Preface on the Iberian Languages Evaluation Forum (IberLEF 2021)

IberLEF is a comparative evaluation campaign for Natural Language Processing Systems in Spanish and other Iberian languages. 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 its third edition in 2021, IberLEF has again been a remarkable collective effort for the advancement of Natural Language Processing in Spanish and other Iberian languages: with 12 main tasks and 359 researchers involved, from institutions in 22 countries in Europe, Asia and the Americas, IberLEF 2021 has been the largest up to date, and has contributed to advance the field in the areas of emotions, stance and opinions, harmful information, health-related information extraction and discovery, humour and irony, and lexical acquisition. In a field where Machine 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.

This volume is a collection of papers describing systems that participated in the evaluation activities carried out in IberLEF 2021. Besides system descriptions, the volume opens with an overview of all activities carried out in IberLEF 2021, providing some aggregated figures and insights. Papers with task overviews, however, are not included in these proceedings, and have been published in the journal *Procesamiento del Lenguaje Natural*, in its September 2021 issue.

The tasks undertaken at IberLEF 2021 have been:

Emotions, Stance and Opinions

EmoEvalEs was an emotion classification task, where systems are asked to predict which emotions are present in texts written in Spanish (from this set: anger, disgust, fear, joy, sadness, surprise, others). Twitter was used as textual source, and the dataset consists of 8232 manually annotated tweets. 15 research groups submitted runs for this task, out of which 11 submitted papers to the proceedings.

REST-MEX was an evaluation exercise focused on recommendation tasks using TripAdvisor as textual source, with texts written in several variants of Spanish (Mexican Spanish being the most common). Task 1 (*Recommendation*) consists in predicting the degree of satisfaction (in a 1-5 scale) of a tourist visi-

IberLEF 2021, September 2021, Málaga, Spain.

ting a given Mexican place, given the information available in TripAdvisor about the tourist and about the site. The tourist profile includes gender, place of origin, her textual self-description in TripAdvisor, and her opinions on places she has visited. The information about the place is a brief textual description and a series of representative characteristic of the place for touristic purposes (adventure, beach, family atmosphere, etc.). Task 2 (*Sentiment Polarity*) consists of predicting the polarity (in a 1-5 scale) of a given TripAdvisor opinion.

Overall, the dataset gathers 2263 instances tourist/destination for the first task and 7413 opinions for the second task. 2 groups submitted results for task 1 and 7 for task 2.

VaxxStance focused on predicting the stance of short texts (tweets) with respect to vaccines (in favour, neutral or against). This was a multilingual task including Spanish (2697) and Basque (1384) tweets.

The challenge was addressed in three variants: in Task 1 (*close track*), systems could only use the text of the tweets; in Task 2 (*open track*), systems could use any kind of data (including tweets' metadata); finally, Task 3 (*zero-shot track*) was a cross-lingual stance detection challenge: systems were trained on one of the languages and tested on the other language. Three groups participated in the first task, and one in the second and third tasks.

Harmful Information

There were four challenges around harmful textual information in 2021:

MeOffendES focused on offensive language detection in Spanish, and included two subtasks on a dataset of generic Spanish and two subtasks on a Mexican Spanish corpus. The generic Spanish dataset (OffendES) comprises 30,416 comments collected from Twitter, Instagram and Youtube; the Mexican Spanish dataset (OffendMEX) comprises 7319 annotated tweets.

The tasks on generic Spanish asked systems to predict the right class from OFP (offensive, target person), OFG (offensive, target group), OFO (offensive, target others), NOE (non offensive, but with expletive language), NO (not offensive). Systems were also asked to predict the strenght of the class, taken as the ratio of annotators than concur on the class. Subtask 1 allowed textual data as input, and Subtask 2 allowed metadata as additional input. Four teams submitted results for the first task, and one for the second.

The tasks on Mexican Spanish asked systems to do a binary prediction (offensive / not offensive), using only textual input (Subtask 3) or also metadata (Subtask 4). 10 groups submitted results to Subtask 3 and one to Subtask 4.

EXIST focused on the identification of sexism in Spanish and English texts, asking systems to predict whether a text has sexist content (Subtask 1) and to identify the type of sexism (Ideological and inequality / stereotyping and dominance / objectification / sexual violence / mysogyny and non-sexual violence) in Subtask 2. The dataset comprises 13,000 tweets and 982 gabs. 31 groups submitted results for the first subtask, and 27 for the second.

DETOXIS focused on the identification of toxic content in texts, and prepared a dataset with 4359 comments from news and online forums, annotated with their level of toxicity (in a scale from 0 to 3). Subtask 1 required a binary classification (toxic / non toxic) and Subtask 2 asked systems to predict the level of toxicity in the same scale that was annotated. 31 groups submitted to the first task and 24 to the second.

FakeDeS focused on discovering fake news written in Spanish, and prepared a dataset with 971 news articles written in Spanish from Spain and Mexico. It was designed as a binary classification task (fake or real), and 16 groups submitted results.

Health-Related Information Extraction and Discovery

Health-Related content received special attention in IberLEF 2021, as in previous editions, with two tasks related to the medical domain:

e-HealthKD focused on entity recognition and classification (subtask A) and relation extraction (subtask 2) in both Spanish and English. Systems had to recognize and classify concepts, actions, predicates and references in subtask 1, and to extract relations between them (subtask B). e-HealthKD also contemplated a main, complex task where both entity recognition and relation extraction were evaluated jointly. 8 participants submitted results to subtask A and, out of them, 7 also submitted results to subtask B and to the main challenge.

The organizers performed an exhaustive annotation of 1,800 sentences extracted from MedLinePlus, WikiNews and the CORD-19 corpus.

MEDDOPROF worked on clinical cases (the annotations include 1844 cases extracted from medical literature), and asked systems to annotate information related to occupations/professions. Task 1 (NER) was about finding mentions of occupations and classifying each of them as a profession, an employment status or an activity; Task 2 (CLASS) involved finding mentions of occupations and determining whether they are related to the patient, to a family member, to a health professional or to someone else; and Task 3 (NORM) was about mapping predictions to one of the codes in a list of unique concept identifiers from the European Skills, Competences, Qualifications and Occupations (ESCO) classification and relevant SNOMED-CT terms. 15 groups submitted results to Task1, 11 to Task 2 and 8 to Task 3.

Humour and Irony

There were two tasks related to Humour and Irony in 2021:

HAHA dealt with humour detection and characterization in Spanish texts, and included four subtasks: (1) humour detection, which required determining if a tweet was humorous or not; (2) funniness score prediction, in a 1-5 scale; (3) humour mechanism classification, out of a set of classes such as irony, wordplay, hyperbole or shock; (4) humour content classification: predict the content of the joke from a set of classes such as racist jokes, sexist jokes, dark humour, dirty

jokes, etc. The dataset included 36,000 annotated tweets. 14 groups submitted to the first task, 11 to the second, 9 to the third and 8 to the fourth.

IDPT was a task on irony detection in Portuguese texts, defined as a binary classification problem (is this text ironic or not?). The dataset included 18494 news pieces and 15212 tweets, and 7 groups submitted results for the task.

Lexical Acquisition

ADoBo focused on the acquisition of borrowings into Spanish from other languages (English primarily). Systems were asked to detect expressions (in Spanish news articles) that have been imported from other languages in their raw form. The dataset is an annotated collection of news articles that comprise 372,701 tokens. Four systems submitted results for this task.

September 2021.
The editors