CrisisFACTS: Buidling and Evaluating Crisis Timelines

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

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

ABSTRACT

Between 2018 and 2021, the Incident Streams track (TREC-IS) developed standard approaches for classifying information types and criticality of tweets during crises. While successful in producing substantial collections of labeled data, TREC-IS as a data challenge had several limitations: It only evaluated information at type-level rather than what was reported; it only used Twitter data; and it lacked measures of redundancy in system output. This paper introduces Crisis Facts and Cross-Stream Temporal Summarization (CrisisFACTS), a new data challenge piloted in 2022 and developed to address these limitations. The CrisisFACTS framework recasts TREC-IS into an event-summarization task using multiple disaster-relevant data streams and a new fact-based evaluation scheme, allowing the community to assess state-of-the-art methods for summarizing disaster events Results from CrisisFACTS in 2022 include a new test-collection comprising human-generated disaster summaries along with multi-platform datasets of social media, crisis reports and news coverage for major crisis events.

Keywords

Emergency Management, Crisis Informatics, News, Twitter, Facebook, Reddit, Wikipedia, Summarization

INTRODUCTION

US National Incident Management System (NIMS) guidance directs public information officers (PIOs) to monitor social media channels, answer questions, and report requests for aid to the Incident Commander. Despite this direction, PIOs lack adequate tools and manpower to monitor social media effectively, given the volume of information posted and the need to categorize, cross-reference and verify that Information. To meet these demands, emergency response agencies are increasingly relying on volunteers to manually identify actionable information. While these volunteer groups improve scalability, they are still comparatively slow, with substantial risk that valuable information may be lost or delayed.

Despite substantial effort within crisis informatics to address these issues, this research has had little impact beyond academic communities Reuter and Kaufhold 2018, with recent studies enumerating concerns and questions that remain for disaster-response personnel (e.g., Castillo et al. 2021). These concerns include 1) segregation of information sources (e.g., Twitter is nearly the lone platform of study); 2) "presenting crisis-relevant information from social media in a useful manner" (Castillo et al. 2021); 3) summarizing large volumes of social media; and 4) supporting disaster-response personnel in allocating attention during disaster.

The CrisisFACTS initiative aims to mitigate these issues by supporting event summarization research across multiple data streams and crises. At a high level, this framework, as shown in Figure 1, includes 1) a multi-platform test-collection comprised of news and social media from 8 crisis events; 2) an encoding of disaster-related information needs into queries; 3) ground truth data for evaluating crisis summaries; and 4) evaluation methodologies to assess how well community participants perform at this task. This paper discusses the CrisisFACTS pilot in 2022, launched at the annual Text Retrieval Conference (TREC), its resulting test collections, and its context as an open data challenge–where research groups have submitted output from their event-summary solutions for evaluation. These summaries highlight new developments across crises in the form of *event timelines* (Allan et al. 2001). Through these aspects CrisisFACTS supports research, development, and evaluation in crisis summarization.

*http://gla.ac.uk and http://dcs.gla.ac.uk/~richardm
†corresponding author
‡https://ischool.umd.edu/ and http://cody.bunta.in/

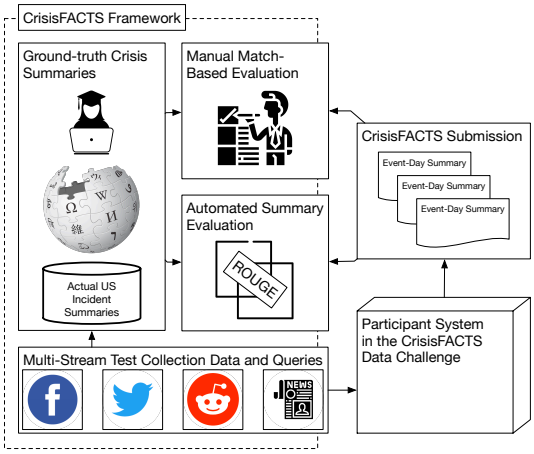


Figure 1. CrisisFACTS Test Collection and Framework. CrisisFACTS provides test collections around crises for participants in the CrisisFACTS community challenge to consume. These systems generate daily event summaries and submit them back to CrisisFACTS for multiple kinds of evaluation, both automated, n-gram-based and manual, match-based assessment.

This paper's overview of the 2022 CrisisFACTS pilot contributes the following:

1. We introduce the CrisisFACTS track, its motivation, and how it fits into the wider context of information systems for crisis response research.
2. We detail the CrisisFACTS test collection, with statistics and descriptions of the data provided, how it should be used, and its limitations.
3. We use the ground truth data developed for this test collection to answer questions regarding information coverage across sources during emergencies.

THE IMPORTANCE OF TEST COLLECTIONS AND DATA CHALLENGES

Over the last decade, considerable research has gone into automated tooling to highlight valuable information from social media platforms for crisis-responders. High-level tasks have included content categorization (Castillo 2016), criticality estimation (McCreadie et al. 2019), or summarization (Rudra et al. 2016). For example, categorization efforts have developed approaches to find affected people via social media (Imran et al. 2013), estimate infrastructure damage (Truelove et al. 2015), identify eyewitness accounts (Olteanu et al. 2014; Diakopoulos et al. 2012) or more broadly assign categories based on the information conveyed (McCreadie et al. 2020).

To enable sound research and development of these technologies, researchers need *test collections*, which are comprised of 1) a corpus of documents (Twitter posts, images, etc.); 2) ground truth (a representation of 'correct' output for a task, e.g. manually defined classification labels or a gold-standard summary); and 3) a standardized evaluation methodology and metrics (that define how to evaluate a system given the corpus and ground truth) (Voorhees, Harman, et al. 2005).[1] For computer-aided tasks, high-quality test collections are critical to ensure new technologies are evaluated and different solutions are compared fairly. Otherwise, it becomes impossible to determine the true "state of the art".

Test collections are often developed by individuals single research groups, then open-sourced for the wider community. In crisis informatics, notable examples include CrisisLex (Olteanu et al. 2014) and CrisisMMD (Ofli et al. 2020). Alternatively, multiple researchers/groups can collaborate on a *data challenge/evaluation campaign*. Such data challenges typically include three tasks: 1) pool resources to develop an effective test collection for a given task; 2) determine effectiveness of state-of-the-art solutions for that task using the new test collection; and 3) promote/advance research on that task.

[1]While the term 'dataset' is often used interchangeably with 'test collection' in academic writing, the terms are distinct in that datasets typically do not specify an evaluation methodology. While norms of a given field may govern evaluation, such ambiguity can lead to differences in experimental settings, making true comparison problematic.

Practically, a data challenge produces test collections for a task, but organisers also organise an event wherein research and industry members can submit solutions to that task. Organisers provide official evaluations for submitted systems, often resulting in a leaderboard. Data challenges are also usually multiple years, where each year produces new test collections. These long-running challenges provide a common place where researchers can come together to compare solutions; identify issues with test collections and revise; and pool resources to produce larger and higher quality test collections than would be reasonable for individual researchers. CrisisFACTS builds on one such test collection/data challenge, the TREC Incident Streams 2018-2021 tracks for crisis content categorization (McCreadie et al. 2019).

CrisisFACTS is a new data challenge, targeting automatic event timeline generation during natural disasters, with its inaugural run in 2022.

INFORMATION NEEDS OF AN EMERGENCY RESPONDER

As the CrisisFACTS framework uses online content to produce and assess timelines of relevant information during an emergency, one must define the *information needs* for disaster-response personnel. In this context, when facing disaster, the affected area will activate a local crisis response team to manage the disaster. This team has two main roles: 1) deploy and manage response efforts (e.g., deploying search-and-rescue or directing supply distribution); and 2) manage communication with stakeholders (e.g., the press, government officials and volunteer groups). To perform this role effectively, members of this response team needs good situational awareness and an up-to-date view of events on the ground during the event.

Hence, CrisisFACTS focuses on fulfilling this information need of 'what is happening on the ground'. Not every message or development is relevant to the response effort, however, and varies across disaster types. As such, we make this need more explicit by enumerating specific kinds of important information:

- Damage to key infrastructure, or evacuations
- Changes to affected areas or damage assessments
- Reports regarding civilians and responders, such as casualties, seeking shelter, needing immunizations, or missing
- Emerging threats to life, property, infrastructure, or response operations
- Critical needs, such as food, water or medicine
- Weather concerns, e.g. high wind, temperatures, humidity, floods, or watches/warnings
- Risks from hazardous materials, e.g., chemicals, fuels, infectious agents or radiation
- Restriction to the use or availability of resources
- Progress made and accomplishments by responders
- Incident-command transitions, e.g., transferring command and control to new teams

These response teams typically need summaries of this information at points throughout the emergency–e.g., at the beginning of a new shift, so new team members can be brought up-to speed. Currently, this crisis summarization is performed manually, e.g. by populating incident-report forms, that are then emailed or stored in shared file-storage.

CORE CONCEPTS AND RELATED WORK

Crisis Summarization

Given the above information needs, concise and complete summaries of emergency are invaluable. Generating these summaries is known as crisis summarization, where a system ingests one or more text documents about a crisis and produces a (usually fixed-length) summary. This research area is situated within the larger text summarization research domain (Nenkova and McKeown 2012), which has two main approaches: extractive and abstractive (Munot and Govilkar 2014). Extractive approaches construct summaries by aggregating text snippets from input documents into the summary, generally without modification (Dutta et al. 2019). Abstractive summarization, in contrast, generates new text using the input documents as a prompt (Li et al. 2018).

Resulting summaries generally have two formats: 1) as a single text unit (e.g., a paragraph) or 2) as an itemized timeline of information. The former is simpler and more common (Nenkova and McKeown 2012), but the latter is more useful in contexts where information changes or becomes stale over time (Allan et al. 2001). The CrisisFACTS data challenge targets timeline summary generation, and participants can submit both extractive and abstractive solutions.

TREC – The Annual Text Retrieval Conference

TREC is a combined conference and evaluation campaign that encourages research into information retrieval (IR) technologies on large test collections, supporting search, categorization, recommendation, summarization, and related IR tasks. Sponsored by the National Institute of Standards and Technology (NIST), TREC has run annually for over 25 years and consists of a set "tracks," where each track provides a unique data challenge. TREC tracks incubate new research areas, where the first year of a track often concretizes the problem and develops necessary infrastructure (test collections, evaluation methodology, etc.) to support the following track year(s). 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 research interests of multiple groups. TREC has been highly influential in the IR domain, resulting in foundational research into search engines (S. E. Robertson et al. 1995) and information extraction from social media (Lin et al. 2016). CrisisFACTS is a new TREC track in 2022 and builds upon prior TREC tracks.

TREC for Crisis Informatics

TREC has two relevant tracks for the crisis informatics domain: Temporal Summarization and Incident Streams, summarized below, as they directly influence CrisisFACTS:

Temporal Summarization (2013-2015) : The Temporal Summarization track (Aslam et al. 2015) was one of the first data challenges to investigate producing an evolving event summary over time. Given a stream of sentences extracted from news articles about an emergency event, participants generated a variable-length summary for that event on a set day. The target user in this case was a generic news consumer, and, in contrast to CrisisFACTS, no further guidance was given about what constitutes relevant or useful information. This track popularised the idea of "nugget-based evaluation" for evaluating a summary (Pavlu et al. 2012). Rather than assessing textual similarity between a gold-standard summary–as was the standard summarization metric at the time–individual summary items were evaluated against a gold-standard list information items ('nuggets') about the event. Participants would be awarded for each nugget their system covered but would only receive that gain once and would be penalized for each item they returned that did not contain new relevant information. In this way, participants were encouraged to produce short summaries while maximising information coverage. The track ran for three years, spanned 46 events, and matched around 20,000 sentences to 3,700 nuggets. While the subject matter and space of relevant information in Temporal Summarization is less relevant for the crisis informatics domain, this 'matching' evaluation is the inspiration for the fact-matching assessment in the CrisisFACTS framework.

Incident Streams (2018-2021) : The Incident Streams track (McCreadie et al. 2020) (TREC-IS) was designed to bring together academia and industry to research technologies for processing social media streams during emergencies. The track aimed to categorize and prioritize information on Twitter for emergency response personnel, motivated by the real-world application of improving situational awareness. TREC-IS produced curated feeds of Twitter posts, where each feed corresponds to a particular type or priority level of information. Information "types" were defined based on existing hierarchical incident management information ontologies, For instance, for a flash flooding event, feeds might include, "requests for food/water", "reports of road blockages", and "evacuation requests". During an emergency, individual emergency management operators and other stakeholders could consume a subset of these curated feeds corresponding to their responsibilities. The track ran for 4 years and 7 editions between 2018 and 2021, with test collections spanning 98 crisis events with over 136,000 tweets labelled into 25 information types and priority levels. CrisisFACTS is run by the same organisation team as TREC-IS and acts as a direct extension of that track. While CrisisFACTS track has notable similarities with TREC-IS–both ingest streams of textual content during crises; both require systems to identify information of interest to emergency responders–TREC-IS did not account for the substantial redundancy in social media streams and was restricted to Twitter. CrisisFACTS re-uses much of the TREC-IS Twitter data and event descriptions but expands the scope of relevant information sources and actively addresses this redundancy issue.

Cranfield II Evaluation Paradigm and Pooling

The above TREC tasks, and CrisisFACTS require ground truth for assessing solutions to their respective tasks, but constructing this ground truth is a non-trivial task given the large scales of data at play. To address this issue, CrisisFACTS builds on lessons from the IR field, inspired by the lessons learned from the Cranfield II experiments (S. Robertson 2008). These experiments used human assessors to judge items returned by participants. In the context of search, assessors judged document relevance given a user's query. In a timeline summarization context, as in CrisisFACTS, assessors judge the informational content of items included within the timeline. Performance is then based on the total value of the top items returned for a range of information needs. A practical consideration when performing this manual assessment, however, is the imbalance between large datasets and limited assessment time. Consequently, the number of items assessed is usually far smaller than the size of the corpus, and selecting the set of content to judge needs to be selective. To solve this issue, IR evaluations make use of a technique called pooling (Buckley and Voorhees 2004) to select items to have judged. The idea underpinning pooling is that, given a set of candidate solutions for a task, items returned in the top ranks can act as 'votes' for that item to be judged. The more participants that return an item, the more likely that item should be included in the "pool" of items to be judged. CrisisFACTS uses this pooling strategy across participants to select timeline items to be manually assessed for fact matches.

CRISISFACTS TASK FORMULATION AND TERMINOLOGY

At a high level, CrisisFACTS is an assessment framework for measuring how well systems perform at creating daily summaries of crisis events. The CrisisFACTS framework, as outlined in Figure 2, provides multiple streams of crisis-relevant data, including Twitter, Reddit, Facebook, and online news sources, broken down by day. Participant systems consume these daily streams and produce daily summaries for a given crisis. Specifically, participant systems consume sentence-length items from these streams ("stream-items") and construct a short timeline of important information from them, filtering and compressing content where appropriate. Participants then return a prioritized list of "facts" for an <event,day> pair, ranked by perceived importance for inclusion in the summary. CrisisFACTS evaluates the top k scored items from each participant's summary, where k is event-specific. After evaluation, CrisisFACTS organizers return participants' performance metrics back to the community, allowing participants to see how well their systems performed relative to their peers.

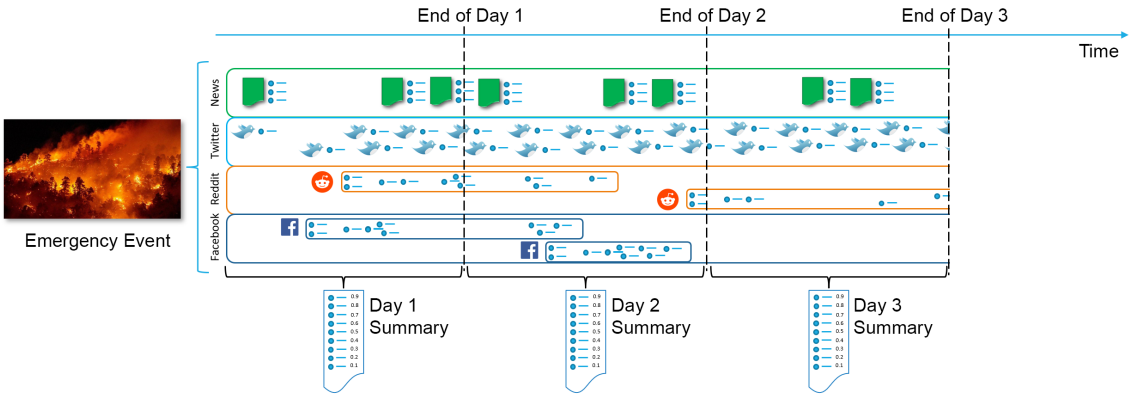


Figure 2. Core Crisis Summarization Task. CrisisFACTS is an evaluation framework for assessing how well systems perform at summarizing the major daily developments of a crisis.

CrisisFACTS Terminology

Event	An event is a period of time where a natural disaster (a hurricane, wildfire or flood) occurred, which we want to summarize. An event runs for multiple days and we want to produce a summary at the end of each day.
Day	A day is a 24 hour period within an event. A participant system generates a summary at the end of a day using only content published during that day.
Streams	The test collection corpus (input documents) are provided in the form of multiple time-stamped sequences of text items collected from different online sources (news providers, Twitter, Facebook and Reddit). Each source is referred to as a stream. A participant system uses the content from these streams for a particular day to produce a summary on that day.
Stream-Item	A stream-item is a single piece of text provided as part of a stream and is the unit of summarization, i.e. if performing extractive summarization it is a set of stream-items that form the output summary. Stream-items are typically around sentence-length, but may vary greatly depending on the source they came from.
System-Item	A system-item is a text item returned by a participant system in their timeline summary. For extractive systems, a system-item is also a stream-item. For abstractive systems, a system-item is a new generated item produced using one or more stream-items as a prompt.
Pool	The pool is the list of system-items selected to be assessed using pooling. Only pooled system-items contribute to a participant system's score.
Fact	A fact is a short text sequence describing a relevant piece of information about the event. Facts are manually defined by human assessors, and form the first part of the ground truth of the text collection.
Fact List	The list of all facts produced for an event.
Fact Match	A fact match is an explicit link between a system-item and a fact, indicating that the system-item contains the same information as that fact. Human assessors manually examine system-items returned by participant systems, compare each to the fact list, and produce fact matches. A system-item may match multiple facts. Fact matches form the second part of the text collection ground truth.
Query	A short piece of text that describes a question that an emergency response officer might ask during an event.
Query List	A set of queries for an event that represents the different information needs of an emergency responder. Participant systems may use the query list for an event as a guide when selecting relevant content.
Participant System	A participant system is a system whose output has been submitted to the CrisisFACTS track for evaluation.

CONSTRUCTING THE CRISISFACTS TEST COLLECTION

Figure 3 outlines the data requires in the CrisisFACTS framework, which broadly consists of a set of social media/news sources (on the left), and a set of ground-truth summaries (on the right). Participant systems consume data from these multi-source test collections of social media and news data to produce daily summaries for each crisis event in the CrisisFACTS collection. CrisisFACTS evaluations then compare these summaries against a set of gold-standard summaries for what actually happened during the event. Below, we first describe the events on which the CrisisFACTS 2022 test collections focus and then how we collect social media, news, and ground-truth data. To lower the barrier to entry, all test-collection data for CrisisFACTS is available via the ir_datasets[2] project (MacAvaney et al. 2021).

Test Collection: Events

For 2022, we selected 8 crisis events for the pilot CrisisFACTS test collection. Our selection criteria for these crises was 1) whether we have social media/news data from multiple streams for an event, and 2) whether external validation data was available from the ICS-209-PLUS All-Hazards dataset (St. Denis et al. 2020; Denis et al. 2022).[3] These 8 events, and relevant statistics, are listed in Table 1.

For each event, we grouped the content within each stream into single-day periods, so we can evaluate summaries for each <event,day> pair. Each pair has a unique identifier, of the form 'CrisisFACTS-<EVENT-ID>-r<DAY>'–e.g.,

[2]ir_datasets is a Python package that provides a common interface to many IR datasets. https://ir-datasets.com/
[3]The ICS-209-PLUS provides actual disaster summaries from the US National Incident Management System, between 1999-2020.

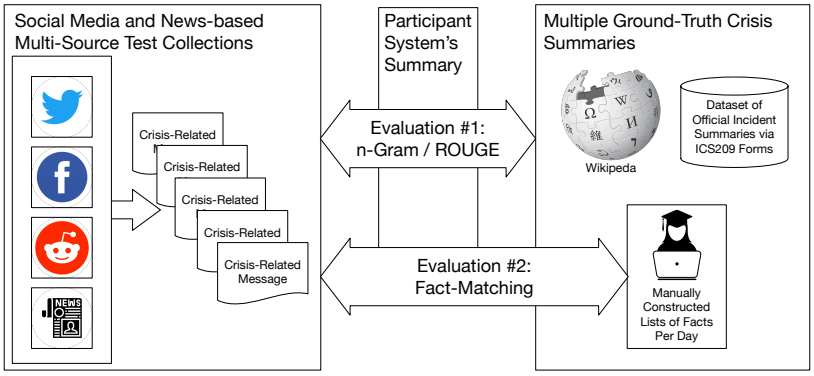


Figure 3. Overview of CrisisFACTS Data Sources. Participant systems consume multi-source test collections (on the left), from Twitter, Facebook, Reddit, and News sources. Systems produce automates summaries, and CrisisFACTS evaluates these outputs against gold-standard/ground-truth summaries from three sources: Wikipedia, a database of real-world ICS 209 reports, and a manually constructed set of daily facts (on the right).

'CrisisFACTS-001-r1' is the first day of event 001.[4] Only events with at least 10 manually identified facts were used for evaluation, and participants did not know in advance what <event,day> pairs would be selected. CrisisFACTS 2022 has released 65 <event,day> pairs to participants, and 31 of which were used in evaluation, as outlined in Table 1.

Table 1. Events and Days Comprising the CrisisFACTS 2022 Test Collection.

ID	Event	Type	ICS-209-PLUS ID	Description	Days Evaluated
CrisisFACTS-001	Lilac Wildfire 2017	Wildfire	2017_7418460_LILAC 5	The Lilac Fire was a fire that burned in northern San Diego County, California, United States, and the second-costliest one of multiple wildfires that erupted in Southern California in December 2017.	Dec 7th, 8th, 9th, 10th, 11th
CrisisFACTS-002	Cranston Wildfire 2018	Wildfire	2018_9217623_CRANSTON	The Cranston Fire was a wildfire that burned in southwest Riverside County, California, in the United States.	July 25th, 26th, 28th
CrisisFACTS-003	Holy Wildfire 2018	Wildfire	2018_9226396_HOLY	The Holy Fire was a wildfire that burned in the Cleveland National Forest in Orange and Riverside Counties, California.	Aug 6th, 7th, 8th, 9th, 11th
CrisisFACTS-004	Hurricane Florence 2018	Hurricane	2018_9085670_HURRICANE FLORENCE RESPONSE, 2018_9247758_HURRICANE FLORENCE NC AG EOC, 2018_9248081_HURRICANE FLORENCE, 2018_9249225_HURRICANE FLORENCE GARNER RSOI	Hurricane Florence was a powerful and long-lived Cape Verde hurricane that caused catastrophic damage in the Carolinas in September 2018.	Sep 9th, 10th, 11th, 12th, 13th, 14th
CrisisFACTS-005	Maryland Flood 2018	Flood	–	In the afternoon of May 27, 2018, after over 8 inches (20 cm) of rain in a span of two hours, the historic Main Street in Ellicott City, Maryland was flooded.	May 27th, 28th
CrisisFACTS-006	Saddleridge Wildfire 2019	Wildfire	2019_10754067_ SADDLERIDGE	The Saddleridge Fire was a wildfire burning near the San Fernando Valley of Los Angeles County, California.	Oct 11th, 12th
CrisisFACTS-007	Hurricane Laura 2020	Hurricane	2020_11820062_KIF HURRICANE LAURA SUPPORT	Hurricane Laura was a deadly and destructive Category 4 hurricane that is tied with the 1856 Last Island hurricane as the strongest hurricane on record to make landfall in the U.S. state of Louisiana.	Aug 27th, 28th
CrisisFACTS-008	Hurricane Sally 2020	Hurricane	2020_11923560_ HURRICANE SALLY FL STATE T1 RED IMT	Hurricane Sally was a destructive Atlantic hurricane which became the first hurricane to make landfall in the U.S. state of Alabama since Ivan in 2004.	Sep 12th, 13th, 14th, 15th, 16th, 17th

Test Collection: Queries

Much of the social media content posted during crises is of little value to disaster-response personnel, as sentiment and social-support messaging are common but uninformative (McCreadie et al. 2019). Therefore, some means for specifying the kinds of information of interest for CrisisFACTS is necessary. Likewise, we expect that a number of groups would want to apply existing search techniques to identify useful and relevant information that *should* be included in a crisis summary. To support these points, the organisers developed a query set for the task, which is a direct translation of information needs extracted from ICS 209 incident summaries. These queries encode information needs that constitute disaster summaries, such as casualties, road closures, etc. While these queries come from fields within the ICS 209 forms, not all fields in the ICS 209 are relevant to all types of disasters–e.g., mass immunizations are not relevant to hurricane events. As such, we constructed a main 'general' set of queries that are relevant to most of the emergency event types. These include queries such as 'How many people are affected' or 'What areas are being evacuated'.

[4]'r1' is not always the day of the event, as numbering is based on when the first related stream-items were collected, which may pre-date the event, (e.g. where a hurricane is forming for instance).

We also created a smaller set of event-type-specific queries. For example, the wildfire type includes the queries 'How quickly is the fire spreading' and 'What is the fire containment level'. Meanwhile, the hurricane set contains queries like 'Where has the hurricane made landfall' and 'How fast is the hurricane travelling'. The query set for a particular event is then the concatenation of the general and event specific sets appropriate for that event type.

Test Collection: Corpora / Streams

Focusing on the left side of Figure 3, for each event, we collected textual content for each across four different data sources. Table 3 shows the frequency of items from each source and event. These data streams provide input for participant systems. We excluded Wikipedia as a source for these input streams, as Wikipedia pages for crisis events are written retrospectively, with information and citations that would not be available during the event.

- **Twitter**: For Twitter, we relied on data collected as part of TREC-IS, which used the Twitter Enterprise API prior to the academic API's release. All Twitter data used for CrisisFACTS 2022 has been collected already for TREC-IS, cleaned, and has a portion of content that has already been manually assessed for type and priority. This re-use served to provide a known base from which participants can work, since relevance and priority labels in TREC-IS could be used to filter irrelevant or low-priority content.

- **Reddit**: For Reddit data, we relied first on the CrowdTangle API, a platform owned by Meta and part of the larger Facebook ecosystem. We relied on CrowdTangle's Reddit search as Reddit's native search is suboptimal; similarly, while we could use the Pushshift.io dataset (Baumgartner et al. 2020) for this search, data sizes for the monthly Reddit submission and comment datasets are large, and we would need to build search indices for them. The CrowdTangle API simplified this need and provided consistency when we use it for data collection from Facebook's public pages. Therefore, we used the CrowdTangle API to search for Reddit submissions, retrieving each submission's Reddit ID. Then, using Reddit's native API, we collected all comments associated with that submission ID.

- **Facebook**: For Facebook, as with Reddit, we used the CrowdTangle API and the same set of queries from TREC-IS. CrowdTangle returns matching posts made by public "pages" and "groups" on Facebook's platform, which are often owned by news and local organizations, as shown in Table 2. This data includes the text of the post, any hyperlink included in the post, the source Facebook page, and metrics on public reactions to and engagement this content. Excluded from this content, however, are all the contents of comments on these posts; CrowdTangle returns only the top-level post itself, no textual response from Facebook users.

- **News**: For collecting news content, several sources provide news coverage (e.g., GDELT Newsdesk from LexisNexis, MIT's MediaCloud, etc.), but the local coverage of these sources is limited. Instead, we observed that many of the sources in the Facebook dataset are classified as journalistic news sources (see Table 2 for a list of the most frequent page categories), where news accounts for at least 44% of these posts. As such, many of the hyperlinks from the posts generated from the Facebook data likely point to news articles about the event. We then extracted these hyperlinks and, using the Newspaper3k library,[5], collected plain-text content of these articles. If Newspaper3k is unable to parse an article's HTML or collect the article's content, that link is discarded.

Table 2. Top-20 Page Categories in Relevant Facebook Posts. The vast majority of content comes from news pages and local organizations.

Category	Count	Category	Count
BROADCASTING_MEDIA_PRODUCTION	31,342	COMMUNITY	12,099
MEDIA_NEWS_COMPANY	31,294	PERSON	68,40
NEWS_SITE	30,038	ACTIVITY_GENERAL	5,736
NEWS_PERSONALITY	29,186	RELIGIOUS_ORGANIZATION	5,663
GOVERNMENT_ORGANIZATION	26,393	LOCAL	4,170
RADIO_STATION	22,539	TV_SHOW	3,489
TV_CHANNEL	22,534	COMMUNITY_ORGANIZATION	3,256
NON_PROFIT	18,124	GOVERNMENT_OFFICIAL	2,833
TOPIC_NEWSPAPER	15,981	CHARITY_ORGANIZATION	2,631
JOURNALIST	12,352	FIRE_STATION	2,599

[5] https://newspaper.readthedocs.io

Table 3. CrisisFACTS 2022 Stream Statistics.

Event	Request ID	Date	# News Items	# Tweets	# Reddit Posts	# Facebook Posts	# Total
Lilac Wildfire 2017	CrisisFACTS-001-r3	Dec 7th	770	5,883	6	629	7,288
	CrisisFACTS-001-r4	Dec 8th	1,125	14,684	1,200	2,222	19,231
	CrisisFACTS-001-r5	Dec 9th	116	4,848	74	801	5,839
	CrisisFACTS-001-r6	Dec 10th	140	3,544	374	349	4,407
	CrisisFACTS-001-r7	Dec 11th	136	2,875	9	374	3,394
Cranston Wildfire 2018	CrisisFACTS-002-r1	July 25th	311	4,200	0	545	5,056
	CrisisFACTS-002-r2	July 26th	694	5,830	108	1,234	7,866
	CrisisFACTS-002-r4	July 28th	298	3,903	16	1,021	5,238
Holy Wildfire 2018	CrisisFACTS-003-r5	Aug 6th	205	3,267	38	892	4,402
	CrisisFACTS-003-r6	Aug 7th	96	4,177	282	1,172	5,727
	CrisisFACTS-003-r7	Aug 8th	425	4,186	60	1,254	5,925
	CrisisFACTS-003-r8	Aug 9th	479	4,142	32	1,331	5,984
	CrisisFACTS-003-r10	Aug 11th	64	3,107	37	815	4,023
Hurricane Florence 2018	CrisisFACTS-004-r13	Sep 9th	852	913	1,723	5,180	8,668
	CrisisFACTS-004-r14	Sep 10th	1,942	3,668	15,119	18,009	38,738
	CrisisFACTS-004-r15	Sep 11th	2,780	6,396	18,854	27,435	55,465
	CrisisFACTS-004-r16	Sep 12th	3,988	8,438	30,942	35,099	78,467
	CrisisFACTS-004-r17	Sep 13th	3,103	8,989	17,243	35,107	64,442
	CrisisFACTS-004-r18	Sep 14th	2,704	11,585	11,151	33,170	58,610
Maryland Flood 2018	CrisisFACTS-005-r3	May 27th	477	5,902	679	322	7,380
	CrisisFACTS-005-r4	May 28th	766	12,576	967	2,098	16,407
Saddleridge Wildfire 2019	CrisisFACTS-006-r5	Oct 11th	1,310	11,171	15,014	2,324	116
	CrisisFACTS-005-r6	Oct 12th	457	7,781	10	1,116	9,364
Hurricane Laura 2020	CrisisFACTS-007-r13	Aug 27th	2,321	35,405	8,295	0	46,021
	CrisisFACTS-007-r14	Aug 28th	4,085	715	1,740	9,621	16,161
Hurricane Sally 2020	CrisisFACTS-007-r4	Sep 12th	674	2,505	3,821	1,656	8,656
	CrisisFACTS-008-r5	Sep 13th	1,180	2,562	528	3,408	7,678
	CrisisFACTS-008-r6	Sep 14th	3,613	4,220	1,955	8,937	18,725
	CrisisFACTS-008-r7	Sep 15th	2,685	7,772	1,218	10,724	22,399
	CrisisFACTS-008-r8	Sep 16th	3,733	15,130	2,144	13,099	34,106
	CrisisFACTS-008-r9	Sep 17th	1,997	6,631	1,568	6,891	17,087

Test Collection: Ground-Truth Summaries and Facts

Moving to the right of Figure 3, this section focuses creating high-quality daily crisis summaries against which we can compare participant output. For this first year of CrisisFACTS, we rely on two types of evaluation: summary- and fact-based assessments. For summarization, we use existing event summaries gathered from Wikipedia and official reports contained in the ICS-209-PLUS dataset (Denis et al. 2022). For fact-based assessments, we rely on TREC assessors to first create lists of facts describing each crisis event and then to match these facts to facts returned by participants.

Gold Standard Event Summaries

As bridge between the traditional summarization methods and Temporal Summarization-inspired matching-based evaluation, the CrisisFACTS pilot year includes event-level summarization as a performance metric. To assess the quality of participants' event summaries, however, we require "gold standard" summaries against which we can compare. We use two sources for these summaries: Wikipedia, wherein the pages for each crisis event has an associated, manually created summary, and an archive of the actual incident summaries extracted from the NIMS database (Denis et al. 2022).

- **Wikipedia Summaries** Every crisis event in the CrisisFACTS dataset has an associated Wikipedia entry, which includes a summary of the event. This "page summary" is available in the `extract` field for a given page in the Wikipedia API or from the `page.summary` field in the Python wrapper for the Wikipedia API. This field generally corresponds to paragraphs in the Wikipedia page's zeroeth section, which appears above the page's table of contents (see Figure 4). Extracting summaries from Wikipedia is simply a matter of using the Wikipedia API to collect the `extract` field from each event's page.
- **Actual Incident Summaries from the ICS-209-PLUS All-Hazards Dataset** While Wikipedia summaries communicate a high-level event summary, these summaries are written for the public and from a historical perspective, not for the utility of emergency-response personnel. For a better-matched alternative, we turned to the NIMS database, as collected in St. Denis's ICS-209-PLUS All-Hazards dataset (Denis et al. 2022). This dataset includes 185,956 official FEMA incident reports between 1999-2020, which we match to CrisisFACTS events (Table 1 shows matched ICS-209-PLUS report IDs in the 'ICS-209-PLUS ID' column). Entries in the ICS-209-PLUS dataset are digitized copies of the ICS 209 incident report form, like that shown in Figure 5.

Figure 4. Event Summary in Wikipedia Page. The red box captures the section of the Wikipedia article returned for `page.summary` in the Wikipedia API wrapper.

Event Fact List

While the above sources provide examples of 'good' crisis summaries, these representations have potential issues. First, it is unclear the recall of the information in these summaries. Useful information may exist in participants' outputs that does not get mentioned in these gold standard summaries, and CrisisFACTS evaluation should reward systems for finding such information. Second, it may be difficult for similarity metrics such as ROUGE to distinguish coverage of textually similar but informationally distinct content. Hence, CrisisFACTS also had human assessors construct high-recall lists of 'facts' for each crisis.

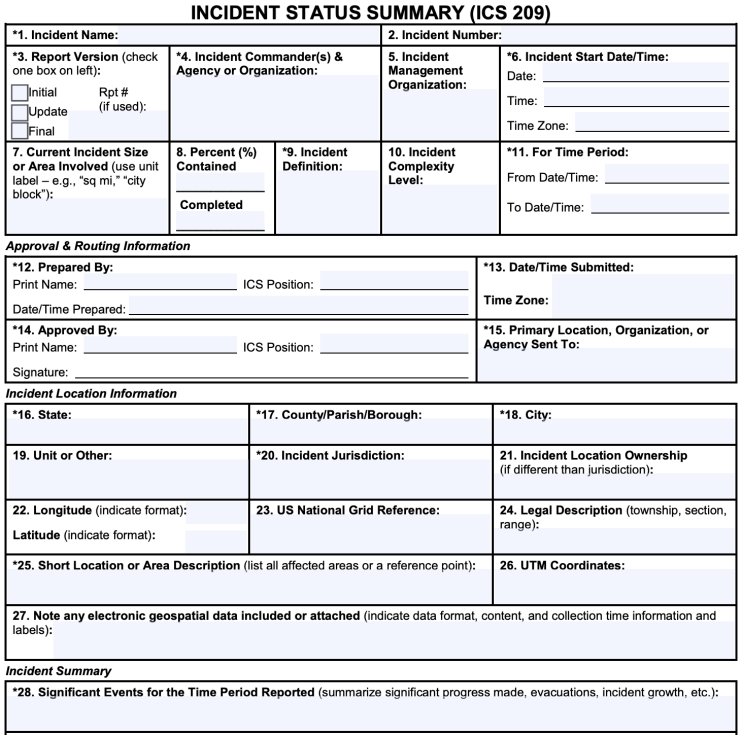
These TREC assessors were directed to read the Wikipedia page for a given crisis, follow links to cited news articles for that event, and record relevant "facts" about this event. Prior to this assignment, assessors underwent training by the CrisisFACTS organisers, who reviewed the ICS 209 forms and crisis-relevant information needs. This training directed assessors to record only information that might be entered into an ICS 209 form and to record "facts" as text snippets describing atomic information that might be relevant to a response officer and the date that information was relevant (usually when the thing happened). Once the Wikipedia page and any associated news articles were analysed, assessors were directed to then analyse all of the tweets marked as 'high' or 'critical' priority from the TREC-IS Twitter datasets for each event. The number of facts per day produced can be found in the Facts column of Table 4.

Note that different TREC assessors analysed each event, so fact reporting style will be constant within an event but may not hold across events. For example, some assessors preferred short facts, such as '50% containment' and 'boil water advisory issued', while others might simply quote the entire source item as a fact like 'RT @NWSWilmingtonNC: The @NHC_Atlantic has indicated that #Florence has made landfall at Wrightsville Beach, NC (34.2 N, -77.8 W) at 7:15 AM near the Wrightsville Beach Water Tower. #ncwx #scwx'. This inconsistency is a potential confounder for automatic similarity metrics, as longer-form facts may introduce noise, while shorter ones may not textually match the item text. This concern is an area for improvement in the future edition(s).

CRISISFACTS MANUAL EVALUATION

The above types of ground truth data provide textual representations of the important information to include within a summary. However, utilizing these ground truths will naturally necessitate a form of automatic similarity calculation between these representations and the actual participant system output, which may be inaccurate. Hence, it would be advantageous to include an alternative effectiveness measure that is based on a manual comparison between these representations and the system output.

This is the role of *Fact-Matching* evaluation. The idea underpinning this approach is that, as we have a manually curated fact list for each event and day, human assessors can label each item returned by a participant system based on which facts those items include. This matching allows us to quantify coverage of all facts that a participant system provides and measure the quantity of information a system missed (a critical metric for identifying directions for system improvement). It also provides us a method to quantify information redundancy within a summary, as we can count the number of times a summary mentions each fact (where an optimal summary would only mention each fact once).



INCIDENT STATUS SUMMARY (ICS 209)

***1. Incident Name:**			**2. Incident Number:**	
***3. Report Version** (check one box on left): ☐ Initial ☐ Update ☐ Final Rpt # (if used):	***4. Incident Commander(s) & Agency or Organization:**		**5. Incident Management Organization:**	***6. Incident Start Date/Time:** Date: Time: Time Zone:
7. Current Incident Size or Area Involved (use unit label – e.g., "sq mi," "city block"):	**8. Percent (%) Contained** **Completed**	***9. Incident Definition:**	**10. Incident Complexity Level:**	***11. For Time Period:** From Date/Time: To Date/Time:

Approval & Routing Information

***12. Prepared By:** Print Name: ICS Position: Date/Time Prepared:	***13. Date/Time Submitted:** **Time Zone:**
***14. Approved By:** Print Name: ICS Position: Signature:	***15. Primary Location, Organization, or Agency Sent To:**

Incident Location Information

***16. State:**	***17. County/Parish/Borough:**	***18. City:**
19. Unit or Other:	***20. Incident Jurisdiction:**	**21. Incident Location Ownership** (if different than jurisdiction):
22. Longitude (indicate format): **Latitude** (indicate format):	**23. US National Grid Reference:**	**24. Legal Description** (township, section, range):
***25. Short Location or Area Description** (list all affected areas or a reference point):		**26. UTM Coordinates:**
27. Note any electronic geospatial data included or attached (indicate data format, content, and collection time information and labels):		

Incident Summary

***28. Significant Events for the Time Period Reported** (summarize significant progress made, evacuations, incident growth, etc.):

Figure 5. A Blank ICS 209 Form. The ICS-209-PLUS dataset (Denis et al. 2022) includes several hundred thousand of these incident summaries, populated by actual emergency-response personnel in the field.

Manual Evaluation Setup: Pooling

Unlike the above ground truth, which are participant system agnostic, *Fact-Matching* requires that we select items from each participant's summary to have our human assessors perform fact matching. In an ideal world, we would be able to have all items returned by a participant summary analysed, however cost and time considerations make this impractical. As such, we employ a system output pooling strategy (Buckley and Voorhees 2004) to select what system-items will be assessed. As we construct pools of output, we count the number of facts identified by the TREC assessors in the "Event Fact List" above for each event-day pair. Using the `priority` field participant runs provide for every fact, we take the top-k highest priority facts for each run's event-day pair. This k value is taken from the number of assessor-identified facts for that event-day, varies across event-day pairs, but remains the same for all systems evaluated on that event-day.

Because CrisisFACTS accepts both extractive and abstractive runs and abstractive runs can combine multiple input-stream elements into a single fact element, pooling strategies must differ between these two types. For extractive runs, we de-duplicate elements based on their `streamID` fields, which refers to a unique text element from the input data stream. For abstractive runs, we rely only on exact matches. In either case, we then compute an average priority for that element's over all participant runs that return that element in its top-k set. Pool depths can also differ between extractive and abstractive runs to allow organisers to balance the number of facts included in the pools from each type.

The pooling columns in Table 4 report the number of each type of system-items that were assessed for each <event,day> pair. Participant system-items could either have been extractive (items directly taken from one of the input streams) or abstractive (a new item generated using the input streams). We report the number of items pooled for the extractive runs based on what stream the item came from, as well as the number of items contributed by the abstractive runs. Notably, only one group submitted abstractive runs resulting in a low proportion of items selected for pooling initially, which was compounded by an issue with loading those items within the assessment system. As a result, items from abstractive runs were only pooled for 3/8 events; assessment coverage is therefore likely insufficient for fact-matching evaluation. In total 8,822 system-items were assessed for fact matches, with the majority of pooled items coming from either the news (32.6%) or tweet (44.9%) streams.

Table 4. CrisisFACTS 2022 Fact List and Pooling Statistics.

Event	Request ID	Date	Facts	Pooling					
			# Created	# News Items	# Tweets	# Reddit Items	# Facebook Items	# Abstractive	# Total
Lilac Wildfire 2017	CrisisFACTS-001-r3	Dec 7th	267	113	217	1	69	-	400
	CrisisFACTS-001-r4	Dec 8th	75	66	261	29	44	-	400
	CrisisFACTS-001-r5	Dec 9th	14	20	130	1	11	-	162
	CrisisFACTS-001-r6	Dec 10th	29	44	208	16	25	-	293
	CrisisFACTS-001-r7	Dec 11th	19	45	141	0	13	-	199
Cranston Wildfire 2018	CrisisFACTS-002-r1	July 25th	27	100	136	0	10	22	268
	CrisisFACTS-002-r2	July 26th	10	51	68	2	0	12	133
	CrisisFACTS-002-r4	July 28th	13	56	90	0	7	8	161
Holy Wildfire 2018	CrisisFACTS-003-r5	Aug 6th	37	92	129	9	30	-	260
	CrisisFACTS-003-r6	Aug 7th	42	36	248	58	37	-	379
	CrisisFACTS-003-r7	Aug 8th	39	157	161	1	30	-	349
	CrisisFACTS-003-r8	Aug 9th	37	152	172	2	29	-	355
	CrisisFACTS-003-r10	Aug 11th	17	31	77	2	17	-	127
Hurricane Florence 2018	CrisisFACTS-004-r13	Sep 9th	15	86	43	34	7	-	170
	CrisisFACTS-004-r14	Sep 10th	55	113	59	191	37	-	400
	CrisisFACTS-004-r15	Sep 11th	26	128	71	137	12	-	348
	CrisisFACTS-004-r16	Sep 12th	14	78	49	69	2	-	198
	CrisisFACTS-004-r17	Sep 13th	37	154	78	138	30	-	400
	CrisisFACTS-004-r18	Sep 14th	46	139	108	116	37	-	400
Maryland Flood 2018	CrisisFACTS-005-r3	May 27th	55	134	187	47	31	43	422
	CrisisFACTS-005-r4	May 28th	15	42	130	21	1	15	209
Saddleridge Wildfire 2019	CrisisFACTS-006-r5	Oct 11th	116	115	201	14	70	35	435
	CrisisFACTS-008-r6	Oct 12th	12	57	77	2	7	12	155
Hurricane Laura 2020	CrisisFACTS-007-r13	Aug 27th	188	27	311	62	0	-	400
	CrisisFACTS-007-r14	Aug 28th	11	118	14	4	10	-	146
Hurricane Sally 2020	CrisisFACTS-008-r4	Sep 12th	12	38	36	62	10	-	146
	CrisisFACTS-008-r5	Sep 13th	14	95	58	9	10	-	172
	CrisisFACTS-008-r6	Sep 14th	26	195	74	27	22	-	318
	CrisisFACTS-008-r7	Sep 15th	30	156	168	19	21	-	364
	CrisisFACTS-008-r8	Sep 16th	72	142	165	29	64	-	400
	CrisisFACTS-008-r9	Sep 17th	19	100	94	22	17	-	233
		Total	1,389	2,880 (32.6%)	3,961 (44.9%)	1,124 (12.7%)	710 (8.0%)	137 (1.6%)	8,822

Manual Evaluation Setup: Fact Matching

To perform the matching process itself, the human assessor was presented with one participant's item at a time in chronological order determined by item timestamp. The assessment interface is shown in Figure 6 and is divided into two columns. The left-hand column contains renderings of the participant system-item, with annotation controls below this rendering and a listing of fact matches the assessor has already selected. In the right-hand column we have the fact search bar and the fact list. Assessors can click any fact on the right to add it to the fact matches rendered on the left, and vice versa. Once the assessor is finished with an item, they can use the annotation controls to move to the next item. Pressing 'Contains no Info' will mark the item as irrelevant, 'Contains Info No Match' marks the item as no match found, 'Skip Item' will add the item to the users skip queue (to which they can return later if desired), and 'Next Item' triggers the loading of the next item.

The fact list shown to the assessor for an item is the fact list for the event. These fact lists can be quite long, with the largest being CrisisFACTS-001 with 404 facts. It would be very difficult and time-consuming for an assessor to perform matching against such a large list without assistance, so we provide an automatic fact-ranking capability within the interface. Two types of search model are provided, namely: 1) Exact, which performs a boolean OR search over the fact texts; and 2) Semantic, which issues a search against a ColBERT (Khattab and Zaharia 2020) index of the facts for the query. By default, when the assessor loads a new item to assess, candidate facts are ranked by their ColBERT similarity scores to the item. The green colouring in the left-had edge of each fact indicates degree of similarity between item and fact. The assessor can also manually enter queries or highlight text from the item and press the 'H-Search' button to search the facts for the highlighted text. We note that as assessors may overly rely on this assistance, which may bias matches.

Table 5 reports the fact matching statistics for each <event,day> pair. The Fact Matches column reports the number of facts that were matched to at least one item, and the proportion of facts from that event that this represented. The Fact Stream Coverage columns report the proportion of the facts that were matched to at least one item that came from each stream or the abstactive runs. The 'All' columns denote the proportion of the facts that were matched to at least one item from that stream, while the 'Uniq.' columns denote the proportion of the facts that were matched to at least one item that were unique to that stream (did not appear in other streams.

As we can see from Table 5, the news and twitter streams are the most likely sources for relevant content, with 54.3% and 66.1% of facts found being present in those streams on average. We also note that these two streams were also contained a sizable proportion of facts not found elsewhere (23.3% and 32.2% on average, respectively). Both Reddit and Facebook streams only provided limited fact coverage on average (12.9% and 3.3%), and generally were redundant with respect to information from other sources. Finally, we note that the assessed items from the abstractive runs appear to have good coverage and find relevant content not surfaced by the more numerous extractive systems, although we note that the sample size is small and for only 3 events.

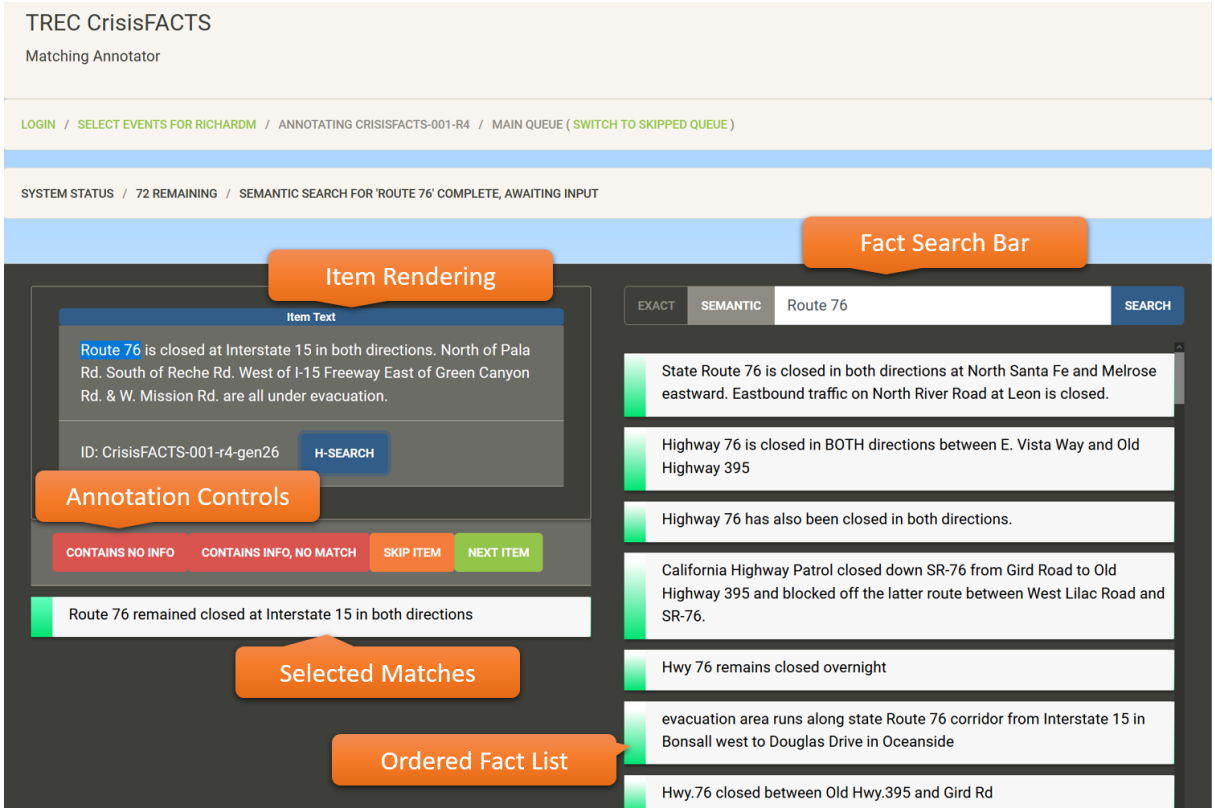


Figure 6. CrisisFACTS 2022 Matching Interface

PARTICIPANT EVALUATION

Above, we have outlined the CrisisFACTS framework, its data sources, and our approach to evaluation. Here, we briefly describe outcomes from the CrisisFACTS inaugural year at TREC.

Summary-to-Summary Similarity

Evaluating summarization performance is a well-researched space with multiple performance metrics, two of which we make use of here: "Recall-Oriented Understudy for Gisting Evaluation", or ROUGE, and BERTScore (Zhang et al. 2020). While ROUGE is commonly used to assesses n-gram overlap between a candidate and reference summary, thus motivating our use of it here, its reliance on exact matching is problematic in the CrisisFACTS context, as social media in general–and Twitter in particular–is stylistically distinct from Wikipedia text, news articles, and professional writing. BERTScore mitigates this issue by assessing summaries via BERT-based contextual embeddings.

Assessing Ground-Truth Summaries

Before evaluating participant performance with these metrics, however, we first use them to compare our three ground-truth summary sources as a robustness check on their quality. By taking pairs of summaries from Wikipedia, the ICS-209-PLUS dataset, and TREC-assessor fact lists and aggregating each into an event-level summary, we can evaluate how well these ground-truth summaries reflect each other, thereby providing context for participant metrics.

Table 6 shows the ROUGE-2 and BERTScore metrics, averaged across each crisis event, for each pair of gold-standard summaries. For ROUGE-2, one can see stronger agreement between NIST-produced fact lists and the ICS 209 summaries,. For BERTScores, TREC assessors' fact-lists perform about equivalently well between ICS 209 and Wikipedia summaries. These results also show that the TREC assessors produce reasonable summaries that perform better at recovering ICS 209-based summaries than Wikipedia, in all cases.

Assessing Participant Summaries

For each participant, we combine the top-k highest priority facts for each event-day pair–where k corresponds to the number of facts in the TREC-assessor fact list for that event-day, as we do with pooling–into a single document,

Table 5. CrisisFACTS 2022 Matching Statistics.

Event	Request ID	Date	Facts Matched	Fact Stream Coverage									
				News		Tweets		Reddit		Facebook		Abstractive	
				All	Uniq.	All	Uniq.	All	Uniq.	All	Uniq.	All	Uniq.
Lilac Wildfire 2017	CrisisFACTS-001-r3	Dec 7th	253 (62.6%)	51.4%	9.1%	89.3%	40.3%	4.0%	0.0%	16.2%	0.8%	-	-
	CrisisFACTS-001-r4	Dec 8th	160 (39.6%)	31.9%	10.0%	86.9%	59.4%	0.0%	0.0%	18.1%	2.5%	-	-
	CrisisFACTS-001-r5	Dec 9th	79 (19.6%)	34.2%	22.8%	77.2%	65.8%	0.0%	0.0%	0.0%	0.0%	-	-
	CrisisFACTS-001-r6	Dec 10th	82 (20.3%)	39.0%	22.0%	62.2%	37.8%	0.0%	0.0%	24.4%	15.9%	-	-
	CrisisFACTS-001-r7	Dec 11th	49 (12.1%)	53.1%	26.5%	71.4%	44.9%	0.0%	0.0%	2.0%	2.0%	-	-
Cranston Wildfire 2018	CrisisFACTS-002-r1	July 25th	29 (58.0%)	82.8%	20.7%	51.7%	3.4%	0.0%	0.0%	6.9%	3.4%	44.8%	10.3%
	CrisisFACTS-002-r2	July 26th	35 (70.0%)	62.9%	0.0%	85.7%	25.7%	8.6%	5.7%	0.0%	0.0%	14.3%	5.7%
	CrisisFACTS-002-r4	July 28th	26 (52.0%)	38.5%	23.1%	69.2%	57.7%	0.0%	0.0%	7.7%	3.8%	7.7%	0.0%
Holy Wildfire 2018	CrisisFACTS-003-r5	Aug 6th	69 (40.1%)	84.1%	49.3%	46.4%	14.5%	4.3%	1.4%	0.0%	0.0%	-	-
	CrisisFACTS-003-r6	Aug 7th	88 (51.2%)	30.7%	8.0%	78.4%	47.7%	29.5%	13.6%	0.0%	0.0%	-	-
	CrisisFACTS-003-r7	Aug 8th	83 (48.3%)	85.5%	51.8%	48.2%	14.5%	1.2%	0.0%	0.0%	0.0%	-	-
	CrisisFACTS-003-r8	Aug 9th	63 (36.6%)	65.1%	27.0%	73.0%	34.9%	0.0%	0.0%	0.0%	0.0%	-	-
Hurricane Florence 2018	CrisisFACTS-004-r13	Sep 9th	34 (17.6%)	64.7%	50.0%	47.1%	32.4%	5.9%	0.0%	0.0%	0.0%	-	-
	CrisisFACTS-004-r14	Sep 10th	92 (47.7%)	53.3%	8.7%	62.0%	17.4%	70.7%	20.7%	0.0%	0.0%	-	-
	CrisisFACTS-004-r15	Sep 11th	93 (48.2%)	55.9%	20.4%	68.8%	34.4%	18.3%	8.6%	0.0%	0.0%	-	-
	CrisisFACTS-004-r16	Sep 12th	75 (38.9%)	37.3%	22.7%	61.3%	49.3%	21.3%	9.3%	0.0%	0.0%	-	-
	CrisisFACTS-004-r17	Sep 13th	55 (28.5%)	52.7%	34.5%	43.6%	27.3%	23.6%	20.0%	0.0%	0.0%	-	-
	CrisisFACTS-004-r18	Sep 14th	28 (14.5%)	53.6%	32.1%	67.9%	42.9%	3.6%	0.0%	0.0%	0.0%	-	-
Maryland Flood 2018	CrisisFACTS-005-r3	May 27th	40 (57.1%)	62.5%	22.5%	32.5%	12.5%	2.5%	0.0%	0.0%	0.0%	55.0%	22.5%
	CrisisFACTS-005-r4	May 28th	31 (44.3%)	12.9%	3.2%	80.6%	38.7%	0.0%	0.0%	0.0%	0.0%	58.1%	16.1%
Saddleridge Wildfire 2019	CrisisFACTS-006-r5	Oct 11th	48 (37.5%)	25.0%	0.0%	77.1%	37.5%	4.2%	0.0%	0.0%	0.0%	56.2%	18.8%
	CrisisFACTS-006-r6	Oct 12th	34 (26.6%)	29.4%	2.9%	73.5%	14.7%	38.2%	0.0%	0.0%	0.0%	61.8%	17.6%
Hurricane Laura 2020	CrisisFACTS-007-r13	Aug 27th	78 (60.5%)	6.4%	1.3%	92.3%	62.8%	33.3%	6.4%	0.0%	0.0%	-	-
	CrisisFACTS-007-r14	Aug 28th	47 (36.4%)	89.4%	66.0%	31.9%	10.6%	0.0%	0.0%	2.1%	0.0%	-	-
Hurricane Sally 2020	CrisisFACTS-008-r4	Sep 12th	34 (19.7%)	17.6%	0.0%	97.1%	64.7%	32.4%	0.0%	0.0%	0.0%	-	-
	CrisisFACTS-008-r5	Sep 13th	95 (54.9%)	73.7%	30.5%	62.1%	22.1%	14.7%	1.1%	0.0%	0.0%	-	-
	CrisisFACTS-008-r6	Sep 14th	115 (66.5%)	99.1%	57.4%	32.2%	0.9%	20.9%	0.0%	0.0%	0.0%	-	-
	CrisisFACTS-008-r7	Sep 15th	131 (75.7%)	85.5%	18.3%	80.9%	13.0%	16.8%	0.8%	0.8%	0.0%	-	-
	CrisisFACTS-008-r8	Sep 16th	113 (65.3%)	64.6%	25.7%	67.3%	26.5%	16.8%	2.7%	10.6%	0.9%	-	-
	CrisisFACTS-008-r9	Sep 17th	90 (52.0%)	85.6%	31.1%	64.4%	12.2%	16.7%	1.1%	10.0%	0.0%	-	-
			Average	54.3%	23.3%	66.1%	32.2%	12.9%	3.0%	3.3%	1.0%	-	-

Table 6. Comparing Pairs of Gold-Standard Summaries. Results show stronger agreement between the ICS 209 reports from actual emergency managers and the TREC-produced fact lists compared to either ICS-209-PLUS or NIST and Wikipedia summaries.

	ICS 209	NIST	Wiki
ICS 209	–	0.5134	0.4885
NIST	0.5134	–	0.5368
Wiki	0.4885	0.5368	–

(a) BERTScore

	ICS 209	NIST	Wiki
ICS 209	–	0.0430	0.0078
NIST	0.0430	–	0.0356
Wiki	0.0078	0.0356	–

(b) ROUGE-2 F1

representing that participant's event summary. We then calculate ROUGE and BERTScore metrics between that summary and the three ground-truth summaries, with results for each participant in Table 7, averaged over all eight crisis events. We exclude ICS-209-based summaries for CrisisFACTS-005, however, as the Maryland floods do not have a corresponding report in the ICS-209-PLUS dataset.

Three major takeaways can be seen from Table 7: First, the top-ranked systems tend to perform well across all three target summaries and both metrics. Second, the two organiser-developed baselines outperform the majority of runs. Third, the maximum values in each column of Table 7 outperform the gold-standard summaries in the corresponding target; that is, the top participant systems produce better summaries, on average, than the other gold-standard summaries.

Fact Matching Evaluation

Moving to the Temporal Summarization-inspired match-based evaluation, we discuss results from TREC assessors' manually matching facts returned by participants to facts in the ground-truth fact list.

Matching Metrics

Given the output from a particular participant system and an <event,day> pair, we ordered the system's output facts by their importance scores (as provided by the participant), against taking the top-k facts as the summary for that day. We denote this set of facts on day d as S_d. After pooling these facts S_d facts, TREC assessors match a sample against the fact list F. This matching process outputs $SystemItem$ -> $[fact1, fact2, ...]$ relationships. To measure effectiveness, we assign a 'gain' for each unique fact $f \in F$ covered by the participant system-items for each day and divide by the ideal gain of a hypothetical a system that covered all facts. This metric is the *comprehensiveness* of the

Table 7. Participant Runs Assessed Against ICS-209-PLUS, TREC-, and Wikipedia-provided Summaries. While we see an order of magnitude difference between BERTScores and ROUGE-2 F1 measures, the runs from `ohm_kiz` and `unicamp` perform consistently well across all target summaries and two metrics. Our baselines, perform strongly as well.

	ICS 209		NIST		Wikipedia	
Run	BERTScore	ROUGE-2	BERTScore	ROUGE-2	BERTScore	ROUGE-2
ohm_kiz.BM25_QAcrisis_ILP	0.4432	0.0464	**0.5642**	**0.1471**	0.5448	0.0337
ohm_kiz.BM25_QAasnq_ILP	0.4477	0.0507	0.5628	0.1468	**0.5646**	**0.0362**
ohm_kiz.BM25_Heuristic_ILP	0.4355	0.0469	0.5596	0.1419	0.5331	0.0303
ohm_kiz.ColBERT_ILP	0.4500	0.0497	0.5460	0.1386	0.5423	0.0307
unicamp.NM-2	**0.4591**	**0.0581**	0.5573	0.1338	0.5321	0.0281
unicamp.NM-1	**0.4591**	**0.0581**	0.5573	0.1338	0.5321	0.0281
baseline.run1	0.4432	0.0418	0.5565	0.1326	0.5296	0.0275
baseline.run2	0.4427	0.0428	0.5565	0.1308	0.5274	0.0267
umcp.combsum	0.4331	0.0390	0.5451	0.1296	0.5198	0.0283
IRIT_IRIS.IRIT_IRIS_mean_USE_INeeds	0.4549	0.0370	0.5515	0.1258	0.5183	0.0241
IRIT_IRIS.IRIT_IRIS_mean_USE	0.4389	0.0344	0.5540	0.1256	0.5328	0.0295
umcp.mrr_sum	0.4372	0.0429	0.5510	0.1245	0.5229	0.0270
umcp.rr_now	0.4303	0.0401	0.5482	0.1237	0.5198	0.0271
umcp.mrr_all	0.4383	0.0400	0.5440	0.1216	0.5216	0.0249
umcp.mrr_nobrf	0.4379	0.0388	0.5460	0.1215	0.5229	0.0290
umcp.mrr_main	0.4356	0.0398	0.5434	0.1209	0.5177	0.0246
umcp.mrr_no_dd	0.4439	0.0398	0.5404	0.1196	0.5113	0.0236
eXSum22.eXSum22_submission_02	0.4308	0.0376	0.5223	0.1014	0.4959	0.0259
IISER22.submission_final.json	0.4363	0.0358	0.5095	0.1013	0.4806	0.0282
IISER22.submission_final_4	0.4381	0.0353	0.5257	0.0997	0.5153	0.0275
IISER22.submission_LM_DS_3	0.4376	0.0374	0.5278	0.0995	0.5270	0.0272
eXSum22.eXSum22_submission_01	0.4204	0.0308	0.5362	0.0980	0.4873	0.0253
IISER22.submission_LM_JM_2	0.4516	0.0360	0.5303	0.0857	0.5133	0.0244
SiPEO.nazmultum11	0.4422	0.0131	0.5529	0.0676	0.5194	0.0259
IRIT_IRIS.IRIT_IRIS_tssubert	0.4301	0.0144	0.5514	0.0651	0.5071	0.0322

summary content, as shown in Eq. 1, where $M(f, S)$ is the set of system-items ($[i^1, i^2, ...]$) in S that matched fact f, and $R(f)$ is the gain assigned to the fact f. For CrisisFACTS 2022, this gain is always 1 and directly equivalent to fact recall–though future iterations could weight facts differently, e.g., based on type, priority, or another factor.

$$Comprehensiveness(S_d) = \frac{1}{\sum_{f \in F} R(f)} \sum_{\{f \in F : M(f,S) \neq \emptyset\}} R(f) \tag{1}$$

Similarly, we measure a participant system's redundancy in S_d by counting unique facts matched and dividing by the total number of fact matches present, as shown in Eq. 2. For all runs, we macro-average these values across days within an event and then average again across all events. This averaging ensures each day has the same weight in its event, and all events have the same weight regardless of their size.

$$RedundancyRatio(S_d) = \frac{\sum_{\{f \in F : M(f,S) \neq \emptyset\}} R(f)}{\sum_{\{f \in F\}} R(f) \cdot |M(f, S)|} \tag{2}$$

Participant Results

Table 8 reports match-based performance for each participant run, sorted by comprehensiveness. This table includes only extractive runs, as an issue emerged after pooling abstractive runs, that leads to non-comparable results in manual annotation. We summarize the main results below:

Assessed@k Before covering the main metrics, we first discuss Assessed@k, which is a measure of confidence in the accuracy of the main metrics. It reports the percent of participant-system-items that were actually assessed for matches by TREC assessors (macro averaged across days and events). The lower this value, the more uncertainty the results, as more items are being excluded from assessment. Runs with higher Assessed@k may also be advantaged, as having more assessed facts increases the probability of a match. For instance, while baseline.run2 has the highest reported comprehensiveness, it also had over 7% more items assessed than the next best run ohm

kiz.BM25_QAasnq_ILP, meaning that we should not draw a strong conclusion about the ordering of those systems. As we can see from Table 8, Assessed@k values average around 60% for most runs, with lows of 33% and highs of 70%. For most runs the spread is quite low, and so should be comparable, however we should be aware that there may be positive or negative bias when comparing runs outwith a +/-0.05 Assessed@k range. In these cases we should look to the matching data to see what affect the disparity in number of assessed items is likely having.

Comprehensiveness Comprehensiveness represents a summary's fact-recall, with higher values being better. Table 8 shows particiapnts 'ohm kiz' and 'SiPEO' achieve the highest comprehensiveness (20-21%). This metric means that, on any day, these participants' summaries covered around 20% of all facts for the event. While this value might seem low, CrisisFACTS events span multiple days, so a system's coverage will accumulate across days during the event. Additionally, the organiser baselines runs (which are simply relevance-focused search systems using the provided query set), were highly effective.

Redundancy Ratio The redundancy ratio measures how often a participant summary repeats information, with lower values being better. Lower redundancy is better, as we don't want to waste our reader's time. As we can see from Table 8, the redundancy ratio is moderate, ranging between 17-30%. For analysis, we should always consider the redundancy ratio within the context of comprehensiveness, as the more facts a system covers, the more likely it is to repeat information. For example, IRIT_IRIS.IRIT_IRIS_tssubert achieved the lower redundancy, but this participant also had the lowest comprehensiveness, meaning that it was a-priori less likely to repeat content. For this reason, participants should only compare redundancy ratios between runs with similar comprehensiveness scores. For example, we can confidently conclude that the ohm kiz.BM25_QAcrisis_ILP run is markedly better than baseline.run1, as the comprehensiveness scores are very close, but ohm kiz.BM25_QAcrisis_ILP has much less redundancy.

Table 8. Match-based run evaluation performances

		Main Metrics		Matching Data			
Run	Assessed@k	Comprehensiveness	Redundancy Ratio	Matched %	Skipped %	Irrelevant %	Unmatched %
baseline.run2	0.6488	**0.2165**	0.2783	0.3510	0.0071	0.2885	0.3534
ohm kiz.BM25_QAasnq_ILP	0.5748	0.2129	0.2263	0.3959	0.0025	0.2393	0.3623
ohm kiz.BM25_QAcrisis_ILP	0.591	0.2101	0.2259	0.3510	0.0006	0.2264	0.4221
baseline.run1	0.6155	0.2094	0.3358	**0.4249**	0.0086	0.2165	0.3501
SiPEO.nazmultum11	0.6617	0.2045	0.2583	0.2616	0.0038	0.4060	0.3286
ohm kiz.ColBERT_ILP	0.568	0.1893	0.2011	0.296	0.0000	0.3268	0.3772
eXSum22.eXSum22_submission_02	0.6725	0.1864	0.3001	0.2249	0.0013	0.5275	0.2464
umcp.mrr_nobrf	0.5806	0.1862	0.1726	0.3226	0.0028	0.3778	0.2968
ohm kiz.BM25_Heuristic_ILP	0.5664	0.1849	0.1761	0.3548	0.0009	0.313	0.3312
umcp.mrr_all	0.5919	0.1700	0.2389	0.3331	0.0034	0.3451	0.3184
IISER22.submission_LM_DS_3	0.6615	0.1665	0.1725	0.2221	0.0005	0.4989	0.2785
umcp.mrr_main	0.5883	0.1660	0.2187	0.334	0.0034	0.3650	0.2977
umcp.combsum	0.6053	0.1642	0.2406	0.3995	0.0042	0.2873	0.309
umcp.mrr_no_dd	0.5983	0.1633	0.2553	0.3443	0.0034	0.3575	0.2948
eXSum22.eXSum22_submission_01	0.4693	0.1607	0.3618	0.4947	0.0004	0.2295	0.2754
IISER22.submission_final.json	0.6329	0.1599	0.1758	0.2216	0.0005	0.4948	0.2832
IISER22.submission_final_4	0.6329	0.1599	0.1758	0.2216	0.0005	0.4948	0.2832
IRIT IRIS.IIRIT_IRIS_mean_USE	0.4288	0.1547	0.1952	0.3887	0.0000	**0.2260**	0.3853
IISER22.submission_LM_JM_2	0.7112	0.1522	0.2219	0.1865	0.0005	0.5562	0.2568
umcp.mrr_sum	0.4903	0.1505	0.1924	0.3415	0.0008	0.3272	0.3305
umcp.rr_now	0.5605	0.1487	0.2547	0.3357	0.0005	0.3641	0.2998
IRIT IRIS.IIRIT_IRIS_mean_USE_INeeds	0.4521	0.1409	0.2851	0.3162	0.0083	0.3174	0.3581
IRIT IRIS.IIRIT_IRIS_tssubert	0.3369	0.1275	**0.1247**	0.2034	0.0071	0.4033	0.3862

Matching Data We now turn to the 'Matching Data' metrics In Table 8 and how they can improve interpretation of the main metrics. The 'Matched %' column is the proportion of system-items that matched at least one fact. 'Skipped %' is the proportion of system-items the assessor skipped (usually because of a rendering issue with the item). 'Irrelevant %' is the proportion of items the assessor marked as lacking relevant information (with respect to the response officer use-case). 'Unmatched %' is the proportion of items the assessor tagged as appearing to have some relevant information but did not match anything in the fact list.

Regarding their implications, we note that 'Skipped %' is very low, demonstrating that TREC assessors rarely skipped items. 'Irrelevant %' is useful, as it tells participants whether they should focus more on removing off-topic content or find more facts as directions for improvement. The best-performing systems have comparably low 'Irrelevant %' values compared to the majority of the mid-table participants, suggesting this factor is one of the reasons these good systems perform well.

Considering 'Matched %' and 'Unmatched %' together, these measures can be summed to produce the percent of assessed items that the assessor thought were relevant. Moreover, 'Matched %' can be compared with Comprehensiveness to analyse the density of facts within a participant system. A high 'Matched %' but a low Comprehensiveness value–as with eXSum22.eXSum22_submission_01–suggest the participant system is producing a set of facts that concentrates on some subset of the total fact list. 'Unmatched %' can also be viewed as an inverse to Assessed@k, where high values of unmatched facts suggest more uncertainty in the Comprehensiveness and Redundancy Ratio calculations. In CrisisFACTS 2022, 'Unmatched %' is high, suggesting participant systems were returning facts not covered by the fact list; this result is a potential consequence of preventing TREC assessors from creating new facts during fact-matching assessment. Future CrisisFACTS iterations will enable assessors to define new facts during matching, which will control for this possibility.

Comparing Across Metrics

Despite the plethora of metrics we have (various versions of ROUGE, BERTScore, comprehensiveness, and others), a key question is how consistent these metrics are in their rankings of system performance. Figure 7 depicts the pairwise Spearman rank correlations across two versions of Comprehensiveness (accumulative and isolated) and F1 measures from ROUGE across several n-gram lengths.

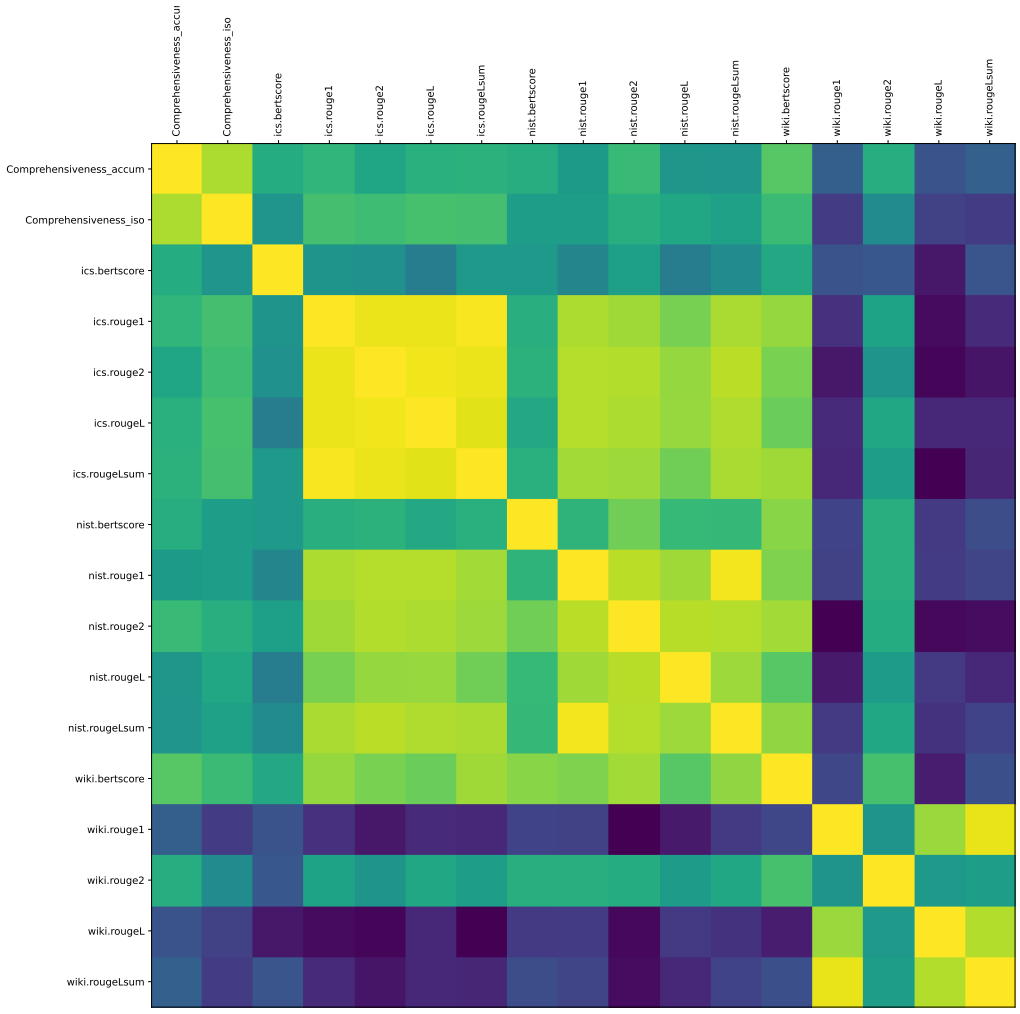


Figure 7. Spearman Rank Correlations Among Metrics. Results suggest scores generated from the NIST and ICS 209 target summaries see the maximum off-axis correlations. The manual fact-matching comprehensiveness metrics are also decently well correlated with the ICS and NIST-based metrics.

Studying these correlations reveal several findings: First, rankings from ROUGE metrics, regardless of n-gram length, are generally stable, often exhibiting Spearman $\rho \geq 0.9$. Second, runs that perform well against the ICS 209 summaries in terms of ROUGE are also highly likely to perform well against the TREC assessors' fact-list summaries (average Spearman $\rho = 0.7936$). Third, manual, fact-matching metrics exhibit positive, weak-to-moderate correlation with the summary-based metrics. Lastly, evaluations against the Wikipedia-based summaries consistently deviate from the manual-, ICS-, and NIST-based rankings, with the exception of the ROUGE-2 metric, which is consistently weakly positive across the other targets. This deviation suggests that summaries from Wikipedia capture a fundamentally different set of information, which is perhaps unsurprising given the different audiences–emergency-response personnel versus the general public.

CONCLUSION

In this paper, we have introduced a new data challenge – CrisisFACTS – that ran its first pilot edition in 2022. CrisisFACTS provides a new test collection and evaluation forum for researchers and practitioners interested in developing and evaluating automated technologies for extracting and summarizing online content posted during emergencies. In its first year, a new crisis corpus was collected and released for the community, comprising News, Twitter, Reddit and Facebook content for 8 crisis events. Furthermore, three types of ground-truth summaries (derived from ICS-209 forms, Wikipedia and manual news analysis, respectively) are included, in addition to manually assessed 'matchings' between items included within a participant summary and a gold-standard fact list, both of which can be used to evaluate the quality of crisis summaries. 23 participant systems were submitted to the first edition from 8 research groups from around the world. Summarization performance was reasonably effective for this first year, with between 70-80% of content returned being assessed as relevant, although much information is still being missed, with reported summary comprehensiveness of the best systems of around 20-21%.

REFERENCES

Allan, J., Gupta, R., and Khandelwal, V. (2001). "Temporal summaries of new topics". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 10–18.

Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., and Sakai, T. (2015). *Trec 2014 temporal summarization track overview*. Tech. rep. NATIONAL INST OF STANDARDS and TECHNOLOGY GAITHERSBURG MD.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (May 2020). "The Pushshift Reddit Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1, pp. 830–839.

Buckley, C. and Voorhees, E. M. (2004). "Retrieval evaluation with incomplete information". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 25–32.

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

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

Denis, L. S., Mietkiewicz, N., Short, K., Buckland, M., and Balch, J. (Dec. 2022). "ICS-209-PLUS - An all-hazards dataset mined from the US National Incident Management System 1999-2014". In.

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.

Dutta, S., Chandra, V., Mehra, K., Ghatak, S., Das, A. K., and Ghosh, S. (2019). "Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms". In: *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*. Springer, pp. 859–872.

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*.

Khattab, O. and Zaharia, M. (2020). "Colbert: Efficient and effective passage search via contextualized late interaction over bert". In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48.

Li, C., Xu, W., Li, S., and Gao, S. (2018). "Guiding generation for abstractive text summarization based on key information guide network". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 55–60.

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.

MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., and Goharian, N. (2021). "Simplified Data Wrangling with ir_datasets". In: *SIGIR*.

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

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

Munot, N. and Govilkar, S. S. (2014). "Comparative study of text summarization methods". In: *International Journal of Computer Applications* 102.12.

Nenkova, A. and McKeown, K. (2012). "A survey of text summarization techniques". In: *Mining text data*, pp. 43–76.

Ofli, F., Alam, F., and Imran, M. (May 2020). "Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response". In: *17th International Conference on Information Systems for Crisis Response and Management*. ISCRAM. ISCRAM.

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*.

Pavlu, V., Rajput, S., Golbus, P. B., and Aslam, J. A. (2012). "IR system evaluation using nugget-based test collections". In: *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 393–402.

Reuter, C. and Kaufhold, M.-A. (2018). "Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics". In: *Journal of Contingencies and Crisis Management* 26.1.

Robertson, S. (2008). "On the history of evaluation in IR". In: *Journal of Information Science* 34.4, pp. 439–456.

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.

Rudra, K., Banerjee, S., Ganguly, N., Goyal, P., Imran, M., and Mitra, P. (2016). "Summarizing situational tweets in crisis scenario". In: *Proceedings of the 27th ACM conference on hypertext and social media*, pp. 137–147.

St. Denis, L. A., Mietkiewicz, N. P., Short, K. C., Buckland, M., and Balch, J. K. (Feb. 2020). "All-hazards dataset mined from the US National Incident Management System 1999–2014". In: *Scientific Data* 7.1, p. 64.

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

Voorhees, E. M., Harman, D. K., et al. (2005). *TREC: Experiment and evaluation in information retrieval*. Vol. 63. Citeseer.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). "BERTScore: Evaluating text generation with BERT". In: *ICLR*.