

# Overview of IberLEF 2022: Natural Language Processing Challenges for Spanish and other Iberian Languages

Julio Gonzalo<sup>1</sup>, Manuel Montes-y-Gómez<sup>2</sup> and Francisco Rangel<sup>3</sup>

<sup>1</sup>*nlp.uned.es, ETSI Informática de la UNED, Madrid, Spain*

<sup>2</sup>*National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico*

<sup>3</sup>*Symanto Research, Valencia, Spain*

## Abstract

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 2022, which included 10 tasks and 19 subtasks dealing with sentiment, stance and opinion analysis, detection and categorization of harmful content, Information Extraction, Paraphrase Identification, and Question Answering. Overall, IberLEF activities were a remarkable collective effort involving 310 researchers from 24 countries in Europe, Asia, Africa, Australia and the Americas.

## Keywords

Natural Language Processing, Artificial Intelligence, Evaluation, Evaluation Challenges

## 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 ten tasks dealing with sentiment, stance and opinion analysis, detection and categorization of harmful content, Information Extraction and Answer Extraction, and Paraphrase Identification. Overall, IberLEF activities were a remarkable collective effort involving 310 researchers from 24 countries in Europe, Asia, Africa, Australia and the Americas. Papers with system descriptions are included in this IberLEF 2022 Proceedings volume, and papers with task overviews are published in the journal *Procesamiento del Lenguaje Natural*, vol. 69 (September 2022 issue).

---

*IberLEF 2022, September 2022, A Coruña, Spain*

✉ [julio@lsi.uned.es](mailto:julio@lsi.uned.es) (J. Gonzalo); [mmontesg@inaoep.mx](mailto:mmontesg@inaoep.mx) (M. Montes-y-Gómez); [kico.rangel@gmail.com](mailto:kico.rangel@gmail.com) (F. Rangel)


🌐 <https://nlp.uned.es/> (J. Gonzalo); <https://ccc.inaoep.mx/~mmontesg/> (M. Montes-y-Gómez);

<https://kicorangel.com/> (F. Rangel)

🆔 0000-0002-5341-9337 (J. Gonzalo); 0000-0002-7601-501X (M. Montes-y-Gómez); 0000-0002-6583-3682 (F. Rangel)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In this paper we summarize the activities carried on in IberLEF 2022, extracting some aggregated figures for a better understanding of this collective effort.

## 2. IberLEF 2022 Tasks

These are the ten tasks successfully run in 2022, grouped thematically:

### 2.1. Sentiment, Stance and Opinions

**ABSAPT** [1] is an aspect-based sentiment analysis task in Portuguese, which used TripAdvisor reviews as target texts. It included (i) a subtask on aspect term extraction devoted to identification of aspects in reviews, and (ii) a subtask on sentiment orientation (polarity) identification about a single aspect mentioned in the review.

**PoliticES** [2] is an author profiling task on Twitter accounts for Spanish politicians and political journalists, where systems must extract gender, profession and political spectrum of each profile.

**Rest-Mex** [3] is a task that works with Mexican Tourist Texts, and addresses three problems: (i) a recommendation subtask where, given a TripAdvisor user and a Mexican tourist destination, the system must predict the degree of satisfaction (1-5) that the user will have when visiting the destination; (ii) a sentiment analysis task where the system must predict the polarity (1-5) of a given TripAdvisor review, and also the type of destination (hotel, restaurant, attraction); (iii) an epidemiological semaphore prediction task, where given covid-related news of a Mexican region, systems must predict the semaphore color of weeks 0, 2, 4 and 8 in the future.

### 2.2. Harmful Content

**DA-VINCIS** [4] is a task where systems must detect and classify tweets (in Spanish) that report violent incidents. It included two subtasks; the first one is a binary classification task in which users had to determine whether tweets were associated to a violent incident or not, and the second one is a multi-label classification task in which the category of the violent incident should be spotted.

**DETESTS** [5] is a task where systems must detect and classify racial stereotypes in comments to online news articles written in Spanish. Subtask 1 is stereotype detection, and systems must identify whether the comment contains at least one stereotype or not. Manual annotations are handled following the learning with disagreement paradigm, where there is not necessarily a single correct label for every example in the dataset. Subtask 2 is a multi-label hierarchical classification problem where systems must detect and classify stereotypes according to this set of categories: victims of xenophobia, suffering victims, economic resources, migration control, cultural and religious differences, people which takes “benefits” of our social policy, problem of public health, security threat, dehumanization, other.

**EXIST** [6] is a task where systems must detect and classify sexist content in Spanish and English tweets and gabs. Task 1 is about identification of sexism-related content: a tweet is positive if it is sexist itself, describes a sexist situation or criticizes a sexist behavior. Task 2 is about sexism categorization: once a message has been classified as sexist, systems must

classify positive tweets in the following categories: ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny and non-sexual violence.

### 2.3. Information Extraction and Paraphrase Identification

**LivingNER** [7] is a task on named entity recognition, normalization and classification of species, pathogens and food. Source texts are medical documents (case reports) annotated by medical experts using the NCBI taxonomy. In Task 1, LivingNER-Species NER track, systems must find all mentions to (human or non-human) species mentioned, such as “hepatitis B”, “virus herpes simple”, “paciente”. In Task 2, LivingNER-Species Norm track, systems have to retrieve all species mentions together with their corresponding NCBI taxonomy concept identifiers. And in Task 3, LivingNER-Clinical Impact track, for each text systems must (i) detect if the text contains information relevant to real-world clinical use cases of high impact; (ii) retrieve the list of NCBI taxonomy identifiers that support such detections; categorize the documents in the following information axes: pets and farm animals, animal causing injuries, food species, and nosocomial entities.

**PAR-MEX** [8] is a paraphrase identification task. Systems must detect sentence-level paraphrase identification in Mexican Spanish food-related texts, which have been manually generated from an original set of texts using *literary creation*, *low paraphrase*, *high paraphrase* and *no paraphrase* methods.

### 2.4. Question Answering and Machine Reading

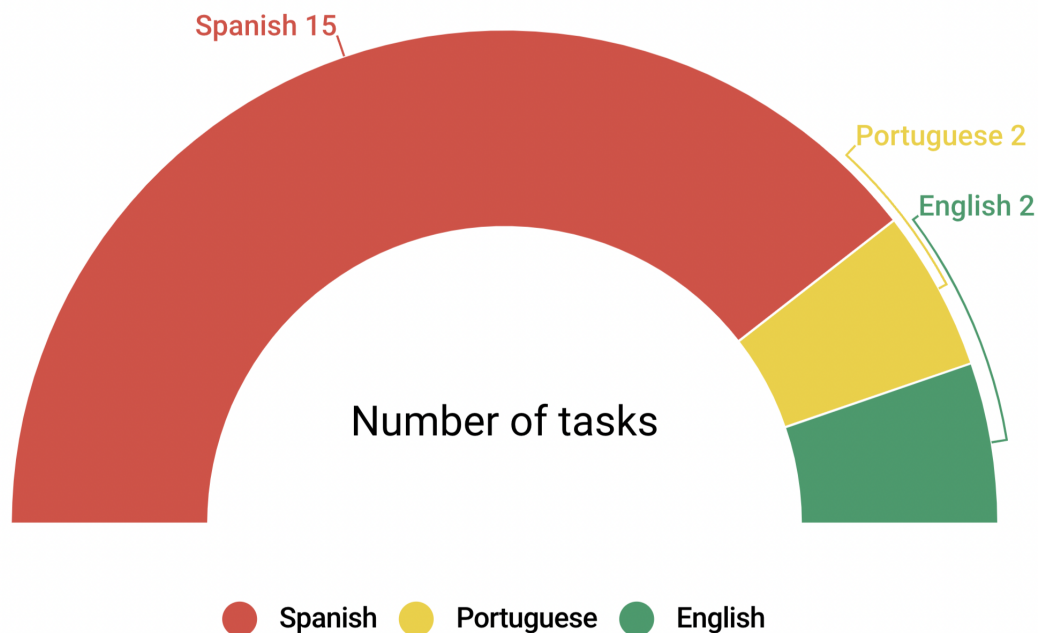
**QuALES** [9] is a Question Answering task where answers must be extracted from news articles written in Spanish. The input for systems is a question and a piece of news, and the system must find the shortest spans of text in the article (if there is any) that answer the question. Most questions (but not all) in the dataset deal with covid-19 issues.

**ReCoRES** [10] is a Reading Comprehension and Reasoning Explanation task for Spanish. Given a passage and a question about its content, Reading Comprehension systems must (1) select the correct answer from a given set of candidates (multiple-choice task); and (2) provide an explanation for why a given candidate was chosen as answer (reasoning explanation). Texts in this dataset are based on university entrance examinations, and explanations are evaluated according to automatic similarity estimations with respect to manual reference explanations, and with manual assessments of their accuracy, fluency and readability.

## 3. Aggregated Analysis of IberLEF 2022 Tasks

### 3.1. Tasks characterization

In terms of **languages**, the distribution per tasks (including subtasks) is shown in Figure 1. Spanish is, one more year, the central language of IberLEF (17 tasks) with Portuguese and English in a secondary role (2 tasks each). Main Spanish variants considered are those from Spain, Mexico, Uruguay and Perú.

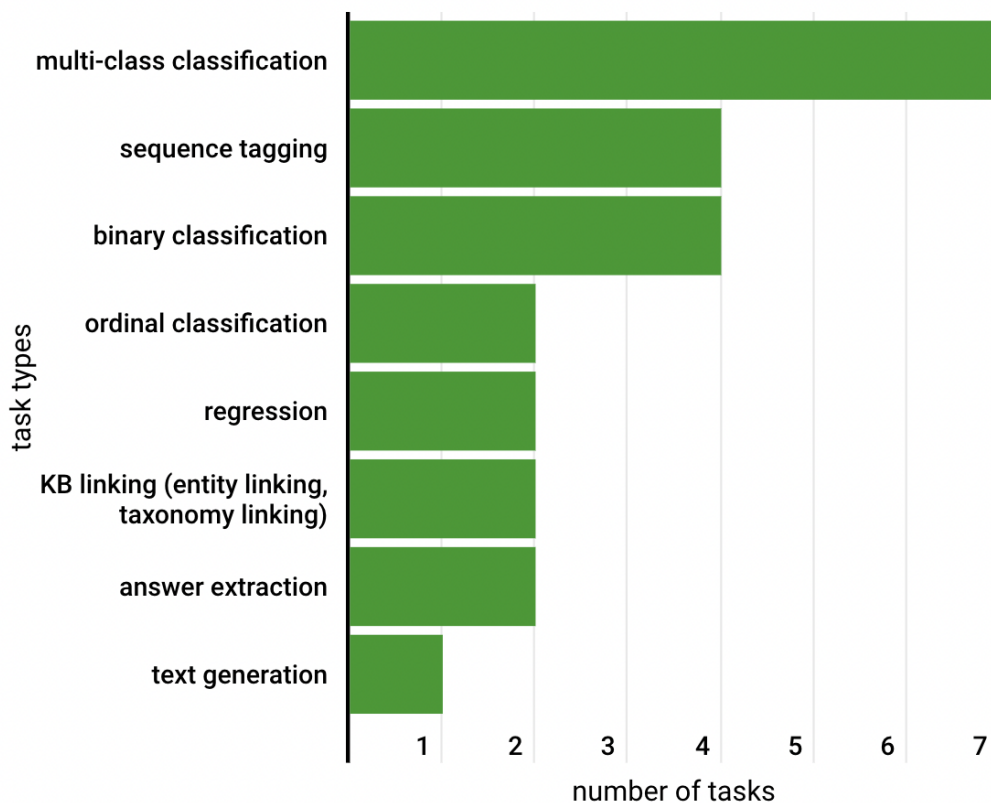


**Figure 1:** Distribution of languages in IberLEF 2022 tasks

In terms of **abstract task types**, the distribution of tasks can be seen in Figure 2. Out of a total of 19 tasks (each subtask is counted as a task here), the most popular type of task is multi-class classification (7 tasks), followed by sequence tagging and binary classification (4 each). There are also two ordinal classification tasks, two regression tasks, two KB linking tasks (one on entity linking and another one on taxonomy linking), two answer extraction tasks (one is multiple choice, which is also counted as classification, and the other one is span selection, which we also count as sequence tagging) and one text generation task. Interestingly, in 2022 there are four complex tasks which involve solving more than one core task at once (for instance, sequence tagging plus entity linking).

Compared with 2021, the trends are towards a less numerous (19 vs 29) but more diverse and more complex set of tasks, where binary classification is no longer the most popular type of task and several tasks imply solving many NLP problems at the same time.

In terms of **evaluation metrics**, the distribution can be seen in Figure 3, which depicts only the main metrics used to rank systems in each task. As in previous years, there is a remarkable predominance of F1 (11 tasks), even if it does not perfectly match the problem considered. Accuracy is used by three tasks, MAE in two regression tasks, and there are other six metrics that are used only in one task. Some of them correspond to the complex tasks which embed subtasks (e.g. the mean of F1 scores for several tasks is used in one occasion, the mean of inverse MAE and F1 scores for different tasks in other, or the average of F1 measures at different points

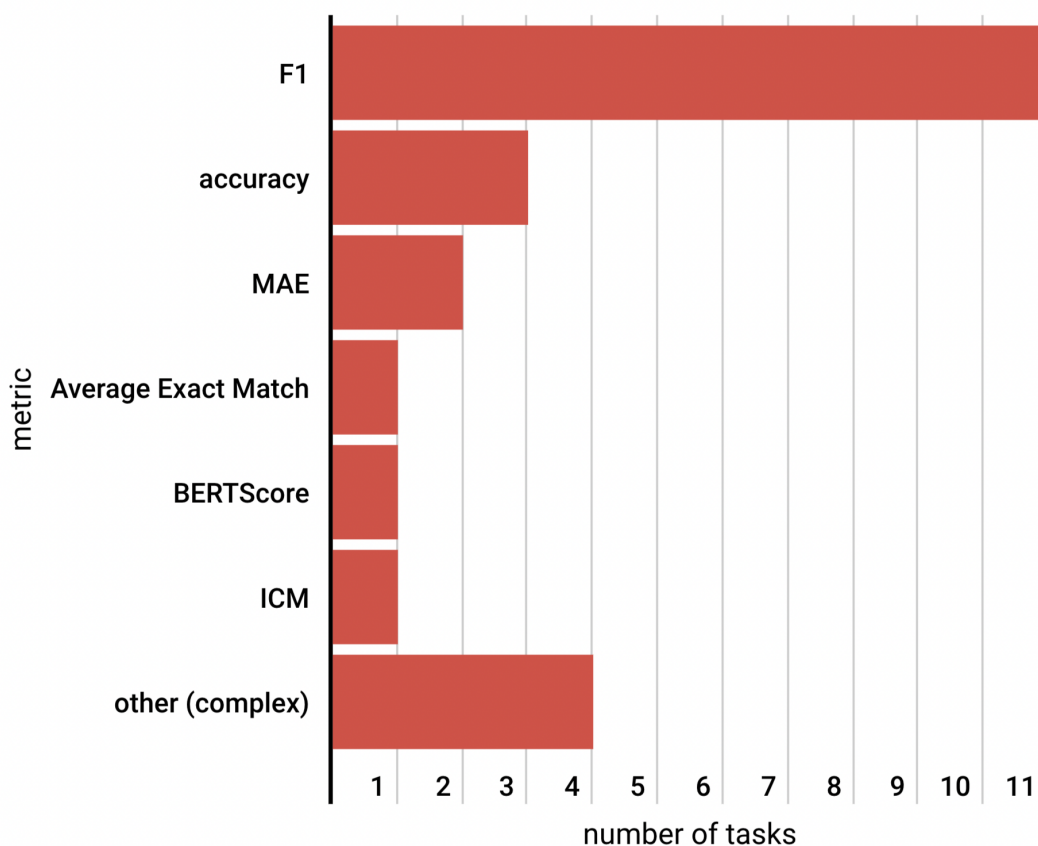


**Figure 2:** Distribution of IberLEF 2022 tasks per abstract task type.

in the future with weights according to the time distance in other. The rest are Average Exact Match [11] for a QA task, BERTScore [12] to compare system and gold standard explanations, and ICM [13] for a hierarchical classification task.

Overall, in IberLEF as in other NLP competitive evaluation challenges we might still be relying too much on averages to combine different quality metrics: it has been common this year to combine F1 measures (which are harmonic averages) with other measures using some other form of averaging. This hides the actual behaviour of systems and give usually no clues on how to improve them. Also, again in 2022 the choice of metrics is, in general, barely justified, particularly in terms of how the system output is going to be used in realistic usage scenarios.

Finally, in terms of novelty/stability IberLEF 2022 has brought many new problems, with seven out of the 10 primary tasks being new this year. Only REST-MEX, EXIST and DETESTS had also been run in 2021.



**Figure 3:** Distribution of official evaluation metrics in IberLEF 2022 tasks.

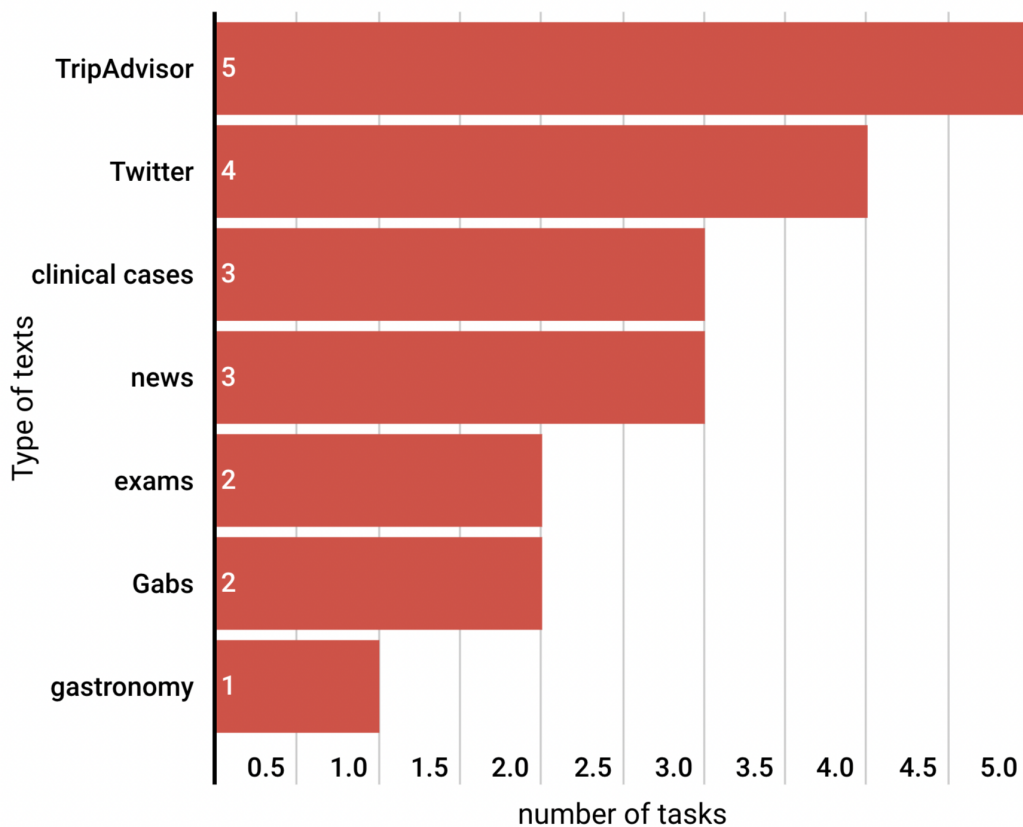
### 3.2. Datasets and results

In terms of **types of textual sources**, Figure 4 shows how they are used in IberLEF 2022 tasks. There is more diversity than previous years, with Twitter being less dominant: TripAdvisor reviews were used in 5 tasks, Twitter in 4, clinical cases in 3 subtasks (all belonging to the same task), exams, news comments and Gabs were used in two subtasks each, and finally news and gastronomy texts were used in one task each.

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. In any case, in the majority of cases (14 tasks) manually annotated datasets were below 6,000 instances. Two other tasks provided annotated collections comprising between 10,000 and 15,000 instances, and there was one task which provided over 40,000 annotated instances.

As for the reliability of the annotations, one useful indicator is inter-annotator agreement, which is reported in 9 out of 19 tasks. In the tasks where it is reported, annotator agreement is high in three cases and mid-low in another six. In general, mid-low agreement indicates the complexity of the task rather than poor annotation guidelines.



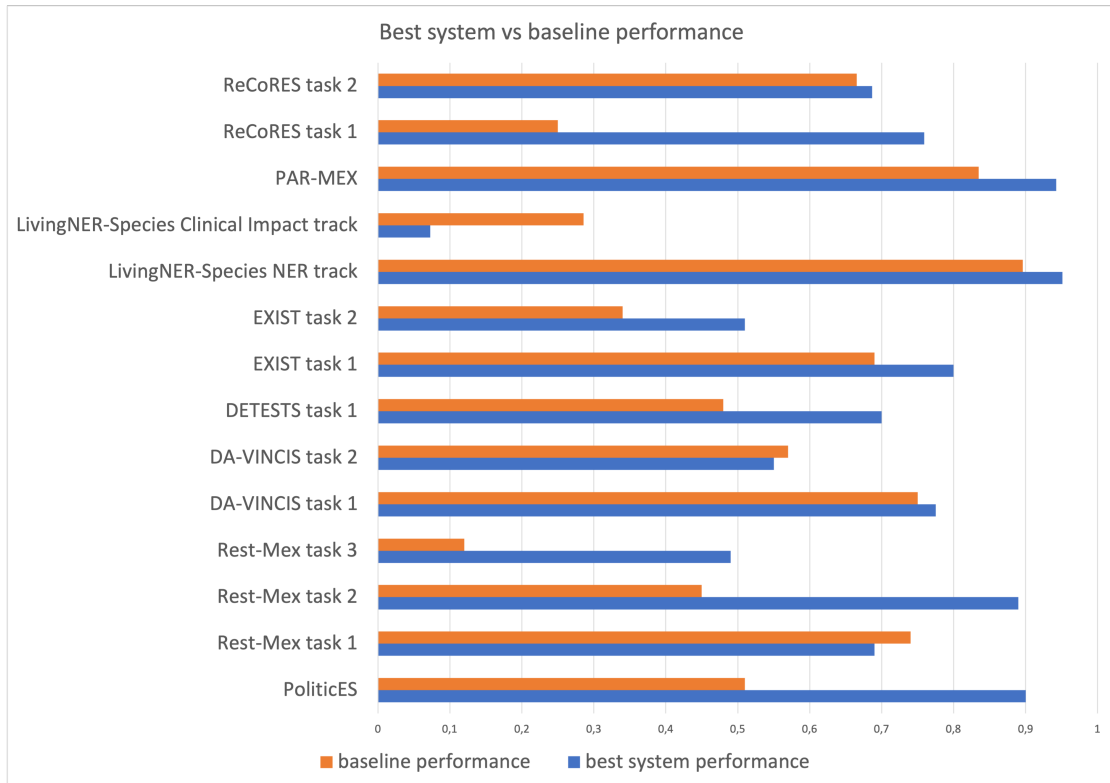


**Figure 4:** Types of textual sources in IberLEF 2022 tasks.

Overall the annotation effort in IberLEF 2022 keeps being a remarkable contribution to enlarge test collections for Spanish (and, less prominently, other languages). One more year, IberLEF has been carried out without specific funding sources (other than those obtained individually by the teams organizing and participating in the tasks). A centralized funding schema could certainly help reaching larger and better annotations in IberLEF as a whole.

In terms of **progress with respect the state of the art**, it is as usual difficult to extract aggregated conclusions for the whole IberLEF effort, in particular given the diversity of approaches for providing task baselines: in five tasks, no baseline was provided. In three, only a trivial baseline was included in the comparisons (e.g. majority class or random baselines in classification). Four tasks used SVM as baseline, and five used some variant of transformers (BETO in two occasions, BERT in another two and T5 in one). Only two used other types of baselines.

In the tasks that used baselines, the baseline was beaten (by a margin larger than 5%) by the best system in eight cases. In two cases, the difference was below 5% (one in favour of the best system, the other in favour of the baseline), and in the last two tasks, the baseline was better than any system. This is an indication that at least some of the tasks



**Figure 5:** Performance of best systems versus baselines in IberLEF 2022 tasks. Only tasks with official evaluation metrics in the range [0-1] that include at least a baseline system are included in this graph.

In Figure 5 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).

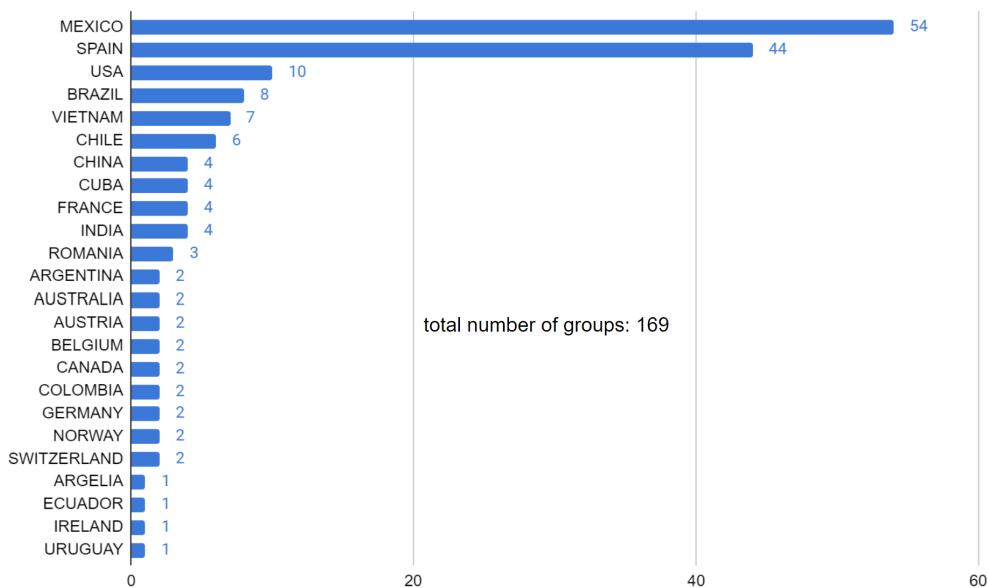
### 3.3. Participation

Given that IberLEF 2022 was not a funded initiative, participation has again 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, 310 researchers representing 169 research groups from 24 countries in Europe, Asia, Africa, Australia and the Americas were involved in IberLEF tasks<sup>1</sup>.

Figure 6 shows the distribution of research groups per country. This year, Mexico has the largest representation, with 54 groups, followed by Spain with 44 groups (note that all figures

<sup>1</sup>Statistics have been compiled from the submitted working notes, meaning two things: *i*) some groups and researchers may be counted twice if they have participated in more than one task; *ii*) real participation may be higher due to the number of teams who submitted runs but did not submit their working notes afterwards, and thus have not been counted in the statistics.



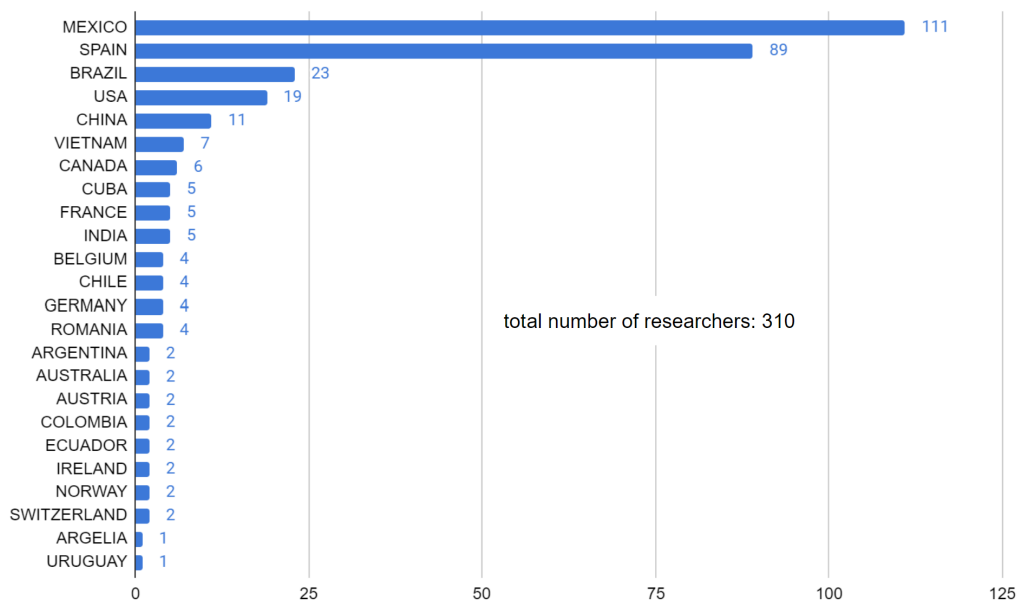


**Figure 6:** Number of research groups participating in IberLEF 2022 tasks per country.

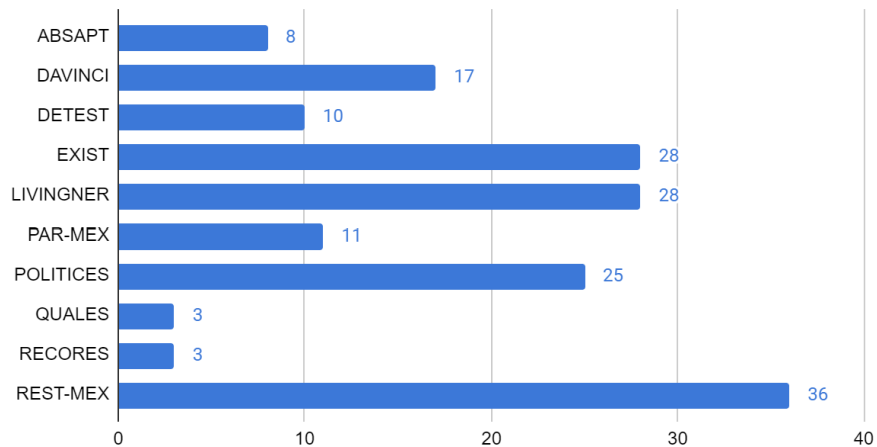
reporting participation do not collapse duplicates: a group or a researcher participating in two tasks is counted twice).

Figure 7 shows the distribution of researchers (appearing as authors in the working notes) per country. The numbers are almost consistent with the distribution of groups per country, with some flips between USA and Brazil, or China, Chile and Vietnam. The top five, with Mexico, Spain, Brazil, USA, and China, represents roughly 80% of the researchers involved. The fact that there are two non-Spanish, non-Portuguese speaking countries in the top five, China and the USA, as well as others such as Vietnam or Canada in the top positions in terms of participation, indicates two things: first, that Spanish attracts the attention of the NLP community at large; and second, that current NLP technologies enable addressing different languages without language-specific machinery, other than pre-trained language models made available to the research community.

The distribution of research groups per task is shown in Figure 8. Participation ranges between 3 and 36 groups. As in other evaluation initiatives, participation seems to be driven not only by the task intrinsic interest, but also by the cost of entry: as usual, 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.



**Figure 7:** Number of researchers participating in IberLEF 2022 tasks per country.



**Figure 8:** Distribution of participant groups per task in IberLEF 2022. The figure displays the number of groups that submitted at least one run.

## 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, comprising 10 main tasks and 310 researchers involved, from institutions in 24 countries in Europe, Asia, Africa, Australia and the Americas. IberLEF 2022 has been one of the most diverse in terms of types of tasks and

application domains, and has contributed to advance the field in the areas of sentiment, stance and opinion analysis, detection and categorization of harmful content, Information Extraction, Answer Extraction, and Paraphrase identification. In a field where machine learning is the ubiquitous approach to solve challenges, the definition of research challenges, the development of high-quality test collections that allow for iterative evaluation and the design of sound evaluation methodologies and metrics are perhaps the most critical aspects of research, and we believe IberLEF keeps making significant contributions to all of them.

## Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation, Project *FairTransNLP* (PID2021-124361OB-C32), and by CONACyT-México, Project CB-2015-01-257383. The work of the third author has been partially funded by CDTI under grant IDI-20210776, IVACE under grant IMINOD/2021/72, and grant PLEC2021-007681 funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

## References

- [1] F. L. V. da Silva, G. d. S. Xavier, H. M. Mensenburg, R. F. Rodrigues, L. P. dos Santos, R. M. Araújo, U. B. Corrêa, L. A. de Freitas, ABSAPT 2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese 69 (2022).
- [2] J. A. García-Díaz, S. M. Jiménez-Zafra, M.-T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology, *Procesamiento del Lenguaje Natural* 69 (2022).
- [3] A. Álvarez Carmona, Miguel A. and Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of Rest-Mex at IberLEF 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts, *Procesamiento del Lenguaje Natural* 69 (2022).
- [4] L. J. Arellano, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, F. Sanchez-Vega, Overview of DA-VINCIS at IberLEF 2022: Detection of Aggressive and Violent Incidents from Social Media in Spanish 69 (2022).
- [5] A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of DETESTS at IberLEF 2022: DETECTION and classification of racial STereotypes in Spanish, *Procesamiento del Lenguaje Natural* 69 (2022).
- [6] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sEXism Identification in Social neTworks, *Procesamiento del Lenguaje Natural* 69 (2022).
- [7] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources, *Procesamiento del Lenguaje Natural* 69 (2022).
- [8] G. Bel-Enguix, G. Sierra, H. Gómez-Adorno, J.-M. Torres-Moreno, J.-G. Ortiz-Barajas,

- J. Vázquez, Overview of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task, *Procesamiento del Lenguaje Natural* 69 (2022).
- [9] A. Rosá, L. Chiruzzo, L. Bouza, A. Dragonetti, S. Castro, M. Etcheverry, S. Góngora, S. Goycochea, J. Machado, G. Moncecchi, J. J. Prada, D. Wonsever, Overview of QuALES at IberLEF 2022: Question Answering Learning from Examples in Spanish, *Procesamiento del Lenguaje Natural* 69 (2022).
- [10] M. A. Sobrevilla Cabezudo, D. Diestra, R. López, E. Gómez, A. Oncevay, F. Alva-Manchego, Overview of ReCoRES at IberLEF 2022: Reading Comprehension and Reasoning Explanation for Spanish, *Procesamiento del Lenguaje Natural* 69 (2022).
- [11] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, 2016. URL: <https://arxiv.org/abs/1606.05250>. doi:10.48550/ARXIV.1606.05250.
- [12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, 2019. URL: <https://arxiv.org/abs/1904.09675>. doi:10.48550/ARXIV.1904.09675.
- [13] E. Amigo, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: <https://aclanthology.org/2022.acl-long.399>. doi:10.18653/v1/2022.acl-long.399.