IberLEF 2021 Overview: Natural Language Processing for Iberian Languages

Julio Gonzalo[1], Manuel Montes-y-Gómez[2], and Paolo Rosso[3]

[1] UNED NLP and IR Research Group, Madrid, Spain
julio@lsi.uned.es
[2] National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico
mmontesg@inaoep.mx
[3] PRHLT Research Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Resumen 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. This paper summarizes the evaluation activities carried out in IberLEF 2021, which included twelve tasks dealing with emotions, stance and opinions, harmful information, health-related information extraction and discovery, humor and irony, and lexical acquisition. Overall, IberLEF activities were a remarkable collective effort involving 359 researchers from 22 countries in Europe, Asia and the Americas.

Keywords: Natural Language Processing · Artificial Intelligence · Evaluation

1. Introduction

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. This paper summarizes the evaluation activities carried out in IberLEF 2021, which included twelve tasks dealing with emotions, stance and opinions, harmful information, health-related information extraction and discovery, humor and irony, and lexical acquisition. Overall, IberLEF activities were a remarkable collective effort involving 359 researchers from 22 countries in Europe, Asia and the Americas. Papers with system descriptions are included in the IberLEF 2021 Proceedings [11], and papers with

IberLEF 2021, September 2021, Málaga, Spain.

task overviews have been published in the journal *Procesamiento del Lenguaje Natural*, vol. 67 (September 2021 issue).

In this paper we summarize the activities carried on in IberLEF 2021, extracting some aggregated figures for a better understanding of this collective effort.

2. IberLEF 2021 Tasks

These are the twelve tasks organized succesfully in 2021, grouped thematically:

2.1. Emotions, Stance and Opinions

EmoEvalEs [4] was an emotion classification task, where systems were 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 [5] 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 visiting 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 characteristics 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 [1] 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.

2.2. Harmful Information

There were four challenges around harmful textual information in 2021:

MeOffendES [3] 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 [14] 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 [15] 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.

Finally, **FakeDeS** [8] 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.

2.3. 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 [12] focused on entity recognition and classification. 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 [9] worked on clinical cases (the annotations include 1844 cases extracted from medical literature), and asked systems to annotate informa-

tion 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 Task 1, 11 to Task 2 and 8 to Task 3.

2.4. Humour and Irony

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

HAHA [6] dealt with humour detection and characterization in Spanish texts, and included four subtasks: (1) humour detection, which required determining whether 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 [7] 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.

2.5. Lexical Acquisition

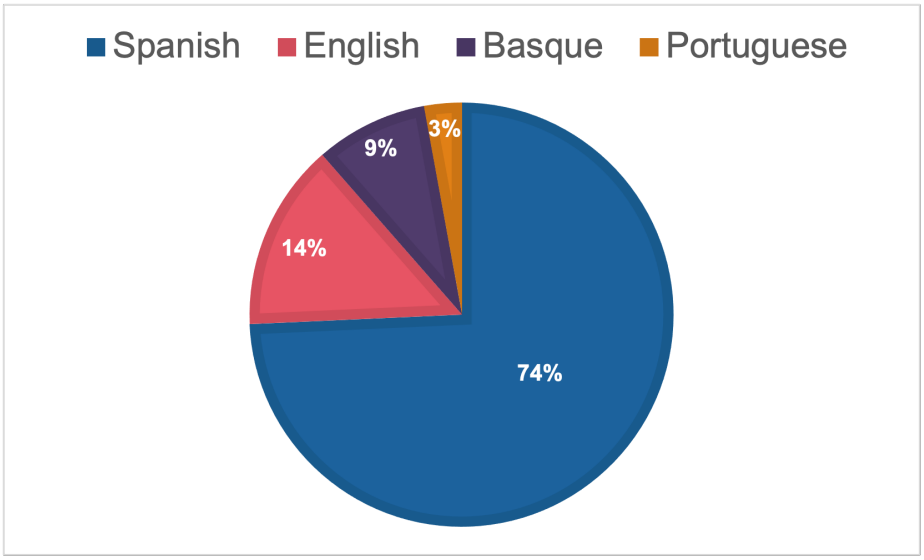
ADoBo [10] focused on the acquisition of borrowings into Spanish from other languages (English primarily). Systems were asked to detect expressions (in Spanish news articles) that have been imported from other languages in their raw form. The dataset is an annotated collection of news articles that comprise 372,701 tokens. Four systems submitted results for this task.

3. Aggregated Analysis of IberLEF 2021 Tasks

3.1. Tasks characterization

In terms of **languages**, the distribution per tasks (including subtasks) is shown in Figure 1. 74 % of the tasks deal at least with Spanish, which is the predominant subject of study in IberLEF. In terms of variants of Spanish, Spain and Mexico are the best represented, with other variants having only anecdotal presence. English is used (never as the main language) in 14 % of the tasks, and this year Basque appears for the first time in IberLEF being present in 9 % of the tasks (all belonging to VaxxStance). Finally, there is also one task dealing with Portuguese.

Figura 1. Distribution of languages in IberLEF 2021 tasks



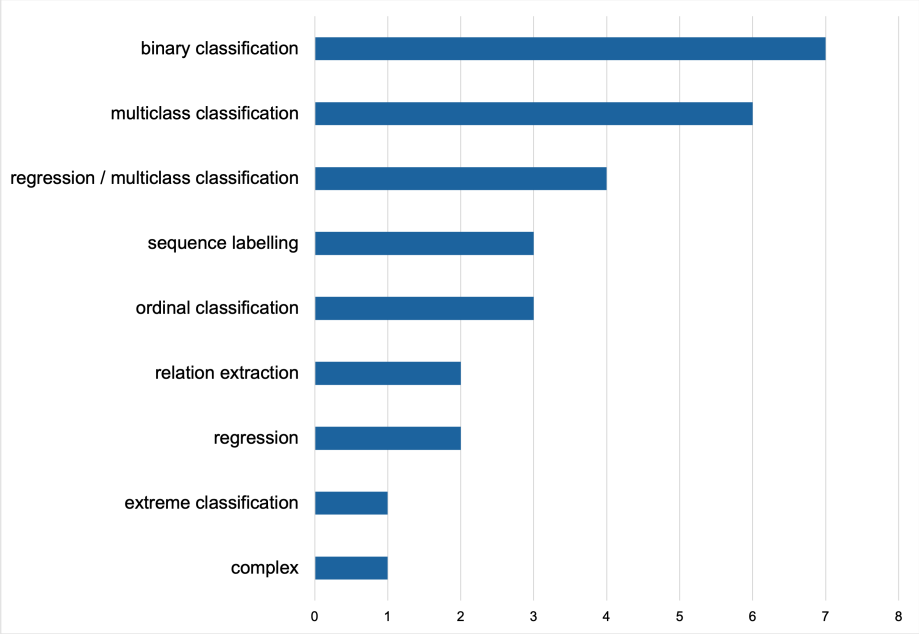
The trend in the number of languages is positive: there were two in IberLEF 2019 (Spanish and Portuguese), only one in 2020 (Spanish) and four languages in 2021.

In terms of **abstract task types**, the distribution of tasks can be seen in Figure 2. Out of a total of 29 tasks (each subtask is counted as a task here), 7 (24 %) are binary classification tasks, which is the most popular choice. Multiclass classification problems are also well represented with 6 tasks. There are also four tasks where classes are ordinals (e.g. 0,1,2,3) that can be interpreted either as a regression or a multiclass classification problem (*regression / multiclass classification* in the figure). Another variant of classification problems is ordinal classification, where classes have a relative ordering (e.g. in favour, neutral or against in stance classification): 3 tasks match this abstract task type. Finally, there is also a normalization task which implies matching profession descriptions in text with standard thesauri / ontologies, which can be seen as an extreme classification task (i.e. a classification problem where the number of classes is extremely large).

There are only 3 sequence labelling tasks, which is perhaps less than expected for an evaluation campaigned focused heavily on Natural Language Problems: tasks that identify specific structures or text chunks in text, such as named entities, fall into this category. Two of them are related to the medical domain, and the other one looks for lexical borrowings (imports from other languages).

Finally, there are two genuine regression tasks, where systems must predict a real number, and only one complex task, where the organizers try to measure the joint performance of systems in two subtasks that build together with a common goal: the e-healthKD main task.

Figura 2. Distribution of IberLEF 2021 tasks per abstract task type.

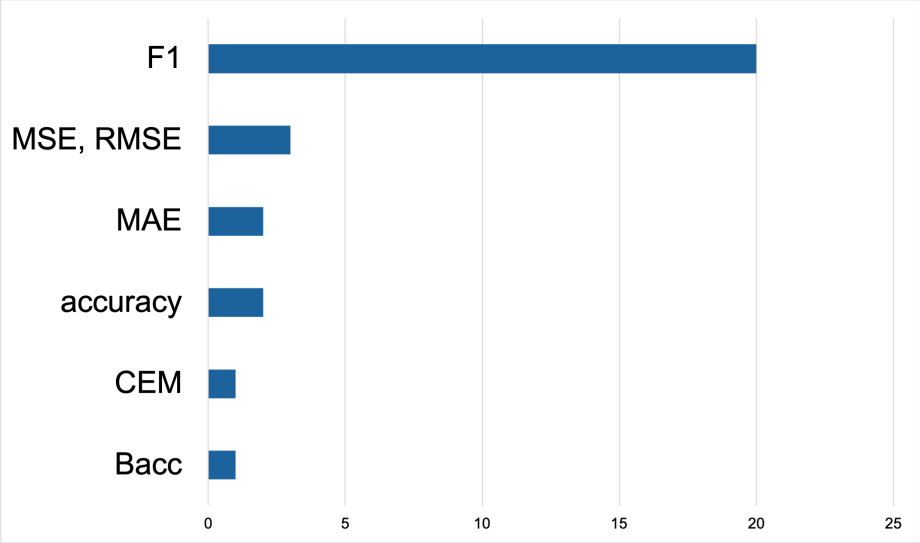


Overall, IberLEF 2021 tasks address a representative sample of abstract task types, covering a wide range of problems. Probably, to get nearer industry needs, in the future we should investigate more how to evaluate complex, end user tasks. IberLEF is also missing tasks that involve text generation, such as text summarization or machine translation problems; and tasks that involve interaction with the users, such as dialogue systems. Finally, we would like to see more application domains in the list of tasks.

In terms of **evaluation metrics**, the distribution can be seen in Figure 3. As in previous years, there is a remarkable predominance of F1 (20 tasks used it as the main evaluation metric to rank systems), which is used for all types of classification tasks (even if it does not perfectly match the problem at hand, as in ordinal classification problems) and for sequence labelling problems. Accuracy is used in a couple of classification tasks, and Bacc (Balanced Accuracy) in another. Finally, CEM (Closeness Evaluation Metric), a metric specially useful for ordinal classification tasks and introduced recently [2] is used for one of the classification/regression tasks. Tasks interpreted as regression problems are evaluated with MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error), with none of these metrics particularly favoured.

Overall, we take this as a hint that the field might be relying too much on F1. It has some desirable properties (particularly, it is robust to the characteristics of the dataset), but it has severe limitations too. Its primary shortcoming is that it hides the actual behaviour of systems, as with all averages (F1 is a harmonic

Figura 3. Distribution of official evaluation metrics in IberLEF 2021 tasks.



average of precision and recall). For multiclass classification, the most common procedure is to compute the arithmetic average of the harmonic averages of precision/recall across classes, which is a way of focusing exclusively on system ranking and giving up on understanding why systems fail and when. We think that the usage of F1 should be accompanied with other metrics.

Most importantly, the choice of metrics does not seem to be made justified on how the system output is going to be used, but rather on mere popularity of the metrics. This is not a shortcoming of IberLEF tasks only: most NLP challenges suffer from the same problems.

Figure 4 shows how IberLEF tasks have evolved in the three years that it has been running on. The number of tasks has increased (from 9 in 2019 to 12 in 2021); and in 2021 the number of new tasks is 9 (75 %), a sign that the scope of problems being studied becomes larger every year. The lower figures in 2020 are due to the irruption of COVID-19: some of the tasks could not be completed and are not depicted in the graph.

3.2. Datasets and results

In terms of **types of textual sources**, Figure 5 shows how they are used in IberLEF 2021 tasks. Twitter is the most popular source, with 15 tasks relying solely or partially on Twitter data. This does not necessarily mean that the field is primarily interested in microblogging communication; it probably reflects that collecting Twitter data is more cost effective given IPR issues and other difficulties in gathering data to redistribute to the scientific community [13]. All other sources are used by at most three tasks. The good news is that there are many

Figura 4. Evolution of IberLEF tasks across time.

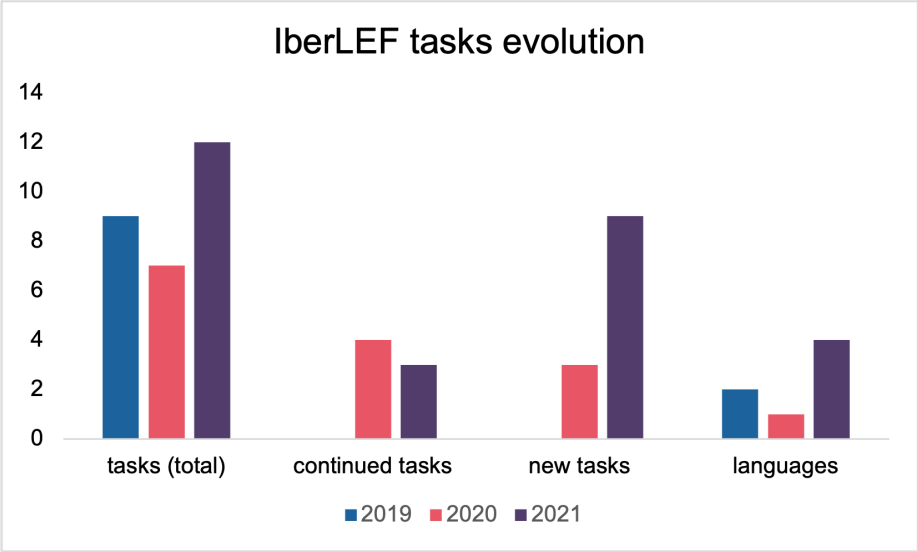
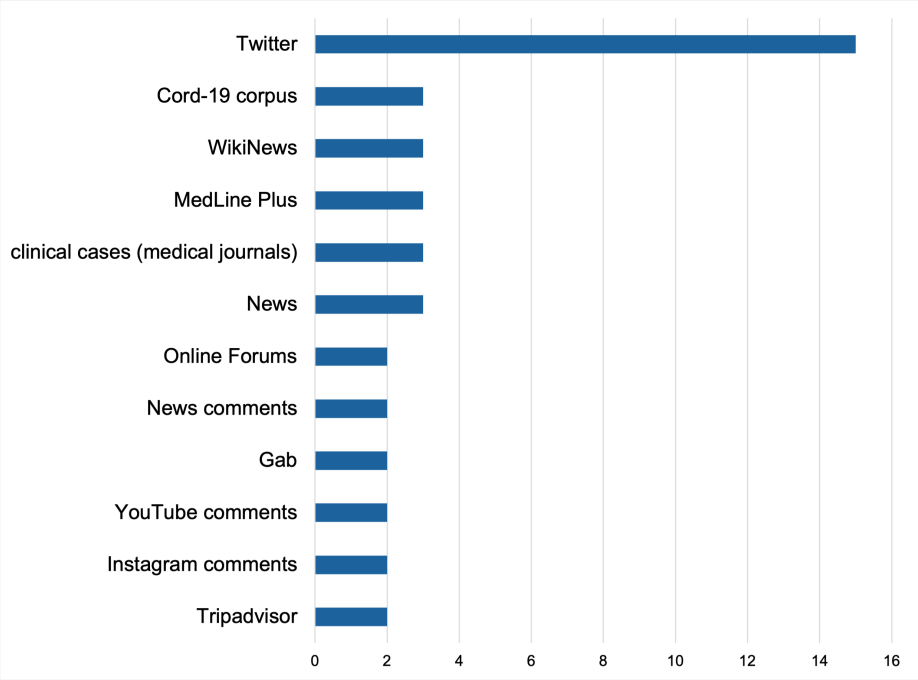
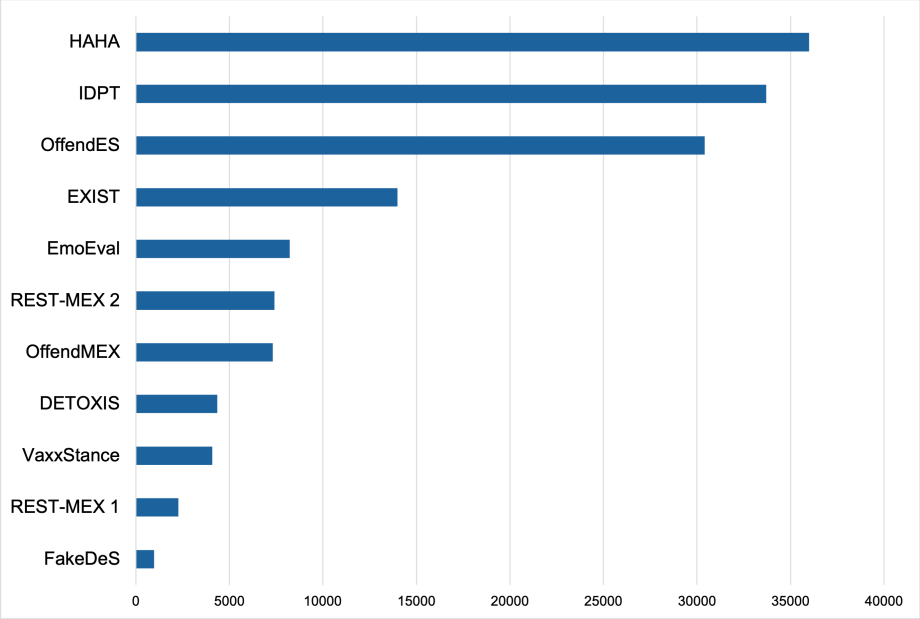


Figura 5. Types of textual sources in IberLEF 2021 tasks.



additional sources used by two or three tasks: news and news comments, medical sources, material from other social networks such as YouTube, Instagram, TripAdvisor and Gab, etc.

Figura 6. Dataset sizes in IberLEF 2021 classification tasks.



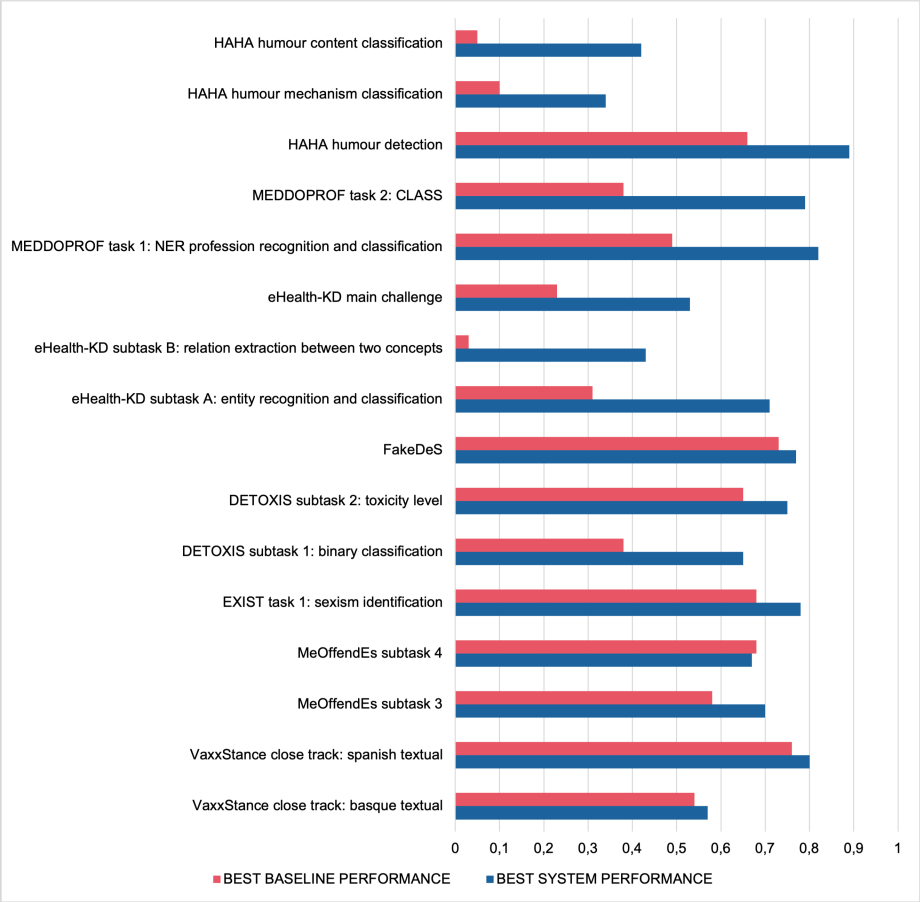
In terms of **dataset sizes** and annotation efforts, it is difficult to establish fair comparisons, because of the diversity of text sizes and the wide variance in terms of annotation difficulty. Figure 6 compares dataset sizes for the classification tasks, where it is more reasonable to establish direct comparisons.

Overall the annotation effort in IberLEF 2021 is remarkable, and it is a significant contribution to enlarge test collections at least for Spanish; and, therefore, to enable significant advances in our field for Spanish and the other languages involved. The number of documents varies substantially, from over 35,000 tweets (HAHA dataset on humour) to 971 news stories for FakeDes (fake news detection). But again, direct comparisons are not fair: for instance, in the case of HAHA, they are expanding annotations on a previously existing dataset (developed in other HAHA editions); and, on the other hand, establishing whether a piece of news is fake or real is probably much more time consuming than classifying humor in tweets.

IberLEF 2021 has been carried out without funding sources (other than those obtained individually by the teams organizing and participating in the tasks). If the IberLEF organization could directly fund the task organizers, this would

probably help reaching large and high quality annotations for all of the tasks accepted each year.

Figura 7. Performance of best systems versus baselines in IberLEF 2021 classification tasks.



In terms of **progress with respect the state of the art**, it is really difficult to extract aggregated conclusions for the whole IberLEF effort. In Figure 7 we display a pairwise comparison between the best system and the best baseline, for each of the tasks where at least one baseline is provided, and with respect to the official ranking metric used in each task. To avoid confusion, we have restricted the chart to tasks where the official metric varies between 0 (worst quality) and 1 (perfect output). Still, it is difficult to extract conclusions, because the effort put by task organizers in providing state-of-the-art baselines varies considerably between tasks. We can say, however, that in a few cases improving the baseline

has proved to be challenging, and there is one case (MeOffendEs subtask 4) where the baseline beats the best system (by a narrow margin). It would probably beneficial for future IberLEF editions to establish some minimum guidelines about the types of baselines to expect in every task; again, this would be easier to implement with dedicated funding.

3.3. Participation

Given that IberLEF 2021 was not a funded initiative, participation has been impressive, with a large fraction of current research groups interested in NLP for Spanish organizing and/or participating in one or more tasks. Overall, 359 researchers representing 173 research groups from 22 countries in Europe, Asia and the Americas were involved in IberLEF tasks.

Figura 8. Number of groups participating in IberLEF 2021 tasks per country

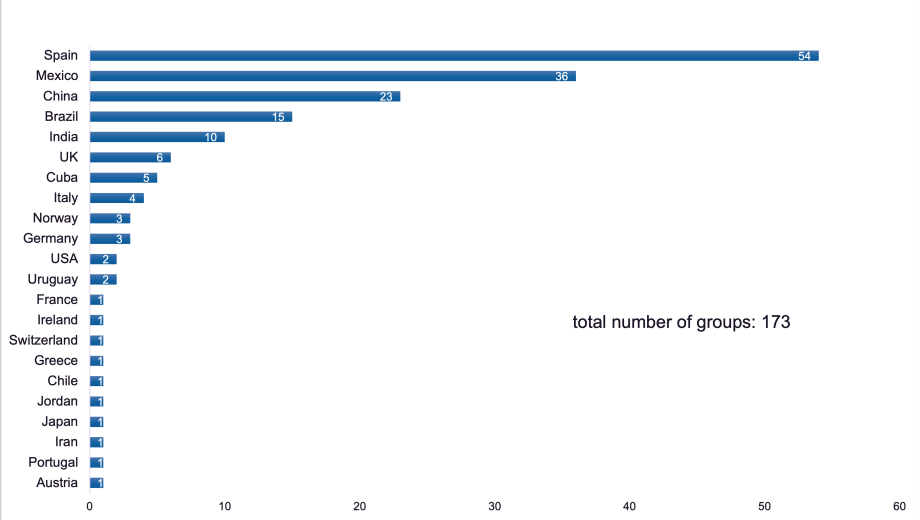
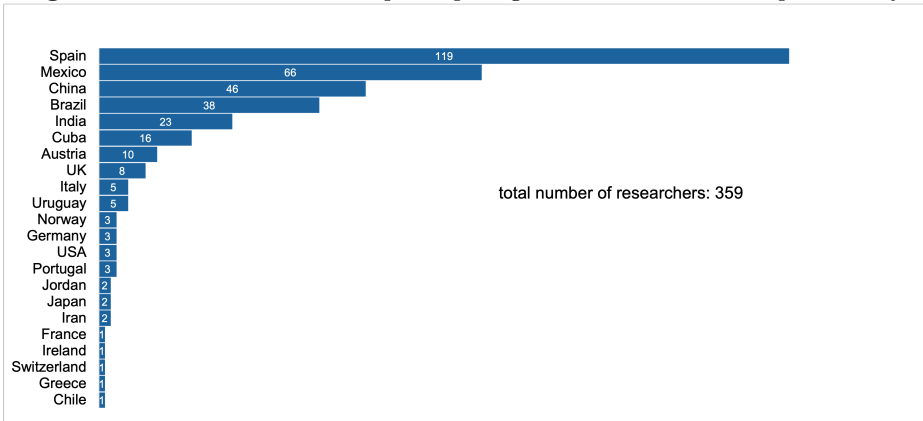


Figure 8 shows the distribution of research groups per country. Unsurprisingly, Spain has the largest representation, with 54 groups (note that all figures reporting participation do not collapse duplicates: a group or a researcher participating in two tasks is counted twice). Mexico has also a remarkable participation, with 36 groups. And there is a suprisingly large number of countries where no Iberian language is used but have at least two groups representing them: China (23 groups), India (10), UK (6, but note that some tasks do use English as an additional source language), Italy (4), Norway (3) and Germany (3).

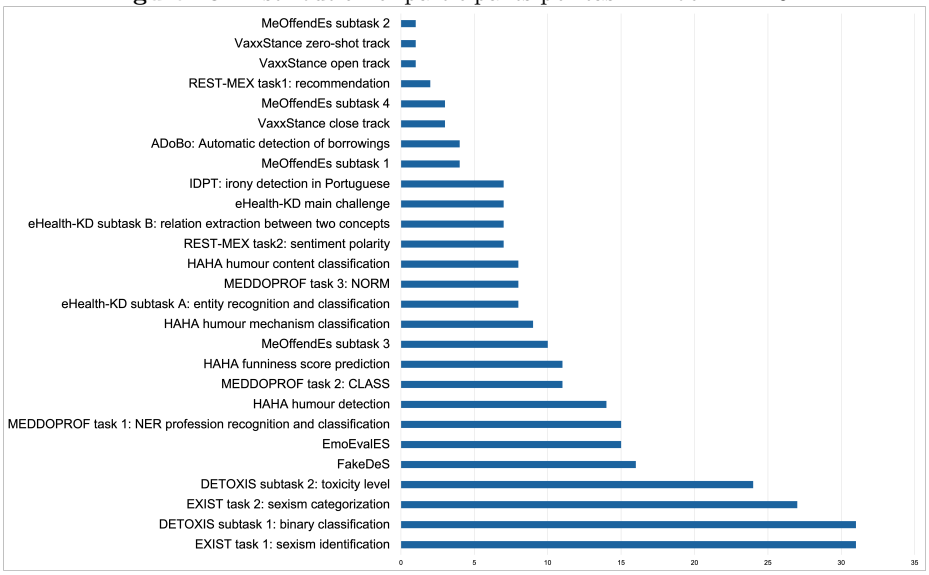
Figure 9 shows the distribution of researchers (appearing as authors in the working notes) per country. The numbers are consistent with the distribution of groups per country, with Spain, Mexico, China, Brazil and India (the top

Figura 9. Number of researchers participating in IberLEF 2021 tasks per country.



five) representing roughly 80 % of the researchers involved. The fact that two countries in the top five, China and India, appear in the top five indicates two things: first, that Spanish attracts the attention of the NLP community at large; and second, that current NLP technologies enable processing dataset without language-specific machinery, other than pretrained language models made available to the research community.

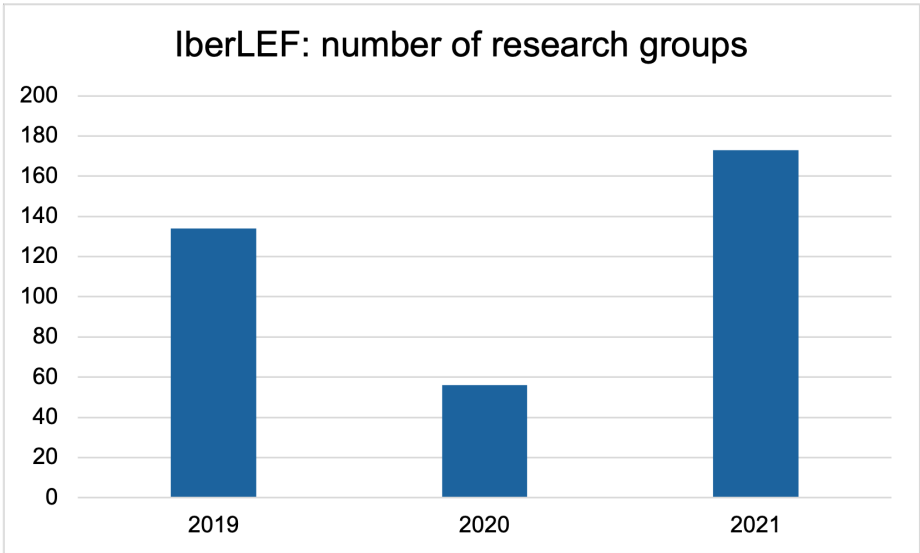
Figura 10. Distribution of participants per task in IberLEF 2021.



The distribution of research groups per task is shown in Figure 10. Participation ranges between 31 groups (EXIST subtask 1 and DETOXIS subtask 1) and one group (MeOffendES subtask 2, VaxxStance zero-shot track and VaxxStance open track). As in other evaluation initiatives, participation seems to be driven not only by the task intrinsic interest, but also by the cost of entry: in general, classification tasks (the most basic machine learning task, for which more plug and play software packages exist) receive more participation than tasks which require more elaborated approaches and more creativity to assemble algorithmic solutions. In the middle of the table we can find most tasks in the medical domain, which attract many groups in spite of being (in general) highly challenging.

Figure 11 shows how participation has evolved in time; while 2020 was a difficult year with the irruption of COVID-19, in 2021 participation has grown considerably, with 173 groups (three times larger than in 2020 and a 30 % increase with respect to 2019). The number of countries involved has also grown from 18 to 22.

Figura 11. Number of research groups participating in IberLEF across time.



4. Conclusions

In its third edition, 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.

Acknowledgements

The authors of this overview have been supported by the Spanish Government, Ministry of Science and Innovation, via research grants MISMIS (PGC2018-096212-B), MISMIS-BIAS (PGC2018-096212-B-C32) and MISMISFAKEnHATE (PGC2018-096212-B-C31); and by CONACyT-Mexico project CB-2015-01-257383 and the thematic networks program (Language Technologies Thematic Network).

Referencias

1. Agerri, R., Centeno, R., Espinosa, M., de Landa, J.F., Álvaro Rodrigo: Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. Procesamiento del Lenguaje Natural **67**, 173–181 (2021)
2. Amigo, E., Gonzalo, J., Mizzaro, S., Carrillo-de Albornoz, J.: An effectiveness metric for ordinal classification: Formal properties and experimental results. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3938–3949. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.363, https://aclanthology.org/2020.acl-main.363
3. del Arco, F.M.P., Casavantes, M., Escalante, H.J., Martín-Valdivia, M.T., Montejo-Ráez, A., y Gómez, M.M., Jarquín-Vásquez, H., Villaseñor-Pineda, L.: Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants **67**, 183–194 (2021)
4. del Arco, F.M.P., Jiménez-Zafra, S.M., Montejo-Ráez, A., Molina-González, M.D., L. Alfonso Ureña-López, M.T.M.V.: Overview of the emoevales task on emotion detection for spanish at iberlef 2021. Procesamiento del Lenguaje Natural **67**, 155–161 (2021)
5. Álvarez Carmona, M., Aranda, R., Arce-Cardenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A.P., Martínez-Miranda, J., Pérez-Espinosa, H., Rodríguez-González, A.Y.: Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism **67**, 163–172 (2021)
6. Chiruzzo, L., Castro, S., Góngora, S., Rosá, A., Meaney, J.A., Mihalcea, R.: Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. Procesamiento del Lenguaje Natural **67**, 257–268 (2021)
7. Corrêa, U.B., Coelho, L., Santos, L., de Freitas, L.A.: Overview of the idpt task on irony detection in portuguese at iberlef 2021. Procesamiento del Lenguaje Natural **67**, 269–276 (2021)

8. Gómez-Adorno, H., Posadas-Durán, J.P., Enguix, G.B., Porto, C.: Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task. Procesamiento del Lenguaje Natural **67**, 223–231 (2021)
9. Lima-López, S., Farré-Maduell, E., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classiffication and normalization of professions and occupations from medical texts. Procesamiento del Lenguaje Natural **67**, 243–256 (2021)
10. Álvarez Mellado, E., Anke, L.E., Arroyo, J.G., Lignos, C., Zamorano, J.P.: Overview of adobo 2021: Automatic detection of unassimilated borrowings in the spanish press. Procesamiento del Lenguaje Natural **67**, 277–285 (2021)
11. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.A., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez-Zafra, S.M., Lima, S., Plaza-del-Arco, F.M., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
12. Piad-Morffis, A., Estevez-Velarde, S., Gutierrez, Y., Almeida-Cruz, Y., Montoyo, A., Muñoz, R.: Overview of the ehealth knowledge discovery challenge at iberlef 2021. Procesamiento del Lenguaje Natural **67**, 233–242 (2021)
13. Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Language and Law / Linguagem e Direito **5**(2), 80–102 (2018), https://ojs.letras.up.pt/index.php/LLLD/article/view/6119
14. Ródriguez-Sánchez, F., de Albornoz, J.C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural **67**, 195–207 (2021)
15. Taulé, M., Ariza, A., Nofre, M., Amigó, E., Rosso, P.: Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. Procesamiento del Lenguaje Natural **67**, 209–221 (2021)