Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages

Alba Bonet Jover[1], José Ángel González-Barba[2] and Luis Chiruzzo[3]

[1]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Spain
[2]TransPerfect, Spain
[3]Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay

Abstract

IberLEF is a shared evaluation campaign for Natural Language Processing systems focused on Spanish and other Iberian languages, organized annually since 2019 as part of the conference for the Spanish Society for Natural Language Processing. Its aim is to inspire the research community to develop and participate in competitive tasks related to text processing, understanding, and generation. These efforts are geared towards defining new research challenges and setting state-of-the-art results in Iberian languages, including Spanish, Portuguese, Catalan, Basque, and Galician. This paper provides an overview of the evaluation activities conducted during IberLEF 2026, which featured 17 tasks and 40 subtasks. These tasks covered various areas such as language comprehension, harmful and inclusive content, clinical NLP, NLP for inclusion, sentiment and figurative analysis, and information extraction and ranking. Overall, the IberLEF 2026 activities represented a significant collaborative effort, involving more than 558 researchers from 26 countries across Europe, Asia, Africa, and the Americas.

Keywords

Natural Language Processing, Artificial Intelligence, Evaluation, Evaluation Challenges

1. Introduction

IberLEF (Iberian Languages Evaluation Forum) is a shared Natural Language Processing (NLP) evaluation campaign focused on Spanish and other Iberian languages. It is organized annually since 2019, as part of the International Conference of the Spanish Society for Natural Language Processing (SEPLN). It aims to inspire the research community to develop and participate in competitive tasks related to processing, understanding, and generation of at least one of the Iberian languages, including: Spanish, Portuguese, Catalan, Basque, and Galician. These efforts are geared towards defining new research challenges and improving the state-of-the-art results in these languages.

In this shared evaluation campaign, the research community defines new challenges and proposes tasks to advance the NLP state of the art. The task proposals are reviewed by the IberLEF steering and program committees, and then evaluated by the IberLEF general chairs. The organizers of the accepted tasks are in charge of setting up the evaluation according to their proposal, promoting the task, and managing the submissions and scientific evaluation of system description papers written by participants. These papers are included in this IberLEF proceedings volume, published at CEUR Workshop Proceedings. In addition, task organizers must prepare and submit an overview of their task and evaluation, which are reviewed by the IberLEF organizing committee and published in the journal *Procesamiento del Lenguaje Natural, vol. 77 (September 2026 issue)*. Finally, the task organizers report the results of the tasks, and selected participants present descriptions of their systems at the IberLEF workshop.

IberLEF 2026 takes place on September 22, 2026, in León (Castilla y León, Spain), as part of the XLII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2026). This

IberLEF 2026 September 2026, León, Spain
alba.bonet@dlsi.ua.es (A. B. Jover); jgonzalez@transperfect.com (J. González-Barba); luis.chiruzzo@gmail.com (L. Chiruzzo)
https://cvnet.cpd.ua.es/curriculum-breve/es/bonet-jover-alba/61154 (A. B. Jover); https://jogonba2.github.io/ (J. González-Barba); https://www.fing.edu.uy/index.php/es/node/40865 (L. Chiruzzo)
0000-0002-7172-0094 (A. B. Jover); 0000-0003-3812-5792 (J. González-Barba); 0000-0002-1697-4614 (L. Chiruzzo)

year, 17 shared tasks were accepted to be organized as part of IberLEF 2026, out of 21 proposals. These tasks focus on a range of NLP challenges, including language comprehension, harmful and inclusive content, clinical NLP, NLP for inclusion, sentiment and figurative analysis, and information extraction and ranking.

In this paper, we provide a summary and analysis of the tasks organized in IberLEF 2026 to offer a clearer understanding of this collective effort.

2. IberLEF 2026 Tasks

The 17 tasks involved in IberLEF 2026 are presented below, grouped by theme.

2.1. Language Comprehension

NiVELE [1], *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 [2], *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.

2.2. Harmful and inclusive content

HOPE-EXP [3], *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 [4], *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 [5], *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.

2.3. Clinical NLP

GRACE [6], *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 feasibility of the proposed tasks while highlighting the challenges posed by clinical argument mining.

MentalRiskES 2026 [7], *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 [8], *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.

2.4. NLP for Inclusion

MER-TRANS [9], *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 [10], *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 [11], *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.

2.5. Sentiment and Figurative Analysis

HAHA-2026 [12], *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 [13], *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 [14], *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 [15], *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.

2.6. Information Extraction & Ranking

GenSIE [16], *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 [17], *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.

3. Aggregated Analysis of IberLEF 2026 Tasks

3.1. Tasks characterization

Figure 1 shows the distribution of **languages** per task used throughout the evaluation campaign. Once again this year Spanish is the dominant language, being used in 17 tasks, followed by English with 2 tasks, and Catalan, Italian, and Arabic with 1 task, meaning that three of the five languages used this year are non-Iberian. Among the Spanish varieties considered this year, the predominant ones were the European (Spain) and Mexican (Mexico) varieties, but varieties from other Latin American countries

were also considered, including Colombia, Puerto Rico, Dominican Republic, Argentina, Chile, Ecuador, Guatemala, Peru, Uruguay, and Venezuela.

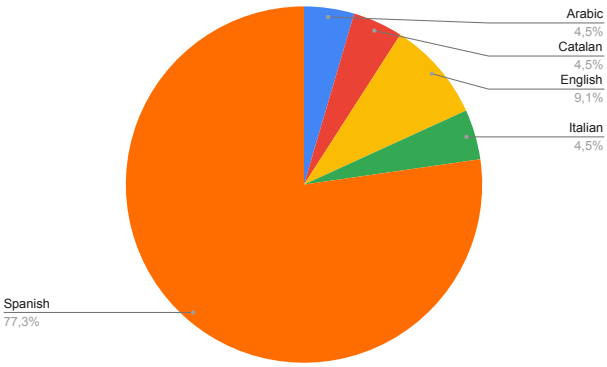

Figure 1: Distribution of languages in IberLEF 2026 tasks.

Looking at the distribution of task types, shown in Figure 2, the most common task type is binary classification, appearing in subtasks across nine shared tasks. Multi-class classification follows, included in seven tasks. In a shift from previous editions, binary classification has overtaken multi-class classification in popularity at IberLEF 2026. Interestingly, text generation tasks are now as prevalent as multi-class classification, with text generation experiencing a 40% increase compared to IberLEF 2025. This is in line with a larger trend in the NLP community as a whole, which is becoming more interested in generative tasks due to the availability of new kinds of generative language models.

The distribution of the main **evaluation metrics** used this year is shown in Figure 3. As in previous years, Macro F_1 remains predominant, being used in nine classification tasks, with five of them also incorporating Precision, Recall, and binary F_1. Excluding Hamming loss, Micro F_1, PA-nDCG, and weighted F_1, almost all metrics in the tail of the distribution correspond to generative tasks, reflecting their prevalence in IberLEF 2026. Interestingly, the diversity of metrics used for text generation tasks is notably higher than for classification tasks, highlighting the ongoing challenges in evaluating generated text, a general reluctance to rely on a single metric, and even the necessity of relying on human evaluation.

IberLEF 2026 incorporated eleven tasks that were not organized before or were considered novel enough to be counted as new (65%), while six of the tasks were new editions of previously run competitions. This strikes a good balance between **novelty and stability**, as successful tasks from previous years such as MiSonGyny, PROFE, HOPE-EXP, MentalRiskES, HAHA, and Rest-MEX, had new editions, while also the campaign introduced new challenges for Iberian language processing which attracted more researchers.

3.2. Datasets and results

Figure 4 shows statistics on the types of **data sources**. As in previous years, the datasets include a wide variety of sources, with news outlets being the most common. Breaking with a long-standing trend, this edition marks the complete absence of Twitter/X, with YouTube, Reddit, and TripAdvisor persisting as the only social media sources. New additions this year include historical letters, sign language resources, a broad range of legal sources (covering justice and public deliberation), and clinical sources (comprising interviews, conversations, reports, and trials).

In terms of **dataset sizes** and annotation efforts[1], making fair comparisons is challenging due to the

[1]Overall, the annotation efforts in IberLEF 2026 continue to make a significant contribution to expanding test collections for Spanish and, to a lesser extent, other languages. Once again, IberLEF has been conducted without specific funding sources, relying instead on the resources obtained individually by the teams organizing and participating in the tasks. Implementing a centralized funding model could undoubtedly help achieve larger and more comprehensive annotations across IberLEF as a whole.

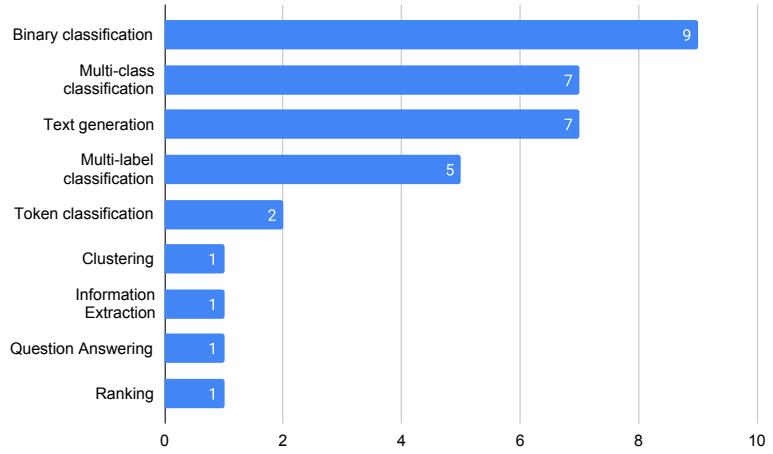

Figure 2: Distribution of task types in IberLEF 2026.

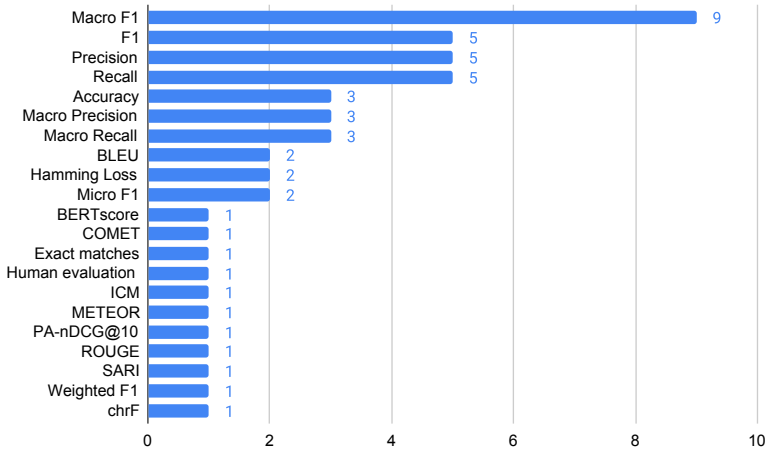

Figure 3: Distribution of official evaluation metrics in IberLEF 2026 tasks.

diversity of data sources, variations in text lengths, and the wide range of annotation difficulties. In most of the tasks (11 out of 17), the datasets do not contain any content automatically generated or labeled without direct human intervention. However, we observed a substantial increase in the number of tasks providing this kind of mixed data compared to previous years. While only one task did so in IberLEF 2025, six tasks include such data this year, representing a 500% increase.

Three datasets contain more than 15,000 instances (Rest-MEX, PoliticHeadlinES, and WomenHelp), one has between 10,000 and 15,000 instances (SSD-2026), two fall between 5,000 and 10,000 instances (HOPE-EXP and SpeechMATICS), and the remaining datasets contain less than 5,000 samples. Regarding annotation reliability, inter-annotator agreement serves as a useful indicator and is reported for 7 out of 17 tasks. Fleiss' Kappa and Cohen's Kappa are the most common statistical measures used to assess agreement reliability, each appearing in ~50% of the cases. Among these, one task show high agreement, one has moderate-high agreement, and five show from low to moderate agreement.[2]

Regarding **progress relative to the state of the art**, it remains challenging to draw overarching conclusions for the entire IberLEF effort due to the varied approaches used for establishing task baselines. Figure 5 shows a pairwise comparison between the best system and the best baseline for each task where at least one baseline is provided and one result was submitted, using the official ranking metric

[2]Generally, moderate agreement may reflect the complexity of the task rather than deficiencies in the annotation guidelines.

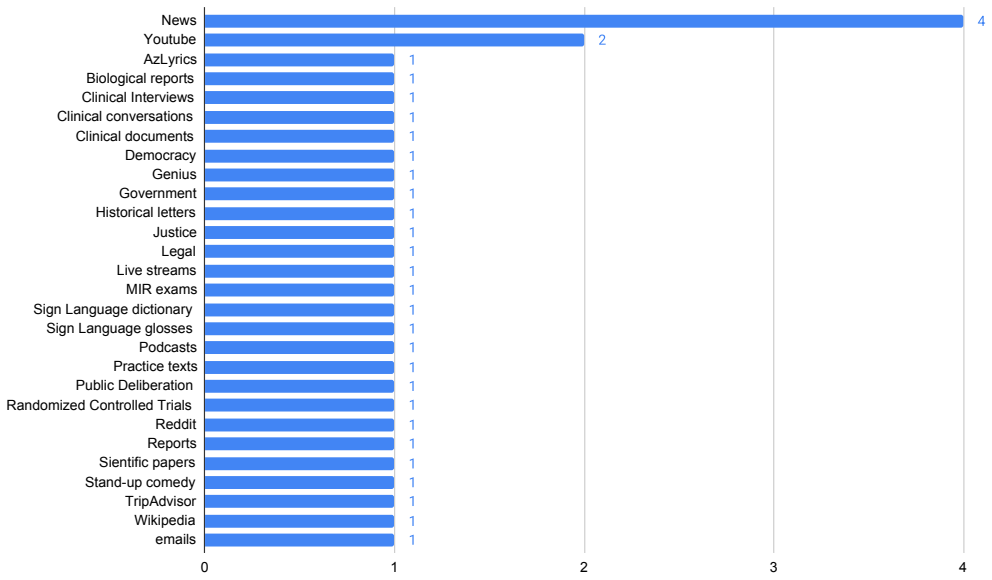

Figure 4: Types of textual sources in IberLEF 2026 tasks.

for each task. To avoid confusion, the chart is limited to tasks where the official metric ranges from 0 (worst quality) to 1 (perfect output).

Five tasks (PROFE, HOPE-EXP, MSLG-SPA, SSD-2026, and GenSIE) did not provide any baseline. Almost every other task included baselines based in pre-trained Transformers, either encoder (29%), encoder-decoder (17%), and decoder-only (29%) models depending on the nature of the task. Majority or random baselines (18%) and classical machine learning including SVM, Random Forest, and Logistic Regression were also used in some tasks (24%). Among the subtasks featuring baselines, the top-performing system outperformed the baseline by a relative margin of over 10% in 20 cases (83% of the total), while the systems could not beat the baseline in only 1 case by a small margin. Examining the results, only 3 subtasks had the top-performing system scoring higher than 0.9, which suggests that there is still room for improvement in most cases. None of the subtasks presented a baseline that performed above this 0.9 level, showing that either there was a preference for weaker baselines, or the tasks were indeed designed to be more challenging.

3.3. Participation

Despite IberLEF 2026 not being a funded initiative, participation was impressive, with a significant portion of current research groups interested in NLP for Spanish and other Iberian languages either organizing or participating in one or more tasks. In total, 558 researchers from 178 research groups across 26 countries in Europe, Asia, Africa, and the Americas were involved in IberLEF tasks[3]. Compared to IberLEF 2025, the number of participating researchers increased by 25%, highlighting the growing interest in NLP evaluation campaigns on Iberian languages.

Figure 6 shows the distribution of research groups by country. Following the trend of the last year, Mexican research groups now have the highest participation, with 69 groups, followed by Spain with 47 groups, and Cuba with 13.

Figure 7 illustrates the distribution of researchers (listed as authors in the working notes) by country. The top five countries —Spain, Mexico, Cuba, Colombia, and India— account for approximately 85% of the participating researchers. There is a noticeable difference between the distributions of researchers

[3]Statistics were compiled from the submitted working notes, which implies two things: *i)* Some groups and researchers may be counted more than once if they participated in multiple tasks; and *ii)* actual participation might be higher because some teams submitted runs but did not submit their working notes, thus not being counted in the statistics.

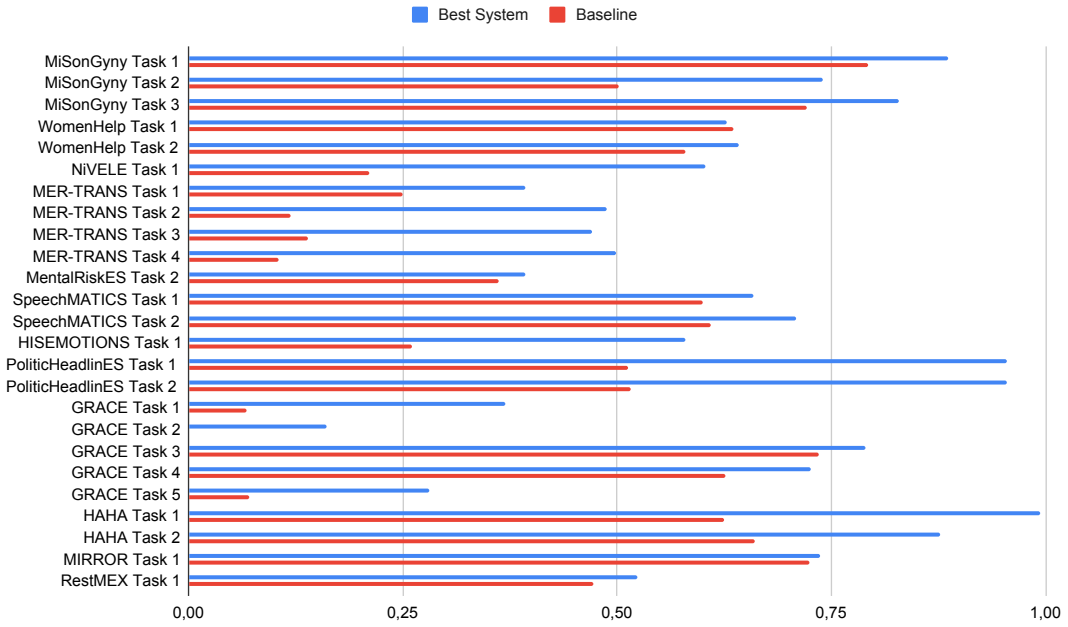

Figure 5: Performance of best systems versus baselines in IberLEF 2026 tasks. Only tasks with official evaluation metrics in the range [0-1] that include at least a baseline system are included in this graph.

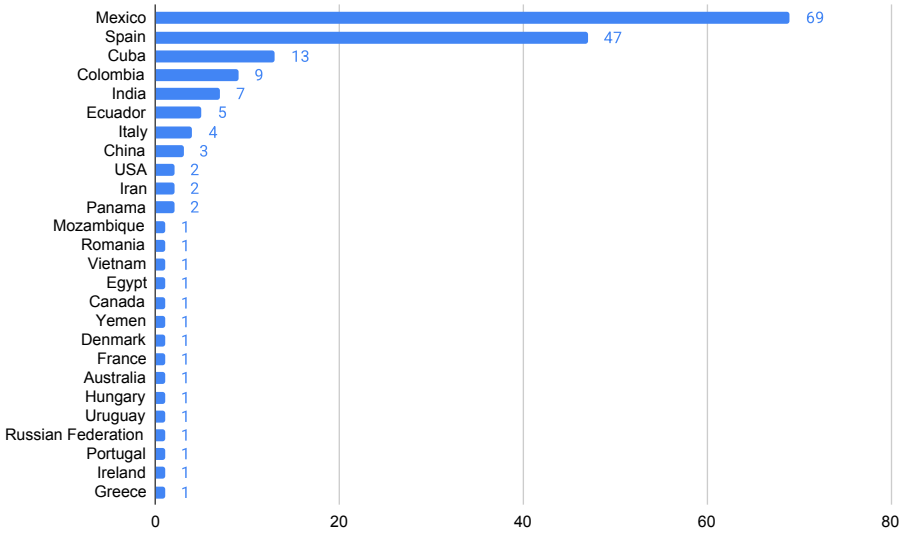

Figure 6: Number of research groups participating in IberLEF 2026 tasks per country.

and research groups. While Mexico has the largest number of participating research groups, Spain contributes the highest number of individual researchers to IberLEF 2026. The presence of non-Spanish-speaking countries such as India, China, Vietnam, Italy or Romania in the top ten highlights two key points: *i)* Spanish captures interest in the broader NLP community; and *ii)* current NLP technologies allow researchers to work with different languages without needing language-specific tools, beyond pre-trained language models available to the research community.

Figure 8 shows the number of teams participating in each of the tasks, considering that they submitted at least one run. Participation ranges between 3 and 30 teams. The tasks with the highest participation are SSD-2026 and Rest-MEX, highlighting the interest in developing approaches for social support detection in social media and generation of synthetic data to train sentiment analysis models.

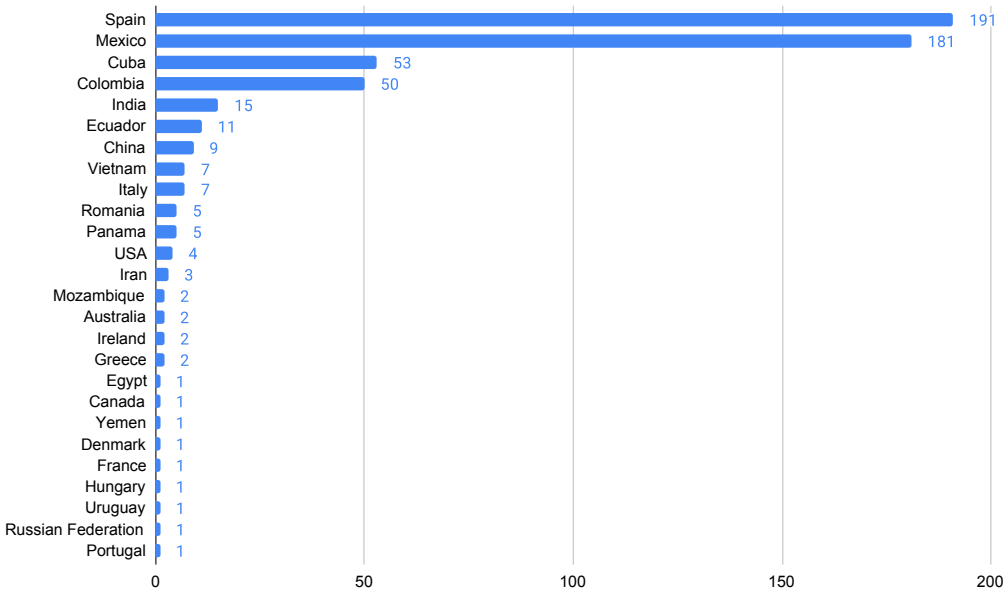

Figure 7: Number of researchers participating in IberLEF 2026 tasks per country.

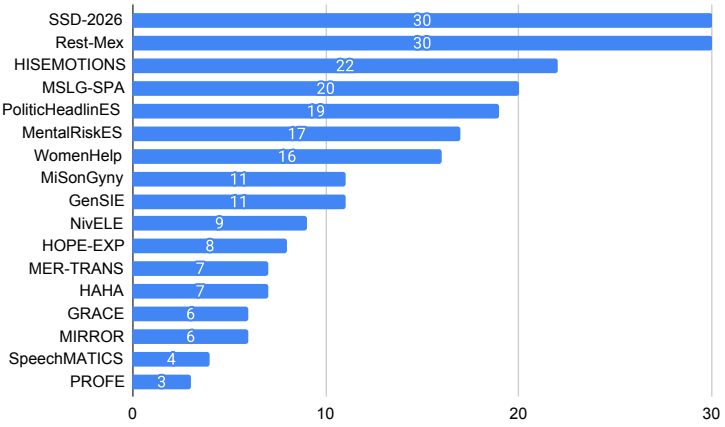

Figure 8: Distribution of participating teams per task in IberLEF 2026. The figure displays the number of teams that submitted at least one run.

The distribution of research groups per task is shown in Figure 9. In this case, participation ranges between 5 and 35 groups[4]. Notably, REST-MEX accounts for the largest share of research group participation (14% of the total), and the top three tasks (REST-MEX, MSLG-SPA, and PoliticHeadlinES) together involve 35% of all research groups.

Notably, participating research groups showed greater interest in tasks involving synthetic data generation (RestMEX), sign language–to–natural language translation (MSLG-SPA), and ranking (PoliticHeadlinES) than in the traditional text classification tasks that dominated previous editions. This trend is particularly noteworthy because these tasks typically present a higher barrier to entry due to their increased complexity. Despite this, they have attracted a larger number of research groups, encouraging the development of more sophisticated methods and innovative algorithmic solutions.

Finally, we tried to analyze the evolution of participation in IberLEF throughout the years. As

[4]A team is composed of researchers from the same or different research groups and entities who collaborate to participate in a shared task. In contrast, a research group typically consists of researchers from the same faculty who specialize in a particular subject and work together officially on that topic, not solely for participating in a shared task.

researchers might be part of different teams, participate in different tasks, and present different system description papers, we decided to measure the number of unique participants as the number of unique authors that took part in at least one system description paper in the proceedings of each year. Please note that this estimation is not perfect, as sometimes authors appear under different names in different papers, and it is also possible that there are two or more unrelated authors that have the same name. Figure 10 presents this estimation of participation since 2019, showing that this year has been the edition with the largest number of unique participants, slightly ahead of the 2023 and 2025 editions.

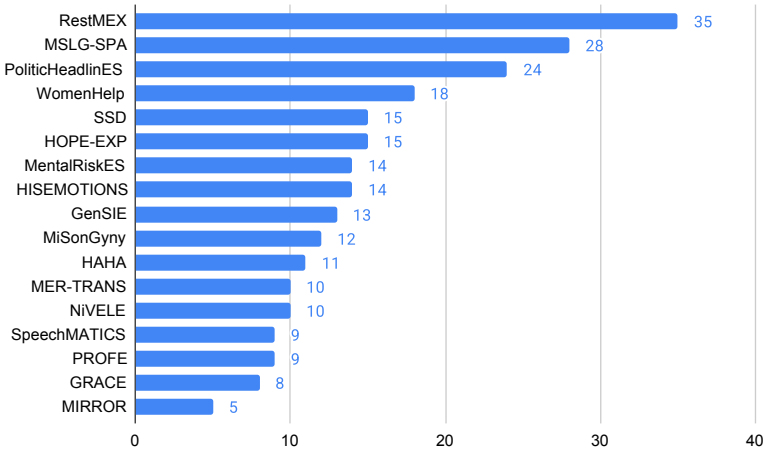

Figure 9: Distribution of participant groups per task in IberLEF 2026. The figure displays the number of groups that submitted at least one run.

4. Conclusions

In this eighth edition, IberLEF has once again demonstrated its significant collective effort to advance Natural Language Processing in Spanish and other Iberian languages. This year's event included 17 main tasks and involved 558 researchers from 178 research groups across 26 countries in Europe, Asia, Africa, and the Americas. Compared to IberLEF 2025, the number of participating researchers increased by 25%, highlighting the growing interest in NLP evaluation campaigns on Iberian languages.

IberLEF 2026 was one of the most diverse editions in terms of application domains, data sources, and task types, with growing interest in multimodal scenarios and text generation tasks where to leverage recent advances in language modeling. It advanced the field in several areas, including language comprehension, harmful and inclusive content, clinical NLP, NLP for inclusion, sentiment and figurative analysis, information extraction, and ranking.

In the realm of Natural Language Processing, where 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.

Declaration on Generative AI

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

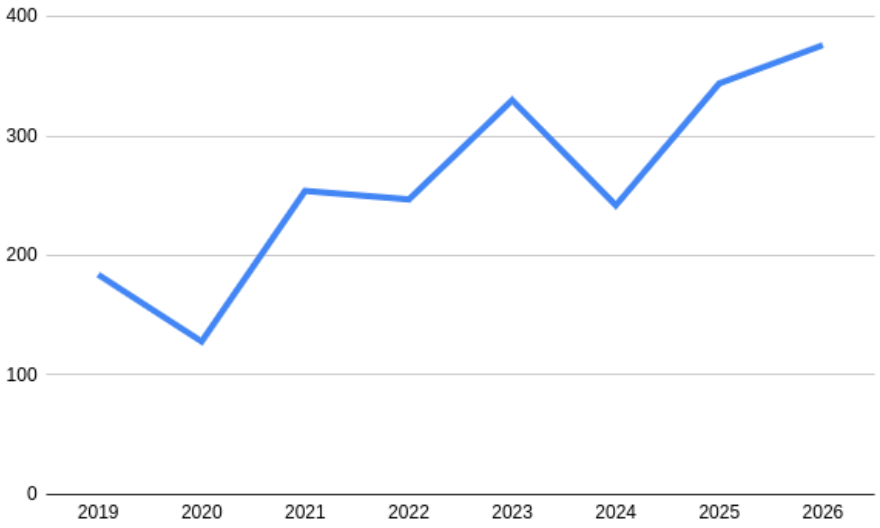

Figure 10: Number of unique participants, measured as unique author names that contributed to the system description papers published in each edition of IberLEF.

References

[1] M. V. Cantero Romero, J. Collado Montañez, I. Cabrera de Castro, J. Fruns Giménez, A. Arjonilla Sampedro, J. Cruz, A. Montejo-Ráez, S. M. Jiménez-Zafra, Resumen de NivELE en IberLEF 2026: Nivelación automática de textos de estudiantes de ELE (Español como Lengua Extranjera), Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Rodrigo, S. Moreno-Álvarez, A. Pérez García-Plaza, A. Peñas, R. Agerri, J. Fruns-Jiménez, I. Soria-Pastor, Resumen de PROFE en IberLEF 2026: Evaluación de la competencia lingüística, Procesamiento del Lenguaje Natural 77 (2026).

[3] F. Balouchzahi, S. Butt, H. Gómez-Adorno, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Resumen de HOPE-EXP en IberLEF 2026: Análisis de Expectativas Orientado a Resultados en Redes Sociales, Procesamiento del Lenguaje Natural 77 (2026).

[4] T. Alcántara, E. U. Alacreu, C. Macías, M. Aloy Mayo, O. Garcia-Vazquez, A. Espinosa-Juarez, M. Soto, P. Rosso, J.-E. Valdez-Rodriguez, H. Calvo, Resumen de MiSonGyny en IberLEF 2026: Detección de discursos misóginos en letras de canciones en español de Latinoamérica y España, Procesamiento del Lenguaje Natural 77 (2026).

[5] N. Martínez-Guevara, G. Bel-Enguix, V. Soto-Mendoza, J. Beltrán, A. Curiel, S. L. Ojeda-Trueba, M. Ramírez-González, H. Gómez-Adorno, Descripción de WomenHelp en IberLEF 2026: Clasificación de denuncias de violencia de género para ayudar a las mujeres amenazadas en el norte de México, Procesamiento del Lenguaje Natural 77 (2026).

[6] I. De la Iglesia, A. Atutxa, A. Barrena, K. Gojenola, R. Martínez, S. Montalvo, M. A. Rodríguez, S. Z. Puig, V. G. Martínez, J. Ramirez-Romero, J. M. Villa-Gonzalez, Resumen de GRACE en IberLEF 2026: Reconocimiento Granular de Evidencia Clínica Argumentativa, Procesamiento del Lenguaje Natural 77 (2026).

[7] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña López, A. Montejo-Ráez, Resumen de la Tarea MentalRiskES en IberLEF 2026: Detección Precoz del Riesgo de Problemas Emocionales en Español, Procesamiento del Lenguaje Natural 77 (2026).

[8] L. J. Arellano, C. Olachea, J. D. Piette, H. J. Escalante, L. Villaseñor Pineda, M. Montes-y Gómez, D. I. Hernández-Farías, Resumen de la Tarea MIRROR en IberLEF 2026: Evaluación de Respuestas en Entrevistas Motivacionales Mediante Turnos Conversacionales Sintéticos, Procesamiento del Lenguaje Natural 77 (2026).

[9] H. Saggion, M. Tareh, N. Khallaf, S. Bott, D. Adanza, A. Rascón Alcaina, N. Pérez Rojas, S. Szasz,

Resumen de MER-TRANS en IberLEF 2026: Primera tarea compartida multilingüe de traducción a lectura fácil, Procesamiento del Lenguaje Natural 77 (2026).
[10] A. Y. Rodríguez-González, D. Fajardo-Delgado, M. A. Álvarez Carmona, M. G. Sánchez-Cervantes, A. Díaz-Pacheco, R. Aranda, S. B. Fajardo-Flores, C. H. Martínez-Rodríguez, Resumen de la tarea MSLG-SPA en IberLEF 2026: Traducción bidireccional entre glosas de lengua de señas mexicana y español, Procesamiento del Lenguaje Natural 77 (2026).
[11] M. Shahiki Tash, Z. Ahani, L. Ramos, F. Balouchzahi, O. Kolesnikova, G. Sidorov, A. Gelbukh, R. Monroy, F. Vargas, Resumen de SSD-2026 en IberLEF 2026: Detección de Apoyo Social e Identificación de Objetivos en Redes Sociales, Procesamiento del Lenguaje Natural 77 (2026).
[12] L. Chiruzzo, I. Sastre, G. Rey, A. Martínez, A. Rosá, S. Góngora, S. Castro, V. Amoroso, J. J. Prada, J. P. Conde, G. Moncecchi, Overview de HAHA en IberLEF 2026: Ánalisis de Humor basado en Anotaciones Humanas y Generación Automática de Humor, Procesamiento del Lenguaje Natural 77 (2026).
[13] A. Sarymsakova, P. Martín Rodilla, E. Martínez Cámara, L. A. Ureña López, Resumen de HISEMOTIONS en IberLEF 2026: Detección de Emociones en Textos Históricos de la Correspondencia Española de la Edad Moderna, Procesamiento del Lenguaje Natural 77 (2026).
[14] M. A. Álvarez Carmona, A. Díaz-Pacheco, R. Aranda, M.-M. Reyes-Sierra, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Resumen de la tarea Rest-Mex en IberLEF 2026: Generación de Texto Sintético para Pueblos Mágicos de México, Procesamiento del Lenguaje Natural 77 (2026).
[15] V. Ahuir-Esteve, A. J. Barceló-Milkova, A. Casamayor-Segarra, L.-F. Hurtado-Oliver, M. J. Castro-Bleda, Resumen de SpeechMATICS en IberLEF 2026: Clasificación Multimodal Audio-Texto de Ironía en Español, Procesamiento del Lenguaje Natural 77 (2026).
[16] A. P. Morffis, Y. A. Cruz, S. E. Velarde, I. E. Zaragoza, M. M. Maestre, L. S. Requena, A. P. Montero, E. E. Valladares, Overview of GenSIE at IberLEF 2026: Schema-Guided Information Extraction with Small Language Models in Spanish, Procesamiento del Lenguaje Natural 77 (2026).
[17] J. Gómez-Navalón, T. Bernal-Beltrán, R. Pan, F. García-Sánchez, J. A. García-Díaz, R. Valencia-García, Resumen de PoliticHeadlinES en IberLEF 2026: Clasificación Multimodal de Titulares en las Noticias Políticas en Español, Procesamiento del Lenguaje Natural 77 (2026).