Incident Streams 2021 off the Deep End: Deeper Annotations and Evaluations in Twitter

Cody Buntain*
University of Maryland, College Park (UMD)[†]
cbuntain@umd.edu

Richard McCreadie
University of Glasgow[‡]
richard.mccreadie@glasgow.ac.uk

Ian Soboroff
National Institute of Standards and Technology (NIST)[§]
ian.soboroff@nist.gov

ABSTRACT

This paper summarizes the final year of the four-year Text REtrieval Conference Incident Streams track (TREC-IS), which has produced a large dataset comprising 136,263 annotated tweets, spanning 98 crisis events. Goals of this final year were twofold: 1) to add new categories for assessing messages, with a focus on characterizing the audience, author, and images associated with these messages, and 2) to enlarge the TREC-IS dataset with new events, with an emphasis of deeper pools for sampling. Beyond these two goals, TREC-IS has nearly doubled the number of annotated messages per event for the 26 crises introduced in 2021 and has released a new parallel dataset of 312,546 images associated with crisis content – with 7,297 tweets having annotations about their embedded images. Our analyses of this new crisis data yields new insights about the context of a tweet; e.g., messages intended for a local audience and those that contain images of weather forecasts and infographics have higher than average assessments of priority but are relatively rare. Tweets containing images, however, have higher perceived priorities than tweets without images. Moving to deeper pools, while tending to lower classification performance, also does not generally impact performance rankings or alter distributions of information-types. We end this paper with a discussion of these datasets, analyses, their implications, and how they contribute both new data and insights to the broader crisis informatics community.

Keywords

Emergency Management, Crisis Informatics, Twitter, Categorization, Prioritization, Multi-Modal, Public Safety, PSCR, TREC

INTRODUCTION

The TREC Incident Streams track was started in 2018 to foster research in assessing, classifying, and notifying public safety about information posted to social media relevant to emergency, crisis, and disaster response. The public has an expectation that the public safety community will monitor and respond to social media postings on common platforms. However, the amount of information the public posts to social media is beyond the ability of the public safety community to manually monitor in real time during emergencies.

Over the four years of the TREC-IS initiative,[1] the track has contributed a dataset of Twitter content and manual assessment of this content's *types* and *priorities* to the crisis informatics community. Coupling this dataset

[*] corresponding author

[†] https://ischool.umd.edu/ and http://cody.bunta.in/

[‡] http://gla.ac.uk and http://dcs.gla.ac.uk/~richardm

[§] https://www.nist.gov/ and https://www.nist.gov/people/ian-soboroff

[1] An overview of the TREC-IS initiative is presented on page 4.

with participation from numerous international research groups, TREC-IS has also demonstrated a frontier in state-of-the-art information retrieval systems for crisis informatics. Much of this progress in TREC-IS and crisis informatics more generally, however, has been limited to text-based information systems despite calls for more wholistic methods from stakeholders in the disaster response space (Castillo, Peterson, et al. 2021). While multimodal resources like CrisisMMD (Alam, Ofli, and Imran 2018) now supplement these text-oriented systems with visual data, limited resources exist to connect this visual media back to information needs, and even fewer resources support integrating a message's broader contextual cues like its intended audience and the credentials of its author. Furthermore, the expense of manually extracting these cues and attributes from crisis-relevant social media content creates a tension between annotating large volumes of content for a few crisis events or relatively few messages for many crisis events. TREC-IS has traditionally prioritized assessment across many events at the expense of fewer assessments per-event, to support evaluation of systems across a wide range of crises, operationalized either via a conservative initial sampling strategy (McCreadie et al. 2020) or via relatively shallow pooling of tweets from participant systems (C. Buntain et al. 2021). This pooling approach, however can lead to incomplete assessment datasets (Voorhees 1998), as important documents might not be shown to assessors; hence if a system returns these omitted documents, the system will receive no credit for them, introducing error into system scoring. Likewise, these pools may introduce bias in favor of content that appears in multiple events, as learning frameworks likely have more training data for such content compared to more crisis-specific messages. In this final year of TREC-IS, we address these open problems via four main enhancements to the track and associated dataset: release of more tweets per event, deeper pools of messages for assessment, more detailed assessments of these messages, and provision of a parallel image dataset:

More Tweets Per Event: To support 2021 and continuing the expanded collections begun in 2020-B, we have collected larger tweet samples for natural disasters and man-made crises. This collection is enabled through an agreement with Twitter, and from it, we have released a larger portion of crisis-event samples to the community. Consequently, the 2021 dataset provides 34,742 tweets on average per event, in contrast to 19,636 in 2020-B and 500 in 2020-A.

Deeper Pools: In 2021, we have assessed more tweets per event, both to improve the robustness of our evaluation and to evaluate the impact of pool depth on our ability to compare systems (particularly in the top ranks). While the track had moved to pooling in 2020 to expand the data available for evaluation, depths of these pools were necessarily shallow to allow for evaluation across 41 crises in two editions; in 2021, however, we have held constant the 26 crisis events across our two editions (2021-A and 2021-B) and instead extract deeper pools of labeled content within these fewer crises. As a result, TREC-IS 2021 has produced nearly double the number of labeled messages per crisis event (averaging 2,555 labeled tweets in 2021 compared to 1,378 in 2020-B, when we first introduced pooling, and 1,374 per event in 2018-2020). Having more tweets assessed makes the evaluation of systems more robust as well as creating a more valuable dataset.

More Labels per Tweet: For more detailed assessments of crisis messages in social media, we have expanded the types of information we solicit from our human assessors. In past iterations of the track, these assessors have provided information only about the type of information and the priority of content posted during crises. In 2021, however, we have expanded this set of annotations, chiefly to include assessments of the *images* shared with crisis content, *intended audience* (local or global) and on the *author account* (whether the account belongs to a politician, private citizen, news organization, etc.). This inclusion of image data is of particular interest as practitioners have consistently called for methods that account for visual media shared during crises (Castillo, Peterson, et al. 2021), as seen in the visually oriented CrisisMMD dataset (Alam, Ofli, and Imran 2018).

Parallel Image Dataset: As part of our collecting additional label categories per tweet, for tweets with images embedded in them, we characterize the types of these images. The 2021 dataset therefore contains manual assessments of 7,297 tweets with images. Additionally, to support larger assessments of images in crisis content, we have collected and released a dataset of all the images associated with the 122 crisis events in the TREC-IS collection (including events for which we have no manual labels), totalling 312,546 images.

In total, the 2021 editions of this track have released 1,530,856 crisis-related Twitter messages across 47 crises. Participants from five different research groups have submitted 33 systems (or "runs" in TREC terminology), and assessment from these systems (after pooling and event down-selection) has produced 66,437 newly labeled tweets from 26 crises – TREC-IS now cumulatively provides 136,263 unique labeled tweets spanning 98 crisis events. Analyses of these datasets suggest participant systems in the 2021-A and 2021-B editions are highly correlated, indicating assessment in our deeper pools broadly maintains rankings, but performance metrics are generally higher for the shallower pools. Following our detailed per-message assessment, including analysis of authors, intended audience, and imagery, we find insights about potentially useful contextual information: On average, tweets from governmental organizations receive higher priority assessments compared to other authors, and content intended for local audiences similarly has higher perceived priority than content produced for general audiences. Regarding imagery, TREC-IS 2021 has produced 10,536 assessments of image-laden tweets, with results showing that tweets with images receive higher average priority assessments (between medium- and high-priority) compared to text-only messages, with images of weather forecasts having the highest average priority. Surprisingly, despite work on damage estimation from social media imagery (e.g., (Alam, Ofli, and Imran 2018)) assessors have rated images of damage as the lowest priority.

Before diving into these results, however, this paper firsts outlines the TREC-IS initiative and how this final year fits with extant crisis informatics research, which then continues into a description of substantive changes in TREC-IS 2021. After describing this year's results and discussing their implications for the space, we then end with a brief retrospective at the end of the TREC-IS initiative.

Related Work in Crisis Informatics

As crisis informatics has matured, numerous datasets and data resources have become available for study. Early multi-crisis datasets like CrisisLex (Olteanu, Castillo, et al. 2014) provide crisis content collected through text-based queries that describe the crisis (e.g., "powerful storms", "victims", and others). Many messages in these datasets may not be relevant to a crisis event, however, as lower-signal query terms like "cleanup" may occur in many irrelevant messages, so the datasets often require additional filtering. Prior TREC-IS editions have expanded these query-based methods with human assessors, who label social media content according to the TREC-IS information-type ontology, so researchers can filter out content tagged as "Irrelevant" – in fact, the first TREC-IS edition in 2018 leveraged events from the CrisisLex collections as an initial training set to bootstrap the track. Despite these foundations and their extensions, resources like CrisisLex and early TREC-IS datasets are fundamentally limited, relying only on these messages' textual characteristics for both collection and characterization. Studies of stakeholder information needs in the crisis informatics space demonstrate such text-only views are too narrow, as disaster-response personnel need information about the multiple modalities and cues in social media (Castillo, Mendoza, et al. 2013), such as imagery and author identity.

While other datasets do provide these *multi-modal* resources, these sources are either small-scale – such as the Social Media for Emergency Response and Preparedness (SMERP) initiative (Ghosh et al. 2019) – or focus primarily on imagery and information needs related to images shared during crises – such as the CrisisMMD multi-modal dataset (Alam, Ofli, and Imran 2018). Similar to CrisisLex and TREC-IS, the SMERP initiative has released a dataset of crisis-related social media content around the 2015 Nepalese earthquake, but SMERP departs from these sources by augmenting this dataset with a collection of images posted alongside crisis-related Twitter messages. Though the SMERP is therefore multi-modal, its focus on a single event and therefore single event *type* leaves questions about how well visual analytics models might generalize to other crisis events – as we know text-based methods do (C. Buntain et al. 2021). The CrisisMMD dataset, on the other hand, is closer to an ideal multi-modal collection, as it contains – like SMERP – social media messages *and* imagery from numerous crisis events, and these messages/images include human annotations – like TREC-IS. CrisisMMD is also part of a larger initiative called CrisisNLP, which maintains many datasets related to social media and crises (e.g., Alam, Ofli, Imran, et al. 2020; Alam, Sajjad, et al. 2021). CrisisMMD's annotations, and many of the datasets in CrisisNLP, focus primarily on damage assessment, informativeness, and identifying humanitarian needs, however, which may omit other critical information about the event, such as alerts about new threats, calls for evacuation, or shelter availability (all elements identified in prior TREC-IS years).

For the 2021 edition of TREC-IS, we have attempted to fill this need for multi-modal crisis data by providing both the standard information-type and priority annotations that TREC-IS has used in prior editions *and* more details assessments of the messages and visual images embedded therein. As a result, researchers can connect the types of images embedded in, authors of, and audiences for social media messages to the types and criticality of information in the message, to provide a more wholistic view of how imagery augments the information in a message.

Beyond this need for additional detail about the social media content we have, as the field of crisis informatics grows, we increasingly need insight about cross-event versus event-specific information. With the above datasets, collections are either focused on single events (as with SMERP) or broadly cover multiple events (CrisisLex, CrisisMMD or TREC-IS). While the latter cross-crisis collections might prove more valuable in understanding general phenomena of crises and social media, these methods prioritize generalizability at the expense of deeper insights within the crisis events studied. Consequently, understanding content in the "long tail" of an event might be difficult with systems trained on these general datasets. CrisisLex, for example, has developed a lexicon for collection across many crises, explicitly prioritizing precision over recall (Olteanu, Castillo, et al. 2014). Likewise, in pooling-based approaches like that in TREC-IS 2020, the construction of these pools – and therefore evaluation – relies on identifying messages for which systems have provided "confident" labels. Messages for which systems

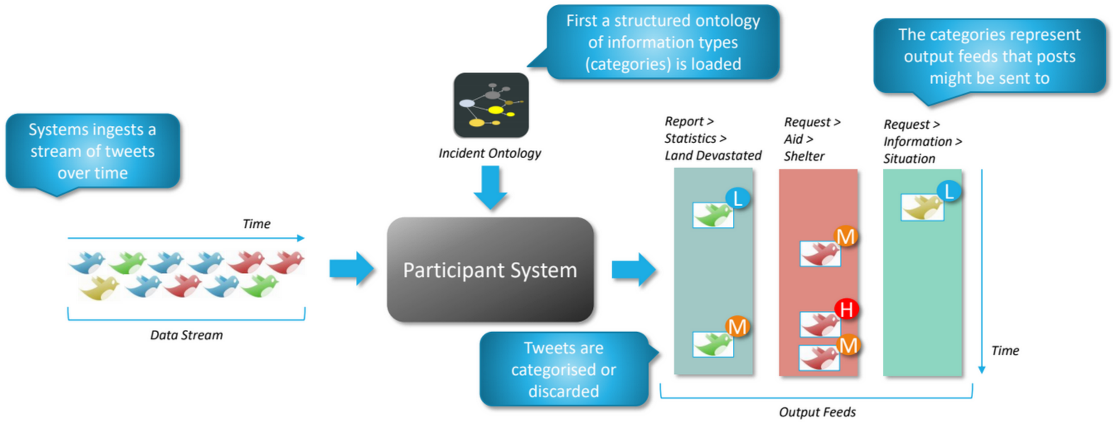

First a structured ontology of information types (categories) is loaded
The categories represent output feeds that posts might be sent to
Systems ingests a stream of tweets over time
Incident Ontology
Report > Statistics > Land Devastated
Request > Aid > Shelter
Request > Information > Situation
Time
Participant System
Data Stream
Tweets are categorised or discarded
L
M
H
Time
Output Feeds

Figure 1. TREC-IS Task Visualization and Concept of Operations

are confident are likely those messages that appear similarly in multiple crises (e.g., location- and sentiment-laden messages are common in crises), meaning that pooling evaluations may miss content that is unique to particular crises or event-types, as common, cross-crisis content has more samples and systems can therefore generalize better. This possibility is particularly concerning in spaces like Twitter, where prior research on the news value of the platform shows Twitter is better at long-term and long-tail coverage that mainstream news media might miss (Petrovic et al. 2013). In 2020, we partially addressed this issue by pooling within specific information types, but in 2021, we have sought to enable more event-specific content analysis by doubling the depths of our evaluation pools.

AN OVERVIEW OF TREC-IS

Having established the context for this work, we now turn to an overview of the TREC-IS initiative as a whole. Prior studies have consistently shown useful, actionable, and critical information about an ongoing crisis is shared on social media in the lead-up to and aftermath of the crisis (McCreadie et al. 2019; McCreadie et al. 2020; Purohit et al. 2018). TREC-IS is a data challenge that aims to support the development of tools to extract this actionable information. To this end, TREC-IS defines a task, to which researchers and practitioners can submit solutions, where the effectiveness of those solutions can be compared. *The core TREC-IS task is to build a system to analyse a stream of social media posts and assign 'information type' and 'priority' labels to each, as illustrated in Figure 1.*

The role of a data challenge organiser is primarily to create a re-usable and robust dataset and evaluation methodology that facilities sound evaluation of participant solutions. By doing so, it becomes possible to compare solutions, identify what strategies are effective, and track progress toward the ultimate goal of building a deployable solution for the task. However, through the process of dataset construction (usually based on significant human-labelling efforts), it is often possible to gain additional valuable insights into the task itself.

Below, we provide a brief overview of the mechanics of the TREC-IS track as context and foundations for the subsequent discussion on changes we have made in 2021. For more details, we refer the reader to the overview papers for each year of the track: 2018 (McCreadie et al. 2019); 2019 (McCreadie et al. 2020); 2020 (C. Buntain et al. 2021).

Foundations of the TREC-IS Framework

Crisis-Related Information Types: A core component of the TREC-IS task is the assignment of 'information type' labels to each tweet. To support this assignment, the track defines an ontology of 25 information types, shown in Table 1. Some of these labels represent categories of information that may be of interest to emergency response officers, such as 'Reports of Road Blockages' or 'Calls for Help'. Other labels represent information that may be valuable to researchers and practitioners engaged in a variety of other crisis-related tasks: e.g., public-health surveillance, volunteer coordination, or other support functions, as outlined in the Federal Emergency Management Agency's Emergency Support Functions (ESFs).

This ontology is similar to other crisis-related ontologies, as it overlaps with the annotation scheme introduced in Imran et al. 2016, which itself builds on the scheme used by the United Nations Office for the Coordination of Humanitarian Affairs. Likewise, this ontology overlaps with the "information type" category in Olteanu, Vieweg, et al. 2015. The TREC-IS ontology subsumes several of these information types such that the annotations used herein should be directly applicable to these related annotation schemes as well.

Based on analysis of this ontology in preceding editions of the track (McCreadie et al. 2020), we have further tagged six of these information types as "actionable", or of high importance (based on a ranking of types by mean message priority): *'Request-GoodsServices'; 'Request-SearchAndRescue'; 'CallToAction-MovePeople', 'Report-EmergingThreats'; 'Report-NewSubEvent', and 'Report-ServiceAvailable'.*

High-Level Information Type	Description	Example Low Level Types
Request-GoodsServices⋆†	The user is asking for a particular service or physical good.	PsychiatricNeed, Equipment, ShelterNeeded
Request-SearchAndRescue⋆†‡	The user is requesting a rescue (for themselves or others)	SelfRescue, OtherRescue
Request-InformationWanted†‡	The user is requesting information	PersonsNews, MissingPersons, EventStatus
CallToAction-Volunteer†‡	The user is asking people to volunteer to help the response effort	RegisterNow
CallToAction-Donations	The user is asking people to donate goods/money	DonateMoney, DonateGoods
CallToAction-MovePeople⋆†‡	The user is asking people to leave an area or go to another area	EvacuateNow, GatherAt
Report-FirstPartyObservation†	The user is giving an eye-witness account	CollapsedStructure, PeopleEvacuating
Report-ThirdPartyObservation	The user is reporting a information from someone else	CollapsedStructure, PeopleEvacuating
Report-Weather	The user is providing a weather report (current or forecast)	Current, Forecast
Report-EmergingThreats⋆†‡	The user is reporting a potential problem that may cause future loss of life or damage	BuildingsAtRisk, PowerOutage, Looting
Report-MultimediaShare†	The user is sharing images or video	Video, Images, Map
Report-ServiceAvailable⋆†‡	The user is reporting that someone is providing a service	HospitalOperating, ShelterOffered
Report-Factoid	The user is relating some facts, typically numerical	LandDevastated, InjuriesCount, KilledCount
Report-Official	An official report by a government or public safety representative	OfficialStatement, RegionalWarning, PublicAlert
Report-CleanUp	A report of the clean up after the event	CleanUpAction
Report-Hashtags	Reporting which hashtags correspond to each event	SuggestHashtags
Other-News	The post providing/linking to continuous coverage of the event	NewsHeadline, SelfPromotion
Report-NewSubEvent⋆†‡	The user is reporting a new occurrence that public safety officers need to respond to.	PeopleTrapped, UnexplodedBombFound
Report-Location†	The post contains information about the user or observation location.	Locations, GPS coordinates
Other-Advice‡	The author is providing some advice to the public	SuggestBestPractices, CallHotline
Other-Sentiment	The post is expressing some sentiment about the event	Sadness, Hope, Wellwishing
Other-Discussion	Users are discussing the event	Causes, Blame, Rumors
Other-Irrelevant	The post is irrelevant, contains no information	Irrelevant
Other-ContextualInformation	The post is generic news, e.g. reporting that the event occurred	NewsHeadline
Other-OriginalEvent	The Responder already knows this information	KnownAlready

⋆ – "Actionable" Information Types, † – Task-2 Information Types, ‡ – COVID-19 Information Types (2020-A only)

Table 1. Ontology of High-level Information Types

Measuring Message Criticality: The second component of the TREC-IS task is to assign scores of importance – or critically – to messages, representing how imperative it is that a given message be seen by emergency response officer quickly. Hence, for each tweet, participant systems submit a criticality score (a number between 0 and 1), and this score is later mapped into a discrete criticality label: *'Low' (<=0.25), 'Medium' (>0.25 & <=0.5), 'High' (>0.5 & <=0.75), or 'Critical' (>0.75).*

TREC-IS Editions: Each year, TREC-IS issues one or more 'editions', wherein participants can submit the output of their solution(s) – or systems – to the TREC-IS organisers. Each edition has a set of training events and a set of test events. For training, track coordinators provide a collection of tweets and their respective information type/criticality labels (typically all events from prior editions). For testing and evaluation, coordinators provide only the tweet stream, and participants must then produce a set of labels for each tweet in the stream; i.e., the predicted information types and critically labels. Collectively, TREC refers to the files listing these tweets and their labels as a 'run'. The organisers then evaluate each run and return effectiveness scores to each participant.

Evaluating a Run: To evaluate participant systems, we require ground-truth data on information-type and criticality labels for the above tweets. To create this ground truth, TREC coordinators at the National Institute of Standards and Technology (NIST) pay human assessors to analyse tweets and assign information-type and criticality labels, mirroring the TREC-IS task that participants perform. TREC-IS coordinators then calculate each run's effectiveness by comparing human labels against the labels produced by the participant solution (see the Metrics section below).

Run Pooling for an Edition: As discussed above, to facilitate the evaluation of a run, we need human assessment to produce labels for tweets. This assessment process is time-consuming and costly, and given the large volume of tweets published during large-scale crises, it is impossible to label the whole tweet stream. Since 2020, TREC-IS has used a *pooling* strategy to decide what tweets will be labeled (Zobel 1998). Each of these pools is associated with a particular crisis event, which we construct by selecting a subset of that event's tweets from each submitted run and "pooling" them together. To build these subsets, we construct several rankings of tweets in each run, one ranking for each information type and an additional ranking by priority score. For information types, a ranking is determined by the participant system's "confidence" score in the associated label such that the top tweet in the ranking denotes the tweet-label pair in which the participant run is most confident – where the meaning of "confidence" is left to the participant. For priority, these rankings correspond to a tweet's importance such that higher-ranking tweets receive higher priority scores. From each of these type-confidence and priority-based rankings, we collect the top-k tweets from the target event into the pool for that run, where k denotes this pool's *depth*. Taking the union of all the per-run pools for this event then produces the total pool of tweets from this event that human assessors will review. In effect, for each event, this pooling strategy creates an ordering of tweets to be assessed, where tweets predicted to be of higher priority by multiple systems or have a high probability for particular information types will be assessed first. These rankings are then divided into 500-tweet blocks for assessment.

Assessment Process: For each event, track coordinators generate 8 assessment blocks (4,000 tweets) from the pool ranking and assign these blocks to individual assessors. Where possible, one assessor will label all assessment blocks for an event to maintain consistency. Each assessor has a time-budget for assessment and attempts to label as many tweets within that time-budget as possible. Blocks are assessed in order based on the pool ranking (higher ranked first), and within a block, tweets are ordered by time. While budget and time constraints have not allowed for multiple assessments per tweet, precluding agreement evaluations, this manual assessment has been used consistently in prior TREC-IS iterations (McCreadie et al. 2019; McCreadie et al. 2020; C. Buntain et al. 2021).

Metrics Once the pooled tweets have been labeled, evaluations for these submitted runs use two main metrics: *Information Feed* and *Information Priority*. Information feed metrics measure performance in terms of information type categorization. For each tweet, a participant run will have assigned (potentially multiple) information-type labels, which are then evaluated using standard precision, recall, and F1-score against the human labels, macro-averaged across the test events. As mentioned previously, not all information types are equally important (i.e., messages of sentiment are generally less critical than requests for search and rescue), so we evaluate these metrics both across all 25 information types and the six actionable information types highlighted above in isolation. Meanwhile, the information priority metrics evaluate run performance when assigning criticality scores to each tweet. We compare these predicted criticality scores to assessor-provided categorical labels using both F1 score as a measure of priority classification (discretizing runs' scores to the above 'Low', 'Medium', 'High', and 'Critical' priority labels) and Pearson *r* as a measure of the correlation between predicted scores and numerically mapped priority assessments.

Main Conclusions from TREC-IS 2018, 2019, and 2020

As 2021 is the fourth and last year of the TREC-IS initiative, we briefly describe the key organizational aspects and findings of prior TREC-IS editions, as they provide the foundation for TREC-IS in 2021.

A crucial finding in TREC-IS's 2018 pilot run and confirmed in the 2019 initiative is that a non-trivial (approximately 10% post-filtering) amount of actionable and high-priority information exists in Twitter during emergency events. The average priority score for an emergency-related tweet appears relatively stable between low- and medium-importance across eight years of emergency events (2011-2019) (McCreadie et al. 2020). Further, we identify six stable "actionable" information types such that tweets with these information-types are consistently perceived to be of higher-priority than the other types. An analysis of manually labeled tweets and systems participating also yielded insights into what information is actionable and critical for emergency response officers, as messages that are perceived as high priority also often contain references to particular locations – in line with Purohit et al. (2018) – and also include hyperlinks to external information sources.

While results from 2018 focused on the distribution of information within crises, participant-system evaluations from prior TREC-IS iterations also found that state of the art systems of the time were insufficient for end users' needs in classifying information type and priority. Participants were relatively effective at identifying news reports and sentiment, but they struggled to identify critical information like search and rescue requests (McCreadie et al. 2019). To this end, in 2019, we sought to expand training data availability and develop a clearer picture of system performance and shortcomings. After the 2019 edition, we have found that systems employing neural language models are increasingly performant compared to traditional learning models (McCreadie et al. 2020). Despite this trend, performance in identifying actionable information has lagged behind the identifying all relevant information types, suggesting significant room for improvement remains.

In 2020, TREC-IS expanded its scope to include large-scale public-health emergencies, following the outbreak of COVID-19. The pandemic has introduced an over-abundance of potential data, and open questions remained about how approaches to crisis informatics and datasets built on other emergencies could adapt to this new context. TREC-IS made available 478,110 COVID-related messages and 282,444 crisis-related messages for participant systems to analyze, of which 14,835 COVID-related and 19,784 crisis-related messages have been manually annotated. Analyses of these new datasets and participant systems demonstrated first that both the distributions of information type and priority of information vary between general crises and COVID-19-related discussion. Secondly, despite these differences, results suggest leveraging general crisis data in the COVID-19 context improves performance over baselines.

CHANGES TO TREC-IS IN 2021

In the final year of TREC-IS, we have instituted several changes to gain deeper insights into the types of information contained in social media during crises and address the gaps outlined above. These changes fall across two major axes: surfacing more critical content through deeper evaluation pools and expanding the types of attributes requested from human assessors.

48 New Events with Higher Volume Streams: At the start of 2021, we have released tweet streams for 48 events (TRECIS-CTIT-H-076 to TRECIS-CTIT-H-122), along with their descriptions. These crisis events, names, event-types, sample sizes, and number of manually assessed tweets are shown in Table 2 (some events also include redundant labels, as listed in the "# Multi-Labeled" column). Unlike previous years, only a subset of these events are used to evaluate participant systems. Participants are not told what subset of these events are to be used for evaluation. Instead, the track requires participants to submit runs containing predictions for all 48 events. Thanks to an agreement with Twitter for access to their commercial Historical PowerTrack API (distinct from the public APIs and the new v2 public/academic APIs), we have been able to collect larger samples for each event (typically in the range of millions of tweets), and collect content from historical events rather than only new or contemporary crises. This historical collection is clear in the list of TREC-IS events shown in Appendix A, where crisis events in 2020-A/B were limited to 2019 and 2020, as our collection infrastructure could only collect content from these crises *as they happened*.

In the same way as TREC-IS 2020, we perform further diversity sampling to filter these high-volume tweet-sets, de-duplicating the data using locality-sensitive hashing – via the Nilsimsa digest (Damiani et al. 2004) – to filter out highly similar tweets and retweets. From these filtered sets, the track has released a sample of up to 50,000 tweets per event to participants.

Edition	Identifier	Event Name	Event Type	2021-A	2021-LB	2021-B	Tweet Sample	# Unique Labeled	# Multi-Labeled
2021-A/B	TRECIS-CTIT-H-076	Houston Explosion 2020	explosion	✓		✓	32,204	2,795	-
	TRECIS-CTIT-H-080	Mideast Tornadoes Day 1 Mississippi 2020	tornado			✓	834	772	-
	TRECIS-CTIT-H-081	Mideast Tornadoes Day 2 Alabama 2020	tornado		✓	✓	9,594	3,829	2,185
	TRECIS-CTIT-H-082	Mideast Tornadoes Day 3 MD 2020	tornado		✓	✓	13,641	1,109	-
	TRECIS-CTIT-H-083	Tennessee Derecho 2020	storm		✓	✓	26,946	2,745	6
	TRECIS-CTIT-H-085	Brooklyn Block Party Shooting 2019	shooting			✓	49,593	3,592	-
	TRECIS-CTIT-H-089	2016 Puttingal Temple	explosion	✓		✓	18,888	3,182	384
	TRECIS-CTIT-H-091	Thomas Wildfire 2017	wildfire	✓		✓	50,000	3,000	-
	TRECIS-CTIT-H-092	Lilac Wildfire 2017	wildfire	✓		✓	50,000	2,342	-
	TRECIS-CTIT-H-096	Klamathon Wildfire 2018	wildfire			✓	2,094	1,583	-
	TRECIS-CTIT-H-097	Holy Wildfire 2018	wildfire			✓	49,664	3,336	-
	TRECIS-CTIT-H-100	Woolsey Wildfire 2018	wildfire			✓	49,167	1,588	-
	TRECIS-CTIT-H-101	2018 Maryland Flood	flood	✓		✓	50,000	2,975	552
	TRECIS-CTIT-H-102	2018 Pittsburgh Synagogue Shooting	shooting			✓	49,341	1,755	-
	TRECIS-CTIT-H-103	Alberta Wildfire 2019	wildfire			✓	5,612	2,322	-
	TRECIS-CTIT-H-104	Hurricane Dorian 2019	hurricane			✓	49,100	2,219	-
	TRECIS-CTIT-H-106	Saddleridge Wildfire 2019	wildfire			✓	49,623	2,168	-
	TRECIS-CTIT-H-107	Kincade Wildfire 2019	wildfire		✓	✓	49,495	4,177	1,034
	TRECIS-CTIT-H-108	2019 Durham Gas Explosion	explosion		✓	✓	26,891	3,476	1,061
	TRECIS-CTIT-H-110	2019 Saugus High School Shooting	shooting	✓		✓	43,363	2,174	-
	TRECIS-CTIT-H-112	2019 Townsville Flood	flood			✓	16,542	2,925	-
	TRECIS-CTIT-H-116	2020 Easter Tornado Outbreak	tornado	✓		✓	49,403	2,183	12
	TRECIS-CTIT-H-119	2020 Tornado Outbreak of April	tornado		✓	✓	41,184	3,421	-
	TRECIS-CTIT-H-120	2020 Tornado Outbreak of March	tornado			✓	48,511	1,717	-
	TRECIS-CTIT-H-121	2020 Visakhapatnam Gas Leak	explosion		✓	✓	48,766	3,657	998
	TRECIS-CTIT-H-122	Tornado Outbreak of December 2018	tornado			✓	22,832	1,395	-

Table 2. TREC-IS 2021 Events Labelled for 2021-A, 2021-B and the Online Leaderboard (2021-LB).

Two New Editions (2021-A and 2021-B) and a Leaderboard: As with previous years, we have organised two official editions for the track in 2021, denoted 2021-A and 2021-B along with an additional leaderboard release. 2021-A uses 7 events for evaluation, which we then augment with an additional 7 events for a public leaderboard-based evaluation – these additional events are necessary for true held-out test data since 2021-A assessments are returned to the track participants. 2021-B then uses 26 events, 14 of which are expanded assessments from the 2021-A and leaderboard events. Crisis events comprising the 2021-A, leaderboard (LB), and 2021-B are shown in Table 2. Unlike previous years, however, events are shared across these editions. Motivation for this event sharing is twofold: First, it enables like-for-like comparison of 2020-A systems with 2020-B systems, allowing for the tracking of progress of systems throughout the year. Second, it allows us to investigate the effect that increasing the depth of assessment for some events has on system rankings, an important factor when considering dataset robustness.

Beyond the two traditional TREC-IS 2021 editions, we have released a public leaderboard for the track. This leaderboard is meant to expand participation and provide a mechanism for ongoing participation after the track officially ends at TREC. This online leaderboard is hosted at GitHub, and in this new evaluation approach, participants need only submit pull requests to the official TREC-IS GitHub repository (along with metadata describing their submission). The pull request triggers a notification to TREC-IS organizers, and we can quickly run evaluations using private evaluation data from the 2021-B and leaderboard crisis events. Organizers post these resulting evaluation metrics to the public leaderboard data file in the GitHub repository. This leaderboard is available at:

- `https://trecis.github.io/`

26 New Events Labelled: Following previous years, we have had human assessors manually label tweets for these events according to the TREC-IS information type ontology, as well as assign criticality labels. As mentioned earlier, a pooling strategy is applied over the submitted runs to each edition of the track to determine which tweets will be labeled - focusing on tweets predicted to be of higher priority as scored by multiple runs and/or have a high confidence score for particular information types. Over the year, we have had 26 of those events labelled by human assessors, with assessment statistics shown in Table 2. As the last year of the track, we also allocated a larger assessment budget for 2021, resulting in a markedly larger set of labelled tweets than in previous years (see Appendix A). In particular, 66,437 unique tweets have been labelled in 2021, which also translates into a higher number of tweets per-event being assessed (averaging 2,555 labeled tweets in 2021 compared to 1,378 in 2020-B).

Variation in Pool Depth: For 2021, we have re-used the pooling strategy described above but have altered the pool depths. Following from larger event datasets and an increased number of candidate crises released (approximately 50,000 In both 2021-A and 2021-B across 26 crises), we have started with shallow pools for 2021-A and the leaderboard, pooling to a depth of 10 tweets per information type and priority per system. This limited depth has provided approximately 1,500-2,300 messages to evaluate per crisis. To examined pooling depth's impact of evaluation, for 2021-B, we have increased pool depth to 20 for a majority of our crisis events, approximately doubling the volume of evaluated tweets per event.

21 New Categories of Labels Collected: As discussed in the introduction and related work, a crucial gap in TREC-IS assessment and similar resources concerns limited detail around the context of individual messages. For example, prior TREC-IS editions (McCreadie et al. 2019) have suggested information about the author of a message can provide useful insight into the criticality of information and, potentially, its veracity (e.g., content from a government organization is likely more trustworthy than content from a private citizen). Prior TREC-IS editions have not recorded such information, however, despite its potential value. As a result, in TREC-IS 2021-A, we have expanded the set of annotations we ask our assessors to provide with the goal of expanding the contextual information around these messages. In particular, we have added three new types of assessment information regarding 1) intended audiences of a particular message, 2) insights about the author writing the message, and 3) assessments of images associated with the message. These types are detailed in Table 3.

Our inclusion of annotations for imagery shared with crisis-related messaging is intended to shed light on the broad class of images that appear in crisis content. For example, preliminary studies from data in earlier TREC-IS editions has suggested infographics and screen shots are quite common, yet the value of such images is unclear. Based on preliminary analysis and integrating studies like CrisisMMD (Alam, Ofli, and Imran 2018), we have developed a set of 10 image-related categories that we ask our assessors to include for image-laden tweets. We note here though that these image labels are associated with the tweet rather than the individual image, as a tweet can contain multiple images. As such, a single tweet may have multiple relevant image labels. This additional assessment has produced image labels for 7,297 tweets.

Insight into intended audience and author are meant to provide context about the source and target of this content. We anticipate content intended for local audiences (e.g., requests for search and rescue, notices about new services, or calls for evacuation) to be potentially more actionable than content for a more global/general audience (e.g., sentiment, news coverage, factoids, etc.). Likewise, the source of a message may provide substantial insight into its veracity and actionability, as information from government organizations, officials, and journalists likely has different connotations than content from private citizens or bot-like automated weather reports. These additional assessments are time-intensive for assessors, however, so we have only included them in the 2021-A round of assessment. This additional assessment has produced audience labels for 1,021 tweets and author labels for 1,104 tweets.

New Parallel Image Corpus: Finally, to support the inclusion of multi-modal data, and particularly visual media, into TREC-IS and crisis informatics, we have made available a large-scale collection of images extracted from all 122 TREC-IS events. The vast majority of these images have no labels associated with them but can be used for unsupervised learning and clustering to evaluate the kinds of images present in crisis-related content. Instead, we have simply iterated through the sample tweets for every event in our dataset (see Appendix A for this list) and, where a tweet's metadata includes image content, we have collected these images. In total, this parallel dataset

High-Level Information Type	Description
Image-Satellite	At least one image appears taken from a satellite
Image-WeatherForcast	At least one image shows a weather forecast, either graphic or satellite projection
Image-Infographic	At least one image shows textual graphic with information or statistics about the event
Image-WeatherPhoto	At least one image shows a photograph of some weather phenomenon (e.g., tornado, wildfire, etc), separate from weather forecasts
Image-Victims	At least one image is a photograph showing victims of or someone who was directly impacted by the crisis
Image-Damage	At least one image is a photograph of damage of infrastructure, homes, vehicles, landscape, or similar
Image-Crowds	At least one image shows a group of people
Image-RecoveryEfforts	At least one image is a photograph of recovery efforts, such as a rescue operation, donations, or volunteering efforts
Image-Screenshot	At least one image shows a screenshot taken from a phone, computer, or video
Image-Informative	A meta label indicating the assessor found seeing the image to be informative
Audience-Unknown	This post does not have a clear intended audience
Audience-Local	This post provides community-level information or information for individuals directly impacted by the event
Audience-General	This post provides general information, e.g., statistics, outlook, etc. intended for those not directly impacted by the event
Author-NewsOrg	The account represents media outlets such as local news affiliates, newspapers, or national/global news organizations
Author-GovtOrg	The account represents a local, regional, or federal governmental agency
Author-PrivateOrg	The account represents a corporations, NGO, or community organization
Author-PublicOfficial	The account represents a politician, elected/appointed official, one who hold public office, or spokesperson for a governmental agency
Author-Journalist	The account represents media personnel
Author-Citizen	The account appears to belong to an individual who otherwise does not hold some position of authority
Author-Bot	The account and its content appears to be populated by some sort of automated feed (e.g., Weather warning bots)
Author-Useful	A meta label indicating the assessor found knowing the author to be informative

Table 3. Additional Information Type labels collected in 2021

includes 312,546 images, over 40 gigabytes worth of data, and averages 2,562 images per event. Slightly more than half of these images come from crisis events released in the 2021 data, however, as these crisis have, on average, much larger samples than earlier TREC-IS releases. This image dataset is publicly available at:

- https://trecis.github.io/datasets.html

RESULTS

Having established the datasets and evaluation methods in TREC-IS 2021, we now turn to substantive questions around the impacts of more detailed assessment and deeper pools. We divide research questions around these goals into two segments, first about characterizing these additional manual assessments (RQ1) and second about the impact of deeper pools (RQ2):

RQ1.1. What are the distributions of different author-, target-audience-, and image-types in crisis data?

RQ1.2. For author-, target-audience-, and image-types in crisis data, which annotations have the highest perceived priority?

RQ2.1. How do the distributions of information type and priority vary between shallow and deep assessment pools?

RQ2.2. How correlated are *performance metrics* and *rankings* between shallow and deep assessment pools?

RQ2.3. How impacted are *per-information-type* and *per-event* performance metrics across depths?

More Detailed Manual Assessments

Figure 2 shows the frequencies of each category type across all the TREC-IS 2021 data. As one may expect, irrelevant content written by a private citizen for a general audience are the most common types of messages sent. Of the images shared, visuals of damage are the most common, closely followed by "informative" images and infographics. Followed by private citizens, content written by news organizations and journalists appear the most frequent. Tweets with location information – as determined by our human assessors – are similarly the second-most common types of information shared during the 2021 TREC-IS crises.

For RQ1.2., Figure 3 shows average priority across the various labels. As in prior editions, several of the six "actionable" information-types exhibit, on average, high priority (MovePeople and SearchAndRescue in particular, with GoodsAndServices nearly high-priority as well). Two of the three remaining actionable information types (EmergingThreats and NewSubEvent) are the next highest priority information types, on average, while the last actionable type – ServiceAvailable – has dropped two spaces, with Official and MultimediaShare types now exceeding ServiceAvailable.

Moving to differences in intended audience (top-right panel of Figure 3), as one might expect, content that is intended for local, directly impacted audiences are higher-priority than content for general audiences. When

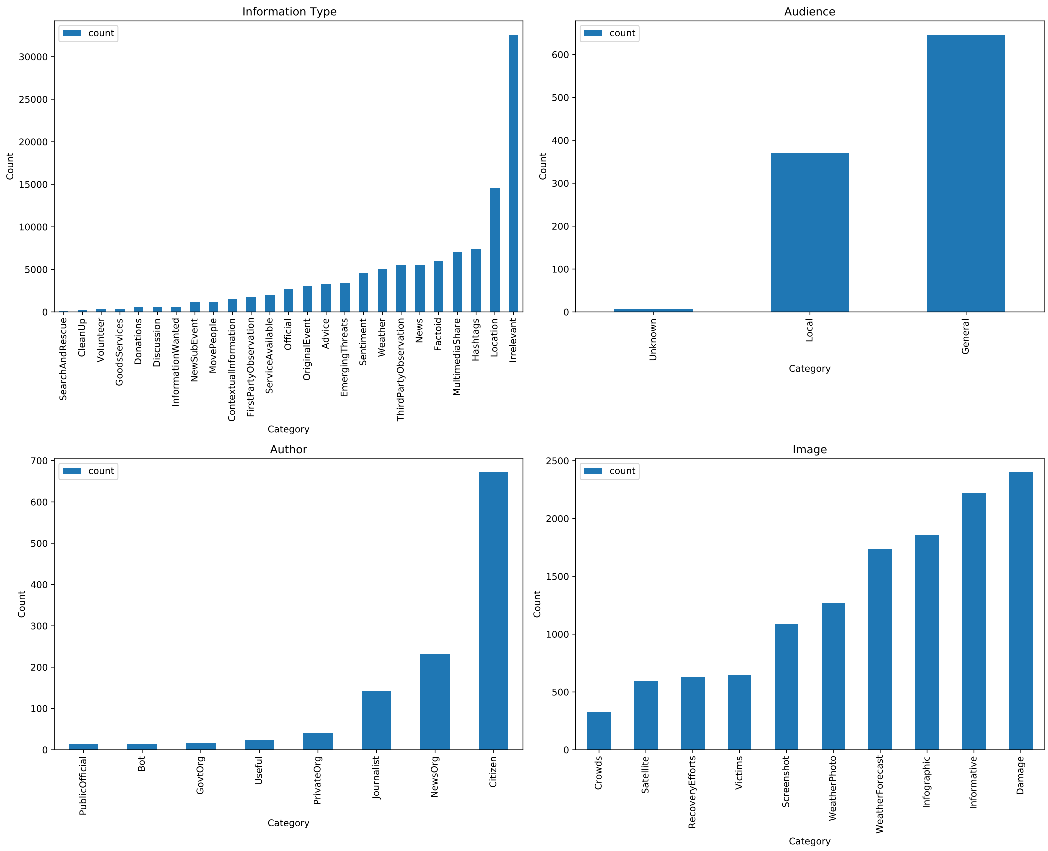


Figure 2. Frequency Across Assessment Categories. Irrelevant content written by a private citizen for a general audience are the most common types of content, as are images of damage. These frequencies represent observed counts in the data, and not an estimate of a wider population.

evaluating aspects of the message author, no single author characteristic appears to exceed an average of high priority or higher, though government organizations do seem to be higher on average than other author characteristics. Interestingly, in the case where the NIST assessor identified that knowledge about the author was *useful* information ("Useful" in the figure), those instances were generally higher priority.

When examining messages with images (bottom right of Figure 3), the four highest-priority image labels exceed the average priority of the top three information types. Specifically, weather-forecast and satellite imagery and infographics having an average priority exceeding 0.7. As with "useful" author information, images found to be informative by assessors were likewise nearly high priority, and as shown in Figure 4, tweets *without* images are below "Medium" priority ($\mu = 0.4899$), whereas tweets *with* are between "Medium" and "High" priority ($\mu = 0.6231$).

Deeper Evaluation Pools and Their Effects

Official TREC-IS performance metrics are shown in the Supplementary Material, specifically Tables 6 - 8. Qualitative assessment of those metrics show that no single system exhibits superior performance across all metrics (though one system does have a lead in the majority of metrics in 2021-B), and evaluations in deeper pools appear to generally lower absolute performance metrics.

Specific to RQ2.1, we examine distributions of information types and priorities across the 13 crisis events for which we have sufficient assessments from both shallow and deep pool depths. These 13 events and the number of assessments in shallow and deep pools are shown in Table 4. Figures 5 and 6 show that differences between these pools exist chiefly in distributions of priority. Regarding information-types, we see location is consistently the most common type across pool depths, and requests for search and rescue and reports of clean-up activities are consistently rare. Calls to action, requests for information, and "actionable" information-types are also consistently rare in both pool depths, suggesting the deeper pools are not surfacing a different distribution of important information. In fact, for information-types, the seven most common and seven more rare types of information are the same across pool depths – though their relative rankings vary slightly.

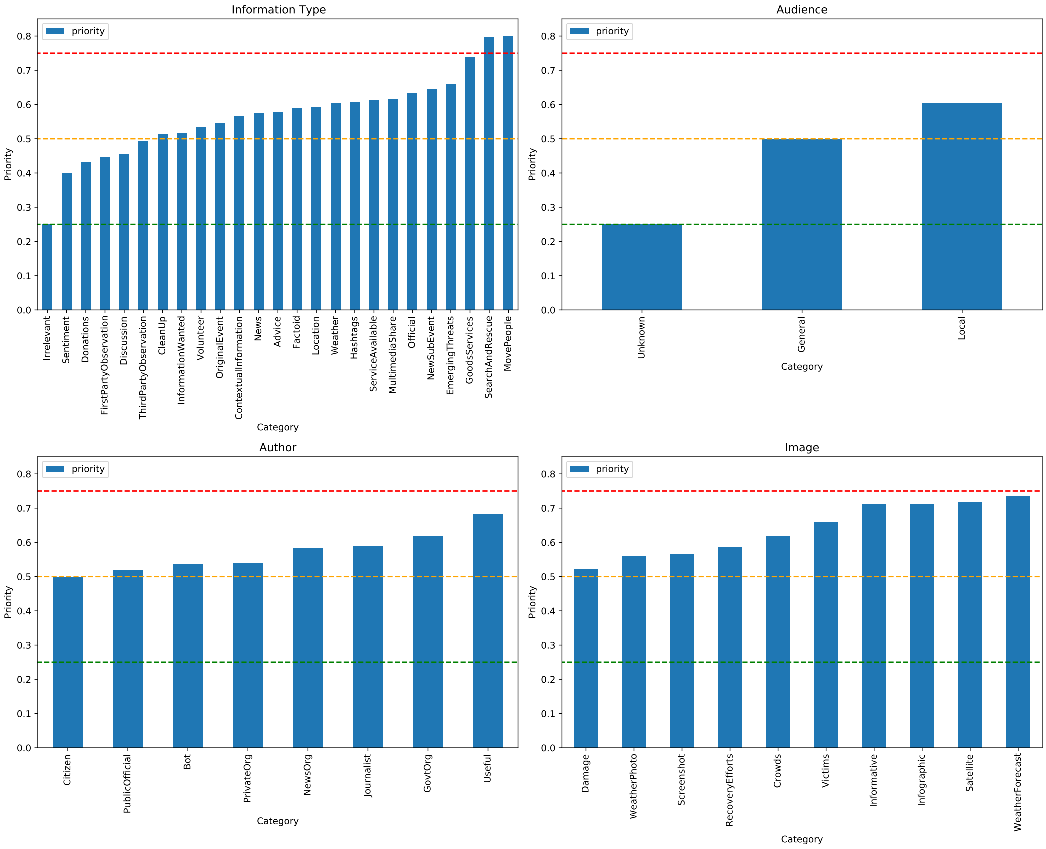


Figure 3. Mean Priority Across Assessment Categories. Only calls for search and rescue and for evacuations consistently receive high-priority assessments.

For priority, surprisingly, we see a *decrease* in high- and critical-priority content in deeper pools compared to the shallow pools in 2021-A. Also unexpectedly, the shallow pools in 2021-A have a vastly different distribution compared to prior TREC-IS tracks, where only about 15% of content was high-priority or higher (or less than half of the prevalence in 2021-A).

Given the variation in priority across pool depths, depth may likewise impact performance metrics, which would have implications for generalization in the crisis informatics space. Alternatively, the deeper pools and additional assessments may uniformly reduce metrics. To examine these possibilities, Figure 7 illustrates performance impact for information-type classification and prioritization for the top-performing system from each participating team's submission prior to July 2021 ($n = 5$) – we use only one run per team to avoid biases from highly similar submissions. Figure 8 likewise shows the scatter plots for all eight metrics across pool depths. Despite the small sample size, we can assess correlations across pool depths using Pearson's correlation and Spearman's rank correlation. In answer to RQ2.2. about whether the performance metrics across pool depths are correlated, correlation in information-type F1 classification performance is significant and very strongly correlated across depths (Pearson's $r > 0.99$). For prioritization metrics, while correlations are still strong ($r > 0.89$), correlation is less consistent than in info-type categorization. Spearman's rank correlations in systems' rankings shows a similar pattern compared to Pearson's correlation among the actual metrics, with the exception of "Priority R".

This variation in efficacy around priority may be driven be the deviation in distributions of priority in the deeper-pool context. We also see consistency in that the move from messages with "Actionable" types to "All" types results in an *increase* in performance – i.e., systems perform better in the "All" context – with the exception of correlations between predicted priority score and actual priority score.

Lastly, we turn to RQ2.3 and variation in per-information-type and per-event performance. Comparing the x- and y-axes from Figures 7 and 8 in RQ2.2 suggests participant runs tend to perform worse in the deeper pools as compared to the shallower pools. To investigate this question, we take the average F1 scores across information-types and events for the most performant systems for each team as above. For information-types, Figure 9a shows that, for many cases, performance in the deeper pools suffers, especially in regards to reports of location, sentiment, etc.; participant runs only outperform the shallow-pool evaluation in seven of the 25 types. Per-event results tell a similar

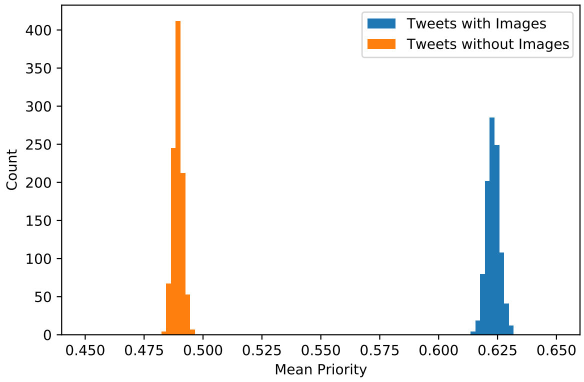


Figure 4. Bootstrapped Sample of Mean Priority for Tweets With and Without Images. Tweets with images are, on average, higher priority.

Identifier	Event Name	Shallow Pool Size	Deep Pool Size
TRECIS-CTIT-H-76	Houston Explosion 2020	1,380	2,795
TRECIS-CTIT-H-81	Mideast Tornadoes Day 2 Alabama 2020	1,600	1,045
TRECIS-CTIT-H-83	Tennessee Derecho 2020	1,838	904
TRECIS-CTIT-H-89	2016 Puttingal Temple	1,490	2,990
TRECIS-CTIT-H-91	Thomas Wildfire 2017	1,521	3,019
TRECIS-CTIT-H-92	Lilac Wildfire 2017	510	2,346
TRECIS-CTIT-H-101	2018 Maryland Flood	1,000	2,699
TRECIS-CTIT-H-107	Kincade Wildfire 2019	2,000	1,643
TRECIS-CTIT-H-108	2019 Durham Gas Explosion	1,500	1,436
TRECIS-CTIT-H-110	2019 Saugus High School Shooting	499	2,174
TRECIS-CTIT-H-116	2020 Easter Tornado Outbreak	499	2,177
TRECIS-CTIT-H-119	2020 Tornado Outbreak of April	1,918	1,503
TRECIS-CTIT-H-121	2020 Visakhapatnam Gas Leak	1,759	1,399

Table 4. Crisis Events Assessed with Shallow and Deep Pools

story: all events exhibit lower F1 scores in deeper pools, some significantly (exceeding a 95% confidence interval), such as the difference in the 2020 tornado outbreak event.

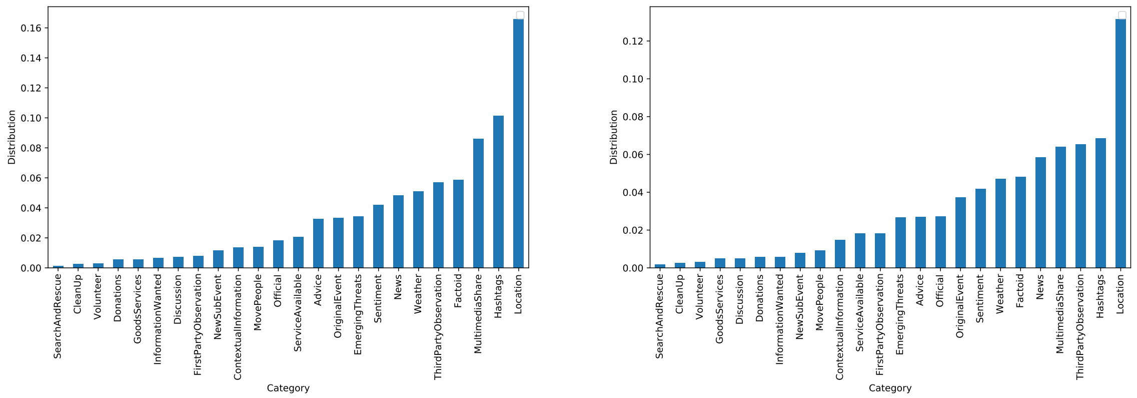
DISCUSSION

Connecting Message Attributes and Importance

A key finding from the above results suggests a connection among content intended for local audiences, the presence of images, and the perceived importance of that content. While interesting in itself and suggestive of future research questions about identifying intended audiences, we also note the limited correlation between the importance and source of content. That is, the identity or social position of the author does not appear to greatly influence the importance of said content. We see large volumes of content from private citizens (who likely outnumber government and journalistic individuals), but these citizens can be from anywhere, and given the occasional global interest in crises (e.g., the Nepalese earthquake, terror attacks in Paris, etc.), the population of private citizens far away from the impacted area is likely relatively much larger. Hence, it is reasonable that private citizens may have high variance in the importance of their content. Other work has suggested methods for identifying *eyewitnesses* of crisis events, which might further separate this broad class of private citizens (Diakopoulos et al. 2012).

Surprisingly, however, accounts with journalistic or governmental roles also appear relatively uncorrelated with importance. Likewise, images depicting damage, while common, are also assessed as having lower priority despite the value associated with this type of imagery in the CrisisMMD datasets (Alam, Ofli, and Imran 2018).

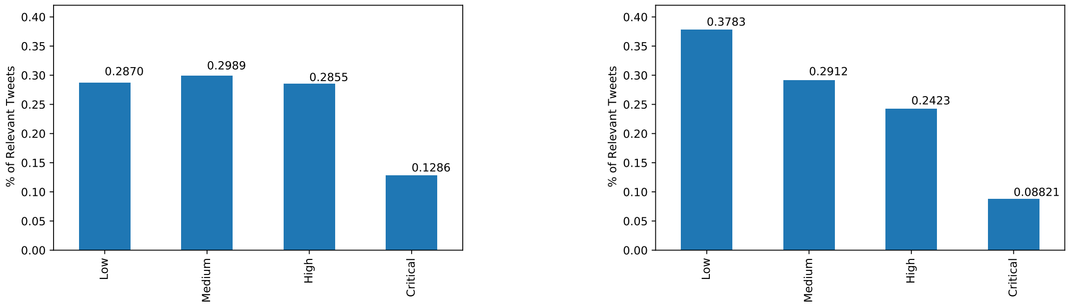
An additional observation we note here is that, across these sets of message attributes, we see that the quantity of a particular class appears inversely related to the perceived criticality of that content. E.g., images of damage are the most common depiction and yet the lowest average priority, and content from private citizens is similarly lower in priority. This observation suggests alternate strategies for identifying high-priority content in social media streams through measuring message similarity. That is, if one can identify outlier messages (i.e., messages that are highly dissimilar) in a particular class, these outliers may be more likely to contain important information about the crisis.



(a) Info-Types in shallow (depth=10) pools, as a fraction of all labeled tweets, 2021-A.

(b) Info-Types in deep (depth=20) pools, as a fraction of all labeled tweets, 2021-B.

Figure 5. Fraction of Tweets labeled with Info-Types Across Pool Depths. Prevalence of various information types appear relatively conserved across pool depths.



(a) Priority distributions in shallow (depth=10) pools, 2021-A.

(b) Priority distributions in deep (depth=20) pools, 2021-B.

Figure 6. Distribution of Priorities Across Pool Depths. In both cases, we see an increase in content labeled as critical compared to prior TREC-IS editions, where only 15% of content was labeled as high or critical. Here, we see approximately between 30-40% of content is high or critical. Across pool depths, however, we see more high- and critical-level content in the shallower pools. These figures are fractions of the entire dataset.

Redundancy and a Motivation for Summarization Stemming from the relative value of rare content, TREC-IS in its current form does not address a core issue of duplication in social media streams. Prior research shows social media contains large volumes of redundant content (C. L. Buntain and Lim 2018; Lin et al. 2016), and providing disaster management personnel with streams of this largely redundant content, even if all that content comes from the same information-type class, is not likely to improve their understanding of the event if they must sift through many repeated messages. Consequently, future efforts in this space should explore methods to collapse these redundant messages and provide more succint summaries of events to disaster management stakeholders. In this alternate scenario, where a system extracts particular "facts" about the crisis event – or automatically extracted points of information about the crisis – and these facts have associated support messages from which the fact was extracted, if the support around a fact is low, this lack of support may indicate an important message. An alternate possibility is that these low-support facts are of dubious veracity, and we know veracity is a crucial concern in the crisis informatics community (and communities around social media more broadly), but as shown in Castillo, Mendoza, et al. 2013, false rumors tend to have a non-trivial amount of contradicting messages sent in reply to try and debunk the rumor.

Implications from the Deep End: TREC-IS and Deeper Pools

While the distributions of information types and priorities appears consistent, evaluation in deeper pools tends to decrease run performance across the submitted systems. As discussed above, these deeper pools are definitionally comprised of messages in which participant systems are less confident of their labels, so it is perhaps expected that system performance decreases. Despite this general trend, we find the overall rankings of systems remains relatively stable in identifying particular types of information shared, and even though prioritization metrics are less stable, correlations in the shallow and deep pools are still quite strong. To our original motivation for deeper

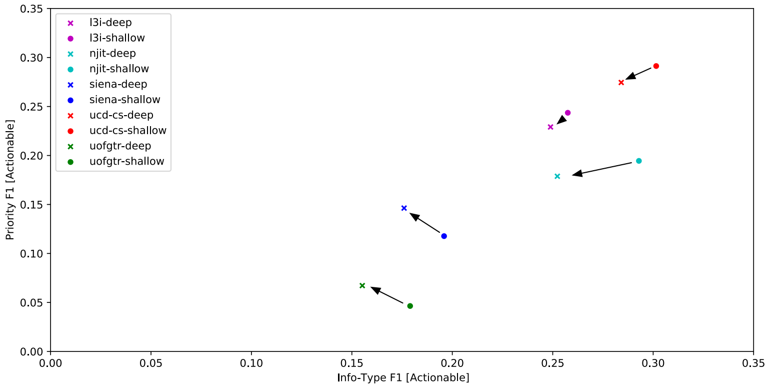


Figure 7. Changes in Information-Type and Prioritization Performance Across Pool Depths. Arrows denote the change in performance moving from shallow to deeper pools. We see all runs experience a decrease in information-type classification performance, and the majority experience a similar drop in prioritization performance as well.

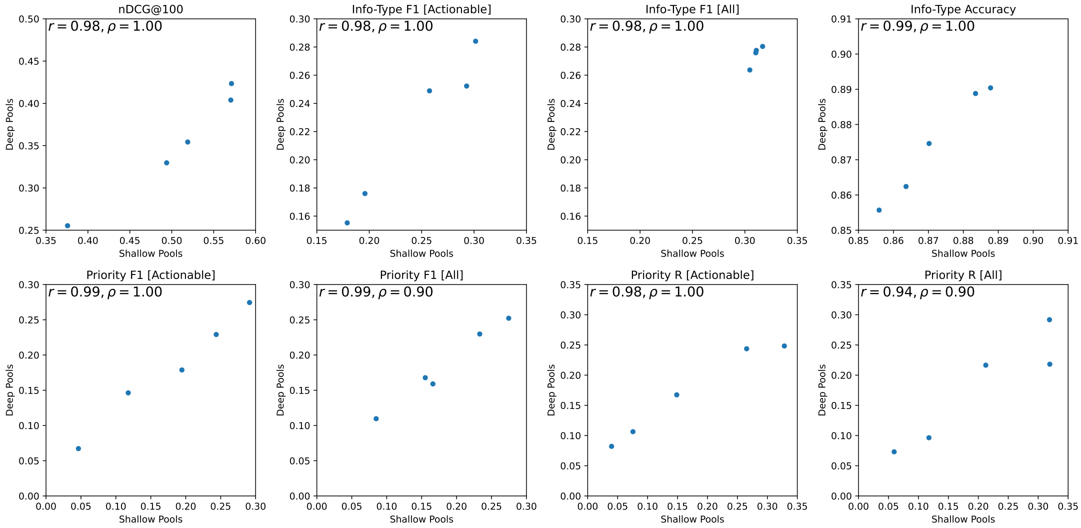
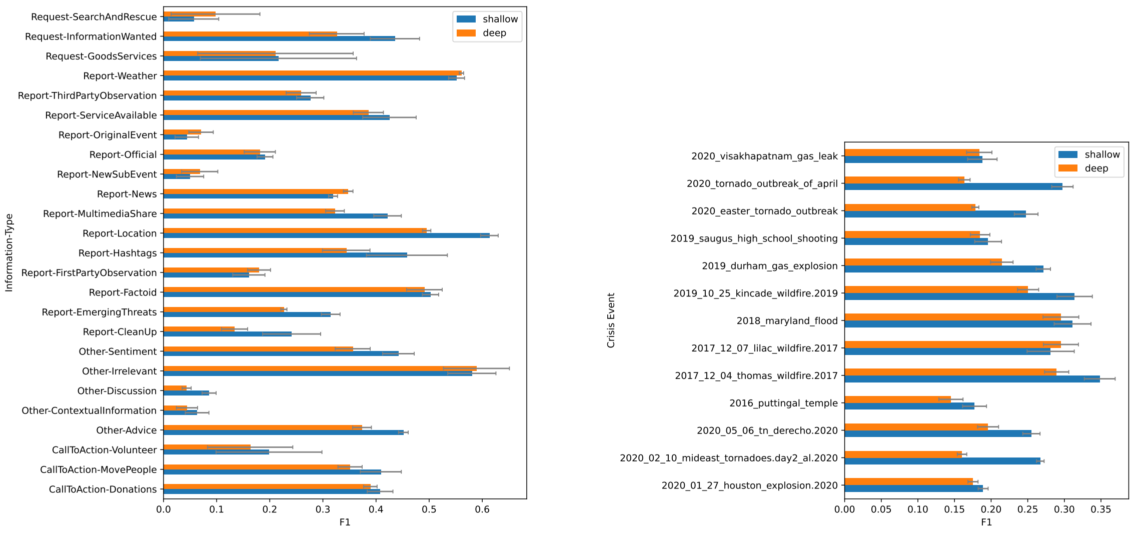


Figure 8. Scatter Plots of Evaluations in Shallow and Deep Pools. F1 metrics of information-type exhibit very strong Pearson's r and Spearman's ρ (shown in plots) across pools depths. Prioritization, however, has less consistency across the editions but still maintains significant correlation.

pools, however, the uniform nature of impact on system performance suggests the driver here is endemic to the task and events and not an issue specific to one or two participant groups. Rather, social media content introduced in the deeper pools may be more unique, event-specific content. As a result, systems that are trained to generalize representations of information-types datasets from many different crisis events of several different types may be fundamentally limited in their ability to correctly label event-specific – or event-type-specific – content. More research is needed here.

Insights into System Performance: The Value of Multi-Task Learning

Participant runs from the UCD and L3i teams clearly establish a frontier in system performance across information-type classification and prioritization. In comparison, runs from the NJIT group produce similar performance metrics in information-type but performance less well in prioritization. In discussions with developers across these teams at the annual TREC meeting, a common theme has emerged: Runs that treat classification and prioritization as completely separate tasks perform less well than architectures that integrate the two tasks. Specifically, high-performing runs from both top-performing teams appear to use some form of multi-task learning, in which parameters in the deep layers are shared across the classification and prioritization tasks, and the final layers are replaced for the output task. In contrast, runs from the NJIT team treat information types and priorities as wholly separate tasks, with no shared structure between them.



(a) F1 Metrics Across Info-Types in Shallow/Deep Pools.

(b) F1 Metrics Across Events in Shallow/Deep Pools.

Figure 9. Average Performance Metrics for Information-Type and Crisis Events in Shallow/Deep Pools. Gray whiskers indicate the 95% standard confidence interval. Results are consistent with an overall reduction in F1 score across nearly all information types and events when assessed against deeper pools. Exceptions for information types include weather and news reports, first-party observations, and reports of the original event.

This multi-task learning approach is likely more in line with the reality of the TREC-IS tasks, as we have previously shown certain information types have higher average priorities. As such, one would expect a shared representation in the layers of the neural networks across these tasks would be beneficial. Coupling this observation with the relationships described above between priority, audience, and imagery suggests that these additional factors could be treated as additional tasks (e.g., audience classification or eyewitness detection), even if the specific task is less useful than the information-type categorization. In addition, a recurring question for TREC-IS has concerned whether the priority task adds much value given the "actionable" information-type category; this observation indicates the priority task may help guide more generalizable representational structure in the learning framework.

THREATS TO VALIDITY

The above analysis raises concerns about the difference seen in the distribution of message priorities in 2021-A's shallow pools compared to prior years' assessments. Specifically, assessments from TREC-IS's 2021 edition (and especially 2021-A) show a much larger proportion of content that has a "critical" priority label and a different shape of low-, medium-, and high-priority content. In previous editions, results show the majority of content is of low priority, with decaying proportions down to critical messages, which have historically occurred in only about 1% of messages. In contrast, 2021-A shows a consistent proportion across low-, medium-, and high-priority content and over 10% receiving a critical rating.

One potential explanation for this divergence from prior years is TREC-IS's change to its data collection methods: As the 2021 edition uses Twitter's enterprise offerings to collect crisis data from retrospective events, the density of high-priority content that remains on the platform long after a crisis may be higher. Alternatively, a change in the assessor population, training, or tasking may have impacted assessment, such that assessors' thresholds for critical content may have fallen. While we have updated training material for 2021, these updates have mainly focused on the additional categories of assessment information (e.g., audience, author, etc.), but these additional categories – which were only present in 2021-A – may have altered assessor perceptions.

In comparing the 2021-A and 2021-B distributions, we see the 2021-B priority distribution better mirrors the step-wise structure observed in prior years, but we still see a larger proportion of critical content than expected. Given overlap in the set of assessors in both the 2021-A and 2021-B editions (many of whom had participated in previous TREC-IS assessments), and 2021-B does not include the more complex author and audience assessments, this divergence may be more closely related to changes in the underlying data, suggesting this data remains valid.

CONCLUSIONS

This paper outlines the core dataset and research contributions for the final year of the TREC-IS initiative. These contributions include datasets from 47 crises – 26 of which contain manual assessments across several categories of information – and several hundred-thousand crisis-relevant images, 7,297 of which also include labels from manual assessment. Taken together, these resources should extend the foundation for research into the multi-modal and context-dependent aspects of crisis informatics, moving us beyond textual dependencies.

Beyond new and larger datasets, this work also contributes new findings to the research community, specifically around the impacts and values of additional detail and deeper pooling methods for sampling and evaluation. In particular, our results suggest more nuanced assessments of a message's *intended audience* provide useful information about the priority of that content, and content with images is consistently perceived as higher priority than messages without images (apparently regardless of information type). As a main takeaway, future efforts could include this audience characterization as an additional learning task, and given participant results showing the value of multi-task learning, this additional task may open new avenues for improving information systems for disaster response personnel and other stakeholders.

Lastly, our evaluation of deeper pools provides additional insights, as they tend to preserve the distributions of actionable information, especially calls to action and information requests, and overall performance rankings of participant runs. These points have particularly valuable implications for the reusability of the 136,263-tweet labeled dataset TREC-IS now provides, as ranking stability suggests pools are sufficiently complete to support system evaluations, though more research is needed to evaluate reusability questions more thoroughly.

Despite several years of running the TREC-IS initiative and the datasets produced, many open questions remain for the space. Such questions include fusing data from multiple streams, dealing with redundancy in crisis-related social media, understanding the veracity of information contained in these streams, and others. In total, we hope coupling the TREC-IS dataset and our findings with other crisis-informatics datasets and resources (e.g., CrisisMMD, CrisisLex, etc.) provides a solid foundation for future crisis informatics work that answers these questions.

ACKNOWLEDGMENTS AND DISCLAIMER

This work was supported by the Social Media Incident Streams Project sponsored by the Public Safety Communication Research Division at NIST. Any mention of products, companies or services in this paper is for explanatory and scientific purposes and does not imply an endorsement or recommendation of those companies, products, or services.

REFERENCES

Alam, F., Ofli, F., and Imran, M. (June 2018). “CrisisMMD: Multimodal Twitter Datasets from Natural Disasters”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 12.1.

Alam, F., Ofli, F., Imran, M., Alam, T., and Qazi, U. (2020). “Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response”. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

Alam, F., Sajjad, H., Imran, M., and Ofli, F. (2021). “CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing”. In: *15th International Conference on Web and Social Media (ICWSM)*.

Buntain, C., Buntain, C., Mccreadie, R., and Soboroff, I. (2021). “Incident Streams 2020 : TRECIS in the Time of COVID-19”. In: *18th International Conference on Information Systems for Crisis Response and Management* April.

Buntain, C. L. and Lim, J. K. ((2018). “#pray4victims : Consistencies In Response To Disaster on Social Media”. In: *Proceedings of the 21st ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. Jersey City, NY, USA: ACM.

Castillo, C., Mendoza, M., and Poblete, B. (2013). “Predicting information credibility in time-sensitive social media”. In: *Internet Research* 23.5.

Castillo, C., Peterson, S., Rufolo, P., Fincons, S. A., Pajarito, D., and Buntain, C. (2021). “Social Media for Emergency Management : Opportunities and Challenges at the Intersection of Research and Practice”. In: *18th International Conference on Information Systems for Crisis Response and Management*. May, pp. 772–777.

Damiani, E., Vimercati, S. D. C. di, Paraboschi, S., and Samarati, P. (2004). “An Open Digest-based Technique for Spam Detection”. In: *Proceedings of the 2004 International Workshop on Security in Parallel and Distributed Systems* 1.1, pp. 559–564.

Diakopoulos, N., De Choudhury, M., and Naaman, M. (2012). “Finding and assessing social media information sources in the context of journalism”. In: *Proceedings of SIGCHI*. New York, NY, USA: ACM.

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G. J. F., and Moens, M.-F. (2019). “Report on the Second Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2018) at the Web Conference (WWW) 2018”. In: *ACM SIGIR Forum* 52.2, pp. 163–168.

Imran, M., Mitra, P., and Castillo, C. (2016). “Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages”. In: *Proceedings of LREC*.

Lin, J., Roegiest, A., Tan, L., McCreadie, R., Voorhees, E., and Diaz, F. (2016). “Overview of the TREC 2016 real-time summarization track”. In: *Proceedings of the 25th text retrieval conference, TREC*. Vol. 16.

McCreadie, R., Buntain, C., and Soboroff, I. (2019). “TREC Incident Streams: Finding Actionable Information on Social Media”. In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

McCreadie, R., Buntain, C., and Soboroff, I. (2020). “Incident Streams 2019 : Actionable Insights and How to Find Them”. In: *Proceedings of the 17th International Conference on Information Systems for Crisis Response And Management*. May.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). “CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises.” In: *Proceedings of ISCRAM*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). “What to expect when the unexpected happens: Social media communications across crises”. In: *Proceedings of CSCW*. ACM.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., and Shrimpton, L. (2013). “Can Twitter replace Newswire for breaking news?” In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. Vol. 2011.

Purohit, H., Castillo, C., Imran, M., and Pandev, R. (2018). “Social-EOC: Serviceability model to rank social media requests for emergency operation centers”. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pp. 119–126.

Voorhees, E. M. (1998). “Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness”. In: *SIGIR 1998*.

Zobel, J. (1998). “How Reliable Are the Results of Large-Scale Information Retrieval Experiments?” In: *Proceedings of SIGIR*. ACM, pp. 307–314.

APPENDIX A: TREC-IS LABELLED EVENT LIST

Edition	Identifier	Event Name	Event Type	Tweet Sample	# Unique Labeled	# Multi-Labeled
2018	TRECIS-CTIT-H-001	2012 Colorado wildfires⋆	wildfire	744	264	2
	TRECIS-CTIT-H-002	2012 Costa Rica Earthquake⋆	earthquake	288	247	-
	TRECIS-CTIT-H-003	2013 Colorado Floods⋆	flood	777	235	-
	TRECIS-CTIT-H-004	2012 Typhoon Pablo⋆	typhoon/hurricane	649	244	-
	TRECIS-CTIT-H-005	2013 LA Airport Shooting⋆	shooting	683	162	-
	TRECIS-CTIT-H-006	2013 West Texas Explosion⋆	bombing	630	184	-
2019-A	TRECIS-CTIT-H-007	2012 Guatemala earthquake⋆	earthquake	178	154	-
	TRECIS-CTIT-H-008	2012 Italy earthquakes⋆	earthquake	118	103	-
	TRECIS-CTIT-H-009	2012 Philippines floods⋆	flood	480	437	-
	TRECIS-CTIT-H-010	2013 Alberta floods⋆	flood	739	722	-
	TRECIS-CTIT-H-011	2013 Australia bushfire⋆	wildfire	710	677	-
	TRECIS-CTIT-H-012	2013 Boston bombings⋆	bombing	543	535	-
	TRECIS-CTIT-H-013	2013 Manila floods⋆	flood	443	411	-
	TRECIS-CTIT-H-014	2013 Queensland floods⋆	flood	744	713	-
	TRECIS-CTIT-H-015	2013 Typhoon Yolanda⋆	typhoon	629	564	-
	TRECIS-CTIT-H-016	2011 Joplin tornado⋆⋆⋆	typhoon	152	96	-
	TRECIS-CTIT-H-017	2014 Chile Earthquake⋆⋆	earthquake	321	311	-
	TRECIS-CTIT-H-018	2014 Typhoon Hagupit⋆⋆	typhoon	6,696	4,193	504
	TRECIS-CTIT-H-019	2015 Nepal Earthquake⋆⋆	earthquake	7,301	7,684	3,359
	TRECIS-CTIT-H-020	2018 FL School Shooting	shooting	1,118	1,118	-
	TRECIS-CTIT-H-021	2015 Paris attacks	bombing	2,066	2066	-
2019-B	TRECIS-CTIT-H-022	2019 Choco Flood	flood	854	389	-
	TRECIS-CTIT-H-024	2014 California Earthquake	earthquake	128	127	-
	TRECIS-CTIT-H-025	2013 Bohol Earthquake	earthquake	646	619	72
	TRECIS-CTIT-H-026	2018 Florence Hurricane†	typhoon	2,500	2,499	1,000
	TRECIS-CTIT-H-027	2017 Dallas Shooting	shooting	2,500	2,500	-
	TRECIS-CTIT-H-028	2016 Fort McMurray Wildfire	wildfire	2,500	2,500	-
	TRECIS-CTIT-H-029	2019 Alberta Wildfires	wildfire	2,500	3,740	2,776
	TRECIS-CTIT-H-030	2019 Cyclone Kenneth	typhoon	2,500	4,527	4,177
	TRECIS-CTIT-H-031	2019 Luzon earthquake	earthquake	2,500	2,939	1,848
	TRECIS-CTIT-H-032	2019 STEM School Highlands Ranch shooting	shooting	2,500	1,239	165
	TRECIS-CTIT-H-033	2019 Durban Easter floods	flood	2,500	2,198	1,698
	TRECIS-CTIT-H-034	2019 Poway synagogue shooting	shooting	2,500	679	83
2020-A	TRECIS-CTIT-H-035	2019 Greece Earthquake	earthquake	500	946	946
	TRECIS-CTIT-H-036	2020 Baltimore Floods	flood	500	1,407	1,407
	TRECIS-CTIT-H-037	2020 Brooklyn block party shooting	shooting	500	958	958
	TRECIS-CTIT-H-038	2020 Dayton Ohio shooting	flood	500	1425	1425
	TRECIS-CTIT-H-039	2020 El Paso Walmart shooting	shooting	500	898	894
	TRECIS-CTIT-H-040	2020 Gilroy Garlic Festival shooting	shooting	500	940	940
	TRECIS-CTIT-H-041	2020 Hurricane Barry	typhoon	500	938	938
	TRECIS-CTIT-H-042	2020 Indonesia Earthquake	earthquake	500	940	940
	TRECIS-CTIT-H-043	2020 Kerala, India floods	flood	500	782	782
	TRECIS-CTIT-H-044	2020 Myanmar floods	flood	500	940	940
	TRECIS-CTIT-H-045	2020 Papua New Guinea earthquake	earthquake	500	447	-
	TRECIS-CTIT-H-046	2020 Siberian Wildfires	wildfire	500	948	948
	TRECIS-CTIT-H-047	2020 Typhoon Krosa	typhoon	500	932	932
	TRECIS-CTIT-H-048	2020 Typhoon Lekima	typhoon	500	938	938
	TRECIS-CTIT-H-049	2020 Whaley Bridge Dam Collapse	accident	500	420	420
	TRECIS-CTIT-H-050	2020 COVID-19 Outbreak in Washington DC	COVID-19	49,893	4,012	3,012
	TRECIS-CTIT-H-051	2020 COVID-19 Outbreak in Washington State	COVID-19	48,498	5,697	5,646
2020-B	TRECIS-CTIT-H-053	2020 Houston explosion	explosion	7876	1,416	-
	TRECIS-CTIT-H-054	2020 Texas A&M-Commerce shooting	shooting	3,489	2,638	2,638
	TRECIS-CTIT-H-055	2020 Southeast US Tornado Outbreak	storm	4,135	1,348	-
	TRECIS-CTIT-H-056	2020 Storm Ciara	storm	2,878	1,075	-
	TRECIS-CTIT-H-057	2020 Storm Dennis	storm	49,999	1,851	-
	TRECIS-CTIT-H-059	2020 Virra Mall Hostage Situation	hostage	2,293	1,149	-
	TRECIS-CTIT-H-060	2020 Storm Jorge	storm	25,776	602	-
	TRECIS-CTIT-H-061	2020 Tennessee tornado outbreak	storm	49,999	1,278	-
	TRECIS-CTIT-H-062	2020 Tennessee derecho	storm	3,852	500	-
	TRECIS-CTIT-H-063	2020 Edenville dam failure	accident	35,464	1,586	-
	TRECIS-CTIT-H-064	2020 San Francisco Pier Fire	structural fire	2,751	1,298	-
	TRECIS-CTIT-H-065	2020 Tropical Storm Cristobal	storm	36,274	1,000	-
	TRECIS-CTIT-H-066	2020 Beirut Explosion	explosion	49,999	500	-
	TRECIS-CTIT-H-068	2020 COVID-19 Outbreak in Jacksonville, FL	COVID-19	13,505	664	-
	TRECIS-CTIT-H-069	2020 COVID-19 Outbreak in Houston, TX	COVID-19	44,296	963	-
	TRECIS-CTIT-H-070	2020 COVID-19 Outbreak in Phoenix, AZ	COVID-19	16,765	871	-
	TRECIS-CTIT-H-071	2020 COVID-19 Outbreak in Atlanta, GA	COVID-19	49,999	966	-
	TRECIS-CTIT-H-072	2020 COVID-19 Outbreak in New York	COVID-19	49,999	6,522	5,172
	TRECIS-CTIT-H-073	2020 COVID-19 Outbreak in Seattle, WA	COVID-19	49,999	966	-
	TRECIS-CTIT-H-074	2020 COVID-19 Outbreak in Melbourne, AU	COVID-19	49,999	957	-
	TRECIS-CTIT-H-075	2020 COVID-19 Outbreak in New Zealand	COVID-19	5,147	783	-
2021-A/B	TRECIS-CTIT-H-076	Houston Explosion 2020	explosion	32,204	2,795	-
	TRECIS-CTIT-H-080	Mideast Tornadoes Day 1 Mississippi 2020	tornado	834	772	-
	TRECIS-CTIT-H-081	Mideast Tornadoes Day 2 Alabama 2020	tornado	9,594	3,829	2,185
	TRECIS-CTIT-H-082	Mideast Tornadoes Day 3 MD 2020	tornado	13,641	1,109	-
	TRECIS-CTIT-H-083	Tennessee Derecho 2020	storm	26,946	2,745	6
	TRECIS-CTIT-H-085	Brooklyn Block Party Shooting 2019	shooting	49,593	3,592	-
	TRECIS-CTIT-H-089	2016 Puttingal Temple	explosion	18,888	3,182	384
	TRECIS-CTIT-H-091	Thomas Wildfire 2017	wildfire	50,000	3,000	-
	TRECIS-CTIT-H-092	Lilac Wildfire 2017	wildfire	50,000	2,342	-
	TRECIS-CTIT-H-096	Klamathon Wildfire 2018	wildfire	2,094	1,583	-
	TRECIS-CTIT-H-097	Holy Wildfire 2018	wildfire	49,664	3,336	-
	TRECIS-CTIT-H-100	Woolsey Wildfire 2018	wildfire	49,167	1,588	-
	TRECIS-CTIT-H-101	2018 Maryland Flood	flood	50,000	2,975	552
	TRECIS-CTIT-H-102	2018 Pittsburgh Synagogue Shooting	shooting	49,341	1,755	-
	TRECIS-CTIT-H-103	Alberta Wildfire 2019	wildfire	5,612	2,322	-
	TRECIS-CTIT-H-104	Hurricane Dorian 2019	hurricane	49,100	2,219	-
	TRECIS-CTIT-H-106	Saddleridge Wildfire 2019	wildfire	49,623	2,168	-
	TRECIS-CTIT-H-107	Kincade Wildfire 2019	wildfire	49,495	4,177	1,034
	TRECIS-CTIT-H-108	2019 Durham Gas Explosion	explosion	26,891	3,476	1,061
	TRECIS-CTIT-H-110	2019 Saugus High School Shooting	shooting	43,363	2,174	-
	TRECIS-CTIT-H-112	2019 Townsville Flood	flood	16,542	2,925	-
	TRECIS-CTIT-H-116	2020 Easter Tornado Outbreak	tornado	49,403	2,183	12
	TRECIS-CTIT-H-119	2020 Tornado Outbreak of April	tornado	41,184	3,421	-
	TRECIS-CTIT-H-120	2020 Tornado Outbreak of March	tornado	48,511	1,717	-
	TRECIS-CTIT-H-121	2020 Visakhapatnam Gas Leak	explosion	48,766	3,657	998
	TRECIS-CTIT-H-122	Tornado Outbreak of December 2018	tornado	22,832	1,395	-

⋆ – Collected from CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
⋆⋆ – CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)
⋆⋆⋆ – CrisisNLP Resource #2 (Imran, Elbassuoni, et al. 2013)
† – Donated by participant group

Table 5. TREC-IS Events

APPENDIX B: PARTICIPANT RESULTS

Below, we report the raw results of TREC-IS system evaluations in 2021-A and 2021-B, and unlike previous iterations, a cross-edition evaluation – possible since the submission data is the same across editions. Numeric scores are not directly comparable between 2021-A and 2021-B because of differences in evaluation data); rather, rankings and trends within the metrics and orderings among participant systems are better axes of comparison.

Run Tag	Date	Team	nDCG @100	Info-Type F1 [Actionable]	Info-Type F1 [All]	Info-Type Accuracy	Priority F1 [Actionable]	Priority F1 [All]	Priority R [Actionable]	Priority R [All]
ucdcs-run3	6/1/21	ucd-cs	0.5441	**0.3287**	**0.3464**	0.8858	0.2242	**0.297**	0.1977	0.2648
njit_roberta	6/1/21	njit	0.4933	0.3138	0.3341	0.8644	0.2203	0.2035	0.2204	0.264
ucdcs-run1	6/1/21	ucd-cs	0.5564	0.3057	0.3369	0.8841	**0.2612**	0.2966	0.2107	0.2967
RB_2Tx2_TTH_280	6/1/21	l3i	0.541	0.2996	0.328	0.887	0.2437	0.2592	**0.2603**	0.2425
RB_2T_MT_H_280	6/1/21	l3i	0.4925	0.2857	0.3065	0.8917	0.2377	0.2742	0.2139	0.2313
ucdcs-run2	6/1/21	ucd-cs	**0.5579**	0.2845	0.3393	0.8828	0.241	0.2845	0.1646	0.2906
RB_2T_TT_280_SVM	6/1/21	l3i	0.3019	0.2679	0.324	0.8905	0.0655	0.1276	0.0229	-0.0024
njit_bert	6/1/21	njit	0.4745	0.2346	0.264	0.8181	0.1791	0.1641	0.0797	0.2353
uogTr-02-pwcoocc	6/1/21	uofgtr	0.3551	0.1884	0.314	0.8847	0.0575	0.1139	0.0514	0.0508
uogTr-04-coocc	6/1/21	uofgtr	0.3552	0.1884	0.309	0.8857	0.0575	0.1139	0.0514	0.0508
sienaRun	6/1/21	siena	0.4285	0.1861	0.3108	0.8565	0.0864	0.1609	0.0231	0.1099
RB_2T_TTH_256_LR5	6/1/21	l3i	0.3678	0.1758	0.2923	**0.8931**	0.2042	0.2401	0.1001	0.0908
uogTr-01-pw	6/1/21	uofgtr	0.3561	0.092	0.2203	0.8841	0.0575	0.1139	0.0514	0.0508
random_run	6/1/21	baseline	0.3287	0.0577	0.1408	0.5001	0.2136	0.2326	0.0153	0.0178
ucdcs-run4	6/1/21	ucd-cs	0.5525	0.0224	0.0615	0.7682	0.2517	0.2825	0.2159	**0.2982**
0rule_run	6/1/21	baseline	0.3105	0	0.0056	0.8673	0.0465	0.136	0	0

Table 6. 2021-A evaluations of submitted runs under the v3.1 evaluation script. For each metric, higher is better. Bolded numbers are maximal over the full column. No single system exhibits superior performance across all metrics, though the `ucdcs-run3` system has the highest performance in the most metrics.

Run Tag	Date	Team	nDCG	Info-Type	Info-Type	Info-Type	Priority	Priority	Priority	Priority
			@100	F1 [Ac-	F1 [All]	Accuracy	F1 [Ac-	F1 [All]	R [Ac-	R [All]
				tionable]			tionable]		tionable]	
ens	10/18/21	ucd-cs	0.4555	**0.251**	**0.2623**	0.8753	0.2783	0.2703	0.2116	0.2604
ens	6/29/21	ucd-cs	0.4521	0.2361	0.2591	0.8708	0.2753	0.2582	**0.2302**	**0.2952**
ucd-cs-run3	6/1/21	ucd-cs	0.4583	0.2218	0.2539	0.8964	0.2543	**0.2756**	0.221	0.2571
njit_eda	7/23/21	njit	0.41	0.2183	0.2419	0.8942	0.1541	0.1528	0.096	0.2007
ucd-cs-run1	6/1/21	ucd-cs	0.4499	0.2177	0.247	0.8966	0.2376	0.2566	0.1547	0.2525
njit_roberta	6/23/21	njit	0.4282	0.2124	0.2416	0.879	0.1646	0.1584	0.1667	0.2214
njit_roberta	6/1/21	njit	0.4273	0.2124	0.2416	0.879	0.1646	0.1584	0.1667	0.2214
ucd-cs-run2	6/1/21	ucd-cs	0.4361	0.2087	0.2433	0.8947	0.242	0.2528	0.207	0.2622
njit_semi_sup	6/27/21	njit	0.4272	0.2071	0.2505	0.9054	0.1646	0.1584	0.1667	0.1932
njit_augly	7/19/21	njit	0.4146	0.2045	0.2417	0.886	0.1715	0.1602	0.1293	0.206
RB_2Tx2_TTH_280	6/1/21	l3i	0.4783	0.1959	0.236	0.8981	0.2237	0.2492	0.2279	0.1993
deberta-mtl-fta	10/14/21	ucd-cs	0.4464	0.1958	0.2369	**0.9067**	0.2309	0.2333	0.2044	0.2454
RB_2T_MT_H_280	6/1/21	l3i	0.4443	0.1885	0.2169	0.9043	0.2049	0.2307	0.2109	0.2322
STrans-GaussianNB	10/10/21	ucd-cs	0.338	0.1861	0.2395	0.8557	0.1733	0.1688	0.0425	0.1003
njit_debly	10/15/21	njit	0.4008	0.1853	0.2275	0.9057	0.2244	0.1722	0.1192	0.1794
njit_label_prop	7/1/21	njit	0.4279	0.1842	0.233	0.9014	0.1646	0.1584	0.1667	0.2214
njit_semi_sup	7/6/21	njit	0.413	0.1842	0.233	0.9014	0.246	0.2345	0.1196	0.1336
njit_label_prop	7/8/21	njit	0.4078	0.1842	0.233	0.9014	0.2271	0.2281	0.1374	0.1374
db-mtl-fta-nla	10/18/21	ucd-cs	0.4069	0.1687	0.2187	0.8945	0.2588	0.2635	0.1861	0.2122
sienaRun	6/1/21	siena	0.3917	0.164	0.2413	0.8628	0.1503	0.1912	0.1148	0.1138
l3i_ttxth_combined	9/28/21	l3i	0.4783	0.1568	0.2393	0.9058	0.2107	0.2447	0.1495	0.1987
RB_2T_TT_280_SVM	6/1/21	l3i	0.2407	0.1568	0.2393	0.9058	0.0889	0.1351	0.0222	-0.0028
l3i_ttxth	9/28/21	l3i	**0.4791**	0.1458	0.2062	0.9046	0.2107	0.2447	0.1495	0.1987
RB_2T_TTH_256_LR5	6/1/21	l3i	0.288	0.1458	0.2062	0.9046	0.2374	0.2325	0.0672	0.0701
njit_deberta	8/2/21	njit	0.4452	0.1426	0.2059	0.7883	0.1471	0.1477	0.1886	0.2742
njit_bert	6/1/21	njit	0.3939	0.1349	0.1809	0.825	0.142	0.1388	0.073	0.1857
uogTr-04-coocc	6/1/21	uofgtr	0.2934	0.134	0.2284	0.8977	0.0727	0.1223	0.1375	0.1096
uogTr-02-pwcoocc	6/1/21	uofgtr	0.2945	0.134	0.228	0.8982	0.0727	0.1223	0.1375	0.1096
ens-fta	10/18/21	ucd-cs	0.4448	0.0946	0.1889	0.8113	**0.2798**	0.2724	0.2149	0.2629
uogTr-01-pw	6/1/21	uofgtr	0.2928	0.0731	0.1532	0.8916	0.0727	0.1223	0.1375	0.1096
random_run	6/1/21	baseline	0.2545	0.0362	0.0976	0.5003	0.2224	0.2031	-0.0115	-0.0002
ucd-cs-run4	6/1/21	ucd-cs	0.452	0.0195	0.046	0.8044	0.2753	0.2582	0.2302	**0.2952**
0rule_run	6/1/21	baseline	0.2516	0	0.0036	0.894	0.0591	0.1222	-0.0054	0.0023

Table 7. 2021-B evaluations of submitted runs under the v3.1 evaluation script. For each metric, higher is better. Bolded numbers are maximal over the full column. No single system exhibits superior performance across all metrics.

Run Tag	Date	Team	nDCG	Info-Type	Info-Type	Info-Type	Priority	Priority	Priority	Priority
			@100	F1 [Ac-	F1 [All]	Accuracy	F1 [Ac-	F1 [All]	R [Ac-	R [All]
				tionable]			tionable]		tionable]	
ens	10/18/21	ucd-cs	0.4643	**0.2784**	**0.2946**	0.8728	**0.2864**	**0.2734**	0.2748	0.2827
ens	6/29/21	ucd-cs	0.4617	0.267	0.2923	0.8685	0.2817	0.2623	0.2886	**0.3182**
ucd-cs-run3	6/1/21	ucd-cs	0.4707	0.2538	0.286	0.893	0.253	0.2694	0.2741	0.3053
njit_eda	7/23/21	njit	0.4292	0.2531	0.2735	0.8892	0.1647	0.1531	0.1553	0.2258
njit_roberta	6/23/21	njit	0.4394	0.2501	0.2753	0.8746	0.1719	0.1557	0.2087	0.2431
njit_roberta	6/1/21	njit	0.4385	0.2501	0.2753	0.8746	0.1719	0.1557	0.2087	0.2431
njit_augly	7/19/21	njit	0.4242	0.2441	0.2729	0.8811	0.1723	0.1565	0.17	0.2172
ucd-cs-run1	6/1/21	ucd-cs	0.4727	0.2433	0.2772	0.8926	0.2657	0.2632	0.259	0.2888
ucd-cs-run2	6/1/21	ucd-cs	0.4569	0.2326	0.2753	0.8911	0.2536	0.2524	0.1995	0.2686
RB_2Tx2_TTH_280	6/1/21	l3i	**0.4904**	0.2289	0.2686	0.8949	0.2247	0.2352	0.302	0.2773
njit_semi_sup	6/27/21	njit	0.4385	0.2275	0.278	0.9003	0.1719	0.1557	0.2087	0.2431
RB_2T_MT_H_280	6/1/21	l3i	0.4549	0.2271	0.2498	0.8999	0.2059	0.2177	0.2258	0.2515
deberta-mtl-fta	10/14/21	ucd-cs	0.463	0.2159	0.2647	**0.9016**	0.2497	0.2361	0.2779	0.2834
STrans-GaussianNB	10/10/21	ucd-cs	0.3368	0.2083	0.2575	0.8474	0.1959	0.1712	0.1096	0.1417
njit_debly	10/15/21	njit	0.407	0.2046	0.2502	0.8994	0.2323	0.1777	0.1608	0.2007
njit_label_prop	7/1/21	njit	0.4381	0.2008	0.26	0.8964	0.1719	0.1557	0.2087	0.2431
njit_semi_sup	7/6/21	njit	0.4219	0.2008	0.26	0.8964	0.2271	0.2323	0.0849	0.1346
njit_label_prop	7/8/21	njit	0.4215	0.2008	0.26	0.8964	0.2268	0.2288	0.1058	0.1225
db-mtl-fta-nla	10/18/21	ucd-cs	0.4193	0.1936	0.2488	0.8907	0.2727	0.2609	0.265	0.2339
l3i_ttxth_combined	9/28/21	l3i	0.4888	0.1928	0.2675	0.9004	0.2086	0.2293	**0.3072**	0.3053
RB_2T_TT_280_SVM	6/1/21	l3i	0.2333	0.1928	0.2675	0.9004	0.0695	0.1151	0.0009	-0.0085
sienaRun	6/1/21	siena	0.4061	0.1788	0.2691	0.8607	0.1328	0.1722	0.0796	0.1128
njit_deberta	8/2/21	njit	0.4541	0.171	0.2312	0.782	0.161	0.1466	0.2587	0.3072
njit_bert	6/1/21	njit	0.4102	0.1662	0.2111	0.8223	0.148	0.1366	0.1057	0.2126
l3i_ttxth	9/28/21	l3i	0.4877	0.1645	0.2368	0.9003	0.2086	0.2293	0.3072	0.3053
RB_2T_TTH_256_LR5	6/1/21	l3i	0.2892	0.1645	0.2368	0.9003	0.2364	0.2294	0.1487	0.1034
uogTr-04-coocc	6/1/21	uofgtr	0.3095	0.1556	0.2587	0.8937	0.0588	0.1035	0.1131	0.0908
uogTr-02-pwcoocc	6/1/21	uofgtr	0.3098	0.1556	0.2606	0.8936	0.0588	0.1035	0.1131	0.0908
ens-fta	10/18/21	ucd-cs	0.4515	0.1131	0.217	0.8073	0.2852	0.2724	0.2479	0.2665
uogTr-01-pw	6/1/21	uofgtr	0.3094	0.0869	0.1798	0.8887	0.0588	0.1035	0.1131	0.0908
random_run	6/1/21	baseline	0.2521	0.0452	0.1132	0.5002	0.2252	0.2044	0.0151	0.0122
ucd-cs-run4	6/1/21	ucd-cs	0.462	0.0181	0.0532	0.7944	0.2817	0.2623	0.2886	**0.3182**
0rule_run	6/1/21	baseline	0.2485	0	0.0044	0.8853	0.0444	0.1043	-0.0041	0.0024

Table 8. Evaluations of submitted runs using *all* 2021 data under the v3.1 evaluation script. For each metric, higher is better. Bolded numbers are maximal over the full column. No single system exhibits superior performance across all metrics, though the ens system submitted on 18 October has the highest performance in the most metrics. NOTE: Systems submitted earlier in the year were at a disadvantage, as a partial set of labeled data from the 2021-ALL dataset was released over the summer to support the leaderboard.