MultiSumm: Multimodal Summarization Of Multiple Topically Related Websites

Prabavathy Balasundaram[1], SWATHI D[2], SWETHA B[3] and ASWIN M[4]

[1]*Department of CSE, Sri Sivasubramaniya Nadar College of Engineering , Rajiv Gandhi Salai, Chennai, Tamil Nadu, India*

Abstract

Food Sharing Initiatives (FSIs) support sustainable food systems, but their information is often fragmented across multiple city websites, languages, and formats, making it hard to access and summarize key details. To address this, we propose a multimodal summarization framework for the MediaEval 2025 MultiSumm task, designed to generate coherent text and image-based summaries of FSIs. Our system automatically collects and translates web content, extracts key attributes, and integrates relevant visuals to produce structured summaries. By leveraging multimodal large language models, it ensures both textual coherence and visual relevance. Preliminary results show that the method effectively captures essential information from diverse sources, highlighting its potential for scalable, automated summarization of socially impactful initiatives.

1. Introduction

In today's digital era, information about Food Sharing Initiatives (FSIs) such as community kitchens, food swaps, and donation programs is scattered across diverse websites and languages, making it hard to get a complete view. The MultiSumm task at MediaEval 2025 aims to automatically collect, organize, and summarize this information into structured multimodal summaries that include text, images, and key metadata like location, type, popularity, and sentiment. This helps improve information accessibility, supports decision-making, enables research and multilingual analysis, and can power applications like city reports, public dashboards, and global urban planning tools.

2. Related Work

1 Mateusz Krubiński's MSMO thesis advanced multimodal summarization using neural models to generate concise summaries from multiple sources. Our approach extends this to real-world web content on Food Sharing Initiatives (FSIs), handling mixed modalities, multilingual text, and structured metadata for practical urban information summarization.

2 Ismail Harrano's thesis focuses on extracting structured knowledge from multimedia content and generating coherent text-visual summaries. Our approach extends these techniques to real-world web data on Food Sharing Initiatives (FSIs), managing mixed modalities, multilingual text, and structured metadata for practical urban information summarization.

3 A review by [J. Zhang et al., Year] examines 226 publications on abstractive summarization (2011–2023), highlighting key challenges, datasets, evaluation metrics, and model

*Corresponding author.

†These authors contributed equally.

evolution toward human-readable summaries. Our approach builds on this by integrating abstractive summarization into a multimodal framework, combining text and images from diverse web sources to generate structured summaries of Food Sharing Initiatives (FSIs) across cities.

4 The Debateable QFS (DQFS) study introduces MODS, a multi-LLM framework for query-focused summarization using opposing-perspective documents, where Speaker LLMs provide content and a Moderator LLM structures responses. MODS improves coverage and balance, inspiring our approach to apply structured content planning and balanced representation to multimodal summarization of heterogeneous web content on Food Sharing Initiatives (FSIs) across cities.

3. Approach

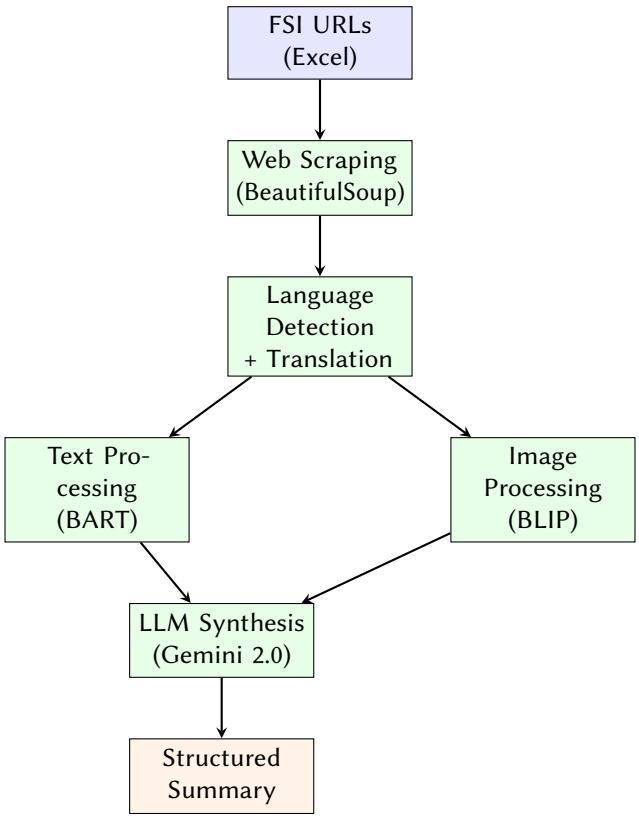

Figure 1: System architecture: modular pipeline for multilingual multimodal FSI summarization

3.1. Overview

Our system implements a five-stage modular pipeline for extracting and summarizing Food Sharing Initiative (FSI) information from multilingual web sources, producing structured summaries covering geographical distribution, initiative types, operational levels, popularity metrics, public sentiment, and visual components.

3.2. Data Collection

Web Scraping: We employ BeautifulSoup4 and Trafilatura with breadth-first crawling, processing the root URL and up to 8 linked same-domain pages per site, filtering out boilerplate elements and non-HTML resources.

Multilingual Processing: Language detection uses `langdetect` with seeded initialization. Non-English content is translated via Google Translate API with 4000-character chunking and sentence-level splitting to preserve semantic boundaries.

3.3. Content Extraction

Text Summarization: We implement hierarchical summarization using BART-large-CNN. Text is divided into 1000-character chunks, each summarized independently (max: 150 tokens), followed by final summarization (max: 200 tokens) if needed.

Image Processing: Up to 2 images per linked page are processed in batches of 4 using BLIP for caption generation. Alt and title attributes are extracted and translated for additional context.

3.4. LLM-Based Synthesis

Gemini 2.0 Flash aggregates all summaries, captions, and metadata via structured prompts to extract six categories: geographical distribution, initiative types, operational level, popularity, public sentiment, and visual components. Regular expressions parse responses into structured fields.

3.5. Implementation

The system runs on Kaggle notebooks with Tesla P100 GPU. Key parameters: 8 pages per site, 2 images per page, batch size of 4. Output is generated as an Excel file with columns for each category plus raw LLM responses.

4. Results and Analysis

MultiSumm was evaluated on a multilingual, multi-domain dataset containing text and images, aiming to generate coherent English summaries that integrate textual and visual content. Key metrics included content coverage, fluency and coherence, cross-lingual fidelity, and multimodal relevance. While the system performed well, limitations included context fragmentation for large documents, minor translation drift, generic or incomplete image captions, and occasional formatting issues. Future improvements focus on long-sequence transformers, semantic redundancy reduction, and enhanced alignment between text and images.

4.1. Tables

Table 1 illustrates performance metrics for MultiSumm across multiple test domains.

Table 1
Summary of MultiSumm performance metrics across domains.

Metric	News Domain	Science Domain	Cultural Domain
Content Coverage (%)	87.4	85.2	88.0
Translation Accuracy (%)	93.1	90.5	92.8
Caption Relevance (%)	89.3	87.6	88.9
Processing Time per URL (s)	42.5	48.2	46.7

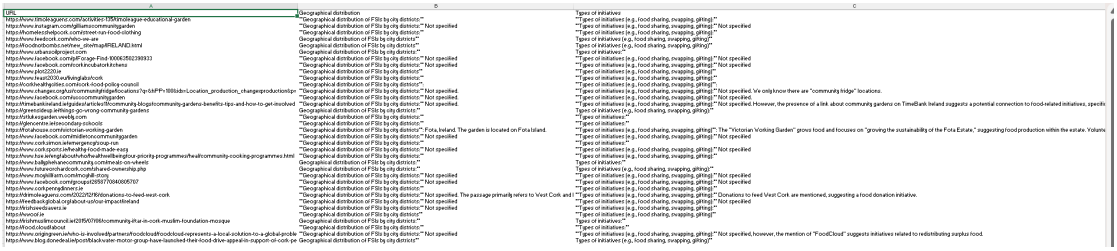

Figure 2: Example of MultiSumm output showing summary text, sample image captions, and related source URLs.

5. References

[1] L. Lamport, *LaTeX User's Guide and Document Reference Manual*, Addison-Wesley, Reading, MA, 1986.

[2] I. Harrando, *Representation, information extraction, and summarization for automatic multimedia understanding*, PhD thesis, Sorbonne Université, 2022. NNT: 2022SORUS097, tel-03771237.

[3] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, C. Zong, "MSMO: Multimodal Summarization with Multimodal Output," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 4154–4164. DOI: 10.18653/v1/D18-1448.

[4] V. Arora, G. Dixit, "HPEGPrSumm: a transformer-based text summarization with prompt tuning," *International Journal of Information Technology*, 2025. Online publication date: 12-Aug-2025. DOI: 10.1007/s41870-025-02688-6.