Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages

José Ángel González-Barba[1], Luis Chiruzzo[2] and Salud María Jiménez-Zafra[3]

[1]*TransPerfect, Spain*
[2]*Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay*
[3]*SINAI, Computer Science Department, CEATIC, Universidad de Jaén, Jaén, Spain*

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 2025, which featured 14 tasks and 33 subtasks. These tasks covered various areas such as language comprehension, harmful and inclusive content, content curation and generation, and sentiment and figurative analysis. Overall, the IberLEF 2025 activities represented a significant collaborative effort, involving more than 440 researchers from 21 countries across Europe, Asia, Africa, and the Americas.

Keywords

Natural Language Processing, Artificial Intelligence, Evaluation, Evaluation Challenges

1. Introduction

IberLEF 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 conference of the Spanish Society for Natural Language Processing. 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. 75 (September 2025 issue)*. Finally, the task organizers report the results of the tasks, and selected participants present descriptions of their systems at the IberLEF workshop.

IberLEF 2025 takes place on September 23, 2025, in Zaragoza (Aragón, Spain), as part of the XLI International Conference of the Spanish Society for Natural Language Processing (SEPLN 2025). This

IberLEF 2025 September 2025, Zaragoza, Spain
jgonzalez@transperfect.com (J. González-Barba); luis.chiruzzo@gmail.com (L. Chiruzzo); sjzafra@ujaen.es (S. M. Jiménez-Zafra)
https://jogonba2.github.io/ (J. González-Barba); https://www.fing.edu.uy/index.php/es/node/40865 (L. Chiruzzo); https://www.ujaen.es/departamentos/dinformatica/contactos/jimenez-zafra-salud-maria (S. M. Jiménez-Zafra)
0000-0003-3812-5792 (J. González-Barba); 0000-0002-1697-4614 (L. Chiruzzo); 0000-0003-3274-8825 (S. M. Jiménez-Zafra)

year, 14 shared tasks were accepted to be organized as part of IberLEF 2025, out of 18 proposals. These tasks focus on a range of NLP challenges, including language comprehension, harmful and inclusive content detection, content curation and generation, and sentiment and figurative analysis.

In this paper, we provide a summary and analysis of the tasks organized in IberLEF 2025 to offer a clearer understanding of this collective effort.

2. IberLEF 2025 Tasks

The 14 tasks involved in IberLEF 2025 are presented below, grouped by theme.

2.1. Language Comprehension

ADoBo 2025 [1], Automatic Detection of Borrowings, addressed the *automatic detection of anglicisms (English lexical borrowings) in Spanish journalistic texts*. This shared task was previously organized in 2021 [2], but with a specific focus on the automatic detection of unassimilated borrowings in the Spanish press. In this edition, 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) [3]. 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 [4], Challenge for Plain Language and Easy-to-Read Adaptation for Spanish texts, 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 [5], Language Proficiency Evaluation, 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.

2.2. Harmful and Inclusive Content

DIMEMEX [6], Detection of Inappropriate Memes from Mexico, is the second edition of DIMEMEX at IberLEF, 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 determine 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 [7], Human-centric polarity detection in Online Messages Oriented to the Latin American-speaking LGBTQ+ population, continues the HOMO-MEX shared tasks from 2023 [8] and 2024 [9], 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 generalization* 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 [10], Early detection of mental disorders risk in Spanish Third edition - Detecting Addiction, is the third edition of this task about early detection of mental risk disorders in Spanish, this time with the particular focus in detecting gambling disorders. Two subtasks were presented: *detection of gambling disorders risk*, and *classification considering different types of addiction.* The task presented a dataset of social media texts annotated with information about the risk of different types of gambling disorders (e.g. betting, online gaming, trading/crypto, lootboxes). A total of 13 teams participated in the task submitting at least one result, and the submissions were evaluated according to overall classification perfromance but also early prediction and efficiency metrics, with the aim of emphasizing the need for sustainable practices in NLP. The best Macro-F1 for task 1 was achieved by team UNSL (0.567), while for task 2 the best Macro-F1 was by team MCDI (0.589). Team PLN_PPM_ISB obtained the best results related to early prediction.

MiSonGyny [11], Misogyny Speech Detection in Spanish Language Song Lyrics, 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 [12], Optimism, Expectation or Sarcasm?, is the continuation of the HOPE tasks that had different editions, two of them in previous IberLEF workshops, and all related to hope speech detection and classification (messages that express optimism, encouragement, or the desire for a better future). In this edition, two subtasks are proposed, with variants in English and Spanish: first *binary hope speech detection*, and second *multiclass categorization* as generalized hope, realistic hope, unrealistic hope, not hope, or a novel sarcasm category meant to detect hopeful language that is used in a misleading way. They presented a dataset of 30,000 tweets labeled with hope speech data, a third of them in English and the rest in Spanish. A total of 31 teams participated in the competition, and 13 of those teams had their papers accepted. The top system for the binary subtask in Spanish was submitted by team teddymas and had 0.852 F1, the best system for English was by michaelibrahim having 0.871 F1; while in the multiclass categorization subtask the top performance for Spanish was 0.742 macro-F1 by lephuquy, and 0.755 for English by supachoke. Some challenges were mentioned by many of the teams, including data imbalance, language mixing, and cultural differences in how people express emotions.

2.3. Content Curation and Generation

PastReader [13], Transcribing Texts from the Past, focuses on the automatic transcription of digitized Spanish historical newspapers. This 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 [14], *Preguntas y Respuestas sobre Tablas en Español* - Questions and Answers about Tables in Spanish, is a question-answering task focusing on answering questions about tabular data in Spanish. Participants were given natural language questions that needed to be interpreted to extract data from tabular sources. The dataset was a collection of 10 different sources totalling 31 thousand data rows, and 300 question-answer pairs over that data were provided, with different expected answer types: boolean, categorical, numeric, or list (either of categories or numbers). The data was split in 200

question-answer pairs for training, and 100 for test, and overall accuracy across all categories was the main metric of the task. The were 7 participant systems, all of them obtaining better results than the GPT-4o baseline. Overall, the ITU NLP and sonrobok4 teams obtained the best performances (87% accuracy), although with slightly different results for the types boolean and list. Both systems use code generation methods with different LLMs and different prompting strategies. One of the conclusions is that although current LLM technologies outperform traditional pipelines, using the largest models is not the only way to go, as small open-source models models when properly used can give good results as well.

TA1C [15], *Te Ahorré Un Click*, focuses on the detection and spoiling of clickbait in Spanish news, particularly in tweets that link to a piece of news. The task consists of two subtasks: *i) Clickbait Detection*, 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`.

One additional task had originally been accepted for IberLEF 2025: **MIMIC**, Multi-Modal AI Content Detection. The goal of this task was to determine whether (image, text) pair, consisting of images, captions, and contexts from English and Spanish Wikipedia, were fully or partially generated by AI. However, due to funding limitations, it was not possible to generate the planned number of instances using Large Multimodal Models to build a high-quality dataset. As a result, the organizers decided to cancel the task.

2.4. Sentiment and Figurative Analysis

ASQP-PT [16], Aspect Sentiment Quad Prediction in Portuguese, is a shared task about aspect based sentiment analysis in Portuguese. It is a continuation of previous IberLEF tasks ABSAPT 2022 and 2024, and this year it consisted of four subtasks: *aspect term extraction*, *opinion term extraction*, *aspect category detection*, and *aspect sentiment quadruple prediction* (ASQP). A corpus of 1236 Portuguese Trip Advisor reviews about hotels in four cities was presented to the participants, with 5749 annotations of *(Category, Aspect, Opinion, Polarity)* quadruples. Out of the two teams that participated in the task, only one attempted solving all four subtasks, while the other only submitted results for the aspect term extraction subtask. The baselines for the first three subtasks were not beaten by any team, but ABCD team outperformed the baseline for the Aspect-Sentiment Quad Prediction subtask, the most complete of the subtasks, obtaining an F1 of 0.46.

REST-MEX 2025 [17], Researching on Evaluating Sentiment and Textual instances selection for Mexican magical towns, 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 [18], *Multimodal Audio-Text Satire Classification in Spanish*, was a novel task proposed for the first time this year, about the automatic recognition of satire in Spanish YouTube videos. The task was divided in two subtasks: one of them only considering the video transcription (text-only), while the other one considered the transcription and the acoustic information in a multimodal setting (text+audio). They presented a dataset of 8000 audio segments no longer than 25 seconds each, together with their transcription. The fragments were obtained from popular satirical shows in YouTube. Each segment was labeled as satirical or non-satirical. Eleven teams participated in both subtasks of the challenge, and almost all of them outperformed the baselines. The best system for the text-only task was from UPV-ELiRF, obtaining 85.6 F1; and the best system for the multimodal text+audio task was from UMU-Ev obtaining 88.3 F1.

3. Aggregated Analysis of IberLEF 2025 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 with 13 tasks, followed by English with 2 tasks, and Portuguese with 1 task. 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 Argentina, Chile, Colombia, Dominican Republic, Ecuador, El Salvador, Guatemala, Peru, Uruguay, and Venezuela.

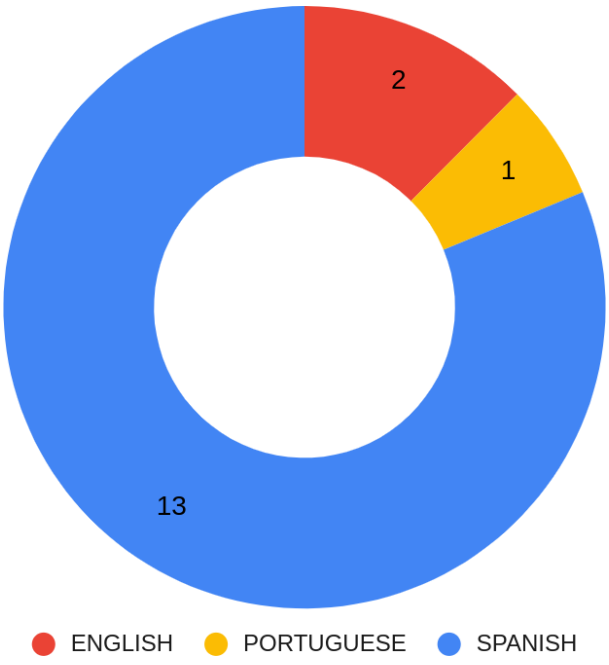

Figure 1: Distribution of languages in IberLEF 2025 tasks.

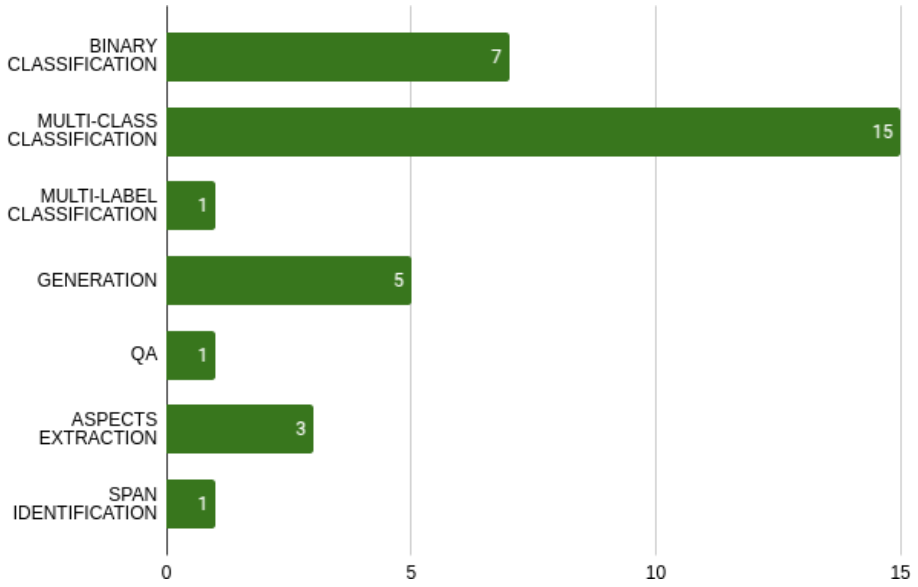

Figure 2: Distribution of IberLEF 2025 tasks per abstract task type.

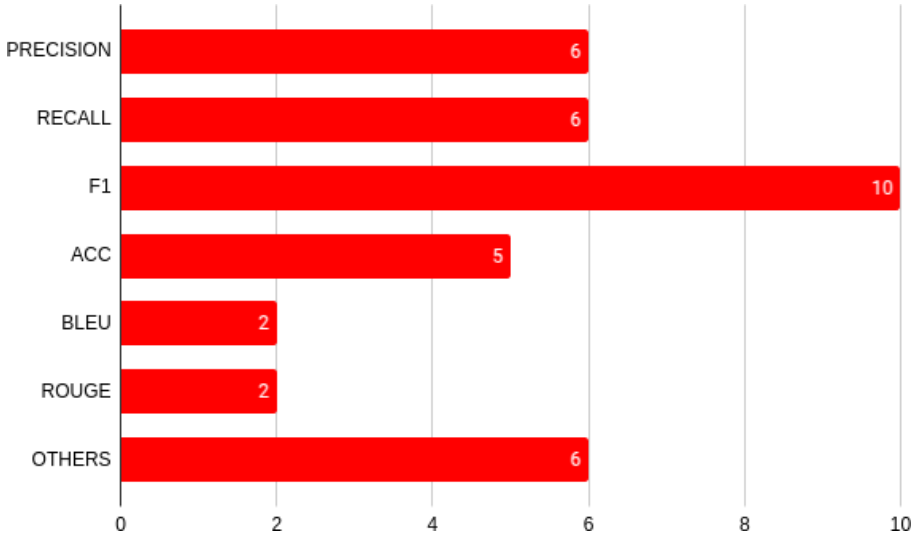

Figure 3: Distribution of official evaluation metrics in IberLEF 2025 tasks.

If we take a look at the distribution of subtasks by **abstract task types**, as shown in Figure 2, we can see that the most common task type is once again multi-class classification with 15 subtasks, followed by binary classification with 7 tasks. Interestingly this year the third most common type of subtasks involve generation (5 tasks plus one QA task, which is also related to generation). 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 are shown in Figure 3. As in previous years, F1 remains predominant, being used in 10 tasks, with 6 of them also incorporating Precision and Recall. Many tasks also include the Accuracy metric, in two cases being the only metric because of the nature of the task. Also note that metrics that correspond mainly to generative tasks, such as BLEU [19] and ROUGE [20], are starting to appear in different tasks as well, together with other metrics like BERT score [21], Levenshtein Distance, Cosine Similarity, and other metrics that are only suitable to some kinds of tasks.

IberLEF 2025 incorporated eight tasks that were not organized before or were considered novel enough to be counted as new (57%), 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 DIMEMEX, HOPE, MentalRiskES, 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 social media platforms and news outlets being the most common. Notably, in this edition, news articles have become dominant, whereas in previous editions Twitter/X was the leading source. New additions this year include advertisements, emails, Genius, social surveys, and Twitch, while some prominent sources from last year, such as PubMed, how-to articles, and Wikipedia, have been not used.

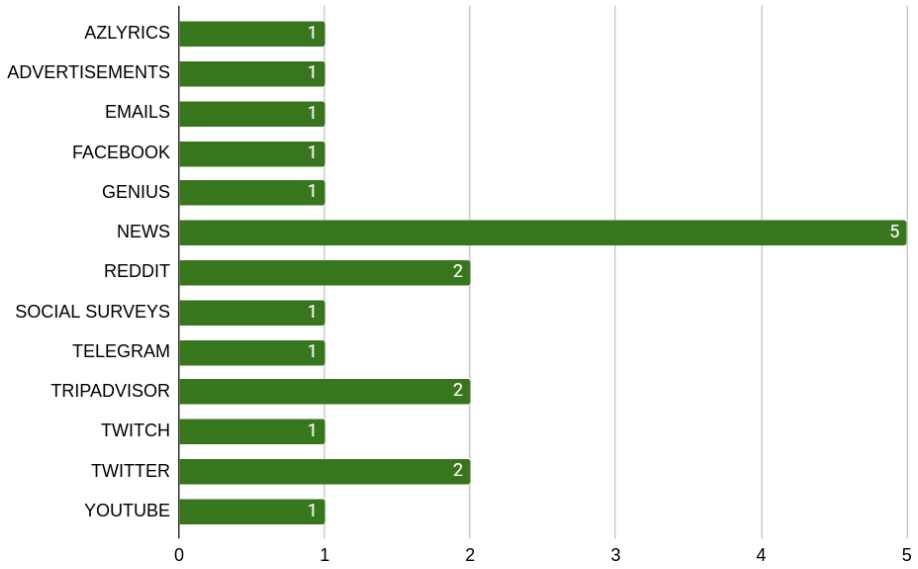

Figure 4: Types of textual sources in IberLEF 2025 tasks.

In terms of **dataset sizes** and annotation efforts[1], making fair comparisons is challenging due to the diversity of data sources, variations in text lengths, and the wide range of annotation difficulties. In almost all the cases (13 out of 14 tasks), the datasets were fully manually annotated. Only one dataset combined synthetic and manually annotated data (PolyHope). Three datasets contain more than 15,000 instances (HOMO-LAT, PolyHope, and REST-MEX), one has between 10,000 and 15,000 instances (PastReader), two fall between 5,000 and 10,000 instances (SatiSpeech and ASQP-PT), and the remaining datasets contain 1,000 to 5,000 samples, most of which are manually annotated. Regarding annotation reliability, inter-annotator agreement serves as a useful indicator and is reported for 6 out of 14 tasks. Fleiss' Kappa and Cohen's Kappa are the only statistical measures used to assess agreement reliability, each appearing in 50% of the cases. Among these, one task show high agreement, three has moderate-high agreement, and two show from low to moderate-low 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 results was submitted, using the official ranking metric

[1]Overall, the annotation efforts in IberLEF 2025 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.

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

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). One task (CLEARS) did not provide any baseline, and almost every other task included baselines based in pre-trained Transformers, either encoder (29%), encoder-decoder (29%), and decoder-only (14%) models depending on the nature of the task. Majority baselines (7%) and classical machine learning including SVM, Random Forest, and Logistic Regression were also used in some tasks (29%). In the subtasks that included baselines and had at least one submission, the best system outperformed the baseline by more than 5% in 18 cases (60% of the total), while the systems could not beat the baseline in only 4 cases. Examining the results, only 5 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.

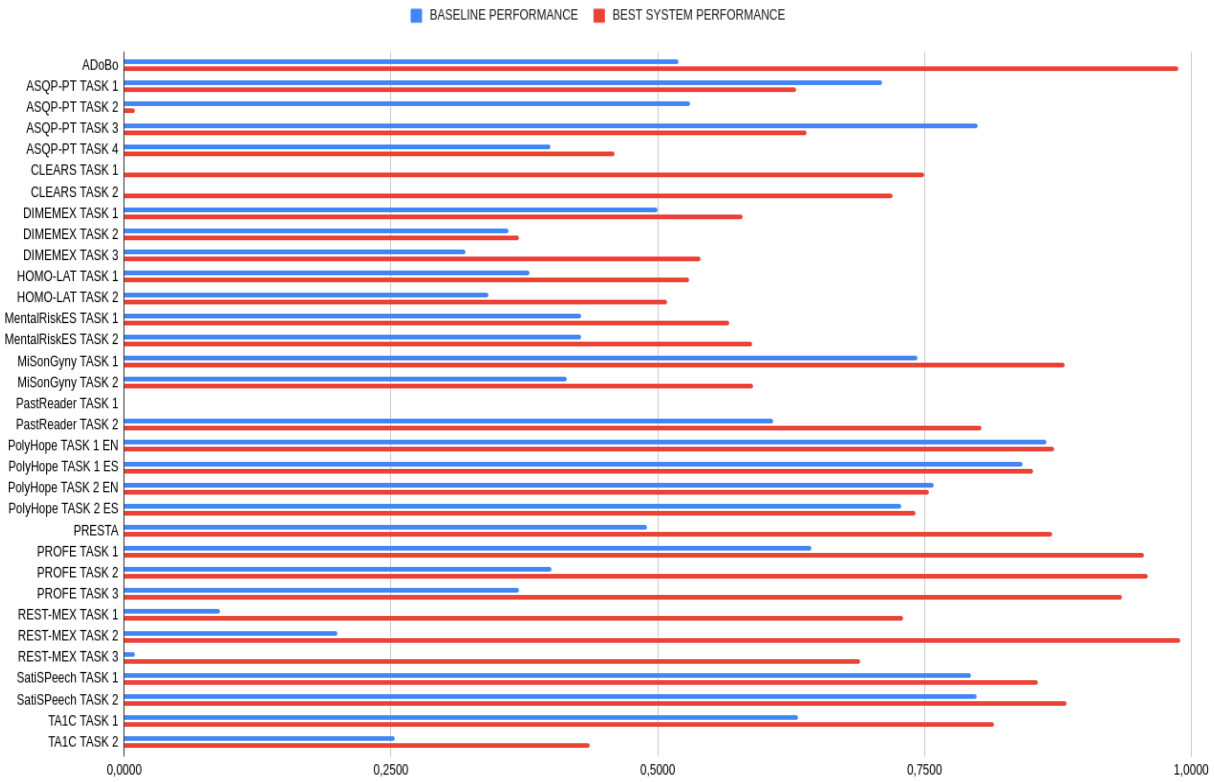

Figure 5: Performance of best systems versus baselines in IberLEF 2025 tasks. Only tasks with official evaluation metrics in the range [0-1] that include at least a baseline system are included in this graph.

3.3. Participation

Despite IberLEF 2025 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, 445 researchers from 196 research groups across 21 countries in Europe, Asia, Africa, and the Americas were involved in IberLEF tasks[3]. Compared to IberLEF 2024, the number of participating researchers increased by 54% and the number of research groups by 46%, highlighting the growing interest in NLP evaluation campaigns on Iberian languages.

Figure 6 shows the distribution of research groups by country. Interestingly, Mexican research groups now have the highest participation, with 71 groups, a shift from previous years when Spanish institutions were dominant. Mexico is followed by Spain with 62 groups, Cuba with 11, the USA with 9, and Vietnam with 7.

[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.

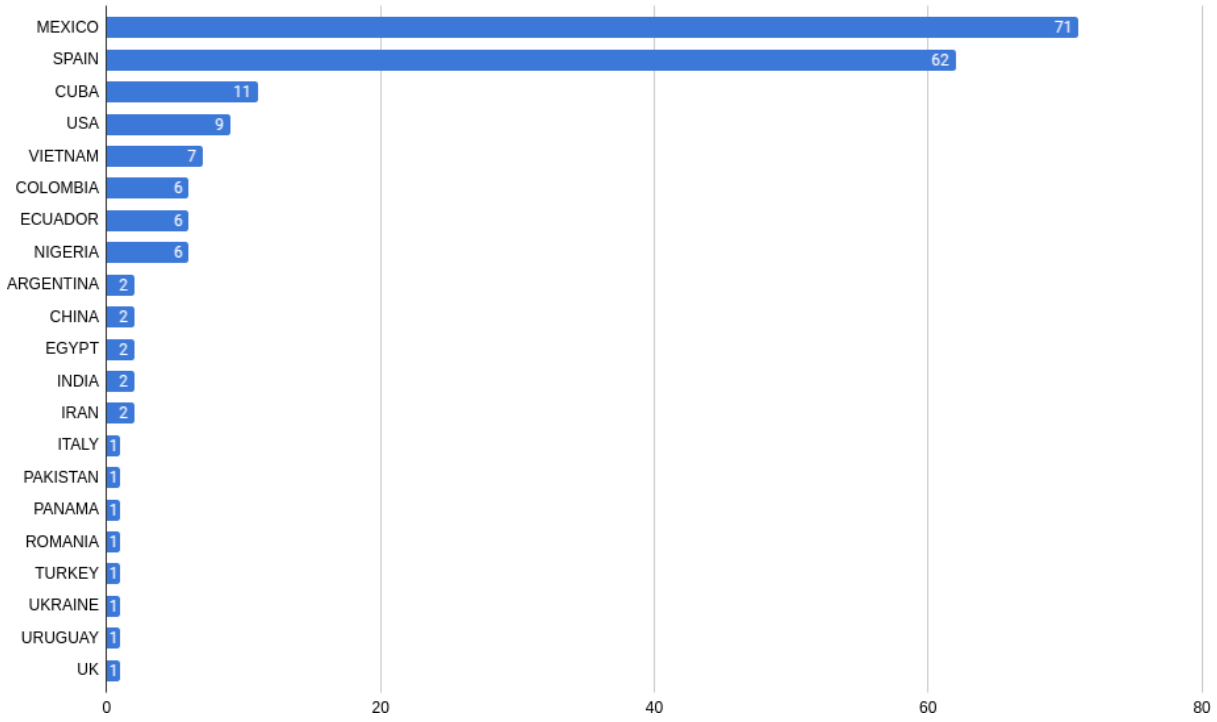

Figure 6: Number of research groups participating in IberLEF 2025 tasks per country.

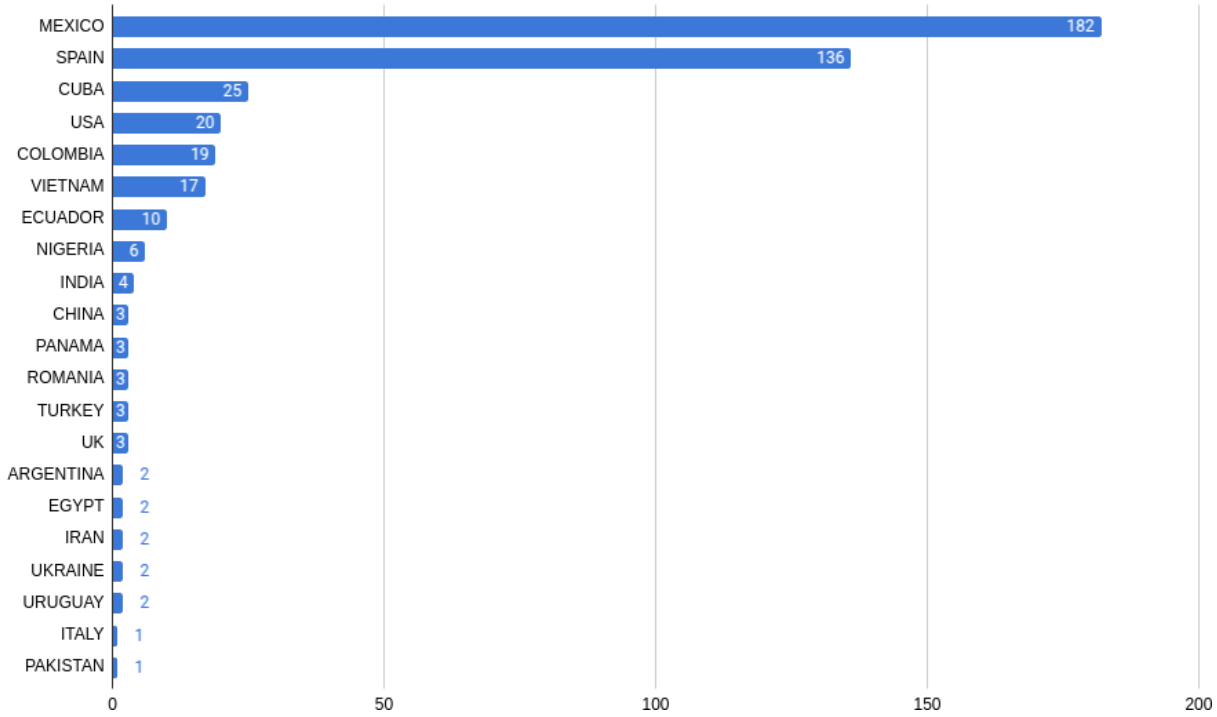

Figure 7: Number of researchers participating in IberLEF 2025 tasks per country.

Figure 7 illustrates the distribution of researchers (listed as authors in the working notes) by country. The top five countries -—Mexico, Spain, Cuba, USA, and Colombia—- account for approximately 85% of the participating researchers. Similarly to the institution-level trends, there has been a shift compared to the previous year, with Mexican researchers now dominating participation. The presence of non-Spanish-speaking countries such as Vietnam, Nigeria, India, and China in the top ten highlights two key points: *i)* Spanish captures interest in the broader NLP community; and *ii)* current NLP technologies

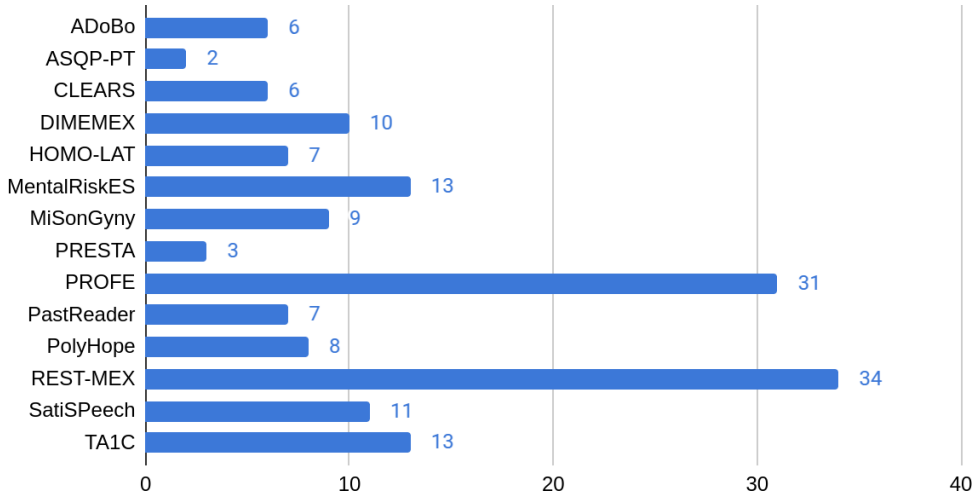

Figure 8: Distribution of participating teams per task in IberLEF 2025. The figure displays the number of teams that submitted at least one run.

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 2 and 34 teams. Notably, the task with the highest participation is REST-MEX, which may have contributed to the increase in Mexican participation, highlighting the interest in sentiment analysis for this Spanish variety. The distribution of research groups per task is shown in Figure 9. In this case, participation ranges between 2 and 52 groups[4]. Notably, REST-MEX accounts for the largest share of research group participation (27% of the total), and the top three tasks (REST-MEX, PolyHope, and SatiSpeech) together involve 50% of all research groups.

As with other evaluation initiatives, participation appears to be influenced not only by the intrinsic interest of the task but also by the cost of entry. Classification tasks, which are the simplest machine learning tasks and have more available plug-and-play software packages, typically attract more participants than tasks that require more complex approaches and creative algorithmic solutions.

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

4. Conclusions

In this seventh 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 14 main tasks and involved 445 researchers from 196 research groups across 21 countries in Europe, Asia, Africa, and the Americas. Compared to IberLEF 2024, the number of participating researchers increased by 54% and the number of research groups by 46%, highlighting the growing interest in NLP evaluation campaigns on Iberian languages.

IberLEF 2025 was one of the most diverse editions in terms of application domains, data sources, and

[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.

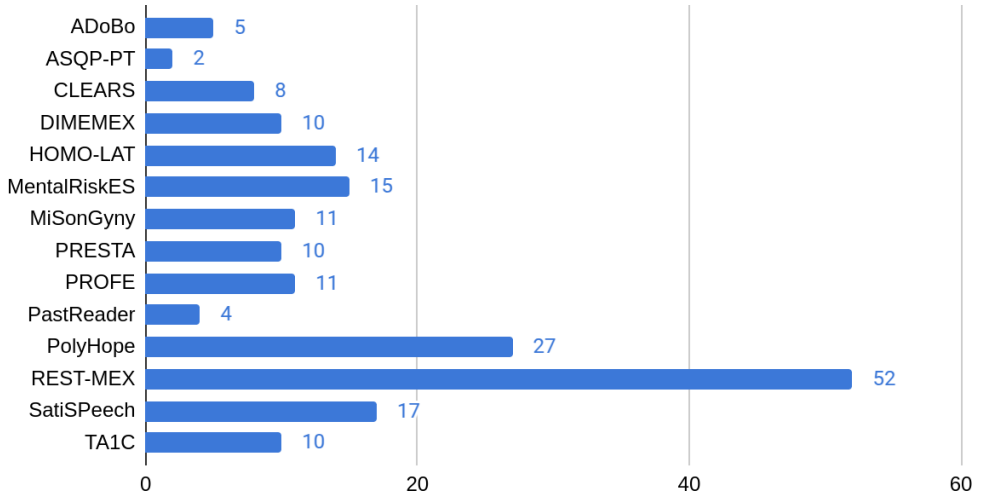

Figure 9: Distribution of participant groups per task in IberLEF 2025. The figure displays the number of groups that submitted at least one run.

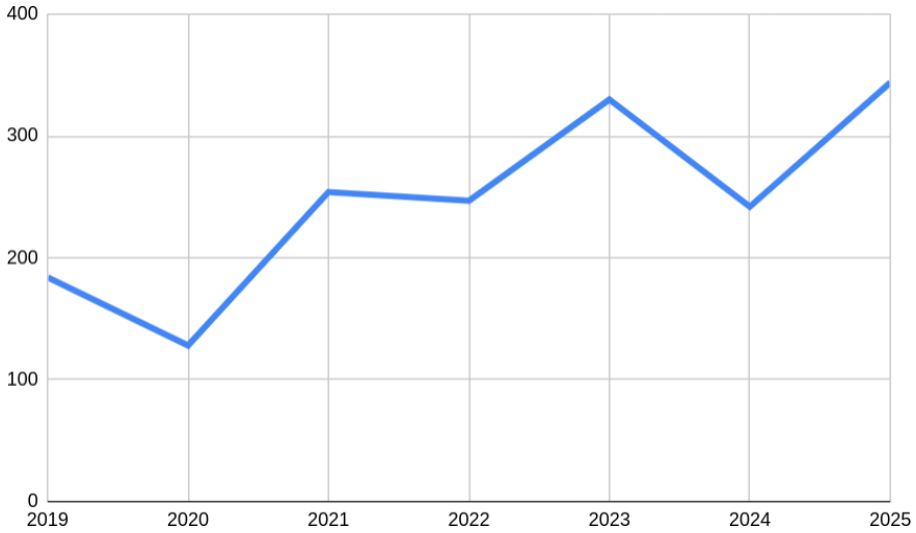

Figure 10: Number of unique participants, measured as unique author names that contributed to the system description papers published in each edition of IberLEF.

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, content curation and generation, and sentiment and figurative analysis.

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.

Acknowledgments

The research work conducted by Salud María Jiménez-Zafra is part of the grant RYC2023-044481-I, supported by MICIU/AEI/10.13039/501100011033 and by ESF+. This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21),

Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and Project FedDAP (PID2020-116118GA-I00) and Project Trust-ReDaS (PID2020-119478GB-I00) supported by MICINN/AEI/10.13039/501100011033.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. Álvarez-Mellado, J. Porta-Zamorano, C. Lignos, J. Gonzalo, Overview of ADoBo at IberLEF 2025: Automatic Detection of Anglicisms in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[2] E. Á. Mellado, L. E. Anke, J. G. Arroyo, C. Lignos, J. P. Zamorano, Overview of adobo 2021: Automatic detection of unassimilated borrowings in the spanish press, Procesamiento del Lenguaje Natural 67 (2021) 277–285.

[3] E. Álvarez Mellado, Lexical borrowing detection as a sequence labeling task: Data, modeling and evaluation methods for anglicism retrieval in Spanish, Phd thesis, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, 2025.

[4] B. Botella-Gil, I. Espinosa-Zaragoza, A. Bonet-Jover, M. Madina, L. Molino Piñar, P. Moreda, I. Gonzalez-Dios, M. T. Martín-Valdivia, L. A. Ureña-López, Overview of CLEARS at IberLEF 2025: Challenge for Plain Language and Easy-to-Read Adaptation for Spanish Texts, Procesamiento del Lenguaje Natural 75 (2025).

[5] A. Rodrigo, S. Moreno-Álvarez, A. Pérez, A. Peñas, R. Agerri, J. Fruns-Jiménez, I. Soria-Pastor, Overview of PROFE at IberLEF 2025: Language Proficiency Evaluation, Procesamiento del Lenguaje Natural 75 (2025).

[6] H. Jarquín-Vásquez, I. Tlelo-Coyotecatl, D. I. Hernández-Farías, H. J. Escalante, L. Villaseñor-Pineda, M. M. y Gómez, Overview of DIMEMEX at IberLEF 2025: Detection of Inappropriate Memes from Mexico, Procesamiento del Lenguaje Natural 75 (2025).

[7] G. Bel-Enguix, H. Gómez-Adorno, S. Ojeda-Trueba, G. Sierra, J. Barco, E. Lee-Romero, J. Dunstan, R. Manrique, Overview of HOMO-LAT at IberLEF 2025: Human-centric polarity detection in Online Messages Oriented to the Latin American-speaking LGBTQ+ populaTion, Procesamiento del Lenguaje Natural 75 (2025).

[8] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population, Procesamiento del lenguaje natural 71 (2023) 361–370.

[9] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.

[10] A. M. Mármol-Romero, P. Álvarez Ojeda, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2025: Early Detection of Addiction Risk in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[11] T. Alcántara, M. Soto, C. Macias, O. Garcia-Vazquez, A. Espinosa-Juarez, H. Calvo, J. E. Valdez-Rodríguez, E. Felipe-Riveron, Overview of MiSonGyny at IberLEF 2025: Misogyny Speech Detection in Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural 75 (2025).

[12] S. Butt, F. Balouchzahi, M. Amjad, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Overview

of PolyHope at IberLEF 2025: Optimism, Expectation or Sarcasm?, Procesamiento del Lenguaje Natural 75 (2025).
[13] A. Montejo-Ráez, E. Sánchez-Nogales, G. Expósito-Álvarez, L. A. Ureña-López, M. T. Martín-Valdivia, J. Collado-Montañez, M. C. Díaz-Galiano, I. C. de Castro, M. V. Cantero-Romero, R. Ortuño-Casanova, Overview of PastReader at IberLEF 2025: Transcribing Texts From the Past, Procesamiento del Lenguaje Natural 75 (2025).
[14] J. Osés Grijalba, L. A. Ureña-López, E. Martínez Cámara, J. Camacho-Collados, Overview of PRESTA at IberLEF 2025: Question Answering Over Tabular Data In Spanish, Procesamiento del Lenguaje Natural 75 (2025).
[15] G. Mordecki, L. Chiruzzo, R. Laguna, J. 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).
[16] E. P. Lopes, G. A. Gomes, A. Thurow Bender, R. M. Araujo, L. A. de Freitas, U. B. Corrêa, Overview of ASQP-PT at IberLEF 2025: Overview of the Task on Aspect-Sentiment Quadruple Prediction in Portuguese, Procesamiento del Lenguaje Natural 75 (2025).
[17] M. A. Álvarez Carmona, A. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of Rest-Mex at IberLEF 2025: Researching Sentiment Evaluation in Text for Mexican Magical Towns, Procesamiento del Lenguaje Natural 75 (2025).
[18] R. Pan, J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, R. Valencia-García, Overview of SatiSPeech at IberLEF 2025: Multimodal Audio-Text Satire Classification in Spanish, Procesamiento del Lenguaje Natural 75 (2025).
[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[20] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: Proceedings of 2020 International Conference on Learning Representations, 2020.