GACOP at PoliticHeadlinES-IberLEF 2026: Listwise Text Ensembles and Agreement-Gated Multimodal Fusion for Spanish Political Headline Reranking

Jesús García-Salmerón[1,*], Pilar González-Férez[1,†] and Gregorio Bernabé[1,†]

[1]*Computer Engineering Department, University of Murcia, Campus de Espinardo, Murcia, 30100, Spain*

Abstract

Political news headlines play a central role in shaping readers' interpretation of public events, making accurate headline selection especially relevant in scenarios where several plausible candidates are available. This work describes our participation in the PoliticHeadlinES-IberLEF 2026 shared task on Spanish political headline reranking. The challenge addresses two tasks: a text-based track, where methods must rank candidate headlines using only the article content, and a multimodal track, where the associated news image is also available. For the text-only task, we propose a listwise reranking framework based on an ensemble of encoder-only models, combining RigoBERTa 2.0, MrBERT-es, and RoBERTa. Each model family is fine-tuned using two different random seeds, and their outputs are internally aggregated before applying a final weighted ensemble. For the multimodal task, we extend this textual ensemble with a SigLIP-based vision-language model that estimates image–headline compatibility. To avoid the negative impact of weakly grounded or generic political news images, we introduce an agreement-gated fusion strategy that incorporates visual evidence only when the text-only ensemble is uncertain and both branches agree on the top-ranked candidate. Our approach achieved an official nDCG@10 of 89.677% in both the text-based and multimodal tasks. These results show that listwise textual ensembles provide a strong and robust signal for Spanish political headline reranking, while multimodal information should be integrated cautiously in this domain.

Keywords

Headlines Reranking, Political News, Multimodal Retrieval, Learning-to-Rank, Natural Language Processing

1. Introduction

News headlines may be understood as representations of journalistic content. They capture the main event, highlight the most relevant actors, and shape the reader's initial interpretation of the story. In political news, such compression is especially consequential, as subtle lexical choices can influence how readers perceive events. Consequently, ranking candidate headlines for a given article goes beyond measuring surface-level similarity. An effective system must determine which candidate is not only thematically related to the article, but also supported by its content and, when available, consistent with the main image.

PoliticHeadlinES 2026 [1], organized within IberLEF 2026 [2], addresses this problem as a shared task on Spanish political news headline reranking. It includes two tracks: a text-only task and a multimodal task where the article image can support headline selection [1]. The training set contains 16,030 articles, each paired with an image and ten candidate headlines. These candidates include real distractors and hard negatives generated using semantic similarity, entity overlap, and temporal proximity [3]. As a result, this challenge requires ranking the headlines that best support the article, rather than simply identifying topical similarity.

The main difficulty is that many candidate headlines initially appear reasonable. Some may refer to the right political party but the wrong minister, the right institution but a different vote, or a related event from the same day. Lexical matching helps capture named entities, dates, acronyms, and institutional

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

jesus.garcias@um.es (J. García-Salmerón); pilargf@um.es (P. González-Férez); gbernabe@um.es (G. Bernabé)

0009-0007-4226-2198 (J. García-Salmerón); 0000-0003-1681-5442 (P. González-Férez); 0000-0002-7265-3508 (G. Bernabé)

terms, but it fails when candidates are paraphrased or when distractors share most entities with the article. Dense semantic encoders can reduce this limitation, although they may blur fine-grained distinctions that are important in political discourse. Multimodal evidence introduces an additional difficulty, since images can be useful when they show a specific politician, protest, parliamentary session, or press conference, but they may also be generic, reused, or only weakly connected to the article.

This work presents a multimodal method for the PoliticHeadlinES [1] challenge. Our proposed approach follows a retrieval-based architecture: each article–candidate pair is represented through several compatibility scores, and candidates are ranked according to a final fused score. For the text-only setting, we combine dense encoder-only models, including Spanish-specific and multilingual architectures such as RigoBERTa 2.0 [4], MrBERT-es [5], and RoBERTa [6], using model-specific sequence lengths. For the multimodal setting, we estimate image–headline compatibility with a vision-language model (VLM) and apply a gated fusion strategy to reduce the impact of noisy or weakly related images. The framework is designed to be robust, interpretable, and modular, allowing individual components to be ablated and fusion weights to be tuned on a validation split without retraining the encoders.

2. Background

2.1. Headline generation, selection and political news understanding

Automatic headline generation has often been studied as a form of extreme summarization, where the model produces a short title from the article body. Recent approaches typically use sequence-to-sequence transformers and evaluate fluency, informativeness, and factuality [7]. PoliticHeadlinES differs from generation: the system is not asked to produce a new headline, but to rank a closed set of candidate headlines. This framing reduces open-ended generation risk while emphasizing discriminative semantic matching. The challenge resembles answer selection, passage retrieval, and learning-to-rank settings, because each article is a query and the ten candidate headlines are the items to be ranked.

Political headline analysis has also been addressed from the perspective of sentiment, framing, stance, and bias. Work on Spanish political headlines has shown that headlines can influence judgment and that target information is useful for modelling political meaning [8]. In this challenge, target awareness is essential even when the objective is not sentiment classification. A distractor may be semantically close because it contains the same candidate, party, or institution, but it can still describe a different action or outcome. This means that approaches must preserve actor-event-role relations rather than relying only on bag-of-words similarity.

2.2. Learning to rank and retrieval metrics

The challenge naturally fits the learning-to-rank paradigm. Learning-to-rank methods are commonly grouped into pointwise, pairwise, and listwise approaches [9]. In a pointwise setting, each article-headline pair receives an independent relevance score. In a pairwise setting, the model learns preferences between candidates. In a listwise setting, the complete ranking is optimized directly.

Ranking quality is better measured with rank-sensitive metrics than with standard classification losses. Mean Reciprocal Rank (MRR) measures the inverse rank of the first correct candidate, making it suitable when a single headline is considered the best answer. Discounted Cumulative Gain (DCG) and normalized DCG (nDCG) assign higher value to relevant candidates placed near the top of the ranking while penalizing lower positions [10]. Although the official task metric is defined by the organizers, using Top-1 accuracy, MRR, and nDCG can offer a more informative assessment of the model's performance.

2.3. Sparse and dense textual representations

Sparse retrieval methods remain strong baselines for news ranking because headlines and articles share named entities, dates, numbers, and institutional terminology. TF-IDF weighting assigns higher

importance to terms that are frequent in a document but less frequent across the corpus, making it useful for detecting distinctive political references [11]. BM25 extends this family of probabilistic retrieval models and remains a standard strong baseline in information retrieval [12].

Dense sentence representations complement sparse retrieval by capturing paraphrases and broader semantic equivalence. Sentence-BERT introduced siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared efficiently with cosine similarity [13]. For Spanish, language-specific encoders are important because political text includes morphology, syntax, idioms, and institutional names that may be underrepresented in English-centric models. BETO was introduced as a BERT-based model pre-trained exclusively on Spanish data [14]. In addition, RigoBERTa 2.0 stands out as a Spanish encoder obtained by further pre-training XLM-RoBERTa-large, a multilingual RoBERTa-based model, on a meticulously curated Spanish corpus [6, 4]. More recently, MrBERT-es has been released as a Spanish-English foundational encoder derived from the MrBERT family and based on the ModernBERT architecture [5]. Because ModernBERT was designed with native long-context support and efficient inference [15], MrBERT-es is well suited to ranking tasks where the full article context may exceed the token limit of older encoders.

2.4. Vision-language models for multimodal ranking

The multimodal task requires a way to compare an image with textual candidates. CLIP demonstrated that image and text encoders trained with natural language supervision on large image-text pairs can learn a shared representation space and transfer to downstream tasks [16]. SigLIP later replaced the softmax contrastive loss with a sigmoid pairwise loss, improving scalability and removing the requirement that each loss term depend on all examples in a global batch [17]. Multilingual and cross-lingual extensions of CLIP, including AltCLIP, aim to expand vision-language retrieval beyond English by replacing or adapting the text encoder [18]. These models are attractive for the challenge because they provide zero-shot or weakly supervised image-headline compatibility scores. However, political news images are often weakly connected to the specific event, since a stock photograph of a parliament building may not distinguish between two different votes, and a politician portrait may not identify the specific legislative action.

3. Methodology

3.1. Text-only module

For Task 1, we design a text-only reranking framework that estimates the compatibility between the news article and each candidate headline. Given an input article and a set of candidate headlines, our module assigns a relevance score to each article–headline pair and returns the final ranking of the candidates. Figure 1 presents our text-only reranking module. The proposed approach is based on an ensemble of three encoder-only language models: RigoBERTa 2.0, MrBERT-es, and RoBERTa. Each model family is trained using two independent versions, employing different random seeds for weights and training-validation splits. This seed-level redundancy reduces the dependence on a single training run and improves the robustness of the final ranking. As shown in the diagram, each version follows the same architecture, consisting of an encoder model, followed by a dropout layer and a linear scoring head.

Formally, for an article a and a set of K candidate headlines,

$$H = \{h_1, h_2, \ldots, h_K\}$$

each candidate pair (a, h_i) is tokenized and passed through the corresponding encoder. The representation of the first token, equivalent to the $[CLS]$ representation for these encoder-only models, is used as the global representation of the article–headline pair. This representation is then regularized with dropout ($p = 0.2$) and projected to a scalar score through a linear layer.

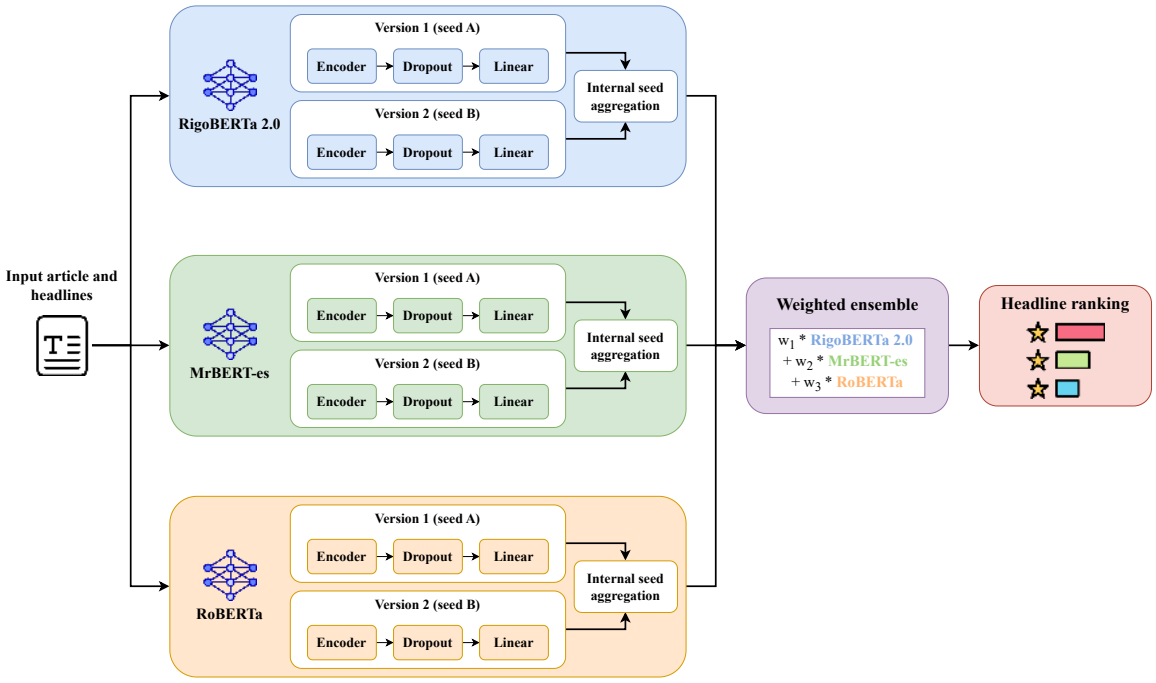

Figure 1: Overview of the text-only reranking module. Each article–headline pair is processed by three encoder-only model families, each trained with two different random seeds. The outputs of each version are internally aggregated and subsequently fused through a weighted ensemble to obtain the final ranking of candidate headlines.

Unlike a binary pairwise classifier, the model is trained in a listwise ranking setting. For each article, the model jointly scores all candidate headlines and produces a score vector. During training, the correct headline, corresponding to the first-ranked headline in the annotated ranking, is used as the target class. The model is optimized using Cross-Entropy (CE) over the complete list of candidates:

$$\mathcal{L}_{\mathcal{CE}} = -\log \frac{\exp(s_y)}{\sum_{j=1}^{K} \exp(s_j)}$$

where y is the index of the correct headline. This formulation encourages the model not only to identify whether a headline is compatible with the article, but also to assign it a higher score than the remaining candidates.

At inference time, each seed-specific model produces a score vector for the same set of candidate headlines. The two versions belonging to the same model family are first combined through an internal seed aggregation step. In our case, this aggregation is implemented as an equally weighted average of the two seed outputs:

$$\mathbf{s}_m = w_a \cdot \mathbf{s}_m^{(A)} + w_b \cdot \mathbf{s}_m^{(B)}$$

where m denotes the model family, $w_a = w_b = 0.5$ are the weights for each seed-specific model, and $\mathbf{s}_m^{(A)}$ and $\mathbf{s}_m^{(B)}$ are the score vectors produced by the two seed-specific versions.

After this internal aggregation, the outputs of the three model families are first normalized using z-score normalization to reduce scale differences across models. The normalized scores are then combined using a weighted ensemble:

$$\mathbf{s}_{\text{final}} = w_1 \cdot \mathbf{s}_{\text{RigoBERTa2.0}} + w_2 \cdot \mathbf{s}_{\text{MrBERT-es}} + w_3 \cdot \mathbf{s}_{\text{RoBERTa}}$$

In our final configuration, the ensemble weights are set to $w_1 = 0.45$, $w_2 = 0.45$, and $w_3 = 0.10$ for RigoBERTa 2.0, MrBERT-es, and RoBERTa, respectively. These values are selected according to the

validation performance of the individual model families and their contribution to the ensemble. In particular, RigoBERTa 2.0 and MrBERT-es receive higher weights due to their stronger validation results, whereas RoBERTa is assigned a lower weight while still contributing complementary information. Finally, the candidate headlines are sorted in descending order according to s_{final}, producing the final headline ranking.

3.2. Multimodal module

Figure 2 shows the multimodal reranking approach proposed for Task 2. Our method builds on the previously described text-only module and incorporates a pre-trained VLM, specifically SigLIP, to rank the candidate headlines according to their compatibility with the main news image. In this framework, the text-only ranking remains the primary signal, while the VLM ranking is used as a complementary source of visual evidence. Accordingly, we perform a weighted ensemble of the text-only and VLM ranking scores only when specific gating conditions are satisfied.

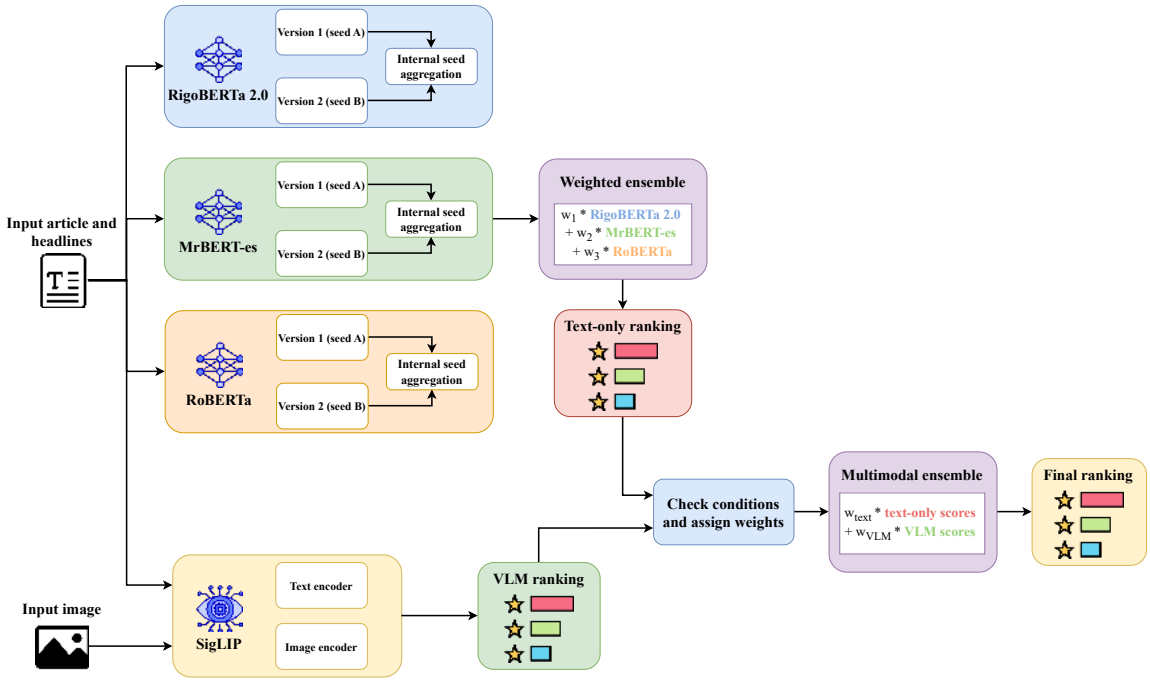

Figure 2: Overview of the multimodal reranking framework. The text-only ensemble module produces a ranking of the candidate headlines, while SigLIP estimates image–headline compatibility scores from the associated news image.

First, we compute the text-only ranking scores (see Section 3.1) and the VLM-based ranking scores in parallel. Specifically, we use a multilingual SigLIP model to handle Spanish headlines, which may also contain English terms. For each candidate headline, we estimate its image–headline compatibility score and then apply z-score normalization to the resulting ranking scores.

Therefore, our approach produces two complementary sets of scores: one from the text-only ensemble and another from the image–headline compatibility estimated by the VLM. However, we observed that directly combining both branches with a fixed weighted average degraded performance. This is mainly because the visual content of political news is often weakly grounded in the specific event described by the headline.

To address this limitation, we adopt a conservative agreement-gated fusion strategy. Instead of always incorporating the VLM scores, the visual branch is only used when the text-only ensemble is uncertain and the VLM agrees with it on the top-ranked headline. In this way, the text-only ensemble

Algorithm 1 Agreement-gated weighting strategy

Require: Normalized text scores $\mathbf{s}^{text}$ and normalized VLM scores $\mathbf{s}^{VLM}$
Require: Gate-gap threshold gap and initial VLM weight $init_{VLM}$
$\Delta_{text} \leftarrow s^{text}_{(1)} - s^{text}_{(2)}$
$i^{text}_1 \leftarrow \text{top}_1(\mathbf{s}^{text})$
$i^{VLM}_1 \leftarrow \text{top}_1(\mathbf{s}^{VLM})$
if $\Delta_{text} \geq gap$ **then**
  $w_{VLM} \leftarrow 0$ ▷ text-only is confident
else if $i^{text}_1 = i^{VLM}_1$ **then**
  $w_{VLM} \leftarrow init_{VLM}$ ▷ text-only and VLM agree
else
  $w_{VLM} \leftarrow 0$ ▷ text-only and VLM disagree
end if
$w_{text} \leftarrow 1 - w_{VLM}$
$\mathbf{s}^{final} \leftarrow w_{text} \cdot \mathbf{s}^{text} + w_{VLM} \cdot \mathbf{s}^{VLM}$

remains the main decision signal, while the VLM is used only as a supporting source of visual evidence. Algorithm 1 summarizes the logic used to assign the weights for the text-only and VLM ranking scores.

The gate-gap threshold controls how confident the text-only ensemble must be before ignoring the visual signal. We define textual confidence as the score margin between the first and second-ranked candidates. When this margin is large, the text branch is assumed to have enough evidence to resolve the ranking, and the VLM weight is set to zero. When the margin is small, the model considers the ranking ambiguous. However, the VLM is only allowed to contribute if it agrees with the text branch on the top candidate. This makes a fusion precision-oriented, where visual evidence cannot override the textual decision, but it can reinforce it and affect the ordering of lower-ranked candidates.

Finally, the headlines are sorted in descending order according to the fused score vector $\mathbf{s}^{final}$, producing the final multimodal ranking. If the image associated with the given article is not available, our approach directly uses the text-only ranking as a fallback, ensuring that a prediction can still be generated for every sample.

4. Experimental Setup

All experiments are implemented in Python 3.11.9 using Transformers 5.6.0.dev0 and Accelerate 1.13.0. The developed code can be found in our GitHub repository (https://github.com/jesussgs/politic_headlines_2026).

4.1. Model configuration

Table 1 summarizes the models and configurations employed in our final submission. For Task 1, we employ three encoder-only language model families: RigoBERTa 2.0, MrBERT-es, and RoBERTa. Most text encoders are trained and evaluated with a maximum sequence length of 512 tokens. However, MrBERT-es is configured with a maximum sequence length of 1024 tokens, since this model supports longer input contexts and can therefore exploit more information from the article body. During tokenization, the article is used as the first sequence and the candidate headline as the second sequence. For Task 2, the text-only ensemble is complemented with a multilingual SigLIP model. This pre-trained VLM is used only at inference time to estimate the compatibility between the main news image and each candidate headline.

Model	Checkpoint	Task	Max. length
RigoBERTa 2.0	`IIC/RigoBERTa-2.0`	Text reranking	512
MrBERT-es	`BSC-LT/MrBERT-es`	Text reranking	1024
RoBERTa	`FacebookAI/xlm-roberta-large`	Text reranking	512
SigLIP	`google/siglip-base-patch16-256-multilingual`	VLM scoring	MAX_LENGTH

Table 1
Configuration of the models used in the proposed framework. Note that MAX_LENGTH refers to the VLM text encoder maximum length.

4.2. Training and inference configuration

For training, the official training data is split into training and validation subsets using an 80/20 split. The validation subset is used to evaluate model performance and select the best-performing checkpoints according to Top-1 accuracy. To reduce the variance associated with a single training run, each text model family is fine-tuned twice using two different random seeds, denoted as seed A and seed B. These seeds affect both the model initialization and the training-validation split.

As described in Section 3.1, the text-only models are trained under a listwise ranking formulation using Cross-Entropy Loss, where the target class corresponds to the correct headline among the candidate set. The main hyperparameters used for training are reported in Table 2.

Hyperparameter	Value
Train/validation split	80/20
Seed A	0
Seed B	42
Loss function	Cross-Entropy Loss
Dropout	0.2
Train batch size	2
Eval batch size	2
Learning rate	$5e^{-6}$
Weight decay	$5e^{-4}$
Optimizer	AdamW
Number of epochs	5

Table 2
Training hyperparameters for the text-only rerankers. Note that batch size values are per device.

At inference time, all models are loaded in evaluation mode and executed using `bfloat16` precision when GPU acceleration is available. Each model produces a score vector over the candidate headlines. The text-only and VLM score vectors are normalized using z-score normalization before the multimodal fusion stage. To determine the final fusion configuration, we tested different values for the agreement-gated strategy, considering $gate\_gap \in \{0.10, 0.15, 0.20, 0.25\}$ and $init_{VLM} \in \{0.10, 0.20, 0.30, 0.40\}$. In the official PoliticHeadlinES 2026 submission, we selected $gate\_gap = 0.15$ and $init_{VLM} = 0.40$ as this configuration achieved the best results.

5. Results

Table 3 shows the official results of the PoliticHeadlinES 2026 challenge, ordered by the Task 2 nDCG@10 score. Since the official ranking of the challenge is mainly determined by the multimodal task, Task 2 is the most relevant setting for comparing the submitted methods. Our approach, submitted under the name GACOP, achieved an nDCG@10 of 89.677% in both tasks, obtaining fifth place in the official

ranking.

Name	Task 1 - Text-based (nDCG@10)	Task 2 - Multimodal (nDCG@10)
Theron	95.436	95.436
LKE-BUAP	92.751	92.751
VerbaNex AI Lab	92.628	92.227
acho_gpt	90.264	90.264
GACOP	**89.677**	**89.677**
HEART-UJA-UA	89.126	89.458
...	...	...
baseline	51.344	51.653
MIMIC	34.796	34.796

Table 3
Official reranking results for PoliticHeadlinES 2026. The table reports the nDCG@10 (%) obtained for the text-based and multimodal tasks.

The obtained results show that our text-only ensemble provides a strong and competitive ranking signal. In Task 1, we reach an nDCG@10 of 89.677%, outperforming the official baseline by a considerable margin. This result supports the effectiveness of the proposed listwise reranking formulation and the use of complementary encoder-only language models.

For Task 2, our method achieves the same score as in Task 1. Although the multimodal approach does not lead to an improvement in the official score, this result remains meaningful, as the proposed agreement-gated fusion strategy is designed to reduce the risk of performance degradation that may occur when textual and visual scores are combined without control. Therefore, these results suggest that the text-only signal is the most reliable component for this task, while the visual signal should be incorporated carefully.

5.1. Limitations

This study has several limitations that help contextualize the results. The first limitation comes from the fine-grained nature of political headline reranking. Several candidate headlines may mention the same political actors, parties, institutions, or temporal context. However, they may describe different actions, decisions, or outcomes. These cases challenge the system because semantically close distractors can receive high compatibility scores, even when the article does not fully support them. As a result, the system may fail to capture subtle actor-event-role distinctions that require more precise factual grounding.

The second limitation concerns the multimodal branch. Images can provide useful evidence in some cases, but political news images often remain generic, archival, or weakly connected to the specific event. For example, images of politicians, parliamentary buildings, press conferences, or institutional scenes may fit several candidate headlines. This reduces the discriminative value of the VLM. It also helps explain why the multimodal task does not improve over the text-only score in the official results.

These limitations motivate the conservative design of the agreement-gated fusion strategy. The visual branch does not override the textual ranking. Instead, it only acts as supporting evidence when the text-only ensemble shows uncertainty and both modalities agree on the top-ranked candidate.

6. Conclusions and Future Work

In this work, we present our approach for the PoliticHeadlinES 2026 challenge on Spanish political headline reranking. For the text-only task, we propose a listwise ensemble composed of three encoder-only model families, namely RigoBERTa 2.0, MrBERT-es, and RoBERTa. Each model family is trained with two different random seeds, and their outputs are internally aggregated before applying a final

weighted ensemble. This design reduces the dependence on a single fine-tuning run and allows the method to combine complementary Spanish-specific and multilingual textual representations.

For the multimodal task, we extend the text-only ensemble with a SigLIP-based image–headline scoring branch. However, instead of directly combining textual and visual scores, we introduce an agreement-gated fusion strategy. This mechanism gives priority to the text-only ranking and incorporates the VLM signal only when the text-only ensemble is uncertain and both branches agree on the top-ranked headline. The official results show that our approach achieves an nDCG@10 of 89.677% in both tasks, ranking fifth on the official leaderboard. These results confirm the competitiveness of the proposed text-only reranking ensemble and suggest that, in this challenge, textual evidence is the most reliable signal for identifying the correct headline.

As future work, we plan to explore more fine-grained multimodal strategies by fine-tuning VLMs on Spanish political news and incorporating event-aware representations to better distinguish similar headlines. Finally, we plan to fine-tune and evaluate established reranking models, such as Qwen3-Reranker [19] and Jina Reranker v3 [20], to determine whether task-specific reranking methods can further improve headline selection.

Acknowledgments

This work has been partially funded by Grant TED2021-129221B-I00 funded by MCIN/AEI/10.13039/501100011033, and by the "European Union NextGenerationEU/PRTR". This work has been also partly supported by Grant PID2022-136315OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", EU.

Declaration on Generative AI

During the preparation of this manuscript, DeepL was employed to improve language accuracy and correct typographical errors. The authors subsequently revised the text and assume full responsibility for the final content of the publication.

References

[1] J. Gómez-Navalón, T. Bernal-Beltrán, R. Pan, G.-S. Francisco, 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).

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: 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.

[3] 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.

[4] Instituto de Ingeniería del Conocimiento, Rigoberta 2.0, https://huggingface.co/IIC/RigoBERTa-2.0, 2025. Hugging Face model card.

[5] D. Tamayo, I. Lacunza, P. Rivera-Hidalgo, S. Da Dalt, J. Aula-Blasco, A. Gonzalez-Agirre, M. Villegas, MrBERT: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation, 2026. doi:10.48550/arXiv.2602.21379.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[7] Z. Li, J. Wu, J. Miao, X. Yu, News headline generation based on improved decoder from transformer, Scientific Reports 12 (2022) 11648. doi:10.1038/s41598-022-15817-z.

[8] T. A. Salgueiro, E. Recart Zapata, D. Furman, J. M. Pérez, P. N. Fernández Larrosa, A spanish dataset for targeted sentiment analysis of political headlines, 2022. doi:10.48550/arXiv.2208.13947.

[9] T.-Y. Liu, Learning to rank for information retrieval, Foundations and Trends in Information Retrieval 3 (2009) 225–331. doi:10.1561/1500000016.

[10] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems 20 (2002) 422–446. doi:10.1145/582415.582418.

[11] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[12] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

[13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[14] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: Proceedings of the Practical Machine Learning for Developing Countries Workshop at ICLR 2020, 2020. doi:10.48550/arXiv.2308.02976.

[15] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 2526–2547. doi:10.18653/v1/2025.acl-long.127.

[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. URL: https://proceedings.mlr.press/v139/radford21a.html.

[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. doi:10.1109/ICCV51070.2023.01100.

[18] Z. Chen, G. Liu, B.-W. Zhang, Q. Yang, L. Wu, AltCLIP: Altering the language encoder in CLIP for extended language capabilities, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 8666–8682. doi:10.18653/v1/2023.findings-acl.552.

[19] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025). doi:10.48550/arXiv.2506.05176.

[20] F. Wang, Y. Li, H. Xiao, Jina-reranker-v3: Last but not late interaction for listwise document reranking, arXiv preprint arXiv:2509.25085 (2025). doi:10.48550/arXiv.2509.25085.