Prompting and Fine-Tuning Approaches for Detection and Generation of Spanish Humor

Osamah H. Alaini[1,*,†], Nasser Thmer[2,†] and Ali Al-Laith[3,†]

[1]Department of Computer Science, Trent University, Peterborough, Ontario, Canada
[2]Department of Computer, College of Education Lawder, University of Abyan, Yemen
[3]Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark

Abstract

We describe our participation in all three subtasks of HAHA@IberLEF 2026, a shared task on humor detection and generation in Spanish. For the two classification subtasks — Humor Detection and LLM-Generated Humor Detection, we experiment with two complementary paradigms: ensemble prompting of locally served open-weight LLMs (Llama 3.2, Mistral 7B, Gemma 3) and fine-tuning of an ensemble of Spanish and multilingual pre-trained language models (RoBERTuito, BETO, XLM-RoBERTa). For Humor Generation we apply a best-of-n prompting strategy with LLM-based reranking. Our few-shot prompting ensemble ranks **5th** on Subtask 1 (F1 = 0.8567), and our Pre-trained Language Models PLMs ensemble ranks **2nd** on Subtask 2 (F1 = 0.8729) — only 0.0035 points behind the top system. On Subtask 3, our generation ensemble ranks **3rd** (Elo = 919), above the official baseline. Results show that the PLMs ensemble generalizes more consistently across both classification subtasks, while few-shot prompting is particularly effective at satirical headline detection but less reliable at detecting LLM-generated humor.

Keywords

computational humor, humor detection, LLM-generated text detection, ensemble methods, Spanish NLP, HAHA 2026

1. Introduction

Humor is one of the most nuanced phenomena in human language, relying on incongruity, irony, wordplay, and shared cultural knowledge — signals that standard NLP models are not well equipped to capture. Despite these challenges, computational humor has attracted growing interest, with applications in dialogue systems, misinformation detection, and creative text generation.

The 2026 edition of HAHA [1], held as part of IberLEF 2026 [2], proposes three subtasks over Spanish text: (1) **Humor Detection**, classifying news headlines as satirical or real; (2) **LLM-Generated Humor Detection**, determining whether a joke was written by a human or generated by an LLM; and (3) **Humor Generation**, producing jokes from news headlines, evaluated via Elo-based human preference judgments.

We participate in all three subtasks, exploring prompting and fine-tuning paradigms, both centered on ensemble strategies. For the classification subtasks we compare zero-shot and few-shot prompt ensembles across three LLMs with a fine-tuned ensemble of RoBERTuito [3], BETO [4], and XLM-RoBERTa [5]. For generation, we use a best-of-n prompting strategy with LLM reranking.

2. Related Work

Computational humor detection has been studied since early work by Mihalcea and Strapparava [6], who showed that classification-based methods could reliably distinguish humorous from non-humorous text. For Spanish, Castro et al. [7] built one of the first supervised humor classifiers from Spanish tweets,

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
osamahalaini@trentu.ca (O. H. Alaini); nasserthmer.net@gmail.com (N. Thmer); alal@di.ku.dk (A. Al-Laith)
0000-0001-6650-3469 (A. Al-Laith)

and later contributed a crowd-annotated corpus [8] that became the foundation for the HAHA shared task series at IberEval 2018, IberLEF 2019, and IberLEF 2021 [9, 10, 11]. Related multilingual efforts include SemEval-2021 Task 7 [12] and the HUHU task at IberLEF 2023. HAHA 2026 [1] extends this line to satirical news headlines, a domain where the satirical–real boundary is especially challenging due to the deliberate emulation of journalistic prose [13].

Distinguishing LLM-generated from human-written text has become an active research area as model quality has improved [14, 15]. Detection approaches span watermarking, zero-shot statistical methods, and supervised neural classifiers [16]. Within creative genres, even human evaluators struggle with this task [17], motivating the second HAHA subtask. For generation, humor remains harder than detection [18], but interest has grown with initiatives such as SemEval-2026 Humor Generation and MWAHAHA [19]. Evaluation via arena-style pairwise preference judgments and Elo ratings [20, 21] provides a reliable proxy for subjective humor quality, as adopted by HAHA 2026 Subtask 3.

3. Methodology

Our approach covers two paradigms for Subtasks 1 and 2: *prompting* via locally served LLMs and *supervised fine-tuning* of Spanish pre-trained language models (PLMs) — here, encoder-based models trained with a masked-language-modeling objective, as distinct from the decoder-only LLMs used for prompting. Both incorporate ensemble strategies as a central design choice. Subtask 3 relies solely on prompting with a best-of-n selection strategy.

3.1. Prompting-Based Systems

We experiment with three open-weight LLMs served locally via Ollama[1]: **Llama 3.2** (3B), **Mistral 7B**, and **Gemma 3** (4B). Temperature is fixed at $T{=}0.0$ for classification and $T{=}0.7$ for generation.

For each classification subtask we design three prompt variants: a **zero-shot** prompt requesting a single-word label, a **few-shot** prompt adding two labeled examples with satirical framing cues, and a **chain-of-thought** prompt that guides the model through plausibility and irony checks before producing structured JSON output. Inputs prepend a country token (`[ES]`, `[MX]`, `[AR]`) and, for Subtask 1, append the full article context field. For our **few-shot** variant, two labeled examples are prepended to each prompt from the training set.

Predictions from all three variants are combined by **majority voting**, ties broken in favor of the positive class (*satirical* / *machine*). Extending the vote pool to all three models yields a **Prompt×Model ensemble** of nine voters per instance, which constitutes our primary prompting submission. For Subtask 3, each model generates one joke per headline; Llama 3.2 then reranks the three candidates by surprise, cultural relevance, and punchline clarity. Full prompt templates are provided in Appendix A.

3.2. Fine-Tuning Systems

We fine-tune three PLMs with complementary pre-training registers. **RoBERTuito**[2] was trained on 500M Spanish tweets, making it well suited to irony-rich text. Inputs are preprocessed with the `pysentimiento` normalization pipeline [3]. **BETO**[3] provides strong coverage of formal written Spanish via whole-word masked pre-training [4]. **XLM-RoBERTa**[4] adds cross-lingual coverage from 100 languages [5].

Training on small labeled sets (540 examples for Subtask 1; 350 for Subtask 2) requires careful regularization. We apply **label smoothing** ($\varepsilon{=}0.1$), **layer freezing** for the first three epochs, and AdamW (lr$=1{\times}10^{-5}$, weight decay $= 0.01$) with 20 % linear warmup for 15 epochs. Performance is estimated under stratified 5-fold cross-validation.

[1] https://ollama.com
[2] https://huggingface.co/pysentimiento/robertuito-base-uncased
[3] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[4] https://huggingface.co/FacebookAI/xlm-roberta-base

The ensemble averages logits element-wise before applying softmax:

$$P(y \mid \mathbf{x}) = \text{softmax}\left(\tfrac{1}{3} \sum_{i=1}^{3} \mathbf{z}_i(\mathbf{x}) \right), \tag{1}$$

where $\mathbf{z}_i(\mathbf{x})$ are the unnormalized logits of model i. We also tune the **decision threshold** τ^* by sweeping $[0.2, 0.8]$ on pooled out-of-fold CV logits to maximize target-class F1, fully contained within cross-validation.

All hyperparameters were selected manually, guided by established best practices for fine-tuning pre-trained transformers on small labeled datasets [22, 23], rather than by automated search, due to the very limited size of the training sets.

4. Results

We submitted three systems per subtask: (S1) zero-shot prompting ensemble, (S2) few-shot prompting ensemble, and (S3) fine-tuned PLMs ensemble. Tables 1–3 show the official leaderboard results; our submissions are highlighted in all results tables.

4.1. Subtask 1: Humor Detection

The **few-shot ensemble** (S2) achieves our best Subtask 1 result (F1 = 0.8567, **5th** out of 13), well above the baseline (F1 = 0.6252). The **PLMs ensemble** (S3) is a close second among our submissions (F1 = 0.8414, **7th** out of 13), indicating that the fine-tuned models transfer reasonably well to the satirical headline domain. The **zero-shot ensemble** (S1) again maximizes recall (0.9500) at the cost of precision (F1 = 0.8237, **9th** out of 13).

Table 1
Subtask 1 results. Primary metric: F1 on *satirical*. Our rows shaded.

Rank	System	Prec.	Rec.	F1	Acc.
1	michaelibrahim	0.9933	0.9933	0.9933	0.9933
2	ymlopez	0.9690	0.9367	0.9525	0.9533
3	namdang_2006	0.9681	0.9100	0.9381	0.9400
4	ymlopez	0.9813	0.8767	0.9261	0.9300
5	**S2: few-shot**	0.8201	0.8967	**0.8567**	0.8500
6	ymlopez	0.7919	0.9133	0.8483	0.8367
7	**S3: PLMs ens.**	0.9042	0.7867	**0.8414**	0.8517
8	namdang_2006	0.8322	0.8267	0.8294	0.8300
9	**S1: zero-shot**	0.7270	0.9500	**0.8237**	0.7967
10	m_buch_135	0.7872	0.8633	0.8235	0.8150
11	Baseline	0.6923	0.5700	0.6252	0.6583
12	edo2002	0.7225	0.5033	0.5933	0.6550
13	arjun108	0.8630	0.4200	0.5650	0.6767

4.2. Subtask 2: LLM-Generated Humor Detection

Subtask 2 is our strongest task. The **PLMs ensemble** (S3) achieves our best result here, ranking **2nd** out of 16 (F1 = 0.8729) — only 0.0035 points behind the top system — with balanced precision (0.8274) and recall (0.9236). The **few-shot ensemble** (S2) ranks **13th** (F1 = 0.6535), over-predicting the positive class (recall = 0.9636) at the cost of precision (0.4944). The **zero-shot ensemble** (S1) performs worst among our submissions and falls below the official baseline, ranking **16th** out of 16 (F1 = 0.5095; precision = 0.4248, recall = 0.6364).

Table 2
Subtask 2 results. Primary metric: F1 on *machine*. Our rows shaded.

Rank	System	Prec.	Rec.	F1	Acc.
1	michaelibrahim	0.8764	0.8764	0.8764	0.8764
2	**S3: PLMs ens.**	0.8274	0.9236	**0.8729**	0.8655
3	michaelibrahim	0.8727	0.8727	0.8727	0.8727
4	m_buch_135	0.8577	0.8327	0.8450	0.8473
5	namdang_2006	0.8339	0.8400	0.8370	0.8364
6	arjun108	0.8605	0.8073	0.8330	0.8382
7	m_buch_135	0.8913	0.7455	0.8119	0.8273
8	michaelibrahim	0.7600	0.7600	0.7600	0.7600
9	ymlopez	0.7472	0.7200	0.7333	0.7382
10	namdang_2006	0.8840	0.5818	0.7018	0.7527
11	Baseline	0.5165	0.9127	0.6597	0.5291
12	ymlopez	0.6593	0.6545	0.6569	0.6582
13	**S2: few-shot**	0.4944	0.9636	**0.6535**	0.4891
14	namdang_2006	0.8993	0.4545	0.6039	0.7018
15	ymlopez	0.6453	0.5491	0.5933	0.6236
16	**S1: zero-shot**	0.4248	0.6364	**0.5095**	0.3873

4.3. Subtask 3: Humor Generation

The S2 and S3 submissions both used the generation ensemble, yielding a single leaderboard entry (Elo = 919, **3rd out of 4**). The non-overlapping confidence interval with the baseline (Elo = 835) confirms that the best-of-n reranking strategy provides a statistically reliable improvement, though a substantial gap remains to the top two systems.

Table 3
Subtask 3 results. Systems ranked by Elo. Our submission shaded.

Rank	System	Elo	95 % CI	Votes
1	ymlopez	1136	[1108, 1163]	344
1	michaelibrahim	1110	[1083, 1138]	344
3	**S2 & S3: gen. ens.**	919	[893, 944]	345
4	Baseline	835	[810, 873]	347

4.4. Discussion

Three patterns emerge across the subtasks. First, the **PLMs ensemble** (S3) is the most consistent of our approaches, ranking 7th out of 13 on Subtask 1 (F1 = 0.8414) and 2nd out of 16 on Subtask 2 (F1 = 0.8729, only 0.0035 points behind the top system). This suggests that the features learned during fine-tuning transfer well across both satirical-headline and LLM-generated-text detection.

Second, **few-shot prompting** (S2) is more task-dependent: it performs well on Subtask 1 (5th place, F1 = 0.8567) but falls to 13th place on Subtask 2 (F1 = 0.6535), where it over-predicts the positive class (recall = 0.9636 at the cost of precision = 0.4944). This indicates that the in-context examples used for satirical-headline detection transfer less effectively to distinguishing human- from LLM-generated humor.

Third, **zero-shot prompting** (S1) is consistently our weakest approach, ranking 9th out of 13 on Subtask 1 (F1 = 0.8237) and last out of 16 on Subtask 2 (F1 = 0.5095), below the official baseline. This highlights the benefit of in-context calibration examples for instruction-tuned models on both tasks.

5. Conclusion

We presented three systems, combining prompting and fine-tuning ensemble strategies for Spanish humor detection and generation. Our key findings are: (i) the PLMs ensemble is the most robust overall approach, achieving competitive results on both classification subtasks (7th on Subtask 1, 2nd on Subtask 2); (ii) few-shot prompting is highly effective for satirical headline detection (5th place) but is less reliable for detecting LLM-generated humor, where it over-predicts the positive class; and (iii) zero-shot prompting is consistently our weakest approach across both classification subtasks, underperforming the official baseline on Subtask 2. For humor generation, best-of-n reranking clearly outperforms the baseline but still leaves a considerable gap to the top systems. Future work will explore hybrid prompting–fine-tuning approaches and the integration of humor-theoretic knowledge into both classification and generation pipelines.

Acknowledgments

The authors thank the HAHA@IberLEF 2026 organizers for providing the datasets and evaluation infrastructure.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with writing and editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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).

[2] 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.

[3] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[4] 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.

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[6] R. Mihalcea, C. Strapparava, Making computers laugh: Investigations in automatic humor recognition, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 531–538. URL: https://aclanthology.org/H05-1067/.

[7] S. Castro, M. Cubero, D. Garat, G. Moncecchi, Is this a joke? detecting humor in spanish tweets, in: Ibero-American Conference on Artificial Intelligence, Springer, 2016, pp. 139–150. doi:10.1007/978-3-319-47955-2_12.

[8] S. Castro, L. Chiruzzo, A. Rosá, D. Garat, G. Moncecchi, A crowd-annotated Spanish corpus for humor analysis, in: Proceedings of the Sixth International Workshop on Natural Language

Processing for Social Media, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 7–11. URL: https://aclanthology.org/W18-3502/. doi:10.18653/v1/W18-3502.

[9] S. Castro, L. Chiruzzo, A. Rosá, Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018, in: CEUR Workshop Proceedings, volume 2150, 2018, pp. 187–194.

[10] 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 the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR-WS, Bilbao, Spain, 2019, pp. 132–144.

[11] 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. doi:10.26342/2021-67-22.

[12] J. A. Meaney, S. R. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, SemEval-2021 task 7: HaHackathon, detecting and rating humor and offense, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Bangkok, Thailand (online), 2021, pp. 105–119.

[13] N. Hossain, J. Krumm, M. Gamon, H. Kautz, SemEval-2020 task 7: Assessing humor in edited news headlines, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 746–758. URL: https://aclanthology.org/2020.semeval-1.98/. doi:10.18653/v1/2020.semeval-1.98.

[14] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, D. F. Wong, A survey on LLM-generated text detection: Necessity, methods, and future directions, Computational Linguistics 51 (2025) 275–338. doi:10.1162/coli_a_00549.

[15] X. Yang, L. Pan, X. Zhao, H. Chen, L. Petzold, W. Y. Wang, W. Cheng, A survey on detection of llms-generated content, arXiv preprint arXiv:2310.15654 (2023).

[16] M. Abassy, K. Elozeiri, A. Aziz, M. N. Ta, R. V. Tomar, B. Adhikari, S. E. D. Ahmed, Y. Wang, O. Mohammed Afzal, Z. Xie, J. Mansurov, E. Artemova, V. Mikhailov, R. Xing, J. Geng, H. Iqbal, Z. M. Mujahid, T. Mahmoud, A. Tsvigun, A. F. Aji, A. Shelmanov, N. Habash, I. Gurevych, P. Nakov, LLM-DetectAIve: a tool for fine-grained machine-generated text detection, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 336–343.

[17] T. Buz, B. Frost, N. Genchev, M. Schneider, L.-A. Kaffée, G. de Melo, Investigating wit, creativity, and detectability of large language models in domain-specific writing style adaptation of reddit's showerthoughts, in: Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024), 2024.

[18] J. Sjöbergh, K. Araki, Recognizing humor without recognizing meaning, in: MICAI 2007: Advances in Artificial Intelligence, volume 4578 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 469–476. doi:10.1007/978-3-540-73400-0_59.

[19] A. Tikhonov, A. Ivanov, lmfaoooo at SemEval-2026 task 1: Humor is an audience. preference modeling for constrained humor generation, in: Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026), 2026.

[20] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, I. Stoica, Chatbot arena: An open platform for evaluating LLMs by human preference, in: Proceedings of the 41st International Conference on Machine Learning, volume 235, PMLR, 2024, pp. 8359–8388.

[21] E. Ajayi, P. Mitra, Humorrank: A tournament-based leaderboard for evaluating humor generation in large language models, arXiv preprint arXiv:2604.19786 (2026).

[22] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines, arXiv preprint arXiv:2006.04884 (2020).

[23] Z. Zhang, X. Wang, Z. Zhang, P. Cui, W. Zhu, Revisiting transformation invariant geometric deep learning: Are initial representations all you need?, arXiv preprint arXiv:2112.12345 (2021).

A. Prompt Templates

This section presents all prompt templates used in zero-shot and few-shot settings for reproducibility. All models are instructed to output only the predicted label without any explanation or additional text.

A.1. Humor Detection

A.1.1. Zero-shot Prompt

```
You are an expert humor and satire detection system.
Your task is to classify the following news headline as:
- satirical
- real
Use both the headline and the context.
Return ONLY a JSON object in the following format:
{
"tag": "satirical" or "real",
"confidence": number_between_0_and_1
}
Do not explain anything.
Headline: {headline}
Context: {context}
```

A.1.2. Few-shot Prompt

```
You are an expert humor and satire detection system.
Your task is to classify the following news headline as:
- satirical
- real
Use both the headline and the context.
Return ONLY a JSON object in the following format:
{
"tag": "satirical" or "real",
"confidence": number_between_0_and_1
}
Do not explain anything.
Headline: {headline}
Context: {context}
Examples:
{FEWSHOT_EXAMPLES}
```

A.2. LLMs-Generated Humor Detection

A.2.1. Zero-shot Prompt

You are an expert detector of machine-generated humor.
Your task is to determine whether the following joke was:
- human
- machine
The joke was inspired by a news headline.
Analyze writing style, creativity, coherence, punchline structure, and signs of LLM-generated text.
Return ONLY a JSON object in the following format:
{
"tag": "human" or "machine",
"confidence": number_between_0_and_1
}
Do not explain anything.
Headline: {headline}
Joke: {joke}

A.2.2. Few-shot Prompt

You are an expert detector of machine-generated humor.
Your task is to determine whether the following joke was:
- human
- machine
The joke was inspired by a news headline.
Analyze writing style, creativity, coherence, punchline structure, and signs of LLM-generated text.
Return ONLY a JSON object in the following format:
{
"tag": "human" or "machine",
"confidence": number_between_0_and_1
}
Do not explain anything.
Headline: {headline}
Joke: {joke}
Examples:
{FEWSHOT_EXAMPLES}

A.3. Humor Generation

A.3.1. Zero-shot Prompt

You are a professional satirical comedy writer.
Generate ONE funny joke in Spanish inspired by the following news headline.
Requirements:
- The joke MUST be written entirely in Spanish
- Do not use English
- Be genuinely funny and creative
- Use irony, absurdity, exaggeration, or wordplay
- Keep it concise (1-3 sentences)
- Avoid explanations
- Avoid offensive content
- Sound natural and human-written
- Do not repeat the headline verbatim
Headline: {headline}
Return ONLY the joke text.