UCO-UC-CICESE-Plenitas at PoliticHeadlinES-IberLEF 2026: Multimodal Headline Ranking in Spanish Political News Using Large Language Models and Embeddings

Yoan Martínez-López[1,2,*], Yanaima Jauriga[3], Mayte Guerra Saborit[3], Julio Madera[3], Luis Quintero Domínguez[4], Ansel Rodríguez-González[5], Carlos de Castro Lozano[1,2] and Jose Miguel Ramirez Uceda[1,2]

[1]*Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[4]*Universidad de Sancti Spíritus, Cuba*
[5]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, México*

Abstract

This paper describes the participation of the **UCO-UC-CICESE-Plenitas** team in PoliticHeadlinES at IberLEF 2026, a shared task on headline ranking for Spanish political news in which, given an article, systems must rank ten candidate headlines, only one of which is the original. We addressed both subtasks proposed by the organisers: Subtask 1 (text-only ranking) and Subtask 2 (multimodal ranking, with the article cover image as an additional signal). Our system is a modular zero-shot pipeline that integrates seven text rankers —including a TF-IDF baseline, multilingual sentence encoders (BGE-M3, multilingual-E5, MiniLM, LaBSE), an NLI cross-encoder (‘nli-deberta-v3-base‘), and pointwise LLM scorers served through the Ollama (local and cloud)— combined with a vision-language module based on CLIP and SigLIP. For the multimodal subtask, textual and visual scores are min-max normalised at article level, which heavily rewards placing the correct headline in first position. On the development set our best run obtained 0.876 PA-nDCG using ‘multilingual-e5-large-instruct‘; on the official evaluation set, our best submission —combining BGE embeddings with the LLM scorer (‘BAAI-LLM‘)— reached 0.705 PA-nDCG, ranking 13th out of 20 teams in Subtask 2 and 12th out of 13 in Subtask 1 (0.856). Our results suggest that, in this benchmark, dense semantic scorers combined with LLM-based pointwise relevance are the most effective single signal, while the visual modality contributes only modestly under the simple late-fusion scheme adopted.

Keywords

Artificial Intelligence, Large Language Models, Headline Ranking, Multimodal Learning, Natural Language Processing

1. Introduction

Political news constitutes one of the main sources of information in contemporary societies and plays a decisive role in shaping public opinion, democratic deliberation, and electoral processes. Within this information ecosystem, headlines occupy a particularly relevant position, as they strongly condition the way in which citizens discover, interpret, and share news [1]. In digital environments, where user attention is scarce and information overload demands rapid decisions, the headline frequently becomes the most visible element —and sometimes the only one— that is read before deciding whether to access the full content, share the article, or engage with it in any other way[2]. This centrality explains why a well-crafted headline can significantly increase the reach and social impact of a news story, whereas a

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
ymlopez2022@gmail.com (Y. Martínez-López); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); mayte.guerra@reduc.edu.cu (M. G. Saborit); julio.madera@reduc.edu.cu (J. Madera); laqdominguez@gmail.com (L. Q. Domínguez); ansel@cicese.mx (A. Rodríguez-González); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-9556-5869 (M. G. Saborit); 0000-0001-5551-690X (J. Madera); 0000-0002-3527-0516 (L. Q. Domínguez); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

poor or misleading headline can distort the perception of the reported facts, even when the body of the article is rigorous. The automatic generation and selection of headlines, however, poses considerable challenges. A good headline must be simultaneously concise, faithful to the article's content, informative, and appealing to the reader, while avoiding sensationalist, ambiguous, or biased framings that may compromise informational quality[3, 4]. This balance is especially delicate in the political domain, where lexical nuances, ideological framings, and contextual references can substantially alter the reading of the same event. For these reasons, headline selection and ranking have become a problem of growing interest for Natural Language Processing (NLP) and Information Retrieval (IR), as well as, increasingly, for multimodal learning when articles are accompanied by visual material that provides complementary information to the text [5, 6]. It is in this context that PoliticHeadlinES is situated, a shared task focused on headline ranking for political news in Spanish [7, 8, 9]. Given a news article, participating systems must rank a set of ten candidate headlines, of which only one corresponds to the actual headline associated with the article, while the remaining nine act as distractors designed to be plausible. The task therefore evaluates the ability of systems to identify the correct headline among competitive alternatives, leveraging both textual information and, in its multimodal variant, the visual cues associated with the news item. This formulation makes it possible to study in a controlled manner the extent to which current models are able to capture the fine-grained semantic relationship between a journalistic text and its headline, a capability with direct implications for applications such as content recommendation, headline-body coherence verification, clickbait detection, and editorial support in digital media. The present work describes our participation in PoliticHeadlinES, presents the approaches developed to address the task, and analyzes the results obtained, with the aim of providing empirical evidence on the behavior of different NLP and IR strategies —and, where applicable, multimodal ones— in the specific domain of political headlines in Spanish.

2. Methodology

This section describes the system developed to address the two subtasks proposed in PoliticHeadlinES 2026: subtask 1, headline ranking based solely on textual information, and subtask 2, multimodal ranking that incorporates the image associated with the article. The system has been designed as a modular and configurable pipeline that integrates multiple strategies from Natural Language Processing (NLP), Information Retrieval (IR), and multimodal learning, allowing for a controlled comparison of different paradigms —ranging from classical approaches based on lexical frequencies to Large Language Models (LLMs) and pretrained vision-language models— on the same dataset and under identical evaluation protocols.

Figure 1 provides an overview of the proposed architecture. The system combines multiple text-based rankers, a vision-language module, score normalization, and a late-fusion strategy to generate the final ranking of candidate headlines.

2.1. Natural Language Processing (NLP)

Natural Language Processing (NLP) provides the textual processing capabilities used by our ranking pipeline. As shown in Figure 1, the NLP module transforms article bodies and candidate headlines into representations from which relevance scores can be computed. In our system, this module combines a lexical scorer based on TF–IDF and a Natural Language Inference (NLI) cross-encoder, providing complementary lexical and semantic signals for headline ranking [10, 11, 12].

Prior to any representation, all texts undergo the same lightweight normalisation step: Unicode NFC normalisation, collapsing of repeated whitespace, removal of HTML residuals from the scraped sources, and lower-casing only for the lexical branch. Punctuation, casing, and diacritics are kept for the semantic and LLM branches, since they carry framing cues that are particularly relevant in political headlines (e.g., quotation marks, question marks, and capitalised proper nouns).

Lexical scorer (TF-IDF). A term-frequency–inverse-document-frequency vectoriser is fitted on the concatenation of all article bodies in the development split, with unigrams and bigrams, sub-linear

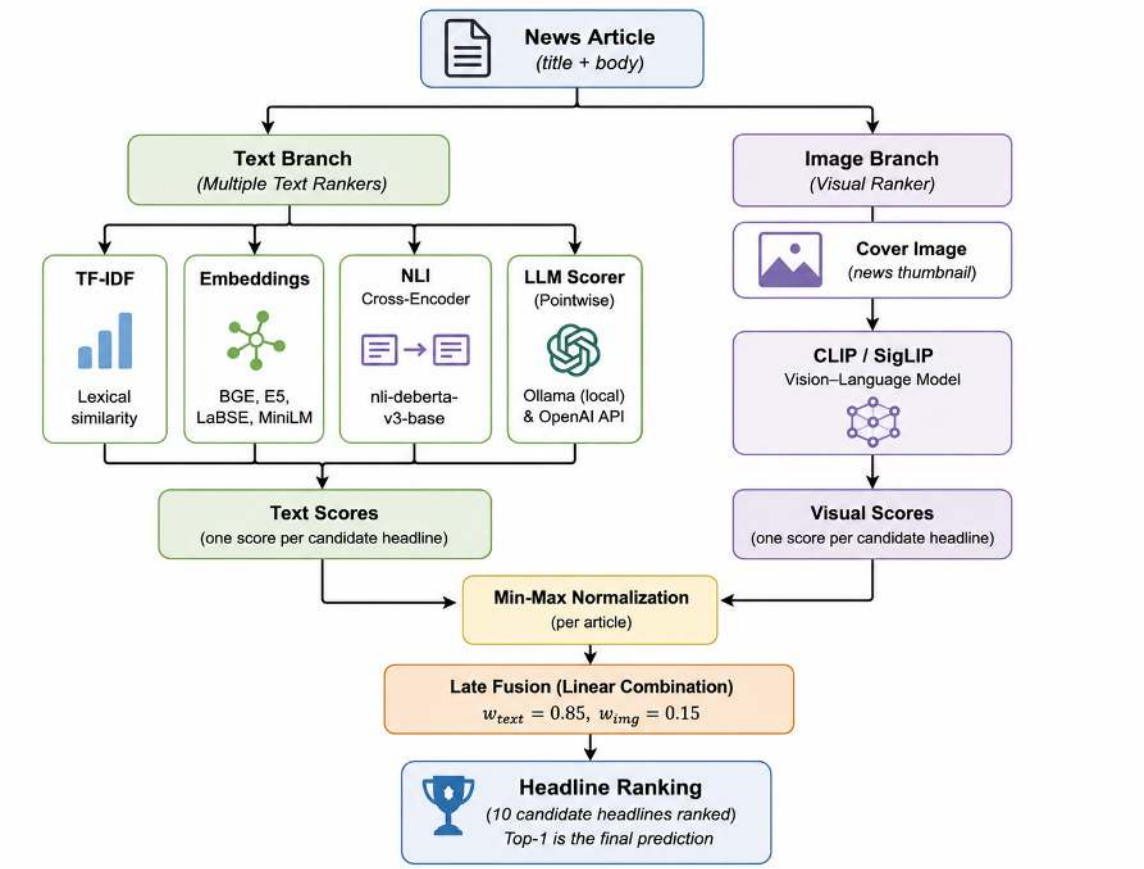

Figure 1: Overview of the proposed PoliticHeadlinES pipeline. Textual relevance is estimated through lexical, embedding-based, NLI, and LLM scorers, while visual relevance is obtained using CLIP/SigLIP. Textual and visual scores are normalized and combined through late fusion to generate the final ranking of candidate headlines.

term-frequency scaling, and a minimum document frequency of two. Stop-word filtering is intentionally disabled: in headline ranking, the presence or absence of function words contributes to surface-level fidelity between article and headline. Both bodies and candidate headlines are projected into the resulting vector space, and the relevance score is computed as the cosine similarity between each headline vector and its article-body vector. This scorer provides a transparent, low-cost reference against which the contribution of the semantic and generative components can be measured.

Natural Language Inference (NLI) As a complementary signal, each (article, headline) pair is scored by a multilingual Natural-Language-Inference cross-encoder [13, 14]. The article body is treated as the premise and the candidate headline as the hypothesis, and the entailment probability returned by the model is used as the relevance score. This formulation matches the task semantics more closely than pure similarity, since a representative headline should be entailed by the article rather than merely share vocabulary with it. Because cross-encoders process both inputs jointly, the body text is truncated from the end to fit the model's context window when necessary, prioritising the lead paragraphs, which empirically carry most of the headline-relevant information in news articles.

2.2. Embeddings

Embeddings are dense, low-dimensional vector representations that map discrete symbols —words, sentences, documents, or, more generally, items from any modality— into a continuous space in which geometric proximity reflects semantic or functional similarity [15]. Their core appeal is that they convert symbolic data, which is intrinsically sparse and combinatorial, into numerical objects on which linear

algebra, similarity metrics, and gradient-based learning can operate efficiently. This shift from symbolic to distributed representations is one of the foundational ideas of modern Natural Language Processing and lies behind most current systems in information retrieval, recommendation, classification, and clustering[16]. The notion of embedding emerged with static word representations such as word2vec, GloVe, and fastText, which learn a single vector per token from large unlabelled corpora using the distributional hypothesis —the assumption that words appearing in similar contexts tend to convey similar meanings. The subsequent introduction of contextualised representations, first with ELMo and then with Transformer-based encoders such as BERT and its multilingual variants, removed the one-vector-per-word limitation and allowed the representation of a token to depend on its surrounding context [17, 18]. More recently, sentence and passage encoders —such as Sentence-BERT, the E5 family, and BGE— have been specifically trained with contrastive objectives so that the cosine similarity between two embeddings is a calibrated estimate of semantic relatedness, which makes them particularly well suited to ranking and retrieval tasks. Beyond text, the same idea extends naturally to other modalities: image, audio, and joint vision–language models such as CLIP and SigLIP project heterogeneous inputs into a shared embedding space, enabling cross-modal retrieval and similarity. In headline ranking, embeddings provide precisely the type of paraphrase-robust similarity signal that lexical features cannot capture, which is why they form one of the central components of our pipeline. The embedding block complements the surface-level NLP scorers with dense, semantic representations of texts. Its goal is to bring together headlines and article bodies that convey the same content despite using different vocabulary —a frequent situation in political news, where headlines often paraphrase, reframe, or condense the body text. Both article bodies and candidate headlines are encoded with a multilingual sentence-encoder. The default backbone is BAAI/bge-m3, which produces 1024-dimensional vectors and offers strong zero-shot performance on Spanish retrieval; intfloat/multilingual-e5-large-instruct is supported as an alternative through the same interface. For E5-family models we follow the prescribed asymmetric protocol, prepending query: to headlines and passage: to article bodies; for BGE-M3 no prefix is used. Embeddings are L2-normalised so that cosine similarity reduces to a dot product, and the relevance score of a (body, headline) pair is taken directly as this similarity value.

Long article bodies are truncated to each model's maximum input length without applying sliding-window chunking: preliminary inspection of the development set indicated that headline-relevant information is concentrated in the first paragraphs, in line with the inverted-pyramid structure of news writing, so chunking and pooling provided little additional signal at the cost of substantially higher inference time.

For the multimodal subtask, the embedding block is extended with SigLIP as a vision–language encoder. SigLIP is chosen over CLIP because its sigmoid-based contrastive objective yields per-pair similarities that are well calibrated without requiring a softmax over batch candidates, which makes its scores directly usable as a standalone signal at article level. For each article, the cover image is encoded into a SigLIP image embedding, and each candidate headline is independently encoded into a SigLIP text embedding; the visual relevance score is then the cosine similarity between the two. The textual and visual scores are subsequently combined in the ranking module by article-level min-max normalisation followed by a convex combination.

2.3. Large Language Model (LLM)

Large Language Models (LLMs) are neural networks —overwhelmingly based on the Transformer architecture— trained on massive amounts of text to predict the next token in a sequence [19]. Despite the simplicity of this self-supervised objective, scaling these models to billions of parameters and trillions of tokens has produced systems that exhibit broad linguistic competence, factual recall, multi-step reasoning, and the ability to follow natural-language instructions across tasks for which they were not explicitly trained. As a result, LLMs have become the dominant paradigm in Natural Language Processing, replacing or subsuming many of the specialised models that previously addressed individual tasks such as classification, summarisation, translation, or question answering [20, 21]. The current LLM landscape includes both proprietary, API-served systems —such as the GPT, Claude, and Gemini

families— and a fast-growing ecosystem of open-weight models, including Llama, Qwen, Mistral, Gemma, DeepSeek, and gpt-oss, which can be served locally through frameworks like Ollama or vLLM. Most modern LLMs are released as base models pretrained on web-scale corpora, and are then aligned through instruction tuning and reinforcement learning from human or AI feedback to produce models capable of dialogue, tool use, and structured-output generation. A more recent line of work introduces reasoning-tuned models, which expose an explicit chain-of-thought (typically delimited by tags such as <think>...</think>) before emitting the final answer, trading higher latency for substantially better performance on tasks that require multi-step inference. In headline ranking, LLMs are particularly attractive because they can act as zero-shot pointwise relevance estimators: rather than computing geometric similarity in an embedding space, they exploit world knowledge and pragmatic competence to judge how well a candidate headline summarises an article from a journalistic standpoint —a capability that complements lexical and semantic scorers and forms the third pillar of our pipeline.

The LLM block performs pointwise zero-shot relevance estimation: the model is queried once per (article, headline) pair with a prompt that frames the decision as a faithfulness and representativeness judgement, and a numeric score is extracted from its response. This component contributes a different kind of evidence from those described above: rather than measuring lexical overlap or geometric proximity in an embedding space, it relies on the world knowledge and pragmatic competence of a generative model to assess how well a candidate headline summarises the article from a political-communication standpoint.

A single prompt template is used across all backbones to ensure comparability. It contains a short role description, a concise restatement of the task, the article body —truncated to a fixed character budget when necessary— and the candidate headline, followed by an instruction to emit the score inside an explicit XML-like tag (e.g., <respuesta>...</respuesta>). This format reduces the variance of the output across families of models with very different default behaviours and makes the response unambiguously machine-parseable. All calls use deterministic decoding (temperature 0, fixed top-p) and a low maximum-tokens budget; few-shot examples are kept short and balanced across the rating scale.

Three backends are integrated behind a common interface, so that any model can be swapped in or out without modifying the rest of the pipeline:

- Ollama (local). Open-weight models served from a local Ollama instance via its /api/chat endpoint. This is the default backend during development and is used for models such as Llama 3.1, Qwen 2.5, and Gemma 2.
- Ollama Cloud. The same Ollama interface is reused against the hosted endpoint at https://ollama.com, with Bearer-token authentication using a per-user API key, enabling access to larger models without local hardware constraints.
- OpenAI API. Used as a strong reference backend with GPT-class models, called through the official Chat Completions interface.

The numeric rating is extracted from the model output with a tolerant parser that (i) strips reasoning traces emitted by reasoning-tuned models, such as <think>...</think> or <thinking>...</thinking> blocks; (ii) prefers content located inside the explicit response tag; and (iii) falls back to the last non-empty line of the answer when the tag is missing. Pairs whose response cannot be parsed unambiguously are assigned a neutral score and logged, ensuring that ranking is never silently corrupted by a malformed generation. This conservative behaviour keeps the LLM scorer robust both to small instruction-tuned models, which occasionally drift from the requested format, and to large reasoning models, which tend to produce verbose chains of thought before reaching the final answer.

2.4. Task formalization

Given a news article represented by its textual body a and, in the multimodal case, an associated image v, the system receives a set of ten candidate headlines $T = \{t_1, t_2, \ldots, t_{10}\}$, of which only one corresponds to the original headline while the remaining nine act as distractors. The expected output is

a permutation π of the ten candidates ordered from highest to lowest likelihood of being the correct headline. The objective of the system is therefore to learn (or compute without explicit training) a scoring function $s(a, t)$ —or $s(a, v, t)$ in the multimodal case— that maximizes the position of the true headline at the top of the ranking.

2.5. Data and preprocessing

The official corpora provided by the organizers were used: `train_corpora.csv` for training/tuning and `PoliticHeadlinES_phase_2_test_public.csv` for evaluation. Each instance contains a unique identifier, the article body `title_1 (article_body)`, an image hash `(image_hash)`, and ten candidate headline columns (`title_1` through `title_10`) [8]. Images were downloaded as separate compressed archives and stored in a local directory, with dynamic path resolution via extension lookup `(.jpg, .jpeg, .png, .webp)`. Textual preprocessing was deliberately kept minimal —whitespace cleaning and null-value handling— in order to preserve the stylistic and lexical information that might be discriminative among plausible headlines.

2.6. System architecture

The system follows a factory pattern that, depending on the configuration, instantiates one of seven text rankers and one of two multimodal rankers. Each component is described below.

2.6.1. Text rankers (subtask 1)

TF-IDF cosine similarity. This component implements the official baseline using scikit-learn's TfidfVectorizer with a maximum vocabulary of 50,000 terms and 1–2 word n-grams. The vectorizer is fitted on the combined corpus of article bodies and training headlines, and the ranking is obtained via cosine similarity between the article vector and each of the ten headline vectors. This approach, fast and not requiring GPU resources, serves as a lexical baseline against which the improvements of the neural approaches can be contextualized. Multilingual dense embeddings. This variant encodes the article and headlines via a Sentence-Transformers model and computes cosine similarity in the embedding space. Several candidate models were explored —including paraphrase-multilingual-mpnet-base-v2, paraphrase-multilingual-MiniLM-L12-v2, sentence-transformers/LaBSE, BAAI/bge-m3, and BAAI/bge-large-en-v1.5— with the aim of evaluating the trade-off between multilingual coverage, dimensionality of the semantic space, and performance in Spanish. The use of `trust_remote_code=True` allowed the integration of architectures with custom code in the model repository.

NLI cross-encoder strategy treats the problem as a Natural Language Inference (NLI) task, using the cross-encoder/nli-deberta-v3-base model through Hugging Face Transformers' text-classification pipeline. For each (article, headline) pair, the model produces probabilities over the entailment, neutral, and contradiction classes; the entailment probability is used as the affinity score, with truncation to 512 tokens and a batch size of 8. The underlying hypothesis is that the correct headline should be more strongly entailed by the article body than the distractors.

The backends based on Ollama are integrated both in local mode (http://localhost:11434) and in remote mode (cloud server), with qwen3:8b as the main model and alternative exploration of llama3.2:3b and ministral-3:14b. The unified interface allows the underlying model to be swapped without modifying the rest of the pipeline. The prompt used is identical to the one employed with the Hugging Face API, in order to guarantee comparability. All rankers implement a common interface with the method `rank_titles (source_text, titles)`, which returns a list of indices ordered by descending score. This enables component swapping without coupling between modules.

2.6.2. Multimodal component (subtask 2)

To integrate visual information, a pretrained vision-language model is used, configured by default as openai/clip-vit-base-patch32, with alternative exploration of openai/clip-vit-large-patch14 and of

the SigLIP family (google/siglip-so400m-patch14-384). Processing is carried out via AutoModel and AutoProcessor from Transformers, allowing the system to operate interchangeably with CLIP and SigLIP without code changes. For each (image, headline) pair, the model produces a logit that quantifies image–text affinity; the ten logits corresponding to the ten headlines are used as visual scores. Inference is performed under `torch.inference_mode()` to avoid building the gradient computation graph and reduce inference time.

2.6.3. Multimodal fusion strategies

Two strategies were implemented for subtask 2:

- `clip_only`. This strategy uses exclusively the image–headline similarity computed by CLIP/SigLIP. It serves as a reference to quantify how much discriminative information the visual modality contributes on its own, as well as to detect cases in which the image is highly predictive of the headline (for instance, photographs portraying the protagonist mentioned in the text).
- `clip_llm_fusion`. This strategy linearly combines the scores of the text ranker with those of the visual component. To guarantee comparability of magnitudes, both score vectors are normalized via min-max normalization to the [0, 1] interval before fusion:

$$s_{\text{final}}(t_i) = w_{\text{text}} \cdot \tilde{s}_{\text{text}}(t_i) + w_{\text{img}} \cdot \tilde{s}_{\text{img}}(t_i) \tag{1}$$

The weights were set to `w_text = 0.85` and `w_img = 0.15` based on empirical observation, reflecting the fact that the textual signal is dominant for discriminating among plausible headlines, while the image provides complementary information of smaller magnitude but potentially decisive in ambiguous cases. When the image is not available, the system automatically falls back to text-only ranking.

2.6.4. Evaluation metrics

Following the official task protocol, the system is evaluated using a primary-aware variant of the Normalized Discounted Cumulative Gain (nDCG), referred to as PA-nDCG (Primary-Aware nDCG). This metric, defined as:

$$\text{PA-nDCG} = \begin{cases} \alpha + (1-\alpha) \cdot \text{nDCG}_{\text{aux}} & \text{if the correct headline is ranked first} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Assigns decisive weight to placing the correct headline in the first position of the ranking (accounting for 90% of the metric), while the remaining 10% comes from the nDCG computed over the ordering of the nine remaining distractors. This formulation reflects the applied nature of the task: in a real-world scenario, the usefulness of the system depends primarily on identifying the correct headline, with the ordering of the distractors being a secondary concern.

3. Results and discussion

The system was developed in Python using the PyTorch ecosystem (Transformers, sentence-transformers, scikit-learn, pandas, NumPy, PIL). Local LLMs were managed via Ollama, launched as a background service. Inference was executed on a CUDA-enabled GPU when available, with automatic device detection and precision selection (float16 on GPU, float32 on CPU). Authentication with the Hugging Face Hub provided access to gated models and to the Inference API. A fixed seed (SEED=42) was used to ensure experimental reproducibility. The final output of the system is materialized in a `results.csv` file containing the columns id, `task_1`, and `task_2` in the format required by the CodaBench evaluation platform, where each prediction is represented as an ordered sequence of t1–t10 tokens.

3.1. Development

Table 1 reports the development PA-nDCG scores obtained with our three dense-embedding configurations. The best result, 0.876, is achieved by intfloat/multilingual-e5-large-instruct, slightly ahead of BAAI/bge-m3 (0.856) and clearly above the paraphrase-multilingual-mpnet-base-v2 baseline (0.816) and LaBSE(0.816). The gap between the two large multilingual encoders is small ($\Delta = 0.020$) but consistent across articles, suggesting that the asymmetric query: / passage: formulation used by the E5 family is well aligned with the structural asymmetry of the task —long article body as passage, short candidate headlines as queries—, whereas the symmetric encoding of BGE-M3 leaves a small but non-negligible amount of signal on the table. The clearer gap with respect to the paraphrase-multilingual baseline ($\Delta = 0.060$) confirms that high-capacity, retrieval-oriented sentence encoders are markedly better suited to Spanish political headline ranking than general-purpose paraphrase models, even when the latter are also multilingual and trained on Spanish data. Based on these observations, multilingual-e5-large-instruct was retained as one of the strongest text-only configurations, while BGE-based variants were further explored in combination with LLM scoring for the final submissions.

Table 1
Development PA-nDCG scores obtained by the explored embedding configurations

Method	Score
baai/bge	0.856
intfloat-multilingual-e5-lar	0.876
paraphrase-multilingual	0.816
LaBSE	0.816

Table 2 shows the official leaderboard for the development phase, with thirteen participating teams ranked by metric_1_3 (the PA-nDCG used for Subtask 1) and metric_2_3 (Subtask 2). Our system, UCO-UC-CICESE-Plenitas, occupies the 12th position with a Subtask 1 score of 0.856 and a Subtask 2 score of 0.916. Three observations are worth highlighting. First, the upper part of the leaderboard is remarkably compressed: the three top teams (padelice, VerbaNex AI Lab, and enrique_lr) tie at 0.976 on both subtasks, and seven of the thirteen systems lie within five hundredths of a PA-nDCG point of each other (between 0.953 and 0.976). This high density at the top suggests that the development set is largely solvable by competitive systems and that the distinguishing factor is, to a large extent, the handling of a small number of genuinely difficult articles rather than overall ranking quality. Second, our system stands out from most of the leaderboard in that its Subtask 2 score (0.916) is markedly higher than its Subtask 1 score (0.856) —a gap of 0.060 in favour of the multimodal subtask. In contrast, the majority of teams in the table report identical or nearly identical scores across both subtasks (e.g., padelice, VerbaNex AI Lab, enrique-lr, deniscedeno, francordel, i2c-uhu-perseus all tie at the same value), which indicates that, for most participants, the visual modality does not modify the ranking induced by the textual signal. The fact that our multimodal pipeline does diverge from the text-only one is consistent with the design of our late-fusion module, and shows that the CLIP / SigLIP signal we integrate is non-trivial; the open question, addressed in our official-evaluation discussion, is whether this divergence translates into a positive contribution on unseen data. Third, the gap of 0.120 PA-nDCG between our Subtask 1 score and the leaders' (0.856 vs. 0.976) corresponds, given the strong weight of $|\alpha| = 0.9$ assigned to the first position by PA-nDCG, to a difference of roughly twelve articles per hundred in which the leading systems place the correct headline in rank one while ours does not. This motivates the analysis carried out in the next subsection on where our system fails to lift the correct headline to the top, and on how the additional visual evidence reshuffles those cases.

3.2. Evaluation

Table 3 reports the PA-nDCG scores obtained by our eight submissions on the official evaluation set, ordered along the methodological progression of our pipeline. Three groups of configurations can be

Table 2
All Teams/Participants

Rank	Participant/Team	metric_2_3	metric_1_3
1	padelice	0.976	0.976
2	VerbaNex AI Lab	0.976	0.976
3	enrique-lr	0.976	0.976
4	deniscedeno	0.956	0.956
5	acho-gpt	0.955	0.955
6	francordel	0.954	0.954
7	davidhzrz	0.953	0.953
8	i2c-uhu-perseus	0.934	0.934
9	haizz	0.916	0.916
10	HEART-UJA-UA	0.936	0.916
11	jesusgs	0.894	0.914
12	**UCO-UC-CICESE-Plenitas**	**0.916**	**0.856**
13	PUCE-UTP	0.875	0.856

clearly distinguished.

First, the visual and lexical baselines cluster tightly in the 0.47–0.53 range. The Nomic-based submission (0.472) and the four CLIP / SigLIP-derived variants —googlemb224 (0.513), baseLine (0.517), BL-patch-32 (0.517), patch-14 (0.522), and patch-336 (0.526)— all hover around the same level, with differences below one hundredth of a PA-nDCG point between most of them. This result is informative on its own: increasing the CLIP/SigLIP patch resolution from 32 to 336 produces no meaningful gain, which suggests that on this benchmark the visual encoder is not bottlenecked by image resolution but rather by the headline–image alignment quality, i.e. by how directly the cover image speaks to the wording of the candidate headlines. The visual modality therefore contributes a roughly constant —and modest— amount of signal, regardless of the specific vision-language backbone employed.

Second, the dense-embedding ranker based on BAAI/bge-large (0.493) lands in the same band as the visual baselines, well below what we observed for the same family of encoders on the development set (0.856–0.876, Section 4.1). This drop is the clearest indicator of the distribution shift between development and evaluation splits: a configuration that was almost saturating the development metric falls back to baseline-level performance on the test set, which suggests that the evaluation distractors are substantially harder to discriminate by surface-level semantic similarity alone.

Third, the configuration that combines BGE embeddings with the pointwise LLM scorer —baai/bge-LLM, which corresponds to the submission BAAI-LLM— reaches 0.705 PA-nDCG, more than 20 PA-nDCG points above every other run we submitted and 42 percentage points above the weakest one (results final at 0.189, omitted from this table). Given that PA-nDCG assigns 90% of its mass to placing the correct headline in first position, this 0.20 absolute gap translates, in practice, into roughly twenty additional articles per hundred in which the LLM-guided system lifts the correct headline to rank one while the dense-embedding-only or CLIP-only systems do not. This is the single most informative observation of our experiments: on this dataset, the complementarity between geometric similarity in a multilingual embedding space and the pragmatic, world-knowledge-based judgement of an LLM is large enough to dominate any contribution coming from the visual modality under our current fusion scheme.

The submitted BAAI-LLM run combined BAAI/bge-large-en-v1.5 embeddings with qwen3:8b served through the Ollama backend, which proved to be the strongest configuration among all submitted runs.

Taken together, these results justify, a posteriori, our choice of BAAI-LLM as the primary submission and motivate the future-work directions outlined in the Conclusions, in particular (i) richer fusion strategies that let the visual modality act as a tie-breaker only on truly ambiguous cases rather than competing linearly with the text, and (ii) listwise rather than pointwise LLM scoring, which is expected to amplify further the gap already observed here.

Table 3
Comparison of submitted runs on the official evaluation set

Method	Score
nomatic	0.472
baai/bge-large	0.493
baai/bge-LLM	0.705
googlemb224	0.513
baseLine	0.517
BL-patch-32	0.517
patch-14	0.522
patch-336	0.526

Tables 4 and 5 report the official leaderboards of the evaluation phase for Task 1 (text-only ranking) and Task 2 (multimodal ranking), respectively, with PA-nDCG expressed as a percentage and twenty participating teams in addition to the organisers' baseline. Our system, UCO-UC-CICESE-Plenitas, ranks 13th out of 20 in both tasks, with 70.49 PA-nDCG on Task 1 and 71.10 on Task 2. Four observations are worth highlighting. First, both leaderboards are dominated by Theron at 95.44, followed by a compact group of seven teams in the 88–93 range (LKE-BUAP, VerbaNex AI Lab, acho-gpt, GACOP, HEART-UJA-UA, franrodrigo). The middle of the table —positions 8 to 12— spans the 77–83 band, and the lower half (13 to 18, where our system sits) ranges from roughly 58 to 71 PA-nDCG. This near-monotonic drop, without large gaps, is consistent with a continuum of system quality rather than a clean separation between strong and weak approaches. Second, our submission clearly outperforms the organisers' baseline (51.65 on Task 1, 51.34 on Task 2): the gap is of approximately +19 PA-nDCG points on Task 1 and +20 points on Task 2. Given that PA-nDCG assigns $|\alpha = 0.9|$ of its mass to placing the correct headline in first position, this gap corresponds —in practical terms— to roughly twenty additional articles per hundred in which our system lifts the correct headline to rank one and the baseline does not. Our system is therefore not only above the official reference, but well above it, even though the upper half of the leaderboard remains out of reach. Third, the absolute gap with respect to the leading system (Theron) is of 24.95 PA-nDCG points on Task 1 and 24.34 on Task 2. Most of this difference is concentrated in the rank-one slot of the metric: under PA-nDCG, the only way to recover that magnitude of score is to promote the correct headline to first position on a substantial number of additional articles, which highlights the central methodological challenge of the task —not ranking quality in a generic sense, but headline-vs-distractor disambiguation in the most contested cases. Fourth, and arguably the most informative point for our pipeline, our Task 2 score (71.10) is marginally higher than our Task 1 score (70.49), with a positive multimodal gap of approximately +0.61 PA-nDCG points. This contrasts with several teams above us in the table, for which Task 1 and Task 2 scores are essentially identical (e.g., Theron, LKE-BUAP, acho-gpt, franrodrigo, I2C-UHU, enrique-lr, UIE, I2C-UHU-Perseus), suggesting that for those systems the visual modality does not modify the text-only ranking. In contrast, four teams —ourselves, francordel, UTP, and (negatively) LYS-UDC, HEART-UJA-UA, GACOP, VerbaNex AI Lab— show measurable differences between the two subtasks. In our case, the difference is small but positive and consistent with the design hypothesis behind our late-fusion module: that the CLIP / SigLIP signal, even with a conservative weight of wimg=0.15, carries information that is not redundant with the textual scorers and acts as a tie-breaker on a small fraction of the test articles. While the magnitude of the gain is modest, it is one of the clearer multimodal effects observable in the leaderboard and motivates the future-work directions outlined in the Conclusions, in particular adaptive per-article fusion weights and image-conditioned LLM scoring.

4. Conclusions

In this work we have described the participation of the UCO-UC-CICESE-Plenitas team in PoliticHeadlinES 2026, a shared task on headline ranking for Spanish political news. We approached the task

Table 4
Results for Task1

Rank	Team	Score
1	Theron	95.43629
2	LKE-BUAP	92.75108
3	VerbaNex AI Lab	92.22656
4	acho-gpt	90.26359
5	GACOP	89.67746
6	HEART-UJA-UA	89.45794
7	franrodrigo	88.41059
8	I2C-UHU	82.59159
9	LYS-UDC	82.41000
10	francordel	80.74166
11	ITO	79.85917
12	UTP	73.94271
13	**UCO-UC-CICESE-Plenitas**	**70.49079**
14	LACELL	65.98162
15	enrique-lr	65.56252
16	UIE	64.47657
17	I2C-UHU-Perseus	62.31144
18	PUCE-UTP	58.06160
19	baseline	51.65320
20	MIMIC	34.79614

Table 5
Results for Task2

Rank	Team	Score
1	Theron	95.43629
2	LKE-BUAP	92.75108
3	VerbaNex AI Lab	92.62804
4	acho-gpt	90.26359
5	GACOP	89.67718
6	HEART-UJA-UA	89.12555
7	franrodrigo	88.41059
8	I2C-UHU	82.59159
9	francordel	81.28402
10	LYS-UDC	80.82452
11	ITO	79.61529
12	UTP	77.04902
13	**UCO-UC-CICESE-Plenitas**	**71.10051**
14	LACELL	65.86319
15	enrique-lr	65.56252
16	UIE	64.47657
17	I2C-UHU-Perseus	62.31144
18	PUCE-UTP	55.45929
19	baseline	51.34444
20	MIMIC	34.79614

as a zero-shot ranking problem and built a modular pipeline that brings together, behind a common interface, lexical (TF-IDF), semantic (multilingual sentence encoders), inferential (NLI cross-encoder), generative (LLMs through HF Inference and Ollama, local and cloud), and visual (CLIP / SigLIP) scorers. Textual and visual evidence were combined by article-level min-max normalisation followed by a linear fusion, with weights $w_{text} = 0.85$ and $w_{img} = 0.15$. Our experiments yield three main observations. First, on the development set the dense semantic scorers clearly outperformed the lexical baseline:

the best development configuration, based on multilingual-e5-large-instruct, reached 0.876 PA-nDCG, against 0.856 for the BGE-based variant and 0.816 for the LLM-based paraphrase-multilingual ranker. This confirms that paraphrase-robust similarity in a multilingual embedding space is well aligned with how Spanish political headlines relate to their article bodies. Second, on the official evaluation set we submitted nine runs covering the main configurations of the pipeline. The best one, BAAI-LLM, which combines BGE embeddings with the LLM pointwise scorer, obtained 0.705 PA-nDCG, well above all the purely visual or visual-text fusion runs we submitted: bge-large (0.493), google224 (0.513), BL-patch-2 (0.517), baseline (0.517), patch-14 (0.522), BL-patch-336 (0.526). This confirms, again, that the textual signal is dominant in this task and that LLM-based relevance estimation is a useful complement to embedding similarity. Third, the simple late-fusion strategy adopted for Subtask 2 did not translate into a clear improvement over the best text-only configuration: the multimodal variants we explored consistently underperformed BAAI-LLM, suggesting that the visual modality, although informative in principle, requires a more sophisticated integration —possibly per-article adaptive weighting, or fusion at the representation level rather than the score level— to actually contribute on top of strong text rankers. In the official rankings, the UCO-UC-CICESE-Plenitas team obtained the 12th position out of 13 in Subtask 1 with a score of 0.856 PA-nDCG (development phase), and the 13th position out of 21 teams in Subtask 2 with 0.705 PA-nDCG on the evaluation phase. The gap between our development result on Subtask 1 (0.876) and the official evaluation on Subtask 2 (0.705) also indicates a non-trivial distribution shift between the development and test splits, an aspect that deserves further analysis in a more controlled setting. Several directions for future work follow naturally from these findings. On the textual side, we plan to explore (i) listwise rather than pointwise LLM scoring, in which the model sees all ten candidates jointly and emits a full ranking in a single call, exploiting the calibration properties of large reasoning models such as Qwen3, gpt-oss, or DeepSeek-R1; (ii) ensembling across heterogeneous scorers via learning-to-rank meta-models trained on the development split; and (iii) the use of instruction-tuned bilingual models specifically adapted to Spanish news. On the multimodal side, the modest contribution of CLIP/SigLIP under simple late fusion motivates moving towards (i) image–article–headline triplet scoring rather than image–headline only, which would let the visual signal disambiguate genuinely ambiguous cases instead of competing with the text; (ii) replacing min-max normalisation with rank-based or temperature-scaled fusion to better calibrate the two modalities; and (iii) exploring vision-language LLMs (e.g., Qwen-VL, Llama 3.2 Vision) as a unified multimodal scorer in place of the CLIP-style dual encoder. We expect that these directions, together with a closer analysis of where the text-only model already places the correct headline at the top of the ranking, will allow the visual modality to contribute the complementary evidence it has been shown to carry in other multimodal news-understanding tasks.

Declaration on Generative AI

During the preparation of this work, the authors used Claude(Opus 4.7) and Grammarly in order to: Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. L. Bach, C. Kern, D. Bonnay, L. Kalaora, Understanding political news media consumption with digital trace data and natural language processing, Journal of the Royal Statistical Society: Series A (Statistics in Society) 185 (2022) S246–S269. doi:10.1111/rssa.12826.

[2] K. L. O'Halloran, G. Pal, M. Jin, Multimodal approach to analysing big social and news media data, Discourse, Context & Media 40 (2021) 100467. doi:10.1016/j.dcm.2021.100467.

[3] J. A. García-Díaz, S. M. Jiménez-Zafra, M.-T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022) 265–272.

[4] J. A. García-Díaz, S. M. Jiménez-Zafra, M.-T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES at IberLEF 2023: Political Ideology Detection in Spanish Texts, Procesamiento del Lenguaje Natural 71 (2023) 409–416.
[5] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sánchez, R. Valencia-García, Overview of EmoSPeech at IberLEF 2024: Multimodal Speech-Text Emotion Recognition in Spanish, Procesamiento del Lenguaje Natural 73 (2024) 359–368.
[6] R. Pan, J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, R. Valencia-García, Overview of SatiSPeech at IberLEF 2025: Multimodal Audio-Text Satire Classification in Spanish, Procesamiento del Lenguaje Natural 75 (2025) 513–522.
[7] 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 Classification of Headlines in Spanish Political News, Procesamiento del Lenguaje Natural 77 (2026).
[8] R. Pan, T. Bernal-Beltrán, J. Gómez-Navalón, J. A. García-Díaz, R. Valencia-García, PoliSUM-2025 Spanish: Evaluating Stylistic and Editorial Faithfulness in Abstractive Summarization of Political News, Procesamiento del Lenguaje Natural 76 (2026) 177–189.
[9] 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 Workshop Proceedings, CEUR-WS.org, 2026.
[10] S. C. Fanni, M. Febi, G. Aghakhanyan, E. Neri, Natural language processing, in: Introduction to Artificial Intelligence, Springer International Publishing, Cham, 2023, pp. 87–99.
[11] P. Shetty, R. Udhayakumar, A. Patil, M. Manwal, P. S. Vadar, Application of natural language processing (NLP) in machine learning, in: 2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE), IEEE, 2023, pp. 949–957.
[12] A. K. Joshi, Natural language processing, Science 253 (1991) 1242–1249.
[13] S. Wang, J. Jiang, Learning natural language inference with LSTM, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1442–1451. URL: https://aclanthology.org/N16-1170. doi:10.18653/v1/N16-1170.
[14] B. MacCartney, Natural Language Inference, Ph.D. thesis, Stanford University, Stanford, CA, 2009.
[15] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: International Conference on Machine Learning, PMLR, 2015, pp. 957–966.
[16] Q. Liu, M. J. Kusner, P. Blunsom, A survey on contextual embeddings, arXiv preprint arXiv:2003.07278 (2020).
[17] F. Almeida, G. Xexéo, Word embeddings: A survey, arXiv preprint arXiv:1901.09069 (2019).
[18] A. Bordes, J. Weston, R. Collobert, Y. Bengio, Learning structured embeddings of knowledge bases, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 25, 2011, pp. 301–306.
[19] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).
[20] A. Zubiaga, Natural language processing in the era of large language models, Frontiers in Artificial Intelligence 6 (2024) 1350306.
[21] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.

A. Online Resources

The results of the PoliticHeadlinES 2026 are available via:

- PoliticHeadlinES 2026,

- Task1 Results,
- Task2 Results,
- PoliticHeadlinES 2026 Code