Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 263-282 recibido 06-04-2025 revisado 04-05-2025 aceptado 21-05-2025

BOE-XSUM: Extreme Summarization in Clear Language of Spanish Legal Decrees and Notifications

BOE-XSUM: Resúmenes Extremos en Lenguaje Claro de Decretos y Notificaciones Legales en Español

Andrés Fernández García,[1,*] Javier de la Rosa,[2,*] Julio Gonzalo,[1] Roser Morante,[1]

Enrique Amigó,[1] Alejandro Benito-Santos,[1] Jorge Carrillo-de-Albornoz,[1]
Víctor Fresno,[1] Adrian Ghajari,[1] Guillermo Marco,[1] Laura Plaza,[1] Eva Sánchez Salido[1]

[1]Universidad Nacional de Educación a Distancia, Spain
[2]The National Library of Norway, Norway
nandezgarcia@gmail.com

Abstract: The ability to summarize long documents succinctly is increasingly important in daily life due to information overload, yet there is a notable lack of such summaries for Spanish documents in general, and in the legal domain in particular. In this work, we present BOE-XSUM, a curated dataset comprising 3,648 concise, plain-language summaries of documents sourced from Spain's "Boletín Oficial del Estado" (BOE), the State Official Gazette. Each entry in the dataset includes a short summary, the original text, and its document type label. We evaluate the performance of medium-sized large language models (LLMs) fine-tuned on BOE-XSUM, comparing them to general-purpose generative models in a zero-shot setting. Results show that fine-tuned models significantly outperform their non-specialized counterparts. Notably, the best-performing model—BERTIN GPT-J 6B (32-bit precision)—achieves a 24% performance gain over the top zero-shot model, DeepSeek-R1 (accuracies of 41.6% vs. 33.5%).
Keywords: Extreme-summarization, legal texts, generative models, evaluation resources.

Resumen: La capacidad de resumir documentos largos de forma concisa es cada vez más importante en la vida cotidiana debido a la sobrecarga de información, pero existe una notable escasez de este tipo de resúmenes para documentos en español en general, y en el ámbito jurídico en particular. En este trabajo, presentamos BOE-XSUM, un conjunto de datos de 3648 resúmenes extremadamente breves en lenguaje claro creados a partir de las entradas del Boletín Oficial del Estado (BOE). El conjunto de datos contiene tanto los resúmenes como los textos originales etiquetados con el tipo de documento. Además, presentamos los resultados de experimentar en modo de fine-tuning y de zero-shot con modelos generativos. Nuestros resultados indican que los modelos generativos supervisados mediante fine tuning funcionan significativamente mejor que los modelos generativos en modo no supervisado, incluso siendo modelos más pequeños. El mejor modelo con finetuning de nuestra experimentación, BERTIN GPT-J 6B (precisión de 32 bits), obtiene resultados un 24% mejores que el mejor modelo no supervisado, DeepSeek-R1 (41,6% vs 33,5%).
Palabras clave: Resumen extremo, texto legal, modelos generativos, recursos de evaluación.

1 Introduction

Among the many capabilities of large language models (LLMs), summarization stands out as one of the most widely used and valued by end users (Budhiraja et al., 2024; Brachman et al., 2025). Recent advances in LLMs have made it possible to summarize extremely long documents with remarkable accuracy and coherence. These break-

*These authors contributed equally to this work.

ISSN 1135-5948 DOI 10.26342/2025-75-20

throughs have been driven in large part by the availability of large-scale, high-quality summarization datasets—particularly in English. However, this progress has not been evenly distributed across languages. Spanish, despite being the second most widely spoken native language in the world with over 485 million speakers (Ethnologue, 2023), remains significantly under-resourced in Natural Language Processing (NLP) (Conde et al., 2024). This scarcity is especially pronounced in summarization tasks: a search on Hugging Face reveals 963 summarization datasets in English,[1] compared to only 75 in Spanish.[2] Moreover, domain-specific resources—such as those tailored for administrative or legal document summarization—are virtually non-existent in Spanish, further limiting the development and applicability of LLMs in these critical areas.

In this work, we introduce BOE-XSUM, a curated dataset comprising 3,648 clear and extremely concise summaries of entries from Spain's "Boletín Oficial del Estado" (BOE)[3], the country's State Official Gazette, along with their original texts. This gazette serves as the primary platform for disseminating legislative decrees, legal notifications, and various official documents, predominantly originating from the national government, but including contributions from regional and local authorities. This is the first dataset of its kind in Spanish and is characterized by two key features: first, it addresses the important challenge of adapting complex domain-specific language, such as legal or administrative texts, into clear, everyday language; and second, it provides extremely concise summaries that are manually curated and verified.

The extreme summaries are built upon social media posts by a Spanish journalist specialized in the analysis and treatment of public information.[4] This is a socially relevant task for the control of governments, whether they are central, autonomous, or municipalities, and other public agencies such as the Constitutional and the Supreme Court. For example, in the context of the floods that devastated several regions in Spain in late 2024,[5] these posts contributed to inform the population about how central and regional governments were managing public resources to alleviate the effects of the disaster.

By compiling and making available the BOE-XSUM dataset, we aim to fill the gap in NLP resources for Spanish, offering a valuable resource for both academic research and practical applications. In addition, we present experiments with generative language models. This allows us to answer the following research question: to what extent can generative language models produce extreme summaries of legal texts, not only capturing their complex meaning but also adapting them into clear language in a manner comparable to that of a human expert?

Our contribution is twofold. First, we release a new publicly available dataset of extreme summaries in Spanish within the legal and administrative domain.[6] Second, we conduct experiments with generative text models, revealing that current systems still struggle to produce high-quality extreme summaries. However, a qualitative analysis shows that in some cases, the generated summaries resemble those written by humans.

This paper is organized as follows. In Section 2 we present the related work. In Section 3 we describe the BOE-XSUM dataset and provide a qualitative analysis of its extreme summaries. Section 4 focuses on the experiments, the results of which are discussed in Section 5. Finally, the conclusions are presented in Section 6.

2 Related work

In general, there are two main approaches to summarization: extractive and abstrac-

[1] `https://huggingface.co/datasets?task_categories=task_categories:summarization&language=language:en&sort=trending`, accessed on April 1st, 2025.

[2] `https://huggingface.co/datasets?task_categories=task_categories:summarization&language=language:es&sort=trending`, accessed on April 1st, 2025.

[3] `https://www.boe.es/`

[4] Eva Belmonte is a journalist and co-director of Civio Foundation (see `https://civio.es/`) an independent, non-profit organization that monitors public authorities through data journalism, and works on three lines of action: journalism, public advocacy and transparency services for public administrations. Belmonte is part of a multidisciplinary group of people who work to improve the democratic quality in Spain. She publishes concise daily summaries of the BOE in X (see `https://x.com/evabelmonte`).

[5] `https://en.wikipedia.org/wiki/2024_Spanish_floods`.

[6] `https://huggingface.co/datasets/bertin-project/BOE-XSUM`

tive summarization (Cajueiro et al., 2023). Within the extractive works, we can find those that compute frequencies, such as those based on spatial vectors (Baeza-Yates et al., 1999; Belwal, Rai, and Gupta, 2021), matrix factorization (Gong and Liu, 2001), graphs (Mihalcea and Tarau, 2004), topics (Haghighi and Vanderwende, 2009) and neural word embeddings (Kågebäck et al., 2014). In addition to frequency-based methods, there are also approaches based on heuristics (Edmundson, 1969), (Dalal and Zaveri, 2011), linguistics (Edmundson and Wyllys, 1961), (Mohamed, 2016), supervised machine learning (Mao et al., 2019), and reinforcement learning (Ryang and Abekawa, 2012; Hyun et al., 2022). Extreme summarization is a form of single-document summarization that aims to generate highly concise summaries—often a single sentence—that capture the core meaning of the source text (Narayan, Cohen, and Lapata, 2018a; Cachola et al., 2020a). Unlike extractive methods, it requires abstractive generation to synthesize information from the input, making it particularly useful in domains like scientific literature and news media (Mao, Zhong, and Han, 2022).

There exist datasets of summaries in many languages such as English (Grusky, Naaman, and Artzi, 2018), Spanish and Catalan (Segarra Soriano et al., 2022), Indonesian (Koto, Lau, and Baldwin, 2020), and Bengali (Khan et al., 2023) among others. And while there are multilingual datasets available, such as MLSUM (Scialom et al., 2020), WikiLingua (Ladhak et al., 2020), EUR-Lex-Sum (Aumiller, Chouhan, and Gertz, 2022), XL-SUM (Hasan et al., 2021) and HumSet (Fekih et al., 2022), they usually share certain characteristics that are not optimal for their use in training generative models. For instance, summaries may be provided as titles preceding the text, or they might conclude with a link to the full text, making it very easy for language models to match the original text and its summary. Such features, particularly given the public nature of these texts, enhance the effectiveness of generative models, which, in the end, perform a memorization task, instead of a meaning abstraction task.

As for extreme summarization datasets in Spanish, the availability of resources is severely limited, with only one available dataset, NoticIA (García-Ferrero and Altuna, 2024), which consists of 850 Spanish news articles with clickbait headlines, each paired with a human-written, single-sentence generative summary. This dataset assesses the ability of models to understand and summarize texts, addressing the challenge of interpreting complex information generated by clickbait headlines. Finding datasets of extreme summaries in languages other than English or Chinese is really difficult (Narayan, Cohen, and Lapata, 2018b; Sotudeh et al., 2021; Cachola et al., 2020b).

The challenge of summarizing complex content is particularly evident when translating from highly specialized technical language—such as legal or scientific discourse—into more accessible forms. This issue has been widely explored in the context of science communication and plain language movements. Studies in popular science writing have shown how lexical and structural simplification strategies can effectively bring technical content closer to general audiences, while still preserving meaning and nuance (Montalt and González Davies, 2007). Similarly, the field of legal communication has highlighted the importance of translating dense, technical legal language into clear and formal language for broader accessibility (Tiersma, 2003). These linguistic transformations not only benefit public understanding but are also central to tasks such as summarization, where preserving intent while changing register is essential. In this sense, research on genre translation—from specialized to general discourse—offers valuable frameworks for evaluating the capacity of language models to handle extreme summarization tasks in real-world scenarios (Baram-Tsabari and Lewenstein, 2017).

3 Dataset

In this section we provide details about the dataset creation: data sources, data extraction, data cleaning, and annotations.

3.1 Sources

The BOE-XSUM dataset originates from the social media posts made by a Spanish journalist who performs daily reviews of the BOE. She selects the articles that considers of greatest social interest and summarizes them in posts that also include a the link to the PDF version of the specific BOE article be-

ing summarized. Depending on the relevance or complexity of the content, the journalist writes one or several short posts summarizing the essence of the resolutions and orders in the article of the BOE. In the case of crucial matters, she elaborates a detailed long post on an external website, which is subsequently linked to the initial post. Each summary in the dataset has two versions: the original social media post written by the journalist, and an edited version crafted for the extreme summarization task. These edited versions have been refined to ensure accurate representation of the BOE content and to meet the standards of clear language summarization. All experiments presented in this work have been conducted using the edited summaries.

We initially collected over 4,500 social media posts. After manual review, we discarded those that lacked a direct link to a BOE article, were not clearly related to the BOE content, or were too subjective. Detailed annotation guidelines are available in Appendix G. The final dataset consists of 3,648 BOE entries, each paired with an original post and its corresponding edited summary. The BOE articles were also annotated by category to enhance their usability for both summarization and classification tasks.

3.2 Data extraction

The next step was to classify the links included in the social media posts into two categories. Given the occasional use of external links, our analysis focused on distinguishing and separating between two main categories of links: those linking to the journalist external website, which contains articles related to the BOE, and those linking directly to BOE documents. This differentiation allowed us to capture both the concise summaries inherent in her posts and, where appropriate, extended summaries along with direct links to BOE documents. In the case of links to BOE documents, mostly in PDF format, a conversion process was employed to facilitate their download in plaint text format, preserving the identifier used originally in the BOE article.[7]

[7]For example, from PDF link from the BOE as `https://boe.es/boe/dias/2024/03/28/pdfs/BOE-A-2024-6273.pdf`, we first take its identifier "BOE-A-2024-6273", `https://boe.es/diario_boe/txt.php?id=BOE-A-2024-6273`, and then use it to generate the official URL that exposes the content in text format, which we subsequently use to download the content of the articles in plain text for our dataset.

3.3 Manual data review

After a preliminary analysis of the dataset, a decision was made to develop a tool for the visualization, tagging, and editing of the dataset, as depicted in Figure 3 included in Appendix A. The main purpose of the tool was to verify the integrity of the data visualization and being able to modify the data if there were any errors. The review of the data was carried out with this tool. We proceeded to verify, one by one, the correspondence between the BOE text and the tweet containing the associated summary. When we identified any inconsistencies, such as a summary that did not match the content of the BOE or that was poorly contextualized, we made the necessary corrections to ensure that the summary accurately reflected the information contained in the official text. The complete annotation guidelines are described in Appendix G.

We also edited a number of summaries to ensure that they better reflected the content of the original BOE document and agreed to our guidelines, improving clarity and accuracy. Both the original and edited versions are part of the dataset. Importantly, all training and evaluation experiments described in this work have been carried out using the edited versions only. This ensures consistency, eliminates ambiguity, and aligns with the goal of producing high-quality summaries in clear language. In Appendix D we show two examples of how the summaries were edited.

To estimate the extent of the editing process, we conducted a similarity analysis based on cosine distance between original and edited versions.[8] Table 8 in Appendix C shows that a third of the dataset (1,154 posts) had a similarity above 90%, while approximately 160 posts had a similarity below 10%. This variation reflects the range of interventions applied–from minor adjustments to full rewrites–to ensure that the summaries do not only provide a consistent view, but also faithfully represent the original BOE content.

Moreover, in a test with two male participants aged between 30 and 45 years, both

[8]We extracted embedding vectors using the original multilingual BERT.

holding university degrees, summaries edited according to the guidelines were chosen 189 out of 200 times (94.5%), 95% CI [90.42%, 96.9%], significantly above chance (two-tailed binomial test, $p < 0.001$).

3.4 Categorization

Each entry in the dataset includes both the original and edited summaries, along with a label indicating the type of BOE article. All categories are disjoint and the labeling of articles follows two main criteria:

1. **Explicit Mention in the BOE article:** If the name of a category is explicitly stated in the BOE article, that label is assigned directly. For example:
   - Articles from `TRIBUNAL_CONSTITUCIONAL` are labeled as *Constitutional Court*.
   - Articles from `CONVENIOS` are labeled as *Agreement*.

   Some categories are formed by combining multiple clearly defined types. For instance, articles related to awards and medals are grouped under the `PREMIOS_Y_MEDALLAS` category.
2. **Frequency-Based Filtering:** Categories with a very low number of articles are grouped into a general category named `OTROS_ANUNCIOS` (*Other announcements*). One exception is the `BANCO_DE_ESPAÑA` (*Bank of Spain*) category, which is retained in the dataset due to its relevance, even though it contains relatively few articles.

A list of all categories with illustrative examples is provided in Table 5 in Appendix B. The distribution of categories across the dataset partitions is shown in Table 6.

As presented in Table 7, the dataset contains several columns. Among the most important are: BOE article texts (text), Original posts (summaries), Edited versions of the posts (edited summaries).

3.5 Content Analysis

Among the BOE documents, there exists a variety of contents, including straightforward articles such as the appointments of ambassadors or designation of official positions. These particular entries, by virtue of their inherently concise nature, offer a less complex summarization task.

Conversely, the dataset also encompasses BOE articles of a more intricate and voluminous nature, including but not limited to decisions from the Constitutional Court, Supreme Court rulings, and various agreements. Such documents are characterized by their extensive length, often spanning thousands of words. To illustrate the complexity and scale of these articles, we show a simplified example, with only the introductory and concluding segments of this representative article:[9]

> El Real Decreto-ley 6/2012, de 9 de marzo, de medidas urgentes de protección de deudores hipotecarios sin recursos, establece una serie de mecanismos conducentes a permitir la reestructuración de la deuda hipotecaria de quienes padecen extraordinarias dificultades para atender su pago. A tal fin, al citado Real Decreto-ley se incorporó un código de buenas prácticas al que podrán adherirse las entidades y cuyo seguimiento será supervisado por una comisión de control, cuya composición ha sido modificada por el artículo 6 de la Ley 1/2013, de 14 de mayo, de medidas para ... ... – Caja Rural San Jaime de Alquerías Niño Perdido, S. Coop. de Crédito V. – Caja Rural San José de Almassora, S. Coop. de Crédito. V. – Caja Rural San José de Burriana, S. Coop. de Crédito V. – Caja Rural San José de Nules, S. Coop. de Crédito V. – Caja Rural San Roque de Almenara, S. Coop. de Crédito V. – Colonya-Caixa D'estalvis de Pollença. – Liberbank, S. A. – Publicredit, S. L. – UNOE Bank, S. A.

The original summary for this BOE article as written by the journalist is as follows:

> Adhesiones a 2 códigos de buenas prácticas en hipotecas Has ahora, poco efectivos. Porque legislar no, ¿verdad? #BOE

The resulting edited summary:

> Lista de adhesiones de bancos a los códigos de buenas prácticas para reforzar la protección a los deudores hipotecarios, reestructuración de deuda y alquiler social

As Table 1 shows, the dataset contains a total of 3,648 BOE texts, with a total 13,304,989 space-delimited words. The average number of words per document is 3,396, while the summaries average a total of 17 words, which gives us a 0.005% compression rate from the original text to the summary. More than 64% of BOE documents have less than 1,000 words and only 2.65% have more than 25,000 words. A histogram for BOE

[9] See English translations in Appendix F.

documents with at least 25,000 words can be seen in Figure 1.[10]

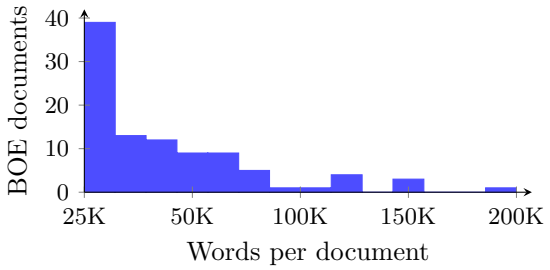

Figure 1: Histogram showing the distribution of BOE documents by word count for BOE documents with at least 25,000 words.

The dataset was divided into three splits: train, development, and test. Initially, we adhered to an 80/10/10 distribution, but adjustments were made to ensure that the annotated categories were well-balanced across the splits. The distribution of categories per split is provided in Table 6 of Appendix C.

4 Experiments

We address the task of extreme summarization using BOE-XSUM, which requires deep textual comprehension to produce highly condensed summaries of no more than 280 characters. This length constraint was chosen to parallel the character limit typical of the social media platform on which the original summaries were written. The degree of compression varies significantly across documents; some entries in the BOE corpus are relatively brief, while others are substantially longer, as detailed in the analysis section. We evaluate language models under two settings: fine-tuning and zero-shot prompting.

Although the dataset also includes the original posts for reference, all experiments are conducted using the edited summaries produced through the manual curation process described in Section 3.3. This ensures a consistent and contextually accurate benchmark for the extreme summarization task.

4.1 Evaluation Metrics

The evaluation of LLMs performed using common metrics in the literature: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007), and BERTScore (Zhang et al., 2020). BLEU focuses on n-gram overlap with reference texts and includes a brevity penalty to discourage overly short outputs. METEOR emphasizes unigram precision and recall with added flexibility through synonym and stem matching, and includes penalties for fragmented matches. ROUGE measures quality using n-gram overlap and the longest common subsequence, with variants tailored to different summarization aspects. BERTScore leverages contextual embeddings from models like BERT to assess semantic similarity via cosine similarity, enabling a more nuanced comparison beyond surface-level word matching.

4.2 Fine-tuning experiments

We used BERTIN GPT-J 6B[11] (De la Rosa et al., 2022; De la Rosa and Fernández, 2022) as the base model. After downloading the entire BOE from 1988 to 2023, we continued pre-training this model for 3 epochs using the same parameters as the original configuration, resulting in a domain-adapted variant, BOLETIN.[12] With these two models in place, we conducted a grid search to train them under various configurations.

To establish a performance baseline, we first fully fine-tuned both models using 32-bit precision. In parallel, we explored parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA introduces trainable low-rank matrices into key components of the model's attention mechanism—specifically the layers responsible for projecting and integrating attention information—allowing the model to adapt to new tasks while updating only a small subset of parameters. This significantly reduces computational and memory requirements. Training was conducted for up to 600 steps using mixed-precision formats (4-, 8-, and 16-bit), as detailed in Table 2. Results in Table 1 illustrate the trade-offs between full and parameter-efficient fine-tuning across model variants.

Given the architectural constraints of these models, the input sequence is limited to a maximum of 2048 tokens. To accommodate summary generation within this limit, a

[10] This is due to the inclusion of complete external texts in the BOE document itself, such as Royal Decrees and Laws. See also Table 8 in Appendix C for the percentages of texts by ranges of increasing size.

[11] https://huggingface.co/bertin-project/bertin-gpt-j-6B

[12] https://huggingface.co/bertin-project/BOLETIN

Model	Precision	Layers	BLEU	METEOR	ROUGE	BERTScore
BOLETIN	4-bit	13	0.050	0.258	0.281	0.262
	8-bit	13	0.065	0.292	0.306	0.289
	16-bit	13	0.066	0.299	0.308	0.294
	4-bit	27	0.054	0.264	0.287	0.269
	8-bit	27	0.062	0.292	0.305	0.286
	16-bit	27	0.074	0.306	0.317	0.298
	32-bit	All	**0.094**	**0.333**	**0.367**	**0.397**
BERTIN GPT-J 6B	4-bit	13	0.062	0.274	0.289	0.264
	8-bit	13	0.057	0.295	0.304	0.287
	16-bit	13	0.061	0.306	0.314	0.299
	4-bit	27	0.068	0.278	0.293	0.267
	8-bit	27	0.056	0.295	0.305	0.291
	16-bit	27	0.057	0.305	0.313	0.300
	32-bit	All	**0.109**	**0.365**	**0.393**	**0.416**

Table 1: Scores of the differet metrics for fine-tuned BOLETIN and BERTIN GPT-J 6B models evaluated on the test split. The models differ in precision (bit-depth) and the number of layers trained (13 or 27 for LoRA training, "All" for full fine-tuning). The last row for each model corresponds to the result for a 32-bit model with all layers trained. Best scores in **bold**.

Parameter	Value
Batch size	4
Gradient accumulation steps	4
Warmup steps	100
Max steps	600
Learning rate	2e-4

Table 2: Fine-tuning hyperparameters for all experiments.

portion of tokens must be reserved for both initiating the summary and for the model's output. During training, we appended the marker "`### RESUMEN:`" ("*`### SUMMARY:`*") after each BOE article, followed by its corresponding summary. In cases where an article exceeded the token limit, we truncated the input as to accommodate for the task marker and allow sufficient space for model output during generation. For especially long documents, we ensured that the summary referred primarily to the initial portion of the article, minimizing the risk that truncation would remove critical information required for accurate summarization. However, these heuristics were not always sufficient as many BOE texts are long and complex. As a result, a non-negligible number of examples likely lacked critical information for the generation of accurate summaries. We did not systematically exclude or mark truncated samples, which means that some training instances may have introduced noise or incomplete context. This could partially explain the issues observed in some generated summaries, which were vague or abruptly cut. Future work should consider integrating models with larger context windows or using strategies such as sliding windows or hierarchical encoding to handle long legal texts more effectively.

Table 1 provides the results of the fine-tuning experiments for both BERTIN GPT-J 6B and BOLETIN. Despite the high level of complexity of the task, most models achieve similar outcomes, with full-precision fine-tuning outperforming any LoRA configuration. We also noted that the number of trained layers is not predictive of performance (see Figure 4 in Appendix E).

	B	M	R	BS
BLEU (B)	1.000	0.872	0.936	0.919
METEOR (M)	0.872	1.000	0.967	0.934
ROUGE (R)	0.936	0.967	1.000	0.991
BERTScore (BS)	0.919	0.934	0.991	1.000

Table 3: Pearson correlation matrix between automatic evaluation metrics.

Moreover, Table 3 shows that all metrics are strongly correlated with each other, suggesting that they measure related aspects of the quality of the generated text. ROUGE and BERTScore show the highest correlation (0.991), indicating that they tend to vary together and reflect very similar characteristics of the content, such as coverage and semantics. BLEU has the lowest correlation with METEOR (0.872), but it is still high. This makes sense, as BLEU is more strict (based on exact n-gram matches), while METEOR is more flexible and semantically oriented. Overall, using these metrics together provides a coherent and consistent evaluation

of model performance.

Interestingly, despite the additional pretraining of the BOLETÍN model on the full content of the BOE (1988–2023) to enhance its legal domain knowledge, this strategy did not lead to improved performance on the extreme summarization task in plain language. Our initial hypothesis was that continued pretraining would help the model better grasp the structure, terminology, and semantics of legal and administrative texts, thereby improving its ability to generate accurate summaries. While this approach may be beneficial for tasks that require formal or technical language, it appears to be less effective when the objective is to produce highly concise summaries in clear, accessible language. The specialized patterns reinforced during domain adaptation may have introduced a linguistic mismatch with the target style. This result underscores the importance of aligning not only the domain but also the linguistic register of the training data with the specific demands of the downstream task.

A deeper analysis revealed that the length constraint during training impacted negatively the generation of summaries. For example, we encountered incomplete summaries like: *'Real Decreto que modifica el Real Decreto 369/1999, de 5 de marzo, sobre términos y...'*, instead of the expected summary: *'Curas evangélicos que acrediten haber ejercido antes de 1999 cobrarán pensión de jubilación.'* Or cases like: *'El CSD notifica a todos los interesados en el recurso del Real Madrid contra las modificaciones de los...'*, where the correct summary should have been: *'El Real Madrid recurre las reformas de reglamento y estatutos sociales de la Liga de Fútbol.'*

4.3 Prompting experiments

For zero-shot experimentation, we selected a variety of open source and proprietary models of different sizes, some multilingual some tailored to Spanish content. Our goal was to analyze the performance of both large and small models.

In this setting, the importance of the prompt is crucial in determining the quality of response of a generative model. After some iterations and a limited number of manual trials, we settled on a prompt that showed promise of being precise and effective in the generation of summaries.

Prompt: *Eres un experto generando resúmenes en lenguaje cotidiano a partir de documentos legales escritos en lenguaje formal. Quiero que me des un resumen en español de entre 15 y 22 palabras del siguiente texto. Recuerda, solo quiero que devuelvas el resumen, nada más. Devuelveme únicamente el resumen, solo el resumen. A continuación te indico el texto que debes resumir: [*BOE DOCUMENT*]*

Table 4 shows the results from the experiments using the edited summaries[13]. On BOE-XSUM, DeepSeek R1 produced the best results with an BERTScore of 0.335, followed by the model ChatGPT 4o with a BERTScore of 0.327 and Llama 3 70b Instruct with a score of 0.285. Gemma 2 27B obtained a 0.266. It is remarkable that the model Gemma 2 9B produced better summaries than Llama 2 70B, given their difference in size.[14] We also observed that Llama 2 70B was prone to generate much longer summaries, while ChatGPT 4o, Llama 3 70B Instruct, and Gemma 2 models produced shorter and more accurate summaries. The three Gemma 2 models produced the shortest summaries, but they got worse as the model got smaller. It is also striking that the average length of the summaries produced by the top-ranked models, ChatGPT 4o and DeepSeek R1, is close to that of those of the ground truth summaries (17 words), with an average of 16.68 and 18.99 words, respectively. Importantly, there seems to exist a strong negative correlation between the average number of words of generated summaries and their BERTScore values across models (Pearson's $r = -0.82$, $p < 0.001$), suggesting that as the length of a generated summary increases, its BERTScore against the ground truth tends to decrease significantly.

5 Discussion and Future Work

Given the evident lack of summarization datasets in Spanish, BOE-XSUM will enable the training of models that generate extreme summaries in clear language, as well as to categorize this type of texts. We have demon-

[13] See Table 10 in Appendix E for results obtained using the original unmodified original posts by the journalist, which are generally worse.

[14] See Table 12 in Appendix E for examples of Gemma 2 2B vs Llama 2 70B.

Model	BLEU	METEOR	ROUGE	BERTScore	Avg. Words
Gemma 2 27b	0.041	0.221	0.245	0.266	12.02
Gemma 2 9B	0.042	0.211	0.238	0.255	14.93
Gemma 2 2B	0.041	0.194	0.222	0.248	16.53
ChatGPT 4o	0.042	0.265	0.314	0.327	16.65
Llama 3 70b Instruct	0.044	0.227	0.269	0.285	18.75
DeepSeek R1	0.041	0.251	0.307	0.335	18.99
Nemotron 70B	0.035	0.183	0.249	0.205	19.93
Llama 3.2 3B	0.035	0.197	0.236	0.245	21.40
Llama 3.2 1B	0.020	0.143	0.187	0.174	31.69
Llama 3.1 70B	0.029	0.229	0.238	0.245	31.69
Llama 2 70B	0.038	0.156	0.216	0.214	33.72
Llama 3.1 8B	0.034	0.199	0.243	0.251	33.72
Salamandra 7B Instruct	0.013	0.113	0.171	0.141	66.24
Solar 10.7B	0.011	0.110	0.172	0.166	60.70
Neural-Chat 7B	0.011	0.125	0.193	0.187	57.77
Mistral 7B	0.015	0.107	0.176	0.158	73.65
Starling 7B	0.007	0.088	0.157	0.149	72.92
Llava 7B	0.007	0.080	0.144	0.126	88.48
BERTIN GPT-J 6B	**0.109**	**0.365**	**0.393**	**0.416**	16.10

Table 4: Performance results of the generative models with prompts across different metrics. Added BERTIN GPT-J 6B for reference. Best scores in **bold**, second best underlined.

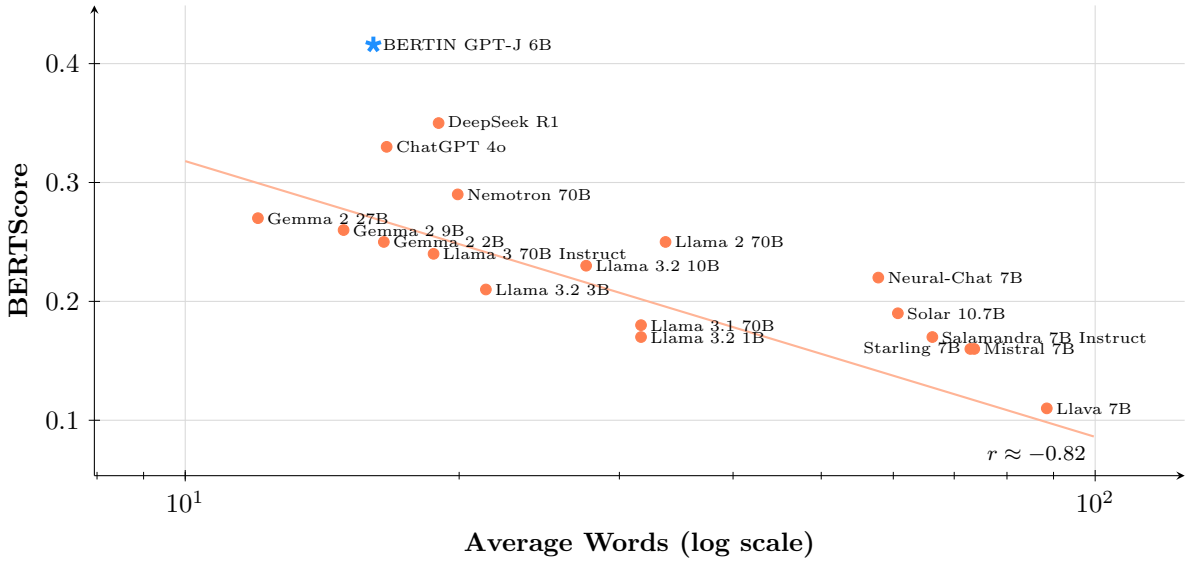

Figure 2: BERTScore against average number of words in generated summaries (log scale). BERTIN GPT-J 6B is highlighted with a bold blue star and excluded from the regression calculation.

strated experimentally that this is a complex task and a major challenge for generative models, regardless of size. The low number of summarized texts in certain categories raises the possibility of discarding them or merging them into the category `OTROS`, but we decided against because these categories are very relevant to the dataset.

Our results show that the task is more effective when fine-tuning is performed. All our fine-tuned models with precision up to 8-bit outperform models in zero-shot settings, except ChatGPT 4o and DeepSeek. It is important to notice that the fine-tuned models have only 6B parameters, in contrast with the 671B parameters of DeepSeek R1.

However, the comparison between fine-tuned and zero-shot models is not entirely symmetrical. Fine-tuned models were specifically trained on the BOE-XSUM dataset, with input-output structures tailored to the task, whereas zero-shot models were evaluated using a single static prompt, without domain-specific tuning or few-shot tech-

niques. Nonetheless, it was exciting to discover that smaller and limited models (6B parameters, 2048-token context window) could outperform much larger frontier models when fine-tuned for a narrowly defined and linguistically constrained task. In that sense, the results showcase the power of targeted fine-tuning in legal summarization scenarios. Future work should continue exploring few-shot learning and prompt optimization for general-purpose models, in order to better understand their potential when properly guided.

Although much remains to be done, due to the complexity of the task, extreme summarization of legal texts as extensive as those in the BOE remains a significant challenge for current models. Key open questions include (i) to what extent automatic metrics are correctly assessing model behavior, and (ii) what is the relative complexity contributed by the two core aspects of the task, i.e., extreme summarization and translation to plain language.

We also acknowledge the limitations of relying on standard automatic metrics such as BLEU, ROUGE, METEOR, and BERTScore for this specific task. These metrics may fail to capture essential qualities such as clarity, usefulness, or alignment with non-specialist expectations. Notably, we observed a strong negative correlation between the length of the generated summaries and the BERTScore (-0.82), which suggests that more informative summaries may be unfairly penalized for being slightly longer. This raises concerns about whether these metrics are truly aligned with the human notion of "better" summaries in this setting. While we conducted a limited human evaluation, a more robust human-centered validation is needed.

Moreover, while the dataset originated from journalistic summaries, we emphasize that all summaries used in our experiments were carefully edited to eliminate stylistic idiosyncrasies and ensure fidelity to the BOE source. This reduces the risk of source bias and provides a more standardized and generalizable input for model training and evaluation.

Future work should thus advance in several directions. First, a robust category classifier could be developed to automatically assign thematic labels to BOE texts, enhancing the utility of the dataset and enabling more nuanced analysis. Second, the introduction of more accurate and task-specific evaluation metrics is crucial. These could include clarity-based judgments or be inspired by natural language inference or question-answering paradigms, aiming to capture the communicative goals of extreme summarization more effectively than traditional n-gram-based or embedding-based scores. Together with expanded human evaluation protocols, these lines of work will help consolidate this task as a benchmark for controlled, high-precision text generation in legal and administrative domains.

6 Conclusions

We have developed a spanish dataset in the highly sought-after legal domain, designed to train and evaluate extreme summary generation models in clear language, suitable for microblogging platforms such as X or Bluesky. The dataset contributes to the increasingly relevant task of bridging the gap between legal language and the citizens affected by legal texts. The dataset presented will facilitate the training of systems that help the supervision of public administrations and allow the detection of relevant BOE entries that are of public interest.

Our preliminary experiments with the dataset indicate that (even small) fine-tuned systems seem a better choice than unsupervised (prompted) frontier models. However, these results are obtained with automatic evaluation measures, and it remains to be verified whether such measures are adequate in this context. Another lines of future work involve applying prompt engineering to optimize the performance of unsupervised models, and determining the relative challenge posed by the summarization factor and the translation to clear language factor. Also, we plan to enhance the dataset with lenghtier summaries from entries in the Civio website.

Finally, given that we have a categorized dataset with extreme summaries of articles published in the BOE, future work could focus on leveraging this dataset to develop a model capable of classifying all daily BOE entries. A generative model could be trained to automatically generate summaries for the classified entries. This would enable the automation and publication of extreme-summaries for BOE entries that are of broad public interest.

Acknowledgments

We thank David Cabo, co-director and CTO of the Civio Foundation, for his continued efforts to promote transparency in public institutions. We are especially grateful to co-director and journalist Eva Belmonte for her outstanding work making the Spanish Official State Gazette (BOE) understandable and relevant to the public. Through her project "El BOE nuestro de cada día," (*"Our daily BOE"*) she has highlighted the Gazette's most impactful content, helping bridge the gap between government actions and citizen awareness.

This work has been partially funded by the European Union - NextGenerationEU through the 'Recovery, Transformation and Resilience Plan', by the Ministry of Economic Affairs and Digital Transformation. Additionally, we thank Google for providing compute resources via the Tensor Research Cloud program, which significantly supported our model training efforts.

References

Aumiller, D., A. Chouhan, and M. Gertz. 2022. Eur-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain.

Baeza-Yates, R., B. Ribeiro-neto, D. Mills, O. Bonn, S. Juan, M. Mexico, C. Taipei, A. Wesley, and L. Limited. 1999. Modern information retrieval. 07.

Baram-Tsabari, A. and B. V. Lewenstein. 2017. Science communication training: What are we trying to teach? *International Journal of Science Education, Part B*, 7(3):285–300.

Belwal, R. C., S. Rai, and A. Gupta. 2021. Text summarization using topic-based vector space model and semantic measure. *Information Processing and Management*, 58(3):102536.

Brachman, M., A. El-Ashry, C. Dugan, and W. Geyer. 2025. Current and future use of large language models for knowledge work.

Budhiraja, R., I. Joshi, J. S. Challa, H. D. Akolekar, and D. Kumar. 2024. "it's not like jarvis, but it's pretty close!" - examining chatgpt's usage among undergraduate students in computer science. In *Proceedings of the 26th Australasian Computing Education Conference*, ACE '24, page 124–133, New York, NY, USA. Association for Computing Machinery.

Cachola, I., K. Lo, A. Cohan, and D. Weld. 2020a. TLDR: Extreme summarization of scientific documents. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online, November. Association for Computational Linguistics.

Cachola, I., K. Lo, A. Cohan, and D. S. Weld. 2020b. TLDR: Extreme summarization of scientific documents.

Cajueiro, D. O., A. G. Nery, I. Tavares, M. K. D. Melo, S. A. dos Reis, L. Weigang, and V. R. R. Celestino. 2023. A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding.

Conde, J., M. Gonzalez, N. Melero, R. Ferrando, G. Martinez, E. Merino-Gomez, J. A. Hernandez, and P. Reviriego. 2024. Open conversational LLMs do not know most spanish words. *Procesamiento De Lenguaje Natural*, 73:95–108.

Dalal, M. K. and M. A. Zaveri. 2011. Heuristics based automatic text summarization of unstructured text. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, ICWET '11, page 690–693, New York, NY, USA. Association for Computing Machinery.

De la Rosa, J. and A. Fernández. 2022. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In *IberLEF@ SEPLN*.

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, and M. Grandury. 2022. BERTIN: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Edmundson, H. P. 1969. New methods in automatic extracting. *J. ACM*, 16:264–285.

Edmundson, H. P. and R. E. Wyllys. 1961. Automatic abstracting and indexing—survey and recommendations. *Commun. ACM*, 4(5):226–234, May.

Ethnologue. 2023. The most spoken languages worldwide. https://www.ethnologue.com/insights/most-spoken-language/. Último acceso: 04/04/2024.

Fekih, S., N. Tamagnone, B. Minixhofer, R. Shrestha, X. Contla, E. Oglethorpe, and N. Rekabsaz. 2022. Humset: Dataset of multilingual information extraction and classification for humanitarian crisis response.

García-Ferrero, I. and B. Altuna. 2024. Noticia: A clickbait article summarization dataset in spanish.

Gong, Y. and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 19–25, New York, NY, USA. Association for Computing Machinery.

Grusky, M., M. Naaman, and Y. Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June. Association for Computational Linguistics.

Haghighi, A. and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *North American Chapter of the Association for Computational Linguistics*.

Hasan, T., A. Bhattacharjee, M. S. Islam, K. Samin, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hyun, D., X. Wang, C. Park, X. Xie, and H. Yu. 2022. Generating multiple-length summaries via reinforcement learning for unsupervised sentence summarization.

Khan, A., F. Kamal, M. A. Chowdhury, T. Ahmed, M. T. R. Laskar, and S. Ahmed. 2023. BanglaCHQ-summ: An abstractive summarization dataset for medical queries in Bangla conversational speech. In F. Alam, S. Kar, S. A. Chowdhury, F. Sadeque, and R. Amin, editors, *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 85–93, Singapore, December. Association for Computational Linguistics.

Koto, F., J. H. Lau, and T. Baldwin. 2020. Liputan6: A large-scale Indonesian dataset for text summarization. In K.-F. Wong, K. Knight, and H. Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China, December. Association for Computational Linguistics.

Kågebäck, M., O. Mogren, N. Tahmasebi, and D. Dubhashi. 2014. Extractive summarization using continuous vector space models. 04.

Ladhak, F., E. Durmus, C. Cardie, and K. McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization.

Lavie, A. and A. Agarwal. 2007. METEOR: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Mao, X., H. Yang, S. Huang, Y. Liu, and R. Li. 2019. Extractive summarization using supervised and unsupervised learning. *Expert Systems with Applications*, 133:173–181.

Mao, Y., M. Zhong, and J. Han. 2022. CiteSum: Citation text-guided scientific ex-

treme summarization and domain adaptation with limited supervision. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10922–10935, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Mihalcea, R. and P. Tarau. 2004. TextRank: Bringing order into text. In D. Lin and D. Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

Mohamed, M. A. 2016. Automatic text summarisation using linguistic knowledge-based semantics.

Montalt, V. and M. González Davies. 2007. Translating scientific and technical texts: Discourse and communication strategies in the popularization of science. In *Scientific and Technical Translation*. Routledge, pages 45–68.

Narayan, S., S. B. Cohen, and M. Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November. Association for Computational Linguistics.

Narayan, S., S. B. Cohen, and M. Lapata. 2018b. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Ryang, S. and T. Abekawa. 2012. Framework of automatic text summarization using reinforcement learning. In J. Tsujii, J. Henderson, and M. Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 256–265, Jeju Island, Korea, July. Association for Computational Linguistics.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. 2020. Mlsum: The multilingual summarization corpus.

Segarra Soriano, E., V. Ahuir, L.-F. Hurtado, and J. González. 2022. DACSA: A large-scale dataset for automatic summarization of Catalan and Spanish newspaper articles. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5931–5943, Seattle, United States, July. Association for Computational Linguistics.

Sotudeh, S., H. Deilamsalehy, F. Dernoncourt, and N. Goharian. 2021. Tldr9+: A large scale resource for extreme summarization of social media posts.

Tiersma, P. M. 2003. The plain language movement. *Law and Contemporary Problems*, 66(1/2):217–240.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Editor server

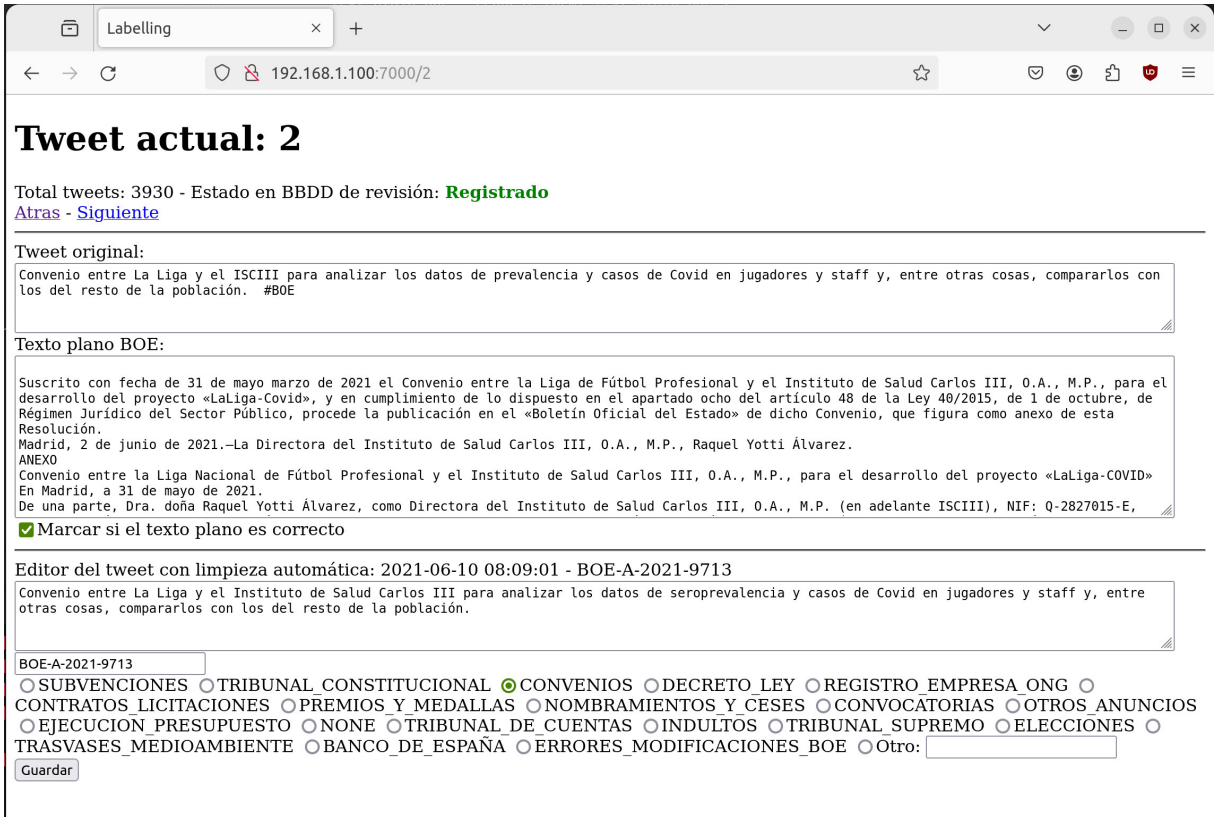

Figure 3: Editor server.

The features of this editing server include:

- Display of the current tweet number, facilitating easy navigation and reference.
- Indication of the data item's status within the new database, distinguishing between Registered and Not registered statuses.
- Navigation functionality, including Back and Forward buttons, which allow users to seamlessly move through the dataset.
- Inclusion of the original tweet authored by Eva Belmonte, providing a direct link to the initial data point.
- Presentation of the plain text version of the BOE documents, essential for content verification and analysis.
- A verification checkbox to confirm the accuracy of the BOE's plain text, addressing instances of incorrect document linkage. Despite the rarity of such occurrences, all affected entries have been excluded from the dataset.
- An automated cleanup editor for the tweet, designed to enhance data cleanliness and usability.
- The identifier of the BOE document, enabling users to verify the relevance and accuracy of the associated tweet.
- Categorization functionality for each data item, facilitating structured analysis and retrieval.
- A Save button to ensure any modifications or tags applied are retained.

B Categories annotated for the text categorization task

Category	Examples
OTROS_ANUNCIOS	*El Gobierno publica la nueva metodología de cálculo precios pequeño consumidor energía*
CONTRATOS_LICITACIONES	*La readjudicación del contrato a Aguas de Valencia por importe total de 7.000.000 euros*
NOMBRAMIENTOS_Y_CESES	*Nombrada comisionada del gobierno frente al reto demográfico a Edelmira Barreira Diz.*
TRIBUNAL_CONSTITUCIONAL	*Sentencia del TC que anula los límites de Cataluña a la libertad de horarios comerciales.*
SUBVENCIONES	*Subvenciones a sindicatos del Ministerio de Empleo: 8.883.890 euros.*
DECRETO_LEY	*Ley de facturación electrónica. Para contratos con la administración: todos obligados a partir de 15 enero 2015*
TRIBUNAL_DE_CUENTAS	*El Tribunal de Cuentas busca más de 400.000 euros perdidos en liquidación en la Cámara Oficial de la Propiedad Urbana de Vizcaya*
CONVOCATORIAS	*El Ministro de Hacienda y Administraciones Públicas ha dispuesto hacer pública la adjudicación de puestos de trabajo especificados en el anexo a la presente orden.*
PREMIOS_Y_MEDALLAS	*Premio Nacional de Historia de España 2013 a José Ángel Sánchez Asiain*
CONVENIOS	*Ratificación y próxima entrada en vigor (el 1 de abril) del convenio de doble nacionalidad Francia-España firmado hace un año. Permite tener ambas nacionalidades (también recuperarla si se perdió para tener la del otro país)*
TRIBUNAL_SUPREMO	*Karl Friedrich Schober recurre contra la Orden HAP/72/2013, la declaración informativa sobre bienes y derechos situados en el extranjero, ante la Audiencia Nacional*
REGISTRO_EMPRESA_ONG	*Exteriores crea una Oficina Consular Honoraria en Incheon (Corea del Sur) para relaciones económicas.*
ERRORES_MODIFICACIONES_BOE	*Corrección de errores para quitar la marca en la casilla 'trabajador sin especialización' sobre la madre*
ELECCIONES	*Resultados definitivos elecciones municipales, hasta la J de Jaen.*
TRASVASES_MEDIOAMBIENTE	*La ley de contaminación y residuos que entra en vigor mañana. Acorta plazo para conseguir autorización ambiental*
EJECUCION_PRESUPUESTO	*Ejecución del presupuesto en junio.*
INDULTOS	*Indulto a María Salmerón Parrilla aprobado viernes en Consejo de Ministros.*
BANCO_DE_ESPAÑA	*Multa del Banco de España a Austrogiros Entidad de Pago, S.A. por un reguero de incumplimientos de la ley (no tener dirección en España, irregularidades contables, etc...): 1.300.000 euros más multas a los dos administradores e inhabilitación.*

Table 5: Categories annotated for the text categorization task and examples.

C Additional Information about the Dataset

Category	All	Train	Development	Test
OTROS_ANUNCIOS	1004	789	111	104
CONTRATOS_LICITACIONES	648	505	73	70
NOMBRAMIENTOS_Y_CESES	324	255	35	35
TRIBUNAL_CONSTITUCIONAL	311	245	31	34
SUBVENCIONES	228	177	25	26
DECRETO_LEY	173	137	19	17
TRIBUNAL_DE_CUENTAS	161	128	18	17
CONVOCATORIAS	155	120	17	17
PREMIOS_Y_MEDALLAS	152	119	16	16
CONVENIOS	141	112	12	15
TRIBUNAL_SUPREMO	97	75	11	11
REGISTRO_EMPRESA_ONG	82	63	9	10
ERRORES_MODIFICACIONES_BOE	52	44	4	5
ELECCIONES	37	30	3	4
TRASVASES_MEDIOAMBIENTE	33	27	3	2
EJECUCION_PRESUPUESTO	25	21	2	2
INDULTOS	20	16	2	2
BANCO_DE_ESPAÑA	5	4	1	0

Table 6: Distribution of text categories annotated in BOE-XSUM.

Column	Description
id	Unique item identifier.
boe_materials	BOE category identifier.
boe_date_publication	Publication date of the BOE article.
boe_previous	Previous BOE articles that are modified by this new BOE.
boe_id	BOE identifier.
boe_title	Title of the BOE article.
boe_soup_xml	Complete scraped web page.
tweet_original	Original tweet by Eva Belmonte.
boe_category	Category to which this item belongs.
boe_alert	BOE classification codes for government areas.
boe_departament	Government department that issued the BOE article.
tweet_text_cleaned	Extreme summary generated from a thorough review of Eva Belmonte's tweet.
boe_subsequent	Subsequent legislation articles modified by this order (Only for articles referring to laws).

Table 7: Dataset Columns.

Range	Count	Range	Count
90%+	1154	40%+	2946
80%+	1624	30%+	3162
70%+	2028	20%+	3341
60%+	2398	10%+	3482
50%+	2720	0%+	3648

Table 8: Distribution of cosine similarity ranges for the pairs of original and modified tweets.

D Examples of edited summaries

As an example, we edited the summary below because the content of the post gave very little information about the BOE article it referred to:

(1) **Original**: *"Es habitual tratar el tema Microsoft en administraciones públicas, pero solemos olvidar la presencia de Oracle #BOE"*
Edited: *"La Agencia Estatal de Seguridad Aérea contrata a Oracle por el servicio de mantenimiento y soporte técnico por un importe total de 1.232.070,40 euros."*

	# pairs doc/sum	# words doc	# words sum	average # words/doc	average # words/sum	compression rate
Train	2,867	10,859,884	49,461	3,787	17	0.0045%
Dev	392	1,155,745	6,733	2,948	17	0.0058%
Test	389	1,367,413	6,605	3,515	17	0.0048%
Total	3,648	13,383,042	62,799	3,668	17	0.0050%

Table 9: Dataset statistics showing the total number of items, word counts in documents and summaries, mean word count in documents and summaries, and the compression ratio for the train, development, and test sets.

(2) **Original**: *"It is common to talk about Microsoft in public administrations, but we tend to forget the presence of Oracle."*
Edited: *"The Aviation Safety State Agency contracts Oracle for the maintenance and technical support service for a total amount of 1,232,070.40 euros."*

In the next example, we edited the content because the post was talking about people and entities not present in the BOE article.

(3) **Original**: *"La Moncloa tiene que licitar su servicio de restauración de forma urgente porque la empresa anterior que daba este servicio, Dulcinea nutrición -que hasta 2018 presidía uno de los bisnietos de Franco y ahora está en concurso-, dejó de pagar impuestos. #BOE".*
Edited: *"Licitación del servicio de restauración, de limpieza de los espacios destinados a dicha finalidad y de máquinas de venta automática en el Complejo de La Moncloa.".*

(4) **Original**: *"La Moncloa has to tender its catering service urgently because the previous company that provided this service, Dulcinea nutrición - which until 2018 was presided over by one of Franco's great-grandchildren and is now in competition - stopped paying taxes.. #BOE".*
Edited: *"Bidding for catering services, cleaning of the areas destined for this purpose and vending machines in the Moncloa Complex.".*

E Additional Information about Results

Model	BLEU	METEOR	ROUGE	BERTScore
Llama 3.2 1B	0.010	0.137	0.171	0.124
Llama 3.1 8B	0.015	0.185	0.220	0.196
Llama 3.1 70B	0.016	**0.220**	0.214	**0.207**
Llama 3 70B Instruct	0.017	0.175	0.204	0.178
Llama 2 70B	0.009	0.110	0.154	0.117
Gemma 2 2B	0.015	0.183	0.211	0.198
Gemma 2 9B	**0.019**	0.207	0.232	0.213
Gemma 2 27B	0.018	0.193	0.214	0.197
Mistral 7B	0.006	0.094	0.152	0.106
Neural-Chat 7B	0.006	0.110	0.167	0.132
Starling 7B	0.003	0.074	0.134	0.092
Llava 7B	0.002	0.063	0.114	0.069
Solar 10.7B	0.005	0.102	0.159	0.118
Nemotron 70B	0.012	0.143	0.185	0.132
Salamandra 7B Instruct	0.004	0.082	0.127	0.070
ChatGPT 4o	**0.019**	0.208	**0.236**	0.206

Table 10: Results on test set with unmodified social media posts by the journalist. Best scores in **bold**.

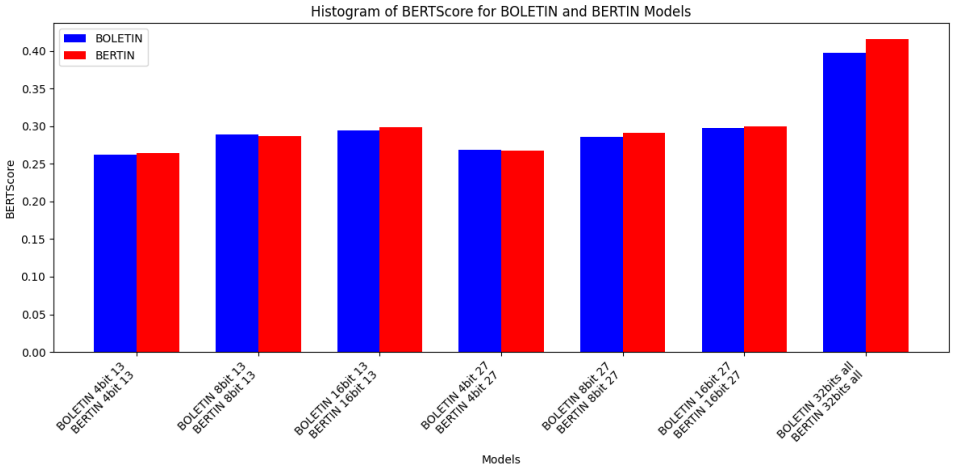

Figure 4: Comparative results between Boletin and Bertin across bit widths and training configurations.

Summary Gemma 2 9B	**F1**	**Summary LLama 2 70B**	**F1**	**Tweet**	**Dif**
José Antonio Pérez Tapias es nombrado Catedrático de Filosofía en la Universidad de Granada.	0.79	La Rectora de la Universidad de Granada, María Pilar Aranda Ramírez, ha nombrado a José Antonio Pérez Tapias como Catedrático de Universidad del área de conocimiento de Filosofía, adscrito al Departamento de Filosofía II, mediante Resolución de 18 de enero de 2017.	0.38	Pérez Tapias ya es Catedrático de Filosofía en la Universidad de Granada.	0.41
Se licita el mantenimiento y soporte técnico de sistemas Oracle de la Agencia Estatal de Seguridad Aérea por valor estimado de 1.018.240 euros.	0.63	El presidente de la mesa de contratación, Eusebio Jiménez Arroyo, anunció que se busca una empresa para el mantenimiento y soporte técnico de los sistemas de información en tecnología Oracle de la Agencia Estatal de Seguridad Aérea. El contrato tendrá una duración de 20 meses y se adjudicará a la empresa que ofrezca los mejores criterios de adjudicación. La fecha límite para la presentación de ofertas es el 3 de junio de 2014.	0.19	La Agencia Estatal de Seguridad Aérea contrata a Oracle por el servicio de mantenimiento y soporte técnico por un importe total de 1.232.070,40 euros.	0.44
El Ayuntamiento de Barcelona licita un sistema para controlar infracciones medioambientales de vehículos en la zona de bajas emisiones.	0.66	El Ayuntamiento de Barcelona invita a licitar para un contrato de suministro e instalación de un sistema de monitorización y control de infracciones medioambientales en la ciudad. El plazo de ejecución es de 11 meses, y se requiere una garantía definitiva del 5% del ... (continue)	0.12	Barcelona licita el contrato para el control de los vehículos en la zona de bajas emisiones.	0.54

Table 12: Table with summaries, F1 scores, and differences.

F English Version of BOE example with original and edited summaries

Royal Decree-Law 6/2012, of March 9, on urgent measures for the protection of mortgage debtors without resources, establishes a series of mechanisms aimed at enabling the restructuring of mortgage debt for those experiencing extraordinary difficulties in making their payments.

To this end, the aforementioned Royal Decree-Law incorporated a Code of Good Practices, to which entities may adhere, and whose compliance will be supervised by a control commission. The composition of this commission was modified by Article 6 of Law 1/2013, of May 14, on measures for...

... – Caja Rural San Jaime de Alquerías Niño Perdido, S. Coop. de Crédito V. – Caja Rural San José de Almassora, S. Coop. de Crédito V. – Caja Rural San José de Burriana, S. Coop. de Crédito V. – Caja Rural San José de Nules, S. Coop. de Crédito V. – Caja Rural San Roque de Almenara, S. Coop. de Crédito V. – Colonya–Caixa d'Estalvis de Pollença – Liberbank, S.A. – Publicredit, S.L. – UNOE Bank, S.A.

The original summary for this BOE article as written by the journalist is as follows:

Adhesions to 2 codes of good practices in mortgages. So far, not very effective. Because legislating is hard, right? #BOE

The resulting edited summary:

List of bank adhesions to the codes of good practices to strengthen the protection of mortgage debtors, debt restructuring, and social rent

G Annotation Guidelines

To improve consistency, objectivity, and reliability, these annotation guidelines will help the annotators create concise, accurate, and legally faithful extreme summaries of BOE texts.

1. General Principles

Each summary must:

- Be concise – Between 15 and 22 words.
- Be factual – Avoid opinions, subjective language, or speculation.
- Preserve critical legal information – Names, amounts, dates, and key legal actions must be included.
- Avoid redundancy – Keep the summary clear and direct.
- Use simple but precise language – The goal is to make legal information more accessible.

2. Formatting Rules

- Use neutral, third-person language (No opinions, interpretations, or assumptions).
- Avoid legal jargon unless necessary – If a legal term is essential, keep it. If it can be simplified without changing meaning, do so.
- No abbreviations unless widely known (e.g., BOE, EU, UN are fine).

Bad Example ✗ (Too vague & informal)

"Another BOE article about a contract for an unnamed company."

Good Example ✔ (Clear, factual, and legally informative)

"The Ministry of Transport awards a €2M contract to Ferrovial for highway maintenance in Madrid."

3. Information Hierarchy

What to include first? If space is limited, prioritize information in this order:

Priority	Include in Summary	Example
①	Main Legal Action	"The government approves a new tax reduction for self-employed workers."
②	Who is involved?	"The Ministry of Defense signs a €5M cybersecurity contract with IBM."
③	Financial or Legal Consequences	"A judge orders a company to pay €1.2M for contract fraud."
④	Date or Location (if critical)	"A new rental law takes effect in Barcelona on January 1, 2025."

If space is limited, drop information from the bottom of the list first.

4. Common Errors & How to Avoid Them

Subjectivity or Opinions

✗ "A useless new law that won't change anything."

✔ "The government introduces a new environmental protection law."

Missing Critical Information

✗ "A contract was signed." (Who signed it? What for?)

✔ "The Ministry of Health awards a €3.5M contract for hospital equipment."

Overly Complex Legal Language

✗ "In accordance with Royal Decree 1892/2024, a framework agreement has been instituted for the execution of infrastructural oversight."

✔ "The government approves a contract for highway inspections."

Misleading Summaries

✗ "The Supreme Court supports corruption!" (Misinterpretation)

✔ "The Supreme Court rules against a corruption conviction in X case."

5. Handling Difficult Cases

A) Summarizing Lengthy Legal Texts

- If a BOE article is very long, focus on the main action in the first few paragraphs.
- **BOE Text (Full Version)**:
  "The government issues a new decree modifying Article 10 of the Labor Code, changing workplace safety regulations."
- **Summary**:
  "New decree modifies workplace safety regulations in Spain."

B) Appointments & Dismissals

- Format: "[Person] appointed/dismissed as [position] at [institution]."
- **BOE Text**: "By Royal Decree 2024/317, José Pérez is appointed Minister of Finance."
- **Summary**: "José Pérez appointed Minister of Finance by Royal Decree."

C) Court Rulings

- Mention court name, ruling outcome, and key details.
- **BOE Text**: "The Constitutional Court overturns a law restricting media freedom."
- **Summary**: "The Constitutional Court annuls media restriction law."

D) Large Financial Figures

- If space is limited, round numbers where possible.
- **Example**:
  "The Ministry of Transport signs a €2,345,678.90 contract for roads."
  Can be summarized as:
  "€2.3M contract for roadworks by Ministry of Transport."

5. Final Checklist for Annotators

- Did you keep the summary between 15–22 words?
- Is the summary neutral and free of opinion?
- Does it capture the key legal action (law, contract, ruling, appointment)?
- Does it include important details (names, amounts, institutions)?
- Did you avoid excessive jargon?
- Would a non-lawyer understand it?