Crisis Event Social Media Summarization with GPT-3 and Neural Reranking

Jayr Pereira*
NeuralMind, Brazil
Centro de Informática, Universidade Federal de Pernambuco, Brazil
jayr.pereira@neuralmind.ai

Robson Fidalgo
Centro de Informática, Universidade Federal de Pernambuco, Brazil
rdnf@cin.ufpe.br

Roberto Lotufo
NeuralMind, Brazil
roberto@neuralmind.ai

Rodrigo Nogueira
NeuralMind, Brazil
rodrigo.nogueira@neuralmind.ai

ABSTRACT

Managing emergency events, such as natural disasters, requires management teams to have an up-to-date view of what is happening throughout the event. In this paper, we demonstrate how a method using a state-of-the-art open-sourced search engine and a large language model can generate accurate and comprehensive summaries by retrieving information from social media and online news sources. We evaluated our method on the TREC CrisisFACTS challenge dataset using automatic summarization metrics (e.g., Rouge-2 and BERTScore) and the manual evaluation performed by the challenge organizers. Our approach is the best in comprehensiveness despite presenting a high redundancy ratio in the generated summaries. In addition, since all pipeline components are few-shot, there is no need to collect training data, allowing us to deploy the system rapidly. Code is available at https://github.com/neuralmind-ai/visconde-crisis-summarization.

Keywords

Crisis Management, Social Media, Multi-document Summarization, Query-based Summarization.

INTRODUCTION

In the information age, managing emergency events, such as natural disasters, requires management teams to communicate efficiently with stakeholders (e.g., the press, government, and the general public). To perform this role, they must be aware of as many facts as possible throughout the event to produce reliable and up-to-date reports. In addition, these reports must meet the needs of people affected or potentially affected by the disaster, which can be quite diverse and distributed over a large area. However, it is humanly impossible to have, for example, a crisis agent at each relevant point affected by a wildfire. In these cases, counting on information from the general public and independent news sources may be an alternative to staying up-to-date.

Nowadays, relevant information about the crisis flows continuously from different sources. Social media and online news sources can offer decision-makers access to up-to-date and relevant information during crisis events. When phone lines are congested due to thousands of people trying to call simultaneously, online social media like Facebook and Twitter can serve the people (Saroj and Pal 2020). One can use social media to seek and share information about food, shelter, transportation, roads, airport situations, etc. In such cases, social media can provide access to timely and relevant information that can empower decision-makers in crisis management. It may remain online even when other forms of communication are affected by disasters. For example, radio and television can be affected if

*corresponding author

electricity is cut off or disconnected, while social media may stay online (Saroj and Pal 2020). Besides, it is widely used due to the increased use of mobile technology. However, the vast volume and variety of data generated from social media can make it difficult for crisis management teams to quickly extract relevant information.

Recent research used social media as a source for information extraction and summarization (Saroj and Pal 2020). The produced summaries can be used to compose reports to keep crisis managers up-to-date and help them make effective decisions promptly. Recent proposals for summarizing disaster-specific social media content rely on a two-step pipeline for document (e.g., tweet) classification and summarization. The classification step involves assigning documents into classes related to user information needs (e.g., situational classes, such as affected regions or airport status). The summarization step involves generating text summaries using the documents from each class. This approach can be classified as topic-based multi-document summarization. Previous studies in this task used clustering algorithms (Kedzie et al. 2015) or supervised classifiers (Rudra, Banerjee, et al. 2016; Rudra, Goyal, Ganguly, Mitra, et al. 2018; Rudra, Goyal, Ganguly, Imran, et al. 2019; Nguyen, Shaltev, et al. 2022) for classification and summarization. However, recent advances in machine learning promoted by pre-trained large language models (LLM) such as GPT-3 (Brown et al. 2020) opened horizons to using few- or zero-shot strategies, which require no or few annotated data.

This paper presents a method that utilizes a search engine and a Language Model (LM) to generate precise and comprehensive summaries of an event situation at a particular time. The proposed method performs a query-based multi-document summarization by leveraging a state-of-the-art search strategy based on NeuralSearchX (Almeida et al. 2022) to retrieve relevant documents catering to the general public or crisis managers' information needs. NeuralSearchX is a metasearch engine based on a multi-purpose large reranking model that uses a single component to merge results from multiple sources. The engine retrieves and reranks documents and outputs a list of items in descending order of their relevance to the query. We summarize the top-k documents using GPT-3 with few-shot learning and Chain of Thought (CoT; Wei et al. 2022). This way, our approach combines the power of NeuralSearchX searching with the contextual understanding of GPT-3 to generate accurate, informative and comprehensive summaries.

We evaluated our method in the TREC CrisisFACTS 2022 challenge dataset using standard summarization metrics (e.g., Rouge-2 and BERTScore) and the manual evaluation performed by track organizers. The results demonstrate that our method performs well in both evaluations. In the automatic evaluation, our method is in the top 3. In the manual assessment, our approach is the best in comprehensiveness despite presenting a high redundancy ratio in the generated summaries. In addition, since all pipeline components are zero-shot or few-shot, there is no need for training data collection, allowing us to deploy the system rapidly. This is a desirable feature for applications where the underlying data changes constantly.

RELATED WORK

In this section, we detail the works related to ours. We divide it into two parts: 1) works which propose using social media summarization for crisis management and 2) works which propose methods for query-based multi-document summarization.

Social Media Summarization for Crisis Management

Social media can offer decision-makers access to up-to-date and relevant information during crisis events. Unlike earth's observational data, which rely on satellites' orbital positioning, or forecasts that can be inaccurate due to a lack of observational data, social media information is available at all times and in real-time (Lorini et al. 2021). Furthermore, social media can handle large amounts of traffic, remain online, and serve as a medium of communication even if electricity is cut off or disconnected, which often affects radio and television communications (Saroj and Pal 2020). The increased usage of smartphones and other mobile technology has caused social media to become a widely used platform across the globe. Therefore, social media can be a valuable source of information during crises and emergencies (Lorini et al. 2021; Saroj and Pal 2020; Phengsuwan et al. 2021). However, the vast volume and variety of data generated by social media can make it challenging for crisis management teams to process it and extract relevant information quickly.

Different approaches have been proposed in the literature to cope with the challenges of using social media for crisis management. According to Saroj and Pal 2020, most research in this field is based on information extraction, summarization, and event classification techniques. The information summarization can help management teams make effective decisions quickly.

Recent studies used different methods to summarize disaster-specific social media content. Most use a pipeline based on two main steps: classification and summarization. The classification step consists of assigning a tweet, for

example, into classes regarding the user or crisis event manager's interest (e.g., situational classes, like affected regions). The summarization step consists of generating summaries based on these classes. Kedzie et al. 2015, for example, performs an unsupervised classification by predicting the salience of documents concerning a crisis event and integrates these predictions into a clustering-based multi-document summarization system. Rudra, Goyal, Ganguly, Mitra, et al. 2018 presented an approach to identify sub-events and generate summaries from them. They used a summarization algorithm based on an Integer Linear Programming (ILP) technique. Rudra, Goyal, Ganguly, Imran, et al. 2019 proposed a classification-summarization framework that assigns tweets into different situational classes and then summarizes them. They used the AIDR classifier (Imran et al. 2014) for tweet classification. A two-step framework was proposed for summarizing tweets. In the first step, important tweets were extracted, and a word graph was constructed. In the second step, an optimization technique based on ILP was used to select the most important tweets and word-graph paths based on informativeness and coverage of content words. Finally, Nguyen, Shaltev, et al. 2022 presented CrisICSum, a platform for classifying and summarizing crisis events' tweets. They propose the BERT2BERT model to classify tweets into different humanitarian classes (i.e., damage, affected people, etc.) and RATSUM (Nguyen and Rudra 2022) for summarization.

Our proposal differs from those cited above by recasting the classification step as a search problem and using few-shot learning for summarization. For both classification and summarization, no domain-specific annotated data is required.

Query-based Multi-Document Summarization

Multi-document summarization (MDS) aims to generate a concise and informative summary from a cluster of topic-related documents. It is a more complex and challenging task than single-document summarization. The goal is to identify the correct text for a summary, remove redundancy, ensure coherence and completeness, and provide novelty. Multi-document summarization has inspired web-based clustering systems, such as news aggregators (Tas and Kiyani 2017). Query-based or query-oriented MDS (qMDS) aims to generate summaries from multiple documents driven by a user interest query (C. Ma et al. 2022; Kulkarni et al. 2020; Abdi et al. 2017). qMDS combines information retrieval and MDS techniques to summarize multiple documents according to the user's needs. The generated summary is sometimes called a query-focused, topic-focused, or user-focused summary (Gambhir and Gupta 2017).

Recent studies used different methods to perform qMDS. Wu et al. 2019 proposed an unsupervised pattern-enhanced approach for representing documents and queries. This approach employs a pattern-enhanced topic model to generate discriminative and semantic-rich representations for topics and documents and a pattern-based relevance model for the query relevance of sentences. Roitman et al. 2020 proposed an unsupervised, query-focused, multi-document extractive summarizer called Dual-CES. This method builds on the Cross-Entropy Summarizer (CES) and is designed to handle the tradeoff between saliency better and focuses on summarization. Lamsiyah et al. 2021 proposed an unsupervised extractive summarization method that leverages transfer learning from pre-trained sentence embedding models to represent documents' sentences and users' queries. The method applies the maximal marginal relevance criterion to re-rank the selected sentences by maintaining query relevance and minimizing redundancy. Chali and Mahmud 2021 proposed an unsupervised extractive summarization method using a reinforcement learning technique and a transformer model. This technique considers the importance of information coverage and diversity under a fixed sentence limit. Popescu et al. 2021 proposed a convex optimization formulation of the extractive text summarization problem and a simple and scalable algorithm to solve it. The optimization program was constructed as a convex relaxation of an integer programming problem, and a specific projected gradient descent algorithm was designed to solve it. Finally, Laskar et al. 2022 proposed a series of domain adaptation techniques using pre-trained transformer models to generate abstractive summaries for the text summarization of single and multiple documents. This included transfer learning, weakly supervised learning, and distant supervision.

These studies demonstrate the diverse range of approaches to qMDS and the ongoing efforts to improve the effectiveness of multi-document summarization. While these qMDS methods have shown promising results, they often require extensive training data, which can be time-consuming and expensive. On the other hand, our proposed method does not require any training data and uses few-shot components, allowing for rapid deployment. Moreover, our approach is well-suited for emergency events, where timely and accurate summaries are critical for effective management.

OUR APPROACH

This section presents our proposed method for generating facts from social media and web news to support crisis event management. The proposed method has two main steps: Document Retrieval and Aggregation, as shown

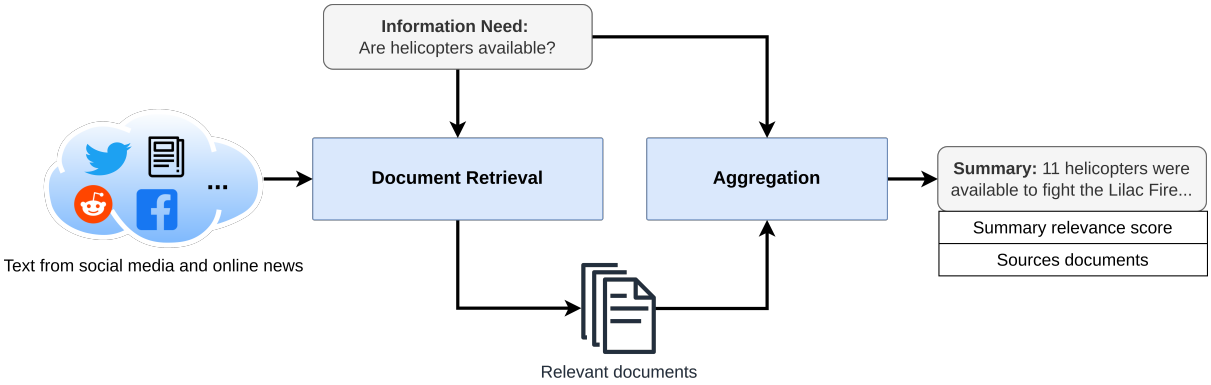

Figure 1. Illustration of the proposed method.

in Figure 1. Our method can be categorized as a query-based multi-document summarization system, as it can summarize text from multiple documents using a query to guide summary generation. Thus, it assumes the preexistence of a set of questions comprising the user information needs. We detail the steps of our approach in the following subsections.

Document retrieval

We use a two-stage pipeline based on NeuralSearchX Almeida et al. 2022 for document retrieval. The first stage consists of retrieving candidate documents using a bag-of-words retriever (i.e., BM25). The second stage consists of reranking candidate documents using a neural reranker. We provide information on implementing these two steps in the following subsections.

Candidate Document Retrieval

This stage aims to retrieve relevant candidate documents given the queries representing user information needs. This is essential to reduce the number of documents processed by the reranking step, which is more computationally expensive. Different search functions can be used to accomplish this goal. We use the Pyserini (Lin, X. Ma, et al. 2021) library implementation of the BM25 algorithm (Robertson et al. 1994).

For performing the experiments presented in this work, we used Pyserini to create a searchable index of the documents provided in the CrisisFACTS challenge dataset. A searchable index is a body of structured data that a search engine refers to when looking for relevant results. Pyserini has a Python interface to the Lucene library [1], which is a high-performance search engine backend. The Lucene library uses the term frequency-inverse document frequency (TF-IDF) weighting scheme for indexing, a widely-used technique for information retrieval that gives more weight to terms that are rare in the collection and less weight to common words.

In real crisis scenarios, the search space can be considerably more significant than the one tested in this work, especially when news websites are used as sources in addition to social media. Maintaining an up-to-date searchable index is challenging in these cases due to the constant information flow. An alternative is replacing the candidate document retrieval step with third-party APIs, such as Bing, Google, or Twitter search. For web search, it is necessary to assume that the snippets returned by the engines are enough to represent the pages' content (Almeida et al. 2022).

Document Reranking

This stage aims to rerank the candidate documents according to the user information needs represented as queries. Given a list of candidate documents from different sources (web news and social media), the reranker should order them according to their relevance to the query. The output of this stage is an ordered list of documents chosen as relevant to each query.

Document reranking consists of training a classifier to estimate the probability of each document being relevant given a question and ordering the documents in decreasing order according to their probabilities (Lin, Nogueira, et al. 2022). Recent studies demonstrated that using Transformers reaches state of the art in this task (Nogueira, Jiang, et al. 2020; Lin, Nogueira, et al. 2022; Nogueira and Cho 2019). Nogueira, Jiang, et al. 2020, for example,

[1] https://lucene.apache.org/

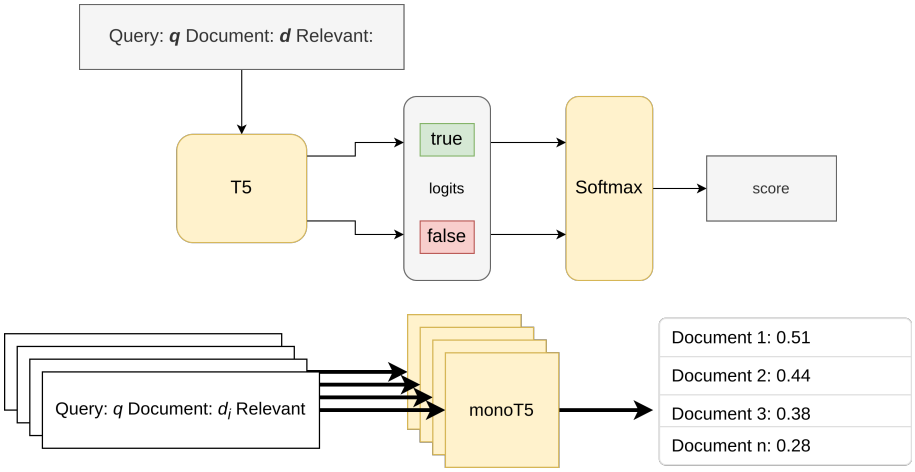

Figure 2. monoT5's training (top) and inference (bottom).

proposed monoT5, an adaptation of the T5 sequence-to-sequence transformer model (Raffel et al. 2020) for reranking. monoT5 is a T5 finetuned to produce the words "true" or "false" whether the document is relevant to the query. As shown in Figure 2 (top), the model receives as input a sequence with the document and the query, and a softmax function is applied only over the logits calculated by T5 to the tokens "true" and "false". The probability of the token "true" is the document relevance score given the question. That is, "true" and "false" are the "target words" (i.e., ground truth predictions in the sequence-to-sequence transformation) (Nogueira, Jiang, et al. 2020). However, only the probability of the word "true" is used at inference time.

Figure 2 (bottom) illustrates how the model reranks a list of documents at inference time. Each query-document pair is independently inputted to the model, and a relevance score is estimated by computing the probability of the token "true" to complete the given sequence. Finally, the documents are ordered in decreasing order according to their relevance scores.

Aggregation

As mentioned before, the method proposed in this paper can be classified as a query-based multi-document summarization system, which involves answering a question using information from multiple documents. The produced answer is generally a sentence or paragraph that summarizes the documents. Recent studies have shown that using large language models (LLMs) as few-shot in-context learners for question answering is an efficient and inexpensive strategy, as only a few labeled training examples are needed (Pereira et al. 2022). Pereira et al. 2022 achieved results over or close to state-of-the-art supervised models using GPT-3 (Brown et al. 2020) for multi-document QA. Before answering, they used a chain-of-thought (CoT) reasoning step, significantly improving LLMs' few-shot performance. CoT can lead to better performances not only in multi-document QA but also on diverse QA benchmarks (Wei et al. 2022; Wang et al. 2022; Kojima et al. 2022; Creswell and Shanahan 2022). We used CoT for summarization with an approach similar to Pereira et al. 2022's. However, we consider the reasoning paragraph generated by the model as the summary instead of the final answer.

Figure 3 presents an example of a prompt used in our method. The prompt has three main characteristics: 1) the header – which we use to explain the task to the model and give instructions; 2) the one-shot example – which has four main elements: the context documents (e.g., [Document 1]), the question, the evidence paragraph that induces CoT, and the final answer; and 3) the target example – which also has context documents and an associated question. When fed with the prompt, the model generates the evidence paragraph and the final answer to the target example. We use the generated evidence paragraph as the summary that will be part of a management report. Note that the model cites the documents supporting the answer in the evidence paragraph, as it was induced to do so by the one-shot example. These references can inform the manager about the source of the information and other metadata of the cited documents, such as publication timestamp and location (if available).

The employed prompt induces the model to answer "Unanswerable" in cases where the information provided in the documents is insufficient to answer the question. In such cases, the produced summary is dismissed as irrelevant or erroneous. An illustration of such a scenario is presented in Figure 4, where the question of whether airports were closed during the Lilac Wildfire of 2017 is attempted to be answered using tweets published on the first event day (2017-12-07). The occurrence of unanswerable queries can be attributed to the absence of relevant documents that

Figure 3. Prompt used in the aggregation step. The bold text is generated by the model; the remaining is the input prompt.

For each example, use the documents to create an "Answer" and an "Evidence" to the "Question". Use "Unanswerable" when not enough information is provided in the documents.

Example 1:

[Document 1]: Giovanni Messe became aide-de-camp to King Victor Emmanuel III, holding this post from 1923 to 1927.

...

[Document 3]: The First World War was global war originating in Europe that lasted from 28 July 1914 to 11 November 1918

Question: How long had the First World War been over when Messe was named aide-de-camp?

Evidence: Giovanni Messe became aide-de-camp in 1923 [Document 1]. The First World War ended in 1918 [Document 3].

Answer: 5 years.

Example 2:

...

[Document 3]: Lilac fire in San Diego County 4,100 acres burned (as of 12 p.m. Tuesday) 92% containment 1,659 firefighters on scene 157 structures destroyed, 64 damaged 10,000 people evacuated 11 helicopters

[Document 4]: Two helicopters continued making water drops after dark, and the Navy has agreed to mobilize military helicopter crews to fight the fire on Friday.

[Document 5]: North County Fire Protection District, Cal Fire and other firefighters from around San Diego County were battling the blaze, using several helicopters, bulldozers and air tankers.

[Document 6]: (Hayne Palmour IV / San Diego Union-Tribune) 9 / 38 A helicopter drops water on flames in the San Luis Rey riverbed in Bonsall.

[Document 7]: Gov. Jerry Brown declared a state of emergency in the county as night-flying helicopters prepared to make water drops.

...

Question: Are helicopters available?

Evidence: **11 helicopters were available to fight the Lilac Fire [Document 3], two helicopters continued making water drops after dark [Document 4], several helicopters were used to battle the blaze [Document 5], a helicopter drops water on flames in the San Luis Rey riverbed [Document 6], night-flying helicopters prepared to make water drops [Document 7].**

Answer: **Yes.**

can answer the user's interest query in a specific time frame. One way to mitigate this issue is to limit the search scope by setting a threshold of document relevance during the search stage. However, determining the optimal values for such hyperparameters requires experimentation and empirical evaluation. It is important to manage unanswerable queries effectively to avoid providing irrelevant or incorrect responses, which can significantly impact user satisfaction and trust in the system.

Summary Relevance

For crisis event management systems, providing the managers a confidence score for each generated summary may be useful as it might help them to make a decision. The ranking stage of the document retrieval component attributes a probability of each document being relevant to the user information need query, which we call relevance score. We calculate the summary relevance by averaging the relevance scores only of the documents referenced in the generated summary. For example, if the relevance scores of the documents referenced in the generated summary shown in Figure 3 are 0.9, 0.8, 0.75, 0.7, and 0.65, respectively, then the summary relevance score is 0.75.

Figure 4. An Unanswerable question example. When not enough information to answer the question is found in the retrieved documents.

...

[Document 1]: Here are the latest updates: Lilac fire in San Diego County Current estimated fire perimeter: Size: 4,100 acres Containment: 92 percent Road closures: All roads have been reopened.

[Document 2]: Advertisement San Luis Rey Downs relief Santa Anita Park, The Stronach Group, and the Del Mar Thoroughbred Club have set up a GoFundMe page with donations going to Lilac Fire recovery efforts.

[Document 3]: Update: All evacuation orders have been lifted (with some restrictions) and roads re-opened as of 4 p.m. Sunday, Dec. 10.

[Document 4]: As of noon Tuesday, Dec. 12, here are the latest facts and figures we have on the blaze, as well as the five other fires that ahve wreaked havoc in Southern California since Dec. 4.

[Document 5]: If you need to evacuate and have no place to go due to the fires please DM me. We live in San Diego. #CaliforniaWildfires

[Document 6]: @LegendaryNeuro Fire is "possible". Schools in East San Diego county are closed tomorrow due to high winds be and fire watch.

[Document 7]: (Figures will be updated as new information becomes available) Fires' impact across Southern California 260,000+ acres have been burned in the Southland.

[Document 8]: The new blazes come as fatigued firefighters in Los Angeles County began to make progress on major fires that together have destroyed or damaged more than 30 homes and prompted the evacuations of more

[Document 9]: On top of California #wildfires + winds, earthquakes tonight have East County San Diego folks on edge. Because of c... https://t.co/ya6qGSKEDz

[Document 10]: Gusty Santa Ana winds that have driven the fire are expected to weaken Friday, though officials noted that the winds were unpredictable and that they expected another challenging day.

Question: Have airports closed?

Evidence: **Unanswerable**
Answer: **Unanswerable**

EXPERIMENT

This section presents the experiment we performed to evaluate the proposed method. This experiment consists of our submission for the 2022 TREC CrisisFACTS Track.[2] The Text REtrieval Conference (TREC)[3] is a conference co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense aiming at supporting research within the information retrieval community. The conference provides the infrastructure for large-scale evaluation of text retrieval methodologies in different tracks. The TREC Crisis Facts and Cross-stream Temporal Summarization (CrisisFACTS) is one of TREC's tracks, which consists of an open data challenge for summarization technologies to support disaster management using online data sources.

Dataset and Task Formulation

CrisisFACTS 2022 provides a multiple-stream dataset with data related to 8 crisis events (cf. Table 1), including data from Twitter, Facebook, Reddit, and online news sources. The event data are broken down by day. There is a list of items from the online streams for each event-day pair. The task consists of consuming these daily multi-platform streams and producing summaries for a given information need. Participant systems must consider a report request for each event-day pair and produce facts using only the data items published on that day. CrisisFACTS refers to summarising social media crisis content as a fact-extraction task. A system should produce a list of atomic facts, with relevance scores denoting how critical the fact is for stakeholders. These facts together form an event-day summary.

The organizers also provide a list of questions that define the information needs of disaster-response stakeholders extracted from the FEMA ICS209 forms.[4] These queries are divided into subsets according to their intent and

[2] https://crisisfacts.github.io/
[3] https://trec.nist.gov/
[4] https://crisisfacts.github.io/assets/pdf/ics209.pdf

Table 1. TREC CrisisFACTS events and available data amount. Table from https://crisisfacts.github.io/#events

Event Name	Type	Tweets	Reddit	News	Facebook
Lilac Wildfire 2017	Wildfire	41,346	1,738	2,494	5,437
Cranston Wildfire 2018	Wildfire	22,974	231	1,967	5,386
Holy Wildfire 2018	Wildfire	23,528	459	1,495	7,016
Hurricane Florence 2018	Hurricane	41,187	120,776	18,323	196,281
Maryland Flood 2018	Flood	33,584	2,006	2,008	4,148
Saddleridge Wildfire 2019	Wildfire	31,969	244	2,267	3,869
Hurricane Laura 2020	Hurricane	36,120	10,035	6,406	9,048
Hurricane Sally 2020	Hurricane	40,695	11,825	15,112	48,492

include 46 general questions (e.g., "Have airports closed?"), six queries addressed to wildfires (e.g., "What area has the wildfire burned?"), five addressed to hurricanes (e.g., "What is the hurricane category?"), and two to flooding events (e.g., "What flood warnings are active?").

Evaluation Metrics

To assess the quality of participant systems, CrisisFACTS relies on two types of evaluation: 1) summary-based assessment, using existing event summaries from Wikipedia and official reports from the National Incident Management System, and 2) fact-based assessment, matching the lists of facts created by NIST assessors to each event with the facts generated by participant systems.

CrisisFACTS 2022 evaluated participant systems using two metric sets. The first comprises standard summarization metrics: Rouge-1, Rouge-2, Rouge-L, and BERTScore (Zhang et al. 2019). This first set compares the generated summaries against three ground-truth summaries: 1) summaries created by NIST assessors, 2) summaries extracted from ICS209 reports, and 3) Wikipedia articles summarising the events.

The second metrics set is used for fact-matching evaluation, which aims to quantify the volume of unique information in the summaries. These metrics consider the top-k item of the lists of facts generated for each event-day pair ordered by fact importance score, denoted S_d. Those facts are pooled and compared against the ground-truth list of facts F for potential matches. The matched facts are assessed in terms of *comprehensiveness*, defined by Equation 1, where $M(f, S)$ it the set of facts ($[i^1, i^2, ...]$) in S that match the fact f, and $R(f)$ is the gain assigned to the fact f, which for CrisisFACTS 2022 is always 1. *Redundancy* is also assessed in terms of facts of S_d by counting the unique facts matched divided by the total number of fact matches (cf. Equation 2). This metric measures how often a participant's event-day summary repeats information.

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

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

Implementation details

To apply our method to CrisisFACTS, first, we split the event-day datasets into four subsets based on the items' publication timestamp. Items published from 00:00 to 05:59 AM are in the first subset, from 06:00 AM to 12:00 PM are in the second, and so on. This split helps us generate more relevant facts along the event day and mimics a system that receives multiple requests on a unique day.

We generate a fact for each query-subset pair. On average, it takes 4 seconds to generate a fact using the complete pipeline, including retrieval and aggregation. To do this, we search candidate documents using BM25 and rerank them using monoT5, leveraging an NVIDIA A100 GPU for efficient reranking.[5] For the aggregation step, we used GPT-3 with a one-shot prompt and chain-of-thought, utilizing the OpenAI API.[6] We feed the model with the top 10 documents returned by the retrieval step. We discard the generated fact when the model's final answer is equal to "Unanswerable". Figure 5 presents an example of a fact generated by our method.

[5]We used the 3 billion parameters version of monoT5, whose checkpoint is available at https://huggingface.co/castorini/monot5-3b-msmarco-10k.

[6]We used the `text-davinci-002` model available via the OpenAI API.

Figure 5. Example of generated fact, with the event details (black), the generated fact text (blue), the documents used as sources (pink), the fact timestamp (purple), and the relevance score (gray).

Event: Lilac Wildfire 2017 (CrisisFACTS-001)
Request: CrisisFACTS-001-r3
Information Need: Are helicopters available? (CrisisFACTS-Wildfire-q003)

Generated fact: 11 helicopters were available to fight the Lilac Fire, two helicopters continued making water drops after dark, several helicopters were used to battle the blaze, a helicopter drops water on flames in the San Luis Rey riverbed, night-flying helicopters prepared to make water drops.

Sources:

CrisisFACTS-001-News-19-13: Lilac fire in San Diego County 4,100 acres burned (as of 12 p.m. Tuesday) 92% containment 1,659 firefighters on scene 157 structures destroyed, 64 damaged 10,000 people evacuated 11 helicopters

CrisisFACTS-001-News-12-12: Two helicopters continued making water drops after dark, and the Navy has agreed to mobilize military helicopter crews to fight the fire on Friday.

CrisisFACTS-001-News-12-14: North County Fire Protection District, Cal Fire and other firefighters from around San Diego County were battling the blaze, using several helicopters, bulldozers and air tankers.

CrisisFACTS-001-News-6-18: (Hayne Palmour IV / San Diego Union-Tribune) 9 / 38 A helicopter drops water on flames in the San Luis Rey riverbed in Bonsall.

CrisisFACTS-001-News-8-19: Gov. Jerry Brown declared a state of emergency in the county as night-flying helicopters prepared to make water drops.

Timestamp: Thursday, 7 December 2017 00:00:00

Relevance score: 0.75

RESULTS

Summarization (automatic metrics)

In Table 2, we present the results of the automatic evaluation. The automatically generated facts are compared with the ground truth facts from ICS-209, NIST assessors, and Wikipedia event pages using BERTScore and Rouge-2 F1 measures. The results are averaged over all eight events of CrisisFACTS 2022. To calculate these scores, organizers pooled the top-k highest relevant facts into a single document representing the event summary. The value of k varies over event-day pairs according to the number of facts produced by NIST assessors. Table 2 presents the results of the top-3 systems, which include the organizers' baseline system. Besides, the table presents the average, median, minimum, and maximum values for each metric in each testing set of all participants.

The results demonstrate that our proposal is competitive concerning other participants. We had the best results regarding ICS-209 reports and the second best in the other datasets with no significant difference from the best system. Also, our system is better than baselines and the average and median of the participants in all metrics. Note that the three methods presented in Table 2 produced better summaries, on average, than the other gold-standard summaries. This demonstrates that the gold standards sets do not agree with each other.

Fact Matching

Table 3 presents the results regarding the matching metrics, sorted by comprehensiveness in descending order. The table also shows other facts-matching evaluation statistics (i.e., Assessed@k and Matching Data). We'll cover the main metrics first and then the additional statistics. Comprehensiveness represents the fact recall of the system's summary over the ground-truth summaries. For this metric, higher is better. Our system produced facts with the

Table 2. Automatic evaluation results comparing participant runs with ICS-209, NIST, and Wikipedia summaries.

	ICS 209		NIST		Wikipedia	
Run	BERTScore	ROUGE - 2	BERTScore	ROUGE - 2	BERTScore	ROUGE - 2
Participant 1.Run 1	0.4432	0.0464	**0.5642**	**0.1471**	0.5448	0.0337
Participant 1.Run 2	0.4477	0.0507	0.5628	0.1468	**0.5646**	**0.0362**
Ours	**0.4591**	**0.0581**	0.5573	0.1338	0.5321	0.0281
Baseline.Run 1	0.4432	0.0418	0.5565	0.1326	0.5296	0.0275
Baseline.Run 2	0.4427	0.0428	0.5565	0.1308	0.5274	0.0267
Mean	0.4407	0.0395	0.5456	0.1175	0.5216	0.0278
Median	0.4383	0.0398	0.5482	0.1237	0.5216	0.0275
Minimum	0.4204	0.0131	0.5095	0.0651	0.4806	0.0236
Maximum	0.4591	0.0581	0.5642	0.1471	0.5646	0.0362
ICS-209	-	-	0.5134	0.0430	0.4885	0.0078
NIST	0.5134	0.0430	-	-	0.5368	0.0356
Wikipedia	0.4885	0.0078	0.5368	0.0356	-	-

highest comprehensiveness among the participants. We achieved 34.25% against 26.34% of the runner-up system. This means that our system generated summaries that covered around 30-35% of all facts on any event day. The redundancy ratio measures how often a participant summary repeats information. For this metric, lower is better. Our system had the highest redundancy ratio. This indicates that the facts produced by the system to a single event-day pair tend to repeat content and information. It occurred because of the dataset split step we performed. Users may post similar statements during the day, which the proposed method used to produce similar facts.

The Assessed@k column in Table 3 is not a performance metric but rather a measure of confidence in the main metrics. It reports the percentage of system items assessed by TREC assessors for fact matching assessment. The Assessed@k value has an inverse relationship with the uncertainty of the results; the lower the value is, the greater the uncertainty. A higher value may be beneficial as more matches are likely to be identified, given that the system provides meaningful items. Our system has the lowest Assessed@k value of the top-3 models. However, it is still above the average and median of the participants.

Regarding the Matching Data presented in Table 3, it helps to better interpret the results of the main metrics. Matched % represents the proportion of facts generated by participant systems that matched at least one gold-standard fact. The results show that our system had a high proportion of matched facts regarding other participants. Irrelevant % is the percentage of items that do not contain relevant information, according to assessors. Our system had the lowest proportion of irrelevant facts among all the participants. Unmatched % is the proportion of facts that the assessors tagged as having potentially relevant information but did not match any fact in the gold-standard list. The table shows that our system had a proportion of unmatched items above the participants' average and median. Matched and Unmatched % can be summed to produce a % of assessed items the assessors consider relevant. In this case, our method has 88.7% of relevant facts while the average proportion of all participants is 55.5%.

We can compare Matched % with Comprehensiveness to analyze the density of facts within a system's run. A high Matched % but a low comprehensiveness value indicates that the system produced summaries containing only a few (even insufficient) facts. An example is the run of Participant 4, which is on the top-left of the scatterplot presented in Figure 6. Runs with high values in both metrics also have more facts on each summary and, most importantly, relevant facts. Our run has the highest values for both metrics, which shows that our system can generate accurate and comprehensive facts. Moreover, looking at how these two metrics are defined, Matched % is equivalent to

Table 3. Match-based (manual) evaluation performance sorted by comprehensiveness in descending order.

		Main Metrics		Matching Data		
Model	Assessed@k	Comprehensiveness	Redundancy Ratio	Matched%	Irrelevant%	Unmatched%
Ours	0.6428	**0.3425**	0.4313	**0.5840**	**0.1134**	0.3027
Runner-up	0.6680	0.2634	0.2574	0.3936	0.3687	0.2377
Baseline 1	0.6533	0.2629	0.2852	0.4926	0.2546	0.2361
Baseline 2	0.6893	0.2528	0.2696	0.3757	0.3639	0.2480
Mean	0.6263	0.1899	0.2304	0.3320	0.4415	0.2233
Median	0.6523	0.1750	0.2424	0.2891	0.4662	0.2090
Maximum	**0.7814**	0.3425	0.4313	0.5840	0.6474	**0.3483**
Minimum	0.3434	0.0795	**0.0794**	0.1960	0.1134	0.1460

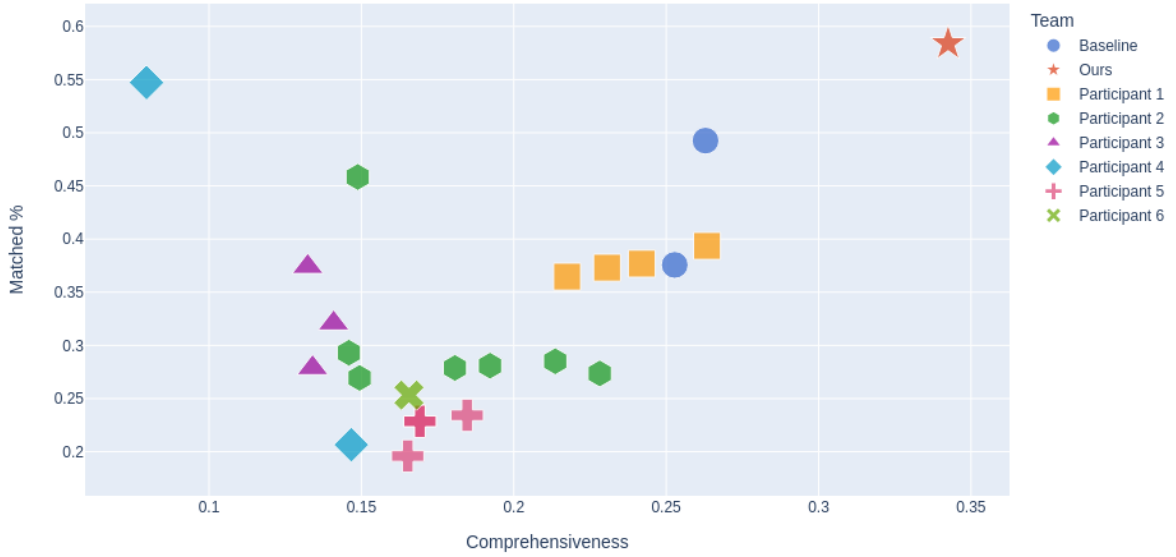

Figure 6. Scatterplot for Matched % and Comprehensiveness.

precision, while Comprehensiveness is equivalent to recall. Thus, in Figure 6, we have a precision-recall analysis, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. Both high scores show that the classifier returns accurate (high precision) results and the most positive (high recall) results.

Comparing Across Metrics

Figure 7 presents a scatter plot of the manual and automatic evaluations. We present Rouge-2 over the ICS-209 test set as the automatic metric and comprehensiveness as the manual metric. Despite two outliers, the proximity of the runs to the overall trendline suggests a moderate-to-strong correlation between the presented metrics. Runs with a high Rouge-2 on the ICS-209 test set may also have a high comprehensiveness. However, our method stands out from the others, considering both metrics.

Limitations

It is essential to acknowledge the limitations of our approach when conducting any research or experiment. Firstly, due to the associated costs, we could not experiment with various hyperparameters, such as changing the number of top-k documents passed to the GPT-3 or using other decoding strategies besides greedy decoding (e.g., nucleus sampling). Secondly, we did not evaluate other large language models like FLAN-T5. Lastly, we only evaluated our approach on a curated dataset, which may not accurately reflect real-world scenarios. Therefore, to establish the generalizability and applicability of our approach, it would be necessary to analyze its performance in a more noisy context where imprecision is present both in time and space. In conclusion, recognizing the limitations of our approach is crucial for the validity and reliability of our research findings.

CONCLUSIONS

This paper presents a method for generating accurate and comprehensive summaries of emergency events by retrieving information from social media and online news sources. The proposed method uses a state-of-the-art search strategy and GPT-3 in a one-shot setting for query-based multi-document summarization of social media and online news sources' content about crisis events. The results of our evaluation on the TREC CrisisFACTS challenge dataset showed that the proposed method performed well both in standard summarization metrics and manual evaluations. Manual evaluations of the generated summaries showed a high comprehensiveness despite a high redundancy ratio.

Our approach can be a valuable tool for crisis management teams, providing them with reliable and up-to-date information during emergencies. Additionally, the proposed method does not require annotated data and thus

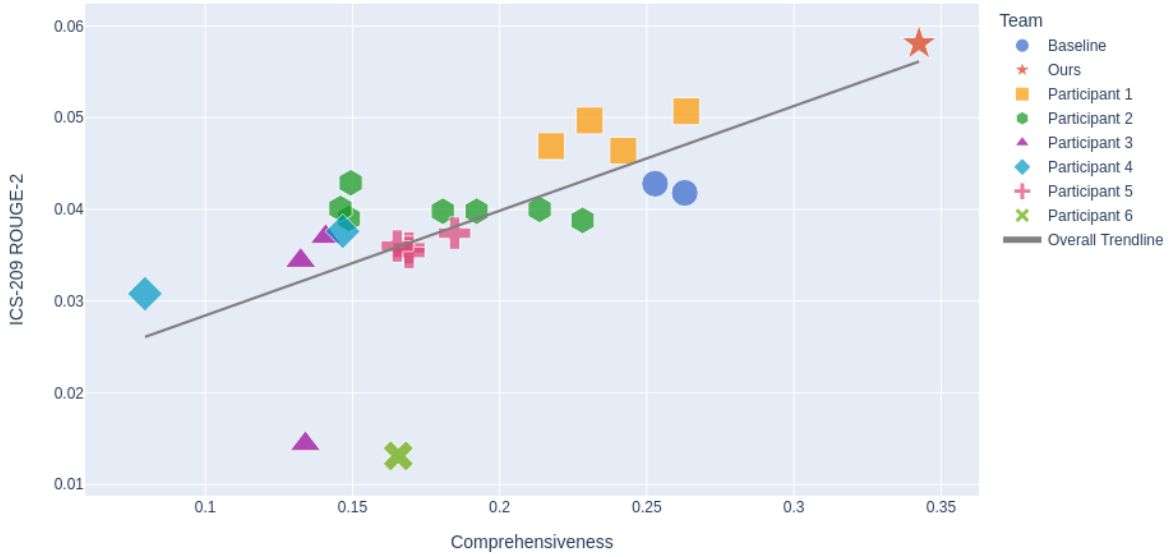

Figure 7. Performance scatter plot for both Similarity (ICS-209, ROUGE-2) and Matching (Comprehensiveness) Evaluations. Runs are colored by organization.

can be rapidly deployed. However, further research is needed to examine the method's scalability and real-time performance in actual crisis scenarios. In future work, we intend to evaluate methods for reducing redundancy in the documents outputted by the retrieval component to produce summaries that encompass more relevant information

ACKNOWLEDGMENTS

This research was partially supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (project id 2022/01640-2) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (Grant code: 88887.481522/2020-00). We also thank Centro Nacional de Processamento de Alto Desempenho (CENAPAD-SP) and Google Cloud for computing credits.

REFERENCES

Abdi, A., Idris, N., Alguliyev, R. M., and Aliguliyev, R. M. (2017). "Query-based multi-documents summarization using linguistic knowledge and content word expansion". In: *Soft Computing* 21.7, pp. 1785–1801.

Almeida, T. S., Laitz, T., Seródio, J., Bonifacio, L. H., Lotufo, R., and Nogueira, R. (2022). "NeuralSearchX: Serving a Multi-billion-parameter Reranker for Multilingual Metasearch at a Low Cost". In: *DESIRES 2022 – 3rd International Conference on Design of Experimental Search & Information REtrieval Systems*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901.

Chali, Y. and Mahmud, A. (2021). "Query-Based Summarization using Reinforcement Learning and Transformer Model". In: vol. 2021-November, pp. 129–136.

Tas, O. and Kiyani, F. (2017). "A SURVEY AUTOMATIC TEXT SUMMARIZATION". In: *PressAcademia Procedia*, pp. 205–213.

Creswell, A. and Shanahan, M. (2022). "Faithful reasoning using large language models". In: *arXiv preprint arXiv:2208.14271*.

Gambhir, M. and Gupta, V. (2017). "Recent automatic text summarization techniques: a survey". In: *Artificial Intelligence Review* 47.1, pp. 1–66.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial Intelligence for Disaster Response". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. Seoul, Korea: Association for Computing Machinery, pp. 159–162.

Kedzie, C., McKeown, K., and Diaz, F. (July 2015). "Predicting Salient Updates for Disaster Summarization". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 1608–1617.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). *Large Language Models are Zero-Shot Reasoners*.

Kulkarni, S., Chammas, S., Zhu, W., Sha, F., and Ie, E. (2020). *AQuaMuSe: Automatically Generating Datasets for Query-Based Multi-Document Summarization*.

Lamsiyah, S., El Mahdaouy, A., Ouatik El Alaoui, S., and Espinasse, B. (2021). "Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, BM25 model, and maximal marginal relevance criterion". In: *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18.

Laskar, M., Hoque, E., and Huang, J. (2022). "Domain Adaptation with Pre-trained Transformers for Query-Focused Abstractive Text Summarization". In: *Computational Linguistics* 48.2, pp. 279–320.

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., and Nogueira, R. (2021). "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. Virtual Event, Canada: Association for Computing Machinery, pp. 2356–2362.

Lin, J., Nogueira, R., and Yates, A. (2022). *Pretrained Transformers for Text Ranking: BERT and Beyond*. Springer Nature.

Lorini, V., Castillo, C., Peterson, S., Rufolo, P., Purohit, H., Pajarito, D., Albuquerque, J. P. de, 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*, pp. 772–777.

Ma, C., Zhang, W. E., Guo, M., Wang, H., and Sheng, Q. Z. (Dec. 2022). "Multi-Document Summarization via Deep Learning Techniques: A Survey". In: *ACM Comput. Surv.* 55.5.

Nguyen, T. H. and Rudra, K. (2022). "Towards an Interpretable Approach to Classify and Summarize Crisis Events from Microblogs". In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, pp. 3641–3650.

Nguyen, T. H., Shaltev, M., and Rudra, K. (2022). "CrisICSum: Interpretable Classification and Summarization Platform for Crisis Events from Microblogs". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, pp. 4941–4945.

Nogueira, R. and Cho, K. (2019). *Passage Re-ranking with BERT*.

Nogueira, R., Jiang, Z., Pradeep, R., and Lin, J. (Nov. 2020). "Document Ranking with a Pretrained Sequence-to-Sequence Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 708–718.

Pereira, J., Fidalgo, R., Lotufo, R., and Nogueira, R. (2022). *Visconde: Multi-document QA with GPT-3 and Neural Reranking*.

Phengsuwan, J., Shah, T., Thekkummal, N. B., Wen, Z., Sun, R., Pullarkatt, D., Thirugnanam, H., Ramesh, M. V., Morgan, G., James, P., et al. (2021). "Use of Social Media Data in Disaster Management: A Survey". In: *Future Internet* 13.2.

Popescu, C., Grama, L., and Rusu, C. (2021). "A highly scalable method for extractive text summarization using convex optimization". In: *Symmetry* 13.10.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *J. Mach. Learn. Res.* 21.140, pp. 1–67.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). "Okapi at TREC-3". In: *TREC*.

Roitman, H., Feigenblat, G., Cohen, D., Boni, O., and Konopnicki, D. (2020). "Unsupervised Dual-Cascade Learning with Pseudo-Feedback Distillation for Query-Focused Extractive Summarization". In: pp. 2577–2584.

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*. HT '16. Halifax, Nova Scotia, Canada: Association for Computing Machinery, pp. 137–147.

Rudra, K., Goyal, P., Ganguly, N., Imran, M., and Mitra, P. (2019). "Summarizing Situational Tweets in Crisis Scenarios: An Extractive-Abstractive Approach". In: *IEEE Transactions on Computational Social Systems* 6.5, pp. 981–993.

Rudra, K., Goyal, P., Ganguly, N., Mitra, P., and Imran, M. (2018). "Identifying Sub-Events and Summarizing Disaster-Related Information from Microblogs". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, pp. 265–274.

Saroj, A. and Pal, S. (2020). "Use of social media in crisis management: A survey". In: *International Journal of Disaster Risk Reduction* 48, p. 101584.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). *Chain of Thought Prompting Elicits Reasoning in Large Language Models*.

Wu, Y., Li, Y., and Xu, Y. (2019). "Dual pattern-enhanced representations model for query-focused multi-document summarisation". In: *Knowledge-Based Systems* 163, pp. 736–748.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). *BERTScore: Evaluating Text Generation with BERT*.