Preface on the Iberian Languages Evaluation Forum (IberLEF 2026)

IberLEF is a shared evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages. This annual cycle begins in November 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 2026 edition of IberLEF takes place in León, Spain, on 22 September 2026. 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 eighth edition, IberLEF 2026 has contributed to the field of NLP in Spanish and other Iberian languages with a total of 558 researchers from 178 research groups in 26 countries participating in 17 NLP challenges in Spanish, Catalan, Italian, Arabic, English.

This volume begins with an overview of all the activities conducted during IberLEF 2026, 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; published in the September 2026 issue of the journal *Procesamiento del Lenguaje Natural*.

1 Shared Tasks

IberLEF 2026 addresses the following 17 tasks, grouped by their overarching themes.

1.1 Language Comprehension

NiVELE, *Nivelación automática de textos de estudiantes de ELE*, focused on predicting Common European Framework of Reference levels from learner written productions. The task, organized on Kaggle, registered 15 enrolled teams, out of which 9 actively participated and 5 submitted working notes describing their systems. The results demonstrate that fine-tuning pre-trained encoder models in Spanish represents the most robust approach, while highlighting the challenge of mitigating thematic overfitting.

PROFE, *Language Proficiency Evaluation*, evaluates automatic Spanish reading comprehension systems using exercises drawn from real Instituto Cervantes exams, preserving the evaluation conditions of human students and without providing task-specific training data. It is the second edition of the task, with the main novelty being the inclusion of exercises with images, which makes it possible to assess the extent to which systems can combine textual and visual information while also deciding when visual cues are relevant or merely decorative. This task registered 9 teams, and 3 of them submitted official runs. The results show strong performance on multiple-choice exercises, notable progress on matching exercises, and greater difficulty in fill-in-the-gap exercises, as well as substantial differences between exercises with and without images depending on the approach used.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-09-29

1.2 Harmful and Inclusive Content

HOPE-EXP, *Outcome-Oriented Expectation Analysis in Social Media*, evaluates social media hope-detection systems by framing the task as an outcome-oriented, multi-level structured prediction problem using bilingual (English/Spanish) posts. The task contains four subtasks: (i) assign one of six primary hope labels, (ii) detect a multi-label set of trigger emotions, (iii) extract up to three spans naming the expected or avoided outcome, and (iv) classify the stance (Desired/Avoided) and actor (Self/Other/World- System/Unclear) of each span. The corpus comprises 4,857 training and 2,082 test posts from Reddit. In this task, 23 participants registered and submitted 195 runs; eight systems obtained a leaderboard score, and 6 teams submitted a system description paper. Approaches ranged from lexical baselines to fine-tuned transformers (XLM-RoBERTa) and parameter-efficient LLMs (QLoRA-tuned Qwen3.5-4B), with the best system reaching 0.785 overall. Primary label classification was comparatively tractable, while span extraction and actor classification proved the hardest components.

MiSonGyny, *Misogynistic Speech Detection in Spanish Song Lyrics from Latin America and Spain*, is the second edition of this shared task. This edition of the task comprises three subtasks: (i) binary misogyny detection at the song level (Misogynistic/Non-Misogynistic), (ii) multi-label classification of misogyny type (Sexualization, Violence, and Hate); and (iii) binary gender stereotype identification (Stereotypical/Non-Stereotypical). The corpus incorporates proportional representation of Latin American and Peninsular Spanish varieties, annotated by native speakers from both regions. A total of 11 teams participated in the three subtasks, submitting 8 working notes. Most approaches employed transformer-based architectures, large language models, and ensemble methods. The best-performing system achieved a macro-F1 of 0.8856 on Subtask 1, 0.7385 on Subtask 2, and 0.8283 on Subtask 3.

WomenHelp-2026, *Classification of Gender-Based Violence Reports to Help Threatened Women in Northern Mexico*, focuses on the analysis and automated classification of real narrative reports documenting gender-based violence, collected by a government institution in northern Mexico between 2016 and 2024. To this end, the Women Help 2026 corpus is introduced, which comprises 21,292 previously anonymized narratives. The task was divided into two subtasks: (i) determining the case's severity levels and (ii) identifying specific types of violence using a multi-label scheme across six distinct categories. In the forum, 16 teams participated, and 10 of them submitted working notes, primarily leveraging architectures based on pre-trained Transformer models—including both Spanish-specific and multilingual approaches—as well as classical machine learning methods. Experimental results show that multi-model systems utilizing Spanish-specific encoders achieved the highest performance in capturing the semantic nuances of severity in the first subtask, whereas fine-tuning single-model setups proved more effective for accurately discriminating the types of violence in the second subtask.

1.3 Clinical NLP

GRACE, *Granular Recognition of Argumentative Clinical Evidence*, is the first shared task on clinical argument mining in Spanish. GRACE comprises two subtasks: (i) focused on argumentative component and relation identification in Randomized Controlled Trial abstracts and (ii) hierarchical argument mining in clinical cases from the Spanish MIR examination. A total of 16 teams registered, of which 6 submitted official runs and system description papers. Participants explored diverse approaches, including encoder-based Transformer models, large language models, ensemble methods, and zero-shot prompting. The results demonstrate the fea-

sibility of the proposed tasks while highlighting the challenges posed by clinical argument mining.

MentalRiskES 2026, *Early Detection of Mental Disorders Risk in Spanish*, is the forth edition of this long-running shared task. The goal of this shared task is to advance the early detection of emotional problems in Spanish. It comprises two subtasks: (i) focuses on the early detection of symptoms in therapeutic conversations and (ii) addresses the selection of the most appropriate therapist response. In addition, participants were asked to report the carbon emissions generated by their systems, emphasizing the importance of sustainable NLP practices. In this fourth edition, 30 teams registered, 17 submitted predictions, and 11 submitted system description papers. Most participants experimented with large language model-based approaches, primarily relying on zero-shot prompting strategies with both open-source and commercial models. In the first subtask, the best system achieved a combined MAE of 0.879. The second subtask proved to be considerably more challenging, with the best Macro-F1 of 0.392, only marginally above the random baseline.

MIRROR, *Motivational Interviewing Response & Rating via Synthetic cOnversational tuRns*, aims to promote research on Generative Artificial Intelligence methods for producing synthetic conversation turns that can improve the performance of models trained to recognize Motivational Interviewing Behavior Codes. The task is formulated as data-centric, asking participants to generate datasets for fine-tuning pre-trained models for distinguishing Behavior Codes as follows: (i) Simple Reflection vs. Complex Reflection; (ii) Open Question vs. Closed Question; and (iii) Persuasion vs. Giving Information. A total of 6 teams submitted their generated datasets for the final evaluation. Two teams outperformed the baseline, and novel sample generation strategies were proposed in the context of this competition. An extensive analysis of the results is presented, proving the potential impact of Generative Artificial Intelligence strategies for robust synthetic data augmentation in the context of Motivational Interviewing analysis.

1.4 NLP for Inclusion

MER-TRANS, *First Shared Task on Multilingual Easy-to-Read Translation*, challenges the NLP community to develop solutions that produce easy-to-read versions of texts and sentences in Arabic, Catalan, Italian, and Spanish. The task features a multilingual corpus of original documents and their easy-to-read adaptations in Catalan, Italian, and Spanish, as well as a sentence-aligned dataset of original sentences and their simplifications in Arabic. MER-TRANS attracted 7 teams that submitted a total of 42 systems, all of which were evaluated using automatic metrics.

MSLG-SPA, *Bidirectional Translation between Mexican Sign Language Glosses and Spanish*, is devoted to bidirectional translation between Mexican Sign Language glosses (MSLG) and Spanish. The task comprised two subtasks: Gloss-to-Spanish and Spanish-to-Gloss translation. 20 teams submitted a total of 100 system runs, exploring a wide range of approaches including multilingual encoder–decoder models, instruction-tuned large language models, parameter-efficient fine-tuning, retrieval-augmented generation, and hybrid rule-based systems. Systems were evaluated using BLEU, METEOR, chrF, and COMET, together with a standardized multi-metric ranking strategy. MSLG-SPA establishes one of the first public benchmarks for Mexican Sign Language translation, providing resources to support future research on low-resource sign language processing and accessibility technologies.

SSD-2026, *Social Support Detection and Target Identification in Social Media*, evaluates

NLP systems for detecting support in English and Spanish YouTube comments, identifying its target, and classifying the supported community. The task comprised three hierarchical subtasks and attracted 72 participants, 387 submissions, and 18 system description papers. Approaches ranged from classical machine learning to transformer-based, multi-task, byte-level, and LLM-assisted methods. Results were strong for binary detection but more limited for fine-grained classification under class imbalance, with Spanish systems generally outperforming English ones.

1.5 Sentiment and Figurative Analysis

HAHA-2026, *Humor Analysis based on Human Annotation and Automatic Humor Generation*, is the fourth edition of this shared task, aiming to assess humor detection systems. This year's edition of HAHA includes the classic task of humor detection, alongside a new track on LLM-generated humor detection. Additionally, it introduces a third task on humor generation subject to human preference obtained from a LLM arena style evaluation. For this task, a total of 50 teams registered, out of which 7 sent results for some of the subtasks, and 6 submitted working notes. Participants obtained almost perfect results for humor detection, and good results for automatic humor identification, but for humor generation systems still generate jokes with inconsistencies.

HISEMOTIONS 2026, *Historical Text-Based Emotion Detection in Early Modern Spanish Correspondence*, is focused on Text-Based Emotion Detection in Early Modern Spanish epistolary texts from the 16th and 17th centuries. The task aims to foster the development and evaluation of computational methods specifically adapted to historical Spanish language varieties, addressing the semantic and diachronic challenges of expressing emotion and affective states in Early Modern Spanish correspondence. The core challenge proposed within HISEMOTIONS is Multi-label Emotion Detection, oriented toward predicting the emotion perceived by a reader when interpreting an author's text. The final ranking encompasses the results of 7 teams, each of which proposed a distinct approach to the problem.

RESTMEX-2026, *Research on Synthetic Text Modeling for Mexican Magical Towns*, is the fifth edition of this shared task. Unlike previous editions, which focused on sentiment classification, REST-MEX 2026 introduced a new evaluation framework centered on synthetic tourism review generation. Participants generated synthetic reviews in Mexican Spanish to improve a domain-adapted RoBERTa sentiment classifier through fine-tuning. All submitted corpora were evaluated on the same hidden test set of authentic reviews using a weighted Macro-F1 metric that emphasizes minority sentiment classes. The shared task attracted 28 participating systems exploring diverse approaches, including prompt engineering, supervised fine-tuning, prototype retrieval, multi-agent architectures, diffusion language models, lexical augmentation, bio-inspired methods, and automatic filtering strategies. The results show that well-designed synthetic corpora can substantially improve downstream sentiment classification, with several systems outperforming the baseline model trained without synthetic data.

SPEECHMATICS, *Speech Multimodal Audio-Text Irony Classification in Spanish*, aims to advance research on automatic irony detection in spontaneous spoken Spanish from both textual and multimodal perspectives. The task is divided into two independent subtasks. The first subtask focuses on irony detection using only manual transcriptions of audio fragments. The second subtask introduces a multimodal setting in which systems can combine textual and acoustic information. SpeechMATICS is based on MESironic, a multimodal corpus of short spoken fragments with manually verified transcriptions and irony annotations. A total

of 14 teams registered for the task, 4 submitted official runs, and 4 submitted working notes describing their systems.

1.6 Information Extraction & Ranking

GenSIE, *General-purpose Schema-guided Information Extraction*, focuses on advancing research for schema-guided information extraction from Spanish text using small language models. Participants build CPU-side agents that consume a (text, instruction, JSON schema) triplet and emit a JSON object that conforms to the schema while staying faithful to the text, including the explicit obligation to return null for fields the text does not support. The benchmark comprises 271 instances in Spanish, spanning eight development domains and sixteen schemas. GenSIE evaluates how much of the baseline-to-perfect error each system removes, averaged across two published models. A total of 11 teams submitted thirty-five pipelines with the best system closing 21.6% of the baseline-to-perfect gap.

PoliticHeadlinES, *Multimodal Headline Ranking in Spanish Political News*, focuses on headline ranking for Spanish political news. Given a news article and ten candidate headlines, systems must rank them by relevance, placing the original headline at the top. The competition comprised two subtasks: a text-only setting and a multimodal setting incorporating the article's image. The dataset contains over 20,000 political news articles from multiple Spanish media outlets, using real-world distractor headlines to create plausible alternatives. A total of 19 teams participated, submitting 168 runs, and 16 of them submitted working notes. Results show that headline ranking remains highly challenging; the inclusion of visual information yielded limited improvements over text-only approaches. Systems leveraging large language models and reasoning-oriented strategies achieved the top overall performance, emphasizing the need for deep contextual and semantic matching.

2 Review Process

At the beginning of the cycle, in November, the organizing committee launches the call for task proposals. Once the proposals have been received, the organizing committee assigns each proposal to two program committee members for review, based on their expertise in the relevant task domain.[1] Each reviewer completes a standard form addressing the relevance of the proposed task, the availability of annotation resources, the feasibility of the proposed timeline, the organizers' experience (e.g., PhD status, previous experience organizing shared tasks), and the suitability of the evaluation metrics. Reviewers are also asked to provide a final decision: "accept as is", "accept subject to improvements", or "reject". If both reviewers recommend rejection, the task is rejected. If both reviewers recommend either form of acceptance, the task is accepted, with the requested changes incorporated where necessary. In the case of mixed recommendations, the organizing committee makes the final decision. Of 21 proposals received, 4 were rejected (20%).

Once decisions on the task proposals have been made, the accepted tasks proceed according to independent schedules, with a single fixed deadline for the submission of both the organizers' overview papers and the participants' working notes (3 July for this edition). Prior to this deadline, task organizers request working notes from the participants. These are then reviewed by the task organizers themselves and, where appropriate, by any external reviewers they choose to involve. The review considers the overall quality of the work, the results obtained in the

[1] The program committee consists of the organizers of shared tasks held during the three years preceding the current edition. It is a heterogeneous group comprising researchers from both academia and industry.

shared task, compliance with the CEUR template, and any other criteria the organizers deem necessary to ensure the quality of the working notes. After the task organizers have reviewed the working notes, they are submitted to the organizing committee for final preparation and submission to CEUR. As part of this process, the organizing committee conducts an additional review of their overall quality and compliance with the CEUR template. If, after submitting their working notes, a participant fails to respond to requests for revisions or other changes during the review process, the working note is rejected.

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 2026.
The editors.