Using Neural Reranking and GPT-3 for Social Media Disaster Content Summarization

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, necessitates real-time situational awareness for management teams. This paper presents a novel approach to obtaining accurate and comprehensive summaries of these events by utilizing a state-of-the-art open-source search engine and a large language model to retrieve and summarize information from social media and online news sources. The efficacy of this approach was evaluated on the TREC CrisisFACTS challenge dataset, utilizing automatic summarization metrics (e.g. ROUGE-2 and BERTScore) and manual evaluation by the challenge organizers. Results indicate that while our approach achieves high comprehensiveness, it also exhibits a high degree of summary redundancy. Importantly, the pipeline components are few-shot, avoiding the need for training data collection and enabling rapid deployment.

Keywords

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

INTRODUCTION

In the digital era, effective management of emergencies, such as natural disasters, requires efficient communication between management teams and various stakeholders, including the media, government agencies, and the general public. To effectively fulfill this role, management teams must have access to comprehensive and up-to-date information during the event. Additionally, these reports must be tailored to meet the needs of individuals and communities affected or potentially impacted by the disaster, who may have diverse and dispersed requirements. However, it is not feasible for a single crisis agent to be physically present at every relevant location during a disaster, such as a wildfire. In these instances, utilizing information from the general public and independent news sources can serve as a means of staying informed.

During a crisis event, the flow of crisis-related information is constant from multiple sources. Decision-makers can utilize online news sources and social media platforms to obtain up-to-date and pertinent information during the event (Saroj and Pal 2020; Phengsuwan et al. 2021; Yan and Pedraza-Martinez 2019; Lorini et al. 2021). When traditional forms of communication, such as telephone lines, are overwhelmed by high volumes of simultaneous usage, social media platforms such as Facebook and Twitter can serve as sources of information. Through these platforms, individuals can access and disseminate information about essential needs, including food, shelter, transportation, road conditions, and airport information. Social media can provide time-sensitive and relevant

*corresponding author

information to support decision-making in crisis management (Saroj and Pal 2020). Additionally, these platforms may continue to be operational even when other forms of communication, such as radio and television, are affected by disaster events (Saroj and Pal 2020). The widespread utilization of social media is attributed to the increased adoption of mobile technology. Despite its potential benefits, the volume and diversity of information generated through social media can pose challenges for crisis management teams in effectively extracting relevant data.

Recent studies have leveraged social media as a source for information extraction and summarization, intending to keep crisis management teams up-to-date and facilitate prompt decision-making (Saroj and Pal 2020). Recently proposed methods for summarizing disaster-specific social media content involve document classification and summarization (Saroj and Pal 2020). The classification step entails assigning documents, such as tweets, into classes corresponding to user information needs, such as situational classes related to affected regions or airport status. The summarization step involves generating text summaries from the documents in each class, resulting in a topic-based multi-document summarization approach. Previous research in this field has utilized 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 document classification and summarization. However, recent advancements in machine learning, particularly the development of pre-trained large language models such as GPT-3 (Brown et al. 2020), have paved the way for few- or zero-shot strategies that require minimal or no annotated data.

This study presents a method for generating comprehensive and accurate summaries of crisis events by retrieving information from social media and online news sources. Our method performs query-based multi-document summarization using the state-of-the-art NeuralSearchX search strategy (Almeida et al. 2022) to find relevant documents and using GPT-3 (Brown et al. 2020) in a one-shot setting to summarize the top-k results. The performance of our method was evaluated on the TREC CrisisFACTS challenge dataset, using standard summarization metrics, such as ROUGE-2 and BERTScore, and a manual evaluation conducted by the challenge organizers. The results show that our method performed well in both evaluations, ranking in the top 3 in the automatic evaluation and being the best in comprehensiveness in the manual evaluation. However, it displayed a high redundancy ratio in the generated summaries. An advantage of our approach is that it does not require any annotated data, making it easy to deploy the system rapidly. This is crucial for applications where the underlying data is constantly changing.

RELATED WORK

This section reviews related literature in crisis management and query-based multi-document summarization. The literature is divided into two main categories: 1) works that focus on utilizing social media summarization for crisis management and 2) works that propose methods for query-based multi-document summarization.

Social Media Summarization for Crisis Management

The utilization of social media as a source of information during crisis events has gained significant attention in recent years. The real-time availability and constant flow of information from various sources make it an indispensable tool for decision-makers during these events (Lorini et al. 2021). Social media's ability to handle heavy traffic and remain functional even in the absence of electricity, unlike traditional forms of communication such as radio and television, has led to its widespread use across the globe (Saroj and Pal 2020). Nevertheless, the abundance and diverse nature of data generated from social media pose challenges for crisis management teams to extract relevant information on time (Lorini et al. 2021; Saroj and Pal 2020; Phengsuwan et al. 2021).

Several methods have been proposed in the literature to address the challenges of utilizing social media information in crisis management. According to Saroj and Pal 2020, most of the research in this area is centered around information extraction, summarization, and event classification techniques. These techniques aim to condense the large volume of data generated by social media into concise and meaningful summaries, which can assist crisis management teams in making informed decisions on time.

In the literature, various methods have been proposed to tackle the challenges of utilizing social media for crisis management. These methods generally involve a two-step pipeline, comprising classification and summarization, to generate useful and concise information from disaster-related social media content. For instance, Kedzie et al. 2015 employed an unsupervised classification approach based on the prediction of the salience of documents in the context of a crisis event and integrated these predictions into a clustering-based multi-document summarization system. Rudra, Goyal, Ganguly, Mitra, et al. 2018 proposed a method that identifies sub-events and summarizes them using an Integer Linear Programming (ILP) based summarization algorithm. Rudra, Goyal, Ganguly, Imran, et al. 2019 presented a classification-summarization framework that assigns tweets into situational classes and then summarizes them, utilizing the AIDR classifier (Imran et al. 2014) for tweet classification. Nguyen, Shaltev,

et al. 2022 presented CrisICSum, a platform for classifying and summarizing crisis events' tweets, utilizing the BERT2BERT model for classification into different humanitarian classes and RATSUM (Nguyen and Rudra 2022) for summarization.

Our proposed approach differs from previous studies by using a search engine and a language model for classification and summarization. Unlike traditional methods that employ supervised classification techniques, our approach recasts the classification step as a search problem, reducing the dependency on domain-specific annotated data. Moreover, using few-shot learning for summarization further avoids the need for extensive annotated data. Furthermore, this approach enables efficient inspection of massive document collections with sublinear computational cost, unlike the linear cost associated with classification, which limits scalability to tens of thousands of documents.

Query-based Multi-Document Summarization

Multi-document Summarization (MDS) task involves generating a succinct and informative summary from a group of related documents. It presents a more complex challenge than single-document summarization, as the objective is to identify the most relevant text, eliminate redundancies, maintain coherence and comprehensiveness, and impart novelty. The concept of MDS has influenced web-based clustering systems such as news aggregators (Tas and Kiyani 2017). Query-based or Query-Oriented Multi-document Summarization (qMDS) aims to produce summaries from multiple documents that cover the user's specific interest (C. Ma et al. 2022; Kulkarni et al. 2020; Abdi et al. 2017). qMDS merges techniques from information retrieval and MDS to generate summaries that align with the user's needs. Such generated summaries are called query-focused, topic-focused, or user-focused summaries (Gambhir and Gupta 2017).

Recent studies have explored various methods for performing qMDS. Wu et al. 2019 proposed an unsupervised pattern-enhanced approach for representing documents and queries, utilizing a pattern-enhanced topic model and a pattern-based relevance model. Roitman et al. 2020 introduced Dual-CES, an unsupervised, query-focused extractive summarizer that builds upon the Cross-Entropy Summarizer (CES). Lamsiyah et al. 2021 presented an unsupervised extractive summarization method based on transfer learning from pre-trained sentence embedding models and the maximal marginal relevance criterion. Chali and Mahmud 2021 proposed a reinforcement learning technique and transformer model for unsupervised extractive summarization, considering information coverage and diversity within a fixed sentence limit. Popescu et al. 2021 formulated extractive text summarization as a convex optimization problem and proposed a simple and scalable algorithm based on convex relaxation and projected gradient descent. Laskar et al. 2022 explored a series of domain adaptation techniques, including transfer learning, weakly supervised learning, and distant supervision, to generate abstractive summaries for single and multiple document text summarization using pre-trained transformer models.

OUR APPROACH

This section presents the method we propose for generating facts from social media and web news to support crisis event management. The method consists of two major steps: Document Retrieval and Aggregation, as depicted in Figure 1. The approach can be classified as query-based multi-document summarization as it summarizes text from multiple documents guided by a user query. The method assumes the availability of questions that reflect the user's information needs. The subsequent subsections provide a detailed explanation of the steps involved in the approach.

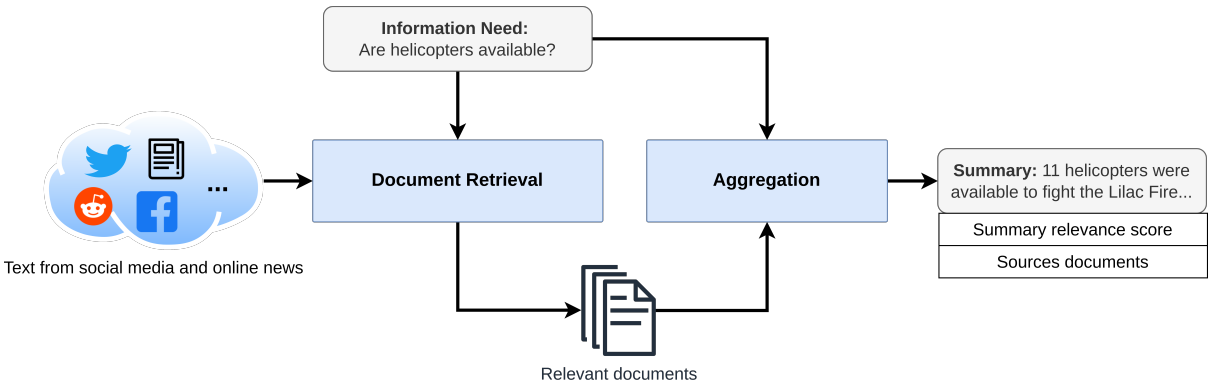

Figure 1. Illustration of the proposed method.

Document retrieval

The document retrieval component of the proposed method is based on the NeuralSearchX engine proposed by Almeida et al. 2022. This approach employs a two-stage pipeline consisting of (1) a bag-of-words retriever (i.e., BM25) for retrieving candidate documents and (2) a neural reranker for reranking the candidate documents. Further implementation details of these two stages are presented in subsequent subsections.

Candidate Document Retrieval

The candidate document retrieval stage aims to identify relevant candidate documents based on the queries that reflect the user's information needs. This step is crucial to minimize the computational burden of the subsequent reranking step. A variety of search functions can be employed to accomplish this aim. In this study, we adopt the Pyserini library (Lin, X. Ma, et al. 2021) implementation of the BM25 algorithm (Robertson et al. 1994).

For the experiment described in this work, we created an index of documents from the CrisisFACTS challenge dataset for searching purposes. However, in real-world crisis situations, the search space can be significantly larger than the one tested, especially when including news websites as sources in addition to social media. Keeping the search index updated in such cases can be challenging due to the constant flow of information. As an alternative, third-party APIs such as Bing, Google, or Twitter search can be utilized for candidate document retrieval instead. For web search, it is necessary to assume that the snippets provided by the search engines accurately represent the content of the pages, as noted by Almeida et al. 2022.

Document Reranking

The objective of the reranking stage is to prioritize the relevance of candidate documents concerning the user information needs expressed as queries. The input for this stage is a set of candidate documents obtained from diverse sources, including web news and social media. The output of this stage is a ranked list of documents deemed to be most relevant to the corresponding query.

For document reranking, we used a classifier trained to estimate the probability of each document being relevant given a question and order the documents in decreasing order according to their probabilities. Research has shown that the use of Transformers results in state-of-the-art performance in this task (Nogueira et al. 2020; Lin, Nogueira, et al. 2022). Nogueira et al. 2020, for example, 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 or not 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. Figure 2 (bottom) illustrates how the model can be utilized for reranking a list of documents during inference. Each query-document pair is processed independently by the model, and a relevance score is estimated. The documents are then ordered according to their estimated relevance scores.

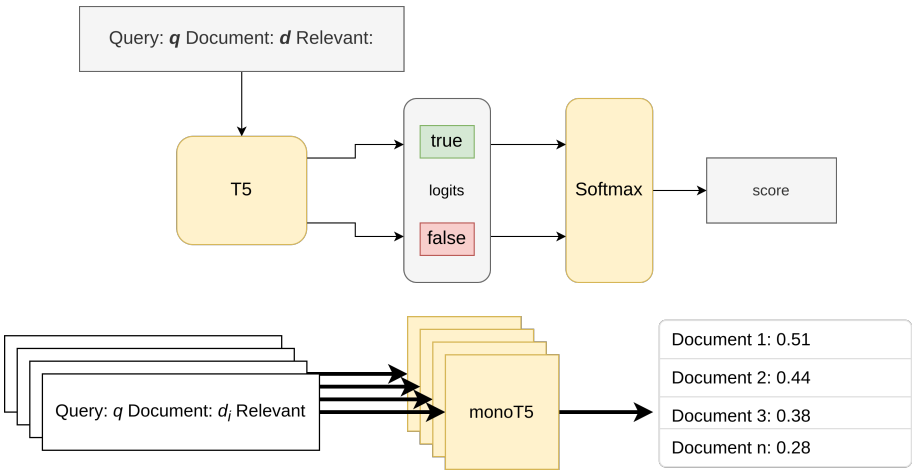

Figure 2. monoT5's training (top) and inference (bottom).

Aggregation

As previously noted, the present method is a multi-document, query-based summarization system that aims to provide a condensed response to a question through information extracted from multiple documents. Recent studies have demonstrated that using Large Language Models (LLMs) as few-shot in-context learners for question answering is an effective and cost-efficient strategy, as it only requires a small number of labeled training examples (Pereira et al. 2022). In Pereira et al. 2022, state-of-the-art performance or near-state-of-the-art performance was achieved in multi-document Question Answering (QA) using GPT-3 (Brown et al. 2020) as a few-shot in-context learner. A reasoning step referred to as chain-of-thought (CoT) was utilized before answering the question, resulting in a substantial improvement in the LLMs' few-shot performance. This technique has also been shown to produce improved results in multi-document QA and in various other QA benchmarks (Wei et al. 2022; Wang et al. 2022; Kojima et al. 2022; Creswell and Shanahan 2022). In this work, we adopted a similar approach to Pereira et al. 2022 by incorporating the CoT reasoning step in our summarization method. However, instead of using the final answer, we consider the reasoning paragraph generated by the model as the summary.

An example of the prompt used in our method is shown in Figure 3. The prompt has three main components: (1) the header, which provides instructions and explains the task to the model; (2) the one-shot example, which includes the context documents ([Document 1]), the question, the evidence paragraph inducing the chain-of-thought (CoT) reasoning, and the final answer; and (3) the target example, which includes context documents and an associated question. The model generates the evidence paragraph and final answer to the target example when provided with the prompt. The generated evidence paragraph is used as the summary in the management report. The model cites supporting documents in the evidence paragraph, as it was instructed to do so by the one-shot example. These citations provide the manager with information about the source of information, such as publication timestamp and location, if available. In cases where insufficient information is provided in the documents, the model generates the answer "Unanswerable" as indicated by the header. In such cases, the generated summary is discarded as it may be deemed irrelevant or incorrect.

Summary Relevance

The confidence level of the generated summaries in crisis event management systems can be important in decision-making processes. The relevance score of each document to the user's information need query is estimated during the ranking stage of the document retrieval component. The summary relevance score is computed as the average of the relevance scores of the documents cited in the produced summary. For instance, if the relevance scores of the documents cited in the generated summary, as illustrated in Figure 3, are 0.9, 0.8, 0.75, 0.7, and 0.65, respectively, then the summary relevance score is 0.75.

EXPERIMENT

The proposed method was evaluated through an experiment submitted to the 2022 TREC CrisisFACTS Track.[1] The Text REtrieval Conference (TREC)[2] is a conference co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense, focused on advancing research within the information retrieval community. TREC provides a platform for large-scale evaluations of text retrieval methodologies across various tracks, including the Crisis Facts and Cross-stream Temporal Summarization (CrisisFACTS) track, an open data challenge for evaluating summarization technologies in disaster management using online data sources.

Dataset and Task Formulation

The TREC CrisisFACTS 2022 evaluation provides a multi-stream dataset related to 8 crisis events, which includes information from various online sources such as Twitter, Facebook, Reddit, and online news (cf. Table 1). The dataset is divided into daily data for each event, and the task is to generate summaries based on the information in the daily data items. The challenge requires systems to produce summaries for each event-day pair, considering only the data published on that day. The summarization of crisis content from social media is defined as a fact-extraction task by CrisisFACTS, requiring systems to generate a list of atomic facts and relevance scores indicating their significance for stakeholders.

The TREC CrisisFACTS organizers provide a set of information needs, represented by questions, extracted from FEMA ICS209 forms[3]. These questions are categorized based on their intent and include 46 general queries (e.g., ""Have airports closed?""), six about wildfires (e.g., ""What area has the wildfire burned?""), five specific

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

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.**

to hurricanes (e.g., ""What is the hurricane category?""), and two regarding flooding events (e.g., ""What flood warnings are active?"").

Evaluation Metrics

The quality of participant systems in CrisisFACTS is evaluated through two methods: 1) Summary-based evaluation, which involves comparing the generated summaries with existing event summaries from Wikipedia and official reports from the National Incident Management System, and 2) Fact-based evaluation, which involves comparing the lists of facts created by NIST assessors with the facts generated by participant systems for each event.

CrisisFACTS 2022 utilized two metrics to evaluate participant systems' performance. The first set included standard summarization metrics, including ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore (Zhang et al. 2019). This first set of metrics compared the generated summaries with three reference summaries: 1) summaries generated by NIST assessors, 2) summaries extracted from ICS209 reports, and 3) Wikipedia articles summarizing the events.

In the second set of evaluation metrics for CrisisFACTS 2022, the focus is on fact-matching assessment. This evaluation aims to determine the volume of unique information present in the summaries. The top-k items from

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

the lists of facts generated for each event-day pair are ranked by their fact importance score (S_d) and compared with the ground-truth list of facts (F) to identify matches. The assessment of the matched facts is done in terms of comprehensiveness, as defined in Equation 1. Where $M(f, S)$ represents the set of facts in S that match the fact f, and $R(f)$ is the gain assigned to the fact f, which is set to 1 for CrisisFACTS 2022. Redundancy is also assessed by counting the number of unique facts matched divided by the total number of fact matches (cf. Equation 2).

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

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

Implementation details

Our method for CrisisFACTS involved dividing the event-day datasets into four subsets based on the publication timestamps of the items. The items were divided into subsets: items published from 00:00 to 05:59 AM were placed in the first subset, items published from 06:00 AM to 11:59 AM were placed in the second subset, and so on. This division was performed to generate relevant facts throughout an event day and simulate a system receiving multiple requests on a single day.

A fact is generated for each query-subset pair, utilizing all provided queries to search all subsets and generate facts based on the relevant documents. The candidate documents are searched using BM25 and reranked using monoT5.[4] The aggregation step involves GPT-3[5] with a one-shot prompt and chain-of-thought process, discarding generated facts if the model's final answer is equal to "Unanswerable". An example of a fact generated by our method is presented in Figure 4.

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 obtained from ICS-209 reports, summaries created by NIST assessors, and Wikipedia articles summarizing the events using the BERTScore and ROUGE-2 F1 measures. The results are averaged over all eight events of CrisisFACTS 2022. To calculate these scores, the organizers combined the top-k most relevant facts into a single document, representing the event summary. The value of k is dependent on the number of facts produced by NIST assessors and varies among event-day pairs. The table shows the results of the top-3 systems, including the baseline system created by the organizers. Additionally, it provides the average, median, minimum, and maximum values for each metric in each testing set for all participants.

In Table 2, the results of the automatic evaluation show that our proposed method is competitive compared to other participants. Our system achieved the best results with respect to ICS-209 reports and had the second best results among the other datasets with no significant difference from the best system. Additionally, our system outperformed baseline runs, mean and median of submissions across all metrics.

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

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

Figure 4. 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

Fact Matching

Table 3 presents the results of the matching metrics for the summary generation system, ordered in descending order by comprehensiveness. The table also includes other fact-matching evaluation statistics, including Assessed@k and Matching Data. Firstly, the comprehensiveness metric represents the recall of the system's generated facts concerning the ground-truth summaries. A higher value for this metric is desirable. Our system demonstrated the highest comprehensiveness among all participants, achieving a value of 34.25%, compared to 26.34% of the second-best system. This result indicates that the system's summaries covered approximately 30-35% of all facts for each event-day pair. Secondly, the redundancy ratio measures the repetition of information in a participant's summary. A lower value for this metric is preferred. Our system showed the highest redundancy ratio, implying that the generated facts for a single event-day pair tend to repeat content and information. This phenomenon can be attributed to the dataset split step, which may result in similar statements being posted during the day, which the proposed method then uses to generate similar facts.

The Assessed@k column in Table 3 represents a measure of confidence in the main metrics rather than a performance metric. It reports the percentage of items produced by the system that was assessed by TREC assessors for fact-matching purposes. Thus, Assessed@k correlates with the level of confidence in the results. The lower this value, the greater the potential for inaccurate results due to the inclusion of irrelevant items in the assessment score A higher value is advantageous as it increases the likelihood of identifying matches when the system outputs meaningful items. The Assessed@k value of our system surpasses the average and median of the participant models.

In Table 3, the Matching Data provides additional context for interpreting the results of the main metrics. The Matched % column reports the proportion of facts generated by the participant systems that have been found to match at least one gold-standard fact. Our system achieved a higher proportion of matched facts than the other

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	-	-

participants. The Irrelevant % column indicates the proportion of items that have been judged by the assessors to contain no relevant information. Our system had the lowest proportion of irrelevant facts among all participants. The Unmatched % column represents the proportion of facts labeled by the assessors as having potentially relevant information but did not match any gold-standard fact. Our system had a higher proportion of unmatched facts compared to the average and median of the participants. The sum of Matched % and Unmatched % yields the proportion of assessed items that were considered relevant by the assessors. In this regard, our system had 88.7% of relevant facts, while the average proportion among all participants was 55.5%.

To evaluate the density of facts within the system's output, we can compare the Matched % to the Comprehensiveness metric. A high Matched % with a low comprehensiveness value implies that the system's summary contains only a limited number of facts, which might not be sufficient. An example of this can be seen in the run of Participant 4, as shown in Figure 5. On the other hand, high scores for Matched % and Comprehensiveness indicate many accurate and relevant facts in the system's summary. In our run, the values for both metrics are the highest among all participants, indicating the ability of our system to generate comprehensive and accurate facts. In Figure 5, we can also observe a precision-recall analysis, where high values for both Matched % and Comprehensiveness equate to high precision (low false positive rate) and high recall (low false negative rate), respectively. Hence, high scores in both metrics demonstrate the accuracy and completeness of the classifier's results.

Comparing Across Metrics

Figure 6 shows the relationship between manual (Comprehensiveness) and automatic evaluations (ROUGE-2 over the ICS-209 test set). The proximity of the data points to the overall trendline indicates a moderate-to-strong correlation between the two metrics. However, the presence of two outliers highlights that the correlation is not perfect. Our method exhibits a unique combination of high scores for Comprehensiveness and ROUGE-2 over the ICS-209 test set, distinguishing it from the other runs.

CONCLUSIONS

This paper introduces a novel approach for generating precise and comprehensive summaries of emergency events by retrieving information from social media and online news sources. The method incorporates a cutting-edge

Table 3. Match-based (manual) evaluation results 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
Maximum	**0.7814**	0.3425	0.4313	0.5840	0.6474	**0.3483**
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
Minimum	0.3434	0.0795	**0.0794**	0.1960	0.1134	0.1460

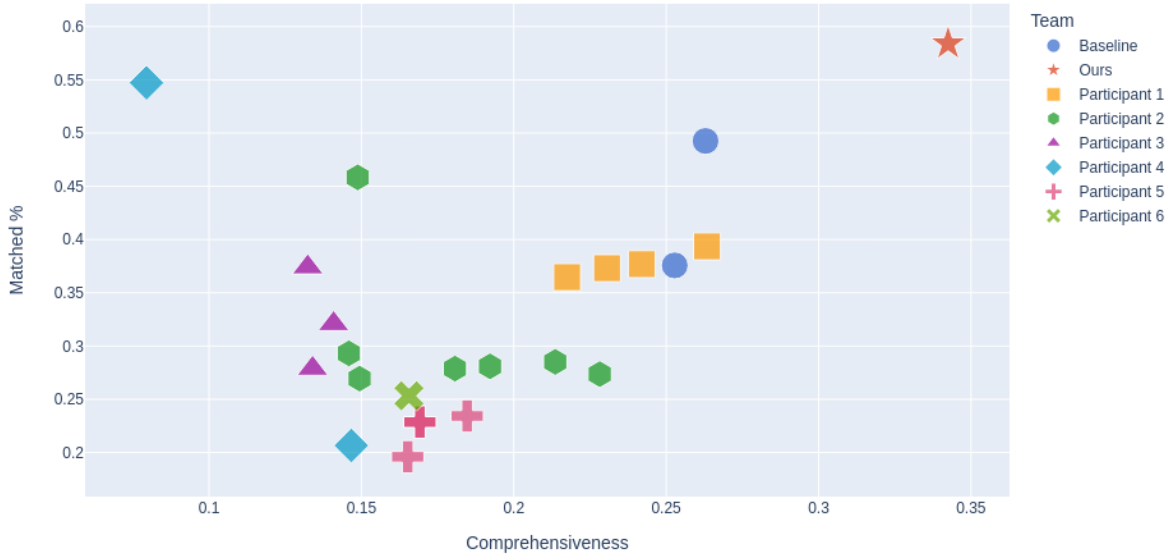

Figure 5. Results for Matched % and Comprehensiveness.

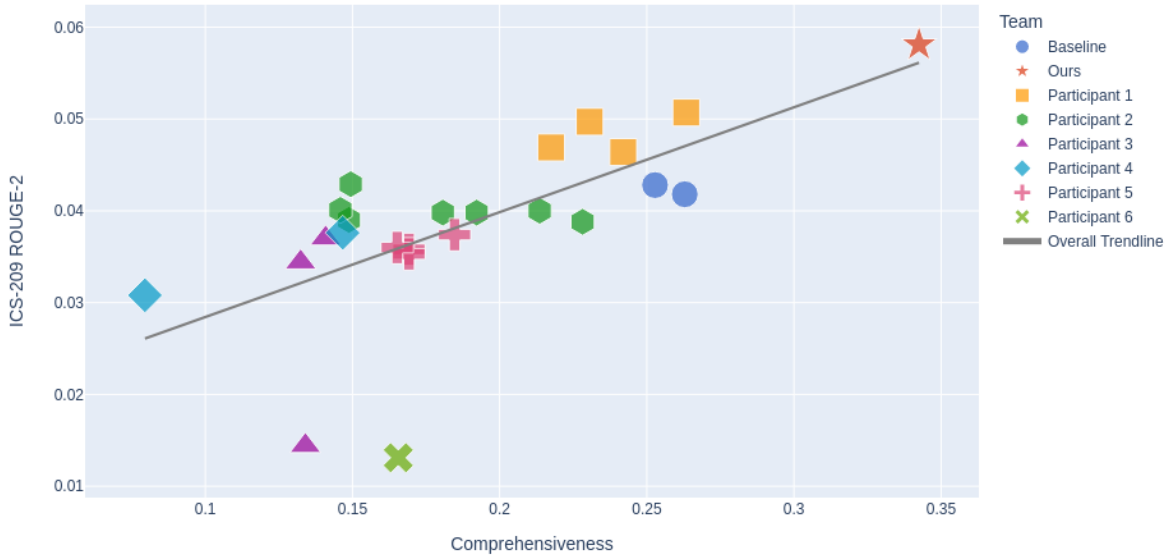

Figure 6. Results for Similarity (ICS-209, ROUGE-2) and Matching (Comprehensiveness) evaluations. Each point represents a submission run, and colors represent the participating teams.

search strategy and utilizes GPT-3 in a one-shot setting to perform query-based multi-document summarization of social media and online news content related to crisis events. The evaluation of the proposed method on the TREC CrisisFACTS challenge dataset demonstrated its effectiveness, with high performance in standard summarization metrics and manual evaluations. Manual evaluations of the generated summaries revealed a high level of comprehensiveness, despite a relatively high redundancy ratio.

Our proposed method has the potential to provide valuable support to crisis management teams by providing them with accurate and up-to-date information during emergency situations. The advantage of this approach is its rapid deployment, as it does not require annotated data. However, further research is necessary to assess its scalability and real-time performance in actual crisis scenarios. As part of our ongoing work, we plan to investigate techniques for reducing redundancy in the retrieval component's output, to generate more comprehensive and relevant summaries.

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., 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.

Yan, L. and Pedraza-Martinez, A. J. (2019). "Social media for disaster management: Operational value of the social conversation". In: *Production and Operations Management* 28.10, pp. 2514–2532.

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