I2C-UHU-Betelgeuse at PoliticHeadlinES-IberLEF 2026: A Hybrid System for Political Headline Ranking

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

[1]*I2C Research Group, University of Huelva, Spain*

Abstract

This paper presents a hybrid system for the headline ranking task proposed in PoliticHeadlinES-IberLEF 2026. The approach combines lexical signals derived from TF-IDF weighting and word overlap with semantic representations generated by bi-encoder neural models, enabling the system to capture both lexical matches and deeper semantic relationships between headlines and news bodies. The system was evaluated in Task 1 of PoliticHeadlinES-IberLEF 2026, which focuses on Spanish political news. Results show that combining lexical and semantic information improves performance compared to single-signal approaches. Specifically, the system achieved a paNDCG score of 0.976 on the development set and 0.656 on the evaluation set. This contrast reveals the impact of distribution shifts across datasets and underscores the challenges of generalising ranking models beyond the development distribution. Notably, lexical signals such as TF-IDF remain competitive under distribution shift, highlighting differences in robustness between lexical and semantic representations. Overall, the contribution of this work is mainly empirical and task-specific: it analyses the behaviour of lexical, semantic, and hybrid ranking signals in the PoliticHeadlinES-IberLEF 2026 setting, showing both the effectiveness of hybrid fusion under development conditions and the need for more robust weighting strategies when the evaluation distribution changes.

Keywords

Hybrid Ranking Pipeline, Semantic Similarity, Information Retrieval, Bi-encoder

1. Introduction

In recent years, natural language processing (NLP) has undergone substantial growth, driven by the development of neural models and, in particular, large pre-trained language models [1] [2]. Within this field, information retrieval and ranking tasks have become increasingly relevant, as they enable ranking documents according to their relevance to a given query. These tasks are essential in applications such as search engines, recommender systems, and textual content analysis.

In evaluation campaigns such as IberLEF, specific tasks are proposed to assess systems under realistic scenarios. In particular, the political headline ranking task poses the challenge of ordering a set of candidate headlines according to their semantic relationship with the remaining content of a news article. This problem is particularly challenging due to linguistic ambiguity, variations in writing style between headlines and article bodies, and differences between the development and evaluation datasets.

Traditionally, these tasks have been addressed using classical information retrieval techniques, such as TF-IDF or BM25 [3], which rely mainly on term matching. However, these approaches present important limitations, as they are unable to capture deeper semantic relationships between texts. By contrast, transformer-based neural models, such as BERT [4], have shown a strong ability to model language, although their standalone use does not always guarantee improved performance in scenarios characterised by high data variability.

This paper proposes a hybrid approach that combines lexical and semantic similarity signals for Task 1 of PoliticHeadlinES-IberLEF 2026. Rather than introducing a new ranking architecture, the main contribution of this work is empirical and task-specific: it analyses how TF-IDF, word overlap, and a

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

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

multilingual bi-encoder behave individually and in combination under development and evaluation conditions. This analysis is especially relevant because the final evaluation results reveal differences in robustness between sparse lexical signals and dense semantic representations.

2. Context

This work is framed within IberLEF 2026, an evaluation forum that proposes a range of tasks related to natural language processing for Spanish and other Iberian languages [5]. More specifically, it addresses Task 1 of PoliticHeadlinES-IberLEF 2026, which focuses on the ranking of Spanish political news headlines [6].

The task consists of ranking several candidate headlines according to their relevance to the content of a news article, using textual information only. Given the body of a news item, the system must determine which headlines are most relevant to it and produce a ranking based on their relevance [6].

One of the main challenges of this task is that headlines may express similar ideas using different wording, which limits the effectiveness of methods based solely on term matching. In addition, since the competition distinguishes between development and evaluation datasets, the system must be able to maintain its performance on previously unseen data. Therefore, the problem depends not only on the model employed, but also on its ability to generalise across variations in writing style, vocabulary, and the distribution of news articles.

It should be noted that this paper reports only the participation of the team in Task 1, which is based exclusively on textual information. No specific system was submitted for Task 2, where image information is also considered. Therefore, the methodology described in this paper, including the TF-IDF, word-overlap, and bi-encoder components, applies only to the text-based headline ranking scenario.

3. Related Work

Traditionally, text ranking tasks have been addressed using information retrieval techniques such as TF-IDF or BM25 [3], which are based on term frequency and term matching across documents. These methods represent texts according to the presence, frequency, or weighting of their terms, making them effective in scenarios where there is a high degree of lexical overlap between the texts being compared. Although these approaches predate current neural models, they remain relevant as strong baselines and as complementary signals in hybrid retrieval and ranking systems [7] [8].

However, methods based solely on term matching present clear limitations when texts convey the same idea using different vocabulary, synonyms, or distinct linguistic structures. To address this issue, the development of transformer-based language models has enabled the extraction of richer semantic representations of texts.

With the progress of language models, new techniques for semantic representation have emerged, including BERT-based models [4], which enable text embeddings to be obtained and compared within a vector space. Building on these models, approaches such as bi-encoders [9] have proven particularly useful for semantic similarity tasks.

Recent work has explored hybrid retrieval strategies that combine sparse and dense representations in order to improve ranking effectiveness across different retrieval scenarios [7] [8] [10]. These studies provide the basis for evaluating whether such combinations are also useful in the specific setting of political headline ranking.

4. Methodology

This paper proposes a hybrid approach to the ranking task, based on the combination of different similarity measures between the source text and the candidate headlines. The system pipeline is described below.

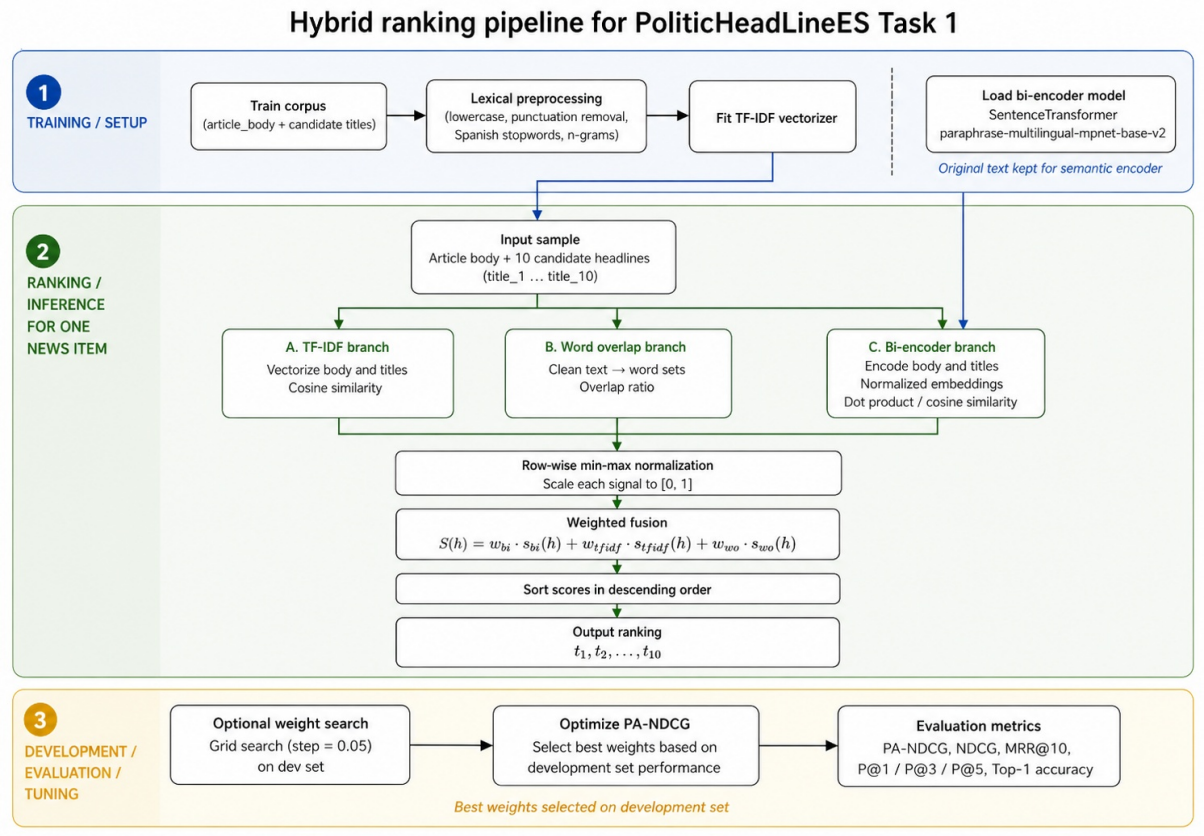

Figure 1: Overview of the proposed hybrid ranking pipeline. The system computes lexical and semantic similarity signals for each article–headline pair, normalises the scores, combines them through weighted fusion, and ranks the candidate headlines according to the final score.

4.1. Proposed Pipeline

The developed system follows an approach based on the combination of different similarity signals between the source text and the candidate headlines to capture both lexical and semantic relationships.

First, the system receives as input a main text together with a set of candidate headlines. From this information, text–headline pairs are generated and evaluated individually.

For each of these pairs, different similarity measures are computed. On the one hand, lexical-content-based techniques are employed, including TF-IDF adapted to this context and word-overlap measures, which identify direct lexical matches between texts. On the other hand, a bi-encoder model is used to transform the texts into vector representations and compute more complex semantic similarities, even when there are no explicit term matches.

Once all scores have been obtained, a fusion strategy is applied to combine the different signals into a single final score for each headline. This combination balances lexical and semantic information, taking advantage of the strengths of both approaches. An overview of the proposed hybrid ranking pipeline is shown in Figure 1.

4.2. Models Used

4.2.1. TF-IDF

As the first similarity signal, TF-IDF (Term Frequency–Inverse Document Frequency) was used, a classical information retrieval technique widely employed to represent texts according to the importance of their terms [11]. Although it is a traditional approach, lexical representations remain relevant in current retrieval and ranking tasks, particularly as robust baselines or as complementary signals to

dense embedding-based models [7].

This method estimates the importance of each term in the news article body with respect to the set of candidate headlines, giving greater weight to terms that are frequent in a given text but less common across the remaining texts. Based on this representation, the similarity between the source text and each candidate headline is computed using cosine similarity.

In this paper, the TF-IDF model was adapted through a specific preprocessing procedure, improving the quality of the textual representation. In particular, text normalisation was applied, including lowercasing, the removal of irrelevant characters, and stopword filtering, to reduce noise in the data. The specific configuration of the TF-IDF vectoriser, including the hyperparameters used, is detailed in the Experimental Setup section.

Although more advanced methods such as BM25 [12] are available, TF-IDF showed better behaviour during the development phase in this study, which motivated its selection as part of the final pipeline.

4.2.2. Word Overlap

As the second signal, a word-overlap measure was employed to capture direct matches between the source text and the candidate headlines.

In this case, the measure is computed from the set of terms present in both texts, estimating the degree of similarity between them. To this end, the texts were previously normalised through a process similar to that applied for TF-IDF, allowing terms to be compared more effectively.

Unlike TF-IDF, this signal relies solely on direct word matching, without applying any weighting scheme. It is therefore useful for reinforcing the presence of shared key terms between the source text and the candidate headlines, such as proper names, political parties, or relevant domain-specific concepts [13].

Furthermore, this measure has a low computational cost and provides a complementary signal to the previous ones, helping to improve system performance in cases where there is clear lexical overlap [14]. Nevertheless, since it depends exclusively on surface-level term matches, its standalone use remains limited when texts express the same idea using different vocabulary [15].

4.2.3. Bi-encoder

As the third similarity signal, a bi-encoder model was employed [9] to model semantic relationships between the body of the news article and the candidate headlines. In contrast to lexical techniques, this approach does not rely solely on direct word matching, but instead produces vector representations, or embeddings, that enable the meaning of texts to be compared.

In this paper, the *sentence-transformers/paraphrase-multilingual-mpnet-base-v2* model was used [9]. This is a multilingual model designed for semantic similarity tasks and sentence-level embedding generation. It is particularly suitable for the task addressed here, as it enables headlines and news bodies to be compared even when they express similar ideas using different words or linguistic structures. Unlike the lexical signals, this component uses the original text, without applying the more aggressive normalisation used for TF-IDF and word overlap. This decision was adopted because, on the development set, the model showed better behaviour when the full contextual information of the sequence was preserved, including word order, punctuation, and morphosyntactic variation.

During feature extraction, the body of the news article and each candidate headline are processed independently by the model, generating their corresponding dense vectors. The semantic proximity between the news article and each candidate headline is then computed using the dot product of their vectors. Since normalised embeddings are used, this is mathematically equivalent to cosine similarity.

This component provides the system with the ability to detect deeper semantic similarities that cannot be captured by the other methods. However, these models may also be affected by shifts in data distribution.

4.3. Combination Strategy

In order to integrate the three signals coherently, a combination strategy based on the weighted fusion of the scores obtained by each component of the system was defined.

Since each signal, namely TF-IDF, word overlap, and the bi-encoder, produces values within different ranges, min–max normalisation was first applied to the scores, as shown in Equation 1.

$$s'_i = \frac{s_i - \min(s)}{\max(s) - \min(s)} \tag{1}$$

where s_i represents the original score assigned by a given signal to a candidate headline, and s'_i denotes its normalised value. In Equation 1, $\min(s)$ and $\max(s)$ refer to the minimum and maximum scores produced by that signal for the set of candidate headlines associated with the same news article. This transformation rescales the scores to the interval $[0, 1]$, allowing all signals to contribute in a balanced manner to the final score and preventing any individual signal from dominating the result solely because of the scale of its values.

In the implementation, a special case was considered for situations in which all candidate headlines receive the same score from a given component. In such cases, the denominator of the min–max normalisation becomes zero. To avoid undefined values, the normalised score for that component is set to zero for all candidate headlines associated with that article. This means that, when a signal cannot discriminate between the candidates, it does not influence their relative ordering, and the final ranking depends on the remaining components.

Once the signals have been normalised, the final score for each headline is computed through a linear combination of the three similarity measures, as shown in Equation 2.

$$S(h) = w_{\text{bi}}\, s_{\text{bi}}(h) + w_{\text{tfidf}}\, s_{\text{tfidf}}(h) + w_{\text{wo}}\, s_{\text{wo}}(h) \tag{2}$$

where $S(h)$ represents the final score assigned to the candidate headline h, and $s_{\text{bi}}(h)$, $s_{\text{tfidf}}(h)$, and $s_{\text{wo}}(h)$ correspond to the normalised scores obtained from the bi-encoder, TF-IDF, and word-overlap components, respectively. In this way, each signal contributes complementary information: TF-IDF captures term relevance, word overlap reinforces exact matches, and the bi-encoder introduces the semantic component.

The weights associated with each signal were tuned using the development set. Specifically, the values of w_{bi}, w_{tfidf}, and w_{wo} were obtained through a grid search over weight combinations with a step size of 0.05, optimising the paNDCG metric on the development set.

Finally, the combination selected during the development phase was kept fixed during evaluation, thereby ensuring a fair evaluation and avoiding any specific adjustment to the evaluation set. The selected configuration assigns 70% of the weight to the bi-encoder component, 5% to the TF-IDF signal, and 25% to word overlap.

This configuration reflects the dominant role of semantic similarity during development, while preserving robustness through lexical signals. The high weight assigned to the bi-encoder indicates that dense semantic representations were especially useful for modelling headline–body relationships in the development distribution. However, the inclusion of TF-IDF and word overlap ensures that explicit lexical evidence, such as named entities, political actors, and domain-specific terms, remains represented in the final score. Therefore, the selected fusion strategy does not rely exclusively on semantic similarity, but incorporates lexical constraints that may improve robustness when the input distribution changes.

5. Experimental Setup

For the experimental setup, this implementation focuses exclusively on Task 1 of PoliticHeadlinES-IberLEF 2026. The development of the system relied on several Python libraries, including Pandas [16], NumPy [17], NLTK [18], scikit-learn [19], Sentence-Transformers [9], and Hugging Face Hub [20].

The experiments were conducted using the dataset released for Task 1 of PoliticHeadlinES-IberLEF 2026, which is associated with Spanish PoliSUM-2025, a resource designed to evaluate stylistic and editorial fidelity in Spanish political news summarisation [21]. Each instance consists of the body of a political news article and ten candidate headlines that must be ranked according to their relevance to the article content. Accordingly, the task is formulated as a ranking problem rather than as a traditional classification problem, with the system assigning a relevance score to each candidate headline and ordering them from highest to lowest score.

Dataset characteristic	Train public	Development	Evaluation
Number of instances	100	50	4912
Average article length (tokens)	507.86	542.42	848.70
Median article length (tokens)	416.50	522.50	757.50
Articles over 512 tokens (%)	43.00	52.00	76.43
Average headline length (tokens)	14.52	14.44	13.67
Vocabulary size	9944	6727	88639
Vocabulary overlap with development (%)	39.21	100.00	7.31
OOV vocabulary with respect to development (%)	60.79	0.00	92.69
Average article–headline lexical overlap (%)	25.93	25.85	27.31
Average unique tokens per article	229.10	238.24	369.53

Table 1
Comparison of selected textual characteristics across the public train, development, and evaluation datasets.

As shown in Table 1, the evaluation set differs substantially from the development set in several textual characteristics. Evaluation articles are considerably longer, with an average length of 848.70 tokens compared with 542.42 tokens in the development set. Moreover, 76.43% of the evaluation articles exceed the 512-token limit used by the bi-encoder, whereas this occurs in 52.00% of the development articles. This suggests that truncation may have affected the semantic component more strongly during evaluation, especially when relevant information appeared after the first 512 tokens.

The evaluation set also presents a much larger vocabulary size and a low vocabulary overlap with the development set. Only 7.31% of the evaluation vocabulary is covered by the development vocabulary, resulting in 92.69% out-of-vocabulary terms with respect to the development set. Although this figure is partly influenced by the larger size of the evaluation set, it still indicates a relevant lexical difference between both phases. By contrast, the average article–headline lexical overlap remains relatively similar, and even slightly higher in evaluation. This may help explain why TF-IDF remains competitive in the evaluation phase, while the bi-encoder suffers a larger performance drop.

Overall, these results support the interpretation that the evaluation phase introduced a distribution shift with respect to the development set. This shift is especially visible in article length, truncation frequency, vocabulary coverage, and the number of unique tokens per article. Therefore, the degradation observed in the final results cannot be attributed only to the ranking model itself, but also to differences in the textual properties of the datasets.

The development set was used to tune the system configuration and select the final combination of signals. In particular, this set enabled different weight configurations for TF-IDF, word overlap, and the bi-encoder to be evaluated through a grid search. Once the best-performing configuration on the development set had been selected, it was kept fixed during the evaluation phase, avoiding any specific adjustment to the evaluation set.

With regard to preprocessing, two differentiated strategies were applied depending on the type of signal used. For the lexical signals, namely TF-IDF and word overlap, the texts were normalised by converting them to lowercase, removing non-alphabetic characters, and filtering Spanish stopwords obtained from the NLTK (Natural Language Toolkit) module [18]. This process reduced textual noise and favoured more homogeneous comparisons between the body of the news article and the candidate headlines. By contrast, for the semantic signal based on the bi-encoder, the original text was preserved, since this configuration achieved better results on the development set by retaining relevant contextual information, such as punctuation, word order, and the full sentence structure.

For the TF-IDF signal, a vectoriser was configured with a maximum of 50,000 features (`max_features=50000`), using unigrams and bigrams (`ngram_range=(1,2)`). Frequency filters were also applied to remove poorly represented terms (`min_df=2`) and excessively frequent terms (`max_df=0.9`), together with sublinear term-frequency scaling (`sublinear_tf=True`) and L2 normalisation. This configuration enables both individual terms and relevant term combinations to be captured, while reducing noise in the representation.

For the semantic signal, the *sentence-transformers/paraphrase-multilingual-mpnet-base-v2* bi-encoder model was used [9], configured with a maximum sequence length of 512 tokens (`max_seq_length=512`). The generated embeddings were normalised (`normalize_embeddings=True`), allowing similarity to be computed using the dot product, which is equivalent to cosine similarity in this setting. Processing was carried out in batches, with a batch size of 64 examples, in order to improve computational efficiency. This strategy reduces the total inference time by exploiting the parallelisation of embedding computation, particularly when GPU acceleration is available. It also provides a suitable balance between computational efficiency and memory consumption, which is especially relevant when a large number of text–headline pairs must be processed.

When the article body exceeded the maximum sequence length supported by the bi-encoder configuration, the input was truncated by the tokenizer to the first 512 tokens. Therefore, the beginning of the article was retained, while information appearing after this limit was not included in the semantic embedding. This decision follows the default processing behaviour of the transformer-based encoder, but it may affect performance in cases where relevant information for distinguishing between candidate headlines appears near the end of the article. This limitation is particularly important for long political news articles and may partly explain the lower robustness of the semantic component in the evaluation phase.

Finally, the system was evaluated using paNDCG, the official metric defined by the organisers for Task 1 of PoliticHeadlinES-IberLEF 2026. This metric is based on NDCG, which is widely used in ranking tasks, but adapted to the competition scenario in order to assess the quality of the ordering generated by the system with respect to the reference ranking. The complete implementation used in this paper is available in the following repository: https://github.com/EnriqueLoboRubio/PoliticHeadLineEs-2026.

6. Results

This section presents the results obtained by the individual signals and by the combined system, using the final weights fixed during development: 70% for the bi-encoder, 5% for TF-IDF, and 25% for word overlap. The main metric used in the competition is paNDCG, a variant of NDCG (Normalized Discounted Cumulative Gain) adapted to the official PoliticHeadlinES-IberLEF 2026 task [6]. NDCG is commonly used in ranking tasks, as it enables the evaluation of not only whether relevant items appear in the ranked list, but also whether they are placed in the highest positions [22].

In general terms, NDCG is computed from DCG (Discounted Cumulative Gain), which accumulates the relevance of the ranked items while applying a logarithmic discount according to their position. Thus, errors made in the highest positions have a greater impact than those made in lower positions. DCG can be expressed as shown in Equation 3.

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \tag{3}$$

where rel_i represents the relevance of the item placed at position i. To normalise this value, it is divided by the ideal DCG, namely the value that would be obtained if the headlines were perfectly ordered according to the reference ranking, as shown in Equation 4.

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{4}$$

In the case of paNDCG, the official adaptation defined by the task organisers is used to evaluate the quality of the generated ranking with respect to the reference order. Therefore, this measure not only rewards placing the most appropriate headlines in high positions, but also penalises deviations from the reference order defined by the task. This makes it particularly suitable for PoliticHeadlinES-IberLEF 2026, where the objective is not to select a single correct headline, but to generate a complete ranking of candidate headlines.

In addition to paNDCG, several complementary metrics were considered in order to obtain a more complete view of the system's behaviour. Specifically, `standard NDCG` was used to analyse the overall ranking order, `MRR@10` to measure the position of the first relevant item, and `Top-1 accuracy` to determine whether the headline placed in the first position matched the most relevant reference headline. `Precision@k` was also used to assess the quality of the first positions of the ranking. In particular, `Precision@1`, `Precision@3`, and `Precision@5` measure the proportion of relevant headlines retrieved among the first 1, 3, and 5 positions, respectively, making it possible to analyse whether the system places relevant candidates near the top of the ranked list.

During the development phase, the combined model achieved the best performance in almost all metrics, supporting the usefulness of the hybrid approach. In the main metric, it reached a value of 0.98, outperforming the bi-encoder model (0.88), TF-IDF (0.83), and word overlap (0.77). This trend was also observed across the remaining metrics, although TF-IDF obtained the best individual result for `Precision@3`.

This result suggests that TF-IDF is particularly effective at identifying several highly relevant candidates among the first positions when there is strong lexical evidence between the article body and the candidate headlines. In political news, relevant headlines often share named entities, institutional terms, party names, or event-specific expressions with the article body. Therefore, although the hybrid model achieves better overall ranking quality according to paNDCG, the sparse lexical representation can be especially competitive in the top-three positions.

During the evaluation phase, a general drop in performance was observed across all signals, suggesting greater difficulty in generalising beyond the development set. However, the magnitude of this decrease was not the same for all approaches. The combined model decreased from 0.98 to 0.66, corresponding to a drop of 0.32 points, or approximately 32.7%. TF-IDF decreased from 0.83 to 0.64, representing a drop of 0.19 points, or 22.9%, while word overlap decreased from 0.77 to 0.60, a drop of 0.17 points, or 22.1%. By contrast, the bi-encoder showed the sharpest degradation, falling from 0.88 to 0.48, which corresponds to a drop of 0.40 points, or 45.5%. This indicates that, although all models were affected by the evaluation distribution, the semantic component was substantially more sensitive to this shift than the lexical signals.

Interestingly, TF-IDF achieves performance comparable to the hybrid model in the evaluation set, with only a small difference in paNDCG. This suggests that lexical signals are more robust under distribution shift than the semantic representation produced by the bi-encoder. In particular, while the bi-encoder contributes substantially to the hybrid model during development, its standalone performance decreases more sharply in evaluation, indicating higher sensitivity to variations in writing style, vocabulary, and headline formulation.

Table 2 summarises the metrics obtained on the development set. Table 3, in turn, reports the results for the evaluation set. Since, at this stage, the organisers only provide the official paNDCG metric, this table includes this measure exclusively.

In addition, the system's position in the official competition ranking was analysed for both phases, as shown in Tables 4 and 5. During the development phase, the system achieved a score of 0.976 in the main metric, ranking third among the participants and tied with the top-ranked systems. In the evaluation phase, however, the system obtained a score of 0.656, placing it fifteenth in the ranking. This phase corresponds to the scenario in which the system showed its lowest performance, as also reflected by the metrics presented above.

Metric	Combined	TF-IDF	Word Overlap	Bi-encoder
paNDCG	0.98	0.83	0.77	0.88
Top-1 accuracy	0.98	0.82	0.76	0.88
standard NDCG	0.98	0.97	0.96	0.97
MRR@10	0.99	0.90	0.87	0.94
Precision@1	0.98	0.82	0.76	0.88
Precision@3	0.82	0.86	0.79	0.79
Precision@5	0.76	0.75	0.71	0.77

Table 2
Performance comparison across individual signals and the hybrid model on the development dataset.

Metric	Combined	TF-IDF	Word Overlap	Bi-encoder
paNDCG	0.66	0.64	0.60	0.48

Table 3
Official paNDCG comparison across individual signals and the hybrid model on the evaluation dataset.

Participant	paNDCG	Ranking
padelice	0.976	1
VerbaNex AI Lab	0.976	2
I2C-UHU-Betelgeuse	**0.976**	**3**
...	...	...
PUCE-UTP	0.856	13
mean	0.934	–

Table 4
Official development results published by the task organisers, considering the official paNDCG score reported for Task 1.

Participant	paNDCG	Ranking
Theron	0.954	1
padelice	0.928	2
...	...	...
I2C-UHU-Betelgeuse	**0.656**	**15**
...	...	...
mimic	0.348	21
mean	0.7572	–

Table 5
Official evaluation results published by the task organisers, considering the official paNDCG score reported for Task 1.

7. Discussion

The results show a marked contrast between the development and evaluation phases, revealing a clear limitation in terms of generalisation. While the hybrid model achieved a paNDCG score of 0.976 in development, its score decreased to 0.656 in evaluation, representing a drop of 0.320 points, or approximately 32.8%. This decrease was also reflected in the official ranking, where the system moved from third position in development to fifteenth position in evaluation.

A plausible explanation for this behaviour is the presence of a distribution shift between the development and evaluation datasets. This interpretation is supported by the quantitative comparison shown in Table 1, where the evaluation set presents longer articles, a higher proportion of texts exceeding the 512-token limit, a larger vocabulary, and lower vocabulary overlap with the development set. However, since the official evaluation labels and full diagnostic information are limited, this explanation should be interpreted as an evidence-based hypothesis rather than as a definitive causal conclusion.

This suggests that lexical signals are more robust to distribution shifts, whereas the bi-encoder is

more sensitive to variations in writing style, vocabulary, and headline formulation. In this task, explicit lexical evidence, such as named entities, political actors, and domain-specific expressions, appears to remain informative even when the evaluation distribution differs from the development data. By contrast, the semantic representations produced by the bi-encoder may have adapted more strongly to the development distribution.

This behaviour indicates a potential overfitting of semantic representations to the development distribution. Although the bi-encoder improves the hybrid model during development, its lower standalone performance in evaluation suggests that dense semantic similarity should be combined with lexical constraints to avoid excessive dependence on a single representation space.

Likewise, combining signals using fixed weights appears to transfer this limitation partially to the hybrid model. Although optimising these weights enables performance to be maximised in a controlled setting, it may also reduce the robustness of the system when applied to unseen data. Consequently, future work should consider more adaptive fusion strategies, capable of adjusting to different input distributions.

A possible way to reduce this limitation would be to tune the fusion weights using more robust validation procedures. For example, cross-validation over different development splits or bootstrap-based resampling could provide a more stable estimate of the contribution of each signal. Another option would be to use a more conservative weighting strategy, assigning a lower weight to the bi-encoder and preserving a stronger lexical component. Although such strategies might slightly reduce the best development score, they could improve robustness when the evaluation distribution differs from the development data.

Overall, the results confirm that, in ranking tasks, achieving high performance under known conditions is not sufficient; it is equally necessary to ensure robustness to data variation. The combination of lexical and semantic signals remains a promising strategy, although its effectiveness depends on incorporating mechanisms that promote better generalisation in real-world scenarios.

8. Conclusions

This paper has presented the participation of the I2C-UHU-Betelgeuse team in Task 1 of PoliticHeadlinES-IberLEF 2026. The proposed system combines TF-IDF, word overlap, and a multilingual bi-encoder through a weighted linear fusion strategy optimised on the development set.

The system achieved strong results in the development phase, with a paNDCG score of 0.976, but its performance decreased to 0.656 in the official evaluation phase. This contrast shows that optimising a hybrid ranking system on a single development distribution may not be sufficient to ensure robust generalisation. The individual signal analysis indicates that TF-IDF remained highly competitive in evaluation, whereas the bi-encoder suffered a larger performance drop, suggesting that dense semantic representations were more sensitive to changes in the evaluation data.

Overall, the results support the usefulness of combining lexical and semantic signals for political headline ranking, while also highlighting the importance of designing fusion strategies with robustness in mind. Future work should explore more conservative or adaptive weighting methods, validation procedures such as cross-validation or bootstrapping, and strategies for handling long articles more effectively, especially when relevant information appears beyond the maximum input length of transformer-based encoders.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools to support the creation of the pipeline figure, adapt the manuscript to the required paper format, and improve grammar, punctuation, and vocabulary in English. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the paper.

References

[1] E. Cambria, B. White, Jumping nlp curves: A review of natural language processing research, IEEE Computational intelligence magazine 9 (2014) 48–57.

[2] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Computing Surveys 56 (2023) 1–40.

[3] D. Marwah, J. Beel, Term-recency for tf-idf, bm25 and use term weighting, in: Proceedings of the 8th International Workshop on Mining Scientific Publications, 2020, pp. 36–41.

[4] 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.

[5] 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-WS. org, 2026.

[6] 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 Headline Ranking in Spanish Political News , Procesamiento del Lenguaje Natural 77 (2026).

[7] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, arXiv preprint arXiv:2104.08663 (2021).

[8] D. Lee, S.-w. Hwang, K. Lee, S. Choi, S. Park, On complementarity objectives for hybrid retrieval, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 13357–13368.

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[10] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[11] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (1988) 513–523.

[12] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1994, pp. 232–241.

[13] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[14] Y. Luan, J. Eisenstein, K. Toutanova, M. Collins, Sparse, dense, and attentional representations for text retrieval, Transactions of the Association for Computational Linguistics 9 (2021) 329–345. URL: https://aclanthology.org/2021.tacl-1.20. doi:10.1162/tacl_a_00369.

[15] H. Schütze, C. D. Manning, P. Raghavan, Introduction to information retrieval, volume 39, Cambridge University Press Cambridge, 2008.

[16] W. McKinney, et al., pandas: a foundational python library for data analysis and statistics, Python for high performance and scientific computing 14 (2011) 1–9.

[17] C. R. Harris, et al., Array programming with numpy, Nature 585 (2020) 357–362.

[18] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, Inc., 2009.

[19] F. Pedregosa, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[20] H. Face, Hugging face hub, 2023. URL: https://huggingface.co.
[21] 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.
[22] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems 20 (2002) 422–446.