MultiSumm: An AI-Powered System for Automated Multimodal News Summarization

Dhannya. S. M[1,†], Lakshmi Priya. S[2,*,†], Shalini. M[3,†], Avanthika Vijayakumar[4,†] and Ranjan S[5,†]

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

This working note reports on MultiSumm, an AI-powered summarization automation system designed for the MediaEval 2025 NewsSumm task. The system collects text and images from a list of URLs corresponding to multiple topically related city-based websites and automatically generates concise summaries accompanied by representative visuals. Using the BART (facebook/bart-large-cnn) Transformer model, our approach integrates web scraping, text summarization, and document generation into a unified workflow. Results demonstrate that employing an AI-based summarization model effectively condenses large volumes of heterogeneous web content into coherent summaries, thereby improving efficiency and consistency in multi-source information analysis. We discuss our findings, encountered challenges, and potential implications for future multimodal summarization research.

1. Introduction

In recent years, the growth of online information sources has created an increasing need for automated summarization systems capable of efficiently condensing large amounts of data into digestible formats. Manual summarization of web content from multiple sources is not only time-consuming but also prone to inconsistency and human bias. Furthermore, existing tools often specialize in either textual or visual content, lacking the capability to integrate both modalities in a structured, unified output. To address this gap, we present MultiSumm, an AI-powered summarization automation pipeline designed to process, summarize, and visually enhance content from multiple related websites. The system takes as input a collection of URLs provided for specific cities - Dublin and Brighton for the main task, and London, Barcelona, and Milan for the subtask - and automatically retrieves relevant web data. The collected text is summarized using the BART model, a transformer-based sequence-to-sequence architecture well-suited for abstractive summarization. Alongside this, corresponding images from the same sources are extracted to visually complement the textual summaries.

2. Background

Research on multimodal summarization (MMSum) has advanced considerably in recent years, driven by the need to integrate information from text, images, and videos into cohesive summaries. One of the early contributions in this area is the MM-AVS dataset by Fu et al. [1], which provided a large-scale benchmark combining video, transcript, and visual metadata to support multimodal learning. Building on this foundation, Mukherjee et al. [2] introduced a topic-aware summarization model that leverages topic information to enhance the relevance and thematic consistency of generated summaries across different modalities. In the domain of news summarization, Krubiński and Pecina [3] proposed MLASK,

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

*Corresponding author.

†These authors contributed equally.

dhannyasm@ssn.edu.in (Dhannya. S. M); lakshmipriyas@ssn.edu.in (L. Priya. S); shalini2470052@ssn.edu.in (Shalini. M); avanthika2470011@ssn.edu.in (A. Vijayakumar); ranjan2410570@ssn.edu.in (R. S)

0000-0002-0302-7458 (Dhannya. S. M); 0000-0002-9923-4020 (L. Priya. S)

a system for summarizing video-based news articles that demonstrated how effectively combining linguistic and visual cues can improve coverage and factual alignment.

Around the same time, Kumbhar et al. [4] reviewed the state of multimodal summarization research, organizing existing approaches by modality type, fusion strategy, and evaluation method, while emphasizing challenges such as limited interpretability and dataset diversity. Jangra et al. [5] offered a more comprehensive survey, examining various model architectures, attention mechanisms, and benchmark datasets, and highlighting persistent issues like modality imbalance and temporal inconsistency. Tang et al. [6] addressed some of these issues by proposing a feature alignment and filtering approach, which selectively attends to salient multimodal features to reduce redundancy and noise.

More recently, Qiu et al. [7] introduced MMSum, a large-scale dataset designed for both multimodal summarization and thumbnail generation, enabling more fine-grained evaluation of visual-textual understanding in videos. The MultiSumm task proposed by Potyagalova and Jones [8] further contributed to the field by establishing a standardized evaluation framework for benchmarking multimodal summarization systems.

3. Approach

Our system, SmartSum, was developed for the MediaEval 2025 MultiSumm task, which focuses on generating multimodal summaries of multiple websites describing Food Sharing Initiatives (FSIs). The goal of our approach is to produce coherent, factual, and visually grounded summaries that combine textual and visual information into a structured English-language format. Unlike traditional summarization methods, which rely on manual curation or separate processing stages, MultiSumm automates the entire workflow - from content extraction to summary compilation - using a combination of web scraping, transformer-based summarization, and document generation. Our system follows a clean, modular pipeline designed to convert raw web pages into structured summaries. First, we collect article URLs and extract the raw HTML using requests, then use BeautifulSoup to isolate meaningful text while removing ads, menus, and scripts. If the source content is non-English, we apply translation to maintain uniform processing. The cleaned text is then split into manageable segments to respect BART's token limit and passed through the BART model for abstractive summarization. Each chunk is summarized separately and later merged using redundancy checks, based on sentence similarity, to maintain consistency and coherence. In parallel, our pipeline scans each webpage for suitable images, filters them by size and quality, and pairs them with the final summary to enhance readability. Finally, we evaluate summary quality using ROUGE scores (lexical overlap) and BERTScore (semantic similarity) against manually drafted reference summaries. This automated feedback loop helped us compare multiple runs and refine our summarisation strategy. This unified approach enhances efficiency, reduces redundancy, and ensures stylistic consistency across summaries. During implementation, we encountered several practical challenges, notably in the URL-fetching phase. Certain URLs were either inaccessible, removed, or blocked by website restrictions, requiring adaptive error-handling and fallback strategies. Despite these hurdles, the system successfully demonstrated that integrating automated summarization with visual context generation leads to more comprehensive and informative outputs.

Our findings indicate that AI-based summarization pipelines such as MultiSumm can play a crucial role in streamlining the process of large-scale content analysis, particularly in domains where textual and visual coherence is essential. In the subsequent sections, we discuss the system architecture, implementation details, and evaluation results, and we outline future directions for enhancing multimodal summarization performance. Given the heterogeneity of the crawled website content and the presence of both English and non-English datasets, SmartSum integrates several sequential processing components -data extraction, text summarization, multilingual translation, image retrieval, and automatic evaluation - all within a lightweight transformer-based pipeline.

Our methodology consists of five experimental configurations (runs), each exploring a different strategy for improving coherence, coverage, and multimodal quality. The following subsections describe each in detail.

Run 1 – Raw Summarization (Baseline)

For the initial baseline run, raw text content was extracted from the crawled URLs using the BeautifulSoup library. The text was then directly summarized using the BART model without additional preprocessing. This run provided a baseline for assessing both fluency and factual alignment of later improvements.

Run 2 – Structured Multi-Section Summarization

In the second run, we introduced a structured summarization design that divided content into thematic clusters corresponding to key FSI categories:

- Food Security & Sustainability
- Community & Non-profits
- Creative & Social Projects

Each category was summarized independently using the same BART model, and the outputs were concatenated to form a unified city-level summary. This approach improved topical clarity and reduced redundancy compared to the baseline.

Run 3 – Multilingual and Cross-Cultural Summarization

For the subtask datasets, which contained content in Italian, Catalan, and Spanish, SmartSum integrated a multilingual translation module using the Google Translator API. The text was first summarized in its original language, then translated into English. This preserved context while ensuring linguistic consistency across multilingual datasets.

Run 4 – Multimodal Enhancement with Image Integration

To satisfy the multimodal requirement of the MultiSumm task, this run focused on visual-text alignment. Images were automatically retrieved from the same web pages as the summarized text using BeautifulSoup-based scraping. Filtering heuristics were applied to retain images above a minimum resolution threshold (100×100 pixels). The selected images were embedded alongside textual summaries within structured Word documents, resulting in comprehensive multimodal summaries per city.

Run 5 – Evaluation-Augmented Summarization (ROUGE + BERTScore)

In the final configuration, SmartSum integrated automatic evaluation directly into the summarization pipeline. The system-generated summaries were compared against the provided human-written references using ROUGE-1, ROUGE-2, and ROUGE-L for lexical overlap, and BERTScore for semantic similarity. This dual evaluation strategy enabled us to quantitatively assess both factual coverage and semantic alignment, ensuring balanced optimization between linguistic fluency and informational completeness.

4. Results and Analysis

Table 1 summarizes the quantitative results obtained from the five SmartSum configurations across both the main and subtask datasets. The evaluation employed ROUGE for lexical coverage and BERTScore for semantic alignment. The results demonstrate that integrating structure-aware summarization and multilingual handling led to notable improvements in both content completeness and coherence.

Across all datasets, the structured summarization model (Run 2) achieved the most balanced performance, producing summaries that aligned closely with human-written references. The multimodal integration run (Run 4) provided visually enriched outputs with slightly lower lexical precision but

Table 1
Average Score for different runs

Run	Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Run 1	Raw Summarization (Baseline)	0.38	0.15	0.35	0.79
Run 2	Structured Multi-Section Summarization	0.47	0.22	0.43	0.83
Run 3	Multilingual + Cross-Cultural Summarization	0.44	0.18	0.41	0.85
Run 4	Multimodal Image-Integrated Summarization	0.42	0.16	0.40	0.82
Run 5	Evaluation-Augmented Summarization	0.46	0.20	0.44	0.86

improved contextual representation. Meanwhile, the evaluation-augmented pipeline (Run 5) yielded the highest overall semantic similarity as measured by BERTScore.

The results reveal several insights into the performance of SmartSum. The structured multi-section model (Run 2) achieved notable gains in ROUGE-1 and ROUGE-L, reflecting improved content coherence and thematic segmentation across domains such as Food Security, Community Engagement, and Creative Initiatives. For non-English datasets like Milan and Barcelona, the multilingual translation pipeline effectively preserved contextual meaning, with BERTScore improvements confirming strong semantic consistency despite minor translation noise. The multimodal integration (Run 4) successfully embedded representative visuals into structured documents, enhancing narrative richness and interpretability even with a slight decrease in lexical overlap. Additionally, incorporating automated evaluation metrics such as ROUGE and BERTScore in Run 5 created an adaptive feedback loop that optimized factual accuracy and readability. Across all datasets—Brighton, Dublin, UK, Milan, and Barcelona—SmartSum maintained stable, high-quality output, producing linguistically fluent and information-dense summaries. Overall, the results demonstrate that SmartSum effectively balances content coverage, semantic depth, and multimodal diversity, showcasing the power of transformer-based summarization enhanced by lightweight multimodal processing for diverse linguistic and cultural contexts.

Leading up to the completion of our project, we explored multiple approaches for extracting, summarizing, and correlating data from diverse web sources. Our experiments included different summarization strategies, such as hierarchical and chunk-based summarization using transformer models, and embedding-based correlation techniques for evaluating semantic alignment between text and related media. While several configurations were tested, none consistently outperformed the baseline summarization pipeline using the standard pre-trained model setup. For future iterations of this work, expanding the dataset with richer and more structured inputs would be crucial for improving accuracy and generalization. Looking ahead, we identify two possible directions for enhancing the system. The first is the integration of a domain adaptive summarization model fine-tuned specifically on urban and cultural datasets, enabling more context-aware outputs. The second is reversing the correlation pipeline, starting from media or image descriptions and aligning them to the most semantically relevant textual summaries.

Finally, we recognize certain limitations in the current evaluation and workflow setup. Treating each text-summary pair as a one-to-one correspondence may oversimplify the multi-perspective nature of real-world data. Future versions should consider graded relevance metrics and support for multiple valid outputs per input. Additionally, handling AI-generated or abstract content remains a complex challenge, often affecting correlation precision and interpretability. Overall, this project demonstrates the foundation for a scalable and adaptive summarization framework capable of handling heterogeneous data sources. Future research focusing on data expansion, adaptive evaluation, and multimodal integration could significantly enhance the reliability and impact of this approach.

5. Declaration on Generative AI

The authors have used Generative AI tools to rephrase some of the sentences in the content. The scientific insights, conclusions, and recommendations have been obtained by human authors.

References

[1] X. Fu, J. Wang, Z. Yang, MM-AVS: A full-scale dataset for multi-modal summarization, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5922–5926.

[2] S. Mukherjee, A. Jangra, S. Saha, A. Jatowt, Topic-aware multimodal summarization, in: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Association for Computational Linguistics, Online only, 2022.

[3] M. Krubiński, P. Pecina, MLASK: Multimodal summarization of video-based news articles, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 910–924.

[4] A. Kumbhar, H. Kulkarni, A. Mali, S. Sonawane, P. Mulay, The current landscape of multimodal summarization, in: Proceedings of the 20th International Conference on Natural Language Processing (ICON), NLP Association of India (NLPAI), 2023, pp. 797–806.

[5] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, M. Hasanuzzaman, A survey on multi-modal summarization, ACM Comput. Surv. 55 (2023).

[6] B. Tang, B. Lin, Z. Chang, S. Li, Multimodal summarization with modality features alignment and features filtering, Neurocomput. 603 (2024).

[7] J. Qiu, J. Zhu, W. Han, A. Kumar, K. Mittal, C. Jin, Z. Yang, L. Li, J. Wang, D. Zhao, B. Li, L. Wang, MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos , in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2024, pp. 21909–21921.

[8] A. Potyagalova, G. J. F. Jones, Multisumm - multimodal summarisation task at mediaeval 2025, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.