TREC Incident Streams: Finding Actionable Information on Social Media

Richard McCreadie*
University of Glasgow†
richard.mccreadie@glasgow.ac.uk

Cody Buntain
New York University‡
cody@bunta.in

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

ABSTRACT

The Text Retrieval Conference (TREC) Incident Streams track is a new initiative that aims to mature social media-based emergency response technology. This initiative advances the state of the art in this area through an evaluation challenge, which attracts researchers and developers from across the globe. The 2018 edition of the track provides a standardized evaluation methodology, an ontology of emergency-relevant social media information types, proposes a scale for information criticality, and releases a dataset containing fifteen test events and approximately 20,000 labeled tweets. Analysis of this dataset reveals a significant amount of actionable information on social media during emergencies (> 10%). While this data is valuable for emergency response efforts, analysis of the 39 state-of-the-art systems demonstrate a performance gap in identifying this data. We therefore find the current state-of-the-art is insufficient for emergency responders' requirements, particularly for rare actionable information for which there is little prior training data available.

Keywords

Emergency Management, Crisis Informatics, Real-time, Twitter, Categorization

INTRODUCTION

Mass adoption of mobile devices paired with wide-spread use of social media platforms has created new ways for the public to contact response services (Castillo 2016). Research has shown that 69% of people expect responders to answer calls for help made from these additional channels (O'Dell 2011). In response, emergency and civil protection services are increasingly looking for ways to monitor these channels, answer questions, provide advice, respond to aid requests, and report other useful information to the Incident Commander (FEMA 2013). Monitoring these social media channels is a nontrivial task, however, given the volume of information posted on these platforms and the need to categorise, cross-reference and verify the information contained therein (Castillo 2016).

The combined issues of data volume and public expectation demonstrate a clear need for computer-supported tools to assist response officers. To this end, researchers have invested significant effort into models for categorizing social media content, such as for finding affected people (Imran, Elbassuoni, et al. 2013), analysing infrastructure damage (Truelove et al. 2015) or identifying eyewitness accounts (Olteanu, Castillo, et al. 2014; Diakopoulos et al. 2012). Additionally, a number of platforms such as AIDR (Imran, Castillo, et al. 2014), CrisisTracker (Rogstadius et al. 2013), Twitcident (Abel et al. 2012) and EPIC Analyse (Barrenechea et al. 2015) have been developed.

These technologies have had limited impact on emergency management practice, primarily due to questions of data quality (Hiltz et al. 2014), insufficient trained staff (Plotnick et al. 2015), fear of commitment to social media

*corresponding author
†http://gla.ac.uk and http://dcs.gla.ac.uk/~richardm
‡http://cody.bunta.in/
§https://www.nist.gov/ and https://www.nist.gov/people/ian-soboroff

as a primary communication channel (Tapia et al. 2013), and incompatibility with organizational policy (Reuter, Heger, et al. 2013). Indeed, a recent study of ISCRAM papers concluded this research has had "a relatively small contribution to actual technology and industry" (Reuter, Backfried, et al. 2018).

While questions of training, commitment, and policy are organizational in nature, we argue the research community can solve *data quality* issues through: 1) standardizing social media analytics tasks; 2) standardizing datasets used to evaluate these tasks; and 3) focusing research efforts on advancing technical readiness and deployability. In particular, by defining a small number of standard social media analytic tasks that are well aligned with the needs of emergency response officers, it becomes feasible to construct datasets with sufficient size and robustness to confidently quantify tool accuracy (and hence resultant data quality). Then, by bringing together international researchers to work on those tasks and datasets, we can incrementally raise the quality of our solutions to a level where potential end users will be willing to use those solutions.

In this paper, we describe the new Incident Streams (TREC-IS) initiative, aimed at achieving this standardization. This initiative is part of the Text REtrieval Conference (TREC)[1] and sponsored by the Public Safety Communications Research program at the US National Institute of Standards and Technology (NIST, US). TREC-IS aims to develop test collections and evaluation methodologies for automatic and semi-automatic filtering approaches that identify and categorize information and aid-requests made on social media during crises. The end goals are: to advance the technology readiness level (TRL) of current social media crisis monitoring solutions; better support social media monitoring by emergency response officers and other stakeholders; and bring together researchers from across the globe to work on this problem by providing an annual evaluation challenge in which they can participate.

This paper's contributions include:

1. Providing the primary overview for the 2018 edition of the track, detailing track design and motivating why core design decisions were made.
2. Providing an analysis of the distribution of information posted on social media during emergency situations, contrasting general information types against actionable information based on labelling efforts undertaken during the track.
3. Giving a high-level examination of systems that participated in the 2018 edition, highlighting state-of-the-art performance for this task and what the challenges are moving forward.

RELATED DATA CHALLENGES AND INITIATIVES

TREC-IS builds upon expertise from previous evaluation initiatives and data challenges and is part of the long-running TREC conference. TREC is a combined conference and evaluation campaign that encourages research into information retrieval technologies on large test collections. Sponsored by NIST, TREC has run annually for over 25 years and consists of a set tracks, where a track is an area of focus in which particular retrieval tasks are defined. Tracks act as incubators for new research areas, in which the first run of a track often concretizes the problem, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups. TREC has been highly influential in the information retrieval domain, resulting in foundational research into search engines (Robertson et al. 1995) and information extraction from social media (Lin, Roegiest, et al. 2016). TREC-IS is a new TREC track[2] started in 2018 and is designed to unite academia and industry around research into automatically processing social media streams during emergency situations and categorizing information and aid requests made on social media for emergency service operators.

TREC-IS also builds on lessons learned from the Exploitation of Social Media for Emergency Relief and Preparedness (SMERP) workshop (Ghosh et al. 2017), which itself was inspired by the TREC Microblog track (Lin, Efron, et al. 2014). The 2017 SMERP workshop examined two challenges: 1) text retrieval for disaster-related queries on Twitter, 2) text summarization of tweets during a disaster event. For the text retrieval task, the challenge defined four information needs: 'what resources are available'; 'what resources are required'; 'what infrastructure damage, restoration and casualties are reported'; and 'what are the rescue activities of various NGOs / government organizations'. We integrated two main lessons from that data challenge: First, the information needs defined by

[1] https://trec.nist.gov/
[2] http://trecis.org

the SMERP challenge was quite broad, making it difficult to map an information need to an activity that might be performed by a response officer. We therefore developed our own information ontology, which we present in more detail later. Second, due to the small number of information needs and only the single event examined during the SMERP challenge, participating systems' generalizability was unclear. Hence, for TREC-IS, we opted for a larger pool of events across six different event types.

INCIDENT STREAMS TASK

2018 was the first year of the track and focused on how to produce a series of curated feeds of social media posts, where each feed corresponds to a particular type of information request, aid request, or report. Each post in these feeds was also given a criticality score, indicating how critical it was that a user be shown that post. This task is referred to as *classifying tweets by information type (high-level)*, which we define below.

Use-case: Enhanced Situational Awareness

As identified by FEMA (FEMA 2013), social media streams can support several valuable use-cases before, during and after emergency events, including: building situational awareness; understanding issues/needs/concerns from the public and survivors; and saving lives through rapid communication. To situate TREC-IS in this space, we first needed to define the use-case the track would satisfy, which would provide a better understanding of how social media could address problems a real user may have (Whipkey and Verity 2015). Based on discussions with officers from civil protection and resilience agencies, we opted for targeting increased *situational awareness*, a use-case that these stakeholders a) believed social media was suitable for and b) currently lacked the means to effectively tackle. Linking this use-case to existing information flow taxonomies during crisis scenarios (Reuter, Marx, et al. 2012), our task involves enhancing the 'citizen to authority' information flow, thereby improving the authority's situational awareness. Specifically, we consider our end-users to be public safety officers in the central command and control centre during an emergency. These users might be responsible for identifying actionable information that other officers can use to direct response efforts and/or answer questions from the public.

Task Formulation

We formally define the TREC-IS *Classifying Tweets by Information Type (high-level)* task as follows: Given a single social media post p published at time t (p_t), assign a high-level information type i (e.g. "call for help") from a set I ($i \in I$) to p_t, and provide a critically score c to p_t, where $0 \leq c \leq 1$. Higher criticality scores indicate that the post contains more important actionable information. Systems may use information that is publicly available prior to t when processing a post p_t.

Mapping this definition to our use-case, the information type (i) assigned to a post indicates the 'feed' to which the post belongs. The criticality score identifies posts that need to be shown to an officer immediately as an alert.

INFORMATION TYPE ONTOLOGY

We next define the information types I that systems may assign to a post. To identify our information types, we build upon past works that examine how to categorize emergency related content. In particular, a survey (Castillo 2016) of previous categorization efforts identified eight main dimensions of categorization, namely: by information provided/contained (Truelove et al. 2015); fact vs. subjective vs. emotional content (Kumar et al. 2013); by information source (Olteanu, Castillo, et al. 2014); by credibility (Castillo et al. 2013); by time (Chowdhury et al. 2013); by location (De Longueville et al. 2009); by embedded links (Shaw et al. 2013); or by environmental relevance (physical, built or social) (Mileti 1999). Given our use-case, our primary requirement is that information types represent categories of information that emergency response officers might be interested in, such as 'Reports of Road Blockages' or 'Calls for Help'. We therefore are mainly concerned with categorization by information provided/contained. However, in some scenarios it may also be valuable to consider categorization by information source (to help find first-hand reports) or credibility.

We also analysed incident management ontologies such as MOAC (Management of a Crisis)[3], response documentation (FEMA 2011) and discussed the challenges with experts. This provided a small number of additional information types, some of which were specific to an event type (e.g. 'Are the Assailants Armed?' for terrorist attacks). Aggregating information categories from prior research with categories derived from emergency management documentation and practitioners provides us with a granular view of what information is valuable during an emergency.

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

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-SignificantEventChange	The user is reporting a new occurrence that public safety officers need to respond to.	PeopleTrapped, UnexplodedBombFound
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-PastNews	The post is generic news, e.g. reporting that the event occurred	NewsHeadline
Other-ContinuingNews	The post providing/linking to continuous coverage of the event	NewsHeadline, SelfPromotion
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-Unknown	Does not fit into any other category	Unknown
Other-KnownAlready	The Responder already knows this information	KnownAlready

Table 1. Ontology High-level Information Types

Such valuable information is rare on social media, as most information shared during emergencies is of lower importance, such as news reports or shocking images. On the other hand, these other types of information may be valuable to researchers studying emergencies in the future when tackling other use-cases. Moreover, providing more resolution on what is likely 'non-relevant' for our use-case may help systems avoid miss-categorizing that content. We then expanded our initial set of information types via a bottom-up analysis of Twitter content from a small sample of events. This resulted in additional information types such as 'Sentiment Expressed', 'Press Releases' and 'Sharing Best Practices'.

In this way, we created a broad ontology of over 100 information types that are either important to emergency response officers or are commonly shared during events.[4] For the purposes of the track, asking participants to categorize posts into over 100 distinct types is unreasonable for the first year of the track. We instead grouped the individual low-level ontology entries into higher-level types. For example, we merged low level entries such as 'Hospital Operating', 'Shelter Offered' and 'Food Distribution Point' into the higher-level type 'Service Available'. In total, we defined 25 high-level types, as shown in Table 1, which form our information type categories I.

DATASET CREATION

A core component of TREC-IS is a set of evaluation datasets. A dataset is comprised of two components: a series of topics, each representing an emergency event; and the social media posts for each of those events. We summarize the creation of these below.

Twitter as a Data Source

The field of social media for emergency management is broad (Palen and Liu 2007; Hughes and Palen 2009; Middleton et al. 2014) in that a wide variety of data types and sources exist that we might desire to analyze. The majority of prior research into social media during emergencies, however, has focused on Twitter data (Reuter, Backfried, et al. 2018), likely due to ease of access. As a new initiative that aims to bring together researchers already working in the field, for the first year of TREC-IS we maintain this trend by using Twitter data, although we are interested in additional sources in future years.

Topics/Events

Our goal is to provide a sound, reproducible experimental environment, where researchers and practitioners can evaluate systems for the TREC-IS task. While most studies in this context have focused on small numbers of events, this choice is problematic; while interesting observations can be made on such datasets, we cannot show how systems generalize across events of the same type or across types. Moreover, with small datasets, illustrating statistically significant differences between systems becomes challenging. Hence, we need to create a dataset that

[4]An early visualization of the ontology can be seen at http://trecis.org/WebVOWL/#trecis

Dataset	Identifier	Event Name	Event Type	Source
Training	TRECIS-CTIT-H-Training-001	2012 Colorado wildfires	wildfire	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Training-002	2012 Costa Rica Earthquake	earthquake	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Training-003	2013 Colorado Floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Training-004	2012 Typhoon Pablo	typhoon/hurricane	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Training-005	2013 LA Airport Shooting	shooting	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Training-006	2013 West Texas Explosion	bombing	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
Test	TRECIS-CTIT-H-Test-007	2012 Guatemala earthquake	earthquake	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-008	2012 Italy earthquakes	earthquake	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-009	2012 Philipintes floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-010	2013 Alberta floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-011	2013 Australia bushfire	wildfire	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-012	2013 Boston bombings	bombing	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-013	2013 Manila floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-014	2013 Queensland floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-015	2013 Typhoon Yolanda	typhoon	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)
	TRECIS-CTIT-H-Test-016	2011 Joplin tornado	typhoon	CrisisNLP Resource #2 (Imran, Elbassuoni, et al. 2013b)
	TRECIS-CTIT-H-Test-017	2014 Chile Earthquake	earthquake	CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)
	TRECIS-CTIT-H-Test-018	2014 Typhoon Hagupit	typhoon	CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)
	TRECIS-CTIT-H-Test-019	2015 Nepal Earthquake	earthquake	CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)
	TRECIS-CTIT-H-Test-020	2018 FL School Shooting	shooting	Crawled by the Organizers via Twitter API
	TRECIS-CTIT-H-Test-021	2015 Paris attacks	bombing	Collected via the GNIP service

Table 2. TREC-IS 2018 Training and Test Events

covers a sufficient number of events for performance patterns to become identifiable. Furthermore, state-of-the-art approaches in information categorization for emergency-related content involve training supervised machine learned models, which need training examples to learn. To ensure a realistic evaluation of these models, we create two datasets, one for training such models and one for testing those models.

As a track, we are interested in emergency events, either natural or man-made. Furthermore, we anticipated that effective systems for the task may require customization for different types of event (McCreadie et al. 2018). For example, there is no point in looking for tweets with information about 'Are the Assailants Armed?' during earthquakes. For this reason, we first selected a small set of six event types to target rather than considering any event type:

Wildfire, Earthquake, Flood, Typhoon/Hurricane, Bombing, Shooting

Given these event types, we needed to decide what events to use. In a real deployment of an information-type categorization system for social media content, one would need to tackle irrelevant or off-topic content within the stream (e.g. due to imperfect data collection (Saleem et al. 2014)). For the first year of the track, we simplified the task and use only on-topic data. For this reason, we collected existing event datasets shared by other emergency management initiatives that had already been manually filtered for relevance. We then combined those datasets together to form the training and test datasets for TREC-IS 2018. In particular, we collected 19 events across the CrisisLex (Olteanu, Vieweg, et al. 2015) and CrisisNLP (Imran, Mitra, et al. 2016) initiatives. We also crawled two additional datasets ourselves for the purposes of comparison. The resultant 21 events and their sources are summarized in Table 2, the first 6 of which form the training dataset, while the latter 15 events form the test dataset.

Datasets From Social Media Streams

Data Collection: For the events listed in Table 3, we provide a stream of tweet statuses (JSON format) collected during that event for download. For the 19 events derived from either CrisisLex or CrisisNLP, we first independently crawled the tweets listed in those datasets by their unique identifier using the Twitter Streaming API . For TRECIS-CTIT-H-Test-020 (a contemporary event from 2018), we also used the Twitter Streaming API to crawl the event while as it occurred. Meanwhile, for TRECIS-CTIT-H-Test-021 we retrospectively collected unique tweets using the GNIP service .

Filtering: To reduce the barrier to entry for the first year, we desired a relevant-only set of tweets for systems to process. For the CrisisLex and CrisisNLP datasets, we pre-filtered using the human annotation labels provided, removing tweets marked with labels such as 'Not Related' or 'Not Relevant'. For the two events collected by the organizers (TRECIS-CTIT-H-Test-020 and TRECIS-CTIT-H-Test-021), we lacked human-generated relevance labels and instead performed KMeans clustering for each event. Each event was clustered into 10k clusters based on textual similarity, and we selected one tweet per cluster (the centroid tweet) to be subsequently labelled. For these two events, due to their size, we also performed additional manual keyword filtering (using terms extracted from an associated Wikipedia page for each event) to reduce it to a diverse and relevant set. We also removed non-English tweets for all events. The statistics of the tweet streams for each event are shown in Table 3.

Dataset	Identifier	Event Name	# Posts Pre-Filtering	# Posts Post-Filtering
Training	TRECIS-CTIT-H-Training-001	2012 Colorado wildfires	3,275	744
	TRECIS-CTIT-H-Training-002	2012 Costa Rica Earthquake	1,880	288
	TRECIS-CTIT-H-Training-003	2013 Colorado Floods	1404	777
	TRECIS-CTIT-H-Training-004	2012 Typhoon Pablo	1,413	649
	TRECIS-CTIT-H-Training-005	2013 LA Airport Shooting	2031	683
	TRECIS-CTIT-H-Training-006	2013 West Texas Explosion	9,877	630
Test	TRECIS-CTIT-H-Test-007	2012 Guatemala earthquake	2,518	178
	TRECIS-CTIT-H-Test-008	2012 Italy earthquakes	5,740	118
	TRECIS-CTIT-H-Test-009	2012 Philipinnes floods	1,658	480
	TRECIS-CTIT-H-Test-010	2013 Alberta floods	4,559	739
	TRECIS-CTIT-H-Test-011	2013 Australia bushfire	1,507	710
	TRECIS-CTIT-H-Test-012	2013 Boston bombings	98,884	543
	TRECIS-CTIT-H-Test-013	2013 Manila floods	1,289	443
	TRECIS-CTIT-H-Test-014	2013 Queensland floods	832	744
	TRECIS-CTIT-H-Test-015	2013 Typhoon Yolanda	27,755	629
	TRECIS-CTIT-H-Test-016	2011 Joplin tornado	206,764	152
	TRECIS-CTIT-H-Test-017	2014 Chile Earthquake	368,630	321
	TRECIS-CTIT-H-Test-018	2014 Typhoon Hagupit	625,976	6,696
	TRECIS-CTIT-H-Test-019	2015 Nepal Earthquake	4,223,937	7,301
	TRECIS-CTIT-H-Test-020	2018 FL School Shooting	12,082,953	1,118
	TRECIS-CTIT-H-Test-021	2015 Paris attacks	1,671,440	2,066

Table 3. TREC-IS 2018 Stream Statistics

LABELING TWEETS WITH INFORMATION TYPES AND PRIORITIES

To evaluate systems' performance, we required ground truth information-type and criticality labels for training and test sets. A team of TREC assessors created this ground truth by reviewing each event and hand-labeling tweets in that event, one tweet per assessor, using the assessment interface shown in Figure 1. These TREC assessors were experienced information generalists with a strong analytical background. For labeling tasks, NIST trains assessors using sample materials and works closely with them to ensure label quality and consistency.

Tweets from the 15 test events were labelled by TREC Assessors during August 2018. TREC assessors were contracted for this assessment and brought into a controlled lab environment for the duration of the job. Six TREC assessors contributed to the labelling of the events. Assessors were trained in following the ontology and identifying critical tweets using the 2012 Colorado, US wildfires event. As a group, the assessment team labeled tweets in the training event and discussed how to make decisions for confusing or ambiguous cases. Furthermore, each event type was accompanied by an information sheet defining the critical types of information expected during an event of that type. This information sheet was also provided to TREC-IS participants in the form of user profiles for each event type.

Two events were large (2014 Typhoon Hagupit and 2015 Nepal Earthquake) and were divided among multiple assessors (the partitions are denoted *DATASET*S1, *DATASET*S2, etc.). Labelling statistics divided by assessor and event are provided in Table 4. Note that as the assessment interface performed automatic textual de-duplication of tweets, the number assessed is lower than the number of tweets in each dataset.

One interesting factor to consider about this labelling task is that it involves markedly more effort by the assessors than a classical tweet categorization task, as can be seen by the high average time spent assessing each tweet across assessors (see column 4 in Table 4). This effort results from the assessors not only reading the text of each tweet but also examining linked content (e.g. news articles) and any attached videos. The level of effort required in labeling prevented a balanced study of agreement. Across the six TREC assessors, the track's test dataset contains 19,784 labeled tweets with 43,514 category labels.

METRICS

Participating organizations could submit up to four *runs*. Each submitted run gives a predicted information type and criticality score to each tweet. Participant runs are evaluated against two criteria: 1) how well did they identify the information type of the tweet (Information Type Categorization); and 2) how accurately did they estimate the criticality of the information within each tweet (Information Criticality).

Information Type Categorization

The first evaluation axis for a TREC-IS system is how effectively it can categorize tweets into the 25 high-level information types in the TREC-IS Ontology. Notably, participant systems for 2018 were tasked with assigning one

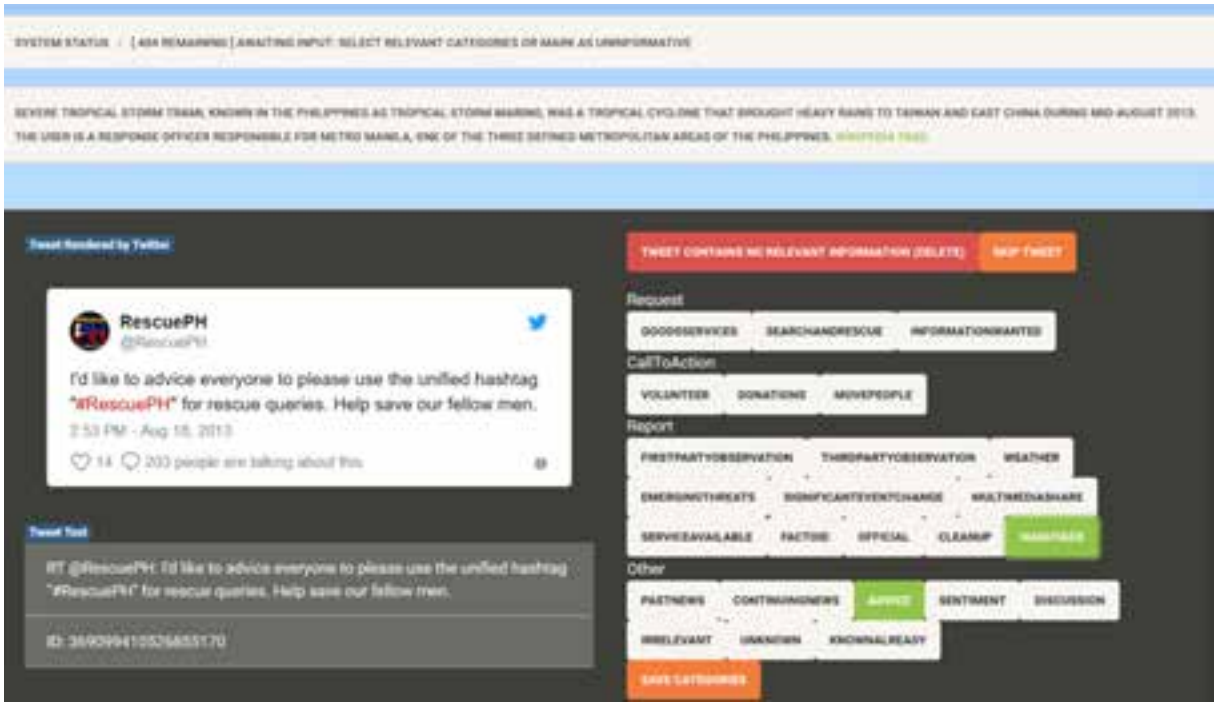


Figure 1. TREC-IS Task Assessment Interface

information type per-tweet (the most representative one). When creating the ground truth, however, TREC assessors were allowed to select multiple information types. For this reason, we evaluate information type categorization in two ways: multi-type and any-type.[5]

- **Multi-Type**: Under multi-type evaluation, we calculate categorization performance per information type in a 1-vs-All manner. A system is considered to have categorized a tweet correctly if both the system and human assessor selected that category. Note that since a system can only select one category per tweet but the assessors can select multiple categories per tweet, system performance across information types cannot be perfect under this metric. For instance, if a tweet had three information types assigned to it by the human assessor, then a system can only receive a maximum of a 1/3 score for that tweet. Multi-type evaluation is primarily useful for contrasting performance between information types or between events. We evaluate multi-type classification performance under four metrics: Precision/Recall/F1 (Positive Class) and Accuracy (Overall).
- **Any-Type**: Under any-type evaluation, a system receives a full score for a tweet if the system assigned any of the categories that the human assessor selected for that tweet. This metric provides absolute classification performance of a TREC-IS system but is more heavily influenced by larger events and more common information types (which may be less valuable for our end-users). For Any-Type, we report Precision/Recall/F1 (micro averaged) and Accuracy (micro averaged).

Information Criticality

The second axis of evaluation for a TREC-IS system is the extent to which it can identify key information that emergency response officer needs to see. This evaluation is operationalized by comparing a tweet's information priority score provided by a system to the criticality label provided by the human assessor.

To enable this comparison, two transformations were first performed. First, four information criticality labels used by the human assessors were mapped into numerical scores as per the guidelines. In this case, low=0.25, medium=0.5, high=0.75 and critical=1.0. Second, as some participant systems did not provide scores within a 0-1 range, all system scores were subject to a max-min normalization, with a minimum score cap of 0.25. Criticality estimation performance was then measured in terms of the *Mean Squared Error* between the human assigned score and the normalized system criticality score (lower is better).

[5]See http://trecis.org/2018/Evaluation.html for details.

Assessor ID	Event Name	Dataset ID	Time Per Tweet (seconds)	# Tweets Labelled
assr1	2012 Guatemala earthquake	guatemalaEarthquake2012	81	137
	2013 Boston bombings	bostonBombings2013	65	488
	2018 FL School Shooting	flSchoolShooting2018	59	1032
assr2	2014 Chile Earthquake	chileEarthquake2014	60	286
	2011 Joplin tornado	joplinTornado2011	41	94
	2013 Typhoon Yolanda	typhoonYolanda2013	45	538
	2013 Queensland floods	queenslandFloods2013	35	689
	2015 Nepal Earthquake	nepalEarthquake2015S3	25	1377
assr3	2013 Australia bushfire	australiaBushfire2013	30	664
	2012 Philipinnes floods	philipinnesFloods2012	24	434
	2013 Alberta floods	albertaFloods2013	24	717
	2015 Nepal Earthquake	nepalEarthquake2015	27	128
	2015 Nepal Earthquake	nepalEarthquake2015S2	23	1393
	2014 Typhoon Hagupit	typhoonHagupit2014S2	22	1346
assr4	2013 Manila floods	manilaFloods2013	23	404
	2015 Paris attacks	parisAttacks2015	16	2037
	2012 Italy earthquakes	italyEarthquakes2012	18	100
assr5	2014 Typhoon Hagupit	typhoonHagupit2014	48	1381
	2014 Typhoon Hagupit	typhoonHagupit2014S1	39	1302
assr6	2015 Nepal Earthquake	nepalEarthquake2015	30	1881
	2015 Nepal Earthquake	nepalEarthquake2015S1	23	1389
	2015 Nepal Earthquake	nepalEarthquake2015S4	19	1392

Table 4. Event Labelling Statistics for the Test Events (TRECIS-CTIT-H 2018 Test).

ANALYSING ASSESSOR AND PARTICIPANT RESULTS

After collecting information-type and criticality labels from assessors and track participants, we investigate two groups of research questions regarding information posted on social media and TREC-IS participants' performance in the track. First, questions that relate specifically to the 19,784 unique tweets labeled by our assessors:

- **RQ1.1** How common are different types of information posted to Twitter during emergencies?
- **RQ1.2** How prevalent is critical/actionable information on Twitter?

Second, by analysing the performance of the 39 systems submitted to the 2018 track we investigate the following:

- **RQ2.1** How well can systems identify different information types?
- **RQ2.2** Are specific information types generally more difficult to identify?
- **RQ2.3** How well do systems perform at information criticality classification?

RQ1.1: How Prevalent is Information on Twitter during Emergencies?

The first question concerns how common different information types are on Twitter based on the TREC-IS test dataset (15 events). This statistic is valuable as it provides insights on how much actionable information one may find on social media.[6] To contrast general and actionable information, for our analysis in this section, we consider the following types of information as actionable: Requests for Goods/Services, Requests for Search and Rescue, Calls to Action for Moving People (Evaluations), Reports of Emerging Threats, Reports of Significant Event Changes and Reports of Services becoming available. This definition of actionability differs from works like (Zade et al. 2018) as it is information type-based. We assume an "actionable" post that generates an immediate alert will be useful to an individual responsible for that type (optionally considering criticality as well). Table 5 reports the number of tweets labeled with each of the 25 information types in the test dataset. Our expectation is that distributions of information types will differ across event types, so we divide these counts across our six event types. The percentage reported after each count is the proportion of tweets for that information type and event type.

[6]Note this is subject to both the sampling strategies employed by the CrisisLex and CrisisNLP original sources and the subsequent filtering we applied.

High-Level Information Type	Wildfire	Earthquake	Flood	Typhoon/Hurricane	Bombing	Shooting
Request-GoodsServices	0 (0%)	72 (0.49%)	48 (0.87%)	5 (0.04%)	1 (0.02%)	0 (0%)
Request-SearchAndRescue	0 (0%)	177 (1.2%)	110 (2%)	6 (0.05%)	5 (0.09%)	0 (0%)
Request-InformationWanted	1 (0.05%)	99 (0.67%)	25 (0.45%)	42 (0.32%)	6 (0.11%)	1 (0.03%)
CallToAction-Volunteer	1 (0.05%)	25 (0.17%)	49 (0.89%)	33 (0.26%)	4 (0.08%)	4 (0.12%)
CallToAction-Donations	15 (0.81%)	424 (2.88%)	173 (3.14%)	160 (1.24%)	9 (0.17%)	15 (0.47%)
CallToAction-MovePeople	1 (0.05%)	5 (0.03%)	11 (0.2%)	8 (0.06%)	0 (0%)	0 (0%)
Report-FirstPartyObservation	102 (5.51%)	1432 (9.74%)	743 (13.5%)	1330 (10.29%)	75 (1.41%)	2 (0.06%)
Report-ThirdPartyObservation	529 (28.58%)	481 (3.27%)	560 (10.17%)	1992 (15.41%)	618 (11.64%)	9 (0.28%)
Report-Weather	7 (0.38%)	31 (0.21%)	66 (1.2%)	1232 (9.53%)	0 (0%)	0 (0%)
Report-EmergingThreats	22 (1.19%)	332 (2.26%)	110 (2%)	113 (0.87%)	86 (1.62%)	33 (1.02%)
Report-SignificantEventChange	25 (1.35%)	85 (0.58%)	72 (1.31%)	74 (0.57%)	161 (3.03%)	0 (0%)
Report-MultimediaShare	99 (5.35%)	1287 (8.76%)	413 (7.5%)	1237 (9.57%)	367 (6.91%)	550 (17.06%)
Report-ServiceAvailable	29 (1.57%)	794 (5.4%)	61 (1.11%)	152 (1.18%)	76 (1.43%)	0 (0%)
Report-Factoid	227 (12.26%)	922 (6.27%)	126 (2.29%)	807 (6.24%)	212 (3.99%)	68 (2.11%)
Report-Official	73 (3.94%)	85 (0.58%)	151 (2.74%)	91 (0.7%)	7 (0.13%)	0 (0%)
Report-CleanUp	0 (0%)	14 (0.1%)	23 (0.42%)	21 (0.16%)	0 (0%)	1 (0.03%)
Report-Hashtags	0 (0%)	631 (4.29%)	488 (8.86%)	1205 (9.32%)	759 (14.29%)	174 (5.4%)
Other-PastNews	2 (0.11%)	83 (0.56%)	41 (0.74%)	23 (0.18%)	255 (4.8%)	946 (29.34%)
Other-ContinuingNews	435 (23.5%)	1440 (9.8%)	840 (15.26%)	1790 (13.85%)	424 (7.98%)	106 (3.29%)
Other-Advice	65 (3.51%)	396 (2.69%)	459 (8.34%)	242 (1.87%)	32 (0.6%)	3 (0.09%)
Other-Sentiment	177 (9.56%)	3326 (22.63%)	501 (9.1%)	1584 (12.26%)	1043 (19.64%)	234 (7.26%)
Other-Discussion	1 (0.05%)	941 (6.4%)	24 (0.44%)	302 (2.34%)	213 (4.01%)	709 (21.99%)
Other-Irrelevant	39 (2.11%)	1192 (8.11%)	379 (6.88%)	341 (2.64%)	652 (12.28%)	86 (2.67%)
Other-Unknown	0 (0%)	53 (0.36%)	12 (0.22%)	9 (0.07%)	4 (0.08%)	0 (0%)
Other-KnownAlready	1 (0.05%)	372 (2.53%)	20 (0.36%)	126 (0.97%)	301 (5.67%)	283 (8.78%)

Table 5. Prevalence of High-level Information Types across Event Types. Information Types highlighted in bold are those we consider 'actionable'.

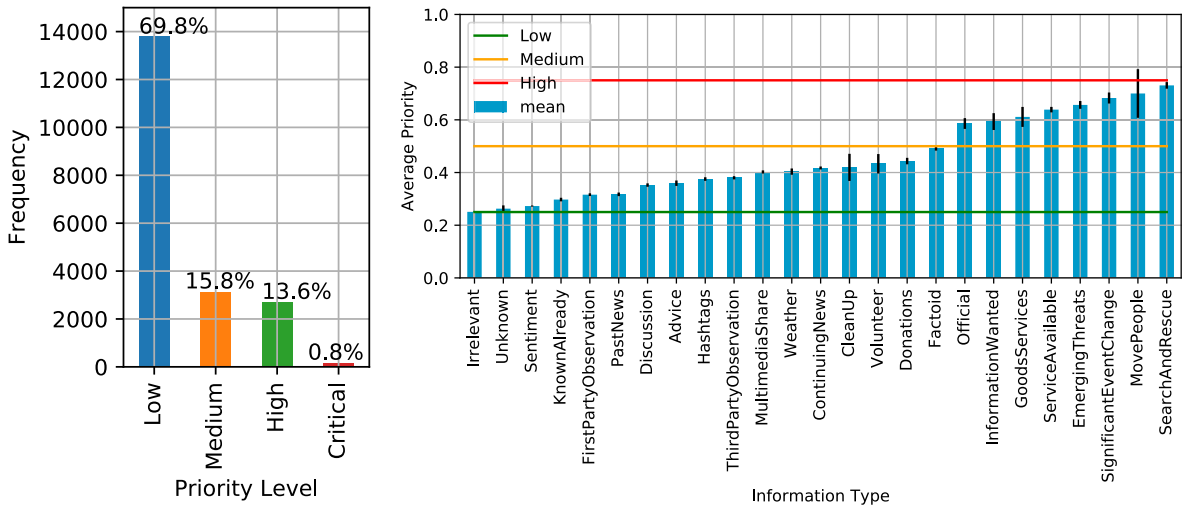
Table 5 shows the following interesting trends: For Wildfire events, the main types of information reported in Twitter are Third Party Observations (29%), Factoids (12%, typically the amount of area destroyed) and News reports (24%). Restricting to actionable information types totals around 4%, which is almost all reports. For Earthquake events, we see a more even spread of information types. The most prevalent information type is Sentiment (23%), with First Party Observations (10%) and Multimedia Sharing (9%) also being popular. In terms of actionable information, this event type has the largest proportion of reports of Services becoming Available again (1.2%) and has the highest proportion of actionable information overall (10%). For Flooding events, First Party Observations (14%), Third Party Observations (10%) and News Reports (16%) are the most common. Meanwhile, this event type has the highest proportion of requests for Search and Rescue (2%), with around 7% of the tweets belonging to our 'actionable' information types. During Typhoon/Hurricane events, we again see a high prevalence of First Party Observations (10%) and Third Party Observations (14%), with Multimedia Sharing (10%), News Reporting (14%) and expressing Sentiment (12%) also being popular. The amount of actionable information here appears to be lower, with very few requests for aid, and only around 2.5% of posts containing potentially valuable reports. Moving to the man-made emergency event types, for Bombings, Third Party Observations (12%) and Sentiment (20%) are the most common information types. Around 6% of tweets contain potentially actionable information, which is almost exclusively reports rather than requests for aid. Finally, for the shooting event (only one exists in the test dataset), we see a markedly different distribution of information types: The most common are Multimedia Sharing (17%), Past News (29%, analysing the shooter's history) and related discussions (22%, about gun control in the U.S.). Almost no actionable information exists on social media during this event (1%).

In conclusion, we can see marked differences between the information types that are prevalent during different types of events. However, in terms of potentially actionable information from the perspective of a response officer, there appears to be valuable information here that can be accessed, with up-to 10% of the tweets containing actionable information according to our definition (although some of that information may be redundant).

RQ1.2: How Much Information is Critical?

Having examined the information type distribution within our test dataset from the perspective of types that we might consider actionable, we next contrast this result against what our assessors judged to be critical information. Recall that for each tweet, our assessors assigned a criticality label (Critical, High, Medium or Low) to each tweet, which we then mapped into a criticality score (Critical=1.0, High=0.75, Medium=0.5 and Low=0.25). In this section we examine the distribution of criticality among the tweets in the test dataset.

We first averaged the criticality scores for all 15 test events, resulting in an average tweet criticality of $\mu = 0.3632$, with a standard deviation $\sigma = 0.1880$. This result suggests, as we might expect, the majority of crisis-related



(a) Criticality Distribution **(b) Average Criticality by Information Type (95% Confidence)**

Figure 2. Criticality Distributions Across and Within Information Types

tweets are between low and medium criticality. Figure 2a shows number of tweets assessed for each criticality level, while 2b reports criticality by information-type. Figure 2a shows the majority of tweets are of Low and Medium criticality (69.8% and 15.8%, respectively), i.e. a response officer would not need to see these tweets. However, we do see a relatively large proportion (14.4%) of tweets our assessors judged as important for a response officer to see promptly. Examples include pictures and videos taken during the emergency or status updates from the ground. For instance, during Typhoon Hagupit, the tweet *"Borongan, Eastern Samar's communication lines and power & water lines are still down. /via @EdwinSevidal #RubyPH"* was judged as having a high priority. Furthermore, while the proportion is small (0.8%), we also see tweets that are critical for an emergency response officer to see (and act on) immediately. These are often timely updates on potentially life-threatening impacts. For example, during Typhoon Hagupit again, the tweet *"Typhoon Hagupit hit Eastern Samar. ACTalliance local partners see transitionsal shelters blown away. Need to assess damage and respond asap"* was judged as critical. This result shows important information gets transmitted via social media during crises, supporting the idea that emergency response agencies should consider it a primary communication channel.

RQ2.1: Can Systems Effectively Identify Information Types?

Having analysed information type labels, we next examine whether current systems can automatically identify these types. Since TREC-IS's overall goal is to advance the TRL of current (automated) social media crisis monitoring solutions, rather than to study how humans annotate social media data, remaining sections focus on systems participating in TREC-IS 2018. Eleven research groups from eight countries participated in this track in 2018, submitting a total of 39 runs (each representing a different implementation).

The first question we examine is to what extent these systems able to automatically identify the 25 information categories. Table 6 reports system performance, ranked by the Any-Type, Micro Avg F1 metric (best score under each metric is highlighted in bold). Our analysis focuses on the Multi-Type metrics, as they provide more insights into categorization performance.

Multi-Type metrics have two sub-types: *Positive Class*, or how effectively a system classifies each information type in a one-versus-all manner; and *Overall*, that calculates (overall) classification accuracy. To perform well under the Multi-Type Positive Class metrics, a system needs to be effective at identifying tweets that belong to each information type (we only consider true positives and true negatives here). Moreover, it needs to perform well for the majority of the information types, as we uniformly average across information types. In contrast, Overall (Accuracy) also considers when a system correctly did not assign a category.[7] Table 6 shows that under the Multi-Type Positive Class metrics, performance is low across all 39 systems, with F1 scores ranging from 4% (lowest) to 15.7% (highest). One should recall this estimate underestimates the true performance of these systems, as systems only assigned a single type per tweet, while the assessors could assign multiple types. Even

[7]Most tweets belong to one or two categories, meaning that true negatives are vastly more common than true positives, hence Overall scores will be much higher than Positive Class scores.

Run	Information Type Categorization								Information Criticality
	Multi-Type, Macro Avg.				Any-Type, Micro Avg.				
	Positive Class			Overall					
	Precision	Recall	F1	Accuracy	Precision	Recall	F1 (Target Metric)	Accuracy	Mean Squared Error (Lower is better)
cbnuS2	**0.2666**	0.1122	0.1262	**0.9059**	0.4559	0.7780	**0.5749**	**0.4213**	0.0943
KDEIS4_DM	0.1483	0.0708	0.0734	0.9035	0.3914	0.9856	0.5603	0.3908	0.0841
umdhcilfasttext	0.1827	0.0962	0.1117	0.9044	0.4534	0.7260	0.5582	0.4022	0.0943
cbnuS1	0.2187	0.1164	0.1254	0.9048	0.4472	0.7402	0.5575	0.4064	0.0943
NHK_run2	0.2104	0.1005	0.1187	0.9042	0.4483	0.7143	0.5509	0.3997	0.0623
DLR_Simple_CNN	0.2446	0.1676	0.1595	0.9036	0.4220	0.7749	0.5464	0.3919	0.1640
NHK_run3	0.2302	0.0952	0.1140	0.9042	0.4701	0.6405	0.5422	0.3996	0.0611
cbnuC2	0.2313	0.1025	0.1164	0.9034	**0.5320**	0.5335	0.5327	0.3893	0.0943
DLR_Baseline	0.1502	0.0710	0.0743	0.9033	0.4813	0.5778	0.5251	0.3881	0.1493
SINAI_run1	0.1729	0.0726	0.0787	0.9030	0.5065	0.5303	0.5181	0.3844	0.0902
SINAI_run3	0.1825	0.0713	0.0767	0.9023	0.5019	0.5130	0.5074	0.3762	0.0894
SINAI_run4	0.1782	0.0754	0.0825	0.9025	0.5297	0.4849	0.5063	0.3786	0.0926
NHK_run1	0.1940	0.1027	0.1196	0.9006	0.4063	0.6414	0.4974	0.3551	**0.0602**
cbnuC1	0.2391	0.1022	0.1152	0.9006	0.5056	0.4797	0.4923	0.3545	0.0943
myrun-10	0.1446	0.0647	0.0673	0.9005	0.4415	0.5502	0.4899	0.3533	0.0944
myrun-11	0.1625	0.0645	0.0672	0.9004	0.4401	0.5502	0.4890	0.3525	0.0941
DLR_Fusion	0.2496	**0.1664**	0.1562	0.8996	0.3811	0.6821	0.4890	0.3425	0.1594
umdhcilbaseline	0.1866	0.0906	0.1096	0.8993	0.3978	0.6251	0.4862	0.3385	0.0943
DLR_Augmented	0.2580	0.1657	**0.1571**	0.8996	0.3965	0.6165	0.4826	0.3419	0.1599
uogTr_R3_asp	0.2159	0.0945	0.1050	0.8973	0.3136	**1.000**	0.4775	0.3136	0.0916
NHK_run4	0.2073	0.0784	0.0878	0.9001	0.4729	0.4575	0.4651	0.3482	0.0663
SINAI_run2	0.1307	0.0697	0.0769	0.8988	0.4189	0.4980	0.4551	0.3317	0.0859
myrun-21	0.1496	0.0712	0.0873	0.8964	0.3514	0.6064	0.4449	0.3018	0.0958
myrun-2	0.1401	0.0674	0.0819	0.8952	0.3339	0.6004	0.4291	0.2876	0.1004
uogTr_R1_asp	0.2295	0.0823	0.0792	0.8963	0.3874	0.4578	0.4196	0.3011	0.0933
KDEIS3_ACSBLSTM	0.1209	0.0577	0.0483	0.8933	0.2631	0.9788	0.4147	0.2635	0.0842
IITBHU1	0.2522	0.1299	0.1369	0.8947	0.4264	0.3926	0.4088	0.2811	0.1244
IITBHU12	0.2522	0.1299	0.1369	0.8947	0.4264	0.3926	0.4088	0.2811	0.1244
KDEIS1_CLSTM	0.1388	0.0607	0.0620	0.8929	0.2575	0.9783	0.4078	0.2580	0.0842
umdhcilfts	0.0993	0.0499	0.0420	0.8932	0.2770	0.7566	0.4055	0.2621	0.0943
umdhcilspread	0.1289	0.0597	0.0663	0.8931	0.2764	0.7124	0.3982	0.2608	0.0943
uogTr_R2_asp	0.2313	0.0970	0.1109	0.8951	0.3623	0.4355	0.3956	0.2861	0.0931
UPB_DICE2	0.0875	0.0406	0.0421	0.8906	0.2348	0.8682	0.3696	0.2299	0.0925
UPB_DICE1	0.1827	0.0721	0.0635	0.8905	0.2713	0.5696	0.3676	0.2286	0.0916
KDEIS2_ACBLSTM	0.1512	0.0689	0.0703	0.8890	0.2089	0.9734	0.3440	0.2098	0.0842
UPB_DICE4	0.0759	0.0403	0.0406	0.8885	0.2155	0.6208	0.3200	0.2031	0.0876
myrun2_fixed	0.2076	0.0829	0.0870	0.8881	0.2076	0.5957	0.3079	0.1986	0.0793
myrun1_fixed	0.2077	0.0826	0.0809	0.8866	0.1810	0.8881	0.3007	0.1790	0.0728
UPB_DICE3	0.0926	0.0403	0.0419	0.8867	0.1949	0.4156	0.2653	0.1804	0.0942
TREC Median	0.1827	0.0784	0.0825	0.8993	0.3978	0.6165	0.4775	0.3385	0.0933

Table 6. TREC-IS 2018 Participant Run Performances.

accounting for this difference, the low performance of these systems indicates they are not yet sufficiently effective at discriminating information types. This result motivates the need for a concerted effort to bring performance up to a level that is deployable by end-users. While performance seems better when examining Overall accuracy, this increase is a reflection of the high frequency of true negatives. Potential users primarily care about getting the true positives correct (e.g., having systems that can correctly identify that urgent request for search and rescue). Hence, in summary, participant systems are getting between 85-90% of categorization decisions correct under the one-versus-all scheme, but the remaining 10-15% are often the cases that our end-users actually care about. It is thus fair to conclude that this technology is currently insufficient for deployment in practice.

RQ2.2: What Information Types are Difficult to Identify?

Insights on where current state-of-the-art systems tend to succeed or fail is important for focusing future research. While a deep analysis of individual systems is beyond this paper's scope, we perform a meta-analysis of systems by information type. In particular, we identify the best performing system for each information type and plot those best per-information type performances. This meta-analysis represents the performance distribution if we could combine all of the best-performing aspects of participant systems into a single '(beyond) state-of-the-art system'. This is informative, as we can observe gaps where participating systems performed poorly, identifying areas where the most effort is needed.

Figure 3 shows the best participant system's performance when ranked by Multi-Type, Macro Avg. Positive Class per information type under the Positive Class metrics. Across information types, the average F1 score is 0.2953, with precision 0.4151 and recall 0.3167, representing a large and statistically significant improvement over the

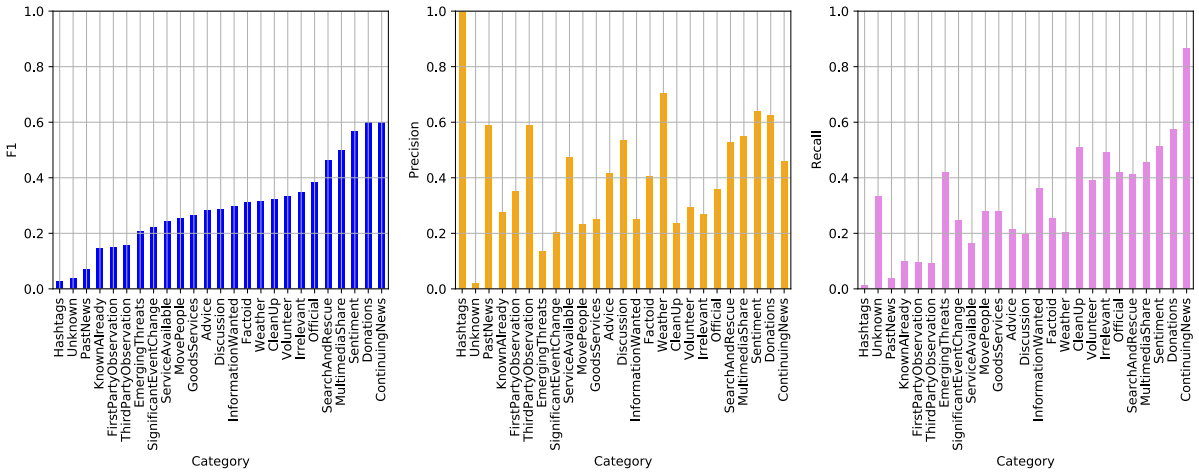


Figure 3. F1, Precision, and Recall for Best Performing Systems by Information Type

0.1571 F1 previously observed for the best single system in Table 6. While this gain is expected, the magnitude of this difference suggests high variance exists in performance among systems for each information type.

Furthermore, from Figure 3, we see the best performing systems are quite effective at identifying continuing news, donations, and expressions of sentiment. Alternatively, of the six actionable information types, all but the search-and-rescue type are in the bottom half of this ranking, suggesting even the best participant systems have difficulty with identifying this important information. We also note a weak correlation exists between the per-type F1 scores and the amount of training data made available for this edition of the track (the training dataset), which accounts for 29% of variation in scores. Given the rarity of critical information in social media data, this result reinforces the necessity of a large-scale and standardized dataset of emergency-related content. Indeed, this gap is a motivation for a continuing initiative that produces datasets with increasing size.

RQ2.3: Are Systems able to Accurately Estimate Information Criticality?

Having examined system performance for information type categorization, we next examine efficacy of criticality assessment. We measure this performance by comparing the (normalized) criticality score provided by a system for a tweet against the criticality label assigned by the assessor for that tweet. The smaller this difference (the lower the error) the better. The final column in Table 6 reports this estimation error for each system, while Figure 4 reports the average criticality scores for each system.

We make the following observations: First, Table 6 shows a more positive picture than when examining information type categorization performance, with prediction errors ranging from 0.06 to 0.16, a fairly low level of error.[8] Second, when examining systems' criticality scores, we see a markedly higher average ($\mu = 0.6067$, $\sigma = 0.2193$) than the assessors criticality scores ($\mu = 0.3632$). This result suggests participating systems tend to over-estimate the priority of messages rather than under-estimate criticality. This over-estimation is especially apparent in eight systems ("cbnu" and "umdhcil") that give all messages a score of "Critical". Within groups variations are also low, with many groups' systems returning the same average criticality score. Removing these eight systems yields an average criticality score of $\mu = 0.5053$, which is closer to the assessors' scores but still overestimating importance. We also note the lack of correlation between systems performing well on information type categorization and those performing well on information criticality estimation.

CONCLUSIONS

This paper describes the 2018 edition of the TREC-IS initiative. TREC-IS 2018 developed test collections and evaluation methodologies for automatic and semi-automatic filtering approaches that identify and categorize information and aid-requests made on social media during crisis situations. It also provides an evaluation challenge in which researchers/developers can participate. This paper also provides an overview of TREC-IS's evaluation methodology, an ontology of emergency information types, and summarizes the creation of a large dataset of 19,784

[8] Although we note the macro averaging across information types being performed here may be masking higher error rates on the rarer information criticality levels (Critical and High).

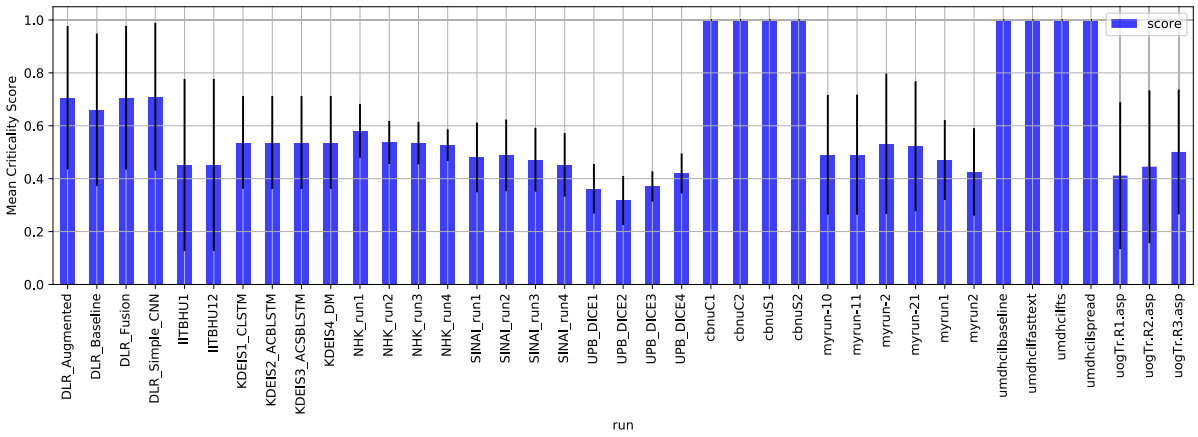


Figure 4. Mean Critical Scores Per Run (+/- one standard deviation)

hand-labeled social media messages from Twitter, in which assessors identified actionable and critical information for each event.

In addition, we examine this labelled Twitter data, showing the amount of actionable information for emergency response officers on Twitter is significant (up-to 10% post-filtering), but varies greatly with event type. This result is supported by information criticality labels produced by our assessors, which indicates that ~15% of tweets were judged as being of 'High' or 'Critical' importance to response officers.

Finally, as an evaluation challenge, eleven teams participated in TREC-IS 2018. This paper summarizes performances of these systems and implications for the maturity of the state-of-the-art. Given the eleven participating teams and 39 submitted systems, we show the current state-of-the-art is not sufficiently effective to be considered deployable by end users. While participants are relatively effective at identifying common information types such as news reports and sentiment, identifying actionable information types like search and rescue requests is still challenging. Alternatively, systems are more accurate at estimating information criticality, although they tend to over-estimate.

TREC-IS has been approved to run at TREC 2019, yielding an opportunity to develop larger, more robust, and standardized assessment datasets spanning multiple years, thereby reducing this reliance on training data and allowing technologists to better serve responders' needs. Up-to-date information and all of the data for the track are available from the TREC-IS track website:

http://trecis.org

ACKNOWLEDGMENTS

This work was supported by the Incident Streams Project sponsored by the Public Safety Communication Research Division at the National Institute of Standards and Technology (NIST, US).

REFERENCES

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., and Tao, K. (2012). "Twitcident: fighting fire with information from social web streams". In: *Proceedings of WWW*. ACM.

Barrenechea, M., Anderson, K. M., Aydin, A. A., Hakeem, M., and Jambi, S. (2015). "Getting the query right: User interface design of analysis platforms for crisis research". In: *Proceedings of ICWE*. Springer.

Castillo, C. (2016). *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

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

Chowdhury, S. R., Imran, M., Asghar, M. R., Amer-Yahia, S., and Castillo, C. (2013). "Tweet4act: Using incident-specific profiles for classifying crisis-related messages." In: *Proceedings of ISCRAM*. Citeseer.

De Longueville, B., Smith, R. S., and Luraschi, G. (2009). "Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires". In: *Proceedings of SIGSPATIAL*. ACM.

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.

FEMA (2011). *FEMA National Incident Support Manual*. Tech. rep. Federal Emergency Management Agency, US.

FEMA (2013). *IS-42: Social Media in Emergency Management*. https://training.fema.gov/is/courseoverview.aspx?code=IS-42.

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G. J., and Moens, M.-F. (2017). "ECIR 2017 Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017)". In: *SIGIR Forum* 51.1.

Hiltz, S. R., Kushma, J. A., and Plotnick, L. (2014). "Use of Social Media by US Public Sector Emergency Managers: Barriers and Wish Lists." In: *Proceedings of ISCRAM*.

Hughes, A. L. and Palen, L. (2009). "Twitter adoption and use in mass convergence and emergency events". In: *International journal of emergency management* 6.3-4.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial intelligence for disaster response". In: *Proceedings of WWW*. ACM.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media." In: *Proceedings of ISCRAM*.

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*.

Kumar, S., Morstatter, F., Zafarani, R., and Liu, H. (2013). "Whom should i follow?: identifying relevant users during crises". In: *Proceedings of Hypertext*. ACM.

Lin, J., Efron, M., Wang, Y., and Sherman, G. (2014). *Overview of the trec-2014 microblog track*. Tech. rep. MARYLAND UNIV COLLEGE PARK.

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., Santos, R. L. T., Macdonald, C., and Ounis, I. (2018). "Explicit Diversification of Event Aspects for Temporal Summarization". In: *ACM Trans. Inf. Syst.* 36.3.

Middleton, S. E., Middleton, L., and Modafferi, S. (2014). "Real-time crisis mapping of natural disasters using social media". In: *IEEE Intelligent Systems* 29.2.

Mileti, D. (1999). *Disasters by design: A reassessment of natural hazards in the United States*. Joseph Henry Press.

O'Dell, J. (2011). *How We Use Social Media During Emergencies*. https://mashable.com/2011/02/11/social-media-in-emergencies.

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.

Palen, L. and Liu, S. B. (2007). "Citizen communications in crisis: anticipating a future of ICT-supported public participation". In: *Proceedings of SIGCHI*. ACM.

Plotnick, L., Hiltz, S. R., Kushma, J. A., and Tapia, A. H. (2015). "Red Tape: Attitudes and Issues Related to Use of Social Media by US County-Level Emergency Managers." In: *Proceedings of ISCRAM*.

Reuter, C., Backfried, G., Kaufhold, M., and Spahr, F. (2018). "ISCRAM turns 15: A Trend Analysis of Social Media Papers 2004-2017". In: *Proceedings of ISCRAM*.

Reuter, C., Heger, O., and Pipek, V. (2013). "Combining real and virtual volunteers through social media." In: *Proceedings of ISCRAM*.

Reuter, C., Marx, A., and Pipek, V. (2012). "Crisis management 2.0: Towards a systematization of social software use in crisis situations". In: *International Journal of Information Systems for Crisis Response and Management* 4.1.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). "Okapi at TREC-3". In: *Nist Special Publication Sp* 109.

Rogstadius, J., Vukovic, M., Teixeira, C., Kostakos, V., Karapanos, E., and Laredo, J. A. (2013). "CrisisTracker: Crowdsourced social media curation for disaster awareness". In: *IBM Journal of Research and Development* 57.5.

Saleem, H. M., Xu, Y., and Ruths, D. (2014). "Novel Situational Information in Mass Emergencies: What does Twitter Provide?" In: *Procedia Engineering* 78.

Shaw, F., Burgess, J., Crawford, K., Bruns, A., et al. (2013). "Sharing news, making sense, saying thanks: Patterns of talk on Twitter during the Queensland floods". In: *Australian Journal of Communication* 40.1.

Tapia, A. H., Moore, K. A., and Johnson, N. J. (2013). "Beyond the trustworthy tweet: A deeper understanding of microblogged data use by disaster response and humanitarian relief organizations." In: *Proceedings of ISCRAM*.

Truelove, M., Vasardani, M., and Winter, S. (2015). "Towards credibility of micro-blogs: characterising witness accounts". In: *GeoJournal* 80.3.

Whipkey, K. and Verity, A. (2015). "Guidance for incorporating big data into humanitarian operations". In: *New York: UN-OCHA and the Digital Humanitarian Network's community of interest on Decision Makers Needs. http://digitalhumanitarians. com/sites/default/files/resource-field_media/IncorporatingBigDataintoHumanitarianOps-2015. pdf. Accessed May* 2, p. 2017.

Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., and Starbird, K. (2018). "From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW.