Preface on the Iberian Languages Evaluation Forum (IberLEF 2022)

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 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 2021, over 100 research groups from 22 countries participated in 12 NLP challenges in Spanish, Portuguese, Basque and English. In its fourth edition, IberLEF 2022 has also contributed to the field of NLP in Spanish and other Iberian languages with the organization of 10 main tasks where over 150 research groups have been involved, from institutions in 24 countries worldwide.

This volume opens with an overview of all the activities carried out in IberLEF 2022 together with some aggregated figures and insights about the different tasks. Furthermore, the collection of papers describing the participating systems are 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 2022 issue.

IberLEF 2022 has tackled the following tasks:

Sentiment, Stance and Opinions

ABSAPT tackled aspect-based sentiment analysis in Portuguese, where the participants were asked to extract aspects (AE sub-task) from reviews about hotels, and classify their sentiment (ASC sub-task). Twelve teams registered to the task, among which five submitted predictions and technical reports, obtaining a best result of 0.67 accuracy in AE sub-task, and 0.82 balanced accuracy in ASC sub-task.

PoliticEs aimed at detecting the gender and profession of Twitter users, and extracting their political ideology from a given set of their tweets. A total of 63 teams registered to the task and 20 submitted their results. Most of the approaches were based on transformers, although also some traditional machine learning algorithms (and their combination) were explored. The best overall macro-F (0.9023) has been achieved by a system based on transformers which combined BETO and MarIA with a multi-layer Perceptron.

Rest-Mex 2022 was divided into three tracks: *(1)* Recommendation System; *(2)* Sentiment Analysis; and *(3)* Covid Semaphore Prediction. The Recommendation System task consisted in predicting the degree of satisfaction that a tourist may have when recommending a destination of Nayarit, Mexico, based on places visited by the tourists and their opinions. On the other hand, the Sentiment Analysis task predicted the polarity of an opinion issued and the attraction by a tourist who traveled to the most representative places in Mexico. The copora consisted of Spanish opinions extracted from TripAdvisor. The Covid Semaphore Prediction task aimed to predict

the color of the Mexican Sempahore for each state, according to the Covid news in the state, using data from the Mexican Ministry of Health. For the three tasks, 18 teams participated, submitting 35 different systems. The best result in the recommendation task (0.69 MAE) was obtained by a system based on BoW, whereas the best overall results respectively for the sentiment analysis task (0.89) and Covid semaphore (0.49) were obtained with BERT-based systems.

Harmful Content

DA-VINCIS@IberLEF2022 challended participants to develop automated solutions for the detection of violent events mentioned in social networks; concretely, using a corpus collected from Twitter and manually labeled with 4 categories of violent incidents (plus the no-incident label). The shared task focused on the Mexican variant of Spanish and it was divided into two tracks: *(1)* a binary classification task in which users had to determine whether tweets were associated to a violent incident or not; and *(2)* a multi-label classification task in which the category of the violent incident should be spotted. More than 40 teams registered for the task and 12 participants submitted predictions for the final phase. In both sub-tasks, transformer-based solutions obtained the best results (F measures of 0.775 and 0.554, respectively)

DETEST 2022 proposed two hierarchical subtasks: For subtask 1, participants had to determine the presence of stereotypes in sentences. For subtask 2, participants had to classify the sentences labeled with stereotypes into ten categories. The DETESTS dataset contains 5,629 sentences in comments in response to newspaper articles related to immigration in Spanish. A total of 51 teams signed up to participate, of which 39 sent runs, and 5 of them sent their working notes. In sub-task 1, the best performing team achieved an F-score of 0.7042 with an ensemble architecture combining different pre-trained language models. In sub-task 2, based on the ICM metric, the best performing team obtained a value of -0.2380 approaching the task as a multi-task learning problem in which a final classification head per stereotype category is stacked on top of a pre-trained RoBERTa model and fine-tuned using a point-wise Cross-Entropy loss function.

EXIST 2022 consisted of two challenges: sexism identification and sexism categorization of tweets and gabs, both in Spanish and English (the dataset consists of more than 12,000 annotated texts from social networks such as Twitter and Gab, manually labeled following two different procedures: external contributors and trained experts). The task received a total of 45 runs for the sexism identification task and 29 runs for the sexism categorization task, submitted by 19 different teams. In sub-task one, the best performing team achieved an overall F1 of 0.7996 with an ensemble of transformers models for different hyper-parameter configurations. In sub-task two, the same team also obtained the best overall F1 (0.5106).

Information Extraction and Paraphrase Identification

LivingNER promotes the development of tools for finding mentions of species, pathogens, or food from medical texts. The task relied on a large Gold Standard corpus of 2,000 carefully selected clinical cases in Spanish covering diverse specialities, which was manually annotated with species mentions that were also carefully mapped to their corresponding NCBI Taxonomy identifiers. Furthermore, the organizers have generated Silver Standard versions of LivingNER for 7 languages: English, Portuguese, Galician, Catalan, Italian, French, and Romanian. LivingNER had three subtasks: *(1)* LivingNERSpecies NER (species mention detection sub-task); *(2)* LivingNER-Species Norm (species mention detection and normalization to NCBI taxonomy Ids); and *(3)* LivingNERClinicalIMPACT (a document classification task related to

the detection of pets, animals causing injuries, food, and nosocomial entities). A total of 62 systems from 20 teams from 11 countries worldwide obtained highly competitive results with successful approaches, typically modifications of pre-trained transformer-like language models (BERT, BETO, RoBERTa, etc.) or embedding distance metrics for entity linking.

PAR-MEX focused on the task of paraphrase detection in Spanish, concretely on a corpus with topics in the domain of Mexican gastronomy (e.g., sushi, molecular cuisine, tequila, kebab, day of the dead, vegan food, street food). Six teams submitted one or more solutions, where half of them submitted transformer-based approaches while the other half approached the task with traditional machine learning. The data set of PAR-MEX included both, low-level and high-level pairs of paraphrases, although they were not distinguished for the participants. A number of 6 teams have participated in the task, and the analysis of the results showed that, whereas low-level paraphrase is currently an easy task for natural language processing (0.90 of average), high-level paraphrase is a problem that has not been conveniently approached yet.

Question Answering and Machine Reading

QuALES addressed the problem of Extractive Question Answering from texts collected from Uruguayan media news about the Covid-19 pandemic and related topics. There were 7 participants submitting their systems, all of them based on different BERT-like language models. The best results (0.5349 exact match metric, 0.7282 F1) were obtained using the multilingual RoBERTa model pre-trained with SQUAD-Es-V2 and a fine tuning on the QuALES corpus.

ReCoRES main goal was to promote the task of Reading Comprehension and Verbal Reasoning. This task was divided into two sub-tasks: *(1)* identifying the correct alternative in reading comprehension questions; and *(2)* generating the reasoning used to select an alternative. Finally, 3 teams participated in this event, mainly proposing transformer-based neural models in conjunction with additional strategies, obtaining results up to 0.7591 of accuracy in sub-task 1 and 0.6867 BERTScore in sub-task 2.

In a field where Machine Learning, and recently Deep Learning, is the ubiquitous approach to solve challenges, the definition of research challenges, their associated evaluation methodologies, 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 2022.
The editors.