UTP at PoliticHeadlinES - IberLEF 2026: A Ranking-Based Approach for Spanish Political News Headline Selection using Bi-Encoders and Cross-Encoders

Denis Cedeno-Moreno[1,*,†], Huriviades Calderon-Gomez[1,†], José Chiru-Figueroa[1,†] and Kexy Rodriguez-Martinez[1,†]

[1]*Universidad Tecnológica de Panamá, Ciudad de Panamá, Panamá*

Abstract

This article describes the participation of the UTP team in the PoliticHeadlinES 2026 shared task, which involved classifying Spanish political news headlines using multimodal data (text and images). Participants are given a news article and must classify a set of ten possible headlines, of which only one is the correct headline associated with that article; the other nine are distractors. We propose a ranking-based approach for the automatic selection of Spanish political news headlines in the PoliticHeadlinES shared task. Our method follows a two-stage architecture that combines bi-encoder models (BGE-M3 and Multilingual E5) for initial semantic retrieval with a fine tuning cross-encoder for supervised re-ranking. Final predictions are obtained through a weighted ensemble, where transformer model contributions are optimized via cross-validation. For the multimodal setting, we incorporate visual information using the Contrastive Language–Image Pre-training (CLIP), adding an image–headline similarity signal as a complementary feature. The system is evaluated using the Normalized Discounted Cumulative Gain at 10 (nDCG@10) metric, achieving a score of 73.94271% in the text-only configuration and 77.04902% in the multimodal setting, highlighting the contribution of visual information in improving ranking performance. Our team ranked 12th out of 19 participants overall, and outperformed the reference model in both tasks. However, the results demonstrate the potential of simple architectures to deliver consistent performance and high interpretability.

Keywords

Natural Language Processing, automatic classification of news headlines, transformers

1. Introduction

News is one of the fundamental bases for an informed and democratic society. It reflects current events immediately and its essential function is to provide citizens with reliable, truthful information about social, economic and political events, enabling them to stay informed about issues that affect their lives [1].

News has a profound impact on society, as it enables the formation of public opinion and is an essential tool for democracy. However, poor structuring or saturation of headlines can also generate anxiety and stress. Constant news consumption can influence mental health, political polarization and citizens' biased perception of reality [2].

Political news is a fundamental source of information for citizens. It influences public perception and decision-making, often generating debates that can cause confusion, distress, or a lack of empathy, particularly when false or biased information is involved. In this sense, news headlines are important because they provide a summary of the key information and act as the first point of contact, influencing the reader's opinion by maximizing visibility and virality. This crucial element determines whether or not the reader engages with the content of a news story [3].

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

denis.cedeno@utp.ac.pa/ (D. Cedeno-Moreno); huriviades.calderon@utp.ac.pa (H. Calderon-Gomez); jose.chiru@utp.ac.pa (J. Chiru-Figueroa); kexy.rodriguez@utp.ac.pa (K. Rodriguez-Martinez)

0000-0002-9640-1284 (D. Cedeno-Moreno); 0000-0002-6118-1154 (H. Calderon-Gomez); 0009-0002-1930-3573 (J. Chiru-Figueroa); 0000-0002-3399-0632 (K. Rodriguez-Martinez)

published 2026-09-29

The interpretive complexity of news increases when it is expressed through multimodal formats combining text and images. This requires models that can process both modalities together in order to effectively detect subtle clues [4].

However, automatically generating or selecting headlines can be challenging as they must be concise, faithful to the article's content, informative and engaging while avoiding a misleading or sensationalist approach. This makes the selection and classification of headlines a relevant problem for natural language processing (NLP), information retrieval (IR) and, increasingly, multimodal learning when articles are accompanied by images [5].

PoliticHeadlinES is a shared task of IberLEF 2026 focused on the ranking of Spanish political news headlines. Given a news article, participants must rank a set of ten candidate headlines, where only one corresponds to the original title and the remaining act as strong distractors. The task evaluates the ability of systems to identify the most relevant headline using both textual information and in the multimodal setting visual cues as well [6].

The first stage, a bi-encoder strategy based on BGE-M3 and Multilingual E5 Large instruct is used to independently encode the article body and candidate headlines. Semantic similarity is computed using cosine similarity to produce an initial relevance-based ranking. In the second stage, this ranking is refined using a cross-encoder model, fine tuning on article–headline pairs derived from the training data. Unlike the bi-encoder setup, the cross-encoder jointly processes each pair, enabling a more accurate estimation of relevance. This component serves as the primary scoring mechanism and substantially improves ranking quality through supervised re-ranking.

The remainder of the article is structured as follows. Section 2 presents some related work; Section 3 describes the dataset used, presents the task, and describes the dataset and evaluation metrics; Section 4 presents the multimodal classification methodology used; Section 5 presents the experimental results and performs an error analysis; and finally, Section 6 concludes the article and discusses future lines of research.

2. Related work

Misleading headlines create a discrepancy between the headline and the news content. A well-designed news headline should summarize the main information and be neutral. However, many of the headlines circulating online use false or distorted information in an attempt to confuse or deceive the reader. Automatically detecting and classifying multimodal news headlines in Spanish poses a complex challenge in NLP. Several studies have addressed this issue by employing linguistic, statistical, machine learning and deep learning methods, which have emphasized the complexity of the task due to the figurative and ambiguous nature of the process. Early approaches the automatic detection and classification of headlines used traditional machine learning methods with manually extracted linguistic features.

For example, in their work [7], presents a generic and flexible multilevel hierarchical classification, based on a two-stage approach that allows for the detection of the stance between the headline and the body of the news text. The proposal uses automatic summarization techniques. Its novelty lies in integrating different summarization approaches (extractive and abstractive) into a generic two-stage classification problem, where both the news headline and its corresponding body text are provided as input, and their relationship is obtained as output. They experimented with the Fake News Challenge (FNC-1) dataset, achieving an accuracy of 94.49%.

With the advent of deep learning, models have been combined using neural networks such as CNN, LSTM, and, more recently, transformers like BERT and RoBERTa, with improved results. As observed in the study [8], L3Cube-IndicNews offers three distinct datasets, designed to handle documents of varying lengths, which classify short news headlines. The datasets were evaluated in Indic languages using four different models: monolingual BERT, multilingual BERT for Indic sentences (IndicSBERT), and IndicBERT. With accuracy exceeding 80%. In addition, within the framework of the study, [9] conducted a classification task of Chinese news headlines based on the Bidirectional Encoder Representations from Transformers (BERT) model. The model was trained and tuned using a collected and organized dataset.

The results of the experiment show that the model has a better classification effect on news headlines, reflecting the advantages and performance of BERT in the text classification of Chinese news headlines. For the study conducted by [10] a specific sentiment analysis is proposed for news headlines related to the 2019 Argentine presidential elections. The research presents a polarity dataset comprising 1976 headlines that mention candidates at the objective level. In the experiment, state-of-the-art classification algorithms based on pre-trained linguistic models (BERTin, BETO, RoBERTa, and RoBERTuito) were used, demonstrating good results with Macro F1 scores of 75.3% and 75.0%, respectively. Meanwhile [11] in their research, evaluate the ability to capture salient features in short textual data. To do this, they systematically compare the performance of five widely used machine learning and deep learning models: logistic regression, Random Forest, Light Gradient Boosting Machine (LightGBM), ALBERT, and Gated Recurrent Units (GRU), on the WELFake dataset, using only news headlines as input. According to the experimental results, ALBERT, based on transformational models, outperforms all other models in both accuracy and retrieval. The ensemble models, particularly Random Forest and LightGBM, also offer solid performance and interpretability advantages for practical implementation. In [12] and others propose a new method for identifying keywords in news headlines and extending their features to the sentence and word levels. They use convolutional neural networks (CNNs) to extract and classify the features. The proposed model was tested on the Sogou News Corpus dataset, using the TextRank Composite Algorithm (TRLDA) LDA model and the enhanced term frequency-inverse document frequency (TF-IDF) algorithm. In the evaluation, it achieved an accuracy of 93.42%. Despite advances observed in this research, most has focused on the textual modality. Recent studies have begun to explore multimodal approaches that incorporate textual and visual information to improve the automatic detection and classification of headlines. However, significant challenges remain, allowing us to evaluate other models. In this context, the IberLEF 2026 shared task PoliticHeadlinES represents a significant step forward by providing a reference dataset for the multimodal detection of Spanish-language news headlines. Furthermore, it enables the systematic evaluation of models that process text and images together, fostering the development of robust and efficient detection systems [13].

3. Dataset Description

Participants were provided with the PoliticHeadlinES at IberLEF 2026 [14] dataset for training, testing and evaluation purposes [15]. This dataset consists of Spanish political news stories collected from digital media. Each instance of the dataset corresponds to a single news story and is associated with a set of headlines for that story. For each news story, ten potential headlines were provided. Only one of these is the original, correct headline; the remaining nine act as carefully selected distractors. These distractor headlines are thematically related to the news story in order to increase the difficulty of the task and prevent trivial solutions. For each news article, systems are provided with: (1) the full text of the news article; and (2) an optional associated image, depending on the subtask.

During an initial stage, we conducted preliminary experiments using the public development subset of the dataset, which includes approximately 100 instances (dev_public.csv). In the final setup, the full training dataset was employed, consisting of approximately 4,912 instances (train_corpora.csv). From this set, we randomly sampled 1,000 instances to construct a local evaluation subset used for validation and weight optimization. This strategy allows us to estimate model performance during development while preserving the majority of the data for training. Table 1 show the split the dataset.

Table 1
Split the dataset

Phase	Dataset	Instances
Development	dev_public	100
Evaluation	train_corpora	4,912

4. Methodology

Figure 1 shows the overall architecture of our approach for the two tasks in the shared PoliticHeadlinES [16]. Each process combines pre-trained feature representations, efficient models and robust deep learning techniques for NLP tasks, aiming to strike a balance between interpretability and competitive performance

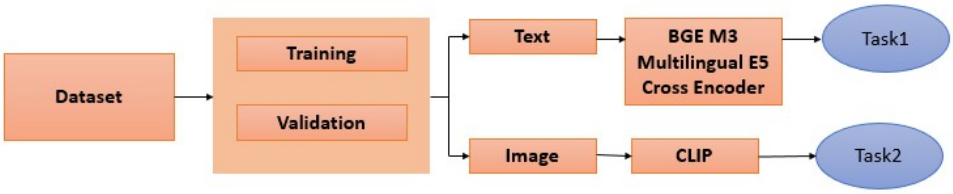

Figure 1: General system architecture for the PoliticHeadlinES tasks.

4.1. Task 1: Text-based Headline

In this subtask, the model was tasked with classifying the optimal headline from a set of ten candidate headlines, using only textual information. The input consisted of the text of the news article and the ten possible candidate headlines. This enabled us to evaluate the model's ability to capture semantic relevance, thematic coherence, and fidelity based solely on language comprehension.

We calculated similarities between the article body and candidate headlines using BI encoders. For this, we used the BGE M3 and Multilingual E5 large instruction text embedding models [17]. The resulting embeddings were compared using cosine similarity and used as the initial ranking signal [18].

A neural network model for NLP called Cross Encoder was applied. This model was trained using article-headline pairs from the provided training set. This model evaluates both texts together and was used to adjust the order obtained in the previous stage. Internal tests proved that this signal was the most relevant for the final decision [19].

4.2. Task 2: Multimodal Headline (Text + Image)

We built on Task 1 by incorporating visual information. Alongside the text of the news article and the candidate headlines, we provided an image associated with the article. Both textual and visual cues were used to generate the headline classification. This is the main subtask of the shared task, and the official classification is determined based on the results obtained in this context.

In Task 2, when an image was present, an additional score was added using CLIP (Contrastive Language-Image Pre-training), a model that connects natural language understanding with computer vision. This score, which was calculated based on the similarity between the image and each headline, was used solely to support the textual cues [20].

4.3. Ensemble and weight optimization strategy

The final ranking was obtained through a weighted combination of the different scoring components, including bi-encoder similarity scores, cross-encoder relevance predictions, and, for Task 2, image-text similarity scores derived from CLIP.

Formally, the final score for each candidate headline was computed as a linear combination of these signals, where each component contributes with a predefined weight. Since the contribution of each model may vary depending on the data distribution, the optimal weights were not fixed a priori.

To address this, we performed weight optimization using cross-validation on a held-out subset of the training data. This procedure allowed us to estimate the relative importance of each component by

maximizing the nDCG@10 metric, which directly reflects ranking quality. In the text-only configuration (Task 1), greater importance was assigned to the cross-encoder, as it provides fine-grained relevance estimation at the pair level. In contrast, for the multimodal configuration (Task 2), the contribution of the image-based component was incorporated in a complementary manner, ensuring that it enhances rather than dominates the textual signal.

This adaptive weighting mechanism enables the model to dynamically balance semantic retrieval, re-ranking precision, and multimodal alignment, resulting in a more robust and effective ensemble.

5. Results

We evaluated our models using the official validation split provided by the organizers of the PoliticHeadlinES 2026 shared task. Since the problem is formulated as a ranking task, system performance was assessed using the nDCG@10 metric. This metric is specifically designed to evaluate ranking quality, as it accounts for both the relevance of the predicted items and their positions within the output list.

Unlike traditional classification metrics such as precision or recall, nDCG provides a more appropriate evaluation in scenarios where the ordering of results is critical. Our approach integrates multiple components, including pre-trained BGE-M3 and Multilingual E5 embeddings for semantic retrieval, a fine-tuning Cross-Encoder for supervised re-ranking, and CLIP for incorporating visual information in the multimodal setting [21]. Table 2 shows the results of the model used in the split test for Task 1 and Task 2.

Table 2
Results of the model used in Task 1 and Task 2.

Phase	Model	Score
Task 1	Bi-encoder strategy	73.94271%
Task 2	CLIP	77.04902%

For Task1, which involves classifying news headlines from text, the model achieved a score of 73.94271%. This score indicates good accuracy and comprehensiveness, suggesting that the model is able to identify headlines from text with some effectiveness, although there is room for improvement [22].

When we compare this result with the official team rankings, we find that our team, 'UTP at PoliticHeadlinES', ranks twelfth out of nine or ten participants, with a score of 73.94271%. This indicates that our model's results are comparable to those of teams in the middle of the table. However, other teams obtained much higher scores, and our team surpassed the baseline [23].

For Task2, which involves classifying news headlines using a combination of images and text, the model achieved a score of 77.04902%. This score is slightly higher than that obtained in Task1, suggesting that the model performs better when both image and text information are included [24]. Table 3 shows the classification report for Task 1, while Table 4 shows the report for the combined image and text approach in Task 2.

Table 3
Task 1: Text-based Headline Ranking

Rank	Team	Score
1	Theron	95.43629%
2	padelice	92.75108%
3	VerbaNex AI Lab	92.22656%
-	–	–
12	UTP	73.94271%
19	baseline	51.65320%

Table 4
Task 2: Multimodal Headline (Text + Image)

Rank	Team	Score
1	Theron	95.43629%
2	padelice	92.75108%
3	VerbaNex AI Lab	92.62804%
-	–	–
12	UTP	77.04902%
19	baseline	51.34444%

Comparing this result with the scores of the teams in the official ranking, our team, "UTP at PoliticHeadlinES," also ranks 12th with a score of 77.04902%. This demonstrates, once again, that the model has achieved results comparable to those of other teams in the ranking, but there is still room for improvement to reach the highest scores and surpass the benchmark level. Overall, the results show that the model performs well in both tasks, but there is still room for improvement, especially in the classification of news headlines from text in Task 1.

6. Conclusion

In this work, we presented our approach to the PoliticHeadlinES 2026 shared task, addressing the problem of ranking candidate headlines for Spanish political news articles using both textual and multimodal information.

Our methodology was based on a two-stage ranking pipeline that combines bi-encoder models for initial retrieval and a fine tuning with cross-encoder for supervised re-ranking. This architecture allows the system to efficiently capture coarse semantic similarity while refining relevance estimation at a pairwise level. In addition, for the multimodal setting, we incorporated visual information using CLIP, enabling the integration of complementary cues derived from image–text alignment.

The system was evaluated using the nDCG@10 metric, which prioritizes the correct ranking of the relevant headline at top positions. The obtained results indicate that the proposed approach achieves stable performance, with the multimodal configuration consistently outperforming the text-only variant. This suggests that visual information can provide additional contextual signals, particularly in cases where textual differences between candidate headlines are subtle.

Overall, the results highlight the effectiveness of combining retrieval-based and re-ranking strategies within a unified framework, while maintaining a balance between computational efficiency and ranking accuracy. However, there remains a gap with respect to top-performing systems, indicating that more advanced learning strategies and fusion mechanisms could further improve performance.

As future work, we plan to explore more sophisticated multimodal fusion techniques, such as attention-based architectures, as well as improved negative sampling strategies for cross-encoder training to better capture hard-to-distinguish headline candidates.

Acknowledgments

The authors would like to express their gratitude for the support provided by the National Secretariat of Science, Technology and Innovation of Panama and the National Research System (SENACYT-SNI) in conducting this research. They would also like to express their gratitude to the Faculty of Computer Systems Engineering at the Technological University of Panama (UTP-FISC).

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL in order to Grammar and spelling check.

References

[1] P. Nickl, M. Moussaï and P. Lorenz-Spreen. (2025). The evolution of online news headlines. *Humanities and Social Sciences Communications*, 12(1), 1–13.

[2] I. Zulfikar, Y. Wibisono and A. Wahyudin. (2025). Intelligent News Aggregation System with Automatic Classification, Clustering, and Summarization. *Brilliance: Research of Artificial Intelligence*, 5(2), 689–696.

[3] P. Pal and D. Das. (2025). The evolution of online news headlines. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, 119–130.

[4] A. Atasoglu and Y. Taspinar. (2025). Embedding-Based Machine Learning Approach for Automatic Classification of Turkish News Articles. In *Proceedings of International Conference on Intelligent Systems and New Applications*, 3(5),103–108.

[5] J. Jim, M. Talukder, P. Malakar, M. Kabir, K. Nur and M. Mridha. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 6.

[6] D. Cedeño-Moreno,M. Vargas-lombardo and A. Delgado-Herrera. (2025). UTP at SatiSpeech IberLEF 2025: FastText and Wav2Vec2 for Satire Detection in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), colocated with the 41th Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*.

[7] R. Sepúlveda-Torres, M. Vicente, E. Saquete and E. Lloret. (2023). Data and Knowledge Engineering Leveraging relevant summarized information and multi-layer classification to generalize the detection of misleading headlines. *Data & Knowledge Engineering*, 145, 102176.

[8] A. Mirashi, S. Sonavane, and R. Joshi. (2023). L3Cube-IndicNews: News-based Short Text and Long Document Classification Datasets in Indic Languages. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, 442–449.

[9] Y. Chen (2024). A Study on News Headline Classification Based on BERT Modeling. In *Proceedings of the 2024 2nd international conference on image, algorithms and artificial intelligence (iciaai 2024)*, 345-355.

[10] F. Valentini, V. Cotik and J. Pérez. (2025). MessIRve: A Large-Scale Spanish Information Retrieval Dataset. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 27752–27769.

[11] S. Xu, Y. Tian, Y. Cao, Z. Wang, and Z. Wei. (2025). Benchmarking Machine Learning and Deep Learning Models for Fake News Detection Using News Headlines. In *5th International Conference on Advanced Algorithms and Neural Networks (AANN)*, 0–10.

[12] K. Miao, X. He, J. Yu, G. Wang and Y. Chen. (2023). A Study of Chinese News Headline Classification Based on Keyword Feature Expansion. *International Journal of Computational Intelligence Systems*, 16(1) 71.

[13] D. Cedeño-Moreno, M. Vargas-Lombardoand, C. Caparrós-Láiz and T. Bernal-Beltrán. (2024). UTP at EmoSPeech–IberLEF2024: Using Random Forest with FastText and Wav2Vec 2.0 for Emotion Detection. In *IberLEF@ SEPLN*, 2—7.

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

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

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

[17] A. Omotoso, H. Shopeju, A. Joshua and S. Oni. (2025). Improving BGE-M3 Multilingual Dense Embeddings for Nigerian Low Resource Languages. In *Proceedings of the 9th Widening NLP Workshop*, 224–229.

[18] S. Papadopoulos, M. Kula and G. Karantaidis. (2025). Multimodal and Multilingual Fact-Checked Article Retrieval. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 1, 1063–1071.

[19] S. Verma, F. Jiang and X. Xue. (2025). Beyond Retrieval: Ensembling Cross-Encoders and GPT Rerankers with LLMs for Biomedical QA. *arXiv preprint arXiv:2507.05577*, 4328.

[20] X. Li, C. Wen, Y. Huand and N. Zhou. (2023). RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124, 103497.

[21] N. Le, T. Mai, N. Hoang, P. Ho, and D. Bui. (2026). Systematic Evaluation of Modern Text Embeddings for Large-Scale Vietnamese News Clustering: E5-Large, BGE-M3, and SimCSE. *National Conference on Artificial Intelligence 2026 (FJCAI 2026)*, 1–8.

[22] L. Wang and N. Yang. (2024) Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

[23] J. Lei and X. Chen and N. Zhang. (2022) Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *arXiv preprint arXiv:2203.05465*.

[24] Q. Sun, Y. Fang, L. Wu , and X. Wang. (2023). EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389*.