Preface on the Iberian Languages Evaluation Forum (IberLEF 2023)

IberLEF is a shared evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages. In an annual cycle that starts in December (with the call for task proposals) and ends in September (with an IberLEF meeting collocated with the SEPLN), several challenges are run with a large international participation from research groups in academia and industry. Its goal is to encourage the research community to organize competitive text processing, understanding and generation tasks in order to define new research challenges and set new state-of-the-art results in those languages.

In 2022, over 150 research groups from 24 countries participated in 10 NLP challenges in Spanish, Portuguese, and English. In its fifth edition, IberLEF 2023 has also contributed to the field of NLP in Spanish and other Iberian languages with the organization of 14 main tasks where over 211 research groups have been involved, from institutions in 35 countries worldwide.

This volume opens with an overview of all the activities carried out in IberLEF 2023 together with some aggregated figures and insights about the different tasks. Furthermore, the collection of papers describing the participating systems is also provided. However, the tasks' overviews are not included in these proceedings and have been published in the journal *Procesamiento del Lenguaje Natural*, in its September 2023 issue.

IberLEF 2023 has addressed the following tasks:

1 Automatically Generated Texts Identification

AuTexTification tackled the multi-domain machine-generated text detection and attribution task in English and Spanish. It consisted 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. A total of 114 teams signed up, with 36 teams participating and sending over 175 runs and 20 working notes. The best-performing team obtained a Macro-F of 80.91 and 70.77 in Subtask 1, and 62.50 and 65.37 in Subtask 2, both in English and Spanish respectively.

2 Clinical Content

ClinAIS addressed the problem of identifying seven section types within unstructured clinical records in the Spanish language. The seven section types were: *i)* Present Illness; *ii)* Past Medical History/ Medical History; *iii)* Family History; *iv)* Exploration; *v)* Evolution; *vi)* Treatment; and, *vii)* Derived from/to. A total of 27 participants registered for the task, with 5 of them submitting their results. The best-performing team obtained 80.22 weighted B2.

MEDDOPLACE has been the first initiative addressing the automatic detection and normalization of all location-relevant entity types present in clinical texts. The task was structured into four subtasks: *i)* Location and place-related entity mention detection; *ii)* Entity normalization (geocoding to GeoNames, PlusCodes and SNOMED CT); *iii)* Location entity classification; and, *iv)* End-to-end evaluation of detection, normalization and classification. A total of 20

teams signed up of which 4 submitted their predictions. The organizers carried out an extensive evaluation based on several metrics.

TESTLINK focused on relation extraction from clinical cases in Spanish and Basque. The task consisted of identifying clinical results and measures and linking them to the tests and measurements from which they were obtained. A total of 3 teams participated in the Spanish track and 2 in the Basque one, obtaining the best-performing ones an F-measure of 68.38 and 72.65 respectively.

3 Code Switch Analysis

GUA-SPA was devoted to detecting and analyzing code-switching in Guarani and Spanish. The challenge consisted of three subtasks: *i)* Identifying the language of a token; *ii)* NER; and, *iii)* A novel task of classifying the way a Spanish span is used in the code-switched context. A total of 20 teams registered to participate, three of them participated both in the development and evaluation phases and two in a single phase. The best-performing teams obtained 93.81 weighted F1 in Subtask 1, and 70.28 and 38.36 labelled F1 in Subtasks 2 and 3 respectively.

4 Early Risk Prediction on the Internet

MentalRiskES aimed at promoting the early detection of mental risk disorders in Spanish. It encompassed three detection subtasks: *i)* Eating disorders; *ii)* Depression; and, *iii)* Undisclosed disorder during the competition (anxiety) to observe the transfer of knowledge among the proposed disorders. A total of 37 teams registered, 18 submitted results, and 16 sent their working notes. The organizers analysed the systems with a wide spectrum of evaluation metrics. Furthermore, they asked participants to submit measurements of carbon emissions of their systems, emphasizing the need for sustainable NLP practices.

5 Harmful and Inclusive Content

DA-VINCIS promoted the research on automatic detection of violent events in social networks. Two subtasks were considered: *i)* A binary classification task aimed to determine whether or not a tweet is about a violent incident; and, *ii)* A multi-label multi-class classification task in which the category(ies) of a violent incident must be identified. A total of 15 systems were submitted for the final evaluation phase. The best-performing teams obtained 92.64 and 87.97 F1 respectively on each subtask.

HOMO-MEX encouraged the development of NLP systems that can detect and classify LGBTIQ+ phobic content in Spanish tweets, regardless of whether it is expressed aggressively or subtly. The task was divided into two tracks: *i)* To determine whether a tweet exhibits LGBTIQ+ phobic content or not; and, *ii)* To classify the LGBTIQ+ phobic tweets as containing Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), and/or other LGBTQ+phobia (O). A total of 8 teams submitted their working notes, and the best-performing teams obtained 84.32 and 68.43 F-score respectively per subtask.

HOPE tackled the detection of hope speech, the speech that is able to relax hostile environments and that helps, inspires and encourages people in times of illness, stress, loneliness or depression. The task has been divided into two subtasks, according to the language in which the texts were written: *i)* Identifying whether a Spanish tweet contains hope speech or not; and, *ii)* Determining whether an English YouTube comment contains hope speech or not. A total of 50 teams registered to participate, 12 submitted their results and 8 of them sent over working notes describing their systems. The best-performing teams obtained 91.61 and 50.12 F1-scores respectively per subtask.

HUHU proposed a frame to study how humour is used to discriminate minorities and to analyse their interplay with the degree of prejudice expressed against specific groups such as women and feminists, LGBTIQ+ community, immigrants and racially discriminated people, and overweight people. This shared task consists of three subtasks: *i)* Determining whether a prejudicial tweet is intended to cause humour; *ii)* Identifying the targeted groups (women and feminists, LGBTIQ+ community, immigrants and racially discriminated people, and overweight people) on each tweet as a multilabel classification task; and, *iii)* Predicting how prejudicial a message is on average to minority groups on a continuous scale from 1 to 5. A total of 46 teams participated in the task, obtaining the top-ranked systems an F-measure of 82.00, 79.60 and 85.50 respectively per subtask.

6 Political Ideology and Propaganda

DIPROMATS addressed the identification and categorization of propagandistic tweets from governmental and diplomatic sources in Spanish and English. Three subtasks were proposed: *i)* To distinguish if a tweet has propaganda techniques or not; *ii)* To classify the tweet into four clusters of propaganda techniques; and, *iii)* A fine-grained categorization of 15 techniques. There was a total of 9 different teams who submitted 34 runs, and the best-performing ones obtained respectively per subtask an F-measure of 80.89, 45.78 and 36.28 in Spanish and 80.90, 55.91 and 48.38 in English.

PoliticES shared the goal with its previous edition to extract political ideology and other psychographic and demographic characteristics of users in social networks. This year's novelty is that the traits were extracted from text clusters of users who shared the same traits. Furthermore, celebrities were included as a type of profession. The task focused on the identification of two demographic traits (self-assigned gender and profession), and one psychographic trait (political ideology), from a binary and multi-class perspective. This edition attracted 43 teams, of which 11 submitted results and 8 sent papers describing their systems. The best-performing approaches obtained an F-measure of 82.96, 86.08, 89.67 and 69.13 respectively per subtask.

7 Sentiment, Stance and Opinions

FinancES extended the challenge of sentiment analysis in Spanish to the financial domain in order to extract the sentiment that a piece of financial information can have for several actors, including the main economic target (i.e., the specific company or asset where the economic fact applies), other companies (i.e., the entities producing the goods and services that others consume) and consumers (i.e., households/ individuals). Two tasks were proposed: *i)* The identification of the main target, determining the sentiment polarity towards such target; and, *ii)* The assessment of the sentiment towards both other companies and consumers. A total of 10 teams submitted their approaches and the best-performing ones obtained an overall F-measure of 79.22 and 64.24 respectively.

Rest-Mex focused on sentiment analysis and text clustering of tourist texts related to tourist destinations in Mexico, Cuba and Colombia. The task consists of two subtasks: *i)* A classification task with the aim of predicting the polarity of opinions expressed by tourists, classifying the type of place visited, whether it's a tourist attraction, hotel, or restaurant, as well as the country it is located in; and *ii)* A text clustering task aiming to classify news articles related to tourism in Mexico. A total of 16 teams submitted 61 solutions for the sentiment analysis tasks, and the organizers evaluated the systems with a wide variety of metrics.

In a field where Machine Learning, and recently Deep Learning, is the ubiquitous approach to solving challenges, the definition of research challenges, their associated evaluation method-

ologies, and the development of high-quality test collections that allow for iterative evaluation is probably the most critical step towards success. We believe IberLEF is making a significant contribution in this direction.

September 2023.
The editors.