- Jacksonville, Florida
- Houston, Texas
- Phoenix, Arizona
- Atlanta, Georgia
- New York City, New York
- Seattle, WA
- Melbourne, Australia
- New Zealand

As in Tasks 1 and 2, 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 seven different types: wildfires, earthquakes, floods, typhoons/hurricanes, bombings, shootings, and public health emergencies (added in 2020-A). We also include data for general storms, hostage situations, (non-wildfire) fires, explosions, and tornadoes in 2020-B.

Departing from prior TREC-IS editions, however, we will release larger datasets for 2020-B 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 2020-B Test’ (Classifying Tweets by Information Type High-Level 2020-B 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. 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 0.7 [0.0,0.0,0.0,1.0,...] myrun
TRECIS-CTIT-H-Test-022 Q0 991855886363541507 2 0.51 [0.1,0.1,0.1,0.7,0.0,...] myrun
...
TRECIS-CTIT-H-Test-022 Q0 991855942093291520 863 0.32 [0.3,0.3,0.3,0.0,...] myrun
TRECIS-CTIT-H-Test-023 Q0 992010886465314816 1 0.637 [0.0,0.0,0.2,0.0,0.2,...] 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 “priority” label 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.
6. The sixth field is a JSON array of predicted probabilities for the **information types** within the ontology, ordered alphabetically. 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.
    b. NOTE: Any 2020-A system output can be converted to this output format by converting the array of labels to an array of binary values, where columns associated with the returned labels are set to 1 and all other columns are set to 0.
7. The seventh field is the **runtag**, 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 should **categorize all tweets for each event** (this is important to enable future analysis of systems).

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). In departure from prior TREC-IS editions, 2020-B will rely solely on pooling for evaluation. 2020-B 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 some 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.

Training Examples for Tasks 1 and 2

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-A 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.

Training Examples for Task 3

Following TREC-IS 2020-A, we released three sets of regionally targeted COVID-19 tweets, a subset of which has been manually assessed in a manner similar to the data from Tasks 1 and 2. We are releasing these assessments, and participants are welcome (and encouraged) to use this data to evaluate or train their systems.

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 are:

- Request-GoodsServices
- Request-SearchAndRescue
- Request-InformationWanted
- CallToAction-Volunteer
- CallToAction-Donations
- CallToAction-MovePeople
- Report-FirstPartyObservation
- Report-ThirdPartyObservation
- Report-Weather
- Report-EmergingThreats
- Report-MultimediaShare
- Report-ServiceAvailable
- Report-Factoid
- Report-Official
- Report-CleanUp
- Report-Hashtags
- Report-News
- Report-NewSubEvent
- Report-Location
- Other-Advice
- Other-Sentiment
- Other-Discussion
- Other-Irrelevant
- Other-ContextualInformation
- Other-OriginalEvent

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	3 August 2020
TRECIS-CTIT-H 2020-B Test release	10 August 2020
TREC-IS 2020-B Participant Workshop	10 August 2020
Runs due	24 September 2020
Scores returned to participants	November 2020