Large Language Models in Summarizing Social Media for Emergency Management

Jayr Pereira*
NeuralMind, Brazil
jayr.pereira@neuralmind.ai

Roberto Lotufo
NeuralMind, Brazil
roberto@neuralmind.ai

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

ABSTRACT

The exponential increase of information during crisis events necessitates efficient and real-time summarization techniques to aid emergency response and coordination. To this end, this study leverages the power of large language models (LLMs) to summarize social media content in the context of crisis management. We introduce a novel method that combines advanced search algorithms with state-of-the-art LLMs to generate concise, relevant summaries based on user queries. Specifically, we utilize the BM25 algorithm and the monoT5 reranker to filter the most pertinent documents, which are then summarized using OpenAI's GPT-3.5-turbo and GPT-4 models. Our submission to the TREC CrisisFACTS Track 2023 demonstrates that integrating the monoT5 reranker with GPT-3.5-turbo significantly reduces redundancy and enhances the comprehensiveness of summaries. This progress indicates a substantial advancement over our previous year's efforts, reflecting the rapid evolution in the field of natural language processing. The capacity of the latest models to process larger contextual inputs without extensive data underpins their utility in streamlining the summarization process, which is vital for effective crisis communication.

Keywords

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

INTRODUCTION

In today's digital age, managing emergencies like natural disasters necessitates streamlined communication among response teams, the media, governmental bodies, and the general populace. For effective coordination, these teams must have access to in-depth, real-time data throughout the crisis. Moreover, the information shared should cater to the unique needs of those directly or potentially affected by the calamity, recognizing their varied concerns.

During crises, there's a continuous influx of information from various sources. Digital news outlets and social media become invaluable tools for accessing real-time updates during such situations (Saroj and Pal 2020; Phengsuwan et al. 2021; Yan and Pedraza-Martinez 2019; Lorini et al. 2021). Research efforts have recently turned to social media as a rich resource for data gathering and synthesis, aiming to keep emergency teams informed for swift decision-making (Saroj and Pal 2020). Modern methods for distilling disaster-specific content from social media involve categorizing and summarizing (Saroj and Pal 2020). This entails grouping content, such as tweets, based on user information needs, including situational updates about affected areas or airport statuses. Once classified, these contents are then summarized according to topic, employing a multi-document summarization technique. Past studies have either employed clustering techniques (Kedzie et al. 2015) or supervised algorithms (Rudra, Banerjee, et al. 2016; Rudra, Goyal, Ganguly, Mitra, et al. 2018; Rudra, Goyal, Ganguly, Imran, et al. 2019; Nguyen et al.

*corresponding author

NeuralMind's submission to CrisisFACTS 2023
Proceedings of the 32nd Text Retrieval Conference (TREC 2023)
Ian Soboroff and Angela Ellis, eds.

2022) for these tasks. However, the emergence of advanced machine learning tools, especially pre-trained models like GPT-3 (Brown et al. 2020), have introduced methods that demand little to no annotated data (Pereira et al. 2023a; Pereira et al. 2023b).

This paper reports our submission to TREC CrisisFACTS Track 2023. We propose a method for using large language models (LLMs) for query-based social media content summarization. This method uses a search engine to select the top-k relevant documents and an LLM to generate the facts about the event that compose the final summary report. We experimented with two search engines: (i) the BM25 algorithm and (ii) the monoT5 reranker (Nogueira et al. 2020). BM25 is a popular text-matching score mechanism that ranks documents based on their content relevance to a query, emphasizing term frequency-inverse document frequency (TF-IDF) weighted terms. On the other hand, the monoT5 reranker is an adaptation of the T5 (Text-to-Text Transfer Transformer) (Raffel et al. 2020) model specifically fine-tuned for reranking tasks in retrieval systems. Also, we employed two versions of OpenAI's language models for the summarization task: GPT-3.5-turbo and GPT-4. These LLMs have shown remarkable capabilities in understanding and generating human-like text, making them ideal candidates for creating coherent and contextually relevant summaries.

Our results indicate that using monoT5 in conjunction with GPT-3.5-turbo significantly improved the quality of the summaries compared to using BM25 alone. The summaries were not only more comprehensive but also contained fewer redundancies. The inclusion of GPT-4 further enhanced the summarization performance, particularly in aligning with the style and quality expected in Wikipedia references, which suggests its potential for applications requiring higher-quality language generation. The evolution from our prior year's submission (Pereira et al. 2023b) to our current methodology showcases the swift progression in natural language processing capabilities. Notably, the models applied in our latest research are adept at processing substantially larger contexts, an advancement that significantly enhances summarization effectiveness. Our results highlight the critical role of combining sophisticated retrieval algorithms with advanced LLMs to refine the automation of social media content summaries. This combination is particularly pivotal in crisis management scenarios, where synthesizing extensive and rapidly evolving information is crucial.

METHOD

Figure 1 presents an overview of the proposed method's workflow. It comprises four main components: the user's query input, the search engine, a prompt, and the large language model (LLM). The query signifies the user's intent and the specific information they seek within the social media content. It is used for searching for relevant documents that are used as a basis for the information in the produced summary. In this context, the search engine acts as a content filter, systematically sifting through vast amounts of data to identify and rank documents based on their relevance to the user's query. Once the search engine has generated a ranked list of documents, the top-k is selected. These are then processed by the prompt mechanism, which formulates a structured instruction to guide the LLM. With the aid of this instruction, the LLM can extract and summarize key facts from the selected documents, ensuring that the final output aligns closely with the user's original information needs. Thus, Figure 1 underscores a synergistic integration of search and summarization processes tailored to deliver accurate and concise responses to user queries from various social media content.

As depicted in Figure 1, the prompt that guides the LLM to extract facts comprises the system instructions, the top-k documents, and the user query. We discussed the documents and the query in the previous paragraph, so now

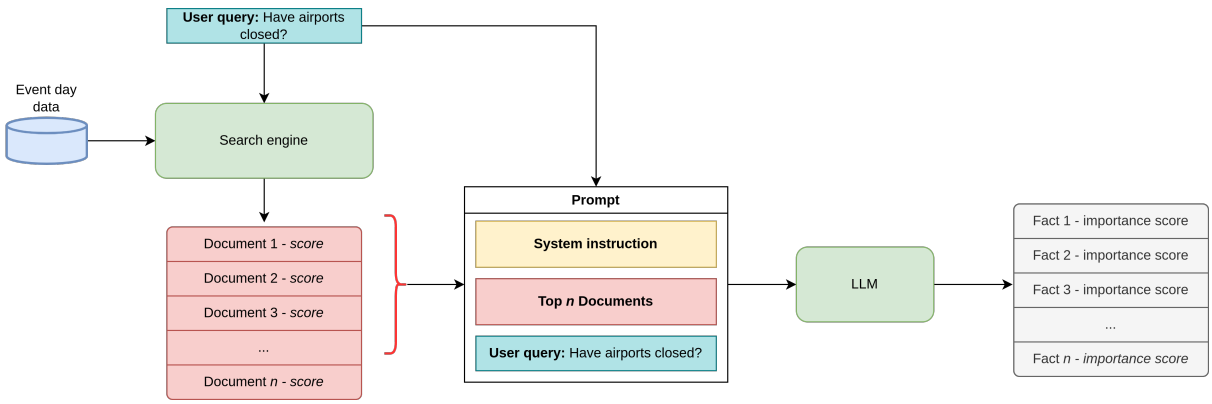

Figure 1. Illustration of the proposed method.

Figure 2. The system instructions

You are a crisis event management assistant. Your task is to summarize the social media information about a crisis event.

The main task focuses on fact extraction, where systems consume a multi-stream dataset for a given disaster, broken into disaster-day pairs. From this stream, the system should produce a minimally redundant list of atomic facts, with importance scores denoting how critical the fact is for responders. CrisisFACTS organizers will aggregate these facts into daily summaries of these disasters.

Consider the user query for guiding the facts construction.

You must produce a list of facts in this format:
Fact number: <>
Fact text: <>
Sources: <The document ids separated by a comma>
Date time: <The earliest point in time the system detected the fact.>
Importance: <A numerical score between 0 and 1 indicates how important you think it is for this fact to be included in the final summary.>

we focus on the system instruction. The system instruction serves as a directive for the LLM, specifying the type of information it should extract from the provided documents. It acts as a bridge between the raw, ranked data and the nuanced needs of the user represented by the query, ensuring that the LLM focuses on the most pertinent details. By combining the context of the user query, the relevance of the top-k documents, and the specificity of the system instruction, the LLM is equipped to generate summaries that are both informative and aligned with the user's intent. This triad of inputs ensures that the LLM's output is not only accurate but also contextually relevant, making the summarized information more actionable for the end user.

Figure 2 presents the text used in the instruction to guide the LLM during the summarization process. This text provides a clear framework, detailing the specific format and structure the summaries should adhere to. By adhering to this structured guideline, the LLM can maintain consistency in its responses and ensure that the generated facts encapsulate the essence of the sourced content while answering the user's query effectively. Notice that the LLM is instructed to extract facts in a standardized manner, with each fact being assigned a unique number for easy reference. The "Fact text" section captures the core information, concisely summarizing the relevant details from the documents. Each fact is also backed by its "Sources," which are identifiable through the provided document IDs, offering a means for users to verify the authenticity and context of the extracted information. The "Date time" field serves to timestamp the earliest moment when a particular fact was detected by the system, providing users with an understanding of the timeliness of the information.

Figure 3 shows how the top-k documents from the search engine are formatted as input. Each document follows a structured layout, beginning with a unique "Document id" that serves as an identifier, making it easier to trace and reference specific content. This ID combines various elements, such as the source platform, a unique number, and other details to ensure its distinctiveness. Following this, the "Text" section provides the actual content of the document, capturing the user's sentiment, statement, or any relevant information shared. The "Source" field specifies the origin platform of the content, in this case, "Twitter", which aids in contextualizing the data and understanding its scope. Lastly, the "Datetime" field registers the exact time when the content was posted or shared, offering insights into the chronological context of the information. This structured format ensures a standardized approach to data intake, simplifying the process of content analysis and fact extraction for the LLM.

Figure 3. Documents input format.

Document id: <>
Text: <>
Source: $< twitter|facebook|... >$
Datetime: <>

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
Beirut Explosion, 2020	Accident	94,892	3,257	1,163	368,866
Houston Explosion, 2020	Accident	58,370	5,704	2,175	6,281
Rutherford TN Floods, 2020	Floods	11,019	475	268	9,116
TN Derecho, 2020	Storm/Flood	49,247	1,496	15,425	13,521
Edenville Dam Fail, 2020	Accident	16,527	2,339	961	8,358
Hurricane Dorian, 2019	Hurricane	86,915	91,173	7,507	370,644
Kincade Wildfire, 2019	Wildfire	91,548	10,174	339	35,011
Easter Tornado Outbreak, 2020	Tornadoes	91,812	5,070	750	34,343
Tornado Outbreak, 2020 Apr	Tornadoes	99,575	1,233	217	19,878
Tornado Outbreak, 2020 March	Tornadoes	95,221	16,911	641	87,242

EXPERIMENTS

This section details the experiment we conducted to assess the efficacy of our proposed method. This particular experiment represents our participation in the 2023 TREC CrisisFACTS Track.[1] The Text REtrieval Conference (TREC)[2] is an esteemed event co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense. Its primary objective is to foster advancements in the information retrieval domain. The conference facilitates comprehensive evaluations of text retrieval methodologies across varied tracks. The TREC Crisis Facts and Cross-stream Temporal Summarization (CrisisFACTS) stands out as a significant track of TREC, posing an open data challenge. This challenge centers on using summarization technologies to aid disaster management by leveraging online data sources.

Dataset and Task Overview

The 2023 CrisisFACTS offers a diversified dataset encompassing data from 18 significant crisis events (refer to Table 1). This dataset amalgamates information from various platforms like Twitter, Facebook, Reddit, and online news sites. Each event's data is categorized daily, comprising items from these online streams for every unique event-day pairing. The proposed task is to analyze these daily streams across platforms and craft summaries based on specific informational requirements. For every event-day combination, participating systems are to address a report request, generating facts solely from the data items released on the respective day. Within CrisisFACTS, the act of condensing social media content related to crises is termed a fact-extraction task. Systems are expected to curate a series of individual facts, each with a relevance score indicating its significance to the stakeholders. Together, these facts constitute the event-day summary.

Furthermore, the organizing team furnishes a catalog of questions, capturing the informational necessities of those involved in disaster response derived from the FEMA ICS209 forms.[3] These inquiries are grouped based on their core intent and comprise 46 universal questions (such as "Have airports closed?"), six tailored for wildfires (like "What region has the wildfire affected?"), five for hurricanes (for instance, "Which category does the hurricane belong to?"), and a couple for flood-related events (e.g., "Which flood alerts are currently in effect?").

Metrics of Evaluation

CrisisFACTS 2023 applies two primary forms of assessment: 1) an overarching summary-based analysis utilizing benchmark event summaries from sources like Wikipedia and official reports procured from the National Incident

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

Management System, and 2) a specific fact-based evaluation that juxtaposes the lists of facts curated by NIST evaluators for each event against the facts enumerated by the participant systems.

For its 2023 edition, CrisisFACTS employed two distinctive metric suites to evaluate the participant submissions. The initial set encompasses conventional summarization metrics, notably Rouge-1, Rouge-2, Rouge-L, and BERTScore (Zhang et al. 2019). This suite benchmarks the participant-generated summaries against three predefined ground truths: 1) those crafted by NIST evaluators, 2) extracts from ICS209 reports, and 3) Wikipedia summaries pertinent to the events.

The subsequent metric suite focuses on a fact-matching evaluation, intending to quantify the richness of unique information present within the summaries. This suite scrutinizes the top-k entries from the fact lists curated for each event-day combination, sequenced by their respective importance scores, symbolized as S_d. These facts are then amalgamated and juxtaposed against a benchmark list of facts, F, to identify potential matches. Recognized matches are then gauged for *comprehensiveness*, as delineated by Equation 1. Here, $M(f, S)$ denotes the collection of facts ($[i^1, i^2, ...]$) within S resonating with fact f, while $R(f)$ signifies the value attributed to fact f, which for CrisisFACTS 2023 is uniformly set to 1. An additional metric, *redundancy*, evaluates the facts of S_d by tallying the distinct facts identified relative to the overall count of fact matches, as elaborated in Equation 2. This metric ascertains the recurrence of information within a participant's event-day summary.

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

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

Implementation details

In the 2023 iteration of TREC CrisisFACTS, our approach was tested across three distinct configurations: 1) deploying the BM25 algorithm as the sole search mechanism coupled with OpenAI's GPT-3.5-turbo-16k as the language model; 2) integrating BM25 with the monoT5 re-ranking process and maintaining OpenAI's GPT-3.5-turbo-16k for language modeling; and 3) combining BM25 with monoT5 for the initial search and retrieval, subsequently employing OpenAI's GPT-4 for language modeling.

For the configurations utilizing both BM25 and monoT5, our method involved using monoT5 to rerank the top 1000 document results from BM25. For searching with BM25, we used Pyserini (Lin et al. 2021) to create a searchable index to each event-day within the CrisisFACTS data corpus. A searchable index refers to an organized dataset a search engine queries to fetch pertinent results. Pyserini offers a Python interface to the Lucene library[4], renowned for its robust performance as a search engine framework. Lucene employs the term frequency-inverse document frequency (TF-IDF) method for its indexing operations. This prevalent approach in the realm of information retrieval assigns higher significance to uncommon terms across the dataset while diminishing the importance of more frequent terms. These tailored indexes were then queried using the predetermined user queries to yield the most pertinent documents.

Leveraging the extended text handling capabilities of GPT-3.5-turbo-16k, we curated the prompts from the top 30 documents returned by the search engine. This selection strategy uses approximately 10,000 tokens, providing a substantial textual context while reserving an additional 6,000 tokens for output generation. In contrast, for GPT-4, we confined the input to the top 10 documents. This decision was dictated by GPT-4's comparatively smaller context window of 8,000 tokens and a higher operational cost, which is approximately 15 times greater than its predecessor.

Summary Relevance

In the 2023 CrisisFACTS challenge, participants must assign a relevance score to every fact generated. In this work, this score is derived during the document retrieval phase, where each document is assessed for its probability of being pertinent to the user's query. To determine the overall relevance of a summary, we compute the average of the relevance scores but only for those documents that are used as a source for the generated fact. For instance, if a generated fact is underpinned by three documents with relevance scores of 0.8, 0.6, and 0.7, the fact's relevance score would be the average of these three numbers, which is 0.7.

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

Table 2. Results for the automatic evaluations.

Configuration	NIST		Wikipedia	
	BERTScore	ROUGE-2	BERTScore	ROUGE-2
BM25 + GPT-3.5-turbo	0.642	0.318	0.481	0.032
MonoT5 + GPT-3.5-turbo	0.668	0.416	0.471	0.035
MonoT5 + GPT-4	0.645	0.353	0.488	0.035
Participants Mean	0.596	0.229	0.462	0.025

RESULTS

Summarization (automatic metrics)

Table 2 presents the outcomes of the automatic evaluation metrics that summarize the crisis events as part of the TREC CrisisFACTS 2023 challenge. The assessment utilized BERTScore and ROUGE-2 metrics to gauge the quality of summaries generated by different system configurations against the NIST standard references and Wikipedia-based summaries. As shown in the table, the MonoT5 + GPT-3.5-turbo setup exhibited the highest degree of semantic accuracy, as reflected in its BERTScore when evaluated against NIST references. It also outperformed other configurations in capturing specific details, evidenced by its ROUGE-2 scores. However, when the same setup is compared with Wikipedia references, there is a slight decline in BERTScore, suggesting differences in the type of content and style between NIST reports and Wikipedia summaries that may affect the outcome.

The BM25 + GPT-3.5-turbo configuration, while not reaching the heights of the MonoT5-enhanced setups, still provided respectable BERTScores and ROUGE-2 scores, indicating it is a viable baseline method for generating crisis event summaries. Including GPT-4 with MonoT5 shows a nuanced improvement in BERTScore against Wikipedia references, though at a significantly higher computational cost. This increase, however, is not reflected in the ROUGE-2 score against the NIST references, where the combination with GPT-3.5-turbo continues to lead. It's noteworthy that for GPT-4, only the top 10 documents from the search engine were utilized due to its smaller context window and higher operational costs, in contrast to the top 30 documents used for GPT-3.5-turbo, which may affect the comparative performance outcomes.

Building on the analysis of the results from the TREC CrisisFACTS 2023 challenge, we now consider the performance of similar system configurations applied solely to events from the 2022 edition. These results provide an opportunity to compare the current models with the previous year's participation and the best-performing system from 2022. Table 3 outlines these comparisons.

Table 3. Automatic evaluation results for the 2022 crisis event summarization.

Configuration	NIST References		Wikipedia References	
	BERTScore	ROUGE-2	BERTScore	ROUGE-2
BM25 + GPT-3.5-turbo	0.680	0.379	0.550	0.037
MonoT5 + GPT-3.5-turbo	0.734	0.483	0.548	0.040
MonoT5 + GPT-4	0.685	0.391	0.564	0.041
NM-2022	0.557	0.134	0.532	0.028
2022 Best	0.564	0.147	0.565	0.036

The MonoT5 + GPT-3.5-turbo configuration leads in BERTScore and ROUGE-2 metrics against NIST references, indicating a significant year-over-year improvement in summarization quality. While the BERTScore of MonoT5 + GPT-3.5-turbo slightly dips compared to Wikipedia references, its ROUGE-2 score slightly increases, suggesting a more nuanced representation of the details important to Wikipedia summaries. The introduction of GPT-4 in combination with MonoT5 does not yield substantial improvements over the GPT-3.5-turbo in terms of BERTScore against NIST references but shows a marked improvement over the NM-2022 system. However, it offers the highest BERTScore and ROUGE-2 scores against Wikipedia references, suggesting that the newer language model may capture nuances in summaries that align more closely with the Wikipedia style. Comparing our participation in 2022 (NM-2022) to the current configurations, there is a clear advancement in performance, highlighting the rapid progression of NLP technology and model training techniques. This improvement also highlights that using long contexts.

The BERTScore improvement of MonoT5 integrated with GPT-4 over the 2022 challenge's best score for Wikipedia summaries demonstrates GPT-4's capability in producing high-caliber content, despite being provided with fewer documents — only the top-10 as opposed to the top-30 documents given to GPT-3.5. Nevertheless, compared to NIST references, the combination of MonoT5 with GPT-3.5-turbo, which utilized a larger document set for summarization, maintains its superiority.

Fact Matching

Table 4 illustrates the outcomes of the manual evaluation for redundancy and comprehensiveness. These results offer insight into the qualitative aspects of the generated summaries. The BM25 configuration demonstrated lower redundancy with a score of 0.426 and a comprehensiveness of 0.238, suggesting a balanced summary with a moderate level of detail. In contrast, the MonoT5 configuration showed significantly higher comprehensiveness at 0.321, coupled with an increase in redundancy, scoring 0.689, indicating a tendency to include more repetitive but relevant information. When compared to the mean scores of all participants, both configurations present substantial improvements, with MonoT5 surpassing the average comprehensiveness score by a notable margin, thereby underlining the effectiveness of the advanced reranking process. These observations are critical for future system enhancements, emphasizing the need for a trade-off between detailed content and information repetition.

Table 4. Manual evaluation metrics for the 2023 Crisis Event Summarization track.

Configuration	Redundancy	Comprehensiveness
BM25 + GPT-3.5	0.409	0.126
MonoT5 + GPT-3.5	0.630	0.201
Participants Mean	0.286	0.078

Examining the results from the 2022 event data, as shown in Table 5, we can discern the performance dynamics of the summarization systems. The BM25 algorithm displayed a redundancy score of 0.476 and a comprehensiveness of 0.282, indicating its summaries had more unique content with adequate detail. The MonoT5 configuration achieved a higher comprehensiveness score of 0.380, with a redundancy of 0.691, reaffirming its capability to produce detailed summaries, albeit with a propensity for including similar information.

Table 5. Manual Evaluation Results for 2022 Event Data.

Run	Redundancy	Comprehensiveness
BM25 + GPT-3.5	0.481	0.168
MonoT5 + GPT-3.5	0.703	0.229
NM-2022	0.431	0.342
2022 Best	0.079	0.342

When these figures are juxtaposed with the performance of our previous system, NM-2022, which had the best comprehensiveness of 2022 but a lower redundancy, it suggests that while our past system was able to match the depth of content of the leading systems, it did so with less repetition. However, the current MonoT5 results exceed both NM-2022 and 2022 best in comprehensiveness, showing that the integration with the GPT models and refinement in methods have led to summaries with richer information.

CONCLUSIONS

This paper reports NeuralMind.ai's participation in the TREC CrisisFACTS 2023 challenge, presenting a comprehensive evaluation of the various configurations of summarization systems, leveraging both established and state-of-the-art models in natural language processing. Our experiments utilized combinations of the BM25 algorithm, the MonoT5 re-ranking method, and two versions of OpenAI's language models: GPT-3.5-turbo and GPT-4. The MonoT5 + GPT-3.5-turbo configuration demonstrated superior performance in terms of BERTScore and ROUGE-2 metrics when compared against NIST references, indicating an appreciable year-over-year improvement in generating succinct and relevant summaries. However, when assessed against Wikipedia references, this configuration exhibited a slight decrease in BERTScore but a small increase in ROUGE-2, suggesting a balance between maintaining comprehensive content and avoiding extraneous information.

Incorporating GPT-4 into our methodology brought refined improvements, especially evident in creating Wikipedia-styled summaries where it outstripped the previous year's best performance. This superior quality was achieved

despite GPT-4 processing fewer documents — the top 10 versus the top 30 for GPT-3.5 — hinting at its enhanced comprehension capabilities. Although its enhancements did not translate into a substantial lead over GPT-3.5-turbo in BERTScore when evaluated against NIST references, GPT-4's advancements are notable, particularly given its reduced data input. These developments suggest GPT-4's potential for applications that require a deeper understanding of context, such as in knowledge synthesis and compiling comprehensive database summaries, in line with Wikipedia's detailed and factual style.

The manual evaluation further illuminated the progress in this domain, focusing on redundancy and comprehensiveness measures. While the systems have become more comprehensive over time, they also tend to produce more redundant information. This suggests a new frontier for optimization, where the balance of informative content and brevity must be further refined.

In conclusion, our findings from the TREC CrisisFACTS 2023 challenge underscore the value of combining retrieval algorithms with large language models to enhance the quality of automatic summarization in crisis scenarios. The remarkable performance improvements from our previous year's participation bear testament to the rapid evolution of NLP technology. Looking ahead, we anticipate that continued advancements in model efficiency and comprehension will open up new frontiers for real-time crisis management and response. Our future work will explore these avenues, focusing on refining our systems to better handle the dynamic and often unpredictable nature of social media data during crises.

REFERENCES

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.

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.

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.

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.

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. (2023a). "Crisis Event Social Media Summarization with GPT-3 and Neural Reranking". In: *Proceedings of the 20th International ISCRAM Conference*. University of Nebraska at Omaha, pp. 371–384.

Pereira, J., Fidalgo, R., Lotufo, R., and Nogueira, R. (2023b). "Using Neural Reranking and GPT-3 for Social Media Disaster Content Summarization". In: *Proceedings of the 31st Text Retrieval Conference (TREC 2022) Ian Soboroff and Angela Ellis, eds.*

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.

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.

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.

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*.