I2C-UHU-Aries at PoliticHeadlinES-IberLEF 2026: Spanish Cross-Encoder Ensembles and Tail Reranking for Political Headline Ranking

David Hernández Ruiz[1,*], Jacinto Mata Vázquez[1] and Victoria Pachón Álvarez[1]

[1]I2C Research Group, University of Huelva, Spain

Abstract

This paper presents the participation of the I2C-UHU-Aries team in the PoliticHeadlinES 2026 shared task at IberLEF 2026, focused on ranking Spanish political news headlines among semantically related distractors. We explored several ranking strategies, including lexical baselines, semantic rankers, transformer-based cross-encoders, ensemble methods, reranking techniques, and lightweight multimodal fusion using pre-trained vision-language models. The final submitted system was based on a weighted ensemble of two Spanish transformer cross-encoders, BETO and BERTIN, followed by a BETO-based reranker trained with graded relevance labels to refine the lower-ranked candidate headlines while preserving the top-ranked prediction of the ensemble system. Lightweight multimodal fusion with CLIP and SigLIP variants was also evaluated, but it provided only marginal and unstable gains under the explored configurations.

The final system achieved an official nDCG@10 score of 0.8259 in both subtasks, obtaining 8th place in the official leaderboard. The results show that Spanish cross-encoder ensembles were the main source of performance improvement, while tail reranking provided a small but positive refinement of the lower-ranked candidate headlines. In contrast, lightweight multimodal fusion did not produce robust official gains, while LLM-based reranking remained exploratory due to its limited effectiveness and higher computational cost.

Keywords

Headline ranking, Political news, Cross-encoders, Transformer models, Multimodal ranking, Reranking strategies, Spanish NLP

1. Introduction

Headline ranking and selection are increasingly relevant tasks in digital journalism and natural language processing, since headlines must summarize the main information of a news article while remaining concise, coherent, and faithful to the original content [1]. Recent advances in transformer architectures have significantly improved performance in news headline ranking and selection tasks [2]. In addition, multimodal approaches combining textual and visual information have gained growing attention in news understanding scenarios, where associated images may provide complementary contextual signals [3].

PoliticHeadlinES 2026 [4], organized as part of IberLEF 2026 [5], is a shared task focused on ranking Spanish political news headlines according to their relevance to a news article. Unlike traditional headline generation or classification tasks, the competition requires systems to order a fixed set of semantically related candidate headlines for each article, making the problem closer to real-world editorial recommendation scenarios. The competition was divided into two subtasks. Task 1 addressed text-based headline ranking using only the textual content of the article, while Task 2 introduced a multimodal setting in which systems could additionally exploit an image associated with the news article. The dataset, derived from the Spanish PoliSUM-2025 corpus [6], provided news articles, images, and multiple candidate headlines for each instance.

This paper presents the participation of the I2C-UHU-Aries team in the shared task. Our work primarily focused on Spanish transformer-based cross-encoder ensembles and reranking strategies for improving complete headline ranking quality. Additional experiments explored multimodal fusion, neural rerankers, and LLM-based approaches under the PoliticHeadlinES setting. Special attention was

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

david.hernandez1@alu.uhu.es (D. H. Ruiz); mata@dti.uhu.es (J. M. Vázquez); vpachon@dti.uhu.es (V. P. Álvarez)

devoted to improving ranking quality through score fusion and tail-reranking strategies, as well as to evaluating the contribution of visual information in the multimodal setting. The paper describes the implemented systems, the experimental setup, and the main findings obtained during both development phase validation and official evaluation.

The main contributions of this work are as follows:

- We evaluate a broad range of ranking approaches for PoliticHeadlinES 2026, including lexical baselines, semantic rankers, transformer-based cross-encoders, multimodal fusion methods, neural rerankers, and LLM-based approaches.
- Our experiments suggest that ensembles of Spanish cross-encoders, particularly BETO and BERTIN, provide the strongest and most stable performance for political headline ranking in Spanish.
- We introduce a tail-reranking strategy based on graded relevance labels that preserves the top-ranked prediction while refining the ordering of the remaining candidate headlines.
- We analyze the contribution of lightweight multimodal fusion using CLIP and SigLIP variants, showing that visual information provides only marginal and unstable improvements under the evaluated setting.
- We compare development phase validation results with official evaluation outcomes across the main submitted systems.

2. Related Work

Headline selection and ranking are closely related to news summarization and headline generation, but differ in that the system must identify or order candidate headlines rather than generate a new one from scratch. In this setting, the main challenge is to estimate the semantic relevance and factual consistency between a news article and a set of candidate headlines. This is especially important in political news, where small differences in wording may affect framing, factuality, or perceived relevance [1, 7].

Traditional information retrieval methods such as TF-IDF and BM25 have been widely used as lexical ranking baselines [8, 9]. These methods are computationally efficient and often competitive when lexical overlap is informative, but they are limited when candidate headlines are semantically related and share many entities or topical terms. Dense semantic representations based on transformer embeddings provide a stronger alternative by comparing article and headline representations in a shared embedding space, although they may still miss fine-grained interactions between the two texts.

Cross-encoder architectures address this limitation by jointly encoding the article and the candidate headline, allowing the model to capture token-level interactions between both inputs [10, 11]. This makes them particularly suitable for ranking tasks with a small number of highly related candidates, as in PoliticHeadlinES. However, cross-encoders are computationally more expensive than bi-encoder or lexical approaches because each article-headline pair must be processed independently.

Multi-stage ranking pipelines combine an initial ranking model with a second reranking stage that refines the candidate ordering [11]. Such approaches are especially relevant when the evaluation metric considers the complete ranking rather than only the top prediction. In candidate headline ranking tasks, this distinction becomes particularly important because semantically similar distractors may require refining the complete candidate ordering rather than only identifying the best headline.

Multimodal approaches have also been explored in news understanding, where images associated with articles may provide complementary contextual information [3]. Vision-language models such as CLIP and SigLIP can be used to estimate image-text similarity and combine it with textual ranking scores. Nevertheless, the usefulness of this signal depends on the degree to which the image is directly informative for distinguishing between candidate headlines.

Finally, large language models have recently been explored for ranking and reranking through prompting and instruction-following strategies [12]. These models can provide flexible semantic judgments, but their computational cost and sensitivity to prompting may limit their practical use

in shared-task settings. In this work, LLM-based reranking was therefore considered an exploratory alternative rather than part of the final submitted system.

3. Task and Dataset Description

PoliticHeadlinES 2026 is a shared task focused on ranking candidate headlines for Spanish political news articles. Unlike traditional headline generation or classification tasks, the objective is to identify and correctly rank candidate headlines according to their relevance to a news article. The task is especially challenging because many distractor headlines share entities, topics, or political context with the target article, requiring systems to capture subtle semantic differences between highly related candidates.

For each news article, systems are provided with the full article text, an associated image, and a fixed set of ten candidate headlines associated with different relevance levels. Although the task defines a single correct headline as the main target candidate, the official annotations also provide a complete ranking of the candidate headlines, enabling ranking-oriented evaluation and experimentation with graded relevance approaches. Systems must therefore generate a complete ranking of the candidates, ordered from most to least relevant according to their semantic relevance to the article.

The task evaluates the ability of systems to model semantic relevance, topical coherence, headline faithfulness, and multimodal consistency. Since candidate headlines are often highly related to the article content, systems must capture fine-grained semantic differences between competing candidates. In the multimodal setting, relevant visual cues from the associated image may additionally help disambiguate between semantically similar headlines.

The dataset is derived from Spanish political news and is related to the Spanish PoliSUM-2025 corpus [6]. Each instance contains the textual content of a news article, an associated image, and ten candidate headlines. The provided data was divided into development, training, and test partitions. The development partition, which included the gold rankings, was used for development phase validation and intermediate experimentation, while the test partition was reserved for official evaluation through the competition platform.

Unlike standard classification tasks focused only on selecting the best candidate, PoliticHeadlinES evaluates the complete ordering of the ten candidate headlines using an nDCG@10-based metric. In addition, the correct top-ranked headline plays a particularly important role in the final score computation, making accurate top-1 selection essential while still rewarding improvements in the overall ranking quality. This ranking-oriented formulation motivated the exploration of ensemble and reranking strategies designed to improve not only the top-ranked prediction, but also the global consistency of the candidate ordering.

For development phase validation, we report the PA-nDCG scores produced by the official evaluation script for Task 1 and Task 2, as well as their mean across both subtasks. Official leaderboard results were reported separately for each subtask using nDCG@10; when a single summary value is shown, we report the mean of the two official task scores.

Table 1 summarizes the main dataset statistics. Articles in the training partition contained an average of 975.60 tokens, while candidate headlines contained an average of 17.68 tokens. Similar headline lengths were observed across all partitions, whereas article lengths showed higher variability, with some articles exceeding 10,000 tokens. A total of 20,942 associated images were provided across all dataset partitions.

3.1. Task 1: Text-Based Headline Ranking

Task 1 focused exclusively on textual information. Systems were required to rank the candidate headlines according to their relevance to the associated news article using only the textual content of the article.

Table 1
Dataset statistics.

Split	Articles	Candidates/article	Avg. article tokens	Avg. headline tokens
Training	16030	10	975.60	17.68
Development	50	10	718.38	18.89
Test	4912	10	1108.45	17.91

3.2. Task 2: Multimodal Headline Ranking

Task 2 introduced a multimodal setting in which systems could additionally exploit an image associated with the article. This task allowed participants to combine textual and visual information in order to improve headline ranking quality and disambiguate between semantically similar candidate headlines.

4. Methodology

Our methodology explored several ranking strategies for political headline ranking, ranging from lexical baselines to transformer-based cross-encoder ensembles and reranking approaches. The final system focused on weighted score fusion between Spanish cross-encoders and a secondary tail-reranking stage designed to refine the lower-ranked candidate headlines while preserving the top-ranked prediction generated by the ensemble systems.

4.1. Preprocessing

Before applying the ranking models, the textual content of the articles and candidate headlines was normalized through basic whitespace cleaning and formatting normalization. Original casing, punctuation, and diacritics were preserved in all experiments in order to maintain the linguistic information available to the transformer models.

For transformer-based ranking systems, article-headline pairs were independently constructed for each candidate headline associated with a news article. Most transformer configurations used a maximum sequence length of 512 tokens, including both the article and the candidate headline.

Since many news articles exceeded the maximum context length supported by transformer-based architectures, several truncation strategies were explored during development. Most experiments used standard head truncation, preserving the initial portion of the article, which typically contains the most informative journalistic content. Tokenization and truncation were applied using the corresponding tokenizer of each transformer model.

Additional experiments evaluated a head-tail truncation strategy designed to preserve both the beginning and ending sections of long articles under the same token budget. In this configuration, the available token budget for the article was computed dynamically for each article-headline pair after reserving space for the candidate headline and the required special tokens. The remaining article budget was then distributed approximately as 75% for the beginning of the article and 25% for the ending section. Although this strategy occasionally improved development phase validation results, it generalized substantially worse on the official evaluation.

In the multimodal setting, images were processed independently from the textual content using frozen vision-language models. Visual similarity scores were later combined with textual ranking scores through lightweight score fusion.

4.2. Lexical and Semantic Baselines

The first experiments focused on lightweight lexical and semantic ranking approaches in order to establish baseline performance references for the task.

The lexical baselines included TF-IDF and BM25. Both approaches ranked candidate headlines according to their textual similarity with the associated news article. Several BM25 configurations were evaluated using different query token limits and retrieval parameters. The strongest lexical configuration used a maximum query length of 512 tokens, matching the maximum sequence length later adopted by the transformer-based cross-encoders.

A semantic ranking baseline based on sentence embeddings was also evaluated as an intermediate approach between lexical methods and transformer-based rankers. In this configuration, article and headline representations were independently encoded into a shared embedding space, and cosine similarity was used to estimate semantic relevance between both texts.

Although semantic ranking substantially improved over lexical baselines during development phase validation, transformer-based cross-encoders consistently achieved stronger and more stable performance in both development phase validation and official evaluation settings. Nevertheless, lightweight semantic and lexical rankers were later explored as auxiliary reranking components for refining lower-ranked candidate headlines.

4.3. Transformer-Based Cross-Encoders

The main ranking systems explored in this work were based on transformer cross-encoders, which jointly encode the news article and each candidate headline in order to predict a relevance score for every article-headline pair.

Several transformer architectures were evaluated, including BETO [13], BERTIN [14], and mDeBERTa-v3 [15]. BETO was implemented using `dccuchile/bert-base-spanish-wwm-cased`, BERTIN using `bertin-project/bertin-roberta-base-spanish`, and mDeBERTa-v3 using `microsoft/mdeberta-v3-base`. Overall, the Spanish-specific models, especially BETO and the BETO+BERTIN ensemble, achieved the strongest official results among the evaluated transformer-based systems.

The primary training formulation treated the task as a binary classification problem. For each article, ten independent article-headline pairs were generated, assigning label 1 to the correct headline and label 0 to the remaining candidates. The models were trained using cross-entropy loss. During inference, candidate headlines were independently scored using the positive-class score, and the final ranking was obtained by sorting candidates according to this score.

Additional experiments explored a ranking-oriented formulation designed to optimize the complete ordering of the candidate headlines rather than only the top-ranked prediction. In this setting, the BETO cross-encoder previously trained under the binary classification formulation was further fine-tuned as a regression model using graded relevance labels derived from the official candidate ranking. Relevance values were normalized according to the gold ranking position, assigning 1.0 to the highest-ranked headline and 0.1 to the lowest-ranked one. The reranker was trained using mean squared error loss and later applied as a secondary ranking stage for refining the ordering of the lower-ranked candidate headlines generated by the ensemble systems.

This formulation allowed the reranker to model relative ordering across all candidate headlines instead of focusing exclusively on binary correct/incorrect classification.

Most transformer configurations used a maximum sequence length of 512 tokens and standard head truncation. Additional experiments evaluated head-tail truncation strategies, although these configurations generalized substantially worse on the official evaluation despite obtaining strong development phase validation scores.

4.4. Multimodal Fusion with Vision-Language Models

To incorporate visual information in the multimodal setting, several pre-trained vision-language models were evaluated, including CLIP [16], SigLIP [17], and SigLIP2 [18]. CLIP experiments used the `openai/clip-vit-base-patch32` model, while SigLIP-based experiments used lightweight pre-trained SigLIP and SigLIP2 encoders from Hugging Face.

For each candidate headline, image and text representations were independently encoded using the corresponding vision-language model. The resulting embeddings were normalized, and cosine similarity was used to obtain a visual relevance score between the article image and the candidate headline.

Visual similarity scores were combined with textual ranking scores through lightweight weighted score fusion, although the final configuration assigned only a very small weight to the visual component. Different fusion weights were explored depending on the underlying textual ranking model and the selected vision-language encoder.

These multimodal scores were treated as auxiliary signals and combined with the textual ranking models through score-level fusion. The purpose of these experiments was to evaluate whether frozen vision-language encoders could provide complementary information without requiring task-specific multimodal fine-tuning.

4.5. Score Fusion and Ensemble Strategies

Several ensemble configurations were explored by combining predictions from multiple ranking systems. The strongest ensembles were based on weighted score fusion between BETO and BERTIN cross-encoders, which consistently achieved the best complementary performance during experimentation.

Before fusion, the prediction scores produced by each model were independently normalized using min-max normalization in order to reduce scale differences between transformer architectures. Final ensemble scores were then obtained through weighted linear fusion:

$$S(h) = \alpha S_{\text{BETO}}(h) + (1 - \alpha) S_{\text{BERTIN}}(h)$$

where $S(h)$ represents the final relevance score assigned to candidate headline h, and α controls the contribution of each cross-encoder model.

Different fusion weights were evaluated during development phase validation. The final submitted ensemble used fusion weights of 0.7 for BETO and 0.3 for BERTIN. The BETO+BERTIN ensemble improved over the individual submitted transformer systems in the official evaluation and provided one of the strongest development phase validation configurations.

Additional experiments also explored score fusion with semantic rerankers, neural rerankers, and multimodal similarity signals. However, the largest and most stable improvements in our experiments were obtained through ensembles of Spanish transformer-based cross-encoders.

4.6. Exploratory Reranking Approaches

Several additional reranking approaches were explored during experimentation, including instruction-tuned large language models and modern neural rerankers.

Among the LLM-based approaches, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct [19], and Llama-3.1-8B-Instruct [20] were evaluated as zero-shot rankers and rerankers. The models received the news article together with the candidate headlines and were prompted to generate relevance-based rankings without additional fine-tuning. Although some configurations produced small development phase validation improvements when used as rerankers over the strongest ensemble systems, overall performance remained clearly below the transformer-based cross-encoder ensembles while requiring substantially higher computational cost and inference time.

Additional experiments evaluated the `bge-reranker-v2-m3` neural reranker [21]. Although the reranker occasionally provided complementary ranking signals when combined with transformer ensembles, the improvements remained small and inconsistent under the limited development setting.

Due to their limited effectiveness and computational cost, these exploratory reranking approaches were not incorporated into the final submitted systems.

4.7. Tail Reranking

Most transformer-based cross-encoder systems were initially optimized as pointwise binary relevance classifiers, independently scoring each article-headline pair in order to identify the most relevant candidate headline. However, since the official evaluation metric considered the complete ranking of the ten candidates, additional experiments explored secondary reranking stages designed to refine the lower-ranked candidates while preserving the strongest top prediction generated by the ensemble systems.

Several reranking strategies were evaluated using lexical rankers, semantic rankers, neural rerankers, and transformer-based regression models. In these experiments, the highest-ranked headline produced by the ensemble system was kept fixed, while the remaining candidate headlines were reranked through a secondary ranking stage.

The strongest reranking configuration used a BETO-based cross-encoder retrained as a regression model using graded relevance labels derived from the official candidate ranking. Unlike the original binary formulation, the reranker assigned continuous relevance scores ranging from 1.0 to 0.1 according to the gold ranking position, allowing the model to better capture relative ordering relationships between semantically similar candidate headlines.

During inference, the reranker preserved the original top-ranked headline produced by the ensemble system and only reranked candidates originally placed between positions 2 and 10. The final ranking was constructed by concatenating the fixed top-ranked candidate with the reranked candidates from the tail. This procedure refined the lower part of the ranking while avoiding modifications to the most important top-1 prediction. Although the official improvement over the base ensemble was relatively small, this strategy produced the strongest official result among our submitted systems.

5. Experimental Setup

This section describes the main training configurations, validation strategy, and official submitted systems used during the shared task evaluation.

5.1. Training Configuration

Most transformer-based cross-encoder models were trained using the Hugging Face Transformers framework under a supervised learning setup. Training was primarily conducted on NVIDIA A100 GPUs available through Google Colab environments, while additional experiments and inference tests were performed on local RTX 4050 hardware.

Table 2 summarizes the main hyperparameter configurations used for the strongest transformer-based systems.

All transformer-based models were optimized using AdamW with a weight decay of 0.01 and a warmup ratio of 0.1. Model checkpoints were selected according to the best development-set PA-nDCG score obtained during development phase validation.

For binary classification models, each article-headline pair was independently processed and assigned a binary relevance label. During inference, candidate headlines were ranked according to the positive-class score predicted by the model. For the regression-based tail reranker, graded relevance labels derived from the official candidate ranking were used together with mean squared error loss.

Most experiments used a maximum sequence length of 512 tokens with standard head truncation. Additional head-tail truncation experiments were also evaluated, although these configurations generalized substantially worse on the official evaluation despite occasionally producing strong development phase validation scores.

Table 2
Main training configurations for transformer-based systems.

Component	Model	Objective	Loss	Max len	Epochs	LR	Batch
Cross-encoder	BETO	Binary cls	Cross-entropy	512	2	$2 \cdot 10^{-5}$	4
Cross-encoder	BERTIN	Binary cls	Cross-entropy	512	2	$2 \cdot 10^{-5}$	16
Cross-encoder	mDeBERTa-v3	Binary cls	Cross-entropy	512	2	$2 \cdot 10^{-5}$	32
Tail reranker	BETO	Graded regression	MSE	512	1	$1 \cdot 10^{-5}$	16

5.2. Implementation and Reproducibility Details

All experiments were implemented in Python using PyTorch and the Hugging Face Transformers ecosystem. A fixed random seed of 42 was used for data splitting, model training, and evaluation whenever stochastic operations were involved. The semantic baseline used the `sentence-transformers/paraphrase-multilingual-mpnet-base-v2` model without task-specific fine-tuning. For the BM25 baseline, texts were lowercased, accents were removed, and tokens were extracted using an alphanumeric regular expression. The BM25 configuration used $k_1 = 1.75$ and $b = 0.5$, with the article query limited to 512 terms.

The transformer cross-encoders were implemented using the following Hugging Face model identifiers: `dccuchile/bert-base-spanish-wwm-cased` for BETO, `bertin project/bertin-roberta-base-spanish` for BERTIN, and `microsoft/mdeberta-v3-base` for mDeBERTa-v3. The neural reranker used in the exploratory reranking experiments was `BAAI/bge-reranker-v2-m3`. The vision-language models used in the multimodal experiments were `openai/clip-vit-base-patch32`, `google/siglip-base-patch16-224`, and `google/siglip2-base-patch16-224`. In all multimodal configurations, image-text scores were computed using frozen encoders and combined with textual scores through min-max normalization followed by weighted linear fusion.

For the LLM-based experiments, the evaluated instruction-tuned models were `Qwen/Qwen2.5-3B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`, and `meta-llama/Llama-3.1-8B-Instruct`. The models were used in a zero-shot setting with deterministic decoding, disabling sampling and limiting the maximum generation length to 512 new tokens. The article text was truncated to 3500 characters before prompting in order to reduce memory usage and inference time. The prompt instructed the model to assign a relevance score from 0 to 10 to each candidate headline and to return only a valid JSON object containing one score per candidate. The final ranking was obtained by sorting the generated scores in descending order.

The implementation is publicly available at: https://github.com/davidHz-Rz/politicheadlines-tfg. The repository contains the scripts used for the official submissions and additional post-competition experiments, including retraining, evaluation, and error analysis.

5.3. Development Phase Validation and Official Evaluation

Development phase experiments were conducted using the official development partition provided by the organizers, while official evaluation was performed on the hidden test partition through the competition platform.

The development phase was used to compare alternative ranking strategies, tune fusion configurations, and select the main systems submitted to the official evaluation.

5.4. Submitted Runs

Among the official submissions, four configurations were considered the main submitted systems:

- BETO + CLIP multimodal fusion.
- BERTIN + SigLIP multimodal fusion.

- BETO + BERTIN weighted ensemble.
- BETO + BERTIN ensemble with BETO-based tail reranking.

Additional official submissions were used to evaluate baseline and auxiliary configurations, as reported in Table 7.

The strongest official results were obtained by the ensemble-based systems. In particular, the final submitted configuration combined weighted score fusion between BETO and BERTIN with a secondary BETO-based tail reranking stage applied only to lower-ranked candidate headlines.

For Task 2, the final submitted configuration included a minimal SigLIP2-based visual component with a visual fusion weight of 0.02. This setup allowed us to submit a multimodal variant while preserving the ranking behavior of the strongest textual ensemble.

6. Results

6.1. Development Phase Validation Results

Table 3 summarizes representative development phase validation results obtained during experimentation.

Additional experiments evaluated different fusion weights for the BETO+BERTIN ensemble in order to study the balance between both cross-encoder models during development phase validation.

Although more balanced fusion weights occasionally achieved slightly higher development phase validation scores, the differences between the strongest ensemble configurations were very small. In the official submissions, the BETO-based configuration outperformed the BERTIN-based one, in contrast to some development phase validation results that favored BERTIN-based systems. Therefore, a slightly BETO-dominated ensemble (0.7/0.3) was selected for the final submitted runs.

Lexical and semantic baselines obtained substantially lower development phase validation scores than transformer-based cross-encoders. Zero-shot LLM approaches also performed poorly as standalone rankers despite their higher computational cost. In contrast, Spanish transformer cross-encoders and their ensembles produced the strongest and most stable development phase validation results.

Some configurations achieved very high development phase validation scores, particularly BERTIN combined with SigLIP and head-tail truncation variants. However, the official evaluation favored the ensemble-based systems, indicating that these configurations provided more stable performance on the hidden test partition.

6.2. Tail Reranking Analysis

Since the official metric evaluated the complete ranking ordering, additional experiments focused on refining lower-ranked candidate headlines through specialized reranking strategies.

Table 5 summarizes the most relevant tail-reranking experiments performed during development phase validation.

The strongest reranking configuration used a BETO-based cross-encoder trained using graded relevance scores and applied only to the lower-ranked candidates, while preserving the top-ranked headline produced by the ensemble. Lightweight lexical and semantic rerankers also produced measurable improvements, although substantially smaller than the specialized transformer-based reranker.

Additional experiments evaluated head-tail token truncation in order to preserve both the beginning and ending sections of long news articles under the same maximum sequence length. Although this approach slightly improved development phase validation performance, it generalized substantially worse on the official leaderboard, suggesting that the beginning of political news articles contained more robust ranking signals than the article ending.

Table 3
Representative development phase validation results during experimentation.

Approach	Mean PA-nDCG	Notes
TF-IDF + CLIP	0.8266	Lexical baseline
BM25 (512 tokens)	0.8561	Strongest lexical baseline
Semantic ranker	0.8759	Sentence embedding baseline
Qwen2.5-3B-Instruct	0.5248	Zero-shot LLM ranking
Llama-3.1-8B-Instruct	0.5550	Zero-shot LLM ranking
Qwen2.5-7B-Instruct	0.6959	Strongest standalone LLM
BGE-reranker-v2-m3	0.7761	Standalone neural reranker
mDeBERTa-v3	0.8330	Initial multilingual cross-encoder
mDeBERTa-v3 (512)	0.9143	Large-scale training
mDeBERTa-v3 (192)	0.9529	Reduced context length
BETO + CLIP	0.9529	Strong textual baseline
BETO + BERTIN Ensemble	0.9739	Best ensemble configuration
Qwen ensemble reranking	0.9746	LLM-assisted ensemble reranking
BGE ensemble reranking	0.9748	Strong neural reranker fusion
Ensemble + BETO tail reranking	0.9767	Best stable development phase validation result
Head-tail token truncation	0.9768	Improved development phase validation performance
BERTIN + SigLIP	0.9841	Best development multimodal result, lower official performance

Table 4
Development phase validation results for different BETO+BERTIN ensemble fusion weights.

BETO / BERTIN weights	Mean PA-nDCG
0.90 / 0.10	0.9533
0.85 / 0.15	0.9536
0.70 / 0.30	0.9739
0.60 / 0.40	0.9739
0.50 / 0.50	0.9740

Table 5
Comparison of tail-reranking strategies on development phase validation.

Tail reranking strategy	Mean PA-nDCG
None	0.9739
TF-IDF	0.9747
BM25	0.9750
Semantic reranker	0.9753
BGE reranker	0.9750
BETO tail reranker	0.9767

6.3. Multimodal Fusion Results

Several multimodal fusion configurations were evaluated using different combinations of textual cross-encoders and vision-language models.

Although some multimodal configurations produced strong development phase validation results, these gains did not generalize reliably to the official evaluation. In particular, the BERTIN+SigLIP configuration achieved the highest development phase validation score (0.9841 Mean PA-nDCG) but dropped to 0.8022 Mean nDCG@10 on the hidden test partition. Overall, visual information contributed only marginal and inconsistent gains compared to the strongest textual ensemble systems.

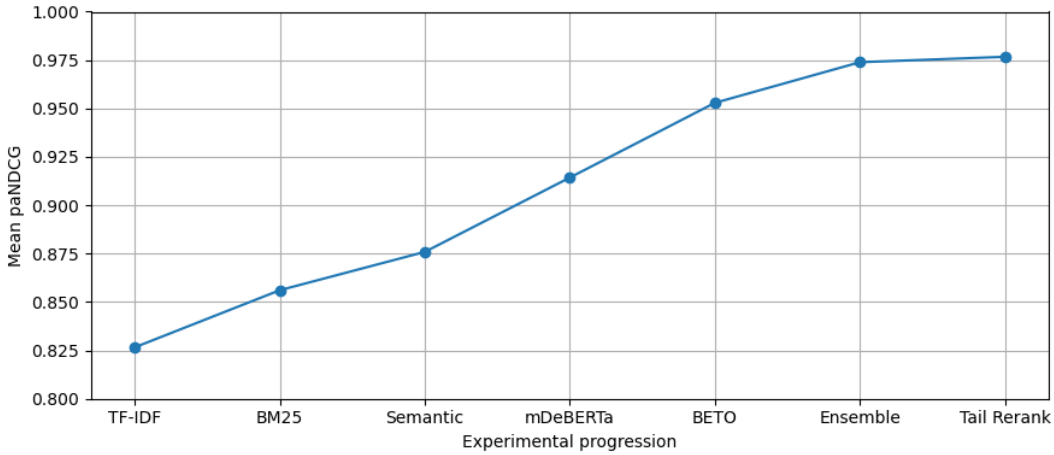

Figure 1: Illustrative summary of representative development phase validation scores across the main experimental stages.

Table 6
Representative multimodal fusion results on development phase validation.

Textual model	Visual model	Weight	Mean PA-nDCG
BETO	CLIP	0.04	0.9529
BETO	SigLIP	0.04	0.9529
BERTIN	SigLIP	0.02	0.9841
mDeBERTa-v3	SigLIP	0.04	0.9529
BETO + BERTIN	SigLIP2	0.02	0.9739

Table 7
Official evaluation results for the submitted systems.

Run	Task 1 text model	Task 2 visual model	Task 2 text/visual weight	T1	T2
Semantic + CLIP	Sentence embeddings	CLIP	0.96/0.04	0.4787	0.4807
BM25 + CLIP	BM25	CLIP	0.96/0.04	0.5613	0.5605
mDeBERTa + SigLIP	mDeBERTa-v3	SigLIP	0.96/0.04	0.7027	0.6998
BETO + CLIP	BETO cross-encoder	CLIP	0.96/0.04	0.8106	0.8110
BETO Head-Tail	BETO head-tail truncation	–	–	0.8101	0.8101
BERTIN + SigLIP	BERTIN cross-encoder	SigLIP	0.98/0.02	0.8017	0.8027
BETO+BERTIN	Weighted ensemble	SigLIP2	0.98/0.02	0.8249	0.8244
Ensemble + BM25	Ensemble + BM25 reranking	SigLIP2	0.98/0.02	0.8230	0.8230
Ensemble + BETO	Ensemble + BETO reranking	SigLIP2	0.98/0.02	0.8259	0.8259

6.4. Official Submission Results

The official competition results for our submitted systems are shown in Table 7. These results correspond to the evaluation performed by the organizers on the hidden test partition.

The strongest official performance was obtained using the BETO+BERTIN ensemble combined with transformer-based tail reranking. The BETO+BERTIN ensemble outperformed the individual submitted cross-encoder systems, while BETO-based tail reranking provided a small additional improvement. In contrast, configurations based on head-tail token truncation obtained lower official scores than the ensemble-based systems despite their strong development phase validation performance. Similarly, the lightweight multimodal fusion strategies evaluated here produced only marginal gains compared to the strongest textual cross-encoder systems.

Table 8 shows an excerpt of the official leaderboard for the ten highest-ranked teams.

Our final system ranked 8th out of 20 participating teams in both subtasks, with an official nDCG@10

Table 8
Excerpt of the official PoliticHeadlinES-IberLEF 2026 leaderboard. Scores correspond to nDCG@10 and are rounded to four decimal places.

Team	Rank	Task 1	Task 2
Theron	1	0.9544	0.9544
LKE-BUAP	2	0.9275	0.9275
VerbaNex AI Lab	3	0.9223	0.9263
acho_gpt	4	0.9026	0.9026
GACOP	5	0.8968	0.8968
HEART-UJA-UA	6	0.8946	0.8913
franrodrigo	7	0.8841	0.8841
I2C-UHU-Aries	8	0.8259	0.8259
LYS-UDC	9	0.8241	0.8082
francordel	10	0.8074	0.8128

score of 0.8259 for Task 1 and Task 2. The best-performing system achieved 0.9544 in both subtasks, while the systems ranked immediately above ours obtained scores between 0.8841 and 0.8946. This indicates that our approach was competitive within the shared task, although there remained a noticeable gap with the highest-ranked systems.

The identical Task 1 and Task 2 scores obtained by our final system are consistent with the minimal visual weight used in the multimodal submission.

7. Discussion

The experiments show that transformer-based cross-encoders were the strongest family of models evaluated in this work. Lexical baselines, semantic rankers, standalone neural rerankers, and zero-shot LLM approaches obtained clearly lower performance, while BETO and BERTIN provided the most useful ranking signals among the evaluated transformer models. These results suggest that Spanish-specific transformer models can be effective for ranking politically related headline candidates, although the comparison is limited to the architectures and configurations explored in this work.

The largest performance improvement came from weighted score fusion between BETO and BERTIN. The ensemble increased the official mean score from 0.8108 for BETO+CLIP and 0.8022 for BERTIN+SigLIP to 0.8247, suggesting that both cross-encoders captured complementary relevance signals. In contrast, the additional gain obtained by tail reranking was relatively small, with the final submitted system reaching 0.8259. This indicates that the ensemble already captured most of the ranking signal, while the BETO-based tail reranker mainly provided a modest refinement of the lower-ranked candidates.

The tail-reranking strategy was nevertheless useful because it was aligned with the structure of the evaluation metric. By preserving the top-ranked prediction of the ensemble and reranking only the remaining candidates, the system avoided disrupting the most important decision while still attempting to improve the lower part of the ranking. This design was motivated by the fact that PA-nDCG gives substantial importance to the highest-ranked positions.

The lightweight multimodal fusion strategies evaluated in this work provided only limited and unstable improvements. Although some combinations of textual cross-encoders with CLIP or SigLIP variants achieved high development phase validation scores, these improvements did not consistently transfer to the official evaluation. The final submitted configuration included only a very small visual fusion weight, and the resulting rankings remained dominated by textual scores. This suggests that simple image–headline similarity using frozen vision-language models was not sufficient to provide robust gains in this task. One possible reason is that the image associated with a news article may not be directly informative enough to distinguish between semantically similar political headlines.

Table 9
Representative examples of failed rankings. Headline fragments are shortened for readability.

Error type	Correct headline	Top predicted headline	Pos.
Entity change	... la abstención del PSOE en Castilla y León	... la abstención del PP en Castilla y León	2
Location change	... acercamiento a China	... acercamiento a Asia	2
Semantic distractor	Xi Jinping ... tercer mandato presidencial	Xi Jinping ... perpetuándose como presidente de China	2
Factual detail	... "¿Margaret Thatcher es tu ídola?"	... "¿Sergio Massa es tu ídola?"	4

7.1. Qualitative Error Analysis

A qualitative inspection of representative errors showed that most mistakes were close misses rather than complete ranking failures. The model usually captured the general topic of the article, but sometimes failed to select the most faithful headline among several highly plausible alternatives.

Table 9 shows representative failed rankings. The most frequent errors involved semantically close headlines differing in small but important factual details, such as entities, locations, quotations, or the specific political action being described. These cases are difficult because the distractor headlines often share most of the political context and lexical structure of the correct headline while changing an essential factual element. These examples suggest that the main errors were not caused by a complete loss of topic, but by the difficulty of distinguishing minimal factual differences between highly related candidate headlines.

8. Conclusion

This paper presented the participation of the I2C-UHU-Aries team in the PoliticHeadlinES 2026 shared task at IberLEF 2026. Several ranking approaches were evaluated, including lexical baselines, semantic rankers, transformer-based cross-encoders, multimodal fusion strategies, ensemble methods, neural rerankers, and LLM-based approaches.

The strongest official results were obtained using a weighted ensemble of BETO and BERTIN cross-encoders combined with a BETO-based tail reranker trained using graded relevance labels. The reranking strategy preserved the top-ranked prediction of the ensemble while refining the ordering of the remaining candidate headlines.

The final submitted system achieved an official nDCG@10 score of 0.8259 in both subtasks, obtaining 8th place in the official leaderboard. Overall, the experiments suggest that Spanish cross-encoder ensembles provided the most robust ranking signals under the evaluated setup, while lightweight multimodal fusion strategies produced limited and unstable gains.

Future work should explore stronger learning-to-rank objectives, task-specific multimodal fine-tuning, and uncertainty-aware reranking strategies for ambiguous candidate sets.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to assist with grammar revision, language refinement, and text improvement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Sepúlveda-Torres, A. Bonet-Jover, E. Saquete, Detecting Misleading Headlines Through the Automatic Recognition of Contradiction in Spanish, IEEE Access 11 (2023) 72007–72026. doi:10.1109/ACCESS.2023.3295781.

[2] P. Voropaev, O. Sopilnyak, Transformers for Headline Selection for Russian News Clusters, arXiv preprint arXiv:2106.10487 (2021). arXiv:2106.10487.

[3] M. Krubiński, P. Pecina, Towards Unified Uni- and Multi-modal News Headline Generation, in: Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 437–450. doi:10.18653/v1/2024.findings-eacl.30.

[4] J. Gómez-Navalón, T. Bernal-Beltrán, R. Pan, F. García-Sánchez, J. A. García-Díaz, R. Valencia-García, Overview of PoliticHeadlinES at IberLEF 2026: Multimodal Headline Ranking in Spanish Political News, Procesamiento del Lenguaje Natural 77 (2026).

[5] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[6] R. Pan, T. Bernal-Beltrán, J. Gómez-Navalón, J. A. García-Díaz, R. Valencia-García, Spanish PoliSUM-2025: Evaluating Stylistic and Editorial Fidelity in Abstractive Summarization of Political News, Procesamiento del Lenguaje Natural 76 (2026) 177–189.

[7] R. Liu, C. Jia, S. Vosoughi, A Transformer-Based Framework for Neutralizing and Reversing the Political Polarity of News Articles, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–26. doi:10.1145/3449139.

[8] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1986.

[9] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: Proceedings of the Third Text REtrieval Conference (TREC-3), volume 500-225 of *NIST Special Publication*, 1995, pp. 109–126.

[10] R. Nogueira, K. Cho, Passage Re-ranking with BERT, arXiv preprint arXiv:1901.04085 (2019). arXiv:1901.04085.

[11] R. Nogueira, W. Yang, K. Cho, J. Lin, Multi-Stage Document Ranking with BERT, arXiv preprint arXiv:1910.14424 (2019). arXiv:1910.14424.

[12] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, M. Bendersky, Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting, in: Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1504–1518. doi:10.18653/v1/2024.findings-naacl.97.

[13] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: PML4DC at ICLR 2020, 2020.

[14] J. de la Rosa, E. G. Ponferrada, P. Villegas, P. González de Prado Salas, M. Romero, M. Grandury, BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. doi:10.26342/2022-68-1.

[15] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, arXiv preprint arXiv:2111.09543 (2021). arXiv:2111.09543.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[17] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid Loss for Language Image Pre-Training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11975–11986.

[18] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, X. Zhai, SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, arXiv preprint arXiv:2502.14786 (2025). arXiv:2502.14786.

[19] Qwen Team, Qwen2.5 Technical Report, arXiv preprint arXiv:2412.15115 (2024).

arXiv:2412.15115.

[20] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783 (2024). arXiv:2407.21783.

[21] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-Pack: Packaged Resources To Advance General Chinese Embedding, arXiv preprint arXiv:2309.07597 (2023). arXiv:2309.07597.