Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages

Luis Chiruzzo[1], Salud María Jiménez-Zafra[2] and Francisco Rangel[3]

[1]*Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay*
[2]*SINAI, Computer Science Department, CEATIC, Universidad de Jaén, Jaén, Spain*
[3]*Symanto Research, Valencia, 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 2024, which featured 12 tasks and 25 subtasks. These tasks covered various areas such as automatic text generation identification, biomedical Natural Language Processing, counter-speech, early risk prediction on the Internet, harmful and inclusive content detection, language reliability, political ideology and propaganda identification, and sentiment and emotion analysis. Overall, the IberLEF 2024 activities represented a significant collaborative effort, involving 289 researchers from 23 countries across Europe, Asia, Africa, Australia, and the Americas.

Keywords

Natural Language Processing, Artificial Intelligence, Evaluation, Evaluation Challenges

1. Introduction

IberLEF is a shared evaluation campaign for Natural Language Processing (NLP) 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.

In this shared evaluation campaign, the research community defines new challenges and proposes tasks to advance the state of the art in NLP. These tasks are reviewed by the steering and program committees of IberLEF and then evaluated by the IberLEF general chairs. The organizers of the accepted tasks set up the evaluation according to their proposal, promote the task, and manage the submission and scientific evaluation of system description papers submitted by participants. These papers are included in this IberLEF proceedings volume published at CEUR-WS.org. Additionally, the task organizers must prepare and submit an overview of their task evaluation exercise. These overviews are reviewed by the IberLEF organizing committee and published in the journal *Procesamiento del Lenguaje Natural, vol. 73 (September 2024 issue)*. Finally, the task organizers report the results of the tasks, and selected participants present descriptions of their systems at the IberLEF workshop.

IberLEF 2024 takes place on September 24, 2024, in Valladolid (Castilla y León, Spain), as part of the XL International Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). This year, 12 shared tasks were accepted for IberLEF 2024 out of 15 proposals. These tasks focus on a range of NLP challenges, including automatic text generation identification, biomedical NLP, counter-speech, early risk prediction on the Internet, harmful and inclusive content detection, language reliability, political ideology and propaganda identification, and sentiment and emotion analysis.

In this paper, we provide a summary and analysis of the tasks organized in IberLEF 2024 to offer a clearer understanding of this collective effort.

2. IberLEF 2024 Tasks

The 12 tasks involved in IberLEF 2024 are presented below, grouped by theme.

2.1. Automatically Generated Texts Identification

Iber AuTexTification [1] extends the previous AuTexTification shared task in three dimensions: *i)* more domains; and, *ii)* more languages from the Iberian Peninsula, adding Portuguese, Galician, Euskera, and Catalan; and *iii)* more prominent LLMs. A total of 21 teams participated in the task, sending over 68 runs. The best-performing team obtained a Macro-F of 80.50 and 69.84 respectively in Subtasks 1 and 2.

2.2. Biomedical NLP

GenoVarDis [2] deals with the problem of detecting names of genomic variants, diseases and symptoms in PubMed scientific articles in Spanish. The participant systems were expected to detect text spans containing named entities and classifying them according to eight different categories like gene, disease, and DNA mutation. 35 teams registered for the task, out of which 7 teams submitted a total of 47 systems and sent 6 working notes. The best team obtained a labeled exact match F1 of 82.10 over the test set.

2.3. Counter Speech

RefutES [3] is a task focused on the generation of counter-narratives or counterspeech in Spanish, i.e. the automatic creation o responses to offensive messages that reject the narratives behind them. The task used a Spanish translation of the CONAN-MT dataset plus a set of 78 Spanish posts with manually generated counter-narratives. The participant systems had to generate a counter-narrative that is neutral and respectful, and were evaluated according to automatic metrics, efficiency metrics, and a subset of results were evaluated manually. Six participant teams registered but only one team submitted results, obtaining a maximum of 89.23 BERTScore-F1 and 63.25 MoverScore.

2.4. Early Risk Prediction on the Internet

MentalRiskES [4] is the second edition of a novel task on early risk identification of mental disorders in Spanish comments. In the first edition [5], the task was resolved as an online problem, that is, the participants had to detect a potential risk as early as possible in a continuous stream of data. For this second edition, it were proposed three novel tasks: i) *Disorder detection*, that is, detect if a user suffers from depression or anxiety, or if there is no detected disorder at all; ii) *Context detection*, consisting of determining the context that may be associated with the disorder; and iii) *Suicidal ideation detection*, for detecting if a user is manifesting symptoms of potential suicidal ideation. As in the first edition, participants were also asked to submit measurements of carbon emissions for their systems, emphasizing the need for sustainable NLP practices. 28 teams registered for the task, 12 submitted results, and 10 presented working notes. The best-performing teams obtained Macro F1-scores of 87.4, 26.8 and 53.4 for disorder detection, context detection and suicidal ideation detection, respectively.

2.5. Harmful and Inclusive Content

DETESTS-Dis [6] is the second edition of the DETESTS task, aimed at detecting the of explicit or implicit stereotypes in social media content. In this edition, participants are given a set of disaggregated annotations so models can use this information to gauge the level of disagreement given the potential subjectivity of the task. 15 teams signed up for the task, of which six sent runs and three sent working notes papers. The best teams obtained 72.4 F1 with hard labels and 84.1 cross-entropy with soft labels for stereotype detection task, and for the implicitness detection task the best results obtained 0.065 ICM with hard labels (not beating the BETO baseline of 0.126), and -0.900 ICM with soft labels, in this case beating the baseline.

DIMEMEX [7] is a multimodal task whose purpose is to distinguish between appropriate, inappropriate content or hate speech in memes using Mexican Spanish. A dataset of 3K manually annotated was presented, and the participants could tackle the problem in two tasks: Classification in the hate speech, inappropriate, or neither categories; and finer-grained classification in categories like classism, sexism, and racism. 19 teams signed up for the competition, seven of them took part in the first task and four in the second task, submitting five working notes in total. The best results were 58 F1 score for task 1 and 44 F1 score for task 2.

HOMO-MEX [8] aims to promote the development of NLP systems for detecting and classifying LGBT+phobic content in Mexican-Spanish digital posts and song lyrics. This shared

task was previously organized [9], but in this new edition it comprises the same subtasks as last year, plus a new subtask to detect hate speech against the LGBT+ community in song lyrics written in Spanish. Specifically, it is composed of three subtasks: i) Task 1 on LGBT+phobia detection on social media posts; ii) Task 2 on fine-grained phobia identification; and iii) Task 3 on LGBT+phobia detection on song lyrics. Task 1 received 19 submissions, Task 2 attracted 10 submissions, and Task 3 got 17 submissions. The best-performing teams obtained F1-scores of 91.43, 97.30 and 57.62 for Task 1, Task 2 and Task 3, respectively.

HOPE [10] is the second edition of a previous shared task [11] related to the inclusion of vulnerable groups. The main novelty of this new edition is the study of hope from two perspectives: i) hope for equality, diversity and inclusion, and ii) hope as expectations. The first perspective was explored in the last edition of IberLEF 2023 [12] for English and Spanish, but this time participants were provided with a Spanish training corpus focused on the LGTBI community, and they had to test their systems with texts belonging to the LGTBI domain and new unknown domains. The second perspective has not been studied previously in any shared task and it were proposed its study from a binary and multi-class perspective for English and Spanish. 19 teams participated in the competition, and 16 submitted their working notes. In the first subtask, the top-ranking team achieved an average Macro F1-score of 71.61. In the second subtask, leading teams achieved F1 scores exceeding 80.00 for binary classification and 78.00 for multiclass classification settings.

2.6. Language Reliability

FLARES [13] aims to detect patterns of reliability in the language used in news that will allow the development of effective techniques for the future detection of misleading information. To this end, the 5W1H journalistic technique for detecting the relevant content of a news item is proposed as a basis, as well as an annotation guideline designed to detect linguistic reliability. Two subtasks are proposed: *i)* the identification of the 5W1H elements; and *ii)* the detection of reliability. A total of 7 teams participated in the shared task. The best-performing systems obtained 0,6613 and 0,6536 respectively in terms of the F measure.

2.7. Political Ideology and Propaganda

DIPROMATS [14] extends the previous edition of the shared task [15] by introducing a refined typology of techniques and a more balanced dataset for propaganda detection, alongside a new task focused on identifying strategic narratives. Specifically, it were proposed two tasks: i) Automatic Detection and Categorization of Propaganda Techniques; and ii) Automatic Detection of Narratives. The dataset for the first task included 12,012 annotated tweets in English and 9,501 in Spanish, posted by authorities from China, Russia, the United States, and the European Union. Participants tackled three subtasks in each language: i) binary classification to detect propagandistic tweets; ii) clustering tweets into three propaganda categories; and iii) fine-grained categorization using seven techniques. The second task presented a multi-class, multi-label classification challenge where systems identified which predefined narratives (associated with each international actor) tweets belong to. This task was supported by narrative descriptions and example tweets in English and Spanish, using few-shot learning techniques. 40

runs from 9 different teams were evaluated. The highest scores for Task 1 on the F1 metric were 81.69, 60.29 and 47.95 for subtasks 1A, 1B and 1C, respectively. For Task 2, the best-performing teams obtained an F1 Avg. of 64.11 and 61.11 for Spanish and English, respectively.

2.8. Sentiment and Emotion

ABSAPT24 [16] tackles the Aspect-Based Sentiment Analysis (ABSA) problem in Portuguese. Two subtasks are presented: *i)* Aspect Extraction, to identify the specific aspects mentioned in a text related to a given entity; and *ii)* Aspect Sentiment Classification, to determine the sentiment polarity associated with each identified aspect. Two teams submitted their results, obtaining the best one's performances of 0.6370 and 0.6530 per subtask.

EmoSPeech [17] addresses the study of Automatic Emotion Recognition (AER) via two subtasks: The first one is AER from text, which focuses on feature extraction and identifying the most representative feature of each emotion in a dataset created from real-life situations. The second task deals with AER from a multimodal perspective, which requires the construction of a more complex architecture to solve this classification problem. A total of 13 teams participated in the task, and the best-performing results in terms of F1 were 67.19 and 86.69 respectively for each subtask.

3. Aggregated Analysis of IberLEF 2024 Tasks

3.1. Tasks characterization

The distribution of **languages** per task (including subtasks) is illustrated in Figure 1. Once again, Spanish is the dominant language in IberLEF with 11 tasks, followed by English with 3 tasks, and Portuguese with 2 tasks. One task also considered Euskara, Galician and Catalan.

The distribution of subtasks by **abstract task types** is shown in Figure2. The most common task type is multi-class classification with 10 subtasks, followed by binary classification with 5 tasks. Additionally, there are two multi-label classification tasks, and one task each for NER, aspects extraction, span identification, and generation. Although the trend towards fewer but more diverse and complex tasks that began a few years ago has continued, binary classification remains one of the most popular task types again this year.

Figure 3 shows the distribution of the used **evaluation metrics**[1], highlighting only the primary metrics used for ranking systems in each task. As in previous years, F1 remains predominant, being used in 11 tasks, with six of those also incorporating Precision and Recall. Accuracy and ICM are used respectively in two tasks, while other metrics such as Cross-Entropy, Hamming Loss, Exact Match Ratio, ERDE5, ERDE30, latencyTP, Speed, Latency-weighted F1, Sentence-mover Score, or BERT score have been used in up to ten tasks.

---

[1]In IberLEF and similar NLP evaluation challenges, we often lean heavily on averages to merge different quality metrics. This year, it was typical to mix F1 scores (harmonic averages) with other metrics using various averaging techniques. Such practices obscure the true performance of systems and provide little insight into how they can be improved. Moreover, in 2024, the selection of metrics has generally lacked justification, especially concerning their relevance to practical usage scenarios.
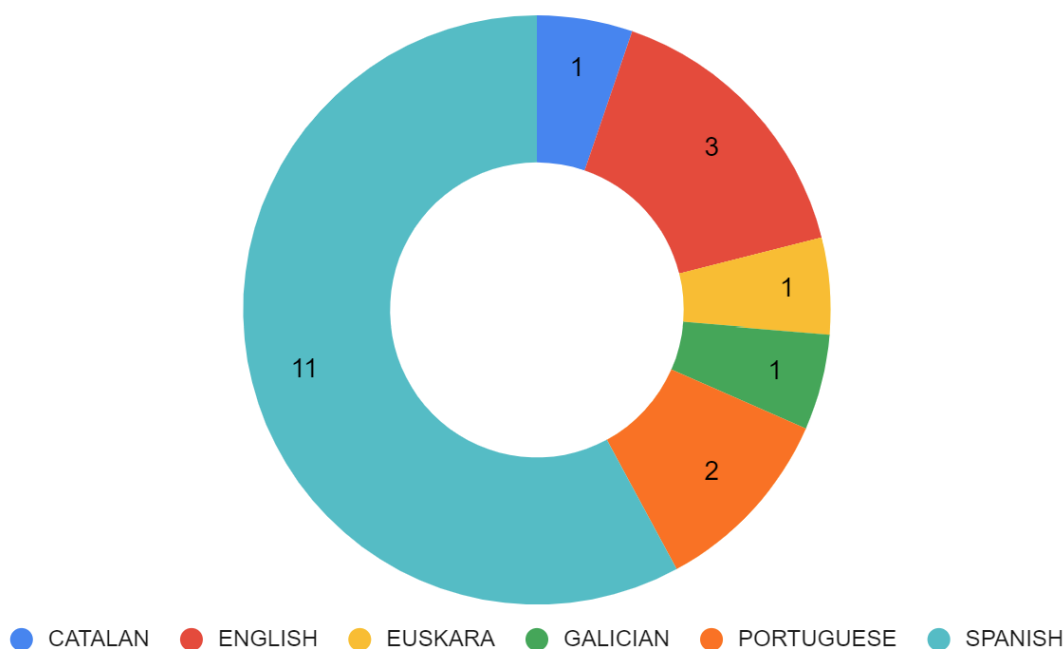
Figure 1: Distribution of languages in IberLEF 2024 tasks.

In terms of **novelty and stability**, IberLEF 2024 introduced a wealth of new challenges, with seven out of the twelve primary tasks (approximately 58%) being new this year. This influx of new problems is balanced by the continued presence of successful tasks from previous years, such as DIPROMATS, HOMO-MEX, HOPE, (Iber)AuTextification, or MentalRiskES, which provide stability and maturity to the competition.

3.2. Datasets and results

In Figure 4, statistics about the **types of data sources** can be seen. As in 2022 and 2023, there is greater diversity compared to previous years, with new sources such as Song Lyrics or MGT being included this year. However, Twitter/X remains the dominant source, used in half of the tasks. News, Reviews and Specialized Websites have been used in three and two tasks respectively, and other media such as Youtube, Telegram, Facebook, Wikis, etc. have been used in one task each.

In terms of **dataset sizes** and annotation efforts[2], making fair comparisons is challenging due to the diversity of data sources, variations in text lengths, and the wide range of annotation difficulties. In most cases (11 out of 12 tasks), datasets have been manually annotated. Of

---

[2]Overall, the annotation efforts in IberLEF 2024 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.
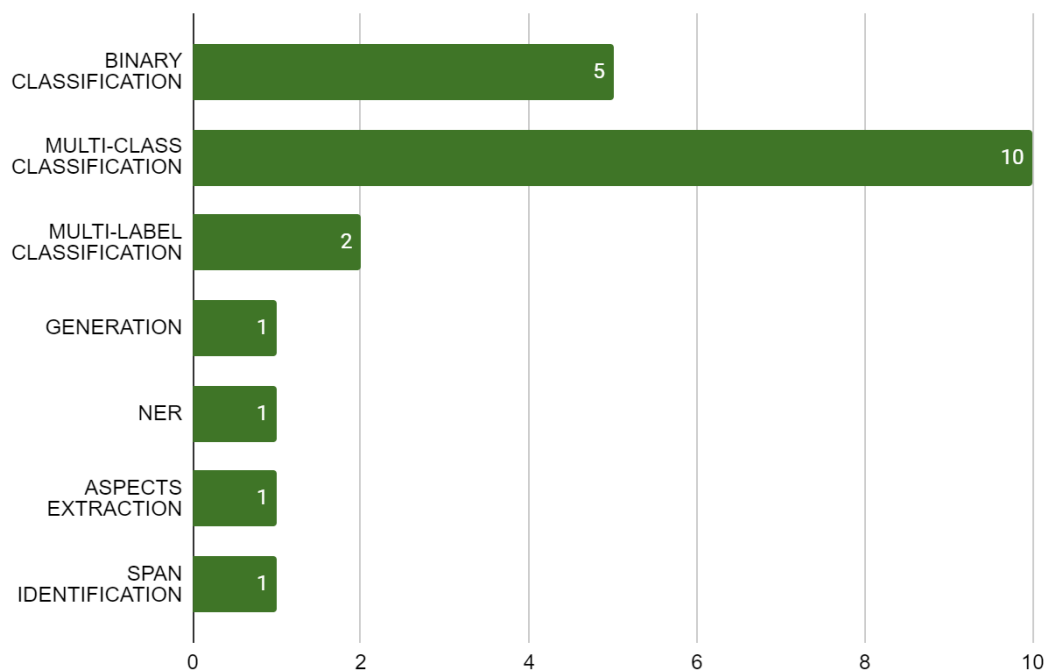
Figure 2: Distribution of IberLEF 2024 tasks per abstract task type.

these, 5 datasets contain fewer than 10,000 instances, while one have a size of approximately 35,000 instances (MentalRiskES), and 5 between 10,000 and 20,000 instances (DETEST-Dis, DIPROMATS, FLARES, HOMO-MEX, and HOPE). The AuTexTification dataset is a combination of self-annotation and human-assisted auto-generation, containing about 168,000 instances. Regarding annotation reliability, inter-annotator agreement serves as a useful indicator and is reported for 6 out of 12 tasks. Among these, two tasks show high agreement, two has moderate-high agreement, and another two show a moderate agreement[3].

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. For example, two tasks did not provide any baseline, and almost in all the rest, only a trivial baseline, such as TF-IDF, SVM, SVC and similar methods (5 subtasks), as well as standard transformers like BETO, RoBERTa and similar models (16 subtasks).

In the subtasks that included baselines, the best system outperformed the baseline by more than 5% in 12 cases, while the baseline achieved better results in 4 cases. Examining the results, only 2 subtasks had the top-performing system scoring higher than 0.9, and just 1 subtask where the baseline reached this level. This suggests that there is still room for improvement in some tasks.

Figure 5 shows a pairwise comparison between the best system and the best baseline for

---

[3]Generally, moderate agreement may reflect the complexity of the task rather than deficiencies in the annotation guidelines.
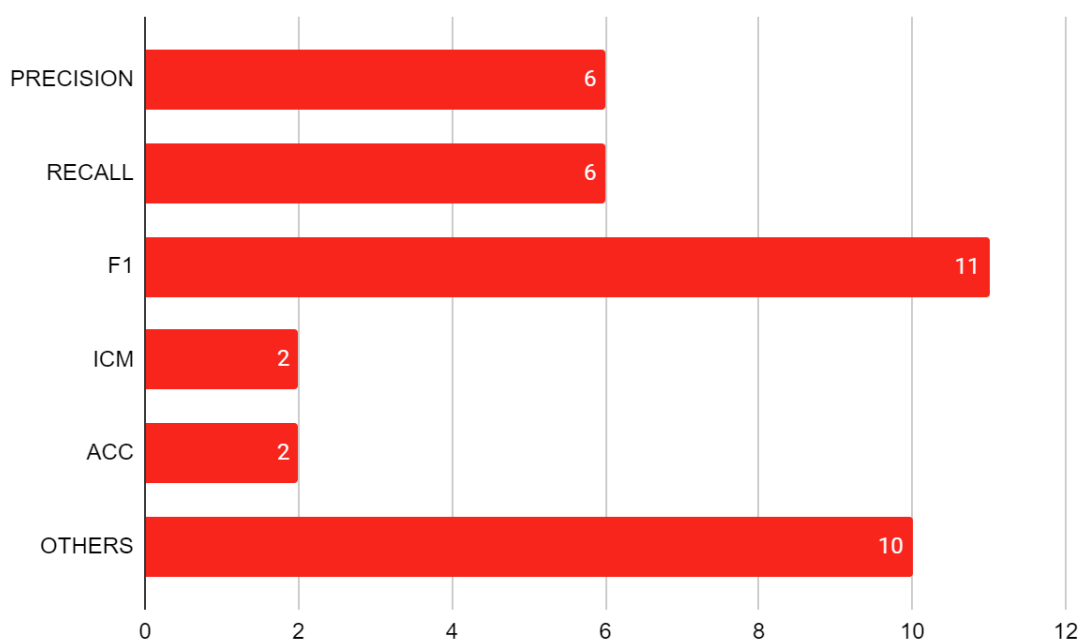
Figure 3: Distribution of official evaluation metrics in IberLEF 2024 tasks.

each task where at least one baseline is provided, using the official ranking metric 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).

3.3. Participation

Despite IberLEF 2024 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, 289 researchers from 134 research groups across 23 countries in Europe, Asia, Africa, Australia, and the Americas were involved in IberLEF tasks[4].

In Figure 6, the distribution of research groups per country is shown. This year agin, Spain has the largest representation, with 47 groups, followed by Mexico with 25 groups, Vietnam with 11, Ireland with 7, India with 6, and so on.

Figure 7 illustrates the distribution of researchers (listed as authors in the working notes) by country. The top five countries—Spain, Mexico, Vietnam, Ireland, and Colombia—account for approximately 80% of the participating researchers. The presence of non-Spanish-speaking countries such as Vietnam, Ireland, India, and Italy in the top ten highlights two key points: *i)*

---

[4]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.
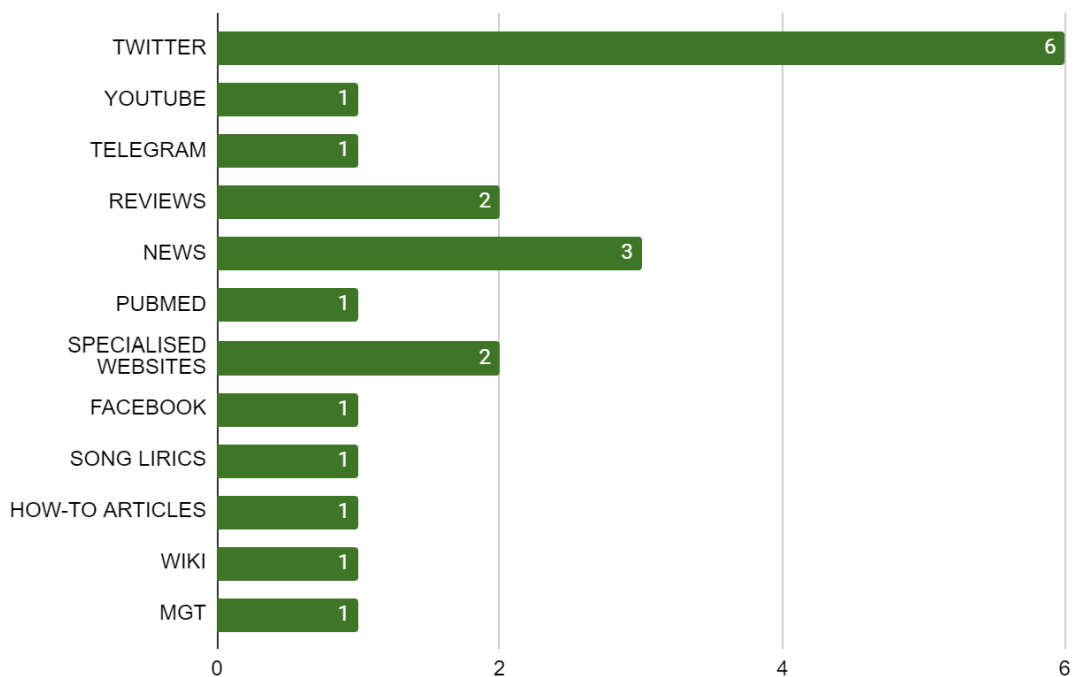
Figure 4: Types of textual sources in IberLEF 2024 tasks.

Spanish captures the interest of 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 1 and 21 teams. The distribution of research groups per task is shown in Figure 9. In this case, participation ranges between 1 and 27 groups[5].

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.

---

[5]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.
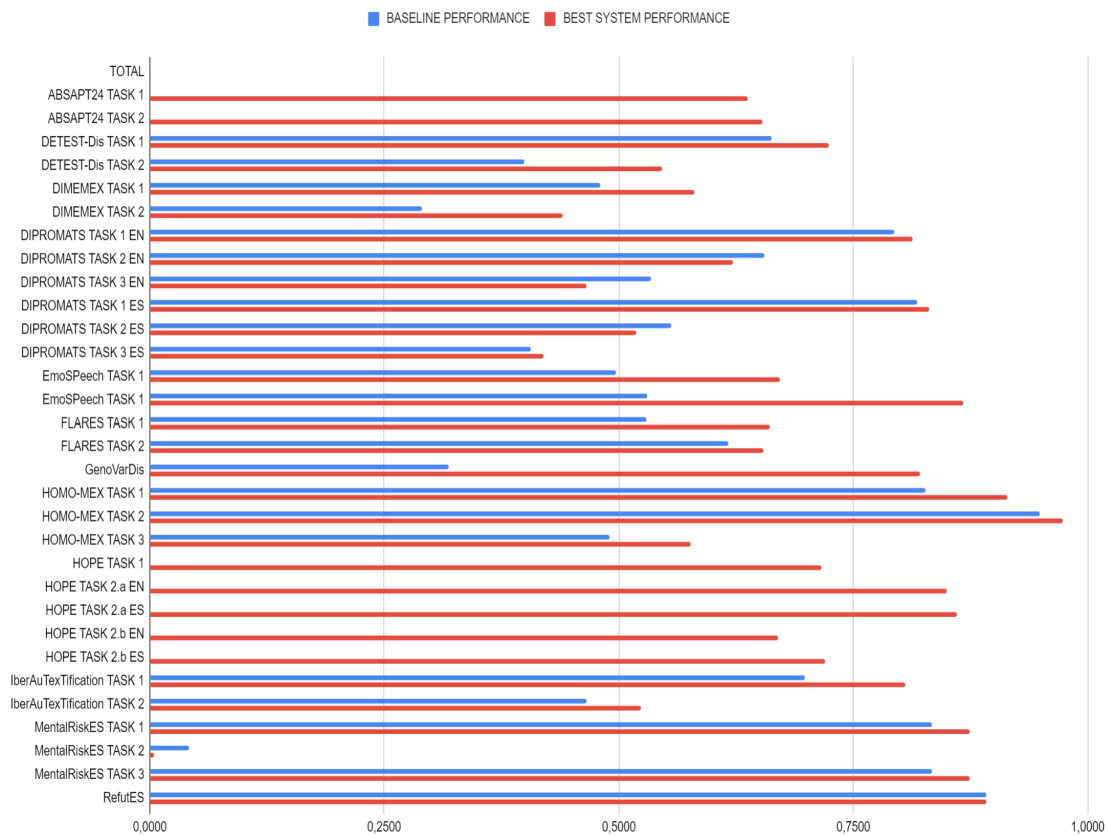
Figure 5: Performance of best systems versus baselines in IberLEF 2024 tasks. Only tasks with official evaluation metrics in the range [0-1] that include at least a baseline system are included in this graph.

4. Conclusions

In its sixth 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 12 main tasks and involved 289 researchers from institutions across 23 countries in Europe, Asia, Africa, Australia, and the Americas. Although there has been a decline in the number of participants (from 432 to 289) and participating countries (from 35 to 23) compared to the previous edition, these numbers still reflect the strong global interest that IberLEF continues to generate.

IberLEF 2024 was one of the most diverse editions in terms of task types and application domains. It advanced the field in several areas, including automatic text generation identification, biomedical NLP, counter-speech, early risk prediction on the Internet, harmful and inclusive content detection, language reliability, political ideology and propaganda identification, as well as sentiment and emotion analysis.

In the realm of Natural Language Processing, where Machine Learning and, more recently, Deep Learning have become the go-to solutions, defining research challenges and creating robust
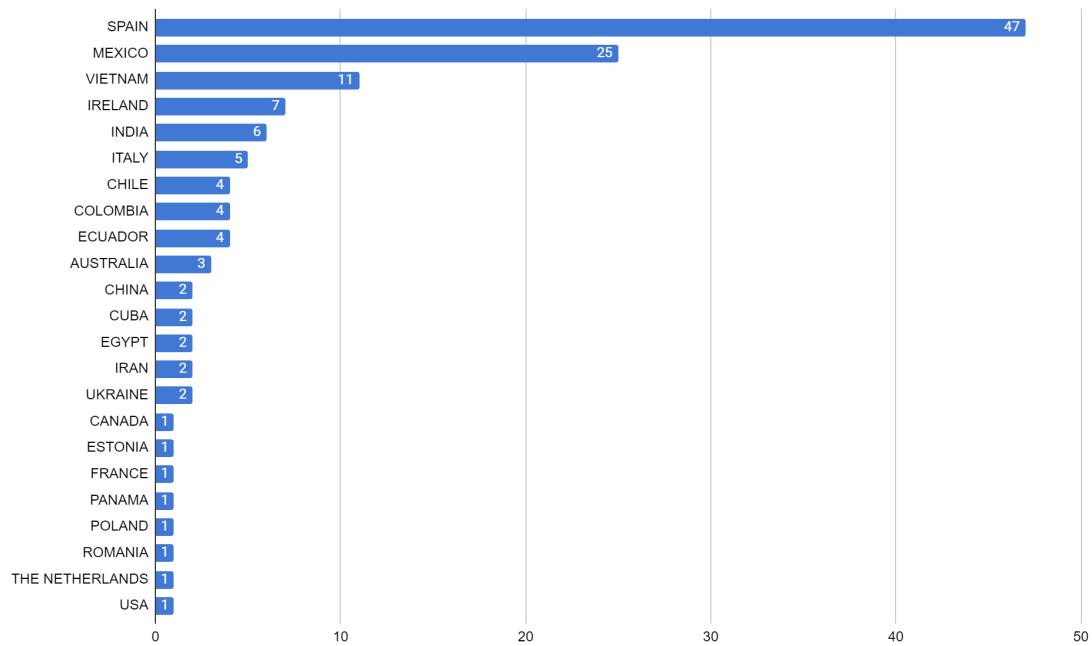
Figure 6: Number of research groups participating in IberLEF 2024 tasks per country.

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.
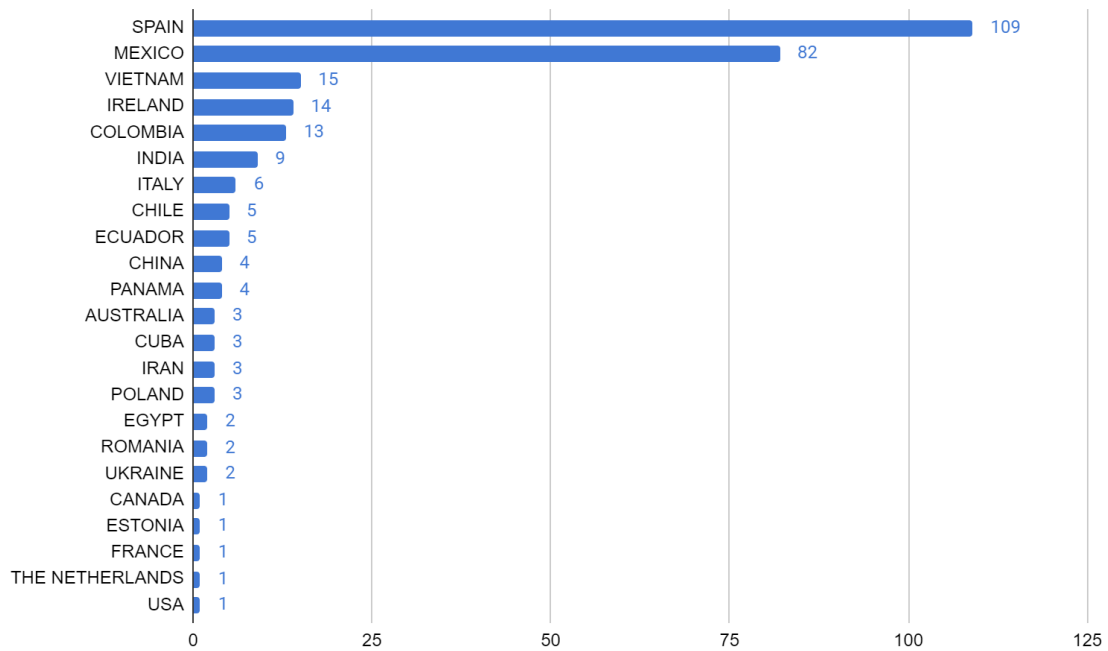
Acknowledgments

Figure 7: Number of researchers participating in IberLEF 2024 tasks per country.
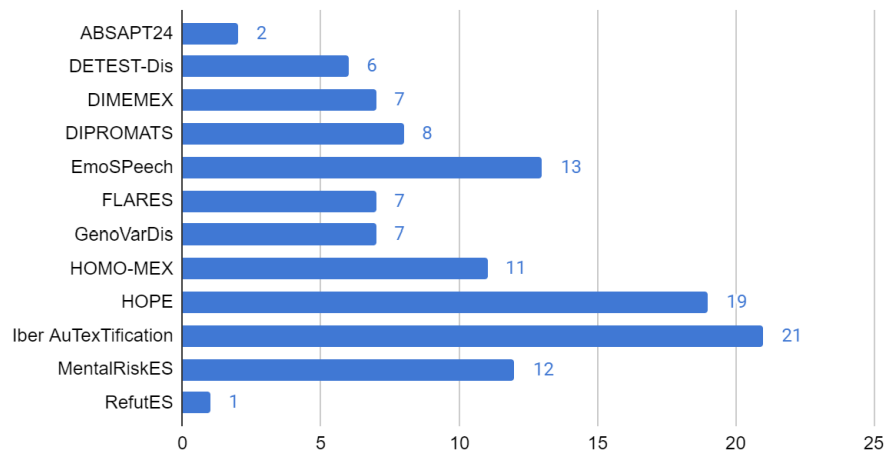

Figure 8: Distribution of participating teams per task in IberLEF 2024. The figure displays the number of teams that submitted at least one run.

References

[1] A. M. Sarvazyan, J. Á. González, F. Rangel, P. Rosso, M. Franco-Salvador, Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula, Procesamiento del Lenguaje Natural 73 (2024).
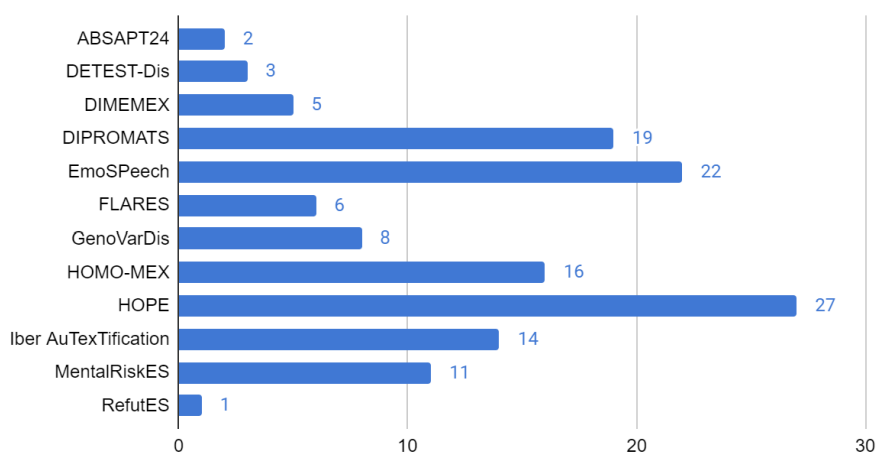
Figure 9: Distribution of participant groups per task in IberLEF 2024. The figure displays the number of groups that submitted at least one run.

[2] M. M. Agüero-Torales, C. Rodríguez Abellán, M. Carcajona Mata, J. I. Díaz Hernández, M. Solís López, A. Miranda-Escalada, S. López-Alvárez, J. Mira Prats, C. Castaño Moraga, D. Vilares, L. Chiruzzo, Overview of GenoVarDis at IberLEF 2024: NER of Genomic Variants and Related Diseases in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[3] M. E. Vallecillo-Rodríguez, M. V. Cantero-Romero, I. C. de Castro, L. A. Ureña-López, A. Montejo-Ráez, M. T. Martín-Valdivia, Overview of RefutES at IberLEF 2024: Automatic Generation of Counter Speech in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[4] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M.-G. M. Dolores, M. T. Martín-Valvidia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2024: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[5] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M.-G. M. Dolores, M. T. Martín-Valvidia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[6] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda, A. Ariza-Casabona, M. Farrús, P. Rosso, M. Taulé, Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement, Procesamiento del Lenguaje Natural 73 (2024).

[7] H. Jarquín-Vásquez, I. Tlelo-Coyotecatl, M. Casavantes, D. I. Hernández-Farías, H. J. Escalante, L. Villaseñor-Pineda, M. M. y Gómez, Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico, Procesamiento del Lenguaje Natural 73 (2024).

[8] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, A. Tania, 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).

[9] 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).

[10] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. García-Cumbreras, A. Lambebo Tonja, J. A. García-Díaz, S. Bozkurt, B. R. Chakravarthi, H. G. Ceballos, V.-G. Rafael, G. Sidorov, L. A. Ureña-López, A. Gelbukh, S. M. Jiménez-Zafra, Overview of HOPE at IberLEF 2024: Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations, Procesamiento del Lenguaje Natural 73 (2024).

[11] S. M. Jiménez-Zafra, M. Á. García-Cumbreras, D. García-Baena, J. A. García-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of HOPE at IberLEF 2023: Multilingual Hope Speech Detection, Procesamiento del Lenguaje Natural 71 (2023).

[12] S. M. Jiménez-Zafra, F. Rangel, M. M.-y. Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEURWS. org, 2023.

[13] R. Sepúlveda-Torres, A. Bonet-Jover, I. Diab, I. Guillén-Pacho, I. C. de Castro, C. Badenes-Olmedo, E. Saquete, M. T. Martín-Valdivia, P. Martínez-Barco, L. A. Ureña-López, Overview of FLARES at IberLEF 2024: Fine-grained Language-based Reliability Detection in Spanish News, Procesamiento del Lenguaje Natural 73 (2024).

[14] P. Moral, J. M. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of DIPROMATS 2024: Detection, Characterization and Tracking of Propaganda in Messages from Diplomats and Authorities of World Powers, Procesamiento del Lenguaje Natural 73 (2024).

[15] P. Moral, G. Marco, J. Gonzalo, J. Carrillo-de Albornoz, I. Gonzalo-Verdugo, Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 71 (2023).

[16] A. Thurow-Bender, G. A. Gomes, E. P. Lopes, R. M. Araujo, L. A. de Freitas, U. B. Corrêa, Overview of ABSAPT at IberLEF 2024: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese, Procesamiento del Lenguaje Natural 73 (2024).

[17] R. Pan, J.-A. García-Díaz, M. Á. Rodríguez-García, F. García-Sánchez, R. Valencia-García, Overview of EmoSPeech at IberLEF 2024: Multimodal Speech-text Emotion Recognition in Spanish, Procesamiento del Lenguaje Natural 73 (2024).