Fine-tuning BERTIN for Detecting Humor in News Headlines for HAHA at IberLEF 2026

Edder O. Quiñones-Burgos[1,*,†], Jairo Serrano Castañeda[1,†]

[1]Universidad Tecnologica de Bolivar, Cartagena; Colombia

Abstract

This article presents the solution we proposed for Task 1 of HAHA at IberLEF 2026: Human Annotation-Based Humor Analysis and Automatic Humor Generation. This task involves detecting whether a Spanish news headline is satirical or factual. For this task, we tuned BERTIN, a RoBERTa-based model pre-trained exclusively on Spanish text, using 5-fold stratified cross-validation. We tested various input configurations, such as learning rate and training epochs, and found that using only the headline text with 15 epochs and a learning rate of 2e-5 yielded the best F1 score. The results showed an F1 score of 0.65 in the development phase, 0.59 in the evaluation phase, and 0.78 in the post-evaluation phase.

Keywords

Natural Language Processing, Deep learning, Pre-trained model, computational humor, humor detection

1. Introduction

Humor is an extremely complex human capacity that requires many cognitive tasks, ranging from linguistic knowledge to the use of cognitive skills that lead to abstract thinking and socio-cultural perception [1]. This same complexity is reflected in attempts to detect it, making it a significant challenge for natural language processing.

Now, if we add to this inherent complexity of humor a more specific characteristic, such as the context of Spanish-language media, we transform this detection into an extremely complex challenge. Distinguishing satirical headlines from real ones presents a major problem, since satirical content often mimics the style and tone of legitimate journalism, making the boundary between the two categories subtle and, at times, ambiguous. Furthermore, Spanish is a very rich and expressive language, where even speakers of the same language perceive humor differently, depending on the context or even geographical location [2].

The shared challenge Humor Analysis Based on Human Annotations and Automatic Humor Generation (HAHA) [3], part of IberLEF 2026 [4], proposed three tasks: Humor Detection, which consisted of classifying whether a news headline was a joke or not; LLM-generated humor detection, which determined whether a joke inspired by a news headline was generated by an LLM or written by a human; and the final task, Humor Generation, which consisted of generating jokes from a news headline using computational methods. This paper presents a system developed to solve only Task 1, humor detection, which uses the F1 score of the "satirical" class as its main performance metric for this subtask.

The HAHA shared challenge has already been carried out in some previous editions (2018, 2019, 2021), where the tasks were applied to humor detection in Twitter data (now called X) [5] [6] [7]. The current edition focuses on a completely different and much more complex area, which is news headlines. This specific area makes the difficulty increase because distinguishing between humorous and non-humorous content can be more difficult, which can even confuse human readers.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

equinones@utb.edu.co (E. O. Quiñones-Burgos); jserrano@utb.edu.co (J. S. Castañeda)

https://github.com/EdoQB (E. O. Quiñones-Burgos); https://github.com/jairoserrano (J. S. Castañeda)

0009-0007-1194-3775 (E. O. Quiñones-Burgos); 0000-0001-8165-7343 (J. S. Castañeda)

Since we were discussing the Spanish language and focusing on specific tasks, the chosen strategy was a transfer learning approach using BERTIN, a language model based on RoBERTa that has already been pre-trained on large Spanish corpora. This was combined with a cross-validation strategy to maximize the use of the available labeled data.

2. Task Description

HAHA proposed a task to automatically classify, analyze, generate, and detect humor in Spanish from news headlines[3]. This task was divided into three subtasks.

- **Humor Detection:** The objective was to determine whether a news headline was a joke or not. The primary performance metric was the F1 score for the satirical class[3].
- **LLM-generated humor detection:** The objective was to determine whether a joke inspired by a news headline was generated by an LLM or written by a human. The primary performance metric for this subtask was the F1 score for the "automatic" class[3].
- **Humor Generation:** The objective was to generate jokes from a news headline using computational methods. Performance was evaluated using human preference judgments[3].

The dataset was provided by the organizers and consisted of Spanish news headlines accompanied by a brief context field. Each data point contained a unique identifier, the country of origin, the publication date, the headline text, and a brief contextual description. The labels were binary: satirical or real, where satirical was the target class. During the development phase, the labeled training data consisted of 24 annotated examples, divided equally between the two classes (12 satirical and 12 real). The development set contained 540 unlabeled instances; subsequently, for the official testing phase, the test set for final evaluation consisted of 600 unlabeled instances[3].

3. System Description

The system is based on fine-tuning a pre-trained Spanish language model for binary sequence classification (Satirical or Real). Given that we are dealing with text and language, and considering that the size of the labeled training data was small compared to larger models, we opted for a transfer learning approach to leverage the linguistic knowledge already encoded in large pre-trained models.

3.1. Pre-trained Model

Two Spanish language models were evaluated: BETO (dccuchile/bert-base-spanish-wwm-cased), a BERT-based model pre-trained on a large Spanish corpus[8], and BERTIN (bertin-project/bertin-roberta-base-spanish), a RoBERTa-based model also pre-trained on Spanish text [9]. In the comparison, BERTIN consistently outperformed BETO in the experiments and was selected as the final model.

3.2. Input Representation

Two input configurations were tested. One option, called "headline_only", used only the headline text as input. The other option, called "headline_context", concatenated the headline with the context field provided in the data. This second option was done by separating them with a [SEP] token and truncating the context to 200 characters. Of the two options, the "headline_only" configuration yielded better results, suggesting that the truncated context introduced more noise than useful information.

3.3. Training Setup

The model was tuned to perform binary classification of the output sequences, which were satirical and real. A 5-fold stratified cross-validation strategy was used to maximize the use of training data

and obtain reliable performance estimates. Two training runs were conducted: one for the official evaluation phase and the other for the post-evaluation phase. The model for the evaluation phase was trained on the 24 examples, while the model for the post-evaluation phase was trained on 540 examples derived from the development data. Both models were trained using the best and same hyperparameter configuration found during experimentation. Table 1 summarizes the final hyperparameters.

Table 1
Hyperparameters used for fine-tuning

Hyperparameter	Value
Base model	bertin-project/bertin-roberta-base-spanish
Max sequence length	128
Batch size	8
Epochs	15
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup ratio	0.1
Optimizer	AdamW

3.4. Threshold Optimization

After training, the optimal classification threshold was sought within the range [0.10, 0.90], using out-of-sample predictions to maximize the F1 score of the satirical class. The optimal threshold found was 0.50, which coincides with the default value, indicating that the model was well-calibrated.

4. Experiments

All experiments described in this section were conducted during the challenge development phase, using the development data provided at this stage, which consisted of 24 labeled examples. These were evaluated using 5-fold stratified cross-validation. The values shown in this section are internal estimates used to guide model selection and do not reflect the official evaluation scores, which are presented separately in Section 5. Table 2 summarizes and shows all the results obtained in this section.

4.1. Model Selection

Initially, two pre-trained Spanish models, BETO and BERTIN, were compared, keeping all other hyperparameters constant in both models (10 epochs, learning rate 2e-5, input mode title only). As shown in Table 2 , BERTIN outperformed BETO by a significant margin, both in the mean internal F1 score and in inter-fold stability.

4.2. Input Configuration

Since BERTIN gave us better results, we kept it as the base model. Subsequently, we compared the two input modes, which were already described in Section 3.2. The headline_only configuration performed better than headline_context, suggesting that the truncated context added noise instead of useful information.

4.3. Number of Training Epochs

The effect of changing the number of training epochs was evaluated using BERTIN with only starters as input. Performance showed improvements from 10 to 15 epochs, but when this value was increased, no further improvement was observed at 20 epochs, indicating that the model converged at 15 epochs.

4.4. Learning Rate

Two learning rates were tested: 2e-5 and 1e-5. Both were tested with the best configuration found to date (BERTIN, headline_only, 15 epochs). The results showed that the 1e-5 rate yielded lower performance, indicating that 2e-5 is the optimal learning rate for this configuration.

4.5. Additional Post-evaluation Experiment

Following the official evaluation period, we conducted an additional experiment using the labeled development set (540 examples) as training data. This setup, which was not used for the official submission, was the same as that used for the official evaluation phase.

Table 2
Model Performance Comparison

Model	Input	Epochs	LR	CV F1	Std
BETO	headline_only	10	2e-5	0.7076	±0.1774
BERTIN	headline_only	10	2e-5	0.8133	±0.1067
BERTIN	headline_context	10	2e-5	0.7267	±0.0523
BERTIN	headline_only	15	2e-5	0.8400	±0.0800
BERTIN	headline_only	20	2e-5	0.8400	±0.0800
BERTIN	headline_only	15	1e-5	0.7867	±0.1222

Note: CV F1 represents the average F1-score from Cross-Validation. The values shown in this section are internal estimates used to guide model selection and do not reflect the official evaluation scores

5. Results

Table 3 presents the official results given by HAHA for our system in the test set; in addition, the result of the post-evaluation experiment performed with additional labeled data is also presented, this result also given by HAHA.

Table 3
Official Test Performance and Training Data Breakdown

System	Phase	Training data	Test F1 (official)
BERTIN + headline_only + 15ep	Development	24 (trial)	0.65*
BERTIN + headline_only + 15ep	Evaluation	24 (trial)	0.59
BERTIN + headline_only + 15ep	Post-Evaluation	540 (dev)	0.78*

Note: *Not part of the official evaluation period.

Table 4
Full results of the official evaluation phase

Rank	Precision	Recall	F1	Accuracy
15/16	0.7225	0.5033	0.5933	0.6550

As shown in Table 3, the system achieved an F1 score of 0.59 on the test set during the official evaluation phase; in addition, Table 4 shows the complete results obtained in this same phase. This increased to 0.65 during the development phase and to 0.78 in the post-evaluation. These changes

in results are primarily attributed to the change in the size of the labeled training data. With only 24 examples available during the development phase, the model tended to overfit the test set. This hypothesis is supported by the post-evaluation experiment, in which, using the exact same model architecture and hyperparameters but now with 540 labeled examples (labeled development set), the results showed an F1 score of 0.78. This 19-point improvement confirms that data scarcity was the main bottleneck for our system, not the model architecture or the training procedure.

6. Conclusions

This article presented a system for Task 1 of the HAHA shared task at IberLEF 2026, which addresses the detection of satirical or truthful news headlines in Spanish. The proposed approach is based on fine-tuning BERTIN, a RoBERTa-based model pre-trained with Spanish text. A 5-fold stratified cross-validation strategy was also used to maximize the use of available labeled data.

Internal tests demonstrated that using only the headline text as input was superior to using a combination of the headline and context. Furthermore, 15 training epochs with a learning rate of 2e-5 yielded the best internal results. The final system achieved an official F1 score of 0.59 on the test set, but with post-assessment adjustments, it reached an F1 score of 0.78, which was also reported by the HAHA Codabench website.

The main findings of this study indicate that the primary limiting factor was not the model architecture (although it can be significantly improved), but rather the size of the labeled training data. The difference is clear: with only 24 annotated examples available during the development phase, the model failed to adequately generalize to the test data. This is demonstrated by the large discrepancy between the internal cross-validation score (0.84) and the official test score (0.59). This theory is reinforced by the post-test experiment, in which, unlike the previous one, training was performed with 540 labeled examples, raising the F1 test score to 0.78.

For future work, other strategies could be explored, such as back-translation or the use of larger pre-trained models. Above all, a key lesson learned from this participation is the importance of using all available labeled data for training, including the labeled development set when provided by the organizers, as this can have a significant impact on the final results.

Acknowledgments

Thanks to the Technological University of Bolívar for the postgraduate scholarship, and to the organizers HAHA and IberLEF 2026 for the idea and data to create the system and to the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) and Gemini (Google) to assist in the writing and editing (text translation, identification and correction of grammatical errors) of the manuscript and to generate LaTeX/Markdown code for tables.

After using these tools/services, the author(s) reviewed and edited the content as necessary and assume full responsibility for the content of the publication.

References

[1] J. Polimeni, J. Reiss, The first joke: Exploring the evolutionary origins of humor, Evolutionary Psychology 4 (2006). doi:10.1177/147470490600400129.

[2] A. Reimann, Intercultural communication and the essence of humour, 2010. URL: https://api.semanticscholar.org/CorpusID:150113779.
[3] L. Chiruzzo, I. Sastre, G. Rey, A. Martínez, A. Rosá, S. Góngora, S. Castro, V. Amoroso, J. Prada, J. Conde, G. Moncecchi, Overview of HAHA at IberLEF 2026: Humor Analysis based on Human Annotation and Automatic Humor Generation, Procesamiento del lenguaje natural 77 (2026).
[4] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.
[5] S. Castro, L. Chiruzzo, A. Rosá, Overview of the haha task: Humor analysis based on human annotation at ibereval 2018, in: Proceedings of IberEval@SEPLN, 2018, pp. 187–194.
[6] L. Chiruzzo, S. Castro, M. Etcheverry, D. Garat, J. J. Prada, A. Rosá, Overview of haha at iberlef 2019: Humor analysis based on human annotation, in: Proceedings of IberLEF@SEPLN, 2019, pp. 132–144.
[7] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. A. Meaney, R. Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish, Procesamiento del Lenguaje Natural 67 (2021) 257–268.
[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[9] J. de la Rosa, E. G. Ponferrada, P. Villegas, P. G. de Prado Salas, M. Romero, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. doi:10.48550/arXiv.2207.06814.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.