AchoGPT at PoliticHeadlinES-IberLEF 2026: Weighted MrBERT Ensembles for Spanish Political Headline Ranking

Alberto Sevilla-Valiente[1,*,†], Claudio Gea-López[1,†] and Javier Hernández-Tercero[1]

[1]*Universidad de Murcia, Spain*

Abstract

This paper describes our system submitted to PoliticHeadlinES-IberLEF 2026, a shared task on ranking candidate headlines for Spanish political news articles. Given a news article and ten candidate headlines, our system estimates the relevance of each article–headline pair and outputs a complete ranking of the candidates. We developed a text-only approach based on MrBERT-es, a bilingual Spanish–English encoder derived from the ModernBERT family. The final submission combines four fine-tuned variants through weighted soft voting: a binary top-1 classifier, a long-context regression/ranking model, a pairwise hard-negative model, and a listwise model trained over full candidate sets. All models encode each article with one candidate headline, truncating only the article body in order to preserve the headline text. The final ensemble weights were selected on the development split by grid search using the official PA-nDCG metric, assigning most of the weight to the long-context MrBERT model while retaining complementary signals from hard-negative and listwise training. Although PoliticHeadlinES includes a multimodal setting, our submitted system relies exclusively on textual information and reuses the same ranking for both official output columns.

Keywords

Headline ranking, Spanish political news, learning to rank, MrBERT, transformer ensembles, IberLEF

1. Introduction

Political news headlines play a central role in shaping how readers interpret public events. In online media, headlines must summarize complex political situations in a short form, often referring to actors, institutions, temporal context, and competing narratives. Automatically selecting the best headline for a political article is therefore not merely a surface-level semantic similarity problem: plausible candidate headlines may mention the same politicians or institutions while describing a different event, stance, or temporal frame.

PoliticHeadlinES-IberLEF 2026 addresses this challenge as a ranking task for Spanish political news [1] [2]. Each instance contains one article and ten candidate headlines. Systems are required to return a complete ranking of the candidates, with special importance assigned to placing the correct headline in the first position. The task includes a text-only track and a multimodal track in which images can be exploited to disambiguate similar political events [3]. In this work we focus on a robust text-only system and submit the same prediction for both official output fields.

Our approach is based on the intuition that different training objectives capture different aspects of the task, a common motivation behind ensemble learning [4]. A pointwise regression model can exploit graded supervision from the full gold ranking, a binary classifier can focus directly on the top-1 decision emphasized by the metric, a pairwise model can learn to separate the correct headline from highly competitive distractors, and a listwise model can optimize scores for the whole set of ten candidates jointly. Rather than selecting a single objective, we fine-tuned several MrBERT-es variants and combined their probability distributions through a weighted ensemble [5].

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

alberto.sevillav@um.es (A. Sevilla-Valiente); c.gealopez@um.es (C. Gea-López); javier.hernandezt@um.es (J. Hernández-Tercero)

0009-0003-7696-6434 (A. Sevilla-Valiente); 0009-0008-5824-790X (C. Gea-López); 0009-0000-2952-4471 (J. Hernández-Tercero)

The main contributions of this system paper are:

- a long-context transformer-based architecture for ranking Spanish political headlines;
- a comparison of pointwise, binary, pairwise, and listwise training formulations over the same base encoder;
- a hard-negative training strategy in which a teacher model selects the most competitive incorrect headline for each article;
- a lightweight weighted soft-voting ensemble optimized directly with the official PA-nDCG metric.

2. Task Description

The PoliticHeadlinES shared task evaluates systems for headline ranking in Spanish political news. Each example is composed of an article body, ten candidate headlines denoted as $t_1, \ldots, t_{10}$, and a gold ranking indicating the preferred ordering of those headlines. The expected submission contains an identifier and two ranking columns, `task_1` and `task_2`. In our scripts, both columns are filled with the same ranking because the submitted system does not use image features.

The official evaluation metric is PA-nDCG, a position-aware variant of nDCG [6] that strongly rewards identifying the correct top headline. Let p be the predicted ranking and g the gold ranking. If the first predicted headline does not match the first gold headline, the score is zero. Otherwise, the score assigns a fixed reward α and uses nDCG over the remaining candidates:

$$\text{PA-nDCG}(p, g) = \begin{cases} 0, & p_1 \neq g_1, \\ \alpha + (1 - \alpha) \cdot \text{nDCG}(p_{\setminus g_1}, g_{\setminus g_1}), & p_1 = g_1. \end{cases} \tag{1}$$

Following the implementation used in the baseline, we set $\alpha = 0.9$ and evaluate rankings over the ten candidates. This metric makes the task highly sensitive to the top-1 decision while still preserving a smaller contribution from the relative order of the remaining headlines.

3. System Overview

Figure 1 summarizes the submitted architecture. For every article, each model scores the ten article–headline pairs independently or as a grouped list, depending on its training objective. At inference time, all model outputs are converted into a probability distribution over the ten candidate headlines using softmax. The final distribution is a weighted linear combination of the individual distributions:

$$P(t_i \mid a) = \sum_{m=1}^{M} w_m P_m(t_i \mid a), \tag{2}$$

where a is the article, t_i is one candidate headline, P_m is the distribution produced by model m, and w_m is the validation-selected ensemble weight. Candidates are sorted in descending order of $P(t_i \mid a)$ to obtain the final ranking.

All four models use `BSC-LT/MrBERT-es` as the base encoder. MrBERT-es is a bilingual Spanish–English encoder pretrained with long-context capacity, making it suitable for pairing full political articles with concise headline candidates. Our implementation uses the HuggingFace Transformers ecosystem [7], following the standard BERT-style sequence-pair formulation [8]. For tokenization, we pass the article as the first sequence and the headline as the second sequence, using `truncation="only_first"`. This choice preserves the full candidate headline whenever the maximum sequence length is reached.

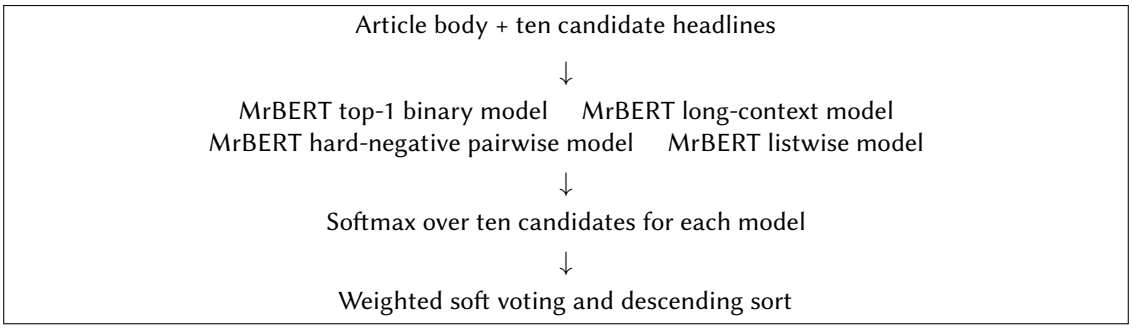

Figure 1: Overview of the text-only ensemble used for PoliticHeadlinES-IberLEF 2026.

4. Model Variants

4.1. Long-Context Pointwise Ranking

Our first ranking model expands every article into ten training pairs, one per candidate headline. The gold ranking is transformed into a graded relevance score. For a candidate appearing at position r in the gold ranking, the target score is proportional to 2^{10-r}. This produces high separation between the true top headline and lower-ranked candidates while still keeping supervision for the complete ranking.

The model is trained as a single-output sequence regression/classification head over article–headline pairs. We use a maximum sequence length of 2048 tokens, a learning rate of 2×10^{-5}, weight decay of 0.01, mixed precision, three epochs, and early stopping based on validation loss [9]. During evaluation, the ten scalar scores for an article are sorted to produce the predicted ranking.

4.2. Binary Top-1 Classifier

Because PA-nDCG returns zero when the first headline is incorrect, we also train a model that focuses exclusively on identifying the best headline. For each article, the headline placed first in the gold ranking is labeled as positive and the remaining nine candidates are labeled as negative. This creates a strong class imbalance, so the binary model uses a weighted binary cross-entropy loss with a positive weight of 12.

The binary classifier is fine-tuned for three epochs with maximum length 2048, batch size 4, learning rate 2×10^{-5}, mixed precision, warmup ratio 0.05, and early stopping. Although this model ignores the lower positions during training, its logits can still be normalized across the ten candidates and used as a distribution for ensemble inference.

4.3. Pairwise Hard-Negative Model

The hard-negative model is designed to improve discrimination between the true headline and highly plausible distractors [10]. First, a previously trained long-context MrBERT model is used as a teacher. For each article, the teacher scores all ten candidate headlines. The highest-scoring incorrect headline is selected as the hard negative, while the top gold headline is used as the positive example.

The student model is then trained with margin ranking loss:

$$\mathcal{L} = \max(0, \gamma - s^{+} + s^{-}), \tag{3}$$

where s^{+} is the score assigned to the positive headline, s^{-} is the score assigned to the selected hard negative, and $\gamma = 0.5$. Training uses maximum length 2048, batch size 4, gradient accumulation of 4 steps, learning rate 2×10^{-5}, three epochs, mixed precision, and early stopping based on pairwise validation accuracy.

Table 1
Model variants included in the final ensemble.

Model	Objective	Max. length	Main loss	Weight
Top-1 binary	Pointwise	2048	Weighted BCE	0.01
Long-context	Pointwise	2048	Regression/eval loss	0.83
Hard negative	Pairwise	2048	Margin ranking	0.05
Listwise	Listwise	1024	KL divergence	0.11

4.4. Listwise Ranking Model

The listwise model treats the ten candidate headlines for each article as a single training instance, following the general motivation of listwise learning-to-rank methods [11]. Each candidate is encoded as an article–headline pair and the ten scalar outputs are reshaped into a score vector. The gold ranking is converted into a probability distribution using exponentially decreasing gains and normalization. The model minimizes the KL divergence between the gold distribution and the log-softmax of the predicted scores:

$$\mathcal{L}_{list} = \mathrm{KL}\left(q(t_i \mid a) \parallel \log P(t_i \mid a)\right). \tag{4}$$

This variant uses maximum length 1024 to reduce memory requirements when processing all ten candidates together. The script is prepared for distributed data parallel training over four GPUs, with batch size 1, gradient accumulation of 16 steps, learning rate 2×10^{-5}, weight decay 0.01, four epochs, and validation every 100 optimization steps.

5. Experimental Setup

5.1. Data Processing

The public training data are split into training and validation partitions using a fixed random seed of 42. All scripts preserve the original ten-candidate structure of the task. Depending on the objective, the data are expanded into independent pairs, positive–negative pairs, or grouped candidate lists. No external training data are used.

For all pointwise and pairwise models, the tokenizer receives the article body and one headline as a sequence pair. Padding is fixed to the selected maximum length. The article body is the only sequence allowed to be truncated, since truncating candidate headlines would remove the most compact signal for the ranking decision.

5.2. Training Configuration

Table 1 summarizes the main configuration of the four models used in the ensemble. All models are initialized from MrBERT-es and trained with the HuggingFace Transformers and PyTorch libraries [12].

5.3. Ensemble Selection

After training the individual models, we precompute their candidate-level probability distributions on the validation split. We then perform grid search over convex combinations of the four models. For each weight tuple (w_1, w_2, w_3, w_4) such that $w_i \geq 0$ and $\sum_i w_i = 1$, the blended distribution is evaluated with PA-nDCG. The best validation configuration assigns weights 0.01, 0.83, 0.05, and 0.11 to the binary, long-context, hard-negative, and listwise models respectively.

The dominance of the long-context model suggests that preserving a large portion of the article is important for this task. However, the non-zero weights assigned to the hard-negative and listwise variants indicate that alternative objectives provide complementary ranking signals. The binary top-1 model receives a very small weight in the final ensemble, which suggests that focusing only on the top headline was less robust than modeling graded relevance across the candidate set.

Table 2
Results of the submitted system.

System	Development PA-nDCG	Top1 Accuracy	Official PA-nDCG
Top-1 binary MrBERT	0.863	–	–
Long-context MrBERT	0.911	–	–
Hard-negative MrBERT	–	0.917	–
Listwise MrBERT	0.874	–	–
Weighted ensemble	0.935	–	0.903

Table 3
Official PoliticHeadlinES-IberLEF 2026 leaderboard (selected entries). PA-nDCG on the test phase; metric_1_3 column. Our submission is highlighted in bold.

Rank	Team	PA-nDCG
1	Theron	0.954
2	padelice	0.928
3	VerbaNex AI Lab	0.922
4	**acho_gpt (ours)**	**0.903**
...	...	...
20	joseagd (baseline)	0.517
21	mimic	0.348

6. Results and Discussion

Table 2 reports the development and official PA-nDCG scores for each model variant and the final ensemble. Development PA-nDCG is reported for all models except the hard-negative variant, which is evaluated using Top-1 Accuracy since its pairwise training objective optimizes a binary decision rather than a complete ranking. Among individual models, the long-context MrBERT achieves the strongest development PA-nDCG at 0.911. The weighted ensemble improves over all individual models on the development split, reaching 0.935, and obtains an official PA-nDCG of 0.903 on the test set.

Table 3 contextualises this result within the official competition leaderboard. Our system, submitted under the team name *acho_gpt*, ranked 4th out of 21 participating teams with a PA-nDCG of 0.903. The difference between our text-only approach and the Top 3 teams could suggest the possibility that incorporating multimodal signals—as discussed in Section 7—could close this gap. The gap between our system and the task baseline is substantial: the official baseline (joseagd, rank 20) scores 0.517, meaning our ensemble nearly doubles the baseline PA-nDCG.

The design of the final ensemble reflects the structure of the metric. Since PA-nDCG gives most of its reward to the first position, models must distinguish the correct headline from very similar distractors. At the same time, the remaining positions still affect the score when the first candidate is correct. This makes the task suitable for combining pointwise, pairwise, and listwise formulations.

The long-context model receives the largest ensemble weight. This is consistent with the nature of political news, where the information that distinguishes two plausible headlines may appear far from the beginning of the article. Names of political actors, parties, institutions, and dates can overlap across candidates, so the model benefits from broader article context.

The hard-negative model addresses a different failure mode. In many cases, the most difficult incorrect headline is not a random distractor but a semantically close headline referring to a related political event. By selecting the highest-scoring incorrect candidate according to a teacher model, the pairwise objective trains the student to focus on fine-grained distinctions between nearly plausible headlines.

The listwise model contributes by seeing the complete set of candidates jointly during training. Although inference still relies on scalar scores per candidate, the loss encourages a distribution over all ten titles rather than independent binary or regression decisions. This can help stabilize lower-ranked

positions, which are not the dominant factor in PA-nDCG but remain relevant after the top-1 condition is satisfied.

Our main limitation is that the submitted system is text-only. PoliticHeadlinES includes a multimodal setting in which the associated news image may help disambiguate actors, events, locations, or temporal context. Because the final submission reuses the same ranking for both output columns, any potential gain from visual evidence is left unexplored. Future work should evaluate vision-language encoders or late fusion strategies that combine image and text signals without weakening the strong long-context textual baseline.

7. Conclusion

We presented the AchoGPT system for PoliticHeadlinES-IberLEF 2026. Our approach frames Spanish political headline selection as a ranking problem over ten candidate headlines and combines four MrBERT-es variants trained with complementary objectives. The final system uses weighted soft voting over candidate-level probability distributions, with ensemble weights selected on the development split using PA-nDCG.

The implementation highlights three practical lessons. First, preserving long textual context is important for political headline ranking. Second, hard negatives are useful because many distractors are semantically close to the correct headline. Third, optimizing only for the top headline is insufficient on its own, even when the official metric heavily rewards top-1 accuracy. In future work, we plan to incorporate multimodal evidence, calibrate model probabilities before ensembling, and analyze errors by political entity overlap, temporal proximity, and event similarity.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT/Codex in order to: code inspection, drafting assistance, grammar and spelling check, and paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

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. 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.

[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] T. G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, pp. 1–15.

[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, arXiv preprint arXiv:2602.21379 (2026).

[6] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems (TOIS) 20 (2002) 422–446.

[7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[9] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al., Mixed precision training, arXiv preprint arXiv:1710.03740 (2017).

[10] J. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, arXiv preprint arXiv:2010.04592 (2020).

[11] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 129–136.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).