Incident Streams 2019: Actionable Insights and How to Find Them

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

Cody Buntain
InfEco Lab, New Jersey Institute of Technology (NJIT)‡
cbuntain@njit.edu

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

ABSTRACT

The ubiquity of mobile internet-enabled devices combined with wide-spread social media use during emergencies is posing new challenges for response personnel. In particular, service operators are now expected to monitor these online channels to extract actionable insights and answer questions from the public. A lack of adequate tools makes this monitoring impractical at the scale of many emergencies. The TREC Incident Streams (TREC-IS) track drives research into solving this technology gap by bringing together academia and industry to develop techniques for extracting actionable insights from social media streams during emergencies. This paper covers the second year of TREC-IS, hosted in 2019 with two editions, 2019-A and 2019-B, contributing 12 new events and approximately 20,000 new tweets across 25 information categories, with 15 research groups participating across the world. This paper provides an overview of these new editions, actionable insights from data labelling, and the automated techniques employed by participant systems that appear most effective.

Keywords

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

INTRODUCTION

Emergency services' ability to respond effectively is highly dependent on their ability to obtain actionable information about the on-the-ground situation. Traditionally, this collection is accomplished via either communication with the public through call-centres or through reports by first responders (FEMA 2011). Wide-spread adoption of smart phones and online networking platforms like Twitter (particularly by the younger generation) has made social media an increasingly common communication channel during emergencies (Castillo 2016). Indeed, given the notional tweet stream about an emergency, like a wildfire, we can imagine a range of information types that might be shared during the incident. While the majority of tweets in such a stream likely express sentiment, solidarity, and wishes to help from around the world (Olteanu, Vieweg, et al. 2015), more valuable content captures reports from news services and government officials that contain useful information for people in the incident area. Meanwhile, the most actionable information in these streams is contained within the small number of tweets by people in the affected region who are reporting first-hand about on-the-ground conditions and immediate safety and health needs.

This shift toward social media has been noted by emergency and civil protection services, who are increasingly searching for effective means to monitor these channels, use them to answer questions, respond to aid requests

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

and more (FEMA 2013). Monitoring these channels, however, is a challenging task given the high volumes of information posted in contrast to the relatively small proportion of actionable information (McCreadie et al. 2019).

Hence, a clear need exists for tools to assist response officers in tackling the deluge of social media content. Researchers have invested significant effort into this task, each addressing different stakeholder needs. Some of these works have also resulted in experimental 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), among others. Despite these efforts, technological solutions have not seen wide-spread adoption by response services, with commonly cited reasons including data quality (Hiltz et al. 2014), limited trained staff (Plotnick et al. 2015), resistance to social media as a primary communication channel (Tapia et al. 2013), and difficulties in integrating social media with current organizational policy/procedures (Reuter, Heger, et al. 2013). Indeed, Reuter, Backfried, et al. 2018 concluded in their study of ISCRAM papers that such research efforts have only made "a relatively small contribution to actual technology and industry".

The research community is poorly positioned to tackle institutional or staffing issues, but questions over *data quality* are soluble given sufficient evidence. We argue that by placing more effort into standardizing task definitions, metrics and datasets, then quantifying the value of automated solutions should be feasible. To this end, in 2018 we founded the Incident Streams (TREC-IS) initiative to establish this standardization (McCreadie et al. 2019). TREC-IS develops test collections and evaluation methodologies to evaluate automated social media monitoring solutions for crisis responders and to provide a re-occurring data challenge in which academic groups and solution providers can participate. Insights from that pilot edition (TREC-IS 2018) are detailed in an associated 2019 ISCRAM paper (McCreadie et al. 2019).

This paper continues that previous work, detailing the design improvements, new datasets, and insights gained from TREC-IS. TREC-IS 2019 includes two editions, 2019-A and 2019-B, and has a release of associated labelled Twitter dataset for each. This paper's primary contributions include:

1. The official overview for TREC-IS 2019, containing the 2019 task description, updated design and motivation, as well as dataset and participant performance statistics;
2. An analysis of actionable information, both in terms of prevalence and the factors that contribute to a social media message being seen as critical for emergency responders; and
3. A detailed examination of participant systems and insights about what learning and featurization techniques perform well in identifying high-priority, actionable information during times of crisis.

RELATED EFFORTS AND PRIOR EDITIONS OF TREC-IS

TREC-IS is part of a broader set of crisis informatics efforts that explore social media's role in the emergency management domain. This section situates TREC-IS in the context of these related efforts before describing relevant background from the pilot edition of TREC-IS from 2018.

Related Crisis Informatics Efforts

Early on, researchers recognized social media's potential as a data source for reactions to and information about emergencies, but making use of this data has proven non-trivial (Palen and Anderson 2016). For example, fragmentation in tasks, datasets, and evaluation metrics hinder assessment of state-of-the-art systems. TREC-IS exists to standardize these issues, with our goals being to construct a clear set of emergency response-relevant tasks, datasets on which these tasks may be executed, and metrics to evaluate systems' performance in these tasks.

Within social media streams, a common task for emergency responders is to classify documents based on the information they contain. A key aspect of TREC-IS is then to support automatic classification of actionable and useful information types that an emergency responder may want to find. To identify these information types, we build upon research into categorizing emergency related content. In particular, a survey (Castillo 2016) of previous categorization efforts identifies eight main dimensions: 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). TREC-IS falls within the space of categorization by information provided.

Crisis informatics researchers have also introduced constrained information tasks beyond classification, instead determining whether crisis-related social media messages contain sufficient information for a responder to take action. Purohit et al. (2018) present a model of message *serviceability* that scores content according to three axes: whether the message contains an explicit request, is addressed to an individual/organization, and whether it contains sufficient detail to identify a location, time, or related markers necessary to direct a response. Similarly, Sachdeva and Kumaraguru (2017) examine text posted to Facebook pages that represent police agencies in India, finding patterns that drive organizational responses. These narrower information tasks align with prior evidence (e.g., as shown in TREC-IS 2018 (McCreadie et al. 2019)) that a large majority of social media content posted during crises are not high-priority or critical messages, meaning simply identifying relevant content is insufficient. TREC-IS maintains a similar classification of *actionable* content, consisting of six information types that are consistently ranked as high-priority messages (e.g., requests for search and rescue or reports of emerging threats). We expand on this research by investigating aspects of social media content that make it particularly important for emergency responders.

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 examined two challenges: 1) disaster-related text retrieval in Twitter, and 2) tweet summarization during a disaster event. We have integrated two lessons from SMERP: First, the information needs defined in the text retrieval task were broad, making it difficult to map an information need to a response officer's activity. We therefore developed our own information ontology, detailed later. Second, SMERP included only four information needs and examined a single event, so participating systems' generalizability is unclear. Hence, for TREC-IS, we have opted for a larger pool of multiple events and event types.

Finally, TREC-IS is both part of and builds upon expertise stemming from the Text Retrieval Conference (TREC). 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 tracks, where a track is an area of focus in which particular retrieval tasks are defined. Tracks act as incubators for new research areas and often result in foundational research into core technologies, such as search engines (Robertson et al. 1995) or information extraction from social media (Lin et al. 2016).

TREC-IS Pilot Effort in 2018

In 2018, TREC-IS ran as a new TREC track and established a pilot set of test collections and evaluation methodology for subsequent editions. This section briefly describes the key organizational aspects of TREC-IS 2018, shown in Figure 1, as they provide the foundation for our discussion of TREC-IS in 2019. At a high level, participant TREC-IS systems can perform two tasks: classifying tweets by information type, and ranking tweets by criticality. As shown in Figure 1, each system receives a stream of filtered, event-relevant tweets and an ontology of information types from TREC-IS; each system then records tweet-level classifications and priority ratings, which they then submit for evaluation. We operationalize these tasks, define information types, and generate ground truth labels for these tasks using human assessors, as follows:

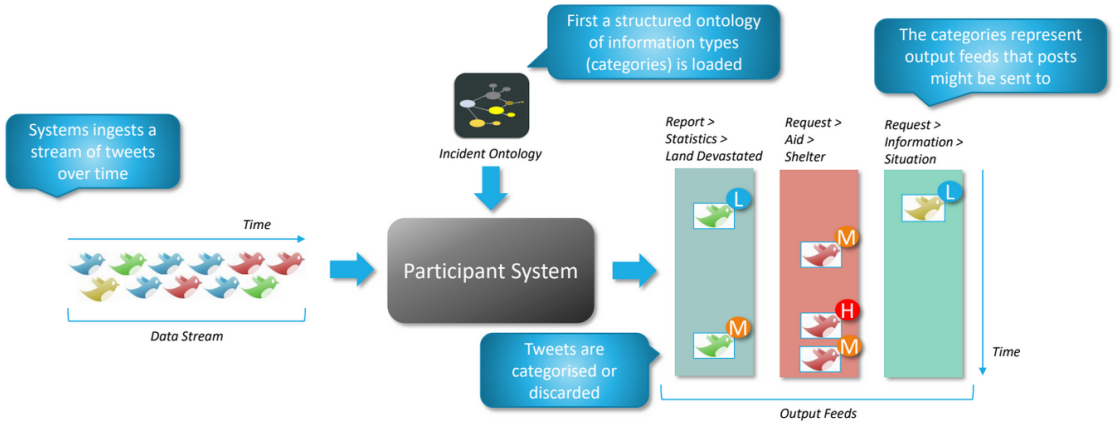


Figure 1. TREC-IS Task Visualization

Crisis-Relevant Information Types and Priorities: TREC-IS 2018 defines information 'types' to represent categories of information that emergency response officers might find interesting, such as 'Reports of Road Blockages' or 'Calls for Help'. We define these types based on a top-down analysis of incident management

ontologies, response documentation and discussion with experts. This valuable information is rare on social media, so to make the track's datasets useful beyond TREC-IS, we expand these information types via a bottom-up analysis of tweets. TREC-IS 2018 then groups the identified information needs into higher-level types shown in Table 1, which we have since modified in TREC-IS 2019, as discussed in later sections (McCreadie et al. 2019).

To capture the importance a given message has to emergency response officers, we also use four information criticality labels: low, medium, high, and critical, where 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.

Event Datasets: TREC-IS 2018 has developed a dataset of past events for training and evaluation. Relevant crises for TREC-IS include six natural and man-made events: wildfires, earthquakes, floods, typhoons/hurricanes, bombings, and shootings. In the pilot TREC-IS run, we have relied on event datasets shared by other emergency management initiatives and augment these sets with two custom crisis-event collections. This original set includes 21 events, 19 from CrisisLex (Olteanu, Castillo, et al. 2014) or CrisisNLP (Imran, Mitra, et al. 2016; Imran, Elbassuoni, et al. 2013). After pre-filtering to remove tweets marked as irrelevant or non-English and downsampling our two custom event sets, TREC-IS 2018 ended with approximately 25,000 tweets (Table 2).

Social Media Labelling: To evaluate participant systems, we require ground-truth data on information-type and criticality labels for the above tweets. For the 2018 effort, six NIST-hired assessors have labeled these tweets, marking priority and all relevant information types per message. After de-duplication, this labeling has resulted in 19,784 labeled tweets with 43,514 information types.

Evaluating Participant Systems: A key aspect of TREC-IS and TREC more generally is that research and professional groups submit systems to the track for evaluation. These research prototypes are instrumental for establishing the state-of-the-art and in driving research agendas in identifying the hard technical problems in these tasks. Each participating research group is allowed to submit up to four runs from candidate systems, each of which is referred to as a *run* and is evaluated separately. In 2018, 11 research groups from 8 countries have participated, submitting a total of 39 runs.

Metrics: The 2018 TREC-IS edition evaluated each participating run across two axes: information-type categorization, and information criticality. For information-type categorization under the 2018 pilot, participant systems were allowed only *one* information type per tweet, but when working with NIST assessors, we realized this constraint was burdensome. We therefore allowed NIST assessors to provide multiple labels per tweet despite the one-label participant constraint and evaluated information-type categorization in two ways: multi-type and any-type (we have since changed this restriction in 2019). The second axis of evaluation for a TREC-IS system was the extent to which it could rank information that emergency response officers need to see. Since our criticality labels are ordered (e.g., "low" is less than "critical"), we assigned numeric scores to these labels and calculated the *Mean Squared Error* between the human-assigned score and a system's score for each tweet.

Summary of TREC-IS 2018 Systems: To summarize what participants in the TREC-IS 2018 pilot attempted, it is worth noting the 2018 systems had far fewer training examples (around 1,300 tweets, denoted 'Boostrap' in Table 2) than systems in TREC-IS 2019, along with a small set of indicator terms for each of the 25 information types. As such, some approaches reported simply using keyword matching between frequently occurring indicator terms and the tweets within the stream for each event (Mehrotra and Pal 2018). This approach was one of the better performing in terms of categorization precision, but its recall was poor. Other approaches focused on using the limited training examples and indicator terms to produce supervised machine learned models, which can be divided along three dimensions: 1) how they expanded the training dataset to obtain sufficient examples for model learning; 2) how they represent tweets; and 3) what type of learning they attempted (classical or deep learning/neural). For example, two groups (Miyazak et al. 2018; García-Cumbreras et al. 2018) used Wordnet to expand the event query and indicator terms to increase recall. Success using this approach was mixed, as care needed to be taken not to expand too far and reduce the discriminative power of the query and indicator terms. Another group collected news articles from the BBC and Fox News that closely matched the event query to collect additional training data.

In terms of tweet representation, participants primarily focused on the tweet text, with some additionally considering the tweet author (Choi et al. 2018) or date information (Miyazak et al. 2018). Date information in particular was reported to add value (Miyazak et al. 2018). When representing the tweet texts, the most common approach was a bag of words representation (usually with term frequency-inverse document frequency weighting). Other groups reported improved performance using word-embeddings (Miyazaki et al. 2019) and entity extraction methods (Choi et al. 2018).

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

* – modified for 2019-A and 2019-B editions

For learning methodology, most groups opted for classical machine learning approaches (e.g., support vector machines or Naive Bayes), although a few groups also tested neural models (convolutional neural nets and multi-layer perceptrons). Where comparisons were made, classical approaches were more effective in 2018, though we will contrast these findings with 2019 participants later in this paper.

Main Conclusions from TREC-IS 2018 TREC-IS's pilot run and the eleven teams that participated provided invaluable information about system evaluation, metrics, and data collection. That effort demonstrated that a non-trivial (approximately 10% post-filtering) amount of actionable information exists in Twitter during emergency events. We 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).

UPDATES TO TREC-IS IN 2019

Conclusions from TREC-IS 2018 motivated us to continue this effort, as the need for computational support in identifying critical information present in social media remains unfulfilled. To expand opportunities to engage with the track, we have run TREC-IS twice in 2019, a first edition (2019-A) in June 2019, and the main TREC edition (2019-B) in September 2019. In both editions, participants have submitted systems for classifying tweets by information type and criticality, though we have modified performance metrics to differentiate between actionable information types and all types.

As before, participant systems are provided training datasets of tweets, labeled by information type and criticality. Unlike the 2018 edition, however, 2019 editions allow participant systems to provide *multiple* information types per tweet, rather than the single-type constraint in 2018. This change brings participants' information categorization task into alignment with the assessors' labeling task. In particular, participant systems now provide *all* categories they consider relevant for a tweet, with the categorization metrics updated to reflect this modification. These metrics, in effect, capture the overlap between assessor-selected categories and system-selected categories, where more overlap indicates better performance.

We have also made minor alterations to the information-type ontology, removing one of the 25 TREC-IS 2018 categories, adding a new one for locations, and refining four of the remaining types.[1] Table 1 presents this updated ontology, with * noting TREC-IS 2019 modifications. Specifically, we have removed the "Other-Unknown" category, as this category was very rare and have merged its contents into the "Other-Irrelevant" category. We have also added *Report-Location* to capture whether a tweet contains information about the location of the tweet subject, as such information is important to responders.

[1] see http://trecis.org/2019/2019Changes.html

Table 2. TREC-IS Datasets and Labelling Statistics

Dataset	Event Name	Event Type	Source	# Sampled	# Assessed
Bootstrap	2012 Colorado wildfires	wildfire	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	744	263
	2012 Costa Rica Earthquake	earthquake	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	288	247
	2013 Colorado Floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	777	235
	2012 Typhoon Pablo	typhoon/hurricane	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	649	244
	2013 LA Airport Shooting	shooting	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	683	162
	2013 West Texas Explosion	bombing	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	630	184
2018	2012 Guatemala earthquake	earthquake	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	178	154
	2012 Italy earthquakes	earthquake	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	118	103
	2012 Philipinnes floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	480	437
	2013 Alberta floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	739	721
	2013 Australia bushfire	wildfire	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	710	677
	2013 Boston bombings	bombing	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	543	535
	2013 Manila floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	443	411
	2013 Queensland floods	flood	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	744	713
	2013 Typhoon Yolanda	typhoon	CrisisLexT26 (Olteanu, Vieweg, et al. 2015)	629	564
	2011 Joplin tornado	typhoon	CrisisNLP Resource #2 (Imran, Elbassuoni, et al. 2013b)	152	96
	2014 Chile Earthquake	earthquake	CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)	321	311
	2014 Typhoon Hagupit	typhoon	CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)	6,696	4,192
	2015 Nepal Earthquake	earthquake	CrisisNLP Resource #1 (Imran, Mitra, et al. 2016)	7,301	7,301
	2018 FL School Shooting	shooting	Crawled(Twitter API)	1,118	1,118
	2015 Paris attacks	bombing	Crawled(GNIP)	2,066	2,066
2019-A	2019 Choco Flood	flood	Crawled(Twitter API)	854	389
	2014 California Earthquake	earthquake	Crawled(Twitter API)	128	127
	2013 Bohol Earthquake	earthquake	Crawled(Twitter API)	646	582
	2018 Florence Hurricane	typhoon	Donated by participant group	2,500	2,499
	2017 Dallas Shooting	shooting	Crawled(Twitter API)	2,500	2,500
	2016 Fort McMurray Wildfire	wildfire	Crawled(Twitter API)	2,500	2,500
2019-B	2019 Alberta Wildfires	wildfire	Crawled(Twitter API)	2,500	2,000
	2019 Cyclone Kenneth	typhoon	Crawled(Twitter API)	2,500	1,999
	2019 Luzon earthquake	earthquake	Crawled(Twitter API)	2,500	1,995
	2019 STEM School Highlands Ranch shooting	shooting	Crawled(Twitter API)	2,500	1,238
	2019 Durban Easter floods	flood	Crawled(Twitter API)	2,500	1,349
	2019 Poway synagogue shooting	shooting	Crawled(Twitter API)	2,500	667

New Datasets

For each TREC-IS edition, we have released a new dataset of tweets, their information types, and their priority labels, as produced by human assessors.As we have run TREC-IS twice in 2019, each edition has a new dataset, and systems are allowed to use datasets from prior editions for training. E.g., for 2019-A, participants could use the 'bootstrap' and '2018' TREC-IS pilot datasets for training. We then evaluate systems over the new 2019-A dataset, which contains 6 new events and around 8,500 tweets. Similarly, for 2019-B participants can leverage the 'bootstrap', '2018' and '2019-A' datasets for training and are evaluated against the '2019-B' dataset, which contains a further 6 events and 9,200 tweets. Over the four datasets created so far, TREC-IS has manually assessed over 35,000 tweets from 33 unique disasters, producing in excess of 125,000 labels. These events, sources and statistics are summarized in Table 2.

We note three additional differences between the datasets in 2018 pilot and TREC-IS 2019:

- We no longer rely on the CrisisLex or CrisisNLP datasets and instead use new events retrospectively crawled by TREC-IS organizers, based on manually selected keywords.[2]
- As our new event collections are both high-volume and noisy, we also downsample them. This filter process first removes non-English tweets, and for any event with more than 1,000 tweets remaining, we apply KMeans clustering ($k = 2,500$) to tweet texts and select one tweet from each cluster as that cluster's sample. Our assessment interface performs a further redundancy check on this downsampled collection, removing very similar tweets using cosine similarity-based fuzzy matching. Assessors then label any remaining tweets.
- For 2019-B, we transitioned fully to contemporary events that occurred in 2019. This move is notable, as Twitter now allows tweets in excess of 140 characters, which may affect participant systems and previously constructed models.

In all cases, assessors label either all sampled tweets per event or the budgeted assessment time expires. This time constraint is notable for 2019-B, where 15,000 were sampled, but only around 9,200 have been assessed (cf. # Sampled and # Assessed columns in Table 2).

[2]With one exception where a participating group donated an event.

Table 3. Six Actionable Information Types

Actionable Information Type	Type Description
CallToAction-MovePeople	The user is asking people to leave an area or go to another area
Report-EmergingThreats	The user is reporting a potential problem that may cause future loss of life or damage
Report-NewSubEvent	The user is reporting a new occurrence to which public safety officers should respond
Report-ServiceAvailable	The user is reporting that someone is providing a service
Request-GoodsServices	The user is asking for a particular service or physical good
Request-SearchAndRescue	The user is requesting a rescue (for themselves or others)

New Metrics Introduced in TREC-IS 2019

An additional conclusion from our 2018 pilot is that calculating performance metrics over all information types is insufficient. I.e., taking the macro average of these metrics over all types biases our results towards common but less critical categories, as these common types often lack actionable information. Consequently, we have modified evaluation metrics as follows:

Information Feed: To evaluate information-type classification performance, we calculate metrics, like accuracy, micro-averaged over all events but macro-averaged over information types. Emergency response officers, however, primarily care about seeing *all* valuable information; i.e. missing actionable information is more serious than seeing irrelevant data. We therefore report performance metrics when considering all information types *and* considering only actionable information types. This meta-class of *actionable* information types (Table 3) captures content that is likely critical for an emergency response officer to see. These constrained metrics therefoer provide better insight into system performance on the most important types of information. Specifically, we measure:

- **Information Feed, Info. Type Accuracy, All**: Overall classification accuracy, micro-averaged across events and macro-averaged across information types. This metric yields a high-level view of categorization performance but does not capture utility to emergency response officers.
- **Information Feed, Info. Type Positive F1, All**: Categorization performance when only considering the target class per information type. E.g., performance for the 'Request-SearchAndRescue' information type only includes the few tweets belonging to that type. This metric measures signal shown to emergency response officers 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 actionable information types shown in Table 3. This metric aims to capture whether systems find actionable information.

Information Priority: TREC-IS metrics also evaluate whether systems can identify key information that emergency response officers need to see, which we operationalize by comparing a system's information priority score for each tweet and that tweet's priority label as given by an assessor.

For this metric, however, we first map assessors' information priority labels into numerical scores (i.e., low=0.25, medium=0.5, high=0.75 and critical=1.0). Second, since some participant systems do not provide scores within an appropriate 0-1 range, we normalize these systems' scores via a max-min normalization, with a minimum score cap of 0.25. Our priority estimation metric then measures the *Mean Squared Error* (RMSE) between assessor score and normalized system priority score. As with information feed metrics, to distinguish between prioritization performance for actionable categories against all categories, we report prioritization error for all information types and over only the actionable types.

- **Prioritization, Priority RMSE, All**: Overall prioritization error, micro-averaged across events and macro-averaged across all information types.

- **Prioritization, Priority RMSE, Actionable**: Prioritization error, micro-averaged across events and macro-averaged across only actionable information types (Table 3).

Tweet Labeling Process

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 assessors. The TREC-IS 2019 budget has not allowed for multiple assessors per tweet, precluding agreement evaluations, but tweets labeled as actionable have been reviewed by track organizers, as described in the next section.

ANALYSING ASSESSOR LABELS AND PARTICIPANT METHODS

Having presented the structure and changes to TREC-IS 2019, we now turn to this paper's remaining contributions: an analysis of tweets assessors labeled as critical or actionable, and an exploration of participant systems' techniques and performance. For analysing assessors' labels, we examine the 35,000 manually assessed tweets to answer the following three research questions:

- **RQ1.1** – What types of information constitute critical messages, and how prevalent are they on Twitter?
- **RQ1.2** – Is the amount of critical information changing as Twitter evolves?
- **RQ1.3** – What features of a social media message make it actionable?

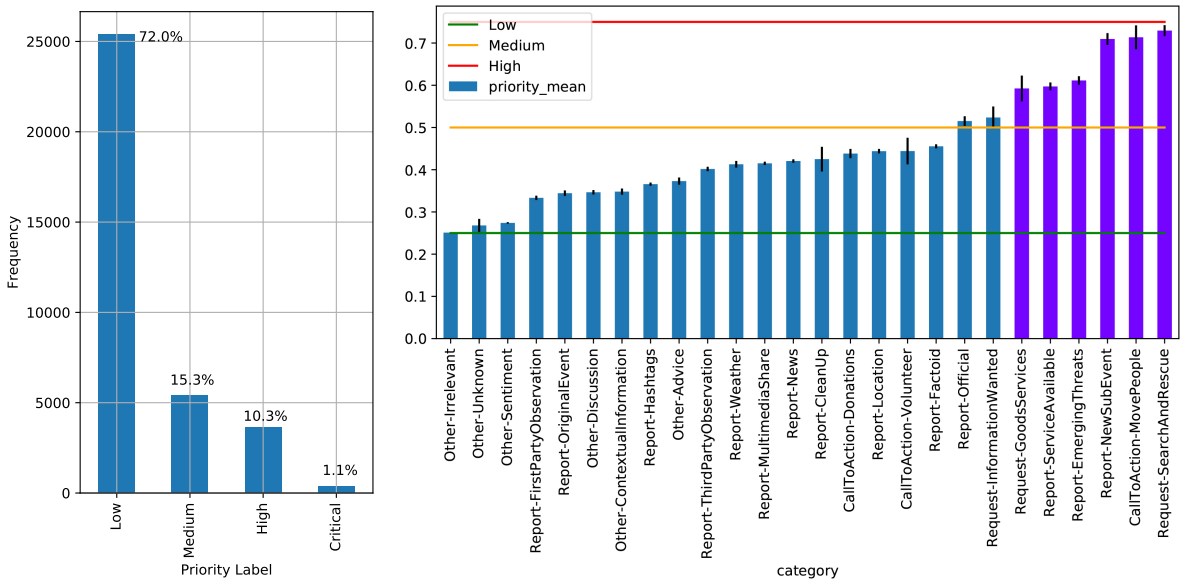
Following discussion of critical and actionable content, we then analyse systems submitted to TREC-IS 2019. This study investigates the 35 participant runs submitted to 2019-A and the 32 runs submitted to 2019-B, which collectively originate from 15 research groups. We answer the following three research questions about these runs:

- **RQ2.1** – What techniques did participants explore?
- **RQ2.2** – How well did these different systems perform?
- **RQ2.3** – Are different techniques yielding different performance results?

RQ1.1 – What Types of Critical Information Appear in Social Media?

A key finding in TREC-IS 2018 is that assessors find a non-trivial number (14.4%) of tweets containing high-priority or critical information across 21 disaster-related event datasets. While this result supports the common belief that social media contains useful and important content during emergencies, we find the majority of these high-priority messages fall into the six "actionable" information types shown in Table 3. Many events from TREC-IS 2018 occurred in 2013-2014, however, whereas TREC-IS 2019 editions are taken primarily from 2019, and the distribution of critical information may have changed over the years. Given this new data and the potential for Twitter's population and popularity to evolve, we revisit this analysis of criticality and information type.

Globally, across TREC-IS editions, the average tweet criticality is $\mu = 0.3507$ with a standard deviation $\sigma = 0.1791$, placing the average tweet in a low-to-medium priority. This metric is slightly lower than but consistent with results from the 2018 edition, wherein we find $\mu_{2018} = 0.3632$. Decomposing this importance by priority level for the 2019 editions (Figure 2a), assessors score 87.3% of tweets as low-to-medium priority (72% and 15.3% respectively, up



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

Figure 2. Criticality Distributions and Criticality by Information Type. Actionable information types are in purple.

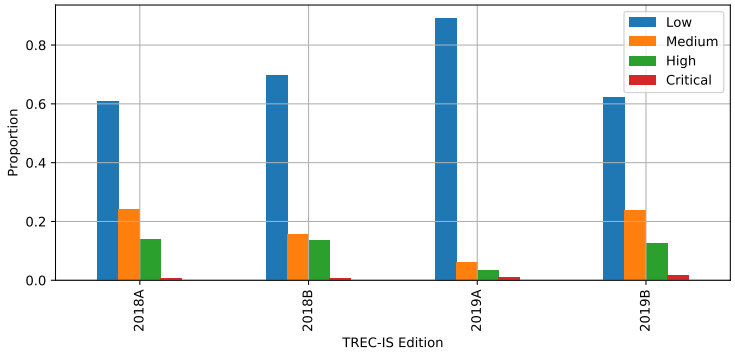


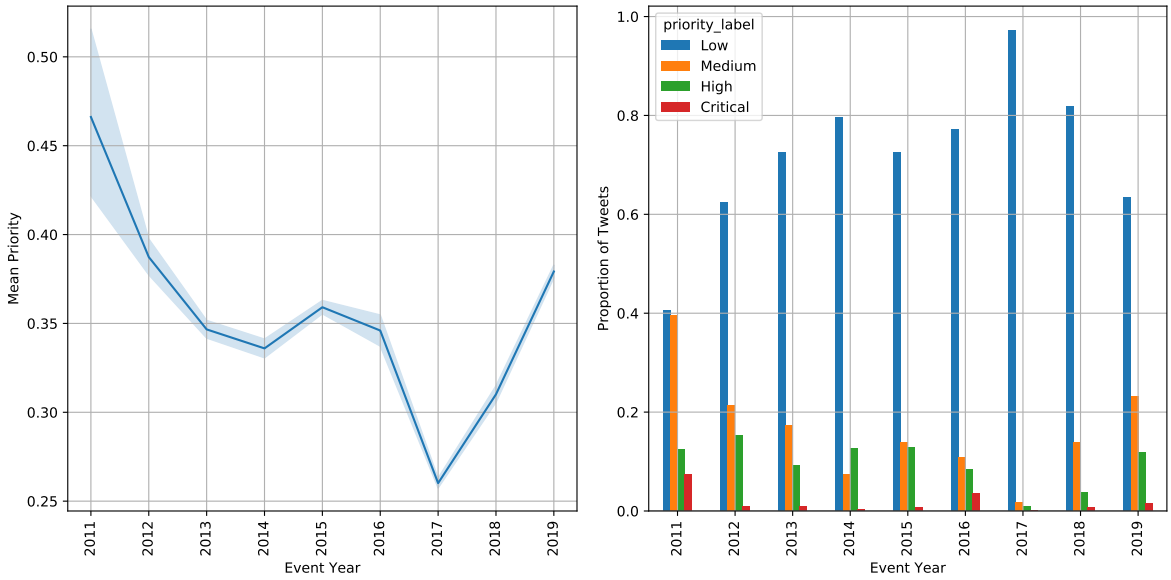
Figure 3. Criticality Distribution by TREC-IS Edition

from 85.6% in 2018). Prevalence of high-importance tweets has decreased in 2019, from 13.6% to 10.3%, but critical tweets have increased from 0.8% to 1.1%. In both cases, assessors see approximately 10% of tweets about disasters as highly important.

Turning to the interaction between criticality and information type in Figure 2b, our results are consistent with TREC-IS 2018: Actionable categories have the highest mean priorities, with a distinct separation between the first three types (search and rescue, calls for relocation, and new sub-events) and the second three (emerging threats, reports of new service availability, and requests for goods/services). This ordering indicates that our actionable information types are consistently more likely to contain critical insights, regardless of edition, answering RQ 1.1.

RQ1.2 – Is Criticality Changing Over Time?

Next, we examine information criticality distributions across TREC-IS editions. We first analyse criticality by edition: Bootstrap/2018A, 2018/2018B, 2019A and 2019B in Figure 3. An important observation apparent from this figure is that 2019A contains proportionally far more low-priority tweets and consequently fewer medium- and high-priority tweets than other editions. For critical tweets, however, we see a consistent increase over the editions, which might indicate the volume of critical information is increasing over time (as later editions use more recent events). Alternatively, our sampling is simply capturing more critical information. Hence, the question arises: As Twitter evolves and its user population grows, is the amount of critical information posted during crises also changing? Using assessors' evaluations from all TREC-IS editions, we can evaluate this yearly message priority. To this end, Table 4 reports event and tweet counts per year in TREC-IS datasets, and Figure 4 depicts annual changes



(a) Mean Priority By Year **(b) Priority Class By Year**

Figure 4. Tweet Priority Over Event Years

between 2011-2019. Figure 4a shows the Joplin tornado in 2011 had a tendency toward high-priority content, with 2012-2016 seeming to regress to the mean. Surprisingly, 2017 and 2018 see a large drop in mean priority. Figure 4b captures this effect as well, with high- and critical-priority content accounting for slightly more than 10% in all years except 2017 and 2018. These aberrations in 2017 and 2018 may result from event paucity in these years or from the over-representation of shootings in those years, which account for two of the three events. To conclude on RQ 1.2, we do not have strong evidence of increasing criticality over time; rather, the volume of critical information seems centered around 10% of the sampled streams, although this is strongly tied to the type of events covered.

Table 4. Yearly Breakdown of Events and Tweets

Event Year	Number of TREC-IS Events	Number of Tweets
2011	1	96
2012	6	1,448
2013	10	4,786
2014	3	4,379
2015	2	7,929
2016	1	2,000
2017	1	2,000
2018	2	3,117
2019	7	9,511

RQ1.3 – What Makes a Tweet Actionable?

Having shown valuable information is consistently available on social media during emergencies, the natural question is: what makes a tweet actionable, or at least valuable for a response officer to see? This question is important for system builders, as its answer indicates what factors a system should consider. While assessors provide both information-type and priority labels for each tweet, these labels do not tell us *why* they were assigned or *what information was used* to make that determination. As such, we require additional information to answer the above question.

While related work discussed above (Sachdeva and Kumaraguru 2017; Purohit et al. 2018) provides insight into the factors that make a post actionable by a response agency, we also perform a smaller-scale labelling study to identify whether additional priority-centric factors are present in critical crisis-related messages. We therefore select 170 tweets labeled as 'Critical' priority by our assessors in 2019-B – forming a set of likely valuable tweets. Next, we render each tweet using the same assessment interface as NIST assessors. A subset of TREC-IS coordinators

review each of these critical tweets, and a new assessor identifies features of the tweet that appear to contribute to actionability. Note, this new assessor was a computer scientist and experienced in the construction of automatic systems for this task (i.e., was a participant in previous TREC-IS editions). Hence, their categorization is reflective of factors one might believe an automatic system should consider when categorizing each tweet. After all tweets are analysed, we aggregate the outcome into three high-level information sources (Tweet Text, Linked Content and Author) with a total of 10 sub-categories, shown in the top part of Table 5 below.

Table 5. Information contained within the critical tweets from 2019-B.

Source	Information	Description	# Tweets	%
Tweet Text	Terms/Phrases	Individual terms or phrases are information bearing, such as 'trapped' or 'lost power'.	164	97%
	Location	The text explicitly mentions a location that is relevant to the event and the information contained is about that location	150	88%
	Event Mention	The text explicitly mentions the event, making it easier to determine that this is relevant	34	20%
	Time Mention	The text explicitly mentions a point in time that helped identify that the information contained was current	10	6%
	Person Mention	The tweet explicitly mentions a person or twitter user account, that increases the tweet's credibility.	10	6%
Linked Content	Article/Web Page	The tweet contains a link to an external web page or news article that contains valuable information and is current.	83	49%
	Tweet	The tweet links to or mentions another tweet, in this case it may the the linked tweet alone that is actionable rather than the source tweet.	21	12%
	Image	The tweet contains a relevant image that is providing valuable information.	8	5%
	Video	The tweet contains an embedded video that provides relevant information.	5	3%
Author	Name/Username	The tweet's username contributed to a belief that the information contained within was trustworthy, i.e. it came from an official source	7	4%
Regional Context Needed		To understand the tweet some additional information (not present in the tweet) is needed, such as an understanding of geographical landmarks in the affected area	42	25%
Tweet is Out of Date		The new assessor noted that based on the time-stamp of the tweet and when the information contained first became available, the information contained could be considered as out of date.	41	24%

Table 5's right-hand columns report critical-tweet counts and proportions of actionable tweets in which that information type was important. First, as we might expect, the most common information source is a tweet's text (97%); the remaining 3% of critical tweets may link to critical information (e.g., an evacuation order) or contain media data without text. Second, the vast majority of critical tweets (88%) also explicitly mention a location, consistent with the idea of serviceability as most emergency responses would need a location to send a response. Location references are also substantially more prevalent than in the overall tweet stream; e.g., assessors identified only 53% of 2019-B tweets as having an explicit location label. Third, unexpectedly, only 20% of critical tweets clearly mention the event, potentially because some events lacked a common descriptor when the critical information appeared, e.g. no agreed upon hashtag, or location(s) were used as a proxy (like 'Highlands Ranch just had a school shooting'). Practically, textual topical-relevance indicators appear less useful for identifying actionable tweets here. About 6% of tweets also contain either a temporal expression (e.g., 'issued at 3:40 a.m.' or '2 min ago') or mention a relevant person (e.g., an official), and these factors aid in contextualizing the content by supporting the reader in determining whether the content is current or a quotation.

Beyond tweet text, most critical tweets also provide some form of linked content (69%, sum of Linked Content sub-categories). The most prevalent type of useful linked content is either an article or web-page (49%); e.g., a news article is just published that contains on-the-ground reporting or a live feed. The next most prevalent is linked tweets (12%), followed by a small number of images and videos (5% and 3%, respectively). This result highlights that systems should not only rely on tweet text but also integrate linked content when identifying critical information. The last category of information that we consider is the name or username of the author of the tweet, which our assessor has mark as useful for only 4% of tweets, suggesting that author identity or veracity is not a strong indicator of actionable information.

Finally, two additional cases that are not strictly related to information within tweets are worth highlighting. First, our assessor noted that for 25% of tweets, extracting mentioned locations would require deeper knowledge of the affected region's geography. E.g., in the extract, 'Three wildfires currently out of control south of High Level, north of Peace River and north of Slave Lake currently', the actual referenced location is implicit and is only mentioned relative to other (unaffected) places. In the social media context, a tweet's author might naturally assume the reader has sufficient local knowledge, but this understanding is difficult for current geolocation technology.

Second, for 24% of tweets, our assessor questioned whether some actionable information was truly useful because, at the time of the tweet, the information was out-of-date. This line of inquiry occurred because the assessor saw the same information earlier in the stream, suggesting the response officer might similarly have seen this information previously. During the general TREC-IS labelling task, NIST assessors are instructed to consider each tweet independently, meaning our labels do not account for whether similar information has been previously seen.

RQ2.1 – What Did Participants Try?

Before analysing participant systems' performance, it is valuable to analyse what techniques are employed by these systems, which can provide insights into common approaches and trends in technology. In this section, we discuss trends in system descriptions for 2019-A and 2019-B based on the 2-3 sentence short summary participants submit with their systems.[3] We describe four main observations below:

- **Machine Learned Categorization is Prevalent**: 89.6% of submitted runs use a form of supervised machine learning to categorise tweets into information types. Two groups opted for a different approach, however: UAGPLSI experimented with a direct application of textual similarity between information-type descriptions and tweets, and IIT-BHU used unsupervised clustering to create vectors for each information type and calculated similarity between those vectors and tweets. Neither of these alternatives were as effective as supervised approaches.
- **Word Embeddings are the Most Common and Effective Text Representation**: All participant systems leveraged tweet text during categorization by converting text into a numerical vector representation. In 2019-A, 46% of systems used classical bag-of-words or n-gram representation, while the remaining 54% used a form of word or character sequence embedding (via GloVe, FastText or SkipThought, or otherwise implicitly generate a text embedding using a neural model like BERT). In 2019-B, embedding-based approaches increased to 56%, while classical representations dropped to 34%.[4] Examining the top 10 approaches for each year, 6/10 and 9/10 used embeddings in 2019-A and 2019-B respectively, indicating that these embeddings are a superior representation to classical approaches. Despite this result, we lack sufficient data to identify a best-performing embedding method.
- **Deep Learning is Becoming More Prevalent and Effective, but Traditional Machine Learning Remains Competitive**: Of the machine learning runs submitted, usage of deep-learning approaches was limited in 2019-A (29%) but increased to 39% in 2019-B. The majority still use traditional machine learning approaches (e.g., Naive Bayes, Logistic Regression or Random Forests). We also observed a large increase in the number of deep-learning based systems in the top 10 across editions (1/10 in 2019-A vs. 6/10 in 2019-B), indicating that participants are learning how to make neural approaches more effective. On the other hand, we note that the most effective system in terms of identifying actionable content for both editions is still a classical rather than deep learned model.
- **Evidence Suggests Participants Are Not Integrating Supporting/Linked Data**: Despite our finding that linked data contributes to making a tweet actionable, no participants mentioned leveraging such data in their system descriptions. While it is possible that participants did make use of such evidence but omitted any reference to it, it seems likely that usage of such data is still largely unexplored for this task.

RQ2.2 – How Well Did Participants Perform?

Tables 6 and 7 present performance metrics for 2019-A and 2019-B respectively for each run submitted by the 15 participant groups, ordered by F1 score across actionable information. While results are not comparable *across* editions (i.e., the underlying events are different), overall trends are comparable. A clear (and expected) result in both 2019-A and 2019-B shows F1 scores for actionable information types are significantly lower than for all information types; on average, runs achieve a two-to-three-fold increase in F1 by expanding to all information types. Across both editions, only runs from UAGPLSI performed as well or better on actionable types than on all types. A similar pattern exists for message prioritization, wherein performance increases by a factor of 1.5-to-2 from prioritizing actionable tweets to all tweets. This result shows actionable types are more difficult to identify than non-actionable types.

[3]Caveat: These summaries only provide us information about what participants thought was important enough to include, and hence will be incomplete. As such, caution should be taken when drawing conclusions from this data.

[4]For 2019-B. 10% of runs did not specify how they represented the tweet text.

Beyond comparing classification performance in actionable versus all information types, we also compare run rankings between actionable versus all information types and between F1 versus priority. In the former case, systems that perform well for actionable types may not perform well for all types, but an analysis of the rankings between these two systems suggests otherwise. That is, Spearman rank correlations between runs ranked by F1-actionable and F1-all are strong (> 0.69 in both editions), suggesting systems that do well at identifying actionable information also do well at identifying other information types. Between F1-actionable and RMSE-actionable, however, we see a different result: Spearman rank correlations are very weak to weak (between 0.19 and 0.4 in both editions), suggesting a disconnect between identifying actionable information type and ranking the most critical messages within these types.

A final key observation over both editions is that the average F1 scores for both actionable and all information types is very low (e.g., mean F1 on actionable types is 0.050 in 2019-A and 0.057 in 2019-B). Further, across 2019-A and 2019-B, only two runs perform beyond two standard deviations away from the mean (the top two irlabISI runs in 2019-A). This result holds for prioritization as well in both 2019-A and 2019-B, suggesting the overall state of the art in these tasks fall far short of addressing the needs of emergency response users. We see this finding as motivation to continue the TREC-IS initiative into 2020.

Table 6. 2019-A Submitted runs under the v2.3 evaluation script. Information Feed metrics range from 0 to 1, higher is better. Prioritization metrics range from 0 to 1, lower is better. The highest performance under each metric is highlighted in bold.

		Information Feed			Prioritization	
		Info. Type Positive F1		Info. Type Accuracy	Priority RMSE	
Group	**Run Name**	**Actionable**	**All**	**All**	**Actionable**	**All**
irlabISI	Base	**0.1695**	**0.2825**	0.8521	0.1559	0.1552
irlabISI	Base3	0.1487	0.2642	0.8775	**0.1132**	0.0833
irlabISI	Base2	0.1284	0.2541	0.8812	0.1145	0.0823
CS-UCD	ELFB3	0.1180	0.1827	0.7853	0.1171	0.0603
CS-UCD	EL1	0.0970	0.1703	0.7784	0.1149	0.0617
DICE_UPB	FastText	0.0922	0.1810	0.8501	0.1752	0.0737
CS-UCD	ELFB4	0.0918	0.1668	0.8519	0.1207	0.0623
CS-UCD	EL2	0.0884	0.1505	0.8324	0.1144	0.0633
DICE_UPB	BERT	0.0868	0.2421	0.8809	0.1514	0.0717
NYU	baseline_multi	0.0858	0.1827	0.8357	0.1223	0.1036
irlabISI	Deep	0.0856	0.1823	0.8765	**0.1132**	0.0833
DICE_UPB	BILSTM	0.0814	0.2041	0.8699	0.1744	0.0757
NYU	baseline	0.0567	0.1287	0.8930	0.1223	0.1036
DICE_UPB	DICE	0.0501	0.2183	0.8740	0.1532	0.0664
CMU	rf-autothre	0.0442	0.1000	0.8911	0.1890	0.0942
NYU	fasttext	0.0440	0.1199	0.8959	0.1406	0.0722
CMU	rf	0.0409	0.1228	0.8734	0.1571	0.0743
IRIT	rf_gb_threshold	0.0398	0.1886	0.7615	0.1134	0.0557
NYU	fasttext_multi	0.0320	0.1672	0.8982	0.1306	0.0911
ubIS	-	0.0312	0.1133	0.3790	NA	NA
BJUTDMS	run2	0.0237	0.0998	0.5565	0.1150	**0.0563**
CMU	xgboost-event	0.0205	0.1516	0.8718	0.1751	0.0724
IRIT	rf_gb_binary	0.0202	0.1513	0.8834	0.0751	0.0694
IRIT	rf_gb_binary_chain	0.0202	0.1513	0.8834	0.0751	0.0694
IRIT	rf_gb	0.0185	0.1756	0.8052	0.1134	0.0557
CMU	xgboost-extra	0.0106	0.1716	0.9017	0.1817	0.0732
BJUTDMS	run1	0.0071	0.0950	0.6740	0.1343	0.0623
CMU	xgboost	0.0049	0.1638	0.8997	0.1864	0.0754
CMU	rf-extra	0.0047	0.1381	0.8982	0.1886	0.0760
ICTNET	-	0.0046	0.0585	0.6844	NA	NA
BJUTDMS	run3	0.0014	0.0066	0.6642	NA	NA
SC	KRun28482low	0	0.0551	0.8962	0.1723	0.0756
SC	KRun68484low	0	0.0447	0.9003	0.1756	0.0747
SC	KRun2624435	0	0.0363	**0.9039**	0.1752	0.0743
SC	KRun60002002410001	0	0.0473	0.8909	0.1731	0.0753

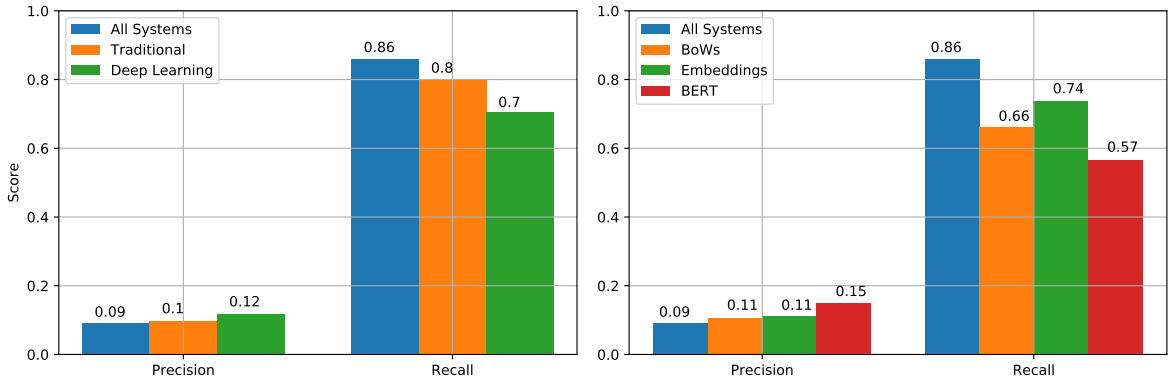
Table 7. 2019-B Submitted runs under the v2.3 evaluation script. Alerting metrics range from -1 to 1, higher is better. Information Feed metrics range from 0 to 1, higher is better. Prioritization metrics range from 0 to 1, lower is better.

		Information Feed			Prioritization	
		Info. Type Positive F1		Info. Type Accuracy	Priority RMSE	
Group	**Run Name**	**Actionable**	**All**	**All**	**Actionable**	**All**
CS-UCD	baseline	**0.1355**	0.2232	0.7495	0.0859	0.0668
DICE_UPB	BERT	0.1338	**0.2343**	0.8139	0.1558	0.0938
CMUInformedia	nb	0.1321	0.2167	0.8605	**0.0788**	**0.0544**
DICE_UPB	FOCAL	0.1287	**0.2343**	0.8159	0.1416	0.0829
CS-UCD	bilstmbeta	0.1269	0.1676	0.8378	0.1004	0.0822
NYU	base.multi	0.1135	0.2437	0.7997	0.1836	0.1104
DLR	USE_R	0.1111	0.2232	0.854	0.1767	0.1019
CS-UCD	bcnelmo	0.1099	0.1721	0.8452	0.1036	0.0769
DLR	BERT_R	0.0998	0.1989	0.856	0.1834	0.1019
NYU	fast.multi	0.0854	0.2256	**0.8808**	0.2153	0.1185
CMUInformedia	rf2	0.0642	0.1382	0.8624	0.1025	0.0683
CS-UCD	bilstmalpha	0.0614	0.171	0.86	0.1521	0.0893
NYU	base.sing	0.0606	0.1373	0.8658	0.1836	0.1104
CMUInformedia	rf3	0.0592	0.0813	0.8434	0.1660	0.2063
NYU	fast.sing	0.0431	0.1228	0.8739	0.2085	0.1169
UAGPLSI	baseline	0.0386	0.0302	0.8753	0.2067	0.1150
UAGPLSI	irn	0.0386	0.0302	0.8753	0.2132	0.1175
UAGPLSI	negative	0.0377	0.0278	0.8758	0.2075	0.1154
UAGPLSI	all	0.0377	0.0278	0.8758	0.2138	0.1177
ICTNET	dl	0.0347	0.0871	0.7285	0.1254	0.1451
CMUInformedia	rf1	0.03	0.1361	0.8638	0.0815	0.0551
IITBHU	run2	0.0275	0.0548	0.7892	NA	NA
DLR	Fusion	0.0249	0.0939	0.8689	0.1916	0.1077
IRIT	run2	0.0248	0.1725	0.8534	0.1175	0.0659
IITBHU	run1	0.0191	0.0893	0.8139	0.1879	0.1128
DLR	SIF_R	0.016	0.1004	0.8605	0.2093	0.1129
IRIT	run1	0.0151	0.1677	0.8418	0.1316	0.0911
DLR	MeanMaxAAE_Regression	0.0071	0.0922	0.8635	0.2111	0.1153
IRIT	run4	0	0.1317	0.7576	0.1461	0.0775
IRIT	run3	0	0.131	0.8565	0.1771	0.1028
CBNU	C1	0	0	0.8788	NA	NA
CBNU	S1	0	0	0.8788	NA	NA

RQ2.3 – How Do Different Learning Methods Perform?

As our earlier analysis of system descriptions illustrates, participants employ a variety of feature engineering and learning methods, and we can use this variation to evaluate how these different approaches impact performance. This question is further motivated by the observation that traditional learning methods are still competitive compared to more sophisticated deep learning approaches. Hence, in this section, we divide 2019-B participant systems into several comparison groups and merge all systems in each group into an exemplar system. We then examine precision and recall for our six actionable information types, comparing performance across these groups to the "meta-system" comprised of combined outputs from all groups. These comparisons illuminate whether the different approaches capture similar dynamics in tweet content because, if the meta-system has similar precision or recall to one or more of the comparison groups, then the alternate groups are providing little new information. Alternatively, if the meta-system deviates significantly from the comparison groups, each comparison group must be capturing different aspects of the data.

In our first comparison, we divide systems by ML paradigm: systems using traditional ML models (e.g., Logistic Regression, Naive Bayes, Random Forests, etc.) versus systems using deep learning methods (e.g., LSTMs, CNNs, etc.). Based on our analysis of system descriptions, 19 of the 2019-B systems use traditional ML approaches, compared to 11 that use deep learning. Figure 5a illustrates the differences in precision and recall for these classes of systems, showing the unified deep learning system obtains a slightly higher precision but lower recall than the traditional ML systems. Crucially though, Figure 5a shows the meta system achieves approximately 8% higher recall than either traditional or deep learning systems, suggesting that while these two approaches generally capture



(a) Learning Paradigm Groups **(b) Text Featurization Groups**

Figure 5. Performance in Actionable Information Types by Comparison Group

similar messages across our actionable types, each group appears able to identify a unique set of content. This result in turn indicates that ensemble approaches may be effective in the future.

An alternative driver for differences in performance may be featurization strategies systems employ. Systems that use recent advances in embeddings, for example, may outperform standard bag-of-words (BoW) methods by integrating context from large, pre-trained models. We examine this possibility in another set of comparisons, wherein we divide systems into three featurization groups: standard BoW (12 systems), word/n-gram embeddings (11 systems), and BERT-based bidirectional embeddings (7 systems). From this analysis, we find embedding-based systems (primarily GloVe and FastText) outperform BoW- and BERT-based models in recall (Figure 5b). As in the ML comparison, the meta-system achieves at least 15% higher recall than any comparison group, suggesting that the different featurization strategies reveal different sets of important messages.

GUIDANCE FOR FUTURE SYSTEMS

We now summarize several considerations for future systems designers:

Importance of Location in Criticality Assessment: As shown in Table 5, the majority of content perceived as critical includes an explicit mention of a location. This result is consistent with the serviceability model presented in Purohit et al. (2018) and suggests future systems would benefit from extracting place names and integrating geolocation pipelines, both for evaluating proximity to an event and for providing responders with actionable information.

Integrate Linked Content for Criticality Assessment: Also in Table 5, the article or web page to which a tweet links often contains valuable information that contributes to a tweet's perceived priority. Future systems could index this linked content or include the domain of the included link to capture this information.

Use of Word Embeddings in Classification: Many participating systems leveraged word embeddings when featurizing textual content. Our analysis suggests these embeddings provide a superior representation to classical bag-of-words approaches in both precision and recall for actionable information types. We lack data to determine which embeddings are superior, however.

Ensemble Learning: A meta-system outperforms individual systems by a significant margin, suggesting the different approaches employed by TREC-IS participants are capturing different aspects of information types. Future systems could therefore integrate a diverse set of featurization strategies and learning models into a single ensemble system.

CONCLUSIONS

In this paper we have provided an overview of the new 2019 editions (2019-A and 2019-B) of TREC-IS. TREC-IS is a standardization initiative that develops test collections and evaluation methodologies for identifying and categorize information and aid-requests made on social media during crisis situations. It also incorporates re-occurring data challenges in which researchers/developers can participate, enabling comparison of state-of-the-art systems. Over

two years and three editions, TREC-IS has manually annotated tweet streams for 33 emergency events, comprising 35,000 tweets and producing over 125,000 labels.

This paper provides analysis of both manually labeled tweets and systems participating in TREC-IS 2019, yielding insights into both what information is actionable and critical for crisis responders, as well as what automated techniques perform well in identifying high-priority, actionable information during times of crisis. From this analysis, we show high-priority information on social media tends to be either calls for aid, warnings about new sub-events or threats, evacuation information and reports of services coming back online, consistent with TREC-IS 2018. Furthermore, we show overall volumes of high or critical information remains near 10% of our samples. Through analysis of these critical tweets, we also show both mentions of location and linked content are more prevalent in critical messages than in the overall stream, indicating that these factors are important when finding high-priority content in a social media stream.

10 organizations participated in both the 2019-A and 2019-B editions (15 unique groups), submitting a total of 67 runs. Through analysis of these systems, we observe that the prevalence and effectiveness of deep learning approaches is increasing (particularly BERT-based systems), but little evidence exists that systems are integrating linked content within the tweets, which may improve effectiveness. In terms of overall performance, we are confident that systems are both becoming more sophisticated and improving in performance, but advances still need to be made before such systems will be ready for live deployment.

The Incident Streams track is slated to continue in TREC 2020 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

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.

Choi, W.-G., Jo, S.-H., and Lee, K.-S. (2018). "CBNU at TREC 2018 Incident Streams Track". In: *Proceedings of the 27th Text REtrieval Conference Proceedings (TREC)*.

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.

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.

García-Cumbreras, M., Díaz-Galiano, M., García-Vega, M., and Jiménez-Zafra, S. (2018). "SINAI at TREC 2018: Experiments in Incident Streams". In: *Proceedings of the 27th Text REtrieval Conference Proceedings (TREC)*.

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*.

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). "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*.

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

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

Mehrotra, H. and Pal, S. (2018). "IIT-BHU In TREC 2018 Incidents Stream Track". In: *Proceedings of the 27th Text REtrieval Conference Proceedings (TREC)*.

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

Miyazak, T., Makino, K., Takei, Y., Okamoto, H., and Goto, J. (2018). "NHK STRL at TREC 2018 Incident Streams track". In: *Proceedings of the 27th Text REtrieval Conference Proceedings (TREC)*.

Miyazaki, T., Makino, K., Takei, Y., Okamoto, H., and Goto, J. (2019). "Label Embedding using Hierarchical Structure of Labels for Twitter Classification". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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 Anderson, K. M. (2016). "Crisis informatics—New data for extraordinary times". In: *Science* 353.6296, pp. 224–225.

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*.

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.

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*.

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.

Sachdeva, N. and Kumaraguru, P. (2017). "Call for service: Characterizing and modeling police response to serviceable requests on facebook". In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pp. 336–352.

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.