Multi Stage Retrieval for Web Search During Crisis

future internet

MDPI

Article

Multi Stage Retrieval for Web Search During Crisis

Claudiu Constantin Tcaciuc, Daniele Rege Cambrin and Paolo Garza *

Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129 Torino, Italy; claudiuconstantin.tcaciuc@studenti.polito.it (C.C.T.); daniele.regecambrin@polito.it (D.R.C.)
* Correspondence: paolo.garza@polito.it

Abstract: During crisis events, digital information volume can increase by over 500% within hours, with social media platforms alone generating millions of crisis-related posts. This volume creates critical challenges for emergency responders who require timely access to the concise subset of accurate information they are interested in. Existing approaches strongly rely on the power of large language models. However, the use of large language models limits the scalability of the retrieval procedure and may introduce hallucinations. This paper introduces a novel multi-stage text retrieval framework to enhance information retrieval during crises. Our framework employs a novel three-stage extractive pipeline where (1) a topic modeling component filters candidates based on thematic relevance, (2) an initial high-recall lexical retriever identifies a broad candidate set, and (3) a dense retriever reranks the remaining documents. This architecture balances computational efficiency with retrieval effectiveness, prioritizing high recall in early stages while refining precision in later stages. The framework avoids the introduction of hallucinations, achieving a 15% improvement in BERT-Score compared to existing solutions without requiring any costly abstractive model. Moreover, our sequential approach accelerates the search process by 5% compared to the use of a single-stage based on a dense retrieval approach, with minimal effect on the performance in terms of BERT-Score.

Keywords: web search; Internet and social media; text retrieval; crisis management

Academic Editor: Gianluigi Ferrari

Received: 17 April 2025
Revised: 27 May 2025
Accepted: 28 May 2025
Published: 29 May 2025

Citation: Tcaciuc, C.C.; Rege Cambrin, D.; Garza, P. Multi Stage Retrieval for Web Search During Crisis. *Future Internet* **2025**, *17*, 239. https://doi.org/10.3390/fi17060239

1. Introduction

The exponential growth of digital information, while enhancing global connectivity and facilitating access to knowledge, presents formidable challenges in efficiently retrieving relevant and accurate data. Users are frequently overwhelmed by the volume of content on various online platforms, such as social media, news outlets, and blogs, particularly when seeking timely insights during rapidly evolving events [1]. Crisis scenarios, such as natural disasters (e.g., wildfires, hurricanes, floods) [2], industrial accidents, or public health emergencies (e.g., COVID-19) [3], exemplify these dynamic contexts. In such situations, the ability to access precise and pertinent information is not only beneficial but can critically influence outcomes, including resource allocation, coordination of rescue efforts, and the effective guidance of affected populations, often making the difference in life-saving decisions.

Despite the critical need, existing information retrieval (IR) methods exhibit significant limitations when applied to crisis management. Traditional lexical search approaches, such as BM25, while computationally efficient and effective for keyword-based queries, often struggle with semantic nuances, synonymy, and the implicit meaning prevalent in crisis-related communication, leading to low precision or missed relevant documents [4]. Conversely, while recent advancements in large pre-trained natural language processing

models [5] and generative models [6,7] have demonstrated improved semantic understanding, they introduce their own set of drawbacks. These pre-trained models, used to implement dense retrieval systems, though powerful, can be computationally intensive for real-time processing of high-volume data streams. Furthermore, these models often require domain-specific fine-tuning to adapt to the specialized vocabulary and context of crisis events, as their performance can degrade when dealing with data uncommon in general-purpose training corpora [8–10]. Generative models, while capable of producing fluent text and employed in other solutions [11,12], pose risks of hallucination and information fabrication that are unacceptable in crisis scenarios [13]. Moreover, their high computational demands or reliance on commercial APIs make them impractical for many emergency response organizations operating under resource constraints.

To overcome these limitations, this paper introduces a novel multi-stage, fully extractive retrieval framework specifically designed for crisis event information management. Our approach integrates the efficiency of traditional lexical search with the semantic understanding of modern dense retrieval methods in a cascaded architecture. This design first employs a lexical retrieval stage for broad candidate generation, followed by a dense retrieval stage for precise reranking, thereby optimizing both recall and precision. This synergy allows our framework to outperform the current best competitor by 13% in BERT-Score F1-Score. Crucially, our system also achieves a 5% reduction in search latency compared to single-stage dense retrieval methods, while concurrently ensuring superior average recall. This enhanced performance provides a robust, scalable, and computationally viable solution tailored to the demands of real-world crisis response operations.

Our contributions can be summarized as follows:

- **Architectural Design for Crisis-Specific Retrieval:** We propose a novel multi-stage retrieval architecture that synergistically combines a lexical retrieval stage for efficient, broad recall with a subsequent dense retrieval stage for high-precision reranking. This cascaded design is specifically tailored to handle the high-volume, redundant, noisy, and time-sensitive data characteristic of crisis events, improving single-stage systems and generic multi-stage designs.
- **LLM Independency:** Our framework achieves state-of-the-art performance without relying on computationally expensive large language models (LLMs). This design choice prioritizes deployability and operational feasibility for resource-constrained emergency response scenarios, offering a practical alternative to LLM-based solutions while surpassing their effectiveness.
- **Algorithmic Design for Extractive Summarization Support:** The framework is designed to be fully extractive, ensuring that all retrieved information is directly traceable to source documents. This is a critical design choice for crisis management, where verifiability and trust are paramount, distinguishing our approach from abstractive methods that risk introducing unsupported information and hallucinations.
- **Comparative Analysis:** We provide a comprehensive empirical evaluation of our proposed pipeline against established single-stage lexical and dense retrieval baselines, as well as state-of-the-art models. Our results highlight a superior balance between retrieval effectiveness and computational efficiency, demonstrating the practical advantages of our architectural and algorithmic choices for the target domain.

We release our code for reproducibility at https://github.com/DarthReca/multistage-crisis-retrieval (accessed on 27 May 2025).

The paper is structured as follows. Section 2 presents the state-of-the-art solutions for the task, Section 3 presents the problem statement, the employed data, and the proposed pipeline, Section 4 presents an experimental comparison with state-of-the-art solutions and

the ablation study; Section 5 summarizes the main findings, the limitations, and the impact on the sector; and Section 6 concludes the paper by discussing future research directions.

2. Related Works

This section presents the related works in text retrieval (see Section 2.1) and deep learning and information retrieval (IR) related to crisis management (see Sections 2.2 and 2.3).

2.1. Text Retrieval

Text retrieval aims to satisfy a user's information need, typically expressed as a query, by identifying and ranking relevant documents from a large collection [14]. The general architecture of such systems involves document processing, the creation of an efficient index, query processing, a matching or ranking function to score documents against the query, and finally, the presentation of results. The inverted index is a central component linked to the performance of traditional information retrieval (IR) systems. It is a data structure that maps terms to the documents containing them. Significant research has focused on optimizing these inverted indexes for both speed and storage, employing techniques such as self-indexing to provide fast access [15] and sophisticated compression strategies coupled with optimized document ordering to manage vast document collections effectively [16].

Early retrieval models, such as the Boolean model, which rely on set theory and Boolean operators, offer precise but often rigid retrieval with no inherent ranking of documents. The Vector Space Model (VSM) [17] was a step forward, representing documents and queries as vectors in a multi-dimensional space where each dimension corresponds to a term [18]. Term weights, often calculated using TF-IDF (term frequency-inverse document frequency), reflect the importance of terms both locally within a document and globally across the collection. Relevance is then typically assessed using similarity measures like cosine similarity, allowing for partial matching and ranked retrieval. Concurrently, probabilistic models offered a different theoretical perspective, aiming to rank documents based on their estimated probability of relevance to a query. The probability ranking principle (PRP) [19] states that optimal retrieval is achieved by ranking documents in decreasing order of their probability of relevance, given all available evidence.

The Okapi BM25 ranking function [20,21] builds directly on these probabilistic foundations. It calculates document scores based on query term presence, considering term frequency saturation (recognizing that repeated terms offer diminishing returns), inverse document frequency (IDF) to emphasize rarer, more informative terms, and document length normalization. The computational efficiency of BM25 is the result of extensive algorithmic engineering for core search operations like list traversal and score accumulation, particularly for in-memory systems [22]. Beyond core term weighting, traditional IR also explored other lexical and structural cues to enhance ranking. For example, the proximity of query terms within documents or the structural zone in which terms appear (e.g., title, abstract) [23].

Given its efficiency, BM25 is well-suited as an initial high-recall retrieval stage in a multi-stage pipeline. This allows for rapid scanning of large document sets to form a broad candidate pool. However, BM25's effectiveness is primarily driven by exact term matches, so it inherently struggles to understand semantic meaning, synonyms, or paraphrased expressions. This limitation is particularly pronounced with short or ambiguous queries or when dealing with diverse and noisy vocabularies, such as those found in social media during crises, potentially leading to lower precision or missing relevant documents that use different terminology [4].

To address such limitations, recent advancements in Natural Language Processing (NLP) and deep learning have led to the development of transformer-based retrieval models [24]. These models leverage deep contextual embeddings to capture semantic relationships, converting queries and documents into high-dimensional vector representations that allow for more effective semantic matching. Transformers are also widely employed as rerankers [25], where they process query-document pairs from an initial candidate set to provide a more accurate final ranking. While these neural network-based approaches often yield significant improvements, they typically require substantial computational resources for both indexing (in the case of dense retrieval) and inference. This can be a challenge for deployment in real-time crisis scenarios. Furthermore, while powerful, if such models were extended or coupled with abstractive components for tasks like summarization, they could introduce the risk of "hallucinations"—generating plausible but factually incorrect information [13].

Hybrid approaches, which combine traditional lexical-based methods with neural retrieval models, have been explored to balance between efficiency and accuracy [26]. A common strategy involves a two-step retrieval process: an initial candidate selection using an efficient lexical model, followed by a more computationally intensive neural reranker to refine the results [25,27]. While this is an established practice, a single pass of retriever and reranker can inherently lose many relevant documents, so we propose a double pass of retriever and reranker to better process a smaller set of potentially relevant documents. Additionally, topic modeling was already explored in retrieval [28], but to assign a topic to a query rather than filtering the corpus from noise, as we conversely propose in our approach. Our extractive multi-stage pipeline, which avoids abstractive generation, offers a promising avenue to meet the demands of crisis responders by maximizing recall of potentially relevant information in early stages while systematically refining precision and ensuring factual grounding in later stages without sacrificing scalability.

2.2. Deep Learning in Crisis Management

Integrating deep learning techniques in crisis management has emerged as a pivotal area of research, driven by the need for rapid, accurate, and scalable solutions to handle complex and dynamic emergency situations. Deep learning has been applied to various domains, including natural disaster response [29,30], public health emergencies [31], and humanitarian crises. These applications often involve the analysis of large-scale, heterogeneous data sources, such as social media feeds [4], satellite imagery [29,30], and sensor data [32] to extract actionable insights and support decision-making.

One of the key advantages of deep learning in crisis management is its ability to process and analyze unstructured data, such as text and images, which are prevalent in emergency situations. However, applying deep learning in crisis management also presents several challenges, including data scarcity and bias. The dynamic and unpredictable nature of crises often results in limited or imbalanced training data, which can affect the performance and generalizability of deep learning models.

For this reason, some ad-hoc models were proposed to face these needs [4,33], and while general-purpose models can still be employed, they can struggle to understand the crisis context fully.

2.3. Retrieval in Crisis Management

Crisis situations like natural disasters, industrial accidents, and public health emergencies require fast and accurate information retrieval to support decision-making. Social media platforms, including Twitter, Facebook, and Reddit, have become crucial real-time information sources [34]. However, the high volume of unstructured, redundant, and often

contradictory data presents significant challenges for retrieval systems. Another challenge in crisis information retrieval is the presence of misinformation and redundant content.

CrisisFACTS [35,36] provides the first attempt at creating a corpus tailored for crisis management with all the mentioned challenges. During the two editions of the challenge, various solutions were proposed to solve a retrieval and summarization task, ranging from extractive [37,38], reinforcement-learning-based [39], and LLM-based [11,12] methods. The first edition highlighted the need to improve the recall of the designed systems [35]. While abstractive solutions can improve the quality of the writing, the retriever should be able to recover all relevant documents to avoid definitively losing important facts during a crisis. This work focuses on a full extractive solution that can surpass even the most recent abstractive solutions thanks to a retrieval pipeline tailored to recover as many relevant documents as possible with a multi-stage approach.

3. Materials and Methods

In this section, we first present the task and dataset. Then, we discuss the employed framework.

3.1. Okapi BM25 Ranking Scheme

Okapi BM25 (Best Match 25) is a widely adopted ranking function used in information retrieval systems to estimate the relevance of documents to a given search query [21]. It is a probabilistic model that extends earlier methods like TF-IDF (Term Frequency-Inverse Document Frequency) by incorporating term frequency saturation and document length normalization. The BM25 score for a document D with respect to a query Q, which is composed of individual query terms $q_1, q_2, \ldots, q_n$, is calculated as follows:

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{1}$$

where:

- $\text{IDF}(q_i)$ is the Inverse Document Frequency of the query term q_i. It measures the general importance of the term in the collection. A common formula for IDF is

$$\text{IDF}(q_i) = \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \tag{2}$$

  Here, N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing the term q_i. The addition of 1 inside the logarithm (or other variants like adding 0.5 to the numerator and denominator) is to prevent division by zero for terms not present in any document or to smooth the IDF values.
- $f(q_i, D)$ represents the frequency of term q_i in document D (i.e., term frequency).
- $|D|$ is the length of document D, measured as the total number of words (or tokens) in it.
- avgdl is the average document length across the entire collection.
- k_1 is a positive tuning parameter that controls the term frequency saturation. It dictates how quickly the contribution of a term's frequency to the score diminishes as the frequency increases. A typical value for k_1 is between 1.2 and 2.0.
- b is another tuning parameter, usually between 0 and 1 (a common default is 0.75), which determines the degree of document length normalization. When $b = 1$, the scaling by document length is fully applied, and when $b = 0$, no length normalization is applied.

Data Structures

Efficient calculation of BM25 scores relies on the following structures:

- **Inverted Index:** This is the primary data structure. It maps each unique term in the collection to a list (postings list) of documents that contain that term. Each entry in the postings list often stores the document ID and the term frequency $f(q_i, D)$ within that document. The document frequencies $n(q_i)$ (number of documents containing term q_i) can be derived from the length of these posting lists.
- **Document Length Storage:** An array or map is typically used to store the length $|D|$ for every document in the collection, indexed by document ID.
- **Corpus Statistics Storage:** Variables to hold global statistics such as the total number of documents N and the average document length avgdl.

3.2. Problem Statement

We address the CrisisFACTS 2023 task first proposed in Buntain et al. [35]. Given a crisis event, it has associated timestamped documents D, each one belonging to a stream $s \in S = \{\text{tweets, Reddit messages, news, Facebook posts}\}$. The CrisisFACTS task aims at selecting a list of documents $R \subseteq D$ relevant to generating a daily summary as shown in Figure 1 for an emergency responder. The summary should cover all facts happening during the day without redundant facts. A set Q of event queries that interest emergency responders is optionally provided as guidance to retrieve R, but they do not necessarily have to be used and can be extended.

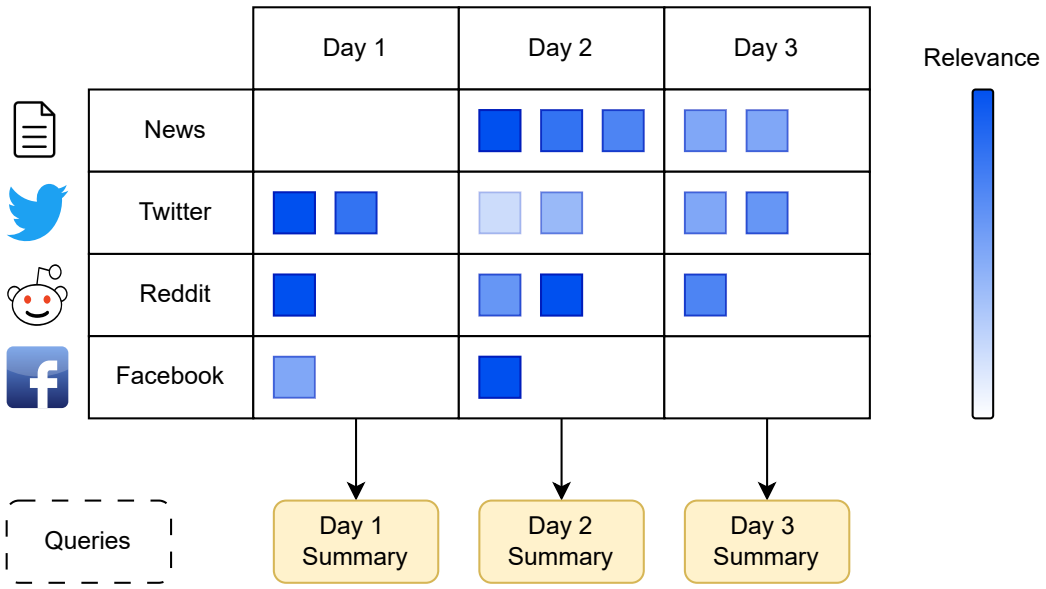


Figure 1. During each day of an event, the system has to select all relevant documents (blue squares in the figure) from the streams (News, Twitter, Reddit, and Facebook) and, at the end of each day, compose a summary using the selected documents.

3.3. Dataset

The CrisisFACTS 2023 dataset [35,36] comprises a series of documents from different sources for different catastrophic events (see Table 1). The events are tornadoes, wildfires, hurricanes, floods, and various accidents. They last for more than one day, and a minimum of 20,000 documents are associated. These documents are produced and deposited on the Internet daily during an event. The data in the collection is only weakly filtered (e.g., based on hashtags and keywords). For this reason, there is a lot of noise and irrelevant documents. The following streams are available:

1. *News*: News articles are an excellent source of information during catastrophic events, and a small number of pieces were included in the dataset.
2. *Twitter*: Twitter posts were collected based on keywords relevant to the events in the analysis.
3. *Reddit*: Top-level posts with all subsequent comments were extracted from Reddit threads relevant to the emergency and included in the dataset.
4. *Facebook*: Facebook posts from public pages are provided based on relevance to each disaster event.

The first eight events were used during the 2022 competitions and can be used for parameter tuning during the 2023 edition. The evaluation set is composed of events from 009 to 018 (see Table 1).

Table 1. Description of CrisisFACTS events. Only the facts from 009 were evaluated in the 2023 edition.

Event ID	Name	Queries	Texts	Days
001	Lilac Wildfire 2017	52	51,015	9
002	Cranston Wildfire 2018	52	30,535	6
003	Holy Wildfire 2018	52	32,489	7
004	Hurricane Florence 2018	51	376,537	15
005	Maryland Flood 2018	48	41,770	4
006	Saddleridge Wildfire 2019	52	38,369	4
007	Hurricane Laura 2020	51	62,182	2
008	Hurricane Sally 2020	51	116,303	8
009	Beirut Explosion, 2020	56	468,178	7
010	Houston Explosion, 2020	56	72,530	7
011	Rutherford TN Floods, 2020	48	20,868	5
012	TN Derecho, 2020	49	79,688	7
013	Edenville Dam Fail, 2020	48	28,185	7
014	Hurricane Dorian, 2019	51	556,233	7
015	Kincade Wildfire, 2019	52	137,072	7
016	Easter Tornado Outbreak, 2020	51	131,954	5
017	Beirut Explosion, 2020	51	120,899	6
018	Tornado Outbreak, 2020 March	51	200,008	6

The dataset is also supplemented with queries (set Q) defining the information needs of disaster-response stakeholders extracted from FEMA ICS 209 forms. As reported in the problem statement, the set Q is optionally used. These queries capture what a responder might consider important, such as emerging/approaching threats, risks from hazardous materials, damage to key infrastructure, evacuations or emerging threats, statistics on casualties or numbers missing, weather concerns, and restrictions on or availability of resources. Some example queries and some related relevant documents are reported in Table 2.

Table 2. Sample queries and relevant documents from the corpus.

Query	Relevant Document
What roads are closed?	Hwy 76 is closed both directions...
How many people have been injured?	Two civilians are being treated for burn injuries...
Where are wind speeds expected to be high?	...San Diego County as winds gust to 75 mph.

As ground truths, two summaries for each event are provided:

1. Wikipedia. It is the Wikipedia page for the given event, providing a high-level description.
2. NIST Summary. It is created by NIST assessors from the documents present in the corpus. It is more detailed and tailored for the task.

It is important to note that the NIST summary is built directly from the corpus and has no direct relation to the provided queries. They are constructed as guidance from the FEMA form.

3.4. Framework

This section presents the Multi-Stage Crisis Retriever (MSCR) as depicted in Figure 2. It comprises the preprocessing, topic filtering, and double retrieval pipeline, before the final summary generation.

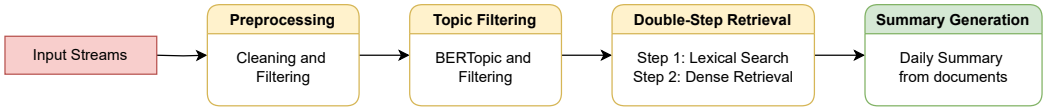


Figure 2. At the end of each day, the streams are preprocessed (e.g., cleaning and duplicate removal). Then, we filter topics relevant to the crisis using BERTopic before feeding the documents into a cascade of lexical and dense searches. The selected documents are used to create the final summary.

3.5. Preprocessing

The corpus is preprocessed as shown in Figure 3. We started by sanitizing each document. Since much of the data come from social networks, hashtags, emojis, and links are prevalent in many texts, so we removed those characters and extra spaces. This helps adapt general-purpose NLP models that were not created for such noisy texts. Sanitizing the text also made similar texts that only differ in different links, emojis, and extra spaces equal to each other.

The initial filter was done by removing duplicated documents, short texts (less than two words), and documents that contain question marks. Short text documents were removed because, in most cases (e.g., event 004 text: "Following... https:..."), they do not provide any important information or contain only the event's name. Questions can be considered uninformative since they do not provide a fact. For this research, we loosely consider all documents containing a question mark as questions. Although an intent model can be used, it is time-consuming in the initial phase, which involves thousands of documents.

Figure 3. Preprocessing pipeline. Each text in the corpus is sanitized, and then duplicates, short texts, and questions are removed.

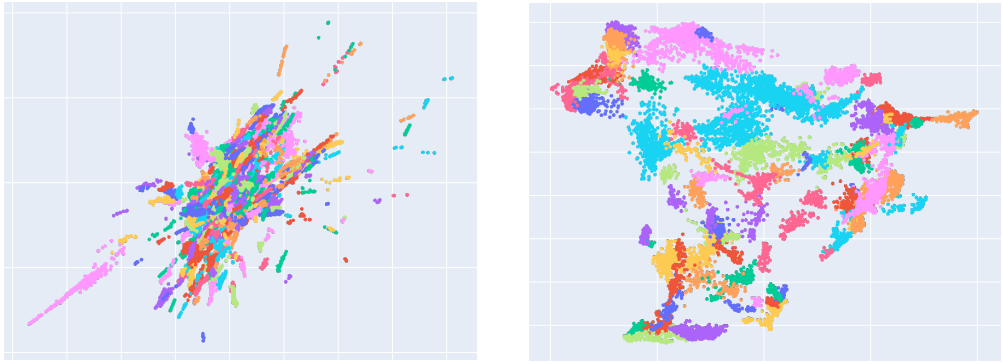
3.6. Topic Filtering

Since the data are unfiltered, there can be unrelated documents for the event under analysis. At this point, we would like to exclude all documents belonging to topics unrelated to the crisis. To this end, we leverage BERTopic [40] as a topic modeling solution to create clusters of documents using dense embeddings as shown in Figure 4. BERTopic represents each cluster/topic through a set of representative keywords.

Each event in the dataset has a short description, so we extracted the keywords using KeyBERT [41] and embedded them, and, finally, we compare these embeddings with the ones of the representative keywords of each cluster. We select a topic only if its similarity score with the short description is higher than the q-th percentile (i.e., 50) of all the scores computed for all topics in the event-day dataset and higher than a predetermined threshold (i.e., 0.5). For the Event 009, for example, we have the following description:

> On 4 August 2020, a large amount of ammonium nitrate stored at the port of the city of Beirut, the capital of Lebanon, accidentally exploded, causing at least 180 deaths, 6000 injuries, US$10–15 billion in property damage, and leaving an estimated 300,000 people homeless.

So we can select topics as seen in Table 3 from the Event 009. This way, we can remove noise clusters, like the ones about fear, recommendations, and news updates. In contrast, we can still retain information about the causes, general information, and ammonium nitrate (linked to the explosion according to the description). This can also be exploited to understand the topics during an event.



(**a**) Topics before filtering (**b**) Topics after filtering

Figure 4. Representation of document embeddings using UMAP grouped by topic by BERTopic before and after the topic filtering. Each color represents a different topic.

Table 3. Example of topic selection for event 009. Topic describes the cluster with keywords obtained from KeyBERT.

Topic	Score (%)	Selected
beirut explosion, beirut blast, beirut, beirut lebanon	75.00	✓
tonnes ammonium nitrate, tons ammonium nitrate	67.34	✓
cause explosion unknown, cause explosion unclear	62.43	✓
news updates, page latest updates, latest updates	42.72	✗
hope stay safe, hope okay safe, hope safe okay	43.71	✗
really scary, damn scary, terrifying scary	44.64	✗

3.7. Additional Queries

The dataset is supplemented with the set of queries Q that defines the basic information needed for a disaster-response stakeholder and is extracted from the FEMA ICS 209 forms. However, the queries in Q are general and lack strong relations with the event under analysis. So, we reformulate the queries by adding more context (e.g., "Have railways closed" becomes "Are railways disrupted or closed because of Hurricane?"). This task was delegated to a language model using the following prompt:

> *I need to refactor some queries in order to better capture the information by adding more context to them <LIST OF QUERIES>.*

Another challenge related to the queries was their inability to cover all aspects of the crisis event. To address this, we generated multiple candidate questions using a language model. We split the set of queries into different groups based on the event type, and we ended up with general and event-specific queries. Afterwards, for each group, we generated additional queries using the following prompt:

Can you analyze this list of queries and add at least 30 more that are related to the event <EVENT TYPE/GENERAL QUERIES> but are not present in the following list? <LIST OF QUERIES>

Subsequently, we manually selected the most fitting queries for each group, which expanded the semantic coverage for a catastrophic event.

In total, we incorporated 20 general queries addressing broad concerns common to crisis situations and 10 event-specific queries related to the characteristics of each particular event. We limit the number of generated queries to avoid dealing with an overwhelming number of them. The following are some examples of general queries (<EVENT> is replaced dynamically based on the event):

- Are there any security risks or criminal activities occurring due to the <EVENT>?
- What medical services are available for affected individuals?
- Are ATMs and banks functioning after the <EVENT>?

The following are some examples of event-specific queries:

- Tornado: "Are there any reports of multiple tornadoes forming during the tornado?"
- Wildfire: "What weather conditions are affecting the wildfire's spread?"
- Hurricane: "How far inland is flooding occurring due to the hurricane?"
- Flood: "Are any dams at risk of overflowing or failure due to the flood?"
- Storm: "How long is the storm expected to last?"
- Accident: "Has the air quality been affected by the accident?"

3.8. Multi Stage Retrieval

Hybrid retrievers combine the strengths of both lexical and dense search methods. In our pipeline (Figure 5), these search methods are used sequentially: first, a lexical search selects the top-n elements for each query, then the selected documents of all queries are aggregated into a filtered corpus, and then a dense search further refines this selection to the top-m elements (where n > m). Each search step comprises a retrieval step (i.e., BM25 or Dense Retriever), followed by a cutoff and a light cross-encoder to rerank the documents, followed by another cutoff, as shown in Figure 5. This way, we reduce the documents we must deal with step by step. The reranker in lexical search ranks the most relevant documents higher and prevents them from being excluded by the second cutoff of the lexical search step. Hence, only the most relevant subset of documents returned by the first phase is provided as input to the dense search.

In the lexical search step, after applying BM25, we selected the top 200 for each query. This threshold was chosen based on empirical observation across development to balance computational costs and performance. After the reranking, we select the top 60 documents for each query to pass them to the next stage. This was done to balance the execution time of the dense search step while maintaining a high recall in the final result. After the lexical search, the filtered corpus aggregates 60 documents times the number of queries in Q in the worst case (duplicates are removed). The dense retriever selects the top 200 (or the top 10% of the corpus size if it is less than 200). This conditional threshold ensures scalability. Moreover, in a small data corpus, it prevents overfitting, while in a large corpus, it ensures computational feasibility. After the final reranking, we keep only the top 10% of the documents. This threshold was determined through ablation experiments to balance retrieval precision and semantic diversity.

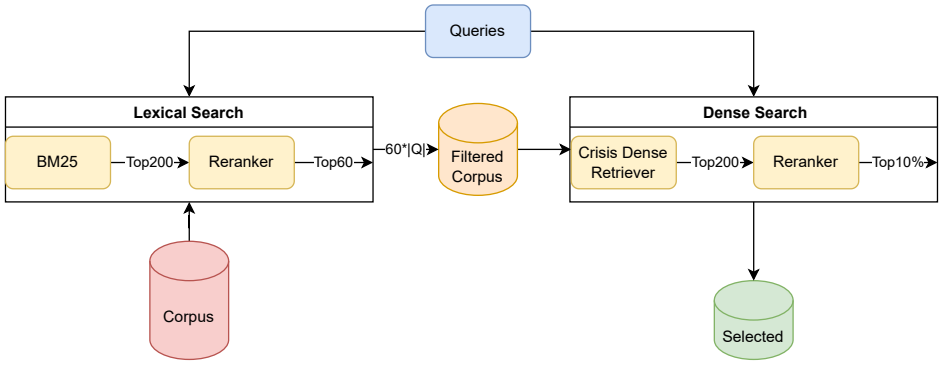


Figure 5. The corpus is initially filtered using BM25 and a general-purpose reranker. Then, a crisis-specific model is used for dense retrieval, followed by a general-purpose reranker.

The reranking of the dense search is essential, as its score is used to construct the final summary, which is detailed in the subsequent section.

This hybrid approach allows us to initially leverage a fast BM25 for broad filtering, followed by a more precise reranking of a subset using a general-purpose cross-encoder. We then employ a crisis-pre-trained model for dense retrieval and rerank a small document set.

While developing a robust reranker specifically for crisis-related documents is challenging, this method enables us to harness the strengths of both a general, powerful reranker and a model with embeddings tailored to the task at hand.

3.9. Summary

To generate the final summary, we choose the highest-ranked 50% of the documents. Then, we apply a similarity filter to reduce redundancy. If two documents have a cosine similarity higher than 0.95, they are considered duplicates, and one is dropped. Each summary comprises a ranked list of facts. In our case, each fact is a maximum of 500 characters and comprises up to the top 7 documents answering the same query. The importance score of the fact is the mean of the documents.

4. Results

In this section, we present the experimental results. We first compare all the available event results to the provided baseline and the other competitors.

4.1. Experimental Settings

We use CT-mBERT-SE [4] to create textual embeddings. BERTopic [40] employs UMAP [42] for dimensionality reduction and HDBSCAN. The reranker is mxbai-rerank-xsmall-v1 (https://huggingface.co/mixedbread-ai/mxbai-rerank-xsmall-v1 (accessed on 27 May 2025)). Only manual evaluation and the BERT score were reported in the original paper [35], and this last metric was used to compare with other solutions. However, as shown in Buntain et al. [35], the BERT score should be sufficient since it correlates well with the other metrics (even the manual ones). We evaluate the experiments over three runs, and we report the mean and standard deviation of the performance. Since competitors do not provide code to reproduce the results, detailed results for each event, or multiple runs, only the mean value is reported for each competitor (The reported values for the competitors are the official results reported by the competition organizer.)

4.2. Comparative Results

In this section, we analyze the performance compared to other solutions using BERTScore, as done in the original paper [35].

4.2.1. Competitors

We evaluated the top two abstractive solutions presented at the competition (nm-gpt35 and llama 13b chat) and also the top two extractive solutions (FM-B and occams). The nm-gpt35 approach [11] comprises BM25, followed by a reranker as a search engine. For each query, the top-k documents are selected using the search engine, and then GPT-3.5 [7] is prompted to aggregate the documents to select and rank the documents to produce a list of facts. The llama 13b chat-based solution [12] is based on a search engine composed of BM25, followed by a reranker. Then, a language model (Llama-13b [6]) is asked to provide the answer for each query and the documents supporting the answer. Regarding the extractive approaches, FM-B [37] employs REBEL [43] and ClausIE [44] for fact extraction, filters facts using indicative terms, and ranks them using an integrated content-graph analysis. The occams extract method [38] approaches extractive summarization as a weighted bounded maximum coverage problem. This means it aims to select the most representative sentences from a set of documents while adhering to a predefined summary length constraint. It uses term weights to determine the importance of different terms in the documents.

4.2.2. Comment on the Results

Table 4 reports the BERT-Score results of the baseline proposed at the competition, our approach, and the top two extractive (E) and the top two abstraction (A) solutions. Looking at Table 4, the extractive solutions perform better than the baseline but still struggle against abstractive solutions. Both extractive solutions get a similar score in both NIST and Wikipedia, surpassing the baseline by +0.8 and +2.5 in terms of F1, respectively. llama 13b chat performs well in both cases (F1: +4.9 on NIST and +2.8 on Wiki compared to the baseline), while nm-gpt35 performs a little worse than llama 13b chat. While using an LLM-based abstraction approach seems appealing, it requires more resources and time than extractive solutions. Additionally, as suggested by our proposal, a good retrieval system is the key to not losing any relevant documents. Once lost, neither a language model can recover them. An example of this can be seen when comparing nm-gpt35 with FMB or occams extract. The extractive solutions perform better than the abstraction nm-gpt35 one. Although simple, our solution (MSCR in Table 4) improves the performance in terms of BERT-Score in both cases (R: +13 on NIST and +19 on Wiki compared to llama 13b chat) without requiring any abstractive step. The lexical search used by our approach is fast and effective in selecting a good candidate set of relevant documents, while the dense search step allows refining the search and highlighting the most relevant documents from the subset returned by the lexical search phase.

Table 4. BERT-Score. Comparison between baseline, top-2 abstractive (A) methods, top-2 extractive (E) methods, and our approach (MSCR). The NIST column is associated with the results achieved using the NIST-based summary as a ground truth. In contrast, the Wiki column is related to the results obtained using the Wikipedia-based summary as a ground truth.

		NIST			**Wiki**		
Run	**Type**	**F1**	**P**	**R**	**F1**	**P**	**R**
baseline.v1 [35,36]	E	59.11	58.95	59.33	50.84	47.58	54.82
FM-B [37]	E	59.94	59.94	59.99	53.31	51.00	55.95
occams extract [38]	E	59.99	59.55	60.48	53.31	48.94	55.45
nm-gpt35 [11]	A	61.45	62.51	60.48	51.14	48.49	54.42
llama 13b chat [12]	A	65.00	64.74	65.29	53.63	50.95	56.69
MSCR	E	**77.76 ± 2.54**	**76.71 ± 2.93**	**78.89 ± 2.58**	**72.69 ± 2.09**	**70.02 ± 3.19**	**75.64 ± 1.72**

4.3. Ablation Study

In this section, we compare our multi-stage approach with two single-stage methods. Specifically, we considered as competitors a single-stage approach based on the lexical

search block only (Lexical) and another one based on the dense search one only (Dense). In Table 5, we report the performance of the single-stage approaches and that of our multi-stage approach. We also report the mean run time per event. Lexical is 10% faster than our multi-stage approach. While the first stage of our solution is still based on the lexical block, our approach can refine all the BERT-Score metrics using the second step. Specifically, the higher recall confirms that our approach, compared to Lexical, can recover some missing documents using the second step. The single-stage Dense approach is 5% slower than our multi-stage approach (and 15% slower than Lexical). While it could be effective in improving the final precision (P), the recall (R) remains practically invariant compared to our solution. From this analysis, we can conclude that our multi-stage approach is beneficial in this case, as it speeds up the search compared to the single-stage Dense approach, maintaining the same quality in terms of Bert-Score and improving the ones of the single-step lexical search only.

Table 5. Execution time and BERT-Score. Comparison between multi-stage and single-stage approaches. * indicates statistically significant difference with $p < 0.05$ compared to MSCR.

Search Mode	Time (s)	NIST			Wiki		
		F1	P	R	F1	P	R
Lexical (single-stage)	913.50 ± 35.83 *	77.47 ± 3.56 *	76.54 ± 4.17	78.46 ± 3.22 *	72.62 ± 3.10	70.00 ± 4.15	75.52 ± 2.59
Dense (single-stage)	1091.89 ± 1.09 *	77.96 ± 2.57	76.83 ± 2.96	79.16 ± 2.68	72.66 ± 1.97	69.95 ± 3.05	75.67 ± 3.94
MSCR	1023.99 ± 4.16	77.76 ± 2.54	76.71 ± 2.93	78.89 ± 2.58	72.69 ± 2.09	70.02 ± 3.19	75.64 ± 1.72

4.4. Results at the Event Level

This section reports detailed information about the performance for each of the test events using BERT-Score (see Table 6). The BERT-Score metrics are generally higher for the NIST-based ground truth than for the Wikipedia-based one. This suggests that the model performs better on NIST in terms of capturing semantic similarity between the generated and reference summaries. This is expected since Wikipedia summaries are generic and not built from the corpus [36]. The scores for NIST are relatively consistent across different events, with F1 scores ranging from 76.19 to 81.72. This indicates stable performance across various scenarios. In contrast, the Wikipedia-based results show more variability in precision. Precision (P) and Recall (R) are balanced for NIST, with recall generally being slightly higher. This suggests that the model is capturing most of the relevant information (high recall) while maintaining good precision. Precision is generally lower than recall for Wikipedia, indicating that the model might include some irrelevant information in the summaries.

Table 6. Comparison of NIST and Wiki BERTScore scores for different events. Wikipedia summaries are missing for events 011 and 012.

Event	NIST			Wikipedia		
	F1	P	R	F1	P	R
009	79.35 ± 1.37	77.10 ± 2.11	81.74 ± 0.65	73.30 ± 0.96	72.52 ± 1.40	74.09 ± 0.63
010	76.17 ± 2.00	74.63 ± 2.92	77.81 ± 1.09	72.72 ± 2.27	70.70 ± 3.06	74.88 ± 1.37
011	78.86 ± 1.05	79.02 ± 1.04	78.70 ± 1.12	—	—	—
012	78.52 ± 1.29	77.63 ± 2.01	79.43 ± 0.77	—	—	—
013	76.22 ± 1.61	75.17 ± 2.13	77.30 ± 1.06	68.67 ± 1.18	65.48 ± 1.51	72.19 ± 0.77
014	76.41 ± 0.51	78.41 ± 0.48	74.51 ± 0.67	75.93 ± 0.24	76.38 ± 0.34	75.48 ± 0.15
015	78.48 ± 0.48	78.30 ± 0.82	78.67 ± 0.14	73.39 ± 0.19	70.27 ± 0.50	76.80 ± 0.24
016	77.01 ± 0.82	74.85 ± 0.96	79.30 ± 0.72	72.44 ± 0.30	70.87 ± 0.50	74.08 ± 0.11
017	80.26 ± 0.38	78.22 ± 0.47	82.40 ± 0.28	71.04 ± 0.47	65.27 ± 0.53	77.95 ± 0.49
018	81.45 ± 0.79	80.96 ± 0.81	81.94 ± 0.78	74.32 ± 0.31	70.93 ± 0.48	78.06 ± 0.38
Mean	78.27 ± 1.71	77.43 ± 1.92	79.18 ± 2.29	72.73 ± 2.03	70.30 ± 3.38	75.44 ± 1.92

5. Discussion

This section summarizes our findings, their impact on the sector, and their limitations.

5.1. Summary of Findings

The results demonstrate that our proposed multi-stage approach enhances the effectiveness of information retrieval and summary generation when applied to streams of crisis documents. Our multi-stage approach, which integrates traditional lexical methods with modern dense retrieval techniques, outperformed existing solutions by achieving a 13% improvement in BERT-Score F1-Score. These findings underscore the framework's ability to efficiently retrieve relevant, non-redundant facts, which is crucial for timely decision-making in crisis management.

5.2. Comparison with Existing Methods

Compared to existing solutions, our extractive methods proved effective without requiring the aid of expensive LLMs, highlighting the need for a strong retriever before trying to improve the quality of the final summary. Compared to a single-stage lexical-based approach, our hybrid approach offers superior semantic understanding and recall, which is essential for capturing the nuances of crisis-related information. While dense retrieval models provide high accuracy, they often require substantial computational resources. Our framework mitigates this by employing a sequential retrieval process, ensuring efficiency and effectiveness. This balance is particularly beneficial in real-time crisis scenarios where rapid information processing is paramount.

5.3. Impact on Crisis Management

Swiftly and accurately retrieving relevant information is critical in crisis management, where decisions can significantly impact outcomes, including life-saving measures. Our framework's enhanced retrieval and summarization capabilities can support emergency responders by providing timely access to precise information, aiding resource allocation, coordinating rescue efforts, and public communication. By reducing misinformation and redundancy, our approach helps make informed decisions, improving crisis response effectiveness.

5.4. Challenges and Limitations

Although the framework optimizes resource usage by combining lexical and dense search methods, the dense search component still requires substantial computational power. This can be a limitation in resource-constrained environments, such as areas with limited infrastructure during a crisis. While the framework improves search speed, real-time processing of large-scale data during rapidly evolving crisis situations remains challenging. As highlighted before, the preprocessing steps, including sanitization and filtering, might not capture all nuances of the data, potentially leading to the loss of some relevant information. The framework leverages pre-trained models. The performance and availability of these models can affect the overall effectiveness of the framework.

6. Conclusions

In conclusion, this paper presents a robust and scalable solution for information retrieval and summarization during crisis events. By leveraging a multi-stage approach, we have demonstrated better retrieval and summarization accuracy and good speed performance, which are critical for effective crisis management. In future work, integrating abstractive models into the retrieval pipeline could enhance the quality of summaries

generated from retrieved information. Exploring more efficient reranking mechanisms and refining the topic-filtering process could also improve the framework's performance.

Author Contributions: Conceptualization, C.C.T. and D.R.C.; methodology, C.C.T.; software, C.C.T.; validation, D.R.C. and P.G.; investigation, C.C.T.; writing—original draft preparation, D.R.C.; writing—review and editing, D.R.C. and P.G.; visualization, D.R.C. and C.C.T.; supervision, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Heiss, R.; Nanz, A.; Matthes, J. Social Media Information Literacy: Conceptualization and Associations with Information Overload, News Avoidance and Conspiracy Mentality. *Comput. Hum. Behav.* **2023**, *148*, 107908. [CrossRef]
2. Magnusson, M. Information Seeking and Sharing During a Flood: A Content Analysis of a Local Government's Facebook Page. In Proceedings of the European Conference on Social Media: ECSM, Brighton, UK, 10–11 July 2014; Academic Conferences International Limited: Oxfordshire, UK, 2014; p. 169.
3. World Health Organization. *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice*; World Health Organization: Geneva, Switzerland, 2018.
4. Lamsal, R.; Read, M.R.; Karunasekera, S. Semantically Enriched Cross-Lingual Sentence Embeddings for Crisis-related Social Media Texts. In Proceedings of the International ISCRAM Conference, Münster, Germany, 25–29 May 2024. [CrossRef]
5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Hyatt Regency, MN, USA, 2–7 June 2019*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
6. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
7. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
8. Huang, D.; Cole, J.M. Cost-efficient domain-adaptive pretraining of language models for optoelectronics applications. *J. Chem. Inf. Model.* **2025**, *65*, 2476–2486. [CrossRef] [PubMed]
9. Lu, R.S.; Lin, C.C.; Tsao, H.Y. Empowering large language models to leverage domain-specific knowledge in E-learning. *Appl. Sci.* **2024**, *14*, 5264. [CrossRef]
10. Priya, S.; Bhanu, M.; Roy, S.; Dandapat, S.K.; Chandra, J. Multi-source domain adaptation approach to classify infrastructure damage tweets during crisis. *Int. J. Data Sci. Anal.* **2025** . [CrossRef]
11. Pereira, J.; Nogueira, R.; Lotufo, R.A. Large Language Models in Summarizing Social Media for Emergency Management. In Proceedings of the Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, 14–17 November 2023; Soboroff, I., Ellis, A., Eds.; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2023; Volume 1328.
12. Seeberger, P.; Riedhammer, K. Multi-Query Focused Disaster Summarization via Instruction-Based Prompting. In Proceedings of the Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, 14–17 November 2023; Soboroff, I., Ellis, A., Eds.; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2023; Volume 1328.
13. Jiao, J.; Park, J.; Xu, Y.; Sussman, K.; Atkinson, L. SafeMate: A Modular RAG-Based Agent for Context-Aware Emergency Guidance. *arXiv* **2025**, arXiv:2505.02306.
14. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008. [CrossRef]
15. Moffat, A.; Zobel, J. Self-indexing inverted files for fast text retrieval. *ACM Trans. Inf. Syst.* **1996**, *14*, 349–379. [CrossRef]

16. Yan, H.; Ding, S.; Suel, T. Inverted index compression and query processing with optimized document ordering. In Proceedings of the 18th International Conference on World Wide Web. ACM, 2009, WWW '09, Madrid, Spain, 20–24 April 2009. [CrossRef]
17. Berry, M.W.; Drmac, Z.; Jessup, E.R. Matrices, Vector Spaces, and Information Retrieval. *SIAM Rev.* **1999**, *41*, 335–362. [CrossRef]
18. James, N.T.; Kannan, R. A survey on information retrieval models, techniques and applications. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2017**, *7*, 16–19. [CrossRef]
19. Robertson, S.E. The probability ranking principle in IR. *J. Doc.* **1977**, *33*, 294–304. [CrossRef]
20. Robertson, S.E.; Walker, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Organised by Dublin City University*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 232–241.
21. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends® Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
22. Transier, F.; Sanders, P. Engineering basic algorithms of an in-memory text search engine. *ACM Trans. Inf. Syst.* **2010**, *29*, 1–37. [CrossRef]
23. Akritidis, L.; Katsaros, D.; Bozanis, P. Improved retrieval effectiveness by efficient combination of term proximity and zone scoring: A simulation-based evaluation. *Simul. Model. Pract. Theory* **2012**, *22*, 74–91. [CrossRef]
24. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.T. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6769–6781. [CrossRef]
25. Nogueira, R.; Yang, W.; Cho, K.; Lin, J. Multi-Stage Document Ranking with BERT. *arXiv* **2019**, arXiv:1910.14424.
26. Lin, J.; Ma, X.; Lin, S.C.; Yang, J.H.; Pradeep, R.; Nogueira, R. 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, New York, NY, USA, 11–15 July 2021; SIGIR '21, pp. 2356–2362. [CrossRef]
27. Bhat, R.A.; Sen, J.; Murthy, R.; P, V. UR2N: Unified Retriever and ReraNker. In Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 595–602.
28. Rezaei, M.R.; Hafezi, M.; Satpathy, A.; Hodge, L.; Pourjafari, E. AT-RAG: An Adaptive RAG Model Enhancing Query Efficiency with Topic Filtering and Iterative Reasoning. *arXiv* **2024**, arXiv:2410.12886.
29. Cambrin, D.R.; Colomba, L.; Garza, P. CaBuAr: California burned areas dataset for delineation [Software and Data Sets]. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 106–113. [CrossRef]
30. Rege Cambrin, D.; Garza, P. QuakeSet: A Dataset and Low-Resource Models to Monitor Earthquakes through Sentinel-1. In Proceedings of the International ISCRAM Conference, Münster, Germany, 25–29 May 2024. [CrossRef]
31. Zhao, X.; Wang, G. Deep Q networks-based optimization of emergency resource scheduling for urban public health events. *Neural Comput. Appl.* **2022**, *35*, 8823–8832. [CrossRef]
32. Chamola, V.; Hassija, V.; Gupta, S.; Goyal, A.; Guizani, M.; Sikdar, B. Disaster and Pandemic Management Using Machine Learning: A Survey. *IEEE Internet Things J.* **2021**, *8*, 16047–16071. [CrossRef]
33. Long, Z.; McCreadie, R.; Imran, M. CrisisViT: A Robust Vision Transformer for Crisis Image Classification. In Proceedings of the 20th International Conference on Information Systems for Crisis Response and Management, Omaha, NE, USA, 28–31 May 2023; University of Nebraska at Omaha (USA): Omaha, NE, USA, 2023. [CrossRef]
34. Acikara, T.; Xia, B.; Yigitcanlar, T.; Hon, C. Contribution of Social Media Analytics to Disaster Response Effectiveness: A Systematic Review of the Literature. *Sustainability* **2023**, *15*, 8860. [CrossRef]
35. Buntain, C.; Hughes, A.L.; McCreadie, R.; Horne, B.D.; Imran, M.; Purohit, H. CrisisFACTS 2023-Overview Paper. In Proceedings of the Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, 14–17 November 2023; Soboroff, I., Ellis, A., Eds.; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2023; Volume 1328.
36. McCreadie, R.; Buntain, C. CrisisFacts: Building and evaluating crisis timelines. In Proceedings of the 20th International Conference on Information Systems for Crisis Response and Management. University of Nebraska at Omaha (USA), Omaha, NE, USA, 28–31 May 2023. [CrossRef]
37. Salemi, H.; Senarath, Y.; Sharika, T.S.; Gupta, A.; Purohit, H. Summarizing Social Media & News Streams for Crisis-related Events by Integrated Content-Graph Analysis: TREC-2023 CrisisFACTS Track. In Proceedings of the Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, 14–17 November 2023; Soboroff, I., Ellis, A., Eds.; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2023; Volume 1328.
38. Burbank, V.; Conroy, J.M.; Lynch, S.; Molino, N.P.; Yang, J.S. Fast Extractive Summarization, Abstractive Summarization, and Hybrid Summarization for CrisisFACTS at TREC 2023. In Proceedings of the Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, 14–17 November 2023; Soboroff, I., Ellis, A., Eds.; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2023; Volume 1328.
39. Rege Cambrin, D.; Cagliero, L.; Garza, P. DQNC2S: DQN-Based Cross-Stream Crisis Event Summarizer. In *Advances in Information Retrieval*; Springer Nature: Cham, Switzerland, 2024; pp. 422–430. [CrossRef]

40. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
41. Grootendorst, M. KeyBERT: Minimal Keyword Extraction with BERT. 2020. Available online: https://zenodo.org/records/4461265 (accessed on 27 May 2025).
42. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861. [CrossRef]
43. Huguet Cabot, P.L.; Navigli, R. REBEL: Relation Extraction By End-to-end Language generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 2370–2381. [CrossRef]
44. Del Corro, L.; Gemulla, R. ClausIE: Clause-based open information extraction. In Proceedings of the 22nd International Conference on World Wide Web, New York, NY, USA, 18 May 2013; WWW '13, pp. 355–366. [CrossRef]