I2C-UHU-PERSEUS at PoliticHeadlinES-IberLEF 2026: RoBERTa Optimization through Context Truncation

Jesús Tejón Carrillo[1,*,†], Enrique Lobo Rubio[1,†], Victoria Pachón Álvarez[1,†] and Jacinto Mata Vázquez[1,†]

[1]I2C Research Group, University of Huelva, Avda. de las Fuerzas Armadas, s/nº, 21007 Huelva, Spain

Abstract

This paper presents the system submitted by the I2C-UHU-PERSEUS team in the PoliticHeadlinES 2026 task, which focuses on the automatic ranking of headlines for political news in Spanish. To address this challenge, we propose a Transformer-based system built on RoBERTa, reformulating the ranking problem as a pointwise binary classification task between the article body and each candidate headline. To handle texts that exceed the model's input limit, we implement an asymmetric truncation strategy, designed to retain as much relevant context as possible within the 512-token limit. In the official evaluation, the system achieved a PA-nDCG score of 0.623, placing 18th overall. These results suggest that Transformer-based pointwise classification combined with task-specific truncation is a viable approach for headline ranking, although they also highlight generalization limitations and the need to better optimize the training process using specific ranking metrics.

Keywords

Natural Language Processing, Headline Selection, RoBERTa-bne, Asymmetric Truncation, Semantic Classification, PA-nDCG

1. Introduction

Automatic headline ranking is an important task in the field of Natural Language Processing (NLP) applied to the journalistic domain. In political news, this challenge is particularly complex since different candidate headlines often share entities, institutional actors, political parties, and topics, although not all of them reflect the specific content of a given news article with the same precision. Therefore, it is essential to model the semantic relationship between the news body and the headline, moving beyond surface-level lexical matching.

The PoliticHeadlinES 2026 task [1], within the context of IberLEF 2026 [2], formulates this problem as a ranking task: given the body of a political news article in Spanish, the system must rank ten candidate headlines according to their suitability for the article content. This formulation requires the system to assign comparable scores among candidates, as the official evaluation employs the PA-nDCG metric, which is designed to measure the quality of the generated ranking.

This paper describes the system submitted by the I2C-UHU-PERSEUS team, which is based on the pre-trained RoBERTa model [3]. The proposed system reformulates the ranking problem as an independent binary classification task (pointwise), evaluating each news-headline pair independently to maximize the model's discrimination capability. Subsequently, the probabilities assigned to the positive class are used to reconstruct the final ranking.

Given that news bodies can exceed the maximum input limit of conventional Transformer models (512 tokens) [4], it is necessary to define a context management strategy. Although there are architectures designed for long sequences such as Longformers [5], they involve higher computational costs. As an efficient alternative, this work proposes an asymmetric truncation strategy, inspired by the inverted pyramid structure commonly found in journalistic writing, where the core information is located at the beginning. The system preserves most of the initial part of the article and a smaller proportion of the

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

jesus.tejon@alu.uhu.es (J. T. Carrillo); enrique.rubio@alu.uhu.es (E. L. Rubio); vpachon@uhu.es (V. P. Álvarez); mata@uhu.es (J. M. Vázquez)

end (head+tail strategy), thereby ensuring that the token limit is not exceeded while minimizing the loss of lexical information.

The remainder of this paper is organized as follows: Section 2 reviews the state of the art and related work; Section 3 details the methodology and system description; Section 4 describes the experimental setup and hyperparameters; Section 5 analyzes the results obtained in the competition; and, finally, Section 6 presents the conclusions and lines for future work.

2. State of the Art

Transformer-based architectures have been shown to be effective for modeling semantic relationships between pairs of sequences. They have been successfully applied to classification, textual similarity, and ranking tasks [3]. In the context of news headline selection, these models are capable of estimating the semantic relatedness that exists between a news body and a headline. To address the ranking problem, pointwise approaches are commonly used as a simple formulation, reformulating ranking as a binary classification task where each news-headline pair is evaluated independently.

However, models such as BERT [6] or RoBERTa have an important practical limitation. The maximum number of input tokens is typically restricted to 512 [4]. This restriction is particularly relevant in the journalistic domain, where the length of news articles often exceeds this limit. To mitigate this problem, several strategies have been proposed, such as the use of sliding windows [7], hierarchical segmentation [8], or architectures specifically designed for long sequences, like Longformer [5], which allow processing contexts of up to 4096 tokens.

Within context management techniques, previous work suggests that truncation strategies can be effective when adapted to the structure of the document. In the journalistic field, many news articles tend to follow an inverted pyramid structure, where the most important information is typically presented at the beginning, followed by additional details and contextual development, and culminating in the final part, which usually provides additional context, reactions, or complementary details. Therefore, several studies indicate that head+tail truncation approaches (preserving the beginning and the end) generally outperform head-only approaches in long document classification tasks [7].

In this work, based on these observations, we choose to use the pre-trained RoBERTa-base-bne model [9] to capture the linguistic characteristics of Spanish. For context management, we propose an asymmetric truncation strategy that seeks to balance semantic richness with computational efficiency.

3. Methodology and System Description

3.1. System Overview

The proposed system follows a text-pair classification architecture using a pointwise approach. The process is divided into four main stages:

- **News-headline pair generation:** Independent instances are constructed by pairing the article body with each of the ten candidate headlines.
- **Context management via truncation:** Application of the head+tail strategy to adapt the text length to the 512-token limit.
- **Relevance estimation:** Use of the RoBERTa-base-bne Transformer model to estimate a relevance score for each pair.
- **Final ranking construction:** Candidates are ranked in descending order based on the probability assigned to the positive class.

Figure 1 illustrates the overall system pipeline.

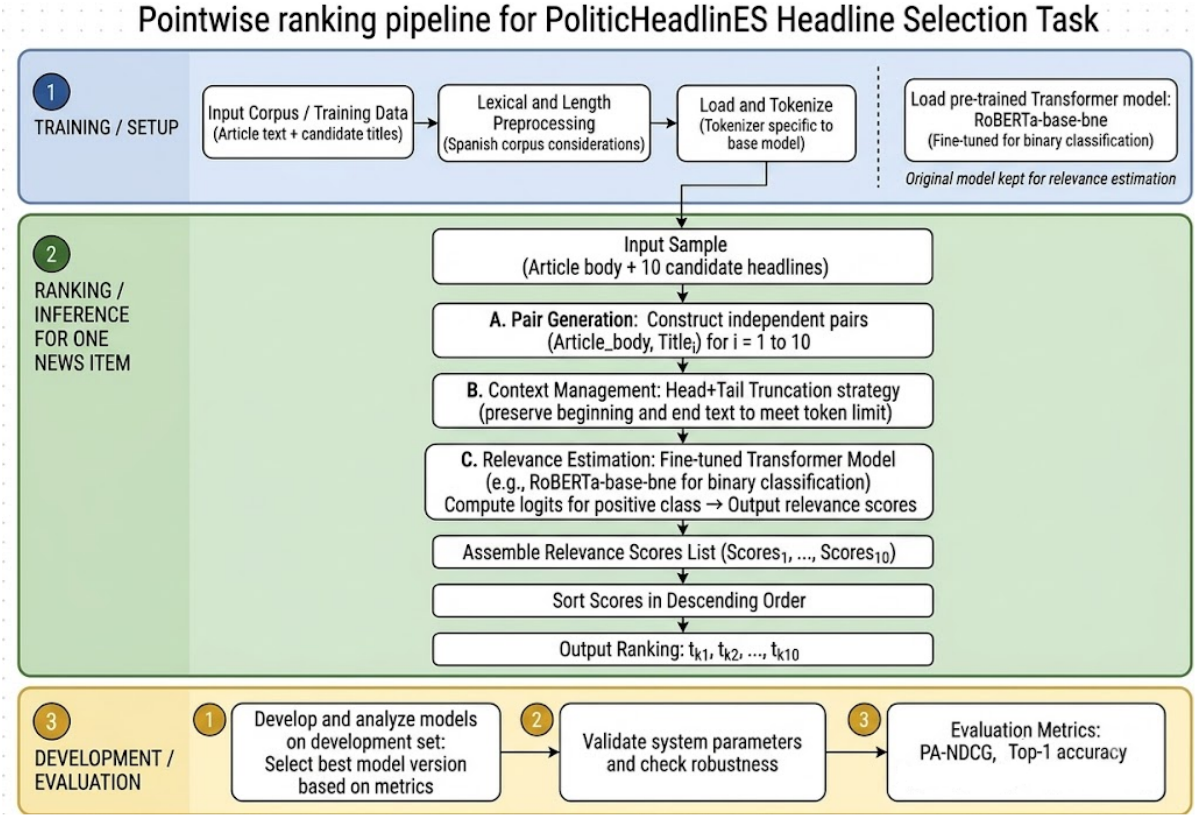

Figure 1: Diagram of the proposed system pipeline.

3.2. Pointwise Problem Reformulation

Although the official evaluation is based on $PA - nDCG$, the task is reformulated as an independent binary classification problem to simplify training and reduce the complexity of modeling the full candidate list jointly. For each news article in the training set, 10 instances are generated, consisting of the news body and a candidate headline.

- **Labeling:** The positive label (1) is assigned to the headline ranked first in the original ranking, and the negative label (0) is assigned to the remaining nine.
- **Limitation**: Although this approach simplifies training and optimizes the accuracy of the top position in the ranking, it entails a loss of information, as the model does not distinguish degrees of relevance among the remaining candidates.

3.3. Context Management: Head+Tail Truncation

We evaluated three truncation strategies for adapting the article body to RoBERTa's 512-token limit: (a) simple truncation, in which the first 512 tokens are selected; (b) symmetric head+tail truncation, in which the same number of tokens from the beginning and the end of the news article are selected; and (c) asymmetric head+tail truncation, selecting 80% from the beginning and 20% from the end of the news article.

Inspired by the "inverted pyramid" structure [10], where the core information resides at the beginning and additional context or reactions may appear toward the end, we apply the following strategy:

- Space is reserved for all tokens of the complete candidate headline.
- The remaining token space for the news body is divided into a proportion of 80% (beginning) and 20% (end).

- The central section of the news article is discarded, under the assumption that it is more likely to contain secondary details or extended quotations that could introduce semantic noise under the token limitation.

3.4. Architecture and Classification Model

After evaluating lexical baselines such as BM25 [11], which is limited by its reliance on exact lexical matching, and long-context architectures such as Longformer, we selected the RoBERTa-base-bne model. This model processes the input as a sequence pair and assigns a score through a classification head. The classification head produces two logits, corresponding to the positive and negative classes. During the inference phase, the score (probability) associated with the positive class is extracted. Finally, the 10 candidate headlines for each news item are sorted in descending order of the positive-class score, thereby generating the final ranking to be evaluated.

4. Experimental Setup

4.1. Data Preparation and Partitioning

For model training, the ranking task was transformed into a set of news-headline pairs. Given that each news item has ten candidates, ten instances were generated for each original article. The headline in the first position was labeled as the positive class (1), while the remaining nine were assigned the negative class (0), resulting in a dataset with an approximate imbalance of $1:9$.

The split into training, validation, and test sets was performed at the news item level, rather than by individual pairs, to prevent data leakage. This ensured that the same news article did not appear in more than one split. The dataset distribution after pair generation is summarized in Table 1.

Set	News Items	Pairs	Positive/Negative
Training	90	900	90 / 810
Validation	10	100	10 / 90
Test	50	500	50 / 450

Table 1
Dataset distribution after pointwise reformulation.

4.2. Preprocessing and Tokenization

For input preparation, the Transformers library [12] was used. For news articles longer than 350 words, an asymmetric truncation strategy (80/20 head+tail) was applied, retaining the first 280 words and the final 70 words of the article body. The two fragments were concatenated using the textual marker `[SALTO]`, designed to signal the sequential discontinuity to the model without introducing excessive structural noise. The following Spanish example illustrates the resulting truncated input:

"El Centro de Investigaciones Sociológicas (CIS), que dirige José Félix Tezanos, ha incorporado una novedad a sus estudios no electorales: formular a los ciudadanos una tanda de preguntas sobre voto, algo que según distintas fuentes del sector consultadas por este periódico no había sucedido antes. El resultado es que en esta legislatura se han realizado casi 89.000 preguntas de intención de voto, un caudal de información electoral que, si se suma a los barómetros específicos en este campo (55.000 desde las últimas generales), multiplica [...] sobre «Cultura y estilos de vida» que incluye 3.700 entrevistas y también preguntas de voto. No hay [SALTO] denomina 'cocina' de los datos, un ejercicio profesional necesario. «El dato en bruto es manipulación. ¿De verdad Tezanos no está haciendo el trabajo de cocina? Si es que no, ¿para qué hace las preguntas? Y [...] voto con estimación. Atentos."

This word-level limit was intended to keep the tokenized sequence within RoBERTa's 512-token limit.

4.3. Training and Hyperparameters

The model was trained using cross-entropy loss, employing a global random seed of 42 to ensure the reproducibility of the experiments.

Hyperparameter search was conducted using the Optuna framework [13]. Specifically, we ran a total of 10 optimization trials. The hyperparameter search space explored the following ranges:

- **Learning rate:** 5×10^{-6} to 4×10^{-5} .
- **Weight decay:** 0.0 to 0.1.
- **Warmup ratio:** 0.0 to 0.2.
- **Number of training epochs:** 3 to 10.

Although the final metric of the competition is based on ranking (PA-nDCG), the fine-tuning process was monitored using the F1-score on the validation set as the exact objective function to maximize. To ensure optimal checkpoint selection and prevent overfitting, models were evaluated and saved at the end of each epoch. We implemented an Early Stopping callback with a patience of 2 epochs, and the checkpoint corresponding to the highest validation F1-score was automatically restored at the end of the training process. The following hyperparameter values were selected for training the RoBERTa-base-bne model using the asymmetric truncation strategy (80/20), which yielded the best results during the development phase:

- **Learning rate**: 2.23×10^{-5}.
- **Batch size**: 8.
- **Weight decay**: 0.0259 (using the AdamW optimizer [14]).
- **Warmup ratio**: 0.0732.

The entire process was executed in a Google Colab environment using an NVIDIA T4 GPU. After training, the checkpoint from the best validation epoch was restored.

4.4. Post-processing and Ranking Construction

To comply with the competition format, a post-processing phase was conducted. The classification head produced two logits, corresponding to the positive and negative classes. A Softmax function was then applied to obtain class probabilities, and the probability assigned to the positive class was taken as an indicator of relevance. Finally, the ten candidate headlines for each news article were sorted in descending order based on this score to generate the final ranking.

5. Results

The evaluation of the system is divided into two phases: an internal validation analysis based on an ablation study to justify the design decisions, and the presentation of the official results obtained in the evaluation phase of the competition.

5.1. Ablation Study

To select the final configuration, several experiments were conducted during the development phase. Table 2 summarizes the performance of the different methodologies explored using the official $PA-nDCG$ metric.

The following key observations are drawn from the ablation study:

- **Baseline Systems**: Lexical matching-based approaches (BM25) showed limited performance, with the best BM25 variant reaching 0.2824. It was observed that reducing the context to 50 words slightly improved the result, suggesting that in purely lexical models, additional text may introduce noise.

Methodology / Configuration	$PA - nDCG$
BM25 / no limit	0.2019
BM25 / 200-word limit	0.2423
BM25 / 50-word limit	0.2824
RoBERTa-base-bne / Simple truncation	0.5146
RoBERTa-base-bne / Symmetric Head+Tail truncation (175+175)	0.7949
Longformer	0.4759
RoBERTa-large-bne / Symmetric Head+Tail truncation (175+175)	0.6542
RoBERTa-base-bne / Asymmetric truncation (80/20)	**0.9340**

Table 2
Results of the different methodologies in the development phase.

- **Impact of Truncation**: The transition from simple truncation to a symmetric head+tail truncation increased the $PA - nDCG$ from 0.5146 to 0.7949. This suggests the importance of retaining the final part of the article to capture information relevant to the headline that may appear near the end of the article.
- **Capacity Analysis**: Models with a larger number of parameters (RoBERTa-large) or architectures for long sequences (Longformer) did not outperform the base model in our tests. This may indicate that larger models require more careful tuning or more training data given the characteristics of the dataset and the training configuration employed.

To further analyze the model's behavior, an error analysis was performed on the 50 news articles from the test partition. The system successfully identified the correct headline in the first position in 41 cases, representing an accuracy of 82.00%. These results indicate that the pointwise approach is promising for top-1 identification of the main headline, suggesting that the approach is suitable as a baseline for the headline selection task.

5.2. Official Results

In the official evaluation phase, the system obtained a score of **0.623**, placing 18th in the overall classification. Table 3 shows the official results.

Rank	Participant	$PA - nDCG$
1	Theron	0.954
2	padelice	0.928
3	VerbaNex AI Lab	0.926
...	...	...
18	**i2c-uhu-perseus**	**0.623**
19	PUCE-UTP	0.555
...	...	...
21	mimic	0.348
-	*Overall Average*	*0.752*

Table 3
Summary of official results in the Test Phase.

5.3. Discussion and Analysis of the Performance Drop

A notable discrepancy is observed between the $PA - nDCG$ in the development phase (0.934) and in the evaluation phase (0.623). This drop suggests that, while the model performs well at identifying the top-ranked headline, it struggles to establish a precise hierarchical order among the remaining nine candidates.

By reformulating the problem as a binary classification, the system maximizes the probability of the correct headline, but treats the rest of the candidates as a uniform negative class. Consequently,

the model lacks information to rank the distractors among themselves according to their degree of semantic closeness, which is essential for optimizing the $PA - nDCG$ metric. This limitation, coupled with possible differences in lexical distribution between the training and test sets, may explain the performance reduction in the final phase of the competition.

6. Conclusions and Future Work

This paper has presented the participation of the I2C-UHU-PERSEUS team in the PoliticHeadlinES 2026 task. The proposed system utilizes the RoBERTa-base-bne Transformer model under a pointwise binary classification formulation, integrating an asymmetric truncation strategy (80/20) to manage the length of political news articles.

The results in the development phase were highly promising, achieving a $PA - nDCG$ of 0.9340 and a Top-1 accuracy of 82%. These data suggest that combining pre-trained models in Spanish with a head+tail preprocessing approach is effective for identifying the main headline. However, performance in the official evaluation dropped to 0.623, falling below the global average of 0.752.

This discrepancy highlights the limitations of the chosen approach. First, the pointwise reformulation simplifies training but does not explicitly model the ordering relationships among candidate headlines, which penalizes the $PA - nDCG$ metric. Second, the significant difference between validation during development and the official evaluation suggests a possible shift in lexical context, which the model failed to adjust to. Ultimately, the system proves to be a computationally efficient solution capable of identifying the correct headline, but it lacks the necessary robustness to optimize a complete ranking on unseen data.

As future work, the following lines of research are proposed:

- **Learning to Rank**: Replacing binary classification with pairwise or listwise loss functions that directly optimize the order of the entire list.
- **Qualitative Error Analysis**: Investigating whether the lowest-performing news articles correspond to those where key information is located in the central section discarded by the truncation.
- **Long Context Models**: Re-evaluating Longformer-type architectures or sliding window techniques with finer hyperparameter tuning to prevent the loss of textual information.
- **LLM Exploration**: Evaluating the use of large-scale generative models via prompting to compare their semantic reasoning capabilities against discriminative models.

Code Availability

The training notebook used in this work is available in the GitHub repository: https://github.com/JesusTejon/PoliticHeadlinES-2026.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3 for grammar and spelling checking, English translation, and the generation of Figure 1. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] 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), colocated with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[2] 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.
[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[5] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).
[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: 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), 2019, pp. 4171–4186.
[7] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China national conference on Chinese computational linguistics, Springer, 2019, pp. 194–206.
[8] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, N. Dehak, Hierarchical transformers for long document classification, in: 2019 IEEE automatic speech recognition and understanding workshop (ASRU), ieee, 2019, pp. 838–844.
[9] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pámies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Maria: Spanish language models, arXiv preprint arXiv:2107.07253 (2021).
[10] H. Pöttker, News and its communicative quality: the inverted pyramid—when and why did it appear?, Journalism Studies 4 (2003) 501–511.
[11] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at trec (1994).
[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.
[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.
[14] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).