Preface on the Iberian Languages Evaluation Forum (IberLEF 2024)

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

In its sixth edition, IberLEF 2024 has contributed to the field of NLP in Spanish and other Iberian languages with a total of 134 research groups from 23 countries participating in 12 NLP challenges in Spanish, English, Portuguese, and other Iberian languages such as Euskara, Galician, or Catalan.

This volume begins with an overview of all the activities conducted during IberLEF 2024, along with aggregated figures and insights about the various tasks. It also includes a collection of papers describing the participating systems. However, the task overviews are not included in these proceedings; they have been published in the September 2024 issue of the journal *Procesamiento del Lenguaje Natural*.

IberLEF 2024 has addressed the following tasks:

1 Automatically Generated Texts Identification

Iber AuTexTification is the second edition of the AuTexTification task that started on 2023. The task consists of two subtasks: *i)* A binary classification task to detect machine- vs. human-generated texts; and, *ii)* A multi-class classification tasks consisting of, given an automatically generated text, identifying which text model generated it. This year's edition extends the task by adding more domains; and, more languages from the Iberian Peninsula (Portuguese, Galician, Euskera, and Catalan); and 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 Biomedical NLP

GenoVarDis aims to encourage work on the biomedical NLP domain for Spanish, in this case with the task of detecting names of genomic variants, diseases and symptoms in PubMed scientific articles in Spanish. The task was to detect spans of named entities and to classify 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.

3 Couter Speech

RefutES focuses on the generation of counter-narratives in Spanish, a very relevant topic in the current social media landscape. Participant systems had to create automatic responses to offensive messages that reject the narratives behind them in a neutral and respectful manner. Besides the usual automatic metrics, this task promoted the use of computational and energy efficiency metrics, and also a subset of results were manually evaluated. Six participant teams registered but only one team submitted results, obtaining a maximum of 89.23 BERTScore-F1 and 63.25 MoverScore.

4 Early Risk Prediction on the Internet

MentalRiskES is the second edition of the MentalRiskES task aimed at promoting the early detection of mental risk disorders in Spanish. This edition proposed three novel tasks: i) detecting if a user suffers from depression or anxiety, or if there is no detected disorder at all; ii) determining the context that may be associated with the disorder; and iii) detecting if a user is manifesting symptoms of potential suicidal ideation. 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.

5 Harmful and Inclusive Content

DETESTS-Dis is the second edition of the DETESTS task, aimed at detecting the of explicit or implicit stereotypes in social media content. This time, given the potential subjectivity of the task, the data is presented as a set of disaggregated annotations and participants can include this information. It is structures as two subtasks: detecting if the text contains stereotypes, and then detecting if it is explicit of implicit. 15 teams signed up for the task, of which six sent runs and three sent working notes.

DIMEMEX is a multimodal task whose purpose is to distinguish between appropriate, inappropriate content or hate speech in memes using Mexican Spanish. It contained two subtasks: i) Classifying a meme as hate speech, inappropriate, or neither; and ii) classify it in according to its content in categories like classism, sexism and racism. 19 teams signed up for the competition, 7 of them took part in the first task and 4 in the second task, submitting five working notes.

HOMO-MEX 2024 is the second edition of the HOMO-MEX task aimed at detecting and classifying LGBT+phobic content in Mexican Spanish texts, this time including digital posts and song lyrics. 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.

HOPE 2024 is the second edition of the HOPE shared task, related to the inclusion of vulnerable groups, defining hope speech as speech that is able to relax hostile environments and that helps, inspires and encourages people in times of illness, stress, loneliness or depression. This new edition includes the study of hope from two perspectives: i) Task 1 - Hope for equality, diversity and inclusion, and ii) Task 2 - Hope as expectations. 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. 19 teams participated in the competition, and 16 submitted their working notes, with the top-ranking systems achieving

71.61 Macro F1 in Task 1 and exceeding 80.00 F1 for binary classification and 78.00 for multiclass classification in Task 2.

6 Language Reliability

FLARES aims to detect reliability patterns in the language used in news that will allow the development of effective techniques for the future detection of misleading information. Two subtasks are proposed: i) the identification of the elements of the 5W1H journalistic approach; and ii) the detection of reliability. A total of 7 teams participated in the shared task, with best-performing systems achieving 66.13 and 65.36 F-score respectively.

7 Political Ideology and Propaganda

DIPROMATS 2024 is the second edition of the DIPROMATS task, refining and extending the tasks presented in the previous edition. Two tasks are proposed: i) Automatic Detection and Categorization of Propaganda Techniques; and ii) Automatic Detection of Narratives; both tasks in Spanish and English. A total of 28 teams registered for the shared task, while 9 of those teams submitted a total of 40 runs and 8 working notes.

8 Sentiment and Emotion

ABSAPT24 explores the problem of Aspect-Based Sentiment Analysis in Portuguese. It has two subtasks: i) identifying specific aspects mentioned in a text related to a given entity; and ii) determining the sentiment polarity associated with each identified aspect. Two teams submitted their results, with the best one achieving a performance of 63.70 and 65.30 per subtask.

EmoSPeech addresses the study of Automatic Emotion Recognition via two subtasks: i) AER from text, focusing on feature extraction and identifying the most representative feature of each emotion, and ii) multimodal AER, which requires more complex architectures to solve. 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.

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 evaluation methods and high-quality test collections are crucial for success. These elements enable iterative testing and refinement. IberLEF is playing an important role in advancing these efforts and moving the field forward.

September 2024.
The editors.