Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 177-189 recibido 05-12-2025 revisado 03-02-2026 aceptado 19-02-2026

Spanish PoliSUM-2025: Evaluating Stylistic and Editorial Fidelity in Abstractive Summarization of Political News

Spanish PoliSUM-2025: Evaluación de la Fidelidad Estilística y Editorial en la Generación de Resúmenes Abstractivos de Noticias Políticas

Ronghao Pan, Tomás Bernal-Beltrán, Jorge Gómez-Navalón, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Murcia, España
{ronghao.pan, tomas.bernalb, jorge.gomezn, joseantonio.garcia8, valencia}@um.es

Abstract: Automatic summarization of political news requires preserving factual content, journalistic style, and editorial orientation. However, research in Spanish remains limited due to the lack of ideologically diverse resources and outlet-sensitive evaluations. We present Spanish PoliSUM-2025, a corpus of 94,832 Spanish political news articles from multiple outlets with different editorial viewpoints and their original summaries. Using this resource, we compare mBART with three instruction-tuned large language models. Results show that instruction-tuned models consistently outperform the encoder–decoder baseline and that model size is the main determinant of quality. Linguistic analysis indicates that LLMs preserve semantic content but diverge from the concise, information-dense journalistic style. Substantial outlet-level variability appears in ROUGE-L but not in BERTScore, while ideological orientation produces no detectable differences under these metrics.
Keywords: Abstractive summarization, political news, large language models, editorial fidelity.

Resumen: La generación automática de resúmenes políticos requiere preservar el contenido factual, el estilo periodístico y la orientación editorial de la fuente. Sin embargo, la investigación en español es limitada por la falta de recursos ideológicamente diversos y de evaluaciones sensibles a la fuente. Presentamos Spanish PoliSUM-2025, un corpus de 94.832 noticias políticas en español procedentes de múltiples medios con distintas líneas editoriales. Con este recurso comparamos mBART con tres modelos ajustados mediante instrucciones. Los resultados muestran que los modelos instruccionales superan de forma consistente al modelo base y que el tamaño es el principal determinante de la calidad. El análisis lingüístico revela que los LLM preservan el contenido semántico, pero se alejan del estilo periodístico conciso y denso en información. La variabilidad entre medios es notable según ROUGE-L pero mínima según BERTScore, mientras que la orientación ideológica no produce diferencias detectables bajo estas métricas.
Palabras clave: Resumen abstractivo, noticias políticas, modelos de lenguaje de gran tamaño, fidelidad editorial.

1 Introduction

Automatic summarization has emerged as a critical natural language processing (NLP) task in fields with high volumes of information and critical timelines. Summarization systems aim to support human decision-making by automatically condensing long documents into concise representations. These systems are useful in domains such as journalism, analysis, and information monitoring.

Recent advances in large language models (LLMs), including multilingual encoder–decoder architectures such as mBART(Liu et al., 2020) and instruction-tuned models such as Llama (Grattafiori et al., 2024), Gemma (Team et al., 2025), and Qwen (Yang et al., 2025) have notably improved systems' ability to understand, select, and condense complex information

ISSN 1135-5948 DOI 10.26342/2026-76-13

into short, coherent, and useful texts. These developments have sparked increasing interest in applying summarization technologies to real editorial workflows to assist journalists, analysts, and readers in processing the growing amount of political information.

Among the different application domains of automatic summarization, political news stands out due to both its high volume and its societal impact. Political journalism embeds factual content together with narrative framing, evaluative tone, and ideological cues that vary systematically across media outlets (Hamborg, Donnay, and Gipp, 2019). Previous studies in media studies and computational journalism has shown that summaries can influence readers' perceptions of political events, and that automatic systems may unintentionally alter the original ideological framing (Deas and McKeown, 2025). Studies on LLM bias further indicate that generative models may introduce stylistic or semantic drift, amplifying or attenuating partisan signals depending on the model and context (Fisher et al., 2025). These observations raise important questions about the reliability of automatic summaries in politically sensitive environments. As a result, summarizing political content introduces challenges that go well beyond length reduction.

Despite recent advances in summarization in politics, rigorous evaluation of systems applied to the Spanish-language political domain remains limited (Helwe, Balalau, and Ceolin, 2025). Moreover, most studies rely on surface-level metrics such as ROUGE or BERTScore, that fail to capture essential dimensions for political journalism, including fidelity to the outlet's narrative style, preservation of its editorial perspective, and stability of ideological orientation during the condensation process. These limitations hinder a deeper understanding of how such systems actually behave in scenarios with high informational impact (Min et al., 2025).

To address these gaps, we compiled the Spanish PoliSUM-2025 corpus, a new collection of political news articles paired with their original, journalist-written summaries. These summaries are annotated with outlet-level editorial orientation. Spanning multiple media outlets with different editorial profiles, this resource provides a realistic basis for studying how summarization systems behave in diverse journalistic environments. Using this corpus, we compared a standard encoder-decoder baseline based on mBART with three instruction-tuned LLMs and evaluated their performance using standard lexical and semantic metrics, as well as a detailed linguistic analysis that addressed the limitations of surface-level metrics highlighted above. Specifically, we examined differences in the linguistic and structural properties of summaries generated by models, including syntactic complexity, lexical diversity, entity retention, and part-of-speech (PoS) features. We also examined how these properties varied across model families and scales.

The following research questions guide this study: **RQ1**: How do instruction-tuned LLMs differ in the linguistic and structural properties of the summaries they generate?; **RQ2**: How do the structural properties of political news articles influence content selection and information retention in automatic summaries?; and **RQ3**: How do the outputs of instruction-tuned LLMs with various editorial orientations differ when evaluated using lexical and semantic metrics?

2 State-of-the-Art

The field of automatic text summarization has evolved through several methodological paradigms, progressing from early neural models to advanced systems powered by large-scale pretraining.

2.1 Neural Abstractive Summarization

Abstractive summarization generates concise texts that paraphrase and compress source content instead of extracting spans verbatim. Early approaches relied on statistical (Mihalcea and Tarau, 2004) (Carbonell and Goldstein, 1998) or rule-based techniques (Kryściński et al., 2020), but performance improved substantially with neural sequence-to-sequence architectures that incorporate attention mechanisms (Bahdanau, Cho, and Bengio, 2014; Rush, Chopra, and Weston, 2015).

More recently, pretrained encoder–decoder models such as BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2019), and T5 (Raffel et al., 2019) have achieved state-of-the-art results in text generation by leveraging large-scale pretraining objectives. These models provide strong generative priors that are particularly

effective for abstractive summarization.

2.2 Large Language Models and Instruction Tuning

Instruction-tuned LLMs such as LLaMA-3 (Grattafiori et al., 2024), Gemma-3 (Team et al., 2025), and Qwen-3 (Yang et al., 2025), have demonstrated robust strong multilingual generation capabilities with limited task-specific fine-tuning. These models often outperform traditional seq2seq architectures in abstractive summarization (Elatiky et al., 2026), though concerns remain about factual consistency and hallucinations, especially in politically sensitive contexts. Despite these advances, however, research on instruction-tuned summarization in Spanish political journalism remains limited, with most prior work focusing on general news domains.

2.3 Summarization of Political and Ideological Content

Political texts present unique challenges for summarization due to their framing, editorial style, and ideological cues, which systems may distort or suppress (Hamborg, Donnay, and Gipp, 2019). Previous research in English indicates that automatic summarizers can alter sentiment, change framing, or perpetuate biases in the training data (Deas and McKeown, 2025), while large-scale LLMs may reflect political biases inherent in web corpora (Fisher et al., 2025).

However, research on political summarization in Spanish remains limited, especially in contexts with diverse editorial traditions and ideological orientations. Existing corpora rarely support systematic, cross-outlet analyses, leaving open questions about whether models preserve or reshape editorial framing. To address this gap, we have assembled a corpus spanning national, regional, and digital media outlets, and we have conducted an initial analysis of summarization behavior across ideological groups. Ideological labels are estimated based on a qualitative analysis of editorial content and should be interpreted as heuristic rather than definitive annotations.

2.4 Multilingual and Spanish Summarization

Multilingual pretrained models such as mBART (Liu et al., 2020), mT5 (Xue et al., 2020), and XLM-R (Conneau et al., 2019) have enabled effective cross-lingual transfer for summarization.

Previous studies on Spanish summarization (García-Ferrero and Altuna, 2024; Li et al., 2024) show that multilingual pretraining can produce high-quality summaries even with limited task-specific data.

Widely used Spanish datasets such as MLSUM (Scialom et al., 2020), Wikilingua (Ladhak et al., 2020), and NoticIA (García-Ferrero and Altuna, 2024) lack editorial diversity, ideological annotations, and political specificity. Consequently, these datasets provide limited insight into how models behave across heterogeneous journalistic styles or politically sensitive content.

3 Spanish PoliSUM-2025 Corpus

The Spanish PoliSUM-2025 corpus was created to enable the systematic evaluation of automatic abstractive summarization models in Spanish political journalism. It contains full news articles alongside their original, journalist-written summaries, which are annotated according to the editorial orientation of the news outlet.

The corpus is gathered through a multistep pipeline designed to systematically collect and filter political news articles. Article selection follows a set of explicit criteria: only political news items are considered, and articles must include valid embedded JSON-LD metadata. To ensure that the summaries are genuine abstractive condensations, we verified that they are neither identical to nor fully contained within the article body. The corpus includes media outlets with diverse editorial orientations, which is essential for analyzing outlet-level variability and ideological stability. No artificial balancing across outlets or ideologies was enforced to preserve the natural distribution of the Spanish media ecosystem.

We developed a dedicated crawler that can extract JSON-LD metadata from article pages on several Spanish news outlets, including *El País*, *ABC.es*, *La Razón*, *ElDiario.es*, *20minutos*, and *La Vanguardia*. For each URL, the crawler downloaded the JSON-LD data of `Articles`, `NewsArticles`, `ReportageNewsArticles`, and `OpinionNewsArticles`. It then retrieved fields including the publication date, headline, article body, and outlet informa-

```
{
  "@type": "NewsArticle",
  "headline": "El Supremo condena a
      Matas a devolver 1,2 millones",
  "author": "infoLibre",
  "datePublished": "2019-06-18",
  "articleSection": "Politica",
  "description": "Condena por
      prevaricacion y trafico de
      influencias.",
  "keywords": ["Politica", "Corrupcion",
       "Tribunal Supremo"],
  "articleBody": "El Supremo condena a
      Jaume Matas por irregularidades en
      la contratacion de Calatrava"
}
```


Figure 1: Example of an item of the dataset.

tion. In this taxonomy, `NewsArticle` typically denotes factual reporting, `OpinionNewsArticle` refers to editorial or opinion pieces, and `ReportageNewsArticle` captures more in-depth journalistic reports. We treated all of these as political news content and did not apply differential processing based on schema type. Additional preprocessing steps included deduplication, boilerplate removal, text normalization, and consistent encoding of whitespace and newline structure.

For each article, we extracted the gold summary from the description or related metadata fields. Since some news outlets use the first paragraph as a lead, we verified that the summary did not match the article body exactly or contain it entirely. This ensured that the summaries functioned as genuine abstractive condensations rather than repeated introductions.

The final corpus contains 94,832 political news articles from 2020 to 2024. It preserves the natural editorial imbalance of the Spanish media landscape, with left-leaning outlets accounting for approximately half of the dataset, right-leaning outlets accounting for slightly over 40%, and neutral outlets accounting for a smaller proportion. It is divided into 76,352 training articles, 8,484 validation articles, and 10,000 test articles. This corresponds to an approximate ratio of 80–10–10.Figure 1 shows a simplified example of the extracted JSON-LD structure.

We performed a detailed linguistic analysis of both articles and summaries using the UMUTextStats tool (García-Díaz et al., 2022). Political news articles tend to be relatively long (mean = 731 words) and syntactically complex (36 words per sentence on average), with high lexical diversity (TTR = 0.47) and low readability (Inflesz = 27.4). Discourse markers appear in 80% of texts. Verb usage patterns further reflect the evaluative and predictive nature of political reporting, with widespread use of the subjunctive, conditional, and future indicative. In contrast, gold summaries are concise (mean = 24 words), typically single-sentence, and lexically dense (TTR = 0.88). Their readability (Inflesz = 40) indicates moderate difficulty due to the high concentration of information.

It should be noted that we release a public version of the corpus that contains only the original article URLs and their corresponding data splits (training, validation, and testing). This allows other researchers to reconstruct the dataset by re-crawling the original sources. Therefore, the complete version of the corpus, including article texts and summaries, is maintained privately.

4 Methods and Experiments

This section describes the training approaches used to adapt different language models for the abstractive summarization of Spanish political news. We use mBART as our encoder-decoder baseline because seq2seq architectures are traditionally well-suited for summarization. These architectures provide explicit encoder representations of long inputs, stable decoding dynamics, and strong performance on multilingual summarization benchmarks (Hasan et al., 2021). In principle, these architectures should, in principle, be highly effective at condensing structured journalistic prose. However, recent instruction-tuned LLMs (Fetahu et al., 2023) have demonstrated strong zero-shot and fine-tuned performance across a wide range of generation tasks, including summarization, despite not relying on an explicit encoder–decoder separation. This discrepancy motivates a direct comparison between the encoder–decoder baseline and instruction-tuned LLMs to evaluate the performance of these different modeling approaches when applied to politically complex content.

4.1 Metrics

To evaluate the quality of the generated summaries, we use standard practices in automatic summarization to combine lexical overlap and semantic similarity metrics. First, we report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L_{sum}, which are computed using the official evaluate implementation with stemming enabled. These metrics serve as indicators of lexical fidelity to the gold references. Second, to capture semantic alignment beyond surface-level overlap, we use the BERTScore F1 metric computed with the Spanish `BETO` model (Cañete et al., 2020). Third, we include a hybrid indicator, which is the mean of ROUGE-L and BERTScore F1. This provides a balanced perspective when models differ in lexical matching or semantic coherence. Finally, to better characterize summarization behavior, we compute the predicted and reference lengths, the word-level compression ratios, and the proportion of empty predictions. These metrics help identify tendencies toward overgeneration, undergeneration, or deviations from the expected compression patterns in journalistic summaries.

4.2 Encoder–Decoder Baseline: Fine-Tuning mBART

We use mBART as our baseline encoder-decoder model because these models offer explicit representations of long encoders and stable decoding mechanisms, making them well-suited for condensing structured journalistic prose. We fine-tuned mBART[1] for monolingual Spanish summarization by setting `es_XX` as both the source and target language.

Each article–summary pair is tokenized and truncated to a maximum of 1,024 tokens for the article body and 256 tokens for the summary. Fine-tuning is performed with the Hugging Face `Seq2SeqTrainer` with beam search decoding for evaluation. Our primary metrics are ROUGE-1, ROUGE-2, and ROUGE-L, and we select the best checkpoint based on ROUGE-L to ensure a robust and easily comparable baseline.

We train the model for five epochs with a learning rate of 2×10^{-5}, weight decay of 0.01, and a per-device batch size of 2. We enable FP16 mixed precision, use a fixed random seed (42), and set a maximum generation length of 256 tokens.

[1]https://huggingface.co/facebook/mbart-large-50

4.3 Instruction-Based Supervised Fine-Tuning

To assess how instruction-based generative models learn to summarize political news, we fine-tuned three instruction-tuned causal LLMs. These models have demonstrated strong performance on a wide range of generation tasks due to their ability to follow natural language instructions and model long-range dependencies. This makes them promising candidates for abstractive summarization.

Each training instance is reformulated into an instruction consisting of an explicit task description and the full article as input and the reference summary as output. The prompts follow a structured template with `Question`, `Input`, and `Response` fields, which stabilizes the behavior of following instructions across model families.

Fine-tuning is performed using the `SFTTrainer` from the `TRL` library, and validation is based on ROUGE scores computed on a held-out subset. Summaries are generated with constrained decoding and evaluated with the same metrics as the mBART baseline, ensuring direct comparability.

We train instruction-tuned models for one epoch with a learning rate of 2×10^{-5}, a per-device batch size of 2, and gradient accumulation of two steps to accommodate long inputs. We apply weight decay of 0.05, 100 warm-up steps, and gradient checkpointing to improve memory efficiency. Generation uses beam search with a maximum output length ranging from 256 to 1,024 tokens depending on the model.

As shown in Figure 2, our instruction template structures each training instance into a Question–Input–Response format.

We evaluated three families of recent instruction-tuned LLMs: LLaMA 3 (`Llama-3.2-1B` and `Llama-3.1-8B`), Gemma 3 (`gemma-3-1b-pt`), and Qwen 3 (`Qwen3-4B` and `Qwen3-8B`). These models span different parameter scales and have demonstrated strong multilingual generation capabilities, making them ideal for assessing summarization performance across various model sizes.

```
### Question:
summarize the following text in Spanish
    clearly and concisely.

### Input:
El presidente del Gobierno anunció hoy
    una nueva propuesta de reforma
    fiscal que será debatida en el
    Congreso durante las próximas
    semanas. La oposición criticó la
    medida por considerarla insuficiente
    ...

### Response:
El Gobierno presentó una reforma fiscal
    que será debatida en el Congreso,
    mientras la oposición la considera
    insuficiente.
```

Figure 2: Example of the instruction-based prompt template used for supervised fine-tuning of LLMs.

5 Results

This section presents the quantitative and qualitative results from the evaluation of mBART and instruction-tuned LLMs.

5.1 Overall Performance

Table 1 summarizes the performance of all models. Contrary to expectations for encoder-decoder architectures, the mBART baseline substantially underperforms all instruction-tuned causal LLMs across every metric, including ROUGE and BERTScore. These results suggest that causal models fine-tuned with instructions are considerably more effective than traditional seq2seq approaches in the political news domain.

Among LLMs, performance consistently increases with model size. Larger models achieve higher ROUGE-1, ROUGE-2, and ROUGE-L scores, which confirms the importance of parameter scale for summarizing long documents. Llama-3.1-8B achieves the highest overall performance, followed closely by Qwen3-8B. Both models also achieve the strongest hybrid scores.

Although instruction-tuned models generate semantically faithful summaries, as reflected by BERTScore values around 75–76, they exhibit greater lexical variability compared to human references. This contributes to their moderate ROUGE-2 scores.

Overall, Llama-3.1-8B is the best-performing model, achieving the highest metrics in ROUGE-1 (39.91), ROUGE-2 (24.71), and ROUGE-L (34.12), as well as the highest hybrid score (55.04). As expected for a compact model with limited representational capacity, the smaller Llama-3.2-1B performs notably worse. It has the lowest ROUGE-2 score (22.68) of all the models, indicating its difficulty in preserving fine-grained relational information.

Gemma-3-1B is a strong competitor for its size, surpassing Llama-3.2-1B on all ROUGE metrics and compression stability. Its hybrid score of 53.46 shows that instruction tuning compensates for its smaller capacity, making it a strong, lightweight option.

Both Qwen models deliver strong, stable results. The Qwen3-8B model approaches the performance of the Llama-3.1-8B model, achieving a ROUGE-L score of 33.89 and one of the highest BERTScore values, at 75.92. Qwen3-4B also performs robustly, surpassing all models from 1B to 3B. The Qwen models demonstrate lower compression ratios than the Llama-1B variants, indicating more controlled and consistent summarization.

In terms of compression, mBART produces extremely high compression values (625.75), indicating a tendency to generate much shorter summaries than human references and often omitting key information. In contrast, LLMs exhibit more stable compression behavior, with values ranging from 248 to 316. Smaller models (e.g., Llama-3.2-1B) tend to over-compress, whereas larger models, such as Llama-3.1-8B and Qwen3-8B, generate summaries that are closer in length to those written by humans. Additionally, we observe a small number of empty or nearly empty predictions for very long articles, particularly in smaller models, though this rate decreases significantly with model size.

6 Discussion

In this section, we analyze the linguistic and structural properties of the gold and generated summaries in order to address RQ1 and RQ2. Then, we examine how model outputs vary across media outlets and ideological groups under lexical and semantic evaluation metrics to address RQ3.

6.1 RQ1. Linguistic and Stylistic Properties

To address RQ1, we conducted an automatic linguistic and stylistic analysis of gold and

Model	R1	R2	RL	BERTScore	Hybrid	Compression
mBART	23.77	12.29	18.88	71.83	45.36	625.75
Gemma-3-1B	37.58	22.71	31.81	75.11	53.46	316.07
Llama-3.2-1B	37.39	22.68	31.77	75.12	53.44	316.08
Llama-3.3-3B	39.20	24.05	33.46	75.68	54.57	281.53
Llama-3.1-8B	**39.91**	**24.71**	**34.12**	**75.97**	**55.04**	**248.65**
Qwen3-4B	38.79	23.71	32.99	75.58	54.28	268.80
Qwen3-8B	39.62	24.65	33.89	75.92	54.91	267.57

Table 1: Evaluation results on the political news summarization test set. We report ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore (F1), a hybrid metric (mean of RL and BERTScore), and the summary compression ratio.

generated summaries using UMUTextStats (García-Díaz et al., 2022), a linguistic analysis tool for Spanish that extracts over 365 linguistic features. The tool performs tokenization, PoS tagging, and morphological analysis to identify syntactic patterns, discourse markers, lexical diversity, and figurative expressions through pattern-based heuristics. It also detects figurative expressions through rule- and lexicon-based heuristics.

First, we analyze the linguistic and structural properties of the generated summaries, focusing on Llama-3-8B, which was the best-performing model in our experiments. Compared to the gold summaries, which are typically short and single-sentence with a very high lexical density, Llama-generated summaries are substantially longer, with an average of 57 words. They contain an average of 2.42 sentences, indicating a tendency to expand rather than compress information. Additionally, the Llama-generated summaries have lower lexical diversity, with a TTR of 0.83. This suggests that Llama repeats lexical material more frequently and distributes information less densely than human writers.

Pragmatic markers and figurative expressions, such as similes and metaphors, appear more frequently in generated text. Discourse connectors also occur in a larger proportion of summaries compared to gold references. This suggests that Llama often produces output that is more explanatory or narrative-style than the concise, fact-focused prose that is typical of journalistic summaries. Morphosyntactic analysis reveals additional differences.

Subjunctive and conditional verb forms, which are common markers of speculation or an evaluative stance, appear more frequently in generated summaries (67% and 5.3% of documents, respectively) than in the gold summaries (60% and 3.3%). This aligns with prior observations that instruction-tuned LLMs may introduce modal nuances or hedging not present in the source material. Together, these stylistic and syntactic differences highlight the discrepancy between human summarization practices and model-generated outputs.

They directly answer RQ1 by showing that although instruction-tuned LLMs generate coherent summaries, they systematically diverge from the compact, information-dense style of professional journalistic writing.

6.2 RQ2. Content Selection and Information Retention

To address RQ2, we examine how the structural properties of source articles relate to the selection of content and the retention of information in generated summaries. We do this by comparing the distributions of entities between the original documents and the model outputs.

To accomplish this, we employed the 5W1H framework typically used in journalism (Sepúlveda-Torres et al., 2024) and examined the distribution of named entities using NER-based lexical features for people, organizations, and locations. In our corpus, the original documents show a balanced distribution of entity mentions with approximately 27% person, 46% organization, and 27% locations. However, summaries tend to preserve institutional references while omitting individual actors and places. The relative proportion of person and location tokens decreases to about 73% of the original value while organizational tokens remain almost unchanged. The median values in the summaries are 0 for people and locations, but

50 for organizations, indicating that many summaries retain only organizational entities and systematically omit people and locations present in the source text. These results suggest that the structural properties of the corpus, such as entity distribution, article length, and syntactic complexity, significantly influence summarization behavior. Smaller models are more sensitive to these constraints, while larger models demonstrate greater robustness.

6.3 RQ3. Outlet and Ideology Sensitivity

The results reveal clear differences in model performance when evaluating the outputs of different models with various editorial orientations using lexical and semantic metrics.

First, as shown in Figure 3, ROUGE-L scores vary substantially across outlets, with some sources consistently yielding higher lexical overlap than others. This variability suggests that outlet-specific stylistic conventions significantly impact lexical alignment. Since gold summaries themselves exhibit stylistic differences across outlets (as shown in RQ1), lexical-overlap metrics such as ROUGE-L naturally reflect this variability. In contrast, BERT-F1 scores remain relatively stable, suggesting that semantic preservation is more consistent even when lexical similarity diverges. These findings imply that models can preserve the core meaning of political articles across outlets, yet they struggle to reproduce outlet-specific stylistic patterns consistentlyLarger models exhibit much stronger stylistic fidelity than smaller ones. Figure 3 also shows that Llama-3-8B (but also the baseline) tend to obtain the highest ROUGE-L values in most outlets. Smaller models such as Gemma-3-1B consistently rank among the lowest across both metrics. The performance gap between models becomes more visible in ROUGE-L than in BERT-F1.

Second, the aggregate results in Figure 4 confirm that Llama-3-8B and the baseline are the best-performing models overall. The mean ROUGE-L scores show greater variability across models, while the BERT-F1 averages are tightly clustered, highlighting the differences in metrics observed earlier. The combined score (ROUGE-L plus BERT-F1) puts Llama-3-8B slightly ahead of the other systems.

Figure 5 shows a heatmap that provides a more detailed view of how model rankings change across outlets. Some outlets show strong consistency in their rankings, with the same systems consistently occupying the top or bottom ranks. Other outlets show a more mixed distribution, where rankings vary depending on the model family. Generally, larger models appear in the top ranks more frequently, while smaller models tend to cluster toward the bottom.

Taken together, the results indicate three consistent patterns: (i) outlet-level differences have a noticeable effect on ROUGE-L but much less on BERT-F1; (ii) model size correlates with performance, especially in lexical metrics; and (iii) despite variation across outlets, the relative ordering of models remains fairly stable, with mBART and Llama-3-8B performing strongest overall.

Finally, Figure 6 summarizes the rankings of the models after the outlets were grouped by ideological orientation. The pattern remains consistent across the three groups: larger models, such as Llama-3-8b and Qwen-3-8b, have the lowest (best) average ranks, while smaller models, including Gemma-3-1b and Llama-3-1b, have the highest (worst) ranks. Mid-sized models (Llama-3-3b and Qwen-3-4b) generally occupy intermediate positions. The similarity of the rankings of left-, neutral-, and right-leaning outlets indicates that ideological orientation does not substantially alter the relative ordering of the models. Instead, model size is the dominant factor affecting performance.

As we have observed, the summarization performance varies across media outlets. ROUGE-L is substantially more sensitive to outlet-specific stylistic differences than BERTScore. This sensitivity is expected given the lexical variability of gold summaries across media sources. Despite this variation, model rankings remain stable. Larger instruction-tuned LLMs and the encoder-decoder baseline consistently outperform smaller models across outlets. Furthermore, clustering outlets by ideological orientation does not change these patterns. This indicates that model size and architectural capacity are dominant performance drivers. At the same time, the outlet-level differences observed in ROUGE-L reflect both the intrinsic variation in the reference summaries and the models' sensitivity to outlet-specific conventions.

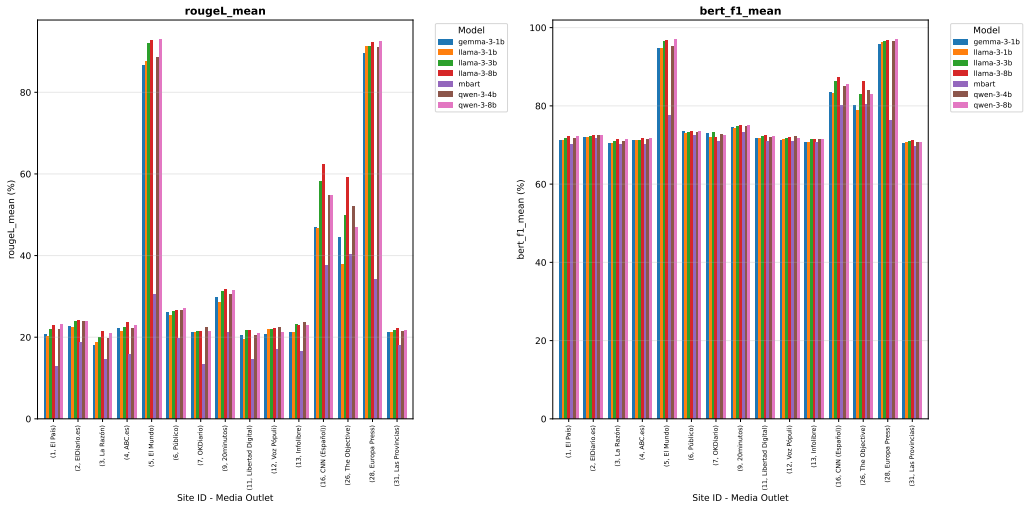

Figure 3: ROUGE-L and BERT-F1 performance of all models across media outlets.

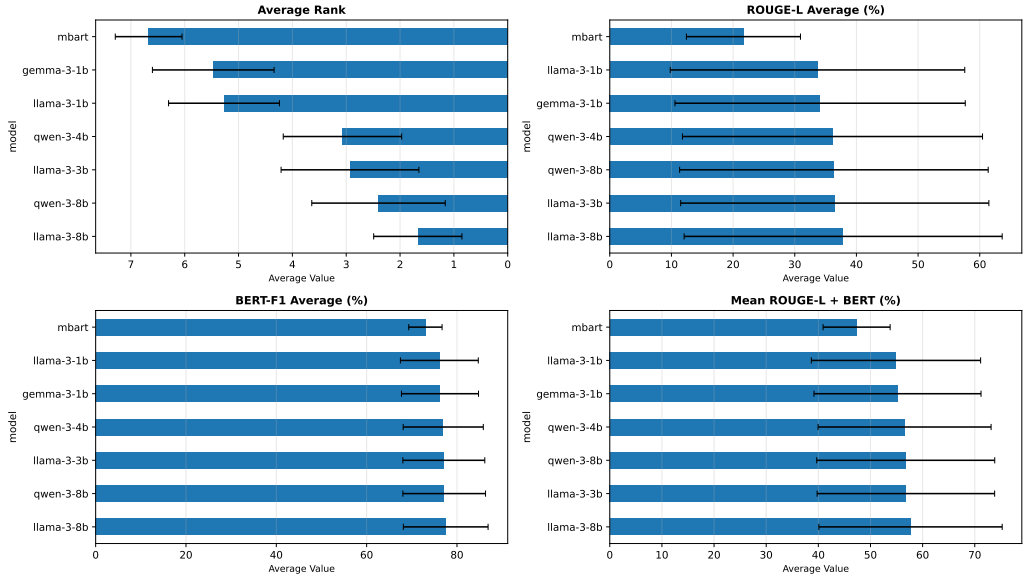

Figure 4: Aggregate evaluation of all models across outlets.

7 *Conclusion and Future Work*

In this work, we present the Spanish PoliSUM-2025 corpus, a new corpus of political news articles from sixteen Spanish outlets with different editorial orientations, alongside their journalist-written summaries.[2] This corpus allows for a thorough analysis of the interaction between journalistic style, editorial stance, and generative models during summarization, an area that remains largely unexplored in Spanish-language political journalism. Using this resource, we evaluated three recent instruction-tuned LLMs alongside an encoder–decoder baseline. This is the first comparative study of these model families in this domain.

Our results demonstrate that instruction-tuned LLMs outperform mBART consistently across lexical and semantic metrics. Additionally, our findings reveal that model size is the strongest predictor of summarization quality. Linguistic analysis reveals that although the best-performing models capture the core meaning of political articles, their summaries diverge from journalistic conventions. They are longer and less lexically dense than human-written references and include more pragmatic markers and modal constructions. Evaluation at the outlet level further shows substantial variability in ROUGE-L due to stylistic differences across news

[2] `https://github.com/NLP-UMUTeam/Spanish-Polisum-2025`

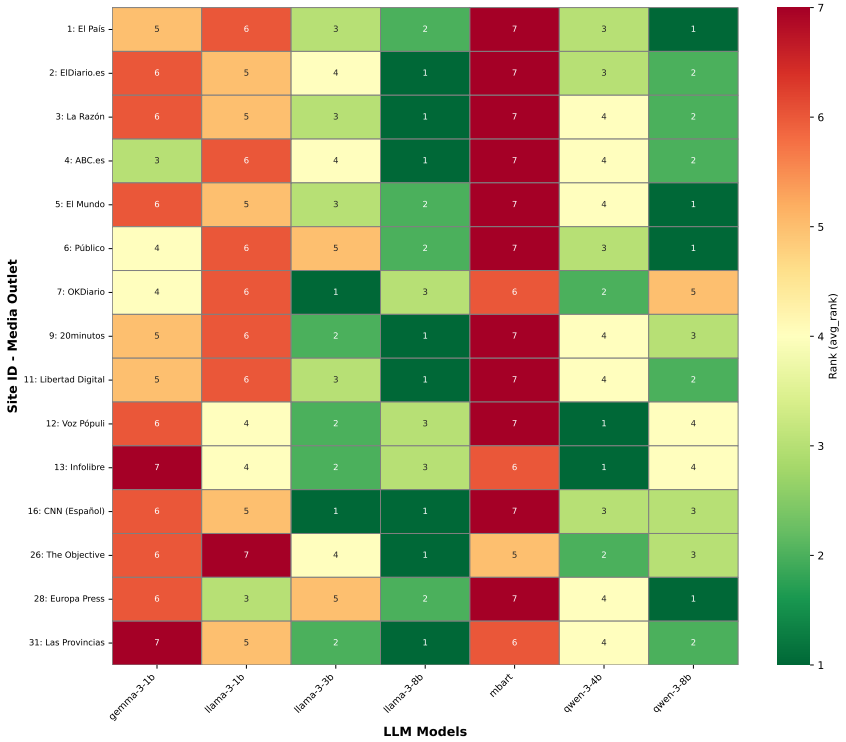

Figure 5: Heatmap of model rankings across media outlets.

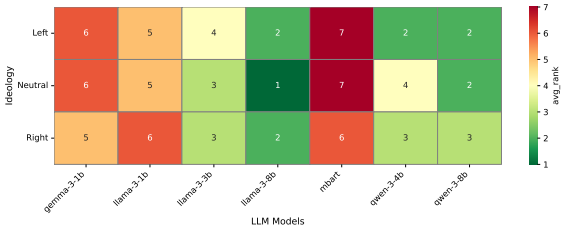

Figure 6: Average ranking of all models grouped by the ideological orientation of the media outlets.

sources. However, BERTScore remains stable, suggesting robust semantic preservation even when lexical alignment diverges.

Although we observed no substantial variation in ROUGE-L or BERTScore across outlets with different ideological orientations, it is important to emphasize that these metrics are not designed to capture ideological drift, framing shifts, or editorial cues. Therefore, our findings should not be interpreted as evidence that models preserve an ideological stance. Rather, they suggest that model performance remains stable across ideological groups under lexical and semantic similarity metrics. Detecting potential ideological drift would require evaluation frameworks dedicated to this task, such as framing-sensitive metrics or human annotation, which we will address in future work.

For future work, we propose: (i) expand the corpus to additional languages and news domains; (ii) examine how alignment techniques and reinforcement-learning methods affect stylistic and ideological stability; (iii) identify which input features, such as framing devices, entity prominence, sentiment cues, most influence model behavior; (iv) explore controllable summarization techniques, such as target-length conditioning or compression-aware decoding, to better align model outputs with journalistic summarization norms; and (v) develop evaluation methods that integrate linguistic, pragmatic, and ideological dimensions to more accurately characterize summarization quality in politically sensitive contexts.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Fund for Regional Development (ERDF/EU)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Carbonell, J. and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Deas, N. and K. McKeown. 2025. Summarization of Opinionated Political Documents with Varied Perspectives. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8088–8108, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Elatiky, A. F., A. M. Hamad, H. Khaled, and M. Fayez. 2026. Instruction-Tuned Decoder-Only Large Language Models for Efficient Extreme Summarization on Consumer-Grade GPUs. *Algorithms*, 19(2):96.

Fetahu, B., Z. Chen, O. Rokhlenko, and S. Malmasi. 2023. InstructPTS: instruction-tuning LLMs for product title summarization. *arXiv preprint arXiv:2310.16361*.

Fisher, J., S. Feng, R. Aron, T. Richardson, Y. Choi, D. W. Fisher, J. Pan, Y. Tsvetkov, and K. Reinecke. 2025. Biased LLMs can Influence Political Decision-Making. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6559–6607, Vienna, Austria, July. Association for Computational Linguistics.

García-Díaz, J. A., P. J. Vivancos-Vicente, A. Almela, and R. Valencia-García. 2022. Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6035–6044.

García-Ferrero, I. and B. Altuna. 2024. NoticIA: A Clickbait Article Summarization Dataset in Spanish. *Proces. del Leng. Natural*, 73:191–207.

Grattafiori, A., A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hamborg, F., K. Donnay, and B. Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Hasan, T., A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Helwe, C., O. Balalau, and D. Ceolin. 2025. Navigating the Political Compass: Evaluating Multilingual LLMs across Languages and Nationalities. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17179–17204, Vienna, Austria, July. Association for Computational Linguistics.

Kryściński, W., B. McCann, C. Xiong, and R. Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Ladhak, F., E. Durmus, C. Cardie, and K. McKeown. 2020. WikiLingua: A

New Benchmark Dataset for Multilingual Abstractive Summarization. *ArXiv*, abs/2010.03093.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Li, J., J. Chen, H. Chen, D. Zhao, and R. Yan. 2024. Multilingual Generation in Abstractive Summarization: A Comparative Study. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11827–11837, Torino, Italia, May. ELRA and ICCL.

Liu, Y., J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mihalcea, R. and P. Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Min, H., Y. Lee, M. Ban, J. Deng, N. H.-Y. Kim, T. Yun, H. Su, J. Cai, and H. Song. 2025. Towards Multi-dimensional Evaluation of LLM Summarization across Domains and Languages. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria, July. Association for Computational Linguistics.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Rush, A. M., S. Chopra, and J. Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. 2020. MLSUM: The Multilingual Summarization Corpus. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November. Association for Computational Linguistics.

Sepúlveda-Torres, R., A. Bonet-Jover, I. Diab, I. Guillén-Pacho, I. Cabrera-de Castro, C. Badenes-Olmedo, E. Saquete, M. T. Martín-Valdivia, P. Martínez-Barco, and L. A. Ureña-López. 2024. Overview of flares at iberlef 2024: Fine-grained language-based reliability detection in spanish news. *Procesamiento del lenguaje natural*, 73:369–379.

Team, G., A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2020. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *North American Chapter of the Association for Computational Linguistics*.

Yang, A., A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K.-P. Yang, L. Yu, L.-C. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S.-Q. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y.-C. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and

Z. Qiu. 2025. Qwen3 Technical Report. *ArXiv*, abs/2505.09388.

Zhang, J., Y. Zhao, M. Saleh, and P. J. Liu. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *ArXiv*, abs/1912.08777.