IReL_IIT(BHU) at HAHA 2026: A Fine-Tuning Pipeline for Satire and LLM-Generated Humor Detection in Spanish News Headlines

Arjun Mukherjee*, Sukomal Pal

Indian Institute of Technology (Banaras Hindu University), Varanasi, India

Abstract

This paper describes the participation of team IReL_IIT(BHU) in Subtasks 1 and 2 of the HAHA 2026 shared task at IberLEF 2026. Subtask 1 (Humor Detection) requires classifying Spanish news headlines as *satirical* or *real*, while Subtask 2 (LLM-generated Humor Detection) requires determining whether a joke written in response to a headline was authored by a human or generated by a large language model (LLM). For both subtasks, we fine-tune BETO as a binary sequence-pair classifier using a shared regularization pipeline designed to mitigate overfitting on the relatively small training sets. Our system obtains an F1 score of **0.5650** on Subtask 1 and an F1 score of **0.8330** on Subtask 2, ranking last among participating teams on Subtask 1 and 6th out of 16 submissions on Subtask 2. We analyze the substantial performance gap between the two subtasks, finding that despite class-balanced training data, our Subtask 1 model exhibited a conservative bias toward the *real* class (precision 0.8630, recall 0.4200), whereas the same pipeline generalized considerably better to the surface-level cues distinguishing human- from LLM-written jokes in Subtask 2.

Keywords

Humor Detection, LLM-Generated Text Detection, Spanish Natural Language Processing, BETO, Computational Humor, IberLEF 2026

1. Introduction

Computational humor processing remains a challenging task in Natural Language Processing, as humor is inherently subjective, culturally grounded, and often relies on figurative or context-dependent language [1]. The HAHA shared task series, organized within the IberLEF evaluation campaign, has provided a recurring benchmark for humor detection and analysis in Spanish [2, 3, 4].

The HAHA 2026 edition [5], part of IberLEF 2026 [6], introduces three subtasks centered on Spanish-language news content: (1) *Humor Detection*, distinguishing satirical from real news headlines; (2) *LLM-generated Humor Detection*, determining whether a joke responding to a headline was written by a human or generated by an LLM; and (3) *Humor Generation*, producing jokes from news headlines, evaluated via pairwise human preference judgments.

This paper describes the participation of team IReL_IIT(BHU) in Subtasks 1 and 2. For both subtasks we adopt a common approach: fine-tuning BETO [7], a BERT model pre-trained on Spanish text, as a binary sequence-pair classifier, combined with a set of regularization techniques aimed at preventing overfitting on the relatively small amount of labeled training data available for each subtask. We report our official results, compare our performance against the competition leaderboard, and provide an error analysis that highlights a marked difference in how well the same pipeline generalized across the two subtasks.

The remainder of this paper is organized as follows. Section 2 reviews related work on Spanish humor detection and LLM-generated text detection. Section 3 describes the datasets used for both subtasks. Section 4 details our experimental setup and methodology. Section 5 describes the evaluation metrics.

IberLEF 2026, September 2026, León, Spain

* Corresponding author.

arjunmukherjee.rs.cse23@itbhu.ac.in (A. Mukherjee); spal.cse@itbhu.ac.in (S. Pal)

0009-0009-4291-9625 (A. Mukherjee); 0000-0001-8743-9830 (S. Pal)

Section 6 and Section 7 report our results and competition standings, respectively. Section 8 presents our error analysis, and Section 9 concludes the paper.

2. Related Work

Automatic humor detection in Spanish has been the focus of several editions of the HAHA shared task. The first edition, HAHA at IberEval 2018 [2], introduced a corpus of 20,000 crowd-annotated Spanish tweets and proposed two subtasks: humor detection and funniness score prediction. The 2019 edition [3] continued with the same subtasks over an expanded corpus, while HAHA at IberLEF 2021 [4] extended the task with two additional subtasks focused on humor mechanism and target classification. Across these editions, transformer-based approaches, including BERT-based models fine-tuned on the provided corpora, have consistently been among the top-performing systems, though humor detection in Spanish remains a difficult task due to the cultural and contextual nature of humor.

Beyond the HAHA series, satire and irony detection in Spanish has also been addressed through neural architectures that combine multiple linguistic perspectives. Ortega-Bueno et al. [8] propose an attentive LSTM informed by lexical, embedding-based, and BERT-based views for irony and satire detection across several Spanish language variants, reporting state-of-the-art results on multiple corpora and highlighting that satire remains substantially harder to identify than more overt forms of figurative language such as irony. This observation is consistent with the conservative behavior of our own Subtask 1 model (Section 8), and suggests that the subtle, context-dependent cues separating satirical from real headlines may require representations beyond what a generic sequence-pair classifier can capture from a few hundred training examples.

For our system, we rely on BETO [7], a BERT-base model pre-trained exclusively on Spanish text, which has been shown to outperform multilingual BERT models on a range of Spanish NLP benchmarks and has been widely adopted as a baseline encoder for Spanish text classification tasks, including prior editions of HAHA. Fine-tuning such pre-trained encoders on small, task-specific corpora is, however, known to be prone to overfitting and catastrophic forgetting: Sun et al. [9] systematically study fine-tuning recipes for BERT-based text classifiers and show that choices such as a low learning rate, a slanted triangular learning rate schedule, and limiting the number of training epochs are important for obtaining stable results on small downstream datasets. This motivates our adoption of an explicit regularization and stabilization pipeline (described in Section 4) when fine-tuning BETO on the relatively small Subtask 1 and Subtask 2 training sets.

Subtask 2 of HAHA 2026 relates to the broader and rapidly growing area of LLM-generated text detection. As large language models have become capable of producing fluent, human-like text across many genres, the need to distinguish human- from machine-authored content has motivated a substantial body of work on detection methods, ranging from fine-tuned classifiers to zero-shot statistical detectors [1]. In the social-media domain, Fagni et al. [10] introduce the TweepFake dataset of real deepfake tweets generated by several text-generation techniques and benchmark a range of fine-tuned detection methods, finding that transformer-based classifiers are among the most effective at separating human- from machine-written short texts. Our approach to Subtask 2 follows this fine-tuned-classifier paradigm, treating LLM-generated humor detection as a binary sequence-pair classification problem analogous to Subtask 1, which allows us to reuse the same model architecture and training pipeline across both subtasks.

3. Dataset

This section discusses the datasets for the corresponding subtasks.

3.1. Subtask 1: Humor Detection

Subtask 1 provides pairs of news headlines and their accompanying context, labeled as either *satirical* or *real*, drawn from news sources across six Spanish-speaking countries: Colombia, Mexico, Puerto Rico, Spain, Uruguay, and Venezuela. The labeled trial and development data, used for training, comprised **564** examples in total and were class-balanced (270 *real* and 270 *satirical* examples in the development set alone). The test set consisted of **600** unlabeled headline–context pairs, on which final predictions were submitted for evaluation.

3.2. Subtask 2: LLM-generated Humor Detection

Subtask 2 provides pairs of news headlines and jokes written in response to them, labeled as either *human*- or *machine*-written. The labeled trial and development data used for training comprised **362** examples, also class-balanced (175 *human* and 175 *machine* examples in the development set alone). The test set consisted of **550** unlabeled headline–joke pairs.

For both subtasks, the balanced nature of the training data ensured that any asymmetries observed in our model's predictions (Section 8) cannot be attributed to label imbalance in training. Table 1 summarizes the key statistics for both datasets.

Table 1
Dataset statistics for HAHA 2026 Subtasks 1 and 2.

	Subtask 1 *Humor Detection*	**Subtask 2** *LLM-gen. Humor Detection*
Input pair	headline + context	headline + joke
Label (0)	real	human
Label (1)	satirical	machine
Train + dev (total)	564	362
Dev set (class 0)	270	175
Dev set (class 1)	270	175
Dev set (total)	540	350
Test set (unlabeled)	600	550
Source countries	CO, MX, PR, ES, UY, VE	
Class balance	balanced (50/50)	

4. Experimental Setup and Methodology

For both subtasks, we fine-tuned BETO [7] as a binary sequence-pair classifier. Figure 1 gives an overview of the shared pipeline: both subtasks reuse the same encoder, pooling, and classification head, differing only in the input fields and label semantics, while a common set of regularization techniques is applied during training of both models.

Input representation. For Subtask 1, the input was constructed as `headline [SEP] context`, with labels *satirical* (1) and *real* (0). For Subtask 2, the input was constructed as `headline [SEP] joke`, with labels *machine* (1) and *human* (0). Aside from the input fields and label semantics, the same pipeline was applied to both subtasks.

Overfitting countermeasures. Given the small size of the training sets (564 examples for Subtask 1 and 362 examples for Subtask 2 relative to the number of parameters in BETO), we applied the following regularization and stabilization techniques [9]:

- Increased dropout (0.3) on the hidden, attention, and classifier-head layers.

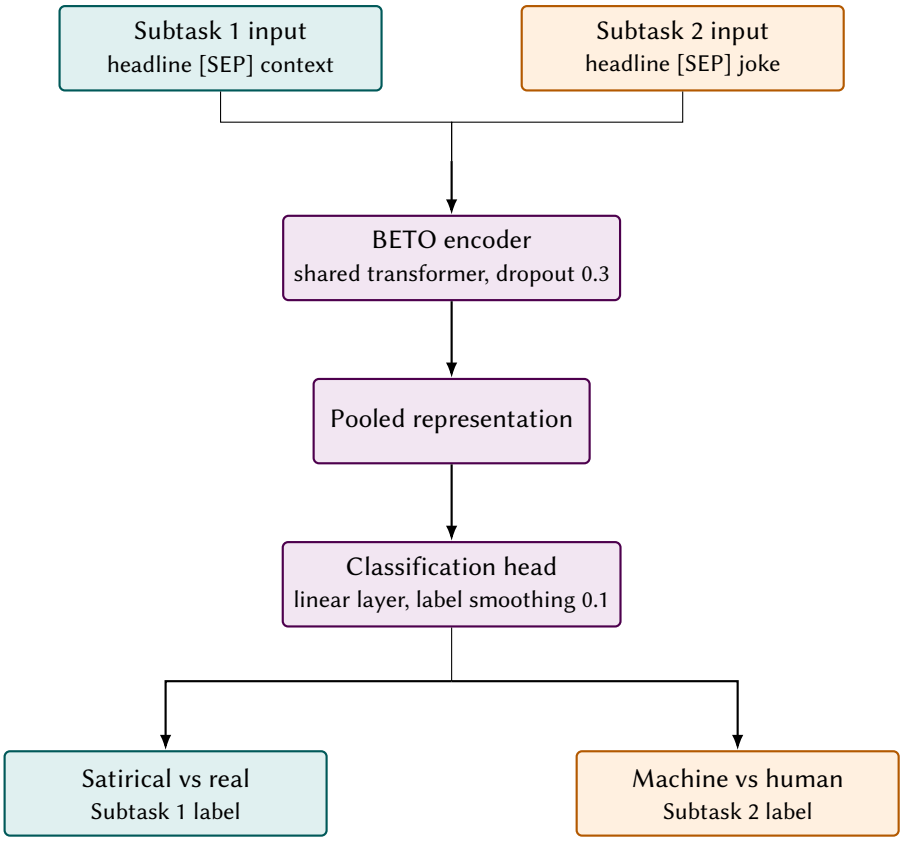

Figure 1: Shared fine-tuning pipeline for Subtasks 1 and 2. Both subtasks reuse the same BETO encoder, pooled representation, and classification head, differing only in the input fields (headline paired with context vs. joke) and label semantics. The regularization techniques are applied during training of both models.

- Label smoothing (0.1) in the cross-entropy loss.
- Early stopping based on validation F1 of the positive class (*satirical* for Subtask 1, *machine* for Subtask 2), with patience of 3 epochs.
- Weight decay of 0.01 via the AdamW optimizer.
- Gradient clipping at a maximum norm of 1.0.
- A linear warmup-and-decay learning rate schedule, with a peak learning rate of 2×10^{-5}.

Training procedure. Models were trained for up to 10 epochs with a batch size of 16. After identifying the best-performing epoch via early stopping on a held-out validation split, the model was retrained on the full combined training data (trial plus development sets) for the corresponding number of steps, and the resulting model was used to generate predictions on the official test sets.

5. Evaluation Metrics

Following the official HAHA 2026 evaluation protocol, Subtask 1 is scored using the F1 score of the *satirical* class, and Subtask 2 is scored using the F1 score of the *machine* (automatic) class. We additionally report precision, recall, and accuracy for both subtasks, as provided by the official leaderboard.

6. Results

Table 2 summarizes our official results for both subtasks. Our system obtained a substantially higher F1 score on Subtask 2 (0.8330) than on Subtask 1 (0.5650). Notably, precision was similar across both subtasks (0.8630 and 0.8605, respectively), while recall differed sharply (0.4200 versus 0.8073), driving the gap in F1 and accuracy. This pattern is examined further in Section 8.

Table 2
Official results for our system (team IReL_IIT(BHU), submission 759419) on HAHA 2026 Subtasks 1 and 2.

Subtask	Precision	Recall	F1	Accuracy
1 (Humor Detection)	0.8630	0.4200	0.5650	0.6767
2 (LLM-generated Humor Detection)	0.8605	0.8073	0.8330	0.8382

7. Competition Standings

On Subtask 1, our submission ranked **last** among the participating teams: our F1 score of 0.5650 fell below the organizer-provided baseline (F1 = 0.6252), and well below the top-performing system (F1 = 0.9933).

On Subtask 2, our submission ranked **6th out of 16 submissions** (approximately 5th among distinct participating teams), with an F1 score of 0.8330 against a top score of 0.8764. This result is also well above the organizer-provided baseline (F1 = 0.6597).

8. Error Analysis

The training data for both subtasks was class-balanced (270 *real* / 270 *satirical* for Subtask 1; 175 *human* / 175 *machine* for Subtask 2), which rules out class imbalance in training as the source of the performance gap between the two subtasks.

For Subtask 1, the precision/recall asymmetry (0.8630 / 0.4200) indicates that, despite balanced supervision, our model learned a conservative decision boundary skewed toward predicting *real*: when it predicted *satirical*, it was usually correct, but it missed the majority of actual satirical headlines. A plausible explanation is that BETO, fine-tuned on only 564 examples spanning six distinct national contexts, struggled to generalize the often-subtle stylistic markers of satire across these varied domains — particularly given that the HAHA 2026 organizers explicitly chose this domain because satirical and real headlines can be difficult to distinguish even for human readers. The combination of label smoothing (0.1) and elevated dropout (0.3), while intended to curb overfitting, may have further flattened the model's confidence near the decision boundary, reinforcing its bias toward the majority-confidence (*real*) prediction at test time.

In contrast, Subtask 2's substantially higher recall (0.8073) alongside comparable precision (0.8605) suggests that distinguishing LLM-generated from human-written jokes relies on more learnable surface-level cues — for example, LLM-generated jokes may exhibit more templated structure or lack the idiomatic, culturally-grounded wordplay typical of human-written Spanish humor. The same fine-tuning pipeline was able to capture these cues effectively even with a smaller training set (362 examples), suggesting that the bottleneck for Subtask 1 lies in the inherent difficulty of the satire-detection task itself, rather than in the training pipeline or data quantity alone.

Overall, this analysis suggests that future iterations of our system on tasks similar to Subtask 1 may benefit from task-specific adaptations — such as data augmentation, ensembling with feature-based models, or country-specific fine-tuning — rather than relying on a generic fine-tuning recipe shared across subtasks.

9. Conclusion

We presented the participation of team IReL_IIT(BHU) in Subtasks 1 and 2 of HAHA 2026, both addressed by fine-tuning BETO with a shared overfitting-mitigation pipeline combining increased dropout, label smoothing, early stopping, weight decay, gradient clipping, and a linear warmup-and-decay learning rate schedule. While this approach achieved competitive results on Subtask 2 (F1 = 0.8330, ranking 6th of 16 submissions), it underperformed substantially on Subtask 1 (F1 = 0.5650, last place), with a pronounced precision/recall asymmetry indicating a conservative bias toward the majority *real* class. Our error analysis suggests that this gap stems primarily from the inherent difficulty of satire detection across multiple national contexts with limited training data, rather than from class imbalance or a flawed training procedure. Future work will explore task-specific architectural adaptations and data augmentation strategies to address the Subtask 1 performance gap.

Declaration on Generative AI

During the preparation of this work, the authors used AI-assisted tools to support code debugging, manuscript drafting, and submission formatting. All content was reviewed and edited by the authors, who take full responsibility for the publication.

References

[1] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions, Computational Linguistics 51 (2025) 275–338.

[2] S. Castro, L. Chiruzzo, A. Rosá, D. Garat, G. Moncecchi, Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval 2018, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), volume 2150 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 187–194.

[3] L. Chiruzzo, S. Castro, A. Rosá, Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR Workshop Proceedings, CEUR-WS.org, 2019, pp. 132–144.

[4] L. Chiruzzo, S. Castro, A. Rosá, J. A. Meaney, R. Mihalcea, Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, CEUR-WS.org, 2021.

[5] 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).

[6] 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.

[7] 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.

[8] R. Ortega-Bueno, P. Rosso, J. E. Medina Pagola, Multi-view Informed Attention-based Model for Irony and Satire Detection in Spanish Variants, Knowledge-Based Systems 235 (2022) 107597. doi:10.1016/j.knosys.2021.107597.

[9] C. Sun, X. Qiu, Y. Xu, X. Huang, How to Fine-Tune BERT for Text Classification?, in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, volume 11856 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 194–206. doi:10.1007/978-3-030-32381-3_16.

[10] T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi, TweepFake: About Detecting Deepfake Tweets, PLOS ONE 16 (2021) e0251415. doi:10.1371/journal.pone.0251415.