Hybrid Rank-MMR: Zero-shot Text Summarization via Weighted Feature Fusion based on Pre-trained Language Models

Original Scientific Paper

Minh-Tuan Luu*
National Economics University,
College of Technology, Faculty of Information Technology
207 Giai Phong Street, Bach Mai Ward, Hanoi, Vietnam
tuanlm@neu.edu.vn

Quoc-Thai Nguyen
National Economics University,
College of Technology, Faculty of Information Technology
207 Giai Phong Street, Bach Mai Ward, Hanoi, Vietnam
quocthai@neu.edu.vn

Thanh-Huong Le
Hanoi University of Science and Technology,
School of Information and Communication Technology, Faculty of Computer Science
1 Dai Co Viet Street, Bach Mai Ward, Hanoi, Vietnam
huonglt@soict.hust.edu.vn

*Corresponding author

Abstract *– Text summarization is a challenging problem in natural language processing (NLP), especially for low-resource languages like Vietnamese or specialized documents (such as scientific articles). Currently, extractive text summarization models often rely on semantic matching using pre-trained language models or select sentences based on their position. However, these methods prove ineffective when applied to different languages or domains in extractive text summarization tasks. Therefore, we propose Hybrid Rank-MMR, a re-ranking model that combines semantic similarity with two features: semantic density and sentence position. This method can handle both news (relying on sentence position) and complex datasets (requiring semantic understanding). Furthermore, we refine the MMR algorithm to automatically control diversity and remove duplicate sentences during summary generation. We test on seven datasets in both English (CNN-DailyMail, Multi-News, XScience, WCEP) and Vietnamese (VNDS, ViMs, AbMuSu). The results show that this model significantly improves performance compared to the basic models. ROUGE-1 increases by +2.56 on Multi-News and +6.68 on ViMs. Finally, we integrate an explainable artificial intelligence (XAI) module based on large language models (LLMs). This module uses natural language to explain to the user why the system selected those sentences for summarization. Compared with prior zero-shot extractive systems such as MatchSum, which are applied without fine-tuning, our framework requires no parameter updates on any benchmark.*

Keywords: *Text summarization, Semantic matching, Centrality density, Positional bias, Maximal Marginal Relevance, Pre-trained language model, Explainable AI.*

Received: March 8, 2026; Received in revised form: June 14, 2026; Accepted: June 15, 2026

1. INTRODUCTION

The explosion of digital information has led to an increasing need for automated summarization systems to distill large amounts of material into concise and coherent summaries. Among these, extractive summarization (i.e., identifying and selecting representative sentences from the original text) remains a popular method because it ensures information accuracy and requires less computational time than abstract summarization methods [1, 2].

Recently, advances in pre-trained language models (PLMs) such as BERT [3] and PhoBERT [4] have signifi-

cantly improved extractive summarization methods. Specifically, the MatchSum model [5] treats summarization as a semantic matching problem. It selects candidate sentences based on their similarity to the original text in a shared embedding space. However, this method often ignores the text's natural structure. Articles often have important sentences at the beginning, while academic papers have the main content scattered throughout. This drawback is evident when we use the model in practice across many different domains and languages.

VNDS [6], a Vietnamese news corpus, is a telling example: even Lead-3 - taking the first three sentences - gets close to what complex AI models produce. Product-review datasets like AbMuSu [7] tell the opposite story; the same approach fails there. Semantic features alone are not enough to cover text types this different.

We propose a Hybrid Rank-MMR model to solve this limitation. This method combines semantic matching with features of article structure. Moreover, it automatically controls diversity through the improved MMR algorithm to remove duplicate sentences.

This research has four main contributions:

- We analyze the relationships among semantic similarity, sentence position, and sentence centrality across multiple datasets. This helps explain why current models yield inconsistent results.
- We propose a hybrid re-ranking mechanism. This method scores sentences by combining meaning, position, and centrality through independent weights.
- We refine the MMR selection algorithm. This method automatically measures sentence similarity, thereby adjusting the penalty level to remove duplicate sentences.
- We integrate an Explainable Artificial Intelligence (XAI) module through chain-of-thought reasoning. This module can explain why it selects those sentences.

The remainder of the paper is structured as follows: Section 2 presents related works. Section 3 introduces our Hybrid Rank-MMR method. Section 4 then details the experimental setup. The evaluation and analysis results are detailed in Sections 5 and 6. Finally, Section 7 summarizes the methodology and suggests some directions for future research.

2. RELATED WORKS

2.1. EXTRACTIVE TEXT SUMMARIZATION

Extractive text summarization is the process of selecting the most representative sentences from a source text. Previous proposals relied on statistics or graphs to find important sentences, such as TextRank [8]. This algorithm represents the text as a graph: each sentence is a point (node), and the lines connecting the points represent the degree of similarity between sentences (e.g., TF-IDF score). While it is fast, this method only sees local connections and does not understand the flow of the entire text. To overcome this, feature-based systems [9] were developed. They scored sentences based on position, length, and the number of proper nouns. Thanks to this, the "lead bias" was discovered – that is, the main idea of the article is often found in the first few sentences. However, the weakness of this method is that when moving to a new field, one has to manually find these features.

Later, the explosion of deep learning brought the sequence model into the summarization problem. For example, SummaRuNNer [10] uses a recurrent neural network to score each sentence as keep-or-discard. A major advance came with the introduction of the Transformer architecture [11] and large language models. For example, BERTSum [12] used BERT to read and understand the two-dimensional context of a sentence, then classified whether the sentence was worth selecting. This sentence-by-sentence scoring method yielded very good results.

However, it scores each sentence independently, ignoring inter-sentence dependencies. To address this, MatchSum [5] takes a different angle. Rather than treating summarization as classification, MatchSum frames it as semantic matching: the system no longer scores each sentence in isolation. Instead, it groups sentences into candidate summaries and chooses the candidate whose meaning sits closest to the source text. This keeps the summary coherent and avoids the redundancy that sentence-level scoring tends to produce. Even so, we found that semantic matching alone is not enough on complex datasets, where the semantic signal is weak or runs against the natural structure of the text [13].

2.2. VIETNAMESE LOW-RESOURCE SUMMARIZATION

Research on English text summarization has matured over the years. In contrast, Vietnamese is a low-resource language with distinct features, meaning it often faces severe data shortages. However, the field has recently made significant progress thanks to new large-scale datasets.

VNDS [6] is a dataset for single-document text summarization. This dataset has also shown the general writing style of news, where important sentences are usually at the beginning of the source text. Next, ViMs [14] is a dataset for multi-document text summarization. This dataset has a major challenge in information synthesis, for example, how to filter duplicate information between sentences and arrange events in the correct chronological order.

Most recently, AbMuSu [7] is a dataset built for the task of abstract text summarization of product reviews. User comments in this dataset are often disorganized

and unstructured. For example, introductory sentences are often off-topic, while the main idea is scattered throughout the paragraph.

2.3. RE-RANKING AND EXPLAINABILITY

Re-ranking is an important step in improving summarization quality in many NLP systems. Maximal Marginal Relevance (MMR) [15] is a fundamental method for selecting diverse candidates for summarization. It helps balance the topic's relevance with the sentence's diversity.

Subsequently, deep learning models such as SummaReranker [16] and SimCSE [17] have significantly improved semantic-based sentence ranking through PLMs. However, most ranking methods select sentences without explaining why.

Explainable artificial intelligence (XAI) addresses this problem [18]. Traditional XAI tools such as LIME [19] and SHAP [20] are designed for classification tasks and are therefore not well-suited for text generation. Instead, recent studies have shown that large language models (LLMs) can perform well for evaluation [21] and inference [22].

Therefore, we use chain-of-thought prompting with LLMs to explain why certain sentences in the summary were selected. This approach increases user confidence in the model's effectiveness.

2.4. SUSTAINABLE SUMMARIZATION

Currently, large AI models require significant computational resources. NLP researchers are concerned about the negative environmental impact of their work. Therefore, some recent studies [23] have begun to call for a shift toward more energy-efficient designs, also known as the "Green AI" trend.

Discussions [24] show that although large language models (LLMs) can generate high-quality abstractive summaries, the energy they consume per run is many times higher than that of standard extractive models. This has led to a renewed interest in hybrid model architectures: using a compact model to select the main content first, and then using a large model to refine it at the final step [25, 26].

Our research follows this direction by combining a BERT encoder with a feature-based re-ranking mechanism. This architecture offers a more computationally efficient alternative to end-to-end LLM-based summarization pipelines.

3. PROPOSED METHODOLOGY

The overall architecture is illustrated in Fig. 1. This method comprises five stages: (i) embedding-based candidate pruning, (ii) dual-encoder semantic matching, (iii) feature-aware score fusion, (iv) MMR-based diversity selection, and (v) rationale generation for explainability. We distinguish two decoding variants: **Hybrid Rank** selects the highest-scoring candidate summary from the candidate pool; **Hybrid Rank-MMR** selects k sentences sequentially via MMR using sentence-level hybrid scores to enforce diversity.

3.1. CANDIDATE GENERATION AND PRUNING

A core difficulty in extractive multi-document summarization is the combinatorial search space. Given a document cluster with M sentences, selecting a summary of k sentences induces $\binom{M}{k}$ possible candidates. For typical values (e.g., $M = 50$ and $k = 3$), this already yields over 19,600 combinations. To enable efficient decoding under a strict training-free setting, we adopt a coarse-to-fine strategy that prunes sentences using only frozen pretrained representations. Concretely, we reuse the shared encoder Φ to pre-compute: (*i*) a sentence embedding h_{s_i} for each sentence s_i, and (ii) a document-cluster embedding h_D representing the overall semantic context. We then define a sentence-level salience proxy by cosine similarity (we use Cos(.) denote the cosine similarity function):

$$\tilde{p}(s_i) = Cos(h_{s_i}, h_D) \quad (1)$$

Intuitively, $\tilde{p}(s_i)$ measures how well a sentence aligns with the global topic of the cluster in the embedding space, providing a lightweight relevance signal without any supervision. We sort all sentences by $\tilde{p}(s_i)$ and retain the top K' sentences, yielding a pruned sentence set

$$S_{pruned} = \{s_{(1)'}.....,s_{(K')}\}.$$

From S_{pruned}, we generate the candidate summary pool by enumerating all k-sentence combinations:

$$C = \{C_1,.....,C_P\}, \quad with \quad P = \binom{K'}{k} \quad (2)$$

where each candidate $C_j (j = \overline{1,P})$ is a set of k sentences. This pruning step reduces the candidate space by orders of magnitude while remaining strictly training-free: we do not train an extractor, and we do not fine-tune the encoder.

Throughout the paper, we use $k = 3$ as the summary budget and set $K' = 15$ as the pruning threshold. The choice of $K' = 15$ balances computational efficiency against oracle coverage: with $K' = 15$, we retain $\binom{15}{3} = 455$ candidates. Empirical analysis indicates that this threshold retains over 90% of oracle-optimal sentences, while $K' \in [12,20]$ yields comparable performance.

3.2. SEMANTIC MATCHING

Once the candidate pool C is constructed, we score each candidate summary against the document cluster in a shared embedding space. This follows a matching-based view of summarization popularized by MatchSum [5]: a high-quality extractive summary should be semantically close to its source documents while being substantially shorter.

3.2.1. Encoder architecture

We employ a Siamese dual-encoder with shared parameters Φ to embed both sentences/candidates and the document cluster. For English datasets, Φ is instantiated as *bert-base-uncased* [3]; for Vietnamese datasets, we use *vinai/phobert-base* [4]. All encoders are used as frozen feature extractors: we do not fine-tune *BERT/PhoBERT*, and we do not train any task-specific head on the evaluated datasets.

Let $\Phi(.)$ map an input token sequence to contextual token representations $H \in R^{L \times D}$, where L is the length of the sequence and D is the hidden dimension. We obtain a fixed-size embedding by mean pooling over the token dimension:

$$MeanPool(H) = \frac{1}{L}\sum_{l=1}^{L} H_l \in R^{L \times D} \quad (3)$$

Sentence embeddings. For each sentence s_i, we compute:

$$h_{s_i} = MeanPool(\Phi(s_i)) \quad (4)$$

Document-cluster embedding. To represent a multi-document cluster D, we aggregate the sentence embeddings:

$$h_D = \frac{1}{M}\sum_{i=1}^{M} h_{s_i} \quad (5)$$

This formulation avoids long-context limitations by avoiding the need to concatenate all documents into a single input, while still capturing the overall semantic centroid of the cluster.

Candidate embedding. For a candidate summary C_j consisting of k selected sentences, we obtain its embedding by aggregating the corresponding sentence embeddings:

$$h_{C_j} = \frac{1}{k}\sum_{s_i \in C_j} h_{s_i} \quad (6)$$

3.2.2. Semantic scoring

We measure the semantic relevance of a candidate summary by cosine similarity between the candidate embedding and the document-cluster embedding:

$$S_{sem}(C_j) = \mathrm{Cos}(h_{C_j}, h_D) = \frac{h_{C_j} \cdot h_D}{\|h_{C_j}\| \|h_D\|} \quad (7)$$

A higher $S_{sem}(C_j)$ score puts the candidate closer to the semantic centroid of the cluster - a compact form of the cluster's meaning.

3.3. FEATURE-AWARE RE-RANKING

While S_{sem} is very effective, it has a weakness: it ignores sentence position (it considers the whole text as just a bunch of unordered sentences). To address this limitation, we incorporate additional discourse-aware features.

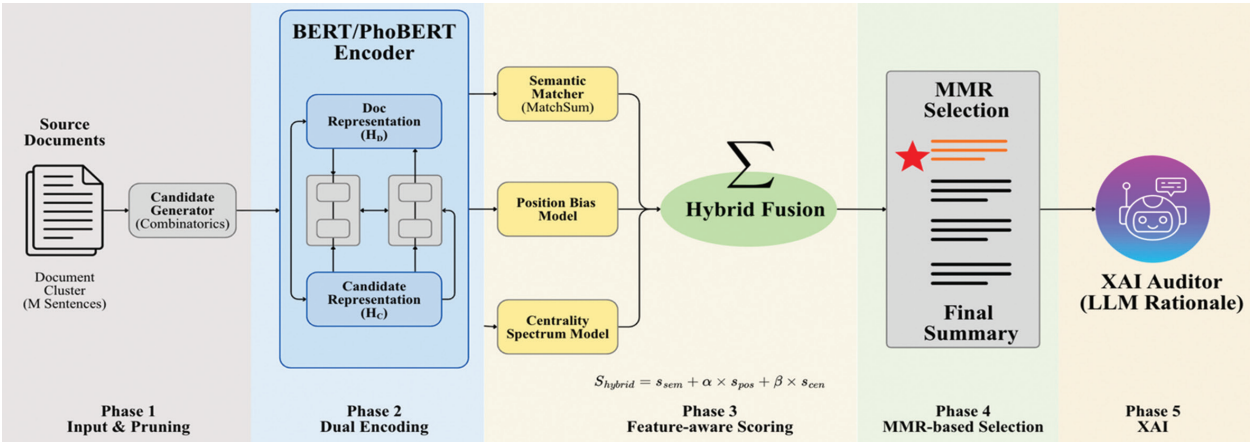


Fig. 1. Hybrid Rank-MMR architecture. (**1**) The input document cluster is encoded using a pre-trained language model to produce potential summary combinations. (**2**) The parallel scoring module computes three orthogonal scores: semantic similarity (green), positional bias (blue), and semantic centrality (red). (**3**) These scores are fused through a weighted combination to rank candidates. (**4**) MMR-based selection with adaptive diversity control. (**5**) The top-ranked candidate is validated by an XAI auditor module

3.3.1. Positional prior (Lead bias)

In domains such as news (CNN-DailyMail (CNN-DM) [27], VNDS [6]), critical information is often found in the beginning of documents. Semantic models can skip this pattern, potentially selecting redundant details from article bodies. We model this bias explicitly using a linear decay function:

$$S_{pos}(s_i) = 1 - \frac{i-1}{M-1}, i \in \{1,...,M\} \quad (8)$$

where i denotes the 1-indexed position of sentence s_i and M is the total number of sentences in the document cluster (the same M introduced in Section 3.1).

The summary-level positional score is computed as the average of the positional scores of its selected sentences.

This mechanism heavily penalises summaries that omit sentences in the introduction, helping the model favour informative early sentences.

3.3.2. Centrality score (Semantic density)

In unstructured domains such as user reviews, lead sentences are often irrelevant. Here, we seek sentences representing the semantic centroid of the cluster. We define the centrality score as the average pairwise cosine similarity between a sentence and all other sentences:

$$S_{cen}(s_i) = \frac{1}{K'-1} \sum_{j \neq i} \text{Cos}(h_{s_i}, h_{s_j}) \quad (9)$$

where K' is the number of sentences in the pruned pool S_{pruned} ($K' = 15$ as defined in Section 3.1).

Unlike the pruning proxy $\tilde{p}(s_i)$ (Eq. 1), which measures similarity to the document centroid, the centrality score captures how well a sentence represents the consensus among all sentences. Unlike S_{sem}, which scores complete summary candidates, the centrality feature aggregates individual sentence-level inter-similarity. This mechanism serves as a noise filter, rejecting outlier sentences that may exhibit high lexical overlap but low semantic centrality.

3.3.3. Hybrid score fusion and parameter estimation

We propose a unified scoring function that combines semantic matching with discourse features:

$$S_{hybrid}(C_j) = S_{sem}(C_j) + \alpha.\bar{S}_{pos}(C_j) + \beta.\bar{S}_{cen}(C_j) \quad (10)$$

where α and β are weighting parameters controlling the influence of positional bias and semantic centrality, respectively. $\bar{S}_{pos}(C_j)$ and $\bar{S}_{cen}(C_j)$ denote the mean positional and centrality scores across sentences in candidate C_j. All component scores are z-normalized per instance across all P candidates (i.e., for each score component, we subtract the mean and divide by the standard deviation computed over $\{C_1, \ldots, C_P\}$) before fusion. The complete scoring procedure is summarized in **Algorithm 1.**

Algorithm 1: Hybrid Rank Scoring

Require: Candidate pool C, document embedding $\boldsymbol{h}_D$, parameters α, β

Ensure: Ranked candidate list R

1. for each candidate $C_j \in C$ do
2. $S_{sem}(C_j) \longleftarrow \text{Cos}(\boldsymbol{h}_{C_j}, \boldsymbol{h}_D)$
3. $\bar{S}_{pos}(C_j) \longleftarrow \frac{1}{k} \sum_{s_i \in C_j} S_{pos}(s_i)$
4. $\bar{S}_{cen}(C_j) \longleftarrow \frac{1}{k} \sum_{s_i \in C_j} S_{cen}(s_i) \quad \{S_{cen}(s_i) = \frac{1}{K'-1} \sum_{j \neq i} \text{Cos}(\boldsymbol{h}_{s_i}, \boldsymbol{h}_{s_j})\}$
5. $S_{hybrid}(C_j) \longleftarrow S_{sem}(C_j) + \alpha.\bar{S}_{pos}(C_j) + \beta.\bar{S}_{cen}(C_j)$
6. **end for**
7. $R \longleftarrow Sort(C, S_{hybrid}, descending)$
8. **return** R

Independent weighting. We do not constrain $\alpha + \beta = 1$. The parameters serve as reliability coefficients indicating the trustworthiness of each feature for a given corpus rather than trade-off weights. Position and centrality capture orthogonal information - structural discourse versus semantic importance - allowing both to contribute simultaneously when appropriate. The semantic score S_{sem} always contributes with unit weight, while α and β modulate the additional contribution of structural features.

Position weight estimation. The parameter α is estimated from lead bias characteristics using only source document statistics. M is the number of sentences in the document cluster. We define the lead set $L = \{s_1, s_2, \ldots, s_{\lfloor M/3 \rfloor}\}$ (first third) and the tail set $T = \{s_{M-\lfloor M/3 \rfloor+1}, \ldots, s_M\}$ (last third).

The lead-bias ratio r_{lead} measures the relative semantic similarity of lead versus tail sentences to the document embedding h_D:

$$r_{lead} = \frac{\frac{1}{|L|} \sum_{s_i \in L} \text{Cos}(h_{s_i}, h_D)}{\frac{1}{|T|} \sum_{s_j \in T} \text{Cos}(h_{s_j}, h_D)} \quad (11)$$

Let $n_L = \lfloor M/3 \rfloor$ and let Q be the set of n_L sentences with the highest cosine similarity $\text{Cos}(h_{s_i}, h_D)$ across the entire document cluster. The first-third salience ratio r_{first} is:

$$r_{first} = \frac{|Q \cap L|}{n_L} \quad (12)$$

The position weight α is then obtained via the sigmoid function $\sigma(x) = (1/(1+e^{-x}))$:

$$\alpha = \sigma\left(\frac{1}{2} r_{lead} + \frac{1}{2} r_{first}\right) \quad (13)$$

High r_{lead} and r_{first} indicate strong lead bias, yielding high α.

This estimation relies only on source-document statistics (no reference summaries) and requires no fine-tuning on the target dataset. We therefore describe our setting as ***a training-free adaptation***.

Centrality weight estimation. The parameter β reflects the semantic diversity of the corpus:

$$\beta = 1 - \bar{\rho} \quad (14)$$

Where $\bar{\rho}$ is the average pairwise cosine similarity among sentences (pool similarity). In practice, mean-pooled embeddings from pre-trained language models yield non-negative cosine similarities, ensuring $\bar{\rho} \in [0,1]$ and consequently $\beta \in [0,1]$. When sentences are semantically diverse (low $\bar{\rho}$), centrality becomes more valuable for identifying representative content; when sentences are homogeneous (high $\bar{\rho}$), centrality provides less discriminative power.

3.4. MMR-BASED DIVERSITY SELECTION

The hybrid score finds salient sentences but says nothing about redundancy. Source documents in summarization often overlap, and a top relevance ranking can

pick three sentences saying almost the same thing. With a 3-sentence budget, that means up to two-thirds of the summary repeats itself. The classic Maximal Marginal Relevance (MMR) framework [15] handles this: it adds a penalty that discounts candidates similar to what is already in the summary, mixing relevance with novelty.

3.4.1. Adaptive diversity control

When evaluating and analyzing the parameters of the datasets, the parameters vary considerably between datasets. Homogeneous datasets (high sentence similarity) benefit from weaker diversity penalties, whereas heterogeneous datasets require stronger penalties to maximize information coverage. We first compute the pool similarity $\bar{\rho}$, representing the average pairwise cosine similarity among all K' candidate sentences:

$$\bar{\rho} = \frac{2}{K'(K'-1)} \sum_{i<j} \text{Cos}(h_{s_i}, h_{s_j}) \tag{15}$$

The diversity parameter γ is then computed as:

$$\gamma = \gamma_{\max} - (\gamma_{\max} - \gamma_{\min}).\bar{\rho} \tag{16}$$

Following the classical MMR formulation [15], where the relevance weight λ typically ranges in [0.5,0.7], we set $\gamma_{\min} = 0.3$ and $\gamma_{\max} = 0.7$ as diversity penalty bounds (since $\gamma = 1-\lambda$). Since $\bar{\rho} \in [0,1]$ (see Section 3.3.3), this formulation ensures $\gamma \in [\gamma_{\min}, \gamma_{\max}]$: homogeneous corpora (high $\bar{\rho}$) receive lower γ, while heterogeneous corpora (low $\bar{\rho}$) receive higher γ to maximize coverage.

3.4.2. Hybrid-MMR selection

The final selection combines hybrid scores with diversity penalties. For the sentence-level variant, we define the hybrid score of an individual sentence as:

$$S_{hybrid}(s_i) = \text{Cos}(h_{s_i}, h_D) + \alpha.S_{pos}(s_i) + \beta.S_{cen}(s_i) \tag{17}$$

Let $\tilde{S}_{hybrid}(s_i)$ denote the z-normalized version of $S_{hybrid}(s_i)$, computed per instance across all sentences in S_{pruned}. For each candidate sentence, the MMR score is:

$$MMR(s_i) = (1-\gamma).\tilde{S}_{hybrid}(s_i) - \gamma.\max_{s_j \in S_{sel}} \text{Cos}(h_{s_i}, h_{s_j}) \tag{18}$$

Where S_{sel} is the set of already-selected sentences. For the first selected sentence ($S_{sel} = \emptyset$), the diversity penalty is omitted. The algorithm proceeds greedily, selecting at each iteration the sentence that maximizes MMR. The complete selection procedure is detailed in **Algorithm 2**.

Algorithm 2: Sentence-level Hybrid Rank-MMR Selection

Require: Pruned sentence pool $\tilde{S}_{hybrid'}$ budget k, diversity γ

Ensure: Selected summary S_{sel}

1. Compute $\tilde{S}_{hybrid}(s_i)$ for all $s_i \in S_{pruned}$ {z-normalize hybrid scores}
2. $S_{sel} \leftarrow \emptyset$
3. **While** $|S_{sel}| < k$ **do**
4. **for** each $s_i \in S_{pruned} \setminus S_{sel}$ **do**
5. **if** $S_{sel} = \emptyset$ **then**
6. $MMR(s_i) \leftarrow \tilde{S}_{hybrid'}(s_i)$
7. **else**
8. $MMR(s_i) \leftarrow (1-\gamma).\tilde{S}_{hybrid}(s_i) - \gamma.\max_{s_j \in S_{sel}} \text{Cos}(\boldsymbol{h}_{s_i}, \boldsymbol{h}_{s_j})$
9. **end if**
10. **end for**
11. $S^* \leftarrow \text{argmax}_{s_i} MMR(s_i)$
12. $S_{sel} \leftarrow S_{sel} \cup \{S^*\}$
13. **end while**
14. **return** S_{sel}

3.5. EXPLAINABLE AI

To address transparency concerns inherent in neural systems, we generate post-hoc rationales for summary selection. We formulate this as a comparative verification task: the language model is prompted not to generate summaries (an abstractive task) but to justify the extractive model's selection relative to a base model (e.g., Lead-3). The rationale generation procedure is formalized in **Algorithm 3**. This approach enables users to verify whether high scores correlate with actual information gain.

Algorithm 3: Chain-of-Thought Rationale Generation

Require: Source text T', selected summary C^*, baseline summary C_{base}

Ensure: Textual rationale R

1. $P \leftarrow$ "You are a professional editor."
2. $P \leftarrow P +$ "[Source]: "$+ T'$
3. $P \leftarrow P +$ "[Summary A]: "$+ C_{base}$
4. $P \leftarrow P +$ "[Summary B]: "$+ C^*$
5. $P \leftarrow P +$ "Compare and explain which provides better coverage."
6. $R \leftarrow$ LLM(P) {Large Language Model}
7. **return** R

4. EXPERIMENTAL SETUP

4.1. DATASETS

We evaluate on seven benchmark datasets spanning two languages. Table 1 summarizes the statistics.

Table 1. Dataset statistics. Avg. Src and Avg. Sum indicate the average number of sentences in source documents and reference summaries, respectively

Dataset	Domain	Test size	Avg. Src	Avg. Sum
English				
CNN-Daily Mail	News	11,489	39.78	3.68
Multi-News	News	5,622	82.73	9.97
XScience	Science	5,093	38.7	5.2
WCEP	Wikipedia	1,022	52.4	1.4
Vietnamese				
VNDS	News	22,644	17.72	1.23
ViMs	News	300	85.3	6.8
AbMuSu	Reviews	300	61.07	4.93

CNN-DailyMail [27] is a single-document benchmark with strong lead bias. Multi-News [28] groups multiple articles into one summary, while WCEP [29] gathers short event summaries from Wikipedia. Scientific papers are the focus of XScience [30]. On the Vietnamese side, VNDS [6] is a news corpus that also leans on position; ViMs [14] covers multi-document Vietnamese news. AbMuSu [7] is the most loosely structured of the seven - product reviews with minimal positional cues.

All parameter estimations (α, β, γ) are computed solely from **source document statistics** - specifically, sentence embeddings and their pairwise similarities. We do not use reference summaries for parameter tuning, nor do we perform grid search on validation sets. This design ensures that the framework can be applied to new datasets immediately without any labeled data.

Vietnamese preprocessing. For Vietnamese datasets, we apply word segmentation using *underthesea*[1] before encoding with PhoBERT, following the standard preprocessing pipeline for Vietnamese NLP. Sentence boundaries are detected using rule-based heuristics.

4.2. BASIC MODELS AND IMPLEMENTATION

We implement two basic models: **Lead-3** selects the first three sentences; **MatchSum (ZS)** applies the matching intuition of MatchSum but without fine-tuning, we only compute cosine similarity in the shared embedding space using a frozen BERT/PhoBERT encoder.

This revision adds six stronger baselines. Two are graph-based, TextRank [8] and LexRank [31]. One is the BERT-based zero-shot extractor BERTSum-ZS [12]. PacSum [32] is an unsupervised position-aware ranker. The last two are large language models: Gemini 2.5 Flash (commercial) and Qwen 2.5-7B (open-weights). None of these methods sees a labelled summary at any point, so the comparison stays apples-to-apples with our training-free pipeline.

We use a fixed extractive budget of $k = 3$ sentences for controlled comparison. Parameters α, β, and γ are estimated automatically from corpus statistics as described in Section 3.3.3. Evaluation employs ROUGE scores [33].

5. RESULTS

This section evaluates our method. We report ROUGE-1, ROUGE-2, and ROUGE-L F1 scores: ROUGE-1 captures unigram overlap with the reference, ROUGE-2 captures bigram overlap, and ROUGE-L captures the longest common subsequence.

5.1. MAIN RESULTS

Table 2 and Table 3 compare our method against the two basic models. We present the incremental contribution of our Hybrid scoring mechanism and the full system with adaptive MMR.

Table 2. Evaluation results on English datasets (ROUGE F1). Best scores are in bold. "Δ vs basic" gives the gain over the stronger of Lead-3 and MatchSum (ZS); "Δ vs strongest baseline" gives the gain over the strongest training-free baseline below Hybrid Rank-MMR among the eight comparators

Method	CNN-DM			Multi-News			XScience			WCEP		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	40.43	15.60	36.67	30.28	8.70	26.06	25.88	3.66	22.56	31.30	9.20	25.60
MatchSum (ZS)	41.48	16.35	37.40	32.07	9.37	27.48	25.23	3.95	23.56	30.37	8.54	24.48
TextRank [9]	37.06	12.71	32.77	32.60	9.84	27.96	26.04	4.02	22.95	31.97	9.64	26.54
LexRank [31]	37.09	12.67	32.93	32.71	9.72	27.78	26.44	4.11	23.23	30.52	9.21	25.19
BERTSum-ZS [12]	35.73	11.94	31.79	32.29	10.04	28.06	27.38	4.35	22.97	28.64	8.17	23.12
PacSum [32]	41.17	16.25	37.07	31.56	9.63	27.24	26.84	4.43	23.17	29.99	8.92	25.09
Gemini 2.5 Flash	39.66	14.57	35.46	34.90	10.68	29.71	27.30	4.10	23.41	31.63	9.92	25.81
Qwen 2.5-7B	39.11	14.56	35.02	33.93	10.68	29.15	27.44	4.28	23.43	31.17	9.49	25.95
Hybrid Rank	42.68	17.42	38.69	33.84	10.17	28.90	27.88	4.55	24.19	31.67	9.39	26.03
Hybrid Rank-MMR (Our)	**42.90**	**17.46**	**38.73**	**34.63**	**10.67**	**29.58**	**28.17**	**4.57**	**24.35**	**33.67**	**10.52**	**28.43**
Δ vs basic (Lead-3 / MatchSum)	**+1.42**	**+1.11**	**+1.33**	**+2.56**	**+1.30**	**+2.10**	**+2.29**	**+0.62**	**+0.79**	**+2.37**	**+1.32**	**+2.83**
Δ vs strongest baseline	**+1.42**	**+1.11**	**+1.33**	**-0.27**	**-0.01**	**-0.13**	**+0.73**	**+0.14**	**+0.79**	**+1.70**	**+0.60**	**+1.89**

[1] https://github.com/undertheseanlp/underthesea

Table 3. Evaluation results on Vietnamese datasets (ROUGE F1). Best scores are in bold. "Δ vs basic" gives the gain over the stronger of Lead-3 and MatchSum (ZS); "Δ vs strongest baseline" gives the gain over the strongest training-free baseline below Hybrid Rank-MMR among the eight comparators

Method	VNDS			ViMs			AbMuSu		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	25.61	10.37	22.30	29.63	11.40	24.40	40.12	24.57	33.41
MatchSum (ZS)	25.31	10.37	22.64	31.86	12.79	26.21	40.99	25.17	34.07
TextRank	23.46	8.04	21.70	33.61	13.06	28.82	40.55	22.41	31.38
LexRank	23.73	8.29	21.04	34.69	14.51	29.40	42.79	25.43	33.50
BERTSum-ZS	25.21	9.56	22.15	36.90	16.34	30.41	41.11	22.90	31.66
PacSum	25.08	9.56	21.96	35.32	15.37	29.45	41.53	25.40	34.41
Gemini 2.5 Flash	24.92	9.49	22.12	35.96	14.31	30.62	43.73	27.48	35.74
Qwen 2.5-7B	24.78	9.32	21.84	34.84	13.38	29.36	40.58	24.03	33.51
Hybrid Rank	26.49	10.40	22.84	35.93	15.77	29.76	42.49	26.40	35.13
Hybrid Rank-MMR (Our)	**26.90**	**10.71**	**23.23**	**38.54**	**17.23**	**31.38**	**43.54**	**27.03**	**35.63**
Δ vs basic models (Lead-3/MatchSum)	**+1.29**	**+0.34**	**+0.59**	**+6.68**	**+4.44**	**+5.17**	**+2.55**	**+1.86**	**+1.56**
Δ vs strongest baseline	**+1.29**	**+0.34**	**+0.59**	**+1.64**	**+0.89**	**+0.76**	**-0.19**	**-0.45**	**-0.11**

5.2. PERFORMANCE ANALYSIS

The results show improvements across all seven datasets. The ROUGE-1 (R-1) scores increase by +1.29 to +6.68 compared to the best baseline model.

To put these gains in context, a one-point ROUGE-1 increase corresponds to recovering roughly one additional informative unigram per reference tokens, which qualitatively manifests as better coverage of named entities, dates and key facts in the produced summaries (the case study in Section 6.4 makes this concrete on a CNN-DM article).

5.2.1. Basic comparison patterns

An interesting pattern emerges when comparing the two basic models. In four datasets (CNN-DM, VNDS, XScience, WCEP), Lead-3 performs competitively with MatchSum (ZS) - the gap is merely 1.05 points on CNN-DM, and Lead-3 actually wins on VNDS (25.61 vs 25.31 on R-1), XScience (25.88 vs 25.23 on R-1), and WCEP (31.30 vs 30.37 on R-1). This result shows that sentence position remains highly effective for structured texts. Conversely, MatchSum (ZS) beats Lead-3 on the remaining datasets: Multi-News (32.07 vs 30.28 on R-1), ViMs (31.86 vs 29.63 on R-1), and AbMuSu (40.99 vs 40.12 on R-1). It shows that in multi-document or unstructured texts, sentence position becomes unreliable.

5.2.2. Magnitude of improvements

The results show that our proposed model clearly outperforms the basic models. The biggest improvement is on the ViMs dataset (+6.68 on R-1), followed by Multi-News (+2.56 on R-1), AbMuSu (+2.55 on R-1), WCEP (+2.37 on R-1), and XScience (+2.29 on R-1). These datasets benefit the most from our hybrid approach because relying only on Lead-3 or MatchSum is not enough to capture the key information.

With the wider baseline set, Hybrid Rank-MMR remains the top method across five of the seven datasets - CNN-DM, XScience, WCEP, VNDS, and ViMs. For the other two, Multi-News and AbMuSu, the gap to Gemini 2.5 Flash is small and negative (−0.27 and −0.19 R-1), well within the noise floor for ROUGE on these sets. WCEP gains the most: +1.70 on R-1 and +1.89 on R-L over the strongest baseline below us. ViMs follows at +1.64 on R-1. Both are multi-document corpora, where the positional prior is weak and feature fusion can add more.

On CNN-DM and VNDS, the increase was more modest (+1.42 and +1.29 on R-1, respectively) because the Lead-3 model performed well. Nevertheless, the hybrid approach still yielded additional value by combining semantic matching with sentence position. The model carries over to several languages without requiring individual adjustments.

5.2.3. Incremental contribution analysis

We examine the significance of the duplicate removal algorithm via MMR using the γ parameter, and the results are presented in Table 4. This component provides a significant improvement for multi-document summarization datasets such as ViMs (+2.61 on R-1), WCEP (+2.00 on R-1), and AbMuSu (+1.05 on R-1), while single-document summarization datasets (CNN-DM, VNDS) show less improvement because there is less duplicate information between sentences. This suggests that when combining information from multiple sources in multi-document summarization, filtering out duplicate content is necessary.

Table 4. The impact of MMR: ROUGE-1 score improvements when upgrading from Hybrid Rank to Hybrid Rank-MMR

Dataset	Hybrid Rank	Hybrid Rank-MMR	Δ
CNN-DM	42.68	42.90	+0.22
Multi-News	33.84	34.63	+0.79
XScience	27.88	28.17	+0.29
WCEP	31.67	33.67	**+2.00**
VNDS	26.49	26.90	+0.41
ViMs	35.93	38.54	**+2.61**
AbMuSu	42.49	43.54	+1.05

5.3. ABLATION STUDY

To understand the effects of each component in our method, we gradually remove features from the Multi-News dataset (Table 5). The results show three findings: First, on this data set, the centrality score is three times more influential than the location score (+1.14 compared to +0.38 on R-1). This fits perfectly with our formula, automatically assigning a very high β weight (0.90). Second, combining features achieves significant results. The total increase when combined (+1.77 on R-1) is even higher than the simple addition of each part (+1.52 on R-1). This shows that location and concentration complement each other and do not overlap. Third, adding MMR increases R-1 by +0.79, showing that even with good rankings, we still need to force the model to choose more diverse sentences.

Table 5. Ablation study on Multi-News (ROUGE-1). Δ: improvement over the basic model. "–" indicates the basic model configuration

Configuration	R-1	Δ
MatchSum (ZS) (*)	32.07	-
(*) + Position ($\alpha.\bar{S}_{pos}$)	32.45	+0.38
(*) + Centrality ($\beta.\bar{S}_{cen}$)	33.21	+1.14
Hybrid Rank	33.84	+1.77
Hybrid Rank-MMR (Our)	**34.63**	**+2.56**

5.4. PARAMETER SENSITIVITY ANALYSIS

A key property of the proposed method is the automatic calculation of parameters from data without manual adjustment. Table 6 shows that these numbers clearly represent the true nature of the text.

For newspaper data (CNN-DM, VNDS), the weight α is much larger than β. This is consistent with newspaper writing habits: the opening sentence summarizes the entire article. On the contrary, with unstructured or multi-document data, β is larger than α. Because it cannot rely on location, the system must rely on semantic focus to find the main idea. On the WCEP data set, the parameters are in the middle (α=0.47, β=0.53). This can be explained by the fact that Wikipedia event summaries are both clearly structured and come from multiple sources, so both signals are equally important.

Table 6. Automatically estimated parameters per dataset. The α, β, and γ indicate positional bias, semantic signals, and diversity penalty, respectively

Dataset	α	β	γ	Characteristic
CNN-DM	0.83	0.17	0.51	Strong lead bias
VNDS	0.76	0.24	0.49	Strong lead bias
Multi-News	0.10	0.90	0.62	Multi-document
XScience	0.12	0.88	0.55	Scientific abstracts
ViMs	0.08	0.84	0.48	Multi-document
AbMuSu	0.17	0.83	0.65	User reviews
WCEP	0.47	0.53	0.63	Event summaries

5.5. VALIDATION OF PARAMETER ESTIMATION

To verify the formula for automatically calculating parameters, we compare it with the grid search method, which scans all values from 0.0 to 1.0 for α and β. This is shown in Fig. 2.

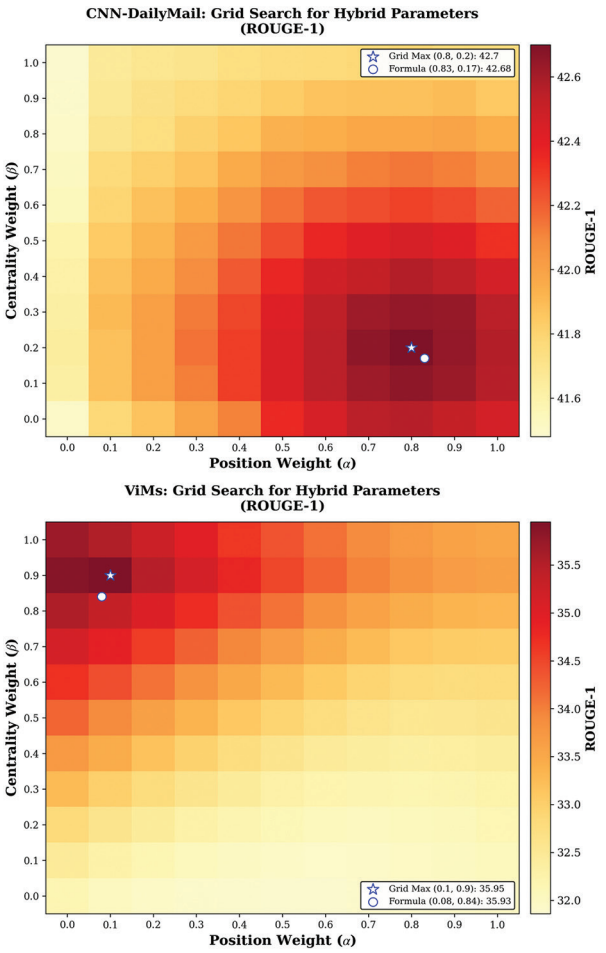


Fig. 2. Grid search validation for formula-estimated parameters. **Top:** CNN-DM. **Bottom:** ViMs. Circles indicate formula estimates

With CNN-Daily Mail, the formula automatically calculates α=0.83, β=0.17 (reaching 42.68 on R-1). The optimal level of exhaustive scanning is α=0.8, β=0.2 (reaching 42.70 on R-1). A difference of 0.02 is completely insignificant. Similar to ViMs, the formula yields

α=0.08, β=0.84 (35.93 on R-1), very consistent with the optimal level of exhaustion of α=0.1, β=0.9 (35.95 on R-1). These results suggest that automatic estimation performs comparably to grid search, while reducing the parameter selection time from $O(n^2)$ to just $O(1)$.

Fig. 3 also verifies the accuracy of the diversity parameter γ. For example, on CNN-DM, the formula yields γ=0.51 (42.90 on R-1), which is only 0.01 points away from the result from the grid search. On the ViMs set, the level γ=0.48 calculated by the formula also matches the actual optimal result.

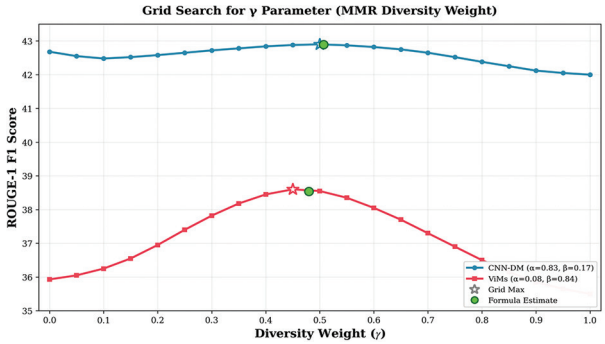


Fig. 3. Grid search validation for γ on CNN-DM and ViMs. Stars indicate grid search maxima; the circles indicate formula estimates

5.6. EFFICIENCY ANALYSIS

Section 2.4 raised the compute-cost question. We measured per-sample inference time on a single NVIDIA RTX 3090, 24 GB. The seven datasets all run on the same hardware. Each reported number is the mean number of sentences in documents.

One encoder per language does the work — BERT-Base at 110M parameters for English, PhoBERT at 135M parameters for Vietnamese. Position and centrality come from closed-form scalar functions of the same embedding matrix. No extra weights are learned, and no second model is loaded.

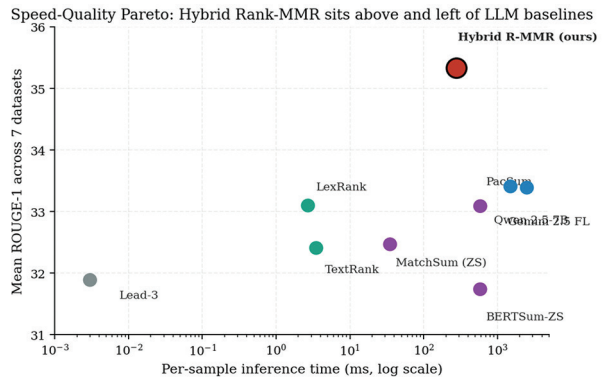


Fig. 4. Speed-quality Pareto. Mean ROUGE-1 across the seven datasets versus per-sample inference time (log scale). Hybrid Rank-MMR (red, ours) lies above and to the left of the LLM baselines and overlaps the latency range of the classical encoder methods

Fig. 4 plots ROUGE-1 against per-sample latency. Hybrid Rank-MMR sits on the upper-left of the Pareto frontier: higher mean ROUGE-1 than the LLM baselines (Gemini 2.5 Flash, Qwen 2.5-7B), and latency in the same band as classical encoder methods. Training cost is zero - no fine-tuning anywhere - against roughly 20 GPU-hours for supervised BERTSum [12]. Few methods in the benchmark match this combination of speed, quality, and no training time.

6. DISCUSSION

This section explains in more detail the internal workings and meanings of the components of the Hybrid Rank-MMR model.

6.1. FEATURE WEIGHT INTERPRETATION

The automatically estimated parameters accurately reflect the nature of each data set. Instead of being randomly chosen numbers, they are actually measures that help us analyze the text's structure.

6.1.1. Position weight

The parameter α measures whether the text focuses on writing important sentences at the beginning. With newspaper data (CNN-DM: α = 0.83, VNDS: α = 0.76), the position weight is very high. This reflects the correct structure for writing news articles, in which the main idea always comes first. On the contrary, when α has a small value below 0.2 (ViMs: 0.08, Multi-News: 0.10, XScience: 0.12, AbMuSu: 0.17), this shows that sentence position is no longer important. In tasks such as multi-document summarization or summarizing scientific articles, important sentences are often scattered throughout the document. These low α values accurately reflect the disorder in the structure of those types of texts.

6.1.2. Centrality weight

Data sets that rely heavily on significance (Multi-News: β = 0.90, XScience: β = 0.88, ViMs: β = 0.84, AbMuSu: β = 0.83) all have $\beta > 0.8$. This shows that when the text does not have a clear structure, concentration (distance to the main idea of the whole article) will become an effective way to select important sentences. On the contrary, when important sentences appear frequently at the beginning (CNN-DM: β = 0.17, VNDS: β = 0.24), adding focus points does not yield much improvement in results. The WCEP event data set falls in the middle (β = 0.53). It shows that Wikipedia summaries need a combination of substantial structure at the beginning of the text and variety in sentence meaning.

6.2. ADAPTIVE DIVERSITY ANALYSIS

The parameter γ is used to balance between choosing sentences that are on topic and choosing diverse sentences. The optimal γ value depends greatly on

the homogeneity of the dataset. Our automatic formulation captured this association through the internal similarity score $\bar{\rho}$. In homogeneous data sets, sentences often have similar meanings ($\bar{\rho}$ is high). If you penalize too many duplicates, the model will eliminate much useful content because the article doesn't have enough distinct ideas to keep. A typical example is the VNDS set ($\bar{\rho}$ = 0.71, γ = 0.49): Vietnamese articles on the same topic often have many overlapping sentences, so applying a moderate diversity penalty is enough. On the contrary, unstructured datasets pose a different problem. AbMuSu ($\bar{\rho}$ = 0.38, γ = 0.65) contains user comments with a variety of opinions, products, and styles. Without a strong duplicate penalty, the model can easily choose all the criticism and forget the praise (or vice versa). Therefore, pushing γ higher will help the summary cover a broader perspective.

The problem of multi-document summarization is different. The ViMs set (γ = 0.48) contains many articles about the same event, so the information is often repeated with minor changes. A moderate γ value here helps balance two requirements: avoiding duplication of information between articles and retaining supporting details to make the summary richer. This automatic adjustment removes the need for manual parameter tuning while maintaining strong performance.

6.3. ERROR ANALYSIS

Although the performance has clearly improved, when analyzing cases where the model runs incorrectly, we find that the extraction method still has certain weaknesses. First is the limit on sentence length. If the reference summary is only 10 words, but the sentence containing the shortest important information is 25 words long, the ROUGE score will be dragged down, no matter how good the sentence is. Second, there is a disadvantage in the preprocessing step. In Vietnamese data, we discovered many errors stemming from incorrect sentence segmentation. For example, in the ViMs collection, sometimes the article title is attached to the body of the article. This creates falsely important sentences that have high position scores but are not the summary sentence sought. For low-resource languages, focusing on considering preprocessing methods is also necessary to increase model efficiency. Third is when concentration is not effective. We hypothesize that a sentence with a high semantic level is a good sentence. However, this assumption is sometimes not appropriate. In some samples in the ViMs set, the label summary appears as distinct sentences rather than a general meaning sentence.

6.4. QUALITATIVE ANALYSIS: XAI CASE STUDY

The case study draws on one real CNN-DM article. The 15-sentence story is a road accident in Nottingham: a £80,000 Porsche 911 was crushed by a builder's lorry while both pulled aside for emergency services responding to a separate cyclist injury. The 3-sentence reference summary names three facts: the Porsche-lorry collision, the cause of the queue, and an onlooker's quote about the driver.

Hybrid Rank-MMR picks sentences 0, 2, 3 - one for each reference fact. Sentence 0 introduces the incident; sentence 2 explains the cyclist-traffic context; sentence 3 reports the onlooker quote. MatchSum (ZS) goes 0, 1, 5 instead. Sentence 1 repeats the lead with different wording; sentence 5 is a bystander aside that the reference never mentions. Hybrid Rank-MMR covers all three reference facts, MatchSum (ZS) covers only one. Table 7 lays out the selection grid. Table 8 reports the feature z-scores for our three picks.

Table 7. Per-sentence selection grid for CNN-DM article 7527. "X" indicates each method's pick. MatchSum (ZS) picks {0, 1, 5}; Hybrid Rank-MMR picks {0, 2, 3}

Index	Sentence	MatchSum (ZS)	Hybrid Rank-MMR
0	This was the heartbreaking sight a Porsche owner had to endure after crashing his pride and joy into a builder's lorry.	X	X
1	The owner was left devastated after his £80,000 911 Carrera was crushed by a builder's lorry as they pulled aside for emergency services in Nottingham.	X	
2	Both vehicles had been caught in traffic after a cyclist was injured on a nearby road and police and ambulance services were rushing to the scene.		X
3	One onlooker, who witnessed the crash, said: "The bloke looked absolutely distraught."		X
4	It was obviously his pride and joy.		
5	"You've got to feel for the man, it was a really nice car."	X	
6	"You would be gutted if you were him."		
7	Heartbreaking site: an onlooker who witnessed this crash in Nottingham said the driver looked absolutely distraught.		
8	It was obviously his pride and joy.		
9	The two-door German sports car was crushed while manoeuvring onto the side road, with the battery hanging out of the crumpled bonnet.		
10	Traffic continued to build up as paramedics attended to the female rider on nearby Radcliffe Road.		
11	Davies Road, where the Porsche 911 Carrera was crushed after colliding with a lorry while pulling aside in traffic queues caused by a nearby cyclist accident.		
12	Police said there were no life-threatening injuries reported and the cyclist was taken to hospital with a minor leg injury.		
13	A Nottinghamshire Police spokesperson said they were called just after 12:05 pm on Wednesday following the two-vehicle collision in Davies Road, West Bridgford.		
14	"No-one was injured during the collision."		

Table 8. Feature z-scores for the three Hybrid Rank-MMR picks. Sentence 2's negative z_{sem} shows the semantic signal alone would not pick it; the MMR step recovers it once sentence 0 has covered the collision fact

Sentence	Role	z_{sem}	z_{pos}	z_{cen}
Sent 0	Lead - collision fact	+1.61	+2.31	+0.70
Sent 2	Traffic-cause context	−0.53	+1.14	+0.15
Sent 3	Onlooker quote	+0.90	+0.71	+1.44

Sentences 0 and 3 come from complementary signals. Sentence 0 leads on position and semantics (z_{pos} = +2.31, z_{sem} = +1.61). Sentence 3 leads on centrality (z_{cen} = +1.44). Sentence 2 is the interesting case. Alone, its semantic z-score is negative - the sentence looks less cluster-central than the lead. The MMR step changes that. Once sentence 0 enters the summary, the diversity term penalises near-duplicates of the collision fact, and sentence 2 - the cyclist-traffic explanation - wins the next slot because nothing in the running summary covers that ground yet. MatchSum (ZS) omits this step and reverts to the lead-style sentence 1. This is the feature-fusion and MMR behaviour described in Sections 3.3 and 3.4, in a single sample.

6.5. QUANTITATIVE XAI EVALUATION

Section 6.4 shows what the explanation module does on a single sample. It does not show whether the rationales are faithful or whether they agree with off-the-shelf saliency methods. We add three quantitative checks on the dataset. The deletion protocol follows [34].

First, faithfulness via deletion (Fig. 5). We remove the K=3 sentences picked by the rationale and re-rank what remains. The drop in ROUGE-1 tells us how much those sentences mattered. Hybrid Rank-MMR's rationale gives a positive drop on every dataset; a random-sentence control gives a negative drop on every dataset. The picks are therefore not arbitrary.

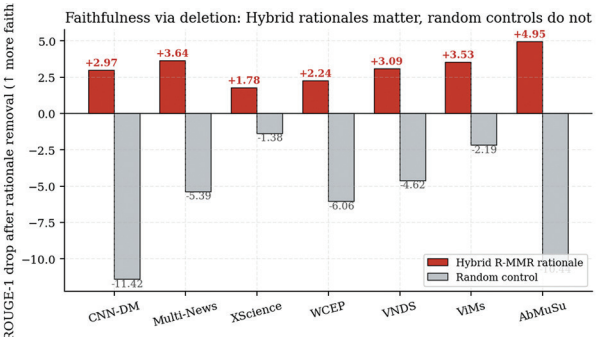


Fig. 5. Faithfulness via deletion. ROUGE-1 drop after removing the k=3 sentences identified by Hybrid Rank-MMR's rationale (red) versus a random-control rationale (grey), per dataset. Positive drops mean the rationale identifies critical content

Second, plausibility against LIME-style saliency. We run leave-one-sentence-out perturbation in the style of [29], then measure Jaccard agreement between LIME's top-k and each method's k=3 picks. Hybrid Rank-MMR beats MatchSum (ZS) on five of the seven datasets. On the other two, the gap is under 0.02. Absolute Jaccard stays in the 0.06-0.18 range, which is what we expect when top-k extractive selections compete.

Third, head-to-head preference from an external judge. We compare Hybrid Rank-MMR against MatchSum (ZS) pairwise, with the two summary positions swapped between runs to absorb order bias. Gemini 3 Flash acts as judge. Hybrid Rank-MMR is preferred on the multi-document and Vietnamese sets - Multi-News 51.5%, WCEP 53.0%, VNDS 52.5%, ViMs 57%, AbMuSu 55%, which are illustrated in Table 9 below.

Table 9. Quantitative XAI summary, three lenses, n=100 per dataset

Dataset	Faithfulness drop	LIME Δ vs MatchSum	Pairwise win % vs MatchSum
CNN-DM	+2.97	-0.014	48.0%
Multi-News	+3.64	+0.025	51.5%
XScience	+1.78	-0.007	49.0%
WCEP	+2.24	+0.002	53.0%
VNDS	+3.09	+0.007	52.5%
ViMs	+3.53	+0.031	57.0%
AbMuSu	+4.95	+0.006	55.0%

7. CONCLUSION

This paper proposes Hybrid Rank-MMR, a training-free extractive summarization framework. This method combines semantic matching with two basic text features: sentence position and semantic focus. The model's strength is that it automatically adjusts its parameters to suit each data set. The results show that, with news text with a clear layout (putting the main idea first), sentence position will play a decisive role. On the contrary, with unstructured data or summarizing many articles at the same time, semantic focus will be more important. In addition, the flexible MMR algorithm integration method automatically evaluates the overlap of information in sentences to filter out duplicate information.

Testing across 7 datasets, both in English and Vietnamese, shows that Hybrid Rank-MMR achieves superior results compared to the base models. The framework applies to many languages without per-dataset tuning. Finally, integrating the explanation module (XAI) helps the system automatically give reasons for choosing sentences in natural language, thereby increasing transparency and reliability for users.

However, this method still has some limitations. First, because the method focuses on extracting summaries, it is limited to sentence length. It cannot create compressed summaries, such as abstract summaries. Second, in a low-resource language like Vietnamese,

errors from the pre-processing step (especially sentence-separation errors) will directly affect the entire downstream calculation process. Third, the mechanism of prioritizing sentences with general representative meanings sometimes runs counter to the preferences of users seeking unique information.

In the future, we plan to develop a hybrid model that extracts and then compresses. This method not only retains the advantage of the extraction method of running light on the machine, but also produces highly compressed summaries. For low-resource languages, upgrading basic sentence processing tools will certainly improve overall results significantly. In addition, adding additional methods to find new and important information will bring superior results. Further, we will extend this model to handle summarization in many other languages.

ACKNOWLEDGMENT

This work is funded by the Vietnam Ministry of Education and Training under Grant B2024-KHA-06. The authors would like to express sincere gratitude.

DECLARATION OF THE USE OF AI TOOLS

The authors used Grammarly and ChatGPT (OpenAI, GPT-5.2) solely for language editing, including assistance with grammar checking, and improving clarity and readability. The AI tools were not used to generate scientific ideas, design the research methodology, conduct experiments, analyze the data, or draw conclusions. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the integrity and accuracy of this work. No AI tool is listed as an author, and AI tools are not cited as primary sources.

REFERENCES

[1] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond", Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, August 2019, pp. 280-290. doi: 10.18653/v1/K16-1028

[2] A. See, P. J. Liu, C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks", Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 30 July - 4 August 2017, pp. 1073-1083. doi: 10.18653/v1/P17-1099

[3] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2-7 June 2019, pp. 4171-4186. doi: 10.18653/v1/N19-1423

[4] D. Q. Nguyen, A. T. Nguyen, "PhoBERT: Pretrained Language Models for Vietnamese", Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16-20 November 2020, pp. 1037-1042. doi: 10.18653/v1/2020.findingsemnlp.92

[5] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, X. Huang, "Extractive Summarization as Text Matching", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5-10 July 2020, pp. 6197-6208. doi: 10.18653/v1/2020.acl-main.552

[6] V. H. Nguyen, T. C. Nguyen, M. T. Nguyen, N. X. Hoai, "VNDS: A Vietnamese Dataset for Summarization", Proceedings of the 6th NAFOSTED Conference on Information and Computer Science, Hanoi, Vietnam, 12-13 December 2019, pp. 375-380. doi: 10.1109/NICS48868.2019.9023886

[7] M. V. Tran, H. Q. Le, D. C. Can, Q. A. Nguyen, "VLSP 2022 AbMuSu Task Dataset: A Resource for Vietnamese Abstractive Multi-Document Summarization", International Journal of Asian Language Processing, Vol. 33, No. 1, 2023, p. 2350003. doi: 10.1142/S2717554523500030

[8] R. Mihalcea, P. Tarau, "TextRank: Bringing Order into Text", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25-26 July 2004, pp. 404-411.

[9] H. P. Edmundson, "New Methods in Automatic Extracting", Journal of the ACM, Vol. 16, No. 2, 1969, pp. 264-285. doi: 10.1145/321510.321519

[10] R. Nallapati, F. Zhai, B. Zhou, "SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents", Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4-9 February 2017, Vol. 31, No. 1, pp. 3075-3081. doi: 10.1609/aaai.v31i1.10958

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need", Advances in Neural Informa-

tion Processing Systems, Long Beach, CA, USA, 4-9 December 2017, Vol. 30, pp. 5998-6008.

[12] Y. Liu, M. Lapata, "Text Summarization with Pretrained Encoders", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3-7 November 2019, pp. 3730-3740. doi: 10.18653/v1/D19-1387

[13] L. Xing, W. Xiao, G. Carenini, "Demoting the Lead Bias in News Summarization via Alternating Adversarial Learning", Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1-6 August 2021, pp. 948-954. doi: 10.18653/v1/2021.aclshort.119

[14] N. T. Tran, M. Q. Nghiem, N. T. H. Nguyen, N. L. T. Nguyen, N. V. Chi, D. Dinh, "ViMs: A High-Quality Vietnamese Dataset for Abstractive Multi-Document Summarization", Language Resources and Evaluation, Vol. 54, No. 4, 2020, pp. 893-920. doi: 10.1007/s10579-020-09495-4

[15] J. Carbonell, J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24-28 August 1998, pp. 335-336. doi: 10.1145/290941.291025

[16] M. Ravaut, S. Joty, N. Chen, "SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization", Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22-27 May 2022, pp. 4504-4524. doi: 10.18653/v1/2022.acllong.309

[17] T. Gao, X. Yao, D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7-11 November 2021, pp. 6894-6910. doi: 10.18653/v1/2021.emnlp-main.552

[18] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences", Artificial Intelligence, Vol. 267, 2019, pp. 1-38. doi: 10.1016/j.artint.2018.07.007

[19] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August 2016, pp. 1135-1144. doi: 10.1145/2939672.2939778

[20] S. M. Lundberg, S. I. Lee, "A Unified Approach to Interpreting Model Predictions", Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017, Vol. 30, pp. 4768-4777.

[21] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore, 6-10 December 2023, pp. 2511-2522. doi: 10.18653/v1/2023.emnlp-main.153

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November - 9 December 2022, Vol. 35, pp. 24824-24837.

[23] J. Xu, W. Zhou, Z. Fu, H. Zhou, L. Li, "A Survey on Green Deep Learning", arXiv:2111.05193, 2021.

[24] A. S. Luccioni, S. Viguier, A. L. Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model", Journal of Machine Learning Research, Vol. 24, No. 253, 2023, pp. 11990-12004.

[25] J. Pilault, R. Li, S. Subramanian, C. Pal, "On Extractive and Abstractive Neural Document Summarization with Transformer Language Models", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online, 16-20 November 2020, pp. 9308-9319. doi: 10.18653/v1/2020.emnlp-main.748

[26] W. T. Hsu, C. K. Lin, M. Y. Lee, K. Min, J. Tang, M. Sun, "A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss", Proceedings of the 56th Annual Meeting of the

Association for Computational Linguistics, Melbourne, Australia, 15-20 July 2018, pp. 132-141. doi: 10.18653/v1/P18-1013

[27] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, "Teaching Machines to Read and Comprehend", Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal, Canada, 7-12 December 2015, Vol. 1, pp. 1693-1701.

[28] A. Fabbri, I. Li, T. She, S. Li, D. Radev, "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, Florence, Italy, 28 July - 2 August 2019, pp. 1074-1084. doi: 10.18653/v1/P19-1102

[29] D. Gholipour Ghalandari, C. Hokamp, N. T. Pham, J. Glover, G. Ifrim, "A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5-10 July 2020, pp. 1302-1308. doi: 10.18653/v1/2020.acl-main.120

[30] Y. Lu, Y. Dong, L. Charlin, "Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16-20 November 2020, pp. 8068-8074. doi: 10.18653/v1/2020.emnlp-main.648

[31] G. Erkan, D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Vol. 22, 2004, pp. 457-479. doi: 10.1613/jair.1523

[32] H. Zheng, M. Lapata, "Sentence Centrality Revisited for Unsupervised Summarization", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July - 2 August 2019, pp. 6236-6247. doi: 10.18653/v1/P19-1628

[33] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out, Proceedings of the ACL-04 Workshop, Barcelona, Spain, 25 July 2004, pp. 74-81.

[34] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, "A Diagnostic Study of Explainability Techniques for Text Classification", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online, 16-20 November 2020, pp. 3256-3274. doi: 10.18653/v1/2020.emnlp-main.263