A Multi-Signal Graph-Based Approach for Real-World Event Detection in News Articles

Juan Ignacio Llaberia

juanillaberia2002@gmail.com

Abstract

Identifying relationships between news articles in order to cluster them into real-world events is a fundamental task for analyzing the news media ecosystem. Many existing approaches rely primarily on semantic similarity, which can lead to incorrect groupings when articles share similar topics but refer to different events. In this work, we propose a multi-signal graph-based pipeline that integrates several sources of information to better model relationships between news articles. Using a gold-standard dataset of worldwide news events, the proposed method extracts multiple similarity signals, including semantic representations and entity-based information, which are combined to compare articles and identify event-level relationships. Experimental evaluation demonstrates that the proposed pipeline significantly improves clustering performance compared to a semantic similarity baseline and traditional approaches to this task. The method achieves 94.7% homogeneity and 85.8% completeness while maintaining 89.4% article coverage in the final clusters. These results indicate that combining multiple signals enables more accurate identification of relationships between articles, leading to more reliable clustering of news into meaningful real-world events.

Introduction

Grouping news articles into meaningful clusters is increasingly important due to the large volume of news content produced every day. Organizing articles around real-world events enables readers and researchers to better understand how events are reported across different media outlets. Such organization also facilitates the analysis of perspectives, biases, and framing in news coverage, which is essential for studying the dynamics of the modern media ecosystem.

However, clustering news articles by event is a challenging task. Articles discussing different events may share similar topics, entities, or vocabulary, while reports describing the same event may differ substantially in wording, emphasis, or narrative framing. As a result, approaches that rely solely on traditional clustering techniques based on semantic similarity or density may incorrectly group together articles that discuss the same topic but refer to different events, or fail to connect articles that describe the same event using different language.

In this work, we propose a multi-signal graph-based pipeline that integrates multiple sources of information to model relationships between news articles. The proposed method combines semantic similarity derived from document embeddings, named entity overlap between articles, and pairwise article scoring using a fine-tuned cross-encoder model. These signals are integrated into a graph representation where nodes correspond to articles and weighted edges represent multi-signal similarity between them. Event clusters are then identified by applying graph-based clustering techniques to this representation.

We evaluate the proposed approach using several clustering quality metrics and compare its performance against a semantic clustering baseline. Specifically, we measure homogeneity, which indicates whether clusters contain articles from a single event; completeness, which evaluates whether all articles describing the same event are grouped together; and V-measure, which combines both metrics. In addition, we measure intra-cluster similarity, computed using cosine similarity between article embeddings, and article coverage, which quantifies the proportion of input articles that are successfully included in the final clusters.

Experimental results show that the proposed method achieves 94.7% homogeneity and 85.8% completeness, improving clustering coverage compared to the baseline approach while maintaining high semantic coherence within clusters. The final clustering retains 89.4% of the articles within the resulting event clusters, demonstrating the effectiveness of combining multiple similarity signals in a graph-based framework.

The main contributions of this work are:

- A multi-signal graph-based pipeline for clustering news articles into real-world events.
- A hybrid similarity framework that combines semantic embeddings, named entity overlap, and cross-encoder pairwise scoring.
- An empirical evaluation demonstrating improvements in clustering completeness and article coverage while maintaining high cluster homogeneity.

Related Work

Automatic detection and clustering of news events has been widely studied as a way to organize large volumes of news articles into coherent real-world stories. Early research in this area emerged from the Topic Detection and Tracking (TDT) program, which aimed to automatically identify emerging events in continuous news streams. These early systems relied primarily on statistical text representations and clustering techniques to group documents describing the same event [1]. The TDT framework established the foundation for much of the later research in news event detection and story tracking.

More recent work has explored methods for organizing large collections of news articles into structured stories. For example, Real-time News Story Identification proposes techniques for identifying chains of semantically related news articles that together describe the development of a story over time [2]. Similarly, Growing Story Forest Online from Massive Breaking News introduces a scalable system capable of organizing large-scale news streams into event clusters and hierarchical story structures, enabling the representation of evolving news narratives [3]. These approaches demonstrate the importance of clustering techniques for grouping articles describing the same real-world event.

Advances in neural language models have further improved document representation methods used in event detection systems. Transformer-based models can generate dense semantic embeddings that capture contextual meaning in text, enabling more accurate similarity comparisons between news articles. These embeddings have been widely used in document clustering and information retrieval tasks involving large textual corpora [4]. However, clustering approaches based solely on semantic similarity may struggle when different events share similar topics or entities, potentially merging distinct events into a single cluster.

Graph-based approaches have also been explored to improve event detection and clustering performance. In these methods, documents are represented as nodes in a graph, while similarity relationships form weighted edges between them. Community detection algorithms can then be applied to identify clusters corresponding to latent events within the dataset [5]. Graph-based frameworks allow the integration of multiple signals beyond simple semantic similarity, making them well-suited for modeling complex relationships between news articles.

Building on these research directions, our work proposes a multi-signal graph-based pipeline for clustering news articles into real-world events. By combining semantic similarity, entity-based similarity, and pairwise scoring through a cross-encoder model, the proposed approach aims to improve clustering completeness and coverage while maintaining high semantic coherence within clusters.

1 Hypothesis

The central hypothesis of this research is that combining multiple layers of processing and analysis—rather than relying solely on semantic or topical similarity—can improve the accuracy of grouping news articles into meaningful real-world events.

To evaluate this hypothesis, it is first necessary to define what constitutes a real-world event in this context. An event is defined as a set of two or more articles that refer to the same underlying incident. These articles may present different perspectives, opinions, or updates that continue previous reporting on the same occurrence.

More formally, the goal is not to cluster articles simply because they mention the same high-level entities or exhibit strong semantic similarity. Instead, the objective is to group articles according to the specific real-world event they describe.

To illustrate this distinction, consider the following two examples referring to the Argentine president:

- "The Argentine president approves a set of laws preventing X, Y, and Z."
- "The Argentine president vetoes a tax reform proposal."

Although both articles mention the same entity (Argentina's president) and relate to legislative actions, they describe different events and therefore should belong to separate clusters.

2 Data

Our work uses a dataset containing 2,500 events and 163,753 news articles, obtained by subsampling the WCEP dataset [6]. This subset is used as our gold-standard dataset, against which we evaluate the performance of our event clustering pipeline.

We do not use the full dataset, which contains approximately 10,000 events. Instead, we limit the dataset size to reduce the computational requirements of the experiments while still maintaining a sufficiently large evaluation set, allowing us to iterate more efficiently during experimentation.

The dataset was then simplified to retain only the fields `id`, `title`, and `content` for each article. These fields contain the textual information used to construct the representations that drive the clustering process.

2.1 Analysis

During the Exploratory Data Analysis (EDA) of the events dataset, we identified several data quality issues and distributional characteristics that informed our preprocessing decisions.

Some article titles contained zero words, while a very small number were extreme outliers (with one exceeding 500 words). These cases were considered anomalous and were removed.

Similarly, some article contents contained zero words, and a small number were extreme outliers (with one exceeding 40,000 words). These instances were also removed as part of the cleaning process.

After applying these filtering steps, the dataset was reduced by 2,133 articles, improving overall consistency while preserving the vast majority of the data.

Additionally, we observed a strong long-tail distribution in source frequency. Approximately 80% of the articles originate from the top ∼400–500 news outlets, while the remaining ∼6,000 sources account for only 20% of the total volume. Rather than removing low-frequency sources—which would bias the dataset toward mainstream media—we chose to retain them in order to preserve source diversity and maintain a broad representation of media coverage.

2.2 Preprocessing

The first preprocessing step involved generating text embeddings for each article using both the title and the content.

To generate these embeddings, we used the `intfloat/multilingual-e5-large` model [7], which produces 1024-dimensional dense embeddings and supports multilingual text while maintaining strong performance across semantic similarity tasks.

One limitation of this model is its maximum context length of 512 tokens. However, this constraint fits most articles in our dataset, as the 75th percentile of article length lies around 500–600 words. Although some articles may slightly exceed the context limit,

news articles typically present their most important information in the opening sentences in order to capture the reader's attention.

For articles exceeding the context length, we applied head truncation, keeping the first portion of the text within the allowed token limit. Finally, embedding normalization was applied to ensure consistent vector magnitudes across all representations.

3 Baseline

As a baseline for this research and for future experiments in the pipeline, we implemented a semantic clustering approach using HDBSCAN (Hierarchical Density-Based Spatial Clustering) [8]. This algorithm clusters data points based on proximity in the embedding space, effectively capturing semantic similarity between news articles.

We selected HDBSCAN because it does not require specifying the number of clusters in advance. This property aligns well with our task, as the true number of real-world events is unknown both during experimentation and in production environments, where no gold dataset is available to indicate the appropriate number of clusters.

Additionally, HDBSCAN is a computationally efficient algorithm and includes the ability to identify noise points, i.e., articles that do not strongly belong to any cluster. While this capability can help avoid incorrect cluster assignments, it may also lead to a significant portion of the data being left unclustered, as discussed later in this section.

3.1 Experiment Structure

For the baseline experiment, we used the previously generated 1024-dimensional embeddings for each article. To improve clustering performance and reduce computational complexity, we applied Principal Component Analysis (PCA) to reduce the dimensionality to 100 components. This dimensionality reduction step helps remove redundant or low-variance components that may introduce noise into the clustering process.

After preparing the data, we applied HDBSCAN with the following hyperparameters:

Parameter	**Value**
Metric	Euclidean
`min_cluster_size`	10
`min_samples`	3
`cluster_selection_epsilon`	0.1
`alpha`	1
`cluster_selection_method`	Leaf

Table 1: Baseline (HDBSCAN) hyperparameters.

3.2 Evaluation & Results

To evaluate the clustering results, we considered five metrics: homogeneity, completeness, V-measure, intra-cluster similarity, and article coverage within clusters.

Homogeneity measures whether clusters contain articles from a single ground-truth event, while completeness measures whether all articles belonging to the same event are assigned to the same cluster. V-measure represents the harmonic mean of homogeneity and completeness, providing a balanced evaluation of both properties. Intra-cluster similarity evaluates the average semantic similarity between articles within the same cluster, and article coverage measures the proportion of articles that are assigned to a cluster rather than being labeled as noise.

The results show strong performance in terms of homogeneity, which reached an almost perfect score of ∼98%. This indicates that the clusters produced by the baseline rarely mix articles from different events.

However, the method presents a significant limitation in terms of article coverage. The generated clusters contain only approximately 61% of the original articles, meaning that nearly 40% of the dataset is labeled as noise and excluded from clusters.

This behavior suggests that HDBSCAN tends to favor conservative clustering decisions, preferring to leave articles unassigned rather than risking incorrect groupings. While this strategy improves homogeneity, it also leads to inflated evaluation scores, since a large portion of the dataset is not considered in the clustering structure.

In addition to the high homogeneity score, the baseline achieved 75% completeness, 85% V-measure, and 88% intra-cluster similarity.

Another potential limitation of this approach arises in scenarios involving closely related events, where articles from different events share strong semantic similarity. In such cases, relying primarily on embedding proximity may make it difficult for the algorithm to clearly separate distinct events.

These limitations motivated the development of a multi-signal, graph-based approach, which aims to improve completeness, V-measure, intra-cluster similarity, and particularly article coverage within clusters, while providing a more reliable representation of real-world events.

Metric	**Score**
Homogeneity	0.98
Completeness	0.75
V-measure	0.85
Intra-cluster similarity	0.88
Article coverage	0.61

Table 2: Results of the baseline model.

4 Methodology

To address the limitations observed in the baseline approach, we propose a multi-signal graph-based pipeline that combines several complementary signals to determine whether two news articles describe the same real-world event. Rather than relying exclusively on semantic similarity between article embeddings, the proposed method incorporates entity-level information and pairwise event likelihood estimation.

The pipeline is composed of four main stages: entity extraction and disambiguation, article pair generation, article pair comparison, and graph-based clustering. Each stage progressively refines the relationships between articles in order to construct clusters that more accurately represent real-world events.

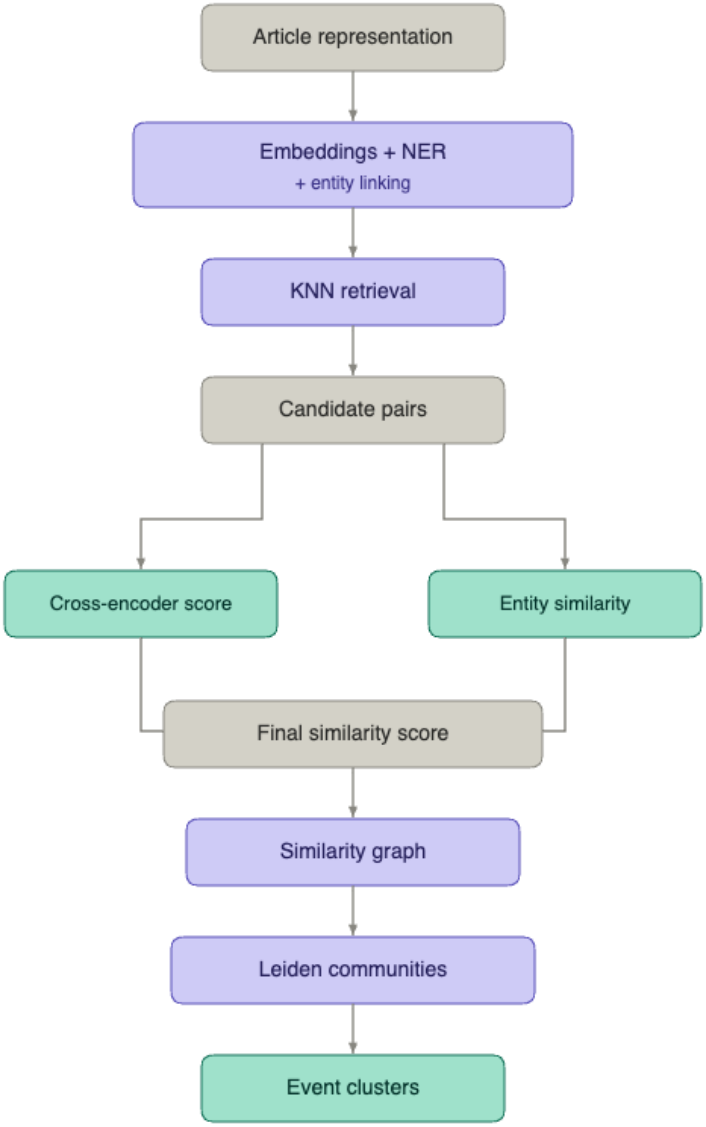

Figure 1: Diagram showing all pipeline stages.

4.1 Entity Extraction & Disambiguation

In this stage, we perform Named Entity Recognition (NER) in order to extract the most relevant entities mentioned in each article. We focus on the following entity types: persons, organizations, locations, and miscellaneous entities, as these categories typically capture key actors and places associated with real-world events.

Before applying the NER model, the raw article text is preprocessed and segmented into individual sentences. This step improves extraction quality, as NER models generally perform more accurately when operating at the sentence level rather than on long documents.

We then apply the `Jean-Baptiste/roberta-large-ner-english` [9] model to extract named entities from each sentence. However, raw entity strings may refer to the same real-world entity in different forms (e.g., abbreviations, alternative spellings, or partial names). To address this issue, we perform entity normalization and disambiguation.

For this purpose, we leverage the `facebook/mgenre-wiki` [10] model, which maps extracted entities to their corresponding Wikipedia titles. This allows us to assign unique identifiers to entities and ensures that different textual mentions referring to the same entity are normalized to a consistent representation. As a result, entity comparisons across articles can be performed in a more robust manner than relying solely on raw string matching.

Future work could extend this process by integrating Wikidata identifiers, enabling even more precise entity linking and richer knowledge-based representations.

4.2 Article Pair Generation

The objective of this stage is to identify candidate article pairs that are likely to describe related events. Instead of comparing every article with every other article, which would be computationally expensive, we restrict comparisons to the k most semantically similar articles for each article.

To retrieve these candidates, we evaluated two similarity search approaches both using $K = 100$:

- **K-Nearest Neighbors (KNN):** Performs exact similarity search in the embedding space, prioritizing retrieval accuracy at the cost of higher computational complexity.
- **Approximate Nearest Neighbors (ANN):** Provides faster and more scalable similarity search by allowing small approximation errors in the retrieved neighbors.

After experimentation, we selected KNN for the final pipeline, as it guarantees exact nearest-neighbor retrieval while maintaining acceptable computational performance for the dataset size used in this study. Although ANN methods scale better for significantly larger datasets, our experiments showed that they introduced an accuracy decrease of approximately 5% in neighbor retrieval.

Using this approach, we retrieve the top-k most similar articles for each article in the dataset. This produces candidate pairs of the form:

$$(A, A_1),\ (A, A_2),\ \ldots,\ (A, A_k)$$

These pairs are then forwarded to the next stage for deeper comparison.

4.3 Article Pair Comparison

In this stage, we compute a final similarity score for each candidate pair by combining two complementary signals: semantic event likelihood and entity overlap similarity.

The first signal is produced by a cross-encoder scoring model, which evaluates whether two articles are likely to describe the same event. The model is a fine-tuned BERT-base architecture [11] based on ModernBERT [12] that receives a pair of articles as input and outputs a similarity score representing the likelihood that the articles refer to the same real-world event.

The cross-encoder processes the two articles jointly by concatenating their textual representations into a single input sequence separated by a special token. Unlike bi-encoder approaches, which compute embeddings for each article independently and compare them afterward, the cross-encoder allows the transformer to attend across both texts simultaneously. This enables the model to capture fine-grained interactions between the two articles, such as shared entities, temporal references, and contextual cues that indicate whether they describe the same real-world occurrence.

The model was trained using the HLGD dataset [13], which contains approximately 20,000 pairs of news headlines labeled according to whether they refer to the same event. Unlike a binary classifier, the model is used here as a continuous scoring function, providing a graded similarity value rather than a strict yes/no decision.

The second signal is derived from entity similarity analysis. For each article, we construct a set of normalized entities extracted during the NER stage. We then compute the Jaccard similarity between the two sets of entities, measuring the degree of entity overlap between the articles.

Finally, we combine these two signals into a single similarity score using a weighted sum:

- Cross-encoder score: 0.85
- Entity similarity score: 0.15

The higher weight assigned to the cross-encoder reflects the fact that entity overlap alone does not necessarily indicate that two articles describe the same event, while semantic context captured by the cross-encoder provides stronger evidence of event-level similarity.

To evaluate the effect of training data balance on downstream clustering quality, we trained a second cross-encoder (v2) on a balanced dataset constructed from a held-out portion of the WCEP dataset, ensuring no event overlap with the evaluation set to prevent data leakage. While both models achieved nearly identical V-Measure scores (0.9010 vs. 0.9034), they represent fundamentally different precision/recall tradeoffs. The balanced model improved completeness from 0.8589 to 0.9306 and article coverage from 89.4% to ∼100%, at the cost of homogeneity dropping from 0.9474 to 0.8776. This demonstrates that training data balance directly controls the precision/recall tradeoff in downstream clustering, independently of the graph construction and community detection stages. Given the pipeline's design philosophy of prioritizing cluster purity, the imbalanced model (v1) was selected for the primary evaluation.

4.4 Graph Generation

After computing similarity scores for all candidate pairs, we construct a similarity graph representing relationships between articles.

- Each node represents a news article.

- An edge between two nodes represents the final similarity score between the corresponding article pair.
- The graph is sparse, since each article is connected only to its top-k nearest neighbors.
- Edges with similarity scores below a predefined threshold (0.5) are removed to reduce noise and prevent weak relationships from influencing the clustering process.

Once the graph is constructed, we apply the Leiden community detection algorithm with a resolution of 200 to identify clusters of strongly connected articles. Leiden is an extension of the Louvain algorithm that improves community detection by ensuring well-connected communities and faster convergence. It has been shown to produce higher-quality and more stable partitions in large networks.

The resulting communities correspond to detected events, grouping together articles that are strongly connected according to the combined similarity signals.

5 Evaluation & Results

To evaluate the pipeline's outcome, we use the same metrics as in the baseline in order to enable direct comparison between both approaches. The gold dataset labels are used as ground truth for all evaluations.

5.1 Metrics

After completing all stages of the proposed pipeline, we obtained the following results:

Metric	**Score**
Homogeneity	0.95
Completeness	0.86
V-measure	0.90
Intra-cluster similarity	0.96
Article coverage	0.89

Table 3: Performance of proposed multi-signal pipeline.

In addition, the final clustering produced 89.4% article coverage, meaning that the vast majority of articles were successfully assigned to clusters.

Compared to the baseline approach, the proposed multi-signal graph-based method maintains very high homogeneity while substantially improving completeness and overall cluster coverage. This indicates that the pipeline is able to group a larger proportion of articles without significantly sacrificing cluster purity.

In particular, the improvement in coverage—from approximately 61% in the baseline to 89.4% in the proposed method—suggests that the graph-based approach is more effective at connecting related articles that may not be sufficiently close in embedding space alone. By combining semantic similarity with entity-based signals and pairwise scoring, the system can recover relationships that purely embedding-based clustering may fail to capture.

The high intra-cluster cosine similarity (0.9573) further indicates that the resulting clusters remain semantically coherent despite the increased coverage.

5.2 Comparison with Baseline

Method	Homogeneity	Completeness	V-Measure	Intra Similarity	Coverage
Baseline	0.98	0.75	0.85	0.88	61%
Multi-Signal	0.95	0.86	0.90	0.96	89%

Table 4: Comparison between baseline and proposed pipeline.

5.3 Discussion

It is important to note that clustering evaluation metrics often present inherent trade-offs. Methods that maximize homogeneity may do so by producing smaller or more conservative clusters, which can negatively impact completeness and overall coverage. This behavior was observed in the baseline approach, where the clustering algorithm preferred leaving many articles unassigned in order to maintain highly pure clusters.

In contrast, the proposed multi-signal graph-based pipeline balances these competing objectives by incorporating additional relational signals between articles. As a result, the system is able to assign a significantly larger proportion of articles to clusters while maintaining high semantic coherence. This balance between purity and coverage suggests that the proposed method provides a more realistic representation of real-world event groupings.

6 Conclusion

We began this research by establishing a strong baseline using HDBSCAN combined with semantic embeddings to cluster news articles into events. The baseline achieved very high homogeneity ($\sim$98%), indicating that most articles within a cluster belonged to the same original event. However, completeness was significantly lower ($\sim$75%), suggesting that a substantial portion of articles belonging to the same event were either assigned to different clusters or marked as noise.

The baseline also achieved an average intra-cluster similarity of $\sim$88%, confirming that articles grouped together shared strong semantic similarity. Nevertheless, an important limitation of this approach was the high proportion of articles labeled as noise. Specifically, the HDBSCAN model left $\sim$38.7% of the articles unassigned, resulting in a coverage of only 61.3% of the dataset.

These results highlight that while a purely semantic clustering approach can produce highly pure clusters, it may struggle to capture the full structure of real-world events. In particular, events that share similar topics or entities may either be merged into a single cluster or remain partially unclustered.

To address these limitations, we proposed a multi-signal pipeline designed to incorporate additional signals beyond semantic similarity. The pipeline combines several stages:

semantic retrieval using K-nearest neighbors, pairwise evaluation with a fine-tuned cross-encoder model, entity similarity analysis, and graph-based clustering using Leiden community detection.

Using the configurations described in this work, the proposed method achieved improved performance across most evaluation metrics:

- Homogeneity: 0.9474
- Completeness: 0.8589
- V-Measure: 0.9010
- Average Intra-cluster Cosine Similarity: 0.9573

Although homogeneity decreased slightly ($\sim$4%), the model achieved notable improvements in completeness, V-measure, and intra-cluster similarity, demonstrating that the clusters capture a larger proportion of event-related articles while maintaining strong semantic coherence.

Perhaps the most significant improvement is observed in article coverage. With the proposed pipeline, only $\sim$10% of articles were labeled as noise, representing a 28.7% reduction compared to the baseline approach. This indicates that the multi-signal method is substantially more effective at assigning articles to meaningful clusters.

Overall, the results suggest that incorporating multiple complementary signals, including semantic similarity, entity overlap, and pairwise event scoring, enables more accurate and robust grouping of news articles into real-world events. The proposed graph-based framework therefore provides a promising direction for improving event detection systems in large-scale news environments.

7 Future Work

Several directions can be explored to further improve the performance and robustness of the proposed event clustering pipeline.

First, future work could incorporate a weighted Jaccard similarity when computing the entity similarity score. Instead of treating all named entities equally, different weights could be assigned to entity types. For example, persons and organizations could receive higher weights, locations could receive a medium weight, and miscellaneous entities a lower weight. This modification would emphasize entities that are typically more informative for identifying real-world events, potentially leading to more precise similarity estimates between articles.

Second, the article retrieval stage could be improved by implementing a hybrid retrieval approach. Currently, candidate articles are retrieved using a K-nearest neighbors (KNN) search over semantic embeddings. A hybrid strategy could combine this semantic retrieval with a BM25-based keyword search, executed in parallel. While KNN retrieval captures semantic similarity, BM25 is particularly effective for matching explicit terms such as names, dates, and quotations. The results from both retrieval methods could then be combined using a fusion strategy to generate the final list of K candidate articles.

References

References

[1] Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic Detection and Tracking Pilot Study: Final Report. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

[2] Škvorc, T., Ivačič, N., Hribar, S., & Robnik-Šikonja, M. (2025). Real-time News Story Identification. *arXiv preprint arXiv:2508.08272*.

[3] Liu, X., Nourbakhsh, A., Li, Q., Fang, R., & Shah, S. (2017). Growing Story Forest Online from Massive Breaking News. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[4] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[5] Newman, M. E. J., & Girvan, M. (2004). Finding and Evaluating Community Structure in Networks. *Physical Review E*, 69(2).

[6] Fabbri, A., Li, I., She, T., Li, S., & Radev, D. (2019). Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Dataset: `https://github.com/complementizer/wcep-mds-dataset`

[7] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

[8] Malzer, C., & Baum, M. (2019). A Hybrid Approach to Hierarchical Density-Based Cluster Selection. *arXiv preprint arXiv:1911.02282*.

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*. Model implementation: `https://huggingface.co/Jean-Baptiste/roberta-large-ner-english`

[10] De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., & Petroni, F. (2022). Multilingual Autoregressive Entity Linking. *Transactions of the Association for Computational Linguistics*.

[11] Llaberia, J. (2026). Articles Pairs Event Detection Model. HuggingFace Model Repository. `https://huggingface.co/Juanillaberia/articles-pairs-event-detection`

[12] Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., & Poli, I. (2024). Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint arXiv:2412.13663*.

[13] Laban, P., et al. (2021). HLGD: A Large-Scale Hierarchical Dataset for

Multi-Document News Analysis. Dataset: https://huggingface.co/datasets/philippelaban/hlgd

Appendix: Supplemental Materials

Appendix A: Supplemental Notebooks

All raw data, handwritten notes (digitized), and preliminary calculations associated with this study are archived and publicly available at the following repository:

Repository Link: https://github.com/JuaniLlaberia/news_articles_grouping_research

Archive Date: March 16, 2026