Incident Streams 2020: TREC-IS in the Time of COVID-19

Cody Buntain*
InfEco Lab, New Jersey Institute of Technology (NJIT)[†]
cbuntain@njit.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

Between 2018 and 2019, the Incident Streams track (TREC-IS) has developed standard approaches for classifying the types and criticality of information shared in online social spaces during crises, but the introduction of SARS-CoV-2 has shifted the landscape of online crises substantially. While prior editions of TREC-IS have lacked data on large-scale public-health emergencies as these events are exceedingly rare, COVID-19 has introduced an over-abundance of potential data, and significant open questions remain about how existing approaches to crisis informatics and datasets built on other emergencies adapt to this new context. This paper describes how the 2020 edition of TREC-IS has addressed these dual issues by introducing a new COVID-19-specific task for evaluating generalization of existing COVID-19 annotation and system performance to this new context, applied to 11 regions across the globe. TREC-IS has also continued expanding its set of target crises, adding 29 new events and expanding the collection of event types to include explosions, fires, and general storms, making for a total of 9 event types in addition to the new COVID-19 events. Across these events, TREC-IS has 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 demonstrate 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. Using these results, we provide guidance on which information types appear most consistent between general crises and COVID-19.

Keywords

Emergency Management, Crisis Informatics, Twitter, Categorization, Prioritization, COVID-19

INTRODUCTION

A core question in crisis informatics concerns how well datasets and models – and in fact whole processes – developed on one emergency event perform when applied to a new, unseen event. This question is especially critical for rare emergencies or short-duration, fast-onset events, where opportunities for data collection and annotation are limited. The COVID-19 pandemic represents such a crisis, leaving open questions about how well crisis-informatics systems developed for non-COVID events adapt to this new, global emergency. Emergency response officers are therefore less likely to trust or use these systems during a critical moment where social-distancing ordinances may drive social media data volumes to new highs, and public health is at significant risk. Likewise, researchers

*corresponding author
[†] https://computing.njit.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

and system developers have limited evidence about what types of information actually transfer to this new crisis context and whether extant datasets on non-COVID crises should be discarded in favor of collecting wholly new COVID-specific data. This paper investigates these concerns through the lens of the 2020 editions of the TREC Incident Streams (TREC-IS) track, wherein we investigate applications of the TREC-IS information-type and priority taxonomies developed on non-COVID-19 crises to COVID-19 data and explicitly evaluate a collection of crisis-informatics systems in both non-COVID and COVID-19 contexts.

TREC-IS 2020 marks this initiative's third year of promoting research and tooling to better support emergency response services' efforts to harness social media, making it an ideal space to investigate these questions. Each year, the track collects, annotates, and publishes datasets of crisis-related social media data and invites researchers from across the globe to submit automated systems for evaluation in crisis-related text retrieval tasks. This framework is built around a text retrieval task, wherein participant systems label social media content collected from numerous emergency events according to a taxonomy of crisis-related information needs (e.g., requests for search and rescue, reports of emerging threats, or calls for people to evacuate an area) and priority levels (from low to critical). At the same time, as the pandemic is an unprecedented event in modern times, and while TREC-IS has an expansive – and growing – volume of annotated social media data for crises, the rarity of large-scale public health crises has precluded their inclusion in prior TREC-IS editions. Consequently, we have expanded the TREC-IS framework in 2020 to introduce new COVID-19-specific evaluations as well as making available 282,444 social media posts and 19,784 manual annotations of this content from 29 new emergency incidents. As in previous years, TREC-IS 2020 includes two editions, 2020-A and 2020-B, with the 2020-A edition (ran in June 2020) including a new task for annotating COVID-19-related social media data and evaluating systems' adaptability from general crises to several COVID-specific regional contexts. For the 2020-B edition, we have extended this initiative by making available annotated social media data around COVID-19 collected from participants in 2020-A and manually annotated by professional assessors.

Using this framework, resulting annotations of crisis- and COVID-19-related social media content, and systems submitted from six research organizations across the globe, we answer two key sets of research questions. First, we examine 45,325 manual annotations of COVID-19 social media data to evaluate how well the TREC-IS information-type and priority taxonomies adapt to the COVID-19 context. Second, we leverage participant systems to compare performance across general crises to regional COVID-19 contexts. This evaluation covers three aspects: 1) how well systems perform in a new crisis context – i.e., COVID-19 – when trained using general crisis information, 2) how much this performance increases when COVID-19 training data is made available, and 3) whether specific information types adapt more easily across the general-to-COVID-19 context.

Together, answers to these questions will highlight how crisis informatics researchers and practitioners can incorporate COVID-19 and future large-scale public health issues with extant frameworks, models, and crisis-related data collections. These answers and the framework for evaluating them provide the following contributions, which will be of value both to emergency-response officers and to researchers and engineers who study and develop crisis-informatics systems:

- The official overview for TREC-IS 2020, containing the 2020 task description, updated metrics, an expanded general-crisis dataset, and participant performance statistics;
- A large dataset of regional COVID-19 social media content, including an analysis of information types and critical information shared on social media during this global event; and
- A detailed examination of participant systems, how well general crisis informatics systems adapt to a previously unseen crisis and crisis *type*, including insights about which types of information are most conserved in new crises.

RELATED EFFORTS, PRIOR EDITIONS OF TREC-IS, AND COVID-19

In 2020, crisis informatics has seen a wholly new global emergency in COVID-19, and TREC-IS is far from the only effort to study this crisis. This section situates TREC-IS in the context of these related efforts before describing relevant background from prior editions.

Related Crisis Informatics Research into COVID-19

Given the global impact COVID-19 has had, naturally significant research effort has since been devoted to questions of social media use during the pandemic, COVID-19-related misinformation, effective communication channels, and many related concerns. Most related to the TREC-IS initiative are several works that have focused on constructing and sharing large-scale Twitter-centric datasets of COVID-19 discussion. Qazi et al., in particular, have published the GeoCov19 dataset (Qazi et al. 2020) as part of the CrisisNLP project[1] (Imran, Mitra, et al. 2016), and this dataset contains in excess of 524 million COVID-19-related tweets posted since February 2020, replete with a subset of geolocation information. Likewise, Chen, Lerman and Ferrara (Chen et al. 2020) and Banda et al. (Banda et al. 2020) both have released significantly sized datasets of COVID-19 discourse on Twitter. While exceedingly valuable for analyzing discourse around COVID-19 across numerous regional contexts, these datasets lack manual assessments around types of information shared, criticality of information, or comparison between COVID-19 discourse and prior studies of crisis-related discussion. Consequently, if a crisis informatics researcher or emergency response officer wanted to leverage computational decision support when searching for discussions of advice, requests for help, or volunteer opportunities during the pandemic, these stakeholders would have to manually search for this content. With the large volume of labeled crisis data TREC-IS makes available, however, practitioners *could* rely on models built on these prior datasets, but given the unprecedented nature of the pandemic and paucity of outbreak-style data, how well such models would adapt to this new context in these COVID-19 datasets is unclear, clearly motivating this present paper.

Other related efforts, such as that by Abd-Alrazaq et al. (Abd-Alrazaq et al. 2020), have studied social media users' COVID-19-related discourse, decomposing these discussions into topics to identify which concerns are receiving the most attention. While results suggest high variance in topical discussions, many are related to speculation around the virus's origin, economic impact, and mitigation efforts. Again, these efforts, while crucial to understanding public concern around the pandemic, these efforts do little to contextualize these discussions with results from prior study of crisis communication. TREC-IS, on the other hand, extends a general taxonomy of crisis-related information types to the COVID-19 pandemic, allowing us to evaluate differences in public concern and topical popularity between COVID-19 and prior crises.

Despite these deviations in studies, however, these related works reach a conclusion consistent with the TREC-IS initiative: Namely that national and international efforts around public health should engage more with social media content for surveillance, monitoring, and correcting misinformation (Abd-Alrazaq et al. 2020).

An Overview Prior TREC-IS 2018 and 2019 Initiatives

As 2020 is the third year of the TREC-IS initiative, this section briefly describes the key organizational aspects of TREC-IS 2018 and 2019, shown in Figure 1, as they provide the foundation for our discussion of TREC-IS in 2020. At a high level, participant TREC-IS systems are intended to produce two outputs for crisis-related social media content: classifying messages by "information type", and ranking these messages by their criticality. To this point, these social media messages have focused on Twitter content – i.e., tweets. As shown in Figure 1, TREC-IS provides participants with a stream of filtered, event-relevant tweets and an ontology of information types; then, each system assigns information-type labels and priority ratings to *all* of these messages, which they then submit back to TREC-IS for evaluation. Below, we operationalize these information types, priorities, ground truth labeling using professional human assessors, and summarize the primary findings of these prior TREC-IS efforts.

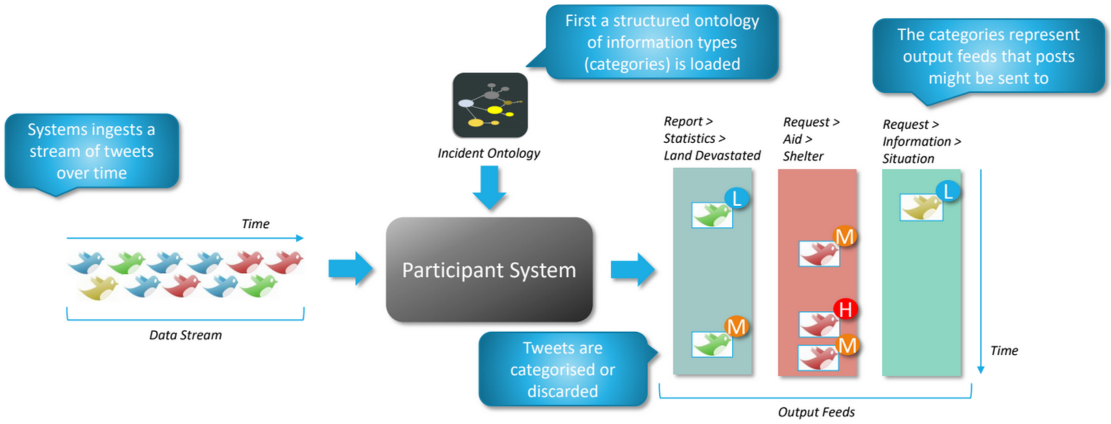

First a structured ontology of information types (categories) is loaded
The categories represent output feeds that posts might be sent to
Systems ingests a stream of tweets over time
Incident Ontology
Report > Statistics > Land Devastated
Request > Aid > Shelter
Request > Information > Situation
Time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

Figure 1. TREC-IS Task Visualization

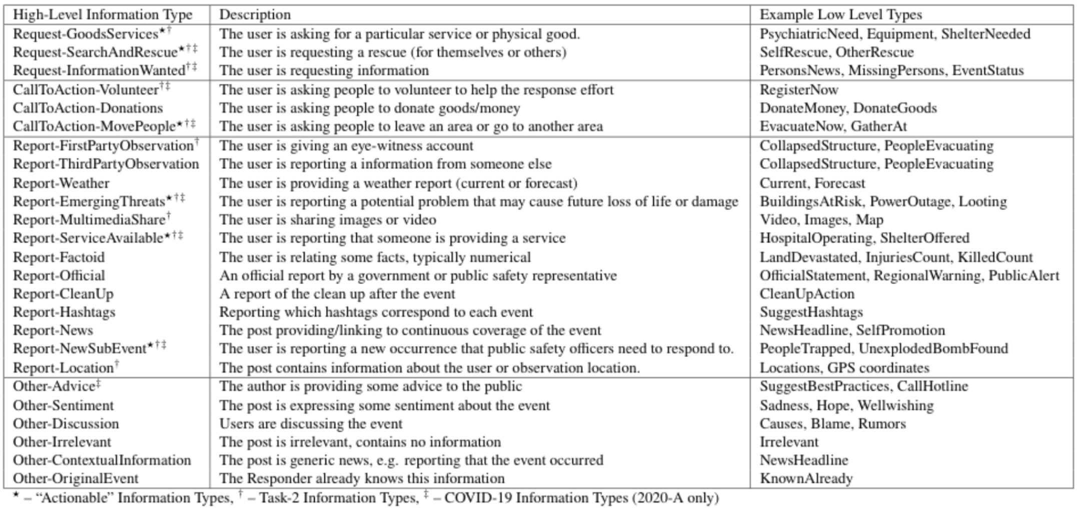
An Ontology of Crisis-Related Information Types Early versions of the TREC-IS track defined information 'types' that represented the categories of information an emergency response officer might be interested in, such

[1] https://crisisnlp.qcri.org/

as ‘Reports of Road Blockages’ or ‘Calls for Help’. We defined these types based on a top-down analysis of incident management ontologies such as MOAC (Management of a Crisis)[2], response documentation and discussion with experts. Prior TREC-IS editions have found this valuable information to be rare on social media, as most information shared during emergencies is of lower importance, such as news reports or shocking images (McCreadie et al. 2020). On the other hand, these other types of information may be valuable to researchers and practitioners engaged in a wide variety of tasks, such as public-health surveillance, volunteer coordination, or other support functions, as outlined in the Federal Emergency Management Agency’s Emergency Support Functions (ESFs). Hence, we expanded the initial set of information types via a bottom-up analysis of tweets from a small sample of events. This resulted in additional information types such as ‘Sentiment Expressed’, ‘Press Releases’ and ‘Sharing Best Practices’. Following the 2018 pilot of TREC-IS, we have defined a total of 25 high-level types, as shown in Table 1. Based on analysis from prior TREC-IS iterations, we also have identified 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.

Table 1. Ontology High-level Information Types



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
Report-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

$\star$ – “Actionable” Information Types, $\dagger$ – Task-2 Information Types, $\ddagger$ – COVID-19 Information Types (2020-A only)

Measuring Message Criticality To capture the importance a given message has to emergency response officers, we also measure the perceived criticality of a given message. For manual assessment, we use four distinct criticality labels: low, medium, high, and critical. High- and critical-level messages require prompt or immediate review and potentially action by an emergency manager. Examples of critical information included calls for search and rescue, emergence of new threats (e.g., a new gunman, aftershock, or secondary event), or calls for evacuation. For participant systems, however, we have experimented with casting criticality measurement as a continuous regression task to predict a criticality score (as in 2018, 2019, and 2020-B) or as a four-label classification task using the same labels as used with assessors (as in 2020-A).

Evaluating Participant Systems To evaluate participant systems, we require ground-truth data on information-type and criticality labels for the above tweets. For the 2018 and 2019 efforts, TREC-IS coordinators collected Twitter data for 33 crisis events, partially from prior crisis informatics efforts, including CrisisLex (Olteanu et al. 2015) and CrisisNLP (Imran, Mitra, et al. 2016; Imran, Elbassuoni, et al. 2013), which coordinators then supplemented with additional TREC-IS-specific collections, primarily released as part of 2019-B (see Table 2). For each event, TREC-IS coordinators de-duplicated and sampled the resulting Twitter collections to create numerous event-specific

[2] http://observedchange.com/moac/ns/

tweet collections from which ground truth could be evaluated. NIST-hired assessors have labeled these tweets, marking priority and *all* relevant information types (taken from Table 1) per message. Following TREC-IS 2018 and 2019, this labeling has resulted in 46,741 manually labeled tweets.

To evaluate the state of the art in crisis informatics, TREC-IS releases subsets of these events, replete with these manual labels. Each TREC-IS edition then releases a new set of events *without* these labels, so participants can develop cutting-edge prototypes using this labeled content and submit predicted information-type labels and priority scores for the new event sets. Each participating research group is then allowed to submit up to four "runs" from their candidate systems, each of which is evaluated separately.

These runs are then evaluated across two axes: information-type categorization, and information criticality. For information-type categorization, participant systems submit (potentially multiple) information-type labels, which are then evaluated using standard precision, recall, and F1-score, macro-averaged across the evaluation events. As mentioned, 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. For the second axis on ranking information criticality, we want to up-weight information that emergency response officers need to see. In the prior 2018 and 2019 iterations of TREC-IS, since our criticality labels are ordered (e.g., "low" is less than "critical"), we assign numeric scores to these labels and calculate the *Mean Squared Error* between the human-assigned score and a system's score for each tweet.

Main Conclusions from TREC-IS 2018 and 2019 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, the six actionable information types we describe above are likewise stable in that tweets with these information-types are consistently perceived to be of higher-priority than the remaining 19 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. These perceptions of criticality are of particular interest in the TREC-IS 2020 edition as the definition of high-priority content may be altered dramatically when faced with long-lived, global events like the COVID-19 pandemic.

While the above results focus on the distribution of information within crises, participant-system evaluations from prior TREC-IS iterations also found that cutting-edge systems of the time were insufficient for end users' needs in classifying information type and priority. While participants were relatively effective at identifying news reports and sentiment, they struggled to identify critical information like search and rescue requests (McCreadie et al. 2019). Though system-performance comparisons across TREC-IS editions are generally difficult to interpret given the variation in crisis events and the ever-expanding volume of training data TREC-IS makes available, we find systems are generally becoming more performant from edition to edition. Likewise, systems relying on neural language models are also increasingly performant compared to traditional learning models (McCreadie et al. 2020). Despite these trend, performance in identifying actionable information lagged behind the identifying all relevant information types, suggesting significant room for improvement remains.

CHANGES TO THE INCIDENT STREAMS TRACK IN 2020

To address open questions around COVID-19, in this third year of the TREC-IS initiative, we have made several modifications and extensions to the track. At the same time, the track has continued its focus on curating feeds of social media posts, where each feed corresponds to a particular type of information request, aid request, or report. As in the previous year, TREC-IS expanded its collection of events, adding 29 new emergency incidents, 282,444 social media posts, and 19,784 manual annotations of this content. Following TREC-IS 2019, we also ran the track twice this year, a first edition in June 2020, which we call 2020-A, and the main TREC edition in September 2020, called 2020-B. In both editions, participant systems could submit systems for the core task, *classifying tweets by information type (high-level), version 2*, which is carried over from TREC-IS 2019 and includes both information-type classification and prioritization.

In 2020, the track has introduced three major changes: First, we have added a COVID-19-specific task in both the 2020-A and 2020-B editions to investigate adaptability of TREC-IS data and processes and how well systems perform with and without COVID-19 training data. Second, we have added a new task to restrict the number of information types to a core set of 11 high-priority classes (intended to allow participants to focus on a core set of information types). And third, to address the growing dataset sizes for events and improve coverage over these events, TREC-IS has moved to post-hoc pooling for evaluation rather than releasing a pre-annotated dataset for each event. We have also instituted smaller changes in run types, formats, and new metrics, which we describe below. In sum, the new TREC-IS 2020 track is comprised of the following three tasks:

Task 1. All High-Level Information Type Classification This task is an extension of the core TREC-IS task used in 2018 and 2019. Systems participating in this task are given tweet streams from a collection of crisis events and should classify each tweet as having one or more of the high-level information types described in the ontology section above. Each tweet should be assigned as many categories as are appropriate. Likewise, each message must be tagged with a numeric measure of priority, on a $[0, 1]$ scale (in 2020-A, we used an ordinal value of "Low", "Medium", "High", or "Critical" but returned to the numeric scale after discussions with participants).

Task 2. Selected High-Level Information Type Classification (Introduced in 2020-A) 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 now includes 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, as shown in Table 1 (marked by a †).

Task 3. COVID-19-Specific Information-Type Classification (Introduced in 2020-A) In this task, systems are intended 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. COVID-19 is different from prior events, however, in that its impacts are global in scale and are difficult to restrict to a single region as we can with wildfires or earthquakes. It is therefore difficult to identify primary affected locations (whereas hurricanes or wildfires are geographically constrained). Hence, for COVID-19 TREC-IS focuses data collections around regional areas of interest, primarily population centers. In 2020-A, this task restricted the set of information types to a subset of eight relevant types (see Table 1 for types marked with a ‡) as the track lacked training data for public health events. For 2020-B, however, we have removed the category restriction as analyses of 2020-A COVID-19 data has found instances across 24 of the 25 information types (weather-related content was not found). Consequently, this Task 3 parallels Task 1 and differs only in the data against which participant systems are evaluated.

New Datasets

For standardized evaluations of systems, TREC-IS provides participants with training and test datasets, comprised of three components: the ontology of high-level information types, a collection of crisis-event descriptions, and the tweets for each event to be categorized. In its original form, the TREC-IS track has aimed to enhance crisis management scenarios across a variety of crisis events, both natural and manmade, so the track has published these three-component datasets over numerous crises. Prior editions have repurposed extant datasets to create these standardized evaluation sets, but in 2020, we have expanded both the collection processes and resulting event collections substantially. To encourage broad applicability of systems, these collections cover the following event types (* are new to TREC-IS 2020):

- Wildfires,
- Earthquakes,
- Floods,
- Typhoons/Hurricanes,
- Shootings,
- Bombing,
- Hostage situations,*
- Accidental explosions,* and
- Structural fires.*

To develop these new crisis datasets for 2020, TREC-IS coordinators have continued with the 2019 effort in constructing custom datasets by tracking ongoing crises in a prospective fashion. Constructing these collections has required coordinators to scour news coverage of global crisis events, and on identification of a sufficiently impactful event, coordinators would begin collecting streams of data around these events using Twitter's public API. Consequently, TREC-IS 2020 relies heavily on current data, but this reliance has enabled coordinators to expand the set of crisis events substantially, as shown in Table 2. This table shows that, compared to prior editions, 2020-A and 2020-B have provided participants with significantly more crisis datasets, totaling 15 and 14 events and associated social media messages respectively, as well as COVID-19 data on 9 regions. Data collection in 2020 has resulted in $45,325$ new, manually assessed social media messages around crises, 22,401 of which are from regions with heavy COVID-19 impact.

An important aspect of this new dataset creation method is that TREC-IS now collects significantly more content than can be assessed manually in reasonable timeframe. For COVID-19 events in particular, this overabundance of data was problematic, as the prior TREC-IS initiatives have first sampled data from large events for manual assessment prior to releasing test sets to participants; in the time of COVID-19, however, the track did not have time to manually assess these large samples prior to release. TREC-IS 2020 has therefore moved to a model where coordinators collect these datasets, de-duplicate data using locality-sensitive hashing – using the Nilsimsa digest (Damiani et al. 2004) – to filter highly similar tweets and retweet, and then release the full datasets to participants for labeling and prioritization. Coordinators then group the resulting participants' runs using a pooling method to select a subset of messages to assess manually. While this process introduces complexity and requires participant systems to label nearly an order of magnitude more messages than will be judged, this method allows coordinators to leverage state-of-the-art technology when identifying valuable or critical content to assess. Otherwise, coordinators would need to develop an internal method for ranking messages according to information type and priority to identify what should be judged, and if this problem were easy to solve, the track would not exist. We discuss this pooling method below.

Pooling

In information retrieval test collections, the standard practice is to label a small sample of the dataset, and to focus that sample where positive examples are likely to be found. The most common approach is *pooling*: take the top-k ranked results from a diverse set of state-of-the-art systems, and manually review the resulting *pool*. This process will yield an unbiased, sufficiently-complete set of labels to measure the pooled systems (Zobel 1998). When the system outputs are not ranked, for example if the system is a binary classifier on a stream where feedback is updating the classifier, stratified sampling is used (Lewis 1995).

In 2020-A, following the practice from earlier editions of TREC-IS, systems returned information type labels for each tweet along with a priority level. Pooling based on priority level alone could bias the pool against information types that are typically low priority (like Report-News) or where priority is hard to predict. Accordingly, we drew a sample stratified on priority levels from the system outputs using reservoir sampling up to a maximum of 2,500 tweets per event. For 2020-B, to support more straightforward pooling approaches, we required systems to return a classification probability for every information type and a numeric priority score. We pooled 2020-B to the top 100 tweets from every system by both priority and by information type probability. This approach resulted in an variation in the number of tweets per event but would theoretically provide sufficient coverage over the 25 TREC-IS information types and over the priority spectrum.

Tweet Labeling Process

Once TREC-IS coordinators have constructed Twitter samples for each event, each tweet in the TREC-IS collections is labeled by a TREC assessor. These assessors all have strong information analysis skills and experience in labeling text and social media for TREC-IS and other TREC retrieval and classification tasks. Prior to performing labeling tasks, each assessor receives a two-hour, in-person training session that includes an overview of the emergency event scenario and guidance on identifying actionable information within each event type. Assessors then exercise their training in a guided, hands-on, group labeling session using the 2012 Colorado wildfires event. Assessors are also allowed to use Wikipedia entries for each event to familiarize themselves with its timeline and geography.

For each labeling task, assessors use an assessment tool that displays the raw text of the tweet and renders it using Twitter's API, which replicates the view (replete with embedded multimedia) a user would see natively on Twitter. Assessors then decide if each tweet is actually relevant to the event, if it contains actionable information, and assigns one or more category labels and a priority level to the tweet. In most cases, all tweets collected for a single event are labeled by one assessor to minimize inconsistencies within an event that could arise from disagreement between

Table 2. TREC-IS Events

Dataset	Identifier	Event Name	Event Type	Provided Tweets	Labeled Tweets
2018 Training	TRECIS-CTIT-H-Training-001	2012 Colorado wildfires⋆	wildfire	744	744
	TRECIS-CTIT-H-Training-002	2012 Costa Rica Earthquake⋆	earthquake	288	288
	TRECIS-CTIT-H-Training-003	2013 Colorado Floods⋆	flood	777	777
	TRECIS-CTIT-H-Training-004	2012 Typhoon Pablo⋆	typhoon/hurricane	649	649
	TRECIS-CTIT-H-Training-005	2013 LA Airport Shooting⋆	shooting	683	683
	TRECIS-CTIT-H-Training-006	2013 West Texas Explosion⋆	bombing	630	630
2019-A	TRECIS-CTIT-H-Test-007	2012 Guatemala earthquake⋆	earthquake	178	178
	TRECIS-CTIT-H-Test-008	2012 Italy earthquakes⋆	earthquake	118	118
	TRECIS-CTIT-H-Test-009	2012 Philippines floods⋆	flood	480	480
	TRECIS-CTIT-H-Test-010	2013 Alberta floods⋆	flood	739	739
	TRECIS-CTIT-H-Test-011	2013 Australia bushfire⋆	wildfire	710	710
	TRECIS-CTIT-H-Test-012	2013 Boston bombings⋆	bombing	543	543
	TRECIS-CTIT-H-Test-013	2013 Manila floods⋆	flood	443	443
	TRECIS-CTIT-H-Test-014	2013 Queensland floods⋆	flood	744	744
	TRECIS-CTIT-H-Test-015	2013 Typhoon Yolanda⋆	typhoon	629	629
	TRECIS-CTIT-H-Test-016	2011 Joplin tornado⋆⋆⋆	typhoon	152	152
	TRECIS-CTIT-H-Test-017	2014 Chile Earthquake⋆⋆	earthquake	321	321
	TRECIS-CTIT-H-Test-018	2014 Typhoon Hagupit⋆⋆	typhoon	6,696	6,696
	TRECIS-CTIT-H-Test-019	2015 Nepal Earthquake⋆⋆	earthquake	7,301	7,301
	TRECIS-CTIT-H-Test-020	2018 FL School Shooting	shooting	1,118	1,118
	TRECIS-CTIT-H-Test-021	2015 Paris attacks	bombing	2,066	2,066
2019-B	TRECIS-CTIT-H-Test-022	2019 Choco Flood	flood	854	854
	TRECIS-CTIT-H-Test-023	2019 Andover Fire	wildfire	375	375
	TRECIS-CTIT-H-Test-024	2014 California Earthquake	earthquake	128	128
	TRECIS-CTIT-H-Test-025	2013 Bohol Earthquake	earthquake	646	646
	TRECIS-CTIT-H-Test-026	2018 Florence Hurricane†	typhoon	2,500	2,500
	TRECIS-CTIT-H-Test-027	2017 Dallas Shooting	shooting	2,500	2,500
	TRECIS-CTIT-H-Test-028	2016 Fort McMurray Wildfire	wildfire	2,500	2,500
	TRECIS-CTIT-H-Test-029	2019 Alberta Wildfires	wildfire	2,500	2,500
	TRECIS-CTIT-H-Test-030	2019 Cyclone Kenneth	typhoon	2,500	2,500
	TRECIS-CTIT-H-Test-031	2019 Luzon earthquake	earthquake	2,500	2,500
	TRECIS-CTIT-H-Test-032	2019 STEM School Highlands Ranch shooting	shooting	2,500	2,500
	TRECIS-CTIT-H-Test-033	2019 Durban Easter floods	flood	2,500	2,500
	TRECIS-CTIT-H-Test-034	2019 Poway synagogue shooting	shooting	2,500	2,500
2020-A	TRECIS-CTIT-H-Test-035	2019 Greece Earthquake	earthquake	501	473
	TRECIS-CTIT-H-Test-036	2020 Baltimore Floods	flood	501	469
	TRECIS-CTIT-H-Test-037	2020 Brooklyn block party shooting	shooting	501	479
	TRECIS-CTIT-H-Test-038	2020 Dayton Ohio shooting	flood	501	475
	TRECIS-CTIT-H-Test-039	2020 El Paso Walmart shooting	shooting	501	451
	TRECIS-CTIT-H-Test-040	2020 Gilroy Garlic Festival shooting	shooting	501	470
	TRECIS-CTIT-H-Test-041	2020 Hurricane Barry	typhoon	501	469
	TRECIS-CTIT-H-Test-042	2020 Indonesia Earthquake	earthquake	501	470
	TRECIS-CTIT-H-Test-043	2020 Kerala, India floods	flood	501	391
	TRECIS-CTIT-H-Test-044	2020 Myanmar floods	flood	501	470
	TRECIS-CTIT-H-Test-045	2020 Papua New Guinea earthquake	earthquake	501	447
	TRECIS-CTIT-H-Test-046	2020 Siberian Wildfires	wildfire	501	474
	TRECIS-CTIT-H-Test-047	2020 Typhoon Krosa	typhoon	501	466
	TRECIS-CTIT-H-Test-048	2020 Typhoon Lekima	typhoon	501	469
	TRECIS-CTIT-H-Test-049	2020 Whaley Bridge Dam Collapse	accident	501	210
	TRECIS-CTIT-H-Test-050	2020 COVID-19 Outbreak in Washington DC	COVID-19	49,894	4,012
	TRECIS-CTIT-H-Test-051	2020 COVID-19 Outbreak in Washington State	COVID-19	48,499	5,697
	TRECIS-CTIT-H-Test-052	2020 COVID-19 Outbreak in New York	COVID-19	50,000	5,126
2020-B	TRECIS-CTIT-H-Test-053	2020 Houston explosion	explosion	7,877	1,416
	TRECIS-CTIT-H-Test-054	2020 Texas A&M-Commerce shooting	shooting	3,490	2,638
	TRECIS-CTIT-H-Test-055	2020 Southeast US Tornado Outbreak	storm	4,136	1,348
	TRECIS-CTIT-H-Test-056	2020 Storm Ciara	storm	2,879	1,075
	TRECIS-CTIT-H-Test-057	2020 Storm Dennis	storm	50,000	1,851
	TRECIS-CTIT-H-Test-058	2020 Porterville library fire	structure fire	131	–
	TRECIS-CTIT-H-Test-059	2020 Virra Mall Hostage Situation	hostage	2,294	1,149
	TRECIS-CTIT-H-Test-060	2020 Storm Jorge	storm	25,777	602
	TRECIS-CTIT-H-Test-061	2020 Tennessee tornado outbreak	storm	50,000	1,278
	TRECIS-CTIT-H-Test-062	2020 Tennessee derecho	storm	3,853	500
	TRECIS-CTIT-H-Test-063	2020 Edenville dam failure	accident	35,465	1,586
	TRECIS-CTIT-H-Test-064	2020 San Francisco Pier Fire	structural fire	2,752	1,298
	TRECIS-CTIT-H-Test-065	2020 Tropical Storm Cristobal	storm	36,275	1,000
	TRECIS-CTIT-H-Test-066	2020 Beirut Explosion	explosion	50,000	500
	TRECIS-CTIT-H-Test-067	2020 COVID-19 Outbreak in Miami, FL	COVID-19	50,000	–
	TRECIS-CTIT-H-Test-068	2020 COVID-19 Outbreak in Jacksonville, FL	COVID-19	13,506	664
	TRECIS-CTIT-H-Test-069	2020 COVID-19 Outbreak in Houston, TX	COVID-19	44,297	963
	TRECIS-CTIT-H-Test-070	2020 COVID-19 Outbreak in Phoenix, AZ	COVID-19	16,766	871
	TRECIS-CTIT-H-Test-071	2020 COVID-19 Outbreak in Atlanta, GA	COVID-19	50,000	966
	TRECIS-CTIT-H-Test-072	2020 COVID-19 Outbreak in New York, part 2	COVID-19	50,000	1,396
	TRECIS-CTIT-H-Test-073	2020 COVID-19 Outbreak in Seattle, WA	COVID-19	50,000	966
	TRECIS-CTIT-H-Test-074	2020 COVID-19 Outbreak in Melbourne, AU	COVID-19	50,000	957
	TRECIS-CTIT-H-Test-075	2020 COVID-19 Outbreak in New Zealand	COVID-19	5,148	783

⋆ – Collected from CrisisLexT26 (**olteanu2015expect**)

⋆⋆ – CrisisNLP Resource #1 (**imran2016twitter**)

⋆⋆⋆ – CrisisNLP Resource #2 (**imran2013practical**)

† – Donated by participant group

assessors. 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).

Metrics for TREC-IS 2020

Information Feed Metrics: TREC-IS 2018 and 2019 primarily used standard accuracy, precision, recall, and F1 as metrics for system evaluation, with TREC-IS 2019 introducing a sub-metric targeting the six high-priority, actionable information types described above. For overall classification accuracy, this metric is micro-averaged over all events (larger events will have more influence[3]), but macro-averaged over information types[4]. However, emergency service operators primarily care about whether they are shown *all* of the valuable information or not, i.e. missing actionable information is a much more serious failure than returning noise. As such, we also report classification F1 only on the positive class, micro-averaged over all events. Further, we report both performance when considering all information types (actionable and non-actionable), as well as when only considering only actionable information types.

- **Information Feed, Info. Type Accuracy, All**: Overall classification accuracy, micro-averaged across events and macro-averaged across information types. Gives a high-level view of categorization performance, but does not capture whether the output is useful to our end-users.
- **Information Feed, Info. Type Positive F1, All**: Categorization performance when only considering the target class per information type. For example, when calculating performance for the 'Request-SearchAndRescue' information type, we only calculate performance over the small number of tweets that belong to that type, ignoring all other tweets. As such, this measures the signal that is shown to the user, while ignoring any noise that the system also produces. The 'All' version of this metric macro-averages over all information types.
- **Information Feed, Info. Type Positive F1, Actionable**: The same as the above metric, but only considers the actionable information types shown in Table 1. This metric aims to capture whether systems are able to find the actionable information within the stream.

Information Priority Metrics: The second aspect of a TREC-IS system that we want to evaluate is to what extent can it identify key information that the emergency response officer needs to see. This is operationalized by comparing the information priority score provided by each system and the priority label provided by the human assessor per tweet. To enable such a comparison and to bring system output in line with the four information priority labels used by the human assessors, we mapped priority scores into discrete priority labels. In this case, low=0.25, medium=0.5, high=0.75 and critical=1.0. We then evaluate participant systems using F1 score around these discrete priority labels, and as with the information feed metrics, to distinguish between prioritization performance for actionable categories against all categories, we report prioritization error for both all information types and only actionable information types. To account for the relative importance of these messages and the general scarcity with which they appear (McCreadie et al. 2020), we also measure normalized discounted cumulative gain nDCG@100 (i.e., over the first one hundred highest-priority tweets), macro-averaged across events.

- **Prioritization F1, All**: Categorization performance when only considering the discretized priority labels. Higher is better.
- **Prioritization F1, Actionable**: The same as the above metric, but only considers the actionable information types shown in Table 1. Higher is better.
- **Prioritization nDCG@100**: This metric ranks the quality of the top-k messages according to their priority and compares this value to the maximum potential gain in this ranking. Here, lower is better.

RESULTS IN COMPARING COVID-19 AND OTHER CRISES

Having established the structure for TREC-IS 2020, we now turn to this paper's core contributions: a comparative analysis of the manual assessments around COVID-19 versus other crises, and an evaluation of how well participant systems perform when integrating extant crisis-informatics datasets. For comparing regional COVID-19 datasets to other crises, we examine 45,325 manually assessed tweets to answer the following research questions:

[3] This is done because some events have very few tweets.

[4] Some information types, such as Continuing News are very common, and we don't want performance on those categories to dominate the performance score.

RQ1.1. How do the distributions of information type and priority compare between COVID-19 and general-crisis events?

RQ1.2. For an individual information type, is the distribution information priority consistent between COVID-19 and general-crisis events?

Second, we leverage participant systems to compare performance across general crises to regional COVID-19 contexts. This evaluation covers three aspects: 1) how well systems perform in a new crisis context – i.e., COVID-19 – when trained using general crisis information, 2) how much this performance increases when COVID-19 training data is made available, and 3) whether specific information types adapt more easily across the general-to-COVID-19 context:

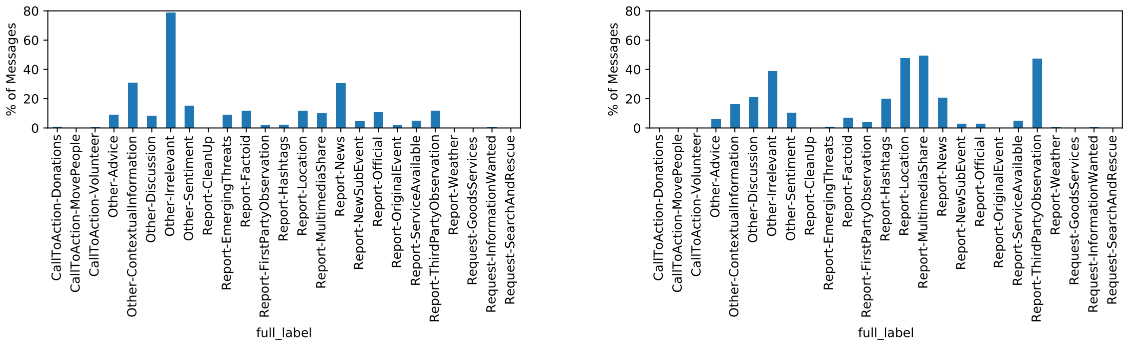
RQ2.1: How well do systems perform information-type classification and prioritization in the context of COVID-19 using only data from prior general-crisis events on which TREC-IS has focused?

RQ2.2: How do systems perform information-type classification and prioritization during COVID-19 when additional COVID-specific training data is made available?

RQ2.3: Do systems perform significantly different for particular information types in the COVID-19 context compared to the general crisis-event context?

RQ1.1 – Comparing Distributions of Information Type and Priority in COVID-19

The COVID-19 construct in 2020-A asked participant systems to classify only a subset of the 25 information types available based on the assumption that many of the information types would not be present in COVID-19 data. During annotation, however, we allowed assessors to evaluate content against all 25 types, and surprisingly, we found content across all but three of the information types. Figure 2 shows the mean distribution of these labels across the three COVID-19 events in 2020-A (Figure 2a) and eight labeled COVID-19 events in 2020-B (Figure 2b).

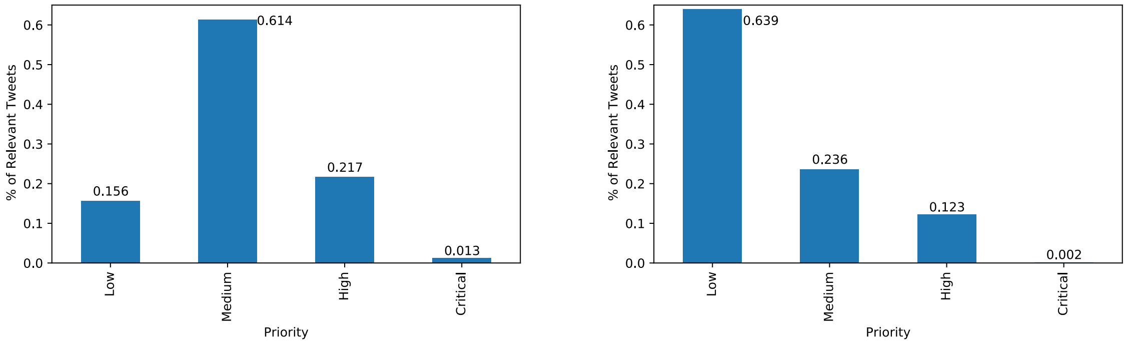


(a) 2020-A Info-Type Distribution **(b) 2020-B Info-Type Distribution**

Figure 2. Distribution of Information Types for 2020-A COVID-19 Events

Regarding priority in COVID-19 content, Figure 3 shows the distributions of priority assessments for relevant tweets in 2020-A and 2020-B. In the three 2020-A events, we find an abundance of "Medium" and "High" priority content, more so than in the general events seen in prior TREC-IS iterations (Figure 3a); critical messages, however, remain consistently rare. In 2020-B, of the eight assessed COVID-19 events, we find a more expected distribution with many "Low" and "Medium" priority content (Figure 3b). We hypothesize that the differences in COVID-19 events in 2020-A and 2020-B are driven by different collection and assessment timeframes, where the relative novelty of the pandemic in early 2020 (when 2020-A was run), assessor perceptions of priority may be artificially higher. In 2020-B, however, as these assessments were completed in November/December 2020, perceptions of criticality may have shifted.

While differences between 2020-A and 2020-B expose some variations and some consistencies in COVID-19 events, the larger question concerns the comparison of information shared between COVID-19 events and other, more general crises, Figure 4 shows the difference in the overall prevalence of information types (measured as a proportion of all tweets) for the 28 crisis events and 11 COVID-19 locations according to our human assessors. In this figure, a negative value indicates information that is more common for COVID-19 events, while positive values indicate information types more common in general crisis events.



(a) 2020-A Priority Distribution

(b) 2020-B Priority Distribution

Figure 3. Distribution of Mean Priority Labels for COVID-19 Events, Macro-Averaged Over Events

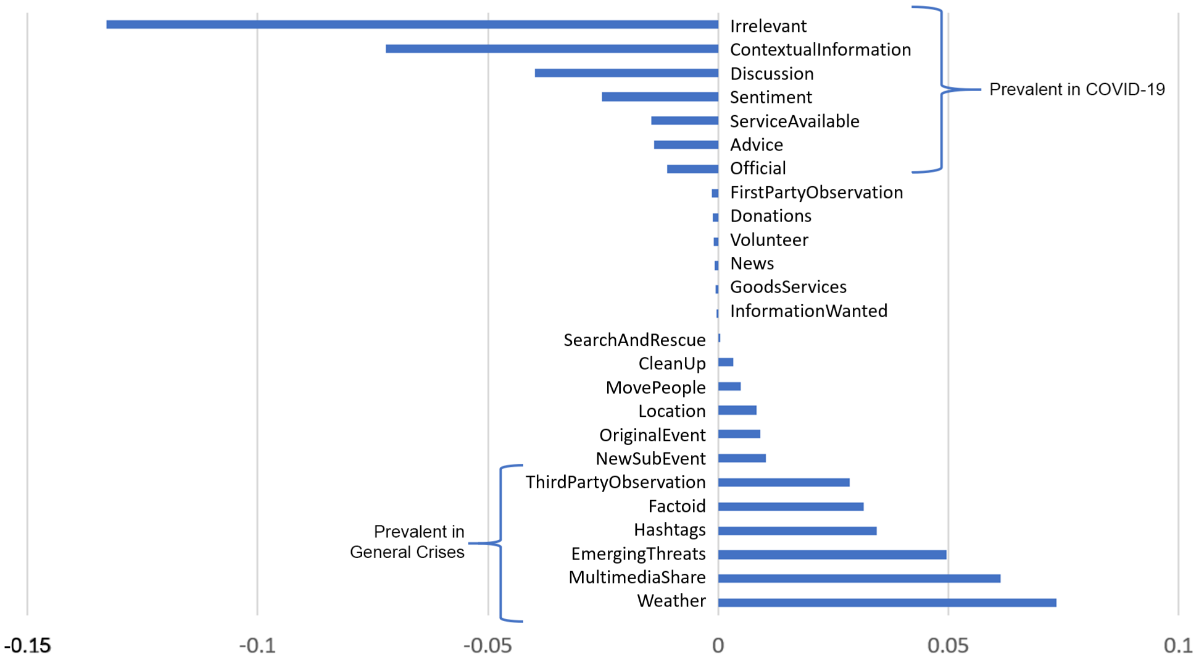


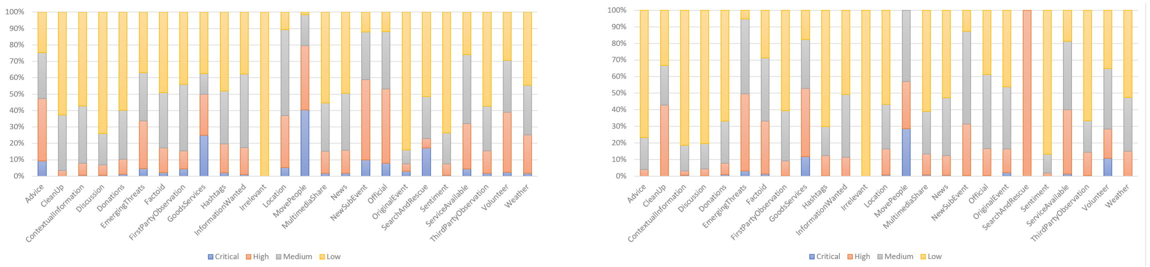
Figure 4. Difference in proportion of Information Types across Crisis and COVID-19 events (2020-A/2020-B).

From this figure, we first observe that overall proportion ranges never exceed around +/-15%, so while the differences are notable, they are not extreme in terms of raw tweet counts. Despite this relative consistency, it should be highlighted that a seemingly small change in the overall proportion might reflect a large change the relative proportion for that information type – e.g., an approximately 7% overall reduction Weather related tweets between general crises and COVID-19 corresponds to a 90% relative reduction (there are almost no weather related tweets for COVID-19 events). Second, focusing in the top and bottom of Figure 4 reveal the most extreme examples: In crisis events, reports of weather, multimedia sharing and emerging threats are more common. For Weather and Multimedia sharing, this result is expected, as there is almost no weather reporting during COVID-19, and less opportunity exists for sharing of disaster imaging in COVID-19 (as oppposed to aftermath of earthquakes or hurricanes, that may have many more instances of demolished buildings). The small volume of emerging-threat content for COVID-19 events is reflective of the long-term nature of the pandemic, where the situation changes slowly. Indeed, what might be considered an “emerging threat” is quite different for COVID-19 events than in general crises. In COVID-19, for example, lack of or behind-schedule testing or reports of quarantine breakers may be considered emerging threats as precursors to outbreaks, but these messages were rare in our sampled data.

Meanwhile, if we examine what information types were more common in COVID-19 events, we see more advice, sentiment, discussion, contextual information, reports of services available and irrelevant content. The higher prevalence of general discussion (discussion, advice, contextual information and sentiment information types) indicates that much of the content for COVID-19 is potentially less useful for direct response than for crises. The higher relative proportion of tweets reporting services available are almost all reports regarding COVID-19 testing centres (when and where they will be available). Hence, to conclude on RQ1.1, there are notable differences in the types of information that are posed during crisis and covid events.

RQ1.2 – Comparing Priorities of Information Types in COVID-19

While we see unsurprising differences in the distributions of information types in COVID-19 events, considering these information types in isolation only tells part of the story. As highlighted above, the meaning of each information type (or at least what types of tweets get labled for each type) differs between crisis events and COVID-19. As such, we next examine how the perceived priority of information for each of the information types varies between the crisis events and our COVID-19-impacted regions. Figures 5 (a) and (b) report the ratio of tweets for each information type labeled as having each priority level. The lower portion of each bar (blue and orange) indicates higher priorities.



(a) *Crisis* Events (2020-A/2020-B). **(b) *COVID-19* Events (2020-A/2020-B).**

Figure 5. Ratio of Information Types across Priority Levels

Comparing Figures 5 (a) and (b), we are looking for differences in the ratio of tweets labeled as either Critical and High. From this comparison, we can make the following observations: First, some information types were perceived to be less important moving from general crises to COVID-19. Most notably, advice, location information, reporting of sub-events, and official reports were marked as less critical. Meanwhile, information types such as reporting of clean-up operations and volunteering were more commonly labeled as either high or critical. Also of note is that emerging threats, while much rarer for COVID-19 (e.g., see above), are consistently perceived as important, just as with general crises. Hence, to answer RQ1.2, we do indeed see marked differences in the criticality of some (but not all) information types. This result, in particular, has important implications for prioritization of content, as prior studies of TREC-IS data shows information type is a strong predictor of criticality.

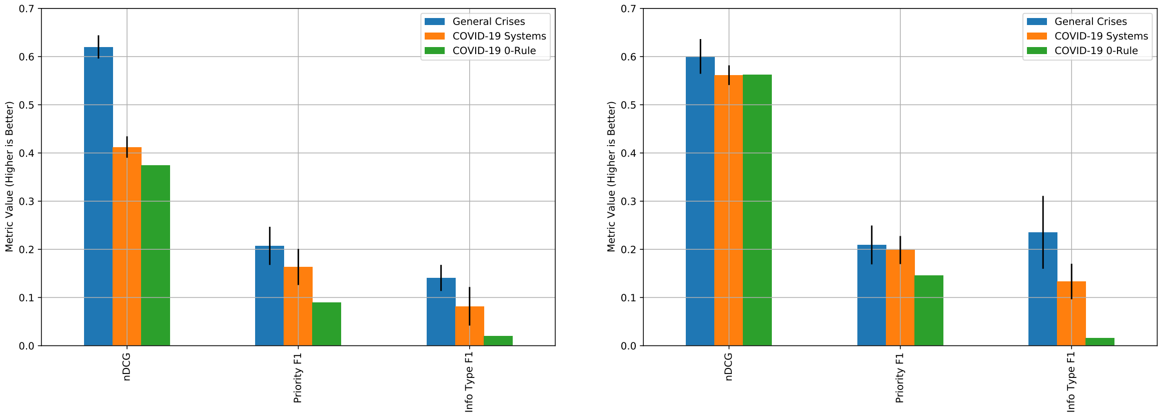
RQ2.1 – Performance in Adapting General-Crisis Datasets to COVID-19

A core question in this analysis is whether and how well automated systems trained on general, non-pandemic crisis data are able to adapt to a new, global event type – namely, the COVID-19 pandemic. Prior to the 2020-A edition, TREC-IS has not focused on large public-health crises, as these events are typically rare, making datasets difficult to construct. Introducing the COVID-19 task into TREC-IS 2020-A has provided an opportunity to evaluate how well such systems can adapt to this new context. While COVID-19 and related public-health crises may be significantly different from other event types (as answers to RQ1 above suggest), several information types may be sufficiently conserved in their language so as to enable this adaptation. To answer RQ2.1, we first evaluate mean system performance across the TREC-IS 2020-A Task 3 ($n = 10$) participant systems, focusing on the seven information types shown in Table 1 (we exclude the “CallToAction-MovePeople” type, as assessors found no instances of this type in the 2020-A instance of Task 3). We first compare this performance against a zero-rule baseline for COVID-19 data, wherein all tweets are labeled as the “Other-Advice” information type and “Low” priority – these labels are the most common information type and priority label calculated across the TREC-IS 2018 and 2019 datasets. We also contextualize this performance by comparing Task 3 system performance to Task 1 ($n = 14$) participant systems. Figure 6a presents results of this analysis, showing that Task-3 systems exhibit significantly higher performance than the zero-rule baseline in 2020-A but lower performance compared to Task 1, significantly so for both nDCG ($p << 0.001$) and Info-Type F1 ($p < 0.05$). Comparing prioritization F1 between Task 1 and Task 3, surprisingly, suggests systems trained on non-COVID data and applied to COVID-19 events do not perform significantly lower in prioritization ($p > 0.05$), implying some aspects of prioritization adapt comparatively well across contexts.

In answer to RQ2.1, we find that crisis datasets from non-COVID-19 events do in fact provide value for both information categorization and (especially) prioritization.

RQ2.2 – Performance in Integrating General-Crisis Datasets with COVID-19-Specific Training data

RQ2.1 focuses on systems that participated in 2020-A, where pandemic-related content was unavailable for training. As the pandemic (and, indeed, TREC-IS) goes on, however, we are amassing more panedmic-related data and have since published a dataset of COVID-19-specific training data developed from 2020-A. Hence, a natural follow-on question is whether this new training data improves models of COVID-19 information-type and priority classification. As COVID-19 impacts regions differently, one might expect relatively little transferability of models among regional contexts; on the other hand, the pandemic's global nature may drive consistent language around topics regardless of location (modulo language differences). RQ2.2 examines this question, and in TREC-IS 2020-B, we run a new comparison of participant systems with the same zero-rule baseline described above, with results shown in Figure 6b. Unsurprisingly, participant systems perform better in 2020-B than in 2020-A, and for prioritization and information-type labeling, these new systems significant outperform the zero-rule baseline. Surprisingly, however, systems appear indistinguishable from the zero-rule baseline in measuring nDCG@100 in 2020-B, which we posit is driven by a change in pooling approaches between 2020-A and 2020-B, where we pooled fewer events more deeply in 2020-A, allowing for the discovery of more high-priority content. In 2020-B, however, the vast majority of content is low-priority content. Hence, in answer to RQ2.2, the availability of COVID-19 data does increase system-level for classifying information-types and priorities.



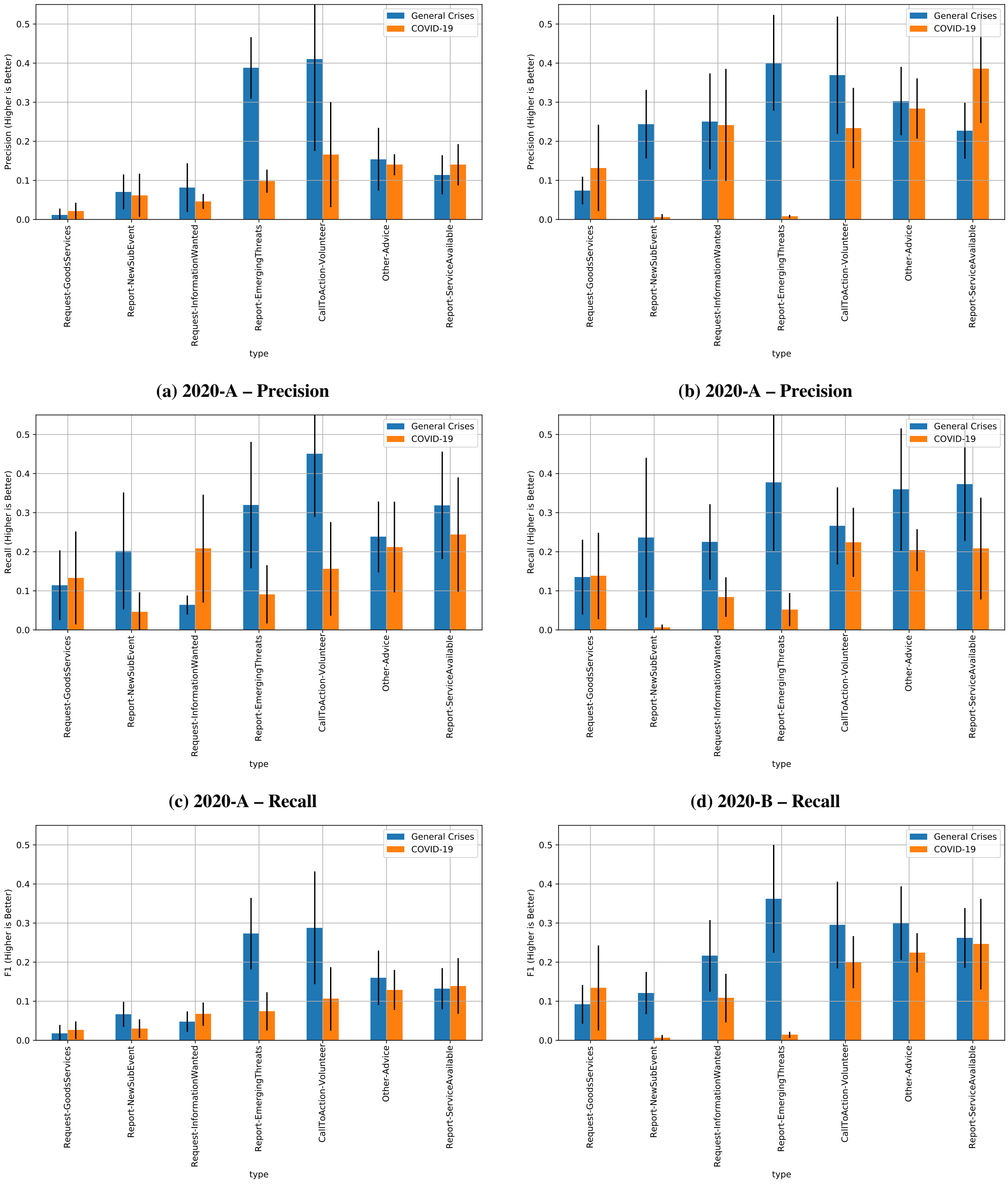
(a) 2020-A **(b) 2020-B**

Figure 6. Average System-Level Performance in 2020-A (a) and 2020-B (b), across Task 1 (General Crises) and Task 3 (COVID-19). In 2020-A, systems perform significantly worse in nDCG and Information-Type F1 in the COVID-19 task (Prioritization F1 is also lower but not significantly so at the $p = 0.05$ level).

RQ2.3 – Performance in Specific Information Types

While RQ2.1 and 2.2 show overall performance across all information types, one might expect certain information types to adapt across contexts better than others (e.g., information about service availability or advice may be sufficiently similar regardless of the underlying event). RQ2.3 investigates this possibility by breaking down participant-system scores for precision, recall, and F1 across the seven information types present in both Task 1 and Task 3, as shown in Figure 7. Of these seven types, in 2020-A, the average system appears to perform similarly for four of the types, while systems perform much worse in identifying reports of emerging threats and calls to volunteer during COVID-19. Likewise, both precision and recall are significantly lower for COVID-19 in these two types. New sub-events also exhibit significantly lower recall in COVID-19 though precision across the two contexts is similar, suggesting COVID-19 has a potentially larger space of discourse around new sub-events compared to general crises. Surprisingly, requests for information exhibit much higher recall in COVID-19 than across the general crises, suggesting a particular consistency in COVID-19 discussion. Results are largely consistent in TREC-IS's 2020-B edition as well.

In answer to RQ2.3, a core set of information types appear to be conserved in the COVID-19 context compared to the general crisis-event context, while reports of emerging threats, requests fore information, and calls for volunteers seem much more difficult to identify in the COVID-19 context. This latter result is consistent with our above analysis of information types, where we suggest emerging threats and related observations differ substantially between COVID-19 and non-COVID events. More research is necessary, however, to understand whether these differences are specific to the COVID-19 pandemic or public-health crises more generally.



(a) 2020-A – Precision **(b) 2020-A – Precision**

(c) 2020-A – Recall **(d) 2020-B – Recall**

(e) 2020-A – F1 **(f) 2020-B – F1**

Figure 7. Average System-Level Performance in 2020-A, across information types in Task 1 (General Crises) and Task 3 (COVID-19). Systems tend to perform similarly for four of the Task 3 information types, while reports of emerging threats and calls for volunteers appear to deviate significantly across contexts.

CONCLUSIONS

In this paper we have provided an overview of the new 2020 editions (2020-A and 2020-B) of TREC-IS and have compared how both social media information and participant systems may change or adapt in the time of COVID-19 and future large-scale pandemics. This paper describes how the 2020 edition of TREC-IS has addressed these dual issues by introducing a new COVID-19-specific task for evaluating generalization of existing COVID-19 annotation and system performance to this new context, applied to 11 regions across the globe. TREC-IS has also continued expanding its set of target crises, adding 29 new events and expanding the collection of event types to include explosions, fires, and general storms, making for a total of 9 event types in addition to the new COVID-19 events. Across these events, TREC-IS has 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 demonstrate 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 from six organizations that participated in both the 2020-A and 2020-B editions, submitting a total of 50 runs, suggest leveraging general crisis data in the COVID-19 context improves performance over baselines. Using these results, we provide guidance on which information types appear most consistent between general crises and COVID-19.

The Incident Streams track is slated to continue in TREC 2021 with a further two editions. All tweet streams, labels and participation details can be found at http://trecis.org.

ACKNOWLEDGMENTS

This work was supported by the 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

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hai, M., and Shah, Z. (2020). “Top concerns of tweeters during the COVID-19 pandemic: A surveillance study”. In: *Journal of Medical Internet Research* 22.4, pp. 1–9.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., and Chowell, G. (2020). “A large-scale COVID-19 twitter chatter dataset for open scientific research - An international collaboration”. In: *arXiv*. arXiv: 2004.03688.

Chen, E., Lerman, K., and Ferrara, E. (2020). “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set”. In: *arXiv* 6. arXiv: 2003.07372.

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.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). “Practical extraction of disaster-relevant information from social media”. In: *Proceedings of WWW*. ACM.

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*.

Lewis, D. D. (1995). “The TREC-4 Filtering Track”. In: *Proceedings of TREC-4*. National Institute of Standards and Technology (NIST).

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

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.

Qazi, U., Imran, M., and Ofli, F. (June 2020). “GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information”. In: *SIGSPATIAL Special* 12.1, pp. 6–15. arXiv: `2005.11177`.

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

SUPPLEMENTARY MATERIAL – PARTICIPANT RESULTS

Below, we report the raw results of TREC-IS system evaluations across all three tasks in 2020-A and 2020-B. For ease of comparison, we group by task rather than edition, but one should note numeric scores are not directly comparable across editions (because of differences in training data and evaluation methods); rather, rankings and trends within the metrics and orderings among participant systems are better axes of comparison.

System Evaluations, Task 1: All High-Level Information Type Classification

Table 3. 2020-A Task 1 Submitted runs under the v2.6 evaluation script. For each metric (Priority-centric nDCG, information-type F1/accuracy, and priority), higher is better. Bolded numbers are maximal over the full column. Results suggest the UCD-based systems outperform other participant systems in the majority of metrics.

Run	nDCG@100	Info-Type F1	Info-Type	Info-Type	Priority F1	Priority
		[Actionable]	F1 [All]	Accuracy	[Actionable]	F1 [All]
baseline.high	0.3701	0.0361	0.1111	0.0715	0.1383	0.0616
baseline.low	0.3696	0	0.0314	**0.9403**	0.0646	0.1579
elmo_all_2020_brf	0.4212	0.0933	0.1438	0.7502	0.2630	0.2575
elmo_all_2020_eec	0.4235	0.0769	0.1371	0.7284	0.2076	0.2351
elmo_all_tfidf	0.4301	0.0884	0.1380	0.7529	0.1651	0.1929
njit-sub01.text	0.4632	0.0792	0.1582	0.9025	0.1524	0.2198
njit-sub02.text+aug	0.4776	0.1466	**0.2089**	0.9056	0.0958	0.2076
sc-rf-021	0.3882	0.0088	0.0857	0.9204	0.0791	0.1791
sc-rf-022	0.3816	0.0041	0.0850	0.9217	0.0688	0.1695
sc-rf-023	0.3824	0.0085	0.0853	0.9209	0.0918	0.1782
sc-rf-024	0.3763	0.0039	0.0861	0.9217	0.0594	0.1574
UCD_CS_R1	**0.4866**	**0.1674**	0.1718	0.8978	0.2291	0.2630
UCD_CS_R2	0.4862	0.1539	0.1712	0.8944	0.2500	**0.2800**
UCD_CS_R3	0.4756	0.0935	0.1517	0.8926	**0.2642**	0.2627
UCD_CS_R4	0.4759	0.1491	0.1719	0.8966	0.2307	0.2364

Table 4. 2020-B Task 1 Submitted runs under the v2.7 evaluation script. For each metric (Priority-centric nDCG, information-type F1/accuracy, and priority), higher is better. Bolded numbers are maximal over the full column. Results suggest the UCD-based systems outperform other participant systems in the majority of metrics.

Run	nDCG@100	Info-Type F1	Info-Type	Info-Type	Priority F1	Priority
		[Actionable]	F1 [All]	Accuracy	[Actionable]	F1 [All]
BJUT-run	0.4346	0.0266	0.0581	0.8321	0.1744	0.0905
njit.s1.aug	0.4480	0.2634	0.3103	**0.8655**	0.2029	0.1518
njit.s2.cmmd.t1	0.4475	0.1879	0.2223	0.8475	0.2029	0.1518
njit.s3.img.t1	0.4222	0.1879	0.2223	0.8475	0.1959	0.1417
njit.s4.cml.t1	0.4164	0.1712	0.1465	0.8445	0.1054	0.1064
ucd-run1	0.5033	**0.3215**	**0.3810**	0.8520	0.2582	0.2009
ucd-run2	0.5022	0.3078	0.3692	0.8316	0.2582	0.2016
ucd-run3	0.5038	0.3001	0.3448	0.8653	**0.2803**	0.3046
ucd-run5	**0.5252**	0.3036	0.3444	0.8601	0.2801	**0.3126**
ufmg-sars-test	0.3634	0.0001	0.0493	0.8337	0.1285	0.1378

System Evaluations, Task 2: Selected High-Level Information Type Classification

Table 5. 2020-A Task 2 Submitted runs under the v2.6 evaluation script. For each metric (Priority-centric nDCG, information-type F1/accuracy, and priority), higher is better. Bolded numbers are maximal over the full column, and italicized numbers are maximal in Task-2 systems. Results suggest the UCD-based systems outperform other participant systems in the majority of metrics. More interestingly, Task-1 systems outperform the type-restricted Task-2 systems, suggesting little benefit comes from restricting info-type label space.

Run	nDCG@100	Info-Type F1 [All]	Info-Type Accuracy	Priority F1 [All]
2020-A Task-1 Systems				
baseline.high (task 1)	0.3694	0.0779	0.1234	0.0988
baseline.low (task 1)	0.3689	0	**0.9535**	0.0887
elmo_all_2020_brf (task 1)	0.4222	0.1113	0.8035	0.1672
elmo_all_2020_eec (task 1)	0.4235	0.0993	0.7835	0.1612
elmo_all_tfidf (task 1)	0.4297	0.1036	0.8072	0.1913
njit-sub01.text (task 1)	0.4631	0.1026	0.9414	0.1679
njit-sub02.text+aug (task 1)	0.478	**0.2079**	0.9407	0.1584
sc-rf-021 (task 1)	0.3879	0.0724	0.9459	0.1241
sc-rf-022 (task 1)	0.3818	0.068	0.9461	0.1251
sc-rf-023 (task 1)	0.3814	0.0708	0.9462	0.1178
sc-rf-024 (task 1)	0.3762	0.0671	0.9467	0.1090
UCD_CS_R1 (task 1)	**0.4864**	0.153	0.9376	**0.2612**
UCD_CS_R2 (task 1)	**0.4864**	0.1551	0.9318	**0.2612**
UCD_CS_R3 (task 1)	0.4764	0.1246	0.9298	0.2414
UCD_CS_R4 (task 1)	0.4762	0.1648	0.9336	0.2089
2020-A Task-2 Systems				
baseline.high (task 2)	0.3691	0.0779	0.0465	0.0988
baseline.low (task 2)	0.3688	0	0.8766	0.0887
njit-sub01.text (task 2)	0.4609	0.1508	0.8612	0.1682
njit-sub02.text+aug (task 2)	*0.4784*	*0.1653*	0.8634	*0.1862*
sc-rf-025 (task 2)	0.3701	0.0075	0.8753	0.0925
sc-rf-026 (task 2)	0.3773	0.0058	0.8752	0.0926
sc-rf-027 (task 2)	0.3696	0.0065	0.8753	0.0920
sc-rf-028 (task 2)	0.374	0.0071	0.8755	0.0916
UCD_CS_T2_R1 (task 2)	0.396	0.1309	*0.8781*	0.1633

Table 6. 2020-B Task 2 Submitted runs under the v2.7 evaluation script. For each metric (Priority-centric nDCG, information-type F1/accuracy, and priority), higher is better. Bolded numbers are maximal over the full column, and italicized numbers are maximal in Task-2 systems. As in 2020-A, Task-1 systems outperform the type-restricted Task-2 systems, suggesting little benefit comes from restricting info-type label space.

Run	nDCG@100	Info-Type F1 [All]	Info-Type Accuracy	Priority F1 [All]
2020-B Task-1 Systems				
BJUT-run	0.4350	0.0472	0.7977	0.1337
njit.s1.aug	0.4487	0.3480	0.8846	0.1838
njit.s2.cmmd.t1	0.4467	0.2494	0.8612	0.1838
njit.s3.img.t1	0.4215	0.2494	0.8612	0.1708
njit.s4.cml.t1	0.4176	0.1278	0.8360	0.1162
ucd-run1	0.5020	**0.4036**	0.8913	0.2320
ucd-run2	0.5027	0.3961	0.8364	0.2322
ucd-run3	0.5032	0.3689	**0.8932**	0.2867
ucd-run5	**0.5240**	0.3674	0.8845	**0.3003**
ufmg-sars-test	0.3630	0.0127	0.8419	0.1480
2020-B Task-2 Systems				
njit.s1.aug.t2	*0.4478*	*0.2548*	*0.8656*	*0.1838*
njit.s2.cmmd.t2	0.4478	*0.2548*	*0.8656*	*0.1838*
njit.s3.img.t2	0.4213	*0.2548*	*0.8656*	0.1708
njit.s4.cml.t2	0.4189	0.1713	0.8327	0.1162
ufmg-sars-test-t2	0.3637	0.0127	0.8419	0.1480

System Evaluations, Task 3: COVID-19-Specific Information-Type Classification

Table 7. 2020-A Task 3 Submitted runs under the v2.4 evaluation script. For each metric (Priority-centric nDCG, information-type F1/accuracy, and priority), higher is better. Bolded numbers are maximal over the full column. Note that not all metrics are presented here, as the restricted information types precludes separating content into actionable categories.

Run	General	Information Type		Prioritization
	NDCG	F1	Accuracy	Priority F1
baseline.high	0.2278	0.0638	0.0405	0.0438
UCD_CS_T3_R1	0.2462	0.0969	0.7995	0.1360
UCD_CS_T3_R2	0.2626	**0.1543**	0.9139	0.2051
UCD_CS_T3_R3	**0.2662**	0.1540	0.9132	0.2309
elmo_textonly_covid	0.2211	0.0523	0.8187	0.2152
njit-sub01.text	0.2289	0.1106	0.9588	**0.2765**
njit-sub02.text+aug	0.2352	0.0993	0.9576	0.2177
sc-rf-029	0.2052	0.0266	**0.9648**	0.1053
sc-rf-030	0.2156	0.0285	0.9642	0.1125
sc-rf-031	0.2077	0.0279	0.9647	0.1060
sc-rf-032	0.2139	0.0262	0.9645	0.1102

Table 8. 2020-B Task 3 Submitted runs under the v2.7 evaluation script. For each metric (Priority-centric nDCG, information-type F1/accuracy, and priority), higher is better. Bolded numbers are maximal over the full column.

Run	nDCG@100	Info-Type F1 [Actionable]	Info-Type F1 [All]	Info-Type Accuracy	Priority F1 [Actionable]	Priority F1 [All]
njit.s1.aug.t3	0.4322	**0.1629**	0.1450	0.8593	0.2551	0.1499
njit.s2.cmmd.t3	0.4329	0.1590	0.1184	0.8586	0.2551	0.1499
njit.s3.img.t3	0.3986	0.1590	0.1184	0.8586	0.2544	0.1562
njit.s4.cml.t3	0.4249	0.0210	0.0650	**0.8626**	0.1375	0.1502
ucd-run4	**0.4497**	0.1425	**0.1817**	0.8541	**0.3443**	**0.2867**