ELiRF-UPV at TA1C-IberLEF 2025: Clickbait Detection in Spanish

Alberto Picazo Pardo[1,†], Vicent Ahuir[2,†] and María José Castro-Bleda[2,3,*,†]

[1] *Department of Computer Systems and Computation, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

[2] *VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

[3] *ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

Abstract

This paper presents our participation in the TA1C: Clickbait Detection and Spoiling in Spanish task at IberLEF 2025, which focuses on the automatic identification and mitigation of clickbait in Spanish-language online news shared on social media. The task comprises two subtasks: (1) a classification task, where the system must determine whether a given tweet constitutes clickbait, and (2) a generative task, in which the system—given a teaser (tweet and headline) along with the corresponding news article—must generate a concise spoiler that resolves the information gap and satisfies the curiosity induced by the teaser. The dataset includes Spanish tweets, headlines, and full news articles, annotated to support both classification and generation objectives. Our team participated in the first subtask, addressing the challenge of clickbait detection by developing several systems based on pre-trained transformer-based language models, applying fine-tuning strategies to improve prediction quality. The results confirm the effectiveness of our approach.

Keywords

Natural Language Processing, Transformers-based Models, Large Language Models, Clickbait Detection

1. Introduction

Clickbait is a prevalent phenomenon in online news media. It refers to the use of sensational, ambiguous, or deliberately incomplete headlines and teasers intended to provoke curiosity and drive user engagement, often at the expense of informativeness. Although its primary function is to increase traffic, clickbait often contributes to the spread of low-quality or misleading information, eroding trust in digital journalism.

The growing use of clickbait is not limited to soft news or dubious sources; increasingly, even reputable media outlets employ clickbait strategies for high-impact stories. The need for automatic tools to detect and mitigate the influence of clickbait content has become critical, not only from a technological standpoint but also from an ethical and communicative perspective.

Although early efforts in clickbait detection focused primarily on English, recent research has begun to address this issue in other languages, including Spanish. However, Spanish remains underexplored in terms of large-scale, annotated datasets, and shared evaluation frameworks.

This work presents our contribution to the shared task TA1C ("Te Ahorré Un Click") at IberLEF 2025 [1], the first shared evaluation campaign specifically focused on clickbait detection and spoiling in Spanish [2]. The goal is twofold: (1) to classify whether a teaser (e.g., tweet + title) constitutes clickbait, and (2) to generate a spoiler that fills the curiosity gap, providing readers with the missing key information. These tasks are relevant not only for the Natural Language Processing (NLP) community but also for researchers in digital communication and journalism studies.

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

apicpar@inf.upv.es (A. Picazo Pardo); vahuir@dsic.upv.es (V. Ahuir); mcastro@dsic.upv.es (M. J. Castro-Bleda)
0000-0001-5636-651X (V. Ahuir); 0000-0003-1001-8258 (M. J. Castro-Bleda)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-10

Table 1
TA1C "Clickbait Detection" dataset sample distribution along the sets. It is also shown the percentage of Clickbait and Non-clickbait samples on Train and Dev sets.

	Total		**Clickbait**		**Non-clickbait**	
Set	**Number**	**(% Total)**	**Number**	**(% Set)**	**Number**	**(% Set)**
Train	2800	(66.66%)	798	(28.50%)	2002	(71.50%)
Dev	700	(16.66%)	203	(29.00%)	497	(71.00%)
Test	700	(16.66%)	-	-	-	-
Total	4200	(100%)				

The remainder of the paper is structured as follows. In Sections 2 and 3, we introduce the task and describe the dataset and evaluation metrics. Section 4 presents our clickbait classification system along with our experimental results. Finally, Section 6 concludes the paper and discusses future work.

2. Task Description: Clickbait Detection

The TA1C shared task consists of two subtasks that address the identification and mitigation of clickbait in Spanish-language online news shared on social networks. The dataset includes tweets, news headlines and full news articles, annotated to support both classification and generation objectives. We have experimented with the first subtask, clickbait detection.

The task is a binary classification problem: Given a teaser consisting of a tweet and the corresponding headline of the news article, the objective is to determine whether the content qualifies as clickbait. The annotation relies on the definition proposed by Mordecki et al. [3], which is grounded in Loewenstein's information gap theory [4]: clickbait deliberately omits crucial information to provoke curiosity and entice users to click. Systems must predict a binary label (clickbait / non-clickbait) for each instance. Performance will be evaluated using standard classification metrics: F1-score (primary metric), Accuracy, Precision and Recall.

3. The Dataset

The dataset for both tasks consists of Spanish-language tweets that link to online news articles. Data were curated for the TA1C shared task [3], with the aim of capturing linguistic diversity across the Spanish-speaking world.

The dataset for Subtask 1 includes 4200 tweets collected between October 2020 and October 2021 from 18 well-known media outlets in 12 Spanish-speaking countries. Each tweet is accompanied by the URL and clean HTML of the linked article, the headline, subheadline, article body (cleaned), images and captions, and embedded external links.

Each tweet was independently labeled by three human annotators to determine if it is clickbait. The 4200 samples of the dataset are split as shown in Table 1. We can appreciate an imbalance towards Non-clickbait labeled samples in the training and development set. Regarding the test set, we could not provide the distribution of Clickbait and Non-clickbait samples for that partition since the participants could not access the test labels. Examples of clickbait and non-clickbait samples are provided in Table 2.

4. Systems for Clickbait Detection

We developed six systems for the Clickbait Detection task, each exploring different pre-training and data augmentation strategies based on the Spanish-language RoBERTa model pre-trained with data from the National Library of Spain (`roberta-base-bne`) [5] publicly available in the HuggingFace hub [6]. The objective was to evaluate the impact of these techniques on classification performance while ensuring comparability across systems.

Table 2
Some examples of the TA1C "Clickbait Detection" dataset.

Sample	Label
La poeta estadounidense Louise Glück gana el Nobel de Literatura 2020	non-clickbait
Con banderas y mascarillas, argentinos salen a las calles a manifestarse contra el Gobierno	non-clickbait
Advierten una falla de seguridad en Google Home y Chromecast: de qué se trata y cómo solucionarlo	clickbait
Un modisto, un abogado y un asesino convicto, los embajadores de la mafia calabresa en la Argentina	clickbait

All systems were trained with the same hyperparameters: 5 epochs, a learning rate of 3×10^{-5}, batch size of 16, 50 warmup steps, and a weight decay of 0.01. The TA1C training split (2800 examples) was used as the training data, and the development set (700 examples) was used for validation, unless otherwise stated. The following configurations were explored.

System T1-1: Baseline Fine-tuning

This system serves as the reference point for evaluating the effectiveness of subsequent strategies. The `roberta-base-bne` model was fine-tuned directly on the TA1C training set. The development set was used to monitor validation performance and select the best checkpoint.

System T1-2: Two-phase Pre-training with Translated Tweets

To enrich the model with more diverse examples of clickbait-like language, we carried out an intermediate pre-training phase using 20,000 English tweets that were machine-translated into Spanish. The 20,000 tweets were randomly sampled from the https://huggingface.co/datasets/christinacdl/clickbait_detection_dataset dataset. Translation was performed using the Helsinki-NLP translation English-Spanish model (https://huggingface.co/Helsinki-NLP/opus-mt-en-es) [7]. This pre-training phase is carried out using the same hyperparameters used for fine-tuning. This phase was followed by fine-tuning on the TA1C training set. The development set was used in both phases for early stopping and evaluation.

System T1-3: Backtranslation-based Augmentation

A data augmentation method was applied using multilingual backtranslation [8, 9]. We generated 1189 synthetic samples from the TA1C training set using a multi-step backtranslation pipeline: *Spanish* → *German* → *French* → *Polish* → *Spanish*, using the Helsinki-NLP translation models[7]. We retained only those samples with a word-level distance between 5 and 35 words relative to their original counterparts, to balance diversity and fidelity. Erroneous or low-quality translations were removed manually. These synthetic samples were then concatenated with the original training data to form an augmented dataset of 3989 examples.

System T1-4: Random Masking Augmentation

In this system, we applied a random masking strategy to simulate Masked Language Model (MLM) noise. We generated five augmented versions of the TA1C training set by randomly masking 15% of the tokens in each sentence. After concatenation and shuffling, the final dataset contained approximately 14,000 augmented samples. This noisy training data was used to fine-tune the `roberta-base-bne` model, again validated on the TA1C development set.

Table 3
Task 1: F1 scores on TA1C development and test sets for each system. The super index marks our best rank in the contest.

System	Description	F1 Dev	F1 Test
T1-1	Baseline Fine-tuning	0.87684	0.79019
T1-2	Pre-training with Translated Tweets	0.86375	0.77976
T1-3	Backtranslation Augmentation	0.87878	**0.81481**^{3}
T1-4	Random Masking	0.85714	0.78651
T1-5	NER Masking	0.84514	0.76471
T1-6	All Data (Train + Dev)	–	0.80342

System T1-5: NER Masking

As an additional strategy to enhance the model's generalization capabilities, Named Entity Recognition (NER) masking was applied using spaCy [10] with the `es_core_news_lg` model. This approach consists of replacing detected named entities (such as organizations, products, or locations) with their corresponding generic labels (e.g., <ORG>, <PRODUCT>, <LOC>), aiming to reduce the model's reliance on specific entity names and encourage a greater focus on linguistic context. Fine-tuning was then done on the masked texts.

System T1-6: Fine-tuning Using All Available Data

This system trained on both the training and validation sets (3500 tweets in total) without a development set for evaluation. The `roberta-base-bne` model was fine-tuned directly on these 3500 tweets. No early stopping or checkpoint selection was applied; the final model corresponds to the last training epoch. All other training parameters were the same.

5. Experimental Results and Discussion

This section summarizes the performance of the systems on both the TA1C development set and the official blind test set provided for the shared task. All systems were evaluated using the F1-score as the primary metric.

Table 3 presents the F1 results. System T1-3, which incorporated backtranslation-based augmentation, achieved the highest F1 score on the test set (0.81481), ranking **third place overall** in the shared task leaderboard. Notably, although System T1-1 (baseline fine-tuning) already achieved a strong performance, augmenting the training data with synthetic samples from multilingual backtranslation in T1-3 offered a meaningful improvement of more than two points in the test F1-score.

System T1-2 (two-phase pre-training) and T1-4 (random masking) showed moderate performance compared to the baseline in the development set, and did not surpass T1-3 on the test set. These findings suggest that while pre-training on loosely aligned translated data or introducing random masking noise may help generalization, carefully filtered backtranslated samples better preserve task-relevant semantics and are more effective for low-resource domain adaptation in Spanish clickbait detection.

System T1-5 (NER masking) showed solid generalization performance, achieving an F1-score of 0.8451 on the validation set and 0.7647 on the test set. Despite a drop from development to test, the results suggest that abstracting entities can still retain useful semantic patterns and may help reduce overfitting.

Finally, System T1-6, trained on all available data, achieved a strong test F1 (0.80342), showing the benefit of leveraging the full dataset at the expense of validation-based checkpoint selection.

Fig. 1 visually summarizes the results from Table 3, with systems ordered by Test performance. System T1-3 (Backtranslation Augmentation) achieved the highest F1 on the test set (0.81481), confirming the effectiveness of multilingual backtranslation in generating useful training samples. System T1-6, which leveraged all available data (training + validation), also performed well on the test set

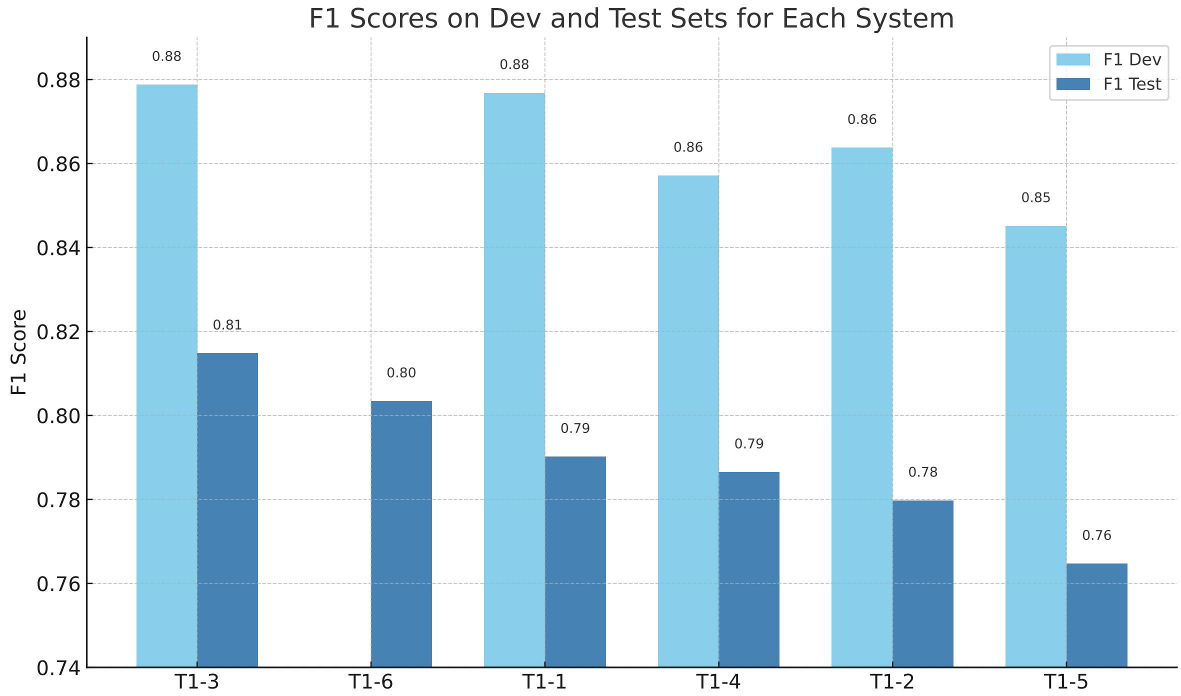

Figure 1: Evolution of the F1-Score of our systems, sorted by Test score (best to worst), for Dev and Test for Task 1.

(0.80342). The baseline system T1-1 already showed strong performance, suggesting that the pre-trained `roberta-base-bne` model is well-suited for the task even without additional augmentation. Systems T1-4 (Random Masking) and T1-2 (Two-phase Pre-training) yielded slightly lower test scores, indicating limited benefits from these strategies in this context. Finally, System T1-5 (NER Masking) had the lowest test F1-score (0.76471), but still demonstrated respectable generalization, highlighting its potential to reduce overfitting by abstracting entity-specific information. Overall, the figure underscores that not all augmentation techniques are equally effective, with carefully designed backtranslation yielding the most significant gains.

6. Conclusions and Future Work

This study explored several approaches for clickbait detection in the TA1C competition in Spanish [2]. We proposed a transformer-based strategy centered on fine-tuning a pre-trained RoBERTa model adapted to Spanish, using the competition's dataset, external clickbait-related corpora, and multiple data augmentation techniques.

Among the techniques evaluated, backtranslation proved particularly effective, significantly enhancing system robustness and contributing to the highest test F1-score (0.81481) across all our models, achieving third place in the competition. While other strategies, such as random masking or NER-based masking, showed more modest gains, they still offer potential for reducing overfitting and warrant further investigation.

Future work will focus on experimenting with larger language models and more sophisticated data augmentation strategies, aiming to further improve performance and generalization in clickbait detection tasks.

7. Ethics Statement

We have not used additional data to those provided by the competition. The pre-trained models used are obtained from HuggingFace models hub, under the Apache License 2.0.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-126061OB-C41. It is also partially supported by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[2] G. Mordecki, L. Chiruzzo, R. Laguna, J. Prada, A. Rosá, I. Sastre, G. Moncecchi, Overview of TA1C at IberLEF 2025: Detecting and Spoiling Clickbait in Spanish-Language News, Procesamiento del Lenguaje Natural 75 (2025).

[3] G. Mordecki, G. Moncecchi, J. Couto, Te Ahorré Un Click: A Revised Definition of Clickbait and Detection in Spanish News, in: Proceedings of Iberamia 2024, 2024.

[4] G. Loewenstein, The Psychology of Curiosity: A Review and Reinterpretation, Psychological Bulletin 116 (1994) 75–98.

[5] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[7] J. Tiedemann, The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT, in: Proc. of the 5th Conference on Machine Translation, ACL, 2020, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.

[8] J. Wei, K. Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/v1/D19-1670.

[9] S. S. Al-Azzawi, G. Kovács, F. Nilsson, T. Adewumi, M. Liwicki, NLP-LTU at SemEval-2023 Task 10: The Impact of Data Augmentation and Semi-Supervised Learning Techniques on Text Classification Performance on an Imbalanced Dataset (2023). arXiv:2304.12847.

[10] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, https://explosion.ai/blog/spacy-2-nlp-updates, 2017.