Spanish Fake News Classifier: A Sequential Fine-Tuning Approach

Juan Ignacio Llaberia

2025

Abstract

Fake news detection in Spanish presents significant challenges due to the scarcity of high-quality annotated data and the variability in article length. While Transformer-based models have demonstrated strong performance on short texts, their effectiveness on long-form news articles remains limited, particularly in low-resource settings where large annotated datasets are unavailable. In this work, we propose a Sequential Fine-Tuning approach for Spanish fake news classification using a BETO variant, `dccuchile/bert-base-spanish-wwm-cased`. Our method consists of a multi-step fine-tuning process in which the model is first adapted to a large dataset of approximately 50k short news articles (pre-fine-tuning) and subsequently fine-tuned on a smaller target dataset of around 2k long-form articles.

This progressive adaptation enables the model to learn task-specific representations while improving generalization to longer documents. Experimental results on the test set show that our approach achieves an accuracy of 0.8205, a precision of 0.7835, a recall of 0.8702, and an F1-score of 0.8246, demonstrating improved performance and stable convergence compared to direct fine-tuning on long documents. These findings highlight the importance of task-specific adaptation and provide a practical strategy for fake news detection in low-resource, long-document Spanish news scenarios.

1 Introduction

The rapid spread of fake news has become a major challenge for modern societies, influencing public opinion, political processes, and trust in journalism. As online news consumption continues to grow, the development of automated systems capable of distinguishing between real and fake news has gained increasing importance. While substantial progress has been made in English, fake news detection in Spanish remains comparatively underexplored, primarily due to the limited availability of high-quality annotated datasets and the linguistic diversity across Spanish-speaking regions.

Recent advances in Transformer-based language models, such as BERT, have significantly improved performance across a wide range of natural language processing tasks, including text classification. Spanish-specific models like BETO further enhance these capabilities by leveraging large-scale pre-training on Spanish corpora. However, most prior work on fake news detection has focused on short texts, such as headlines or social media posts, where contextual complexity is limited. In contrast, long-form news articles introduce additional challenges, including richer discourse structure, higher semantic density, and increased sensitivity to training instability during fine-tuning.

A critical limitation in existing Spanish fake news research is the scarcity of large, high-quality datasets containing long-form articles. As a result, traditional single-stage fine-tuning approaches often struggle to generalize when trained directly on small long-article datasets. Moreover, mixing short and long articles during training can introduce length-based biases, allowing models to rely on superficial cues such as token count rather than meaningful linguistic

or semantic patterns. This issue is particularly problematic in real-world deployment scenarios, where long-form journalistic content is predominant.

To address these challenges, we propose a Sequential Fine-Tuning approach for Spanish fake news classification. Instead of training a model directly on the target long-article dataset, we adopt a multi-stage transfer learning strategy that progressively adapts a Spanish pre-trained Transformer model. Specifically, the model is first fine-tuned on a large corpus of short news articles to learn general fake-news-related linguistic patterns, and subsequently fine-tuned on a smaller dataset of long-form articles to adapt to their structural and semantic complexity. This staged training process enables effective knowledge transfer while mitigating length bias and overfitting.

The main contributions of this work are as follows:

- We introduce a sequential fine-tuning framework for Spanish fake news detection that explicitly addresses the challenges posed by long-form articles and limited labeled data.
- We demonstrate that progressive adaptation from short to long texts improves model stability and classification performance compared to direct fine-tuning.
- We provide a comprehensive empirical evaluation using multiple Spanish fake news datasets, showing strong performance on long-form news articles with limited supervision.

Overall, this work highlights the importance of task-specific and data-aware training strategies for fake news detection in low-resource settings and offers a practical solution for real-world Spanish-language news classification systems.

2 Hypothesis

We hypothesize that a sequential fine-tuning approach, first training on a large dataset of short articles and then adapting to a smaller dataset of long articles as we will talk about in Section 4, will improve classification performance on long-form news. This strategy should reduce length bias and encourage the model to rely on semantic and structural patterns rather than superficial features such as article length allowing us to reach more reliable results.

3 Data

Our work uses two datasets of different sizes and article lengths. The first dataset contains approximately 57,000 short news articles (referred to as the 50k dataset). It was created by merging two sources with similar text length: a dataset developed for a thesis project at the University of Vigo [1], and another publicly available dataset obtained from Kaggle [2].

The second dataset includes a little over 2,000 long-form news articles (the 2k dataset). It combines texts from two main sources: the Spanish Fake and Real News dataset, originally compiled for a master's thesis in cybersecurity at the Polytechnic University of Madrid [3], and the Fake News Corpus Spanish from GitHub [4], which contains articles from various Spanish-speaking regions such as Peru, Uruguay, and Venezuela.

Once all data sources were consolidated, we carried out two main steps: pre-processing and data splitting. For pre-processing, we used the model's tokenizer to encode all articles. Because the model supports a maximum context length of 512 tokens, we had to apply text truncation to longer inputs.

We considered two truncation strategies:

1. **Head Truncation** - Uses the first 512 tokens of the article.
    - *Pros:* Computationally efficient and preserves the introductory context, where the main claim is often presented.

- *Cons:* Removes the ending of the article, potentially losing conclusions or later contradictions.

2. **Head-Tail Truncation** - Keeps the first n tokens and the last m tokens, then concatenates them.
   - *Pros:* Captures both the introduction and the conclusion.
   - *Cons:* More computationally expensive and may disrupt the natural semantic flow of the text, which can affect the model's attention patterns.

For this study, we chose head truncation in order to prioritize computational efficiency. In future work, we plan to experiment with the head-tail strategy to assess whether it provides meaningful performance improvements.

After completing the pre-processing steps, we split the datasets as follows. The 50k dataset was divided into 80% for training and 20% for validation. For the 2k dataset, we adopted a three-way split consisting of 70% for training, 10% for validation, and 20% for testing.

3.1 Why we split it in two

During our analysis, we identified a critical issue: length imbalance between the two datasets. Mixing both datasets directly would introduce a length-based bias. In such a scenario, the model could learn to associate text length with class labels, for example, predicting fake for shorter articles and real for longer ones, rather than focusing on meaningful linguistic or semantic patterns. This is problematic, especially because our target production environment involves primarily long-form articles. Therefore, the model must learn structural and semantic features, not rely on superficial cues such as token count.

With both datasets prepared and this limitation understood, we proceeded to train our model using a sequential fine-tuning approach applied to the chosen transformer architecture.

4 Methodology

To address the challenge of classifying long-form news articles with limited labeled data, we propose a Sequential Fine-Tuning framework. Instead of training a model directly on the target dataset, we adopt a multi-stage transfer learning [5] approach designed to maximize knowledge retention and generalization. Our pipeline begins with a robust pre-trained Spanish language model, proceeds through an intermediate adaptation phase using a large corpus of short articles, and concludes with targeted fine-tuning on the final long-form dataset. This strategy allows the model to incrementally acquire domain-specific knowledge, first learning general veracity cues and subsequently refining its decision boundaries for complex document structures. Thereby ensuring superior performance compared to standard direct training methods, that we have tried and got poor results.

4.1 Base Model

For our foundational architecture, we selected `dccuchile/bert-base-spanish-wwm-cased` [6], a specialized variant of the BETO model designed specifically for the Spanish language. This model uses Whole Word Masking and was pre-trained on a diverse and extensive corpus, including Wikipedia, news articles, and various digital sources. With a vocabulary size of approximately 30,000 tokens, it possesses a robust capability for capturing Spanish syntax and semantic nuances.

The primary benefit of utilizing this pre-trained model is its ability to generate rich contextual embeddings immediately, without the need for training from scratch. Its specific optimization for Spanish, rather than a generic multilingual approach, provides a superior initialization point,

allowing us to achieve higher accuracy on our specific classification task while reducing the computational cost of fine-tuning.

4.2 Sequential Fine-Tuning

We implemented a Sequential Fine-Tuning strategy for our transformer model, consisting of two stages: Pre-Fine-Tuning and Target Fine-Tuning.

Stage 1: Pre-Fine-Tuning

The first stage involves training the model on the 50k short-article dataset using default recommended hyperparameters. The objective of this stage is for the model to learn general patterns and linguistic cues associated with fake news, which can then be transferred to the second stage.

To preserve the pretrained knowledge of `bert-base-spanish-wwm-uncased` [6] while allowing adaptation, we froze the majority of the model parameters, leaving only the following trainable:

- The last 3 encoder layers
- The pooler layer
- The classifier head

This results in approximately 14 million trainable parameters out of a total of ∼110 million of the base model:

Category	**Parameters**
Trainable	14,767,874
Frozen	95,084,544
Total	**109,852,418**

Table 1: Stage 1 Parameter Distribution

This setup allows the model to capture general "fake news" linguistic patterns while limiting overfitting on short-text articles.

Stage 2: Target Fine-Tuning

In this final phase, the model was initialized using the weights derived from Stage 1. To adapt the model specifically to the structure of long-form articles, we applied a selective freezing strategy. We froze all initial layers, keeping only the final two transformer layers, the pooling layer, and the classifier head trainable. This approach allowed us to refine the decision boundary for the target data without distorting the linguistic features learned in previous stages. The distribution of trainable versus frozen parameters for this stage is detailed below:

Category	**Parameters**
Trainable	7,680,002
Frozen	102,172,416
Total	**109,852,418**

Table 2: Stage 2 Parameter Distribution

The primary objective of this sequential transfer learning approach is to leverage the foundational knowledge from the base model combined with the specific adaptations from Stage 1. By restricting updates to the final layers, we transfer these learned patterns effectively, enabling the model to generalize to long-form news articles.

4.3 Training Setup

During Stage 1, as previously discussed, we used the default recommended hyperparameters without performing a dedicated hyperparameter search, as [9] explains that BERT models tend to converge quickly.

In contrast, when training Stage 2, our goal was to ensure optimal convergence on the smaller, domain-specific target dataset. To achieve this, we conducted a comprehensive hyperparameter search to identify the most effective configuration. This optimization process yielded the following values:

- Learning rate: 9.9×10^{-5}
- Batch size: 32
- Training epochs: 4

Additionally, we applied a weight decay of 0.01 and a warmup ratio of 0.1. These regularization strategies were essential for stabilizing the training process and mitigating overfitting.

5 Evaluation & Results

The first stage was trained on the short-text dataset for 3 epochs, following the standard recommendation that BERT [8] models tend to converge quickly and can overfit when exposed to excessive iterations on narrow-domain data. The results showed steady improvements across all metrics:

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.218800	0.193999	0.928802	0.931096	0.947487	0.939220
2	0.199600	0.176485	0.937364	0.949780	0.941920	0.945834
3	0.170900	0.168711	0.943479	0.944700	0.958772	0.951684

Table 3: Stage 1 (Short Text) Training Results

While the short-text domain does not fully reflect production conditions, these results confirm that the model successfully learned generalizable linguistic patterns associated with fake news.

In the second stage, the partially pretrained model was fine-tuned on the long-article dataset, the domain that matches the target real-world deployment environment. Training was extended to 4 epochs, since adaptation to longer, more complex documents requires deeper semantic alignment. The training logs and validation metrics for this stage are shown below:

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.660900	0.544833	0.733645	0.818182	0.637168	0.716418
2	0.466700	0.455799	0.780374	0.761905	0.849558	0.803347
3	0.408400	0.437711	0.803738	0.770992	0.893805	0.827869
4	0.404900	0.425096	0.808411	0.776923	0.893805	0.831276

Table 4: Stage 2 (Long Text) Training Results

Fine-tuning consistently improved performance across epochs, with notable gains in recall, a critical metric for fake-news detection.

5.1 Final Evaluation on the Test Set

After training, the final model was evaluated on an unseen test set to measure its real-world generalization capacity. The results were:

Metric	Value
eval_loss	0.4183
eval_accuracy	0.8205
eval_precision	0.7835
eval_recall	0.8702
eval_f1	0.8246
eval_runtime (s)	14.0762
eval_samples_per_second	30.4770
eval_steps_per_second	1.9180
epoch	4.0

Table 5: Final Test Set Evaluation

The final model reaches an F1 score of 0.8246, with particularly strong recall (0.8702), aligning with the project goal of maximizing the correct detection of fake news in long-form Spanish journalism.

5.2 Extended Evaluation & Model Behavior Analysis

To understand the classifier beyond aggregate metrics, we conducted a deeper set of diagnostic evaluations:

Confusion Matrix Analysis Provided a detailed breakdown of true/false positives and negatives, helping identify systematic tendencies such as over-flagging borderline cases.

Confidence Distribution Allowed us to assess how the model distributes probability scores across both classes, revealing calibration characteristics.

Error Analysis Focused particularly on high-confidence misclassifications (confidence > 0.8), which are the most critical failure mode in production systems.

Article Length Sensitivity Since earlier experiments uncovered strong risks of length bias, we tested the model across different article lengths to ensure stable behavior. Results confirmed that the sequential fine-tuning strategy successfully mitigated length-based misclassification.

5.3 Summary of Findings

The sequential fine-tuning approach proved effective, enabling the model to transfer general fake-news patterns from short texts while adapting to the semantic complexity of long-form articles. The final classifier demonstrates robust recall, balanced precision, and consistent performance across article lengths, providing a strong foundation for future refinements.

6 Conclusion & Future Work

This project demonstrates that a Sequential Fine-Tuning strategy is both viable and effective for developing a Spanish fake news classifier capable of processing long-form journalistic content despite significant data limitations. By dividing the training process into two stages, we successfully mitigated the length bias problem and ensured that the model learned meaningful

semantic patterns rather than relying on superficial correlations such as article length. The final classifier achieved an accuracy of 82.05% and an F1-score of 82.46% on the test set, confirming its robustness when applied to complex, long-form articles.

Future Work

- **Expand the long-article dataset:** Increasing the size of the long-form corpus could significantly improve the model's generalization and reduce overfitting risks.
- **Explore alternative truncation strategies:** Approaches such as head-tail truncation, hierarchical chunking, or sliding-window encoding may help preserve a broader range of contextual information.
- **Evaluate domain robustness:** Testing the classifier on news sources from additional Spanish-speaking regions and across different time periods would help assess its stability under domain shift.
- **Improve interpretability:** Incorporating attention-based or gradient-based explainability techniques may provide deeper insights into the linguistic or structural features influencing predictions.

References

[1] Javier Otero Vizoso, "Spanish Political Fake News", University of Vigo. `https://www.kaggle.com/datasets/javieroterovizoso/spanish-political-fake-news`

[2] Arsenii Tretiakiv, "Noticias falsas en español - Kaggle" `https://www.kaggle.com/datasets/arseniitretiakov/noticias-falsas-en-espaol`

[3] Fabricio A. Zules, "Spanish Fake and Real News" Universidad Politécnica de Madrid. `https://www.kaggle.com/datasets/zulanac/fake-and-real-news`

[4] Posadas-Durán, J. P. (n.d.). Fake News Corpus Spanish [Dataset]. GitHub repository. `https://github.com/jpposadas/FakeNewsCorpusSpanish`

[5] Gómez-Adorno, H., Posadas-Durán, J. P., Enguix, G. B., & Capetillo, C. P. (2021). Overview of FakeDeS at IberLEF 2021: Fake News Detection in Spanish Shared Task. *Procesamiento del Lenguaje Natural*, 67, 223–231.

[6] Aragón, M. E., Jarquín, H., Gómez, M. M. Y., Escalante, H. J., Villaseñor-Pineda, L., Gómez-Adorno, H., et al. (2020). Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. In *Proceedings of the 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF)*, Málaga, Spain.

[7] Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., & Escobar, J. J. M. (2019). Detection of fake news in a new corpus for the Spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5), 4869–4876.

[8] S. Ruder, "An overview of multi-task learning in deep neural networks," arXiv:1706.05098, 2017.

[9] dccuchile/bert-base-spanish-wwm-cased - DCC UChile University, Hugging Face, 2020. `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased`

[10] Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. arXiv: 2308.02976, 2020.

[11] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.