Preface on the Iberian Languages Evaluation Forum (IberLEF 2025)

IberLEF is a shared evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages. This annual cycle begins in December with a call for task proposals and concludes in September with an IberLEF meeting held alongside SEPLN. Throughout this period, various challenges are conducted with significant international participation from academic and industry research groups. The aim is to inspire the research community to organize competitive tasks in text processing, understanding, and generation, thereby establishing new research challenges and advancing state-of-the-art results in these languages.

In its seventh edition, IberLEF 2025 has contributed to the field of NLP in Spanish and other Iberian languages with a total of 445 researchers from 196 research groups in 21 countries participating in 14 NLP challenges in Spanish, Portuguese, and English.

This volume begins with an overview of all the activities conducted during IberLEF 2025, along with aggregated figures and insights about the various tasks. It also includes a collection of papers describing the participating systems. However, the task overviews are not included in these proceedings; they have been published in the September 2025 issue of the journal *Procesamiento del Lenguaje Natural*.

IberLEF 2025 has addressed the following tasks:

1 Language Comprehension

ADoBo 2025 is the second edition of the ADoBo shared task. The first edition was held in 2021 and addressed the automatic detection of unassimilated borrowings in the Spanish press. This second edition of the task focused on the *automatic detection of anglicisms* (i.e., unassimilated lexical borrowings from English) in Spanish journalistic texts. Participants were asked to return annotated spans of anglicisms from a set of Spanish sentences. Unlike the 2021 edition, no training set was provided, although a development set was made available. The development set released was the same used in the 2021 edition of ADoBo, specifically including only sentences that contained anglicisms and no lexical borrowings from other languages. The test set provided was BLAS (Benchmark for Loanwords and Anglicisms in Spanish). BLAS consists of 1,836 annotated sentences in Spanish (37,344 tokens), which contain 2,076 spans labeled as anglicisms. The task was conducted entirely in Spanish and the evaluation was based on strict span-level precision, recall, and F1-score. A total of 14 teams registered for the task, out of which 6 teams submitted results on the test set and 5 teams sent working notes. Participants submitted solutions using LLMs, deep learning models, Transformer-based models, and rule-based systems. The best performing team, qilex, achieved an F1 score of 98.79 using an `OpenAI o3` model with an enriched prompt that included explicit guidelines along with reminders.

CLEARS explores automated techniques for adapting Spanish texts into plain language and easy-to-read formats. The task is divided into two subtasks: one focused on *plain language adaptation* and the other on *easy-to-read adaptation*. The dataset consists of 3,000 news articles from various municipalities in the province of Alicante (Spain), covering a wide range of topics.

Each article was adapted into both plain language and easy-to-read versions following the general guidelines of the Asociación Española de Normalización (UNE), with all adaptations reviewed and validated by a team of field experts. Participants' submissions were evaluated using lexical and semantic similarity measures, along with readability scores. In total, four teams participated in Subtask 1 and five teams in Subtask 2. The top-performing systems in both subtasks used prompting techniques with instruction-tuned LLMs. Team HULAT-UC3M achieved the best results in Subtask 1, reaching a cosine similarity of 0.75 with a method based on prompting a LoRA-adapted `RigoChat-7B-v2` model finetuned on the provided dataset. Team NIL-UCM led Subtask 2 with a cosine similarity of 0.72, using a similar approach based on `Mistral-7B-Instruct-v0.3`.

PROFE is designed to assess the reading comprehension abilities of NLP systems, focusing on their linguistic competence under the same conditions used to evaluate humans. The task includes three subtasks: (i) *Multiple choice*, where systems must select the correct answer from a list of options for each question, (ii) *Matching*, where systems must pair texts from two different sets, similar to natural language inference and semantic textual similarity tasks, and (iii) *Fill-in-the-gap*, where systems must identify the correct position of text fragments within a masked passage. All three subtasks were evaluated using accuracy as the metric. For this task, the organizers created an evaluation dataset based on Spanish proficiency tests developed over the years by the Instituto Cervantes. A total of 19 teams registered for the task, with 8 submitting runs. The multiple-choice subtask received 24 submissions, the matching subtask 11, and the fill-in-the-gap subtask 9. Team Vicomtech achieved the highest accuracy across all three subtasks (above 93%) using ensembles of open-source large language models, such as `Qwen-2.5-14B` and `Phi-4-14B`, operating in a zero-shot setup.

2 Harmful and Inclusive Content

DIMEMEX is the second edition of DIMEMEX at IberLEF 2024, continuing its mission to advance research on automatic detection of inappropriate content in memes, with a particular focus on Mexican Spanish. This year's edition featured three subtasks: (i) Three-way classification to etermine whether a meme contains hate speech, inappropriate content, or neither, (ii) *Fine-grained classification*, where systems must assign memes to specific categories of hate speech, and (iii) *LLM-focused three-way classification*, same as subtask 1, but restricted to using LLMs only. The DIMEMEX 2025 dataset is a refined version of the previous year's, consisting of approximately 3,000 memes manually annotated for abusive content. These memes were collected from public Facebook groups in Mexico known for sharing such material. All subtasks were evaluated using macro-averaged recall, precision, and F_1 score. Ten teams participated in Subtask 1, while Subtasks 2 and 3 each saw three participating teams. Team HARGP-BETO achieved the best performance in Subtask 1 (macro-F_1 score of 0.58), using a text-only gated unit model that fuses local and global attention mechanisms based on OCR and textual descriptions. Team UC-UCO-CICESE led Subtask 2 (macro-F_1 score of 0.37) with a system that combined text and image modalities through a late fusion of `BETO` (for text) and `ViT` (for images).

HOMO-LAT25 continues the HOMO-MEX shared tasks from 2023 and 2024, extending the study of polarity detection toward LGBTQ+ content in online messages to Spanish dialects in Latin America. This year's edition focused on Reddit posts written in Spanish from 19 Latin American countries, annotated with positive, negative or neutral polarity toward specific LGBTQ+ identity keywords. The task comprised two tracks: *(i)* Track 1 evaluated *polarity detection* when training and test data came from the same Spanish dialect (Argentina, Chile, Colombia, and Mexico); and *(ii)* Track 2 evaluated *cross-dialect generalizatio*n by testing on countries unseen during training. 30 teams registered for the task, out of which 7 submitted valid results and 6 presented working notes. All participating teams used Transformer-based models, two also incorporated traditional machine learning and two leveraged large language models (LLMs). The best results were obtained by the PLD team, achieving a macro F1-score of 52.96 in Track 1 and 50.86 in Track 2. Their approach combined translating all input texts into English to take advantage of highly performing pre-trained models and mitigate dialectal variation, with a new "Context Engine" that retrieves semantically similar examples from each sentiment class (positive, negative, neutral) to enrich the model's inference process and improve generalization, especially in a challenging cross-dialect setting.

MentalRiskES aims to promote the early detection of mental risk disorders in Spanish. In this third edition of the task, the focus was on the detection of gambling disorders, with two subtasks: (1) *risk detection of gambling disorders*, (2) *risk detection of gambling disorders and determining the type of addiction.* The task used a dataset of texts from social media, annotated as low-risk and high-risk of different types of gambling disorders. This edition had 13 teams submitting results, 12 of them submitted papers. The submissions were ranked according to overall classification, early prediction, and also with efficiency metrics, emphasizing the need for sustainable practices. Team UNSL obtained the best Macro-F1 for task 1 (0.567), team MCDI obtained the best Macro-F1 for task 2 (0.589), while team PLN_PPM_ISB obtained the best results related to early prediction.

MiSonGyny focused on the automatic detection and classification of misogynistic content in Spanish song lyrics. It was designed to address the underexplored presence of symbolic violence and hate speech in musical texts, which often include subtle and metaphorical expressions of misogyny. The task comprised two subtasks: *(i)* Subtask 1, *binary classification* of song verses as *Misogynistic (M)* or *Non-Misogynistic (NM)*; and *(ii)* Subtask 2, *fine-grained classification* of misogynistic content into *Sexualization (S)*, *Violence (V)*, *Hate (H)*, or *Not Related (NR).* A total of 13 teams participated in Subtask 1 and 9 in Subtask 2, out of which 9 submitted working notes. Most approaches relied on transformer-based architectures, complemented by traditional machine learning, data augmentation and, in some cases, LLMs or hierarchical pipelines. The best-performing team in both subtasks was HULAT UC3M, achieving an F1-score of 0.8811 in Subtask 1 and 0.5895 in Subtask 2. This team developed a comprehensive pipeline that combines data augmentation, transformer-based encoders, and traditional machine learning methods. In addition, it addressed class imbalance through minority class oversampling using back-translation and AEDA techniques.

PolyHope aims to detect hope speech (messages that express optimism, encouragement, or the desire for a better future) in English and Spanish social media. Two subtasks were proposed: (a) *binary hope speech detection*, and (b) *multiclass detection* with the categories generalized hope, realistic hope, unrealistic hope, not hope, or the sarcasm category meant to detect hopeful language that is used in a misleading way. The dataset used contains over 30 thousand tweets annotated with hope speech information, around two thirds in Spanish and the rest in English. 31 participants took part in the competition, with 13 papers accepted. The best performances for the binary task were 0.852 F1 for Spanish by teddymas, and 0.871 F1 for English by michaelibrahim. For the multiclass task, the best performances were 0.742 macro-F1 for Spanish by lephuquy, and 0.755 macro-F1 for English by supachoke. Common challenges mentioned by teams inlcuded imbalanced data, language mixing, and cultural differences in how people express emotions.

3 Content Curation and Generation

PastReader focuses on the automatic transcription of digitized Spanish historical newspapers. The task includes two subtasks: (i) *Error Correction*, where participants receive the output of an OCR system and must generate clean, corrected versions of the extracted texts; and (ii) *End-to-end Extraction*, which explores full pipeline approaches that take scanned pages as input and produce curated transcriptions as output. The corpus used in this task consists of historical newspaper publications from the public domain, digitized by the National Library of Spain (BNE) and available through the Hemeroteca Digital. It includes 298 press titles, 88,748 issues, and a total of 8,302,407 pages in PDF format. For the shared task, the organizers sampled 121,295 documents and transcriptions, which were split into training, validation, and test sets in a 74-4-22 ratio. Evaluation relied on standard text generation metrics such as Word Error Rate, (Normalized) Levenshtein Distance, BLEU, and ROUGE, as well as sustainability metrics, including CO_2 emissions. Only Subtask 2 received participation, with three teams submitting systems. Team OCRTIST achieved the best performance based on the primary ranking metric (Levenshtein distance of 53.30). Their system used `Gemini 2.5 PRO` in a standalone setup, relying on a single prompt to perform OCR directly from scanned images.

PRESTA is a task about *question answering over tabular data in Spanish*, where participants had to interpret natural language questions to obtain data from tabular sources. The dataset contains 31 thousand data rows from 10 different sources, with a total of 200 question-answer pairs for training, and 100 pairs for test. The question-answer pairs could have different expected answer types: boolean, categorical, numeric, or list (either of categories or numbers). The evaluation metric for the task is average accuracy over all categories. Seven teams submitted systems to the competition, with both the ITU NLP and sonrobok4 teams obtaining the best performances of 87% accuracy. ITU NLP uses a zero-shot LLM-based Python code generation method experimenting with several models of the `OpenAI`, `Qwen`, `DeepSeek` and `LLaMA` families. sonrobok4 uses a multi-prompt strategy, also using zero-shot on LLMs with `OpenAI` and `DeepSeek` models. One of the conclusions is that current LLM technologies outperform traditional pipelines, but also small open-source models could work on par with bigger ones if properly used.

TA1C focuses on the detection and spoiling of clickbait in Spanish news, particularly in tweets that links to a piece of news. The task consists of two subtasks: *i) Clickbait Detectio*n, a binary classification task to determine whether a news teaser is clickbait based on the information gap theory where headlines deliberately omit key information to provoke curiosity; and *ii) Clickbait Spoiling*, a generative task that requires producing a concise Spanish text that fills the information gap created by the clickbait. The dataset provided includes 4,200 manually annotated Spanish tweets for the clickbait detection task and 500 human-written spoilers for the spoiling task, all collected from 18 media outlets across 12 Spanish speaking countries and international sources. A total of 27 teams registered for the task, out of which 13 participated in the evaluation phase of the clickbait detection task and 3 teams in the spoiling task. The best-performing team in the detection task, UmuTeam, achieved an F1-score of 0.8156 using an ensemble of fine-tuned transformer models, including `MarIA`, `BERTIN`, `ALBETO`, and the decoder-only `Gemma-2-2B-it` with `QLoRA` fine-tuning. In the spoiling task, the top system in manual evaluation, submitted by CogniCIC, obtained a score of 3.88 out of 5 in Accuracy/Completeness, the highest among all participants, using a few-shot prompting approach with the `Claude Sonnet 4 LLM`.

4 Sentiment and Figurative Analysis

ASQP-PT is a shared task about aspect based sentiment analysis in Portuguese, consisting of four subtasks: (1) *aspect term extraction*, (2) *opinion term extraction*, (3) *aspect category detection*, and (4) *aspect sentiment quadruple prediction* (ASQP). The task presented a corpus of 1236 Trip Advisor reviews in Portuguese about hotels in four cities, annotated with 5749 *(Category, Aspect, Opinion, Polarity)* quadruples. Two teams took part in the task, one of them in all four subtasks, and the other one only in the aspect term extraction subtask. The teams could not beat the finetuned Portuguese BERT baseline for tasks 1, 2 and 3, but for the ASQP task the best result was 0.46 F-Measure by the ABCD team, using an end-to-end encoder-decoder transformers model, beating the baseline of 0.40.

REST-MEX 2025 is the fourth edition of the REST-MEX shared task, aimed at advancing natural language processing for tourism in the Mexican context, with a focus on sentiment analysis and classification of user-generated texts about Mexico's Magical Towns (Pueblos Mágicos). The task is structured into three subtasks: *i) polarity prediction*, a fine-grained classification into five levels of polarity (from 1 to 5); *ii) service type classification*, identifying whether the review refers to a hotel, restaurant, or tourist attraction; and *iii) geographical identification of the visited location*, a multiclass classification task to determine which of the 40 predefined Magical Towns is being reviewed. The corpus consists of 297,217 TripAdvisor reviews shared by tourists who visited representative destinations in Mexico. A total of 32 teams participated in the shared task. The best performing team in all three subtasks was UDENAR, which obtained a macro F1 score of 0.64 in prediction of polarity, 0.99 in classification of type of service, and 0.69 in geographical identification. Their approach transformed each multiclass task into independent binary classification problems (one per class), using centroid-based sampling on balanced datasets to address class imbalance, particularly improving performance on minority classes like negative polarity. They combined fine-tuned transformer models with knowledge transfer techniques to enhance generalization and robustness across the three tasks.

SatiSPeech proposes research on the *automatic recognition of satire in Spanish* in two independent subtasks: the first one is *text-only*, based on YouTube videos transcriptions; the second one is *multimodal*, combining textual and acoustic information. The dataset comprised 8000 short audio segments (no longer than 25 seconds) and their transcription obtained from popular satirical programs in YouTube, each segment was annotated as satirical or non-satirical. There were 11 teams that participated in the task, and all of them submitted systems for both tasks. For subtask 1, the best performance was 85.6 F1, obtained by the UPV-ELiRF team combining several `BERT` and `LLM` models; while for subtask 2, the top performance was 88.3 F1 by the UMU-Ev team using `HuBERT` and prosodic features combined with a `SVM` classifier.

In the realm of Natural Language Processing, where Deep Learning and, more specifically Large Language Models, have become the go-to solutions, defining research challenges and creating robust evaluation methods and high-quality test collections are crucial for success. These elements enable iterative testing and refinement. IberLEF is playing an important role in advancing these efforts and moving the field forward.

September 2025.
The editors.