Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 155-161 recibido 05-07-2021 revisado 12-07-2021 aceptado 16-07-2021

Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021

Resumen de la tarea de detección de emociones en español EmoEvalEs en IberLEF 2021

Flor Miriam Plaza-del-Arco, Salud María Jiménez-Zafra, Arturo Montejo-Ráez, M. Dolores Molina-González, L. Alfonso Ureña-López, M. Teresa Martín-Valdivia
Computer Science Department, SINAI, CEATIC
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, sjzafra, amontejo, mdmolina, laurena, maite}@ujaen.es

Abstract: This paper presents the EmoEvalEs shared task, organized at IberLEF 2021, as part of the 37th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021). The aim of this task is to promote the *Emotion detection and Evaluation for Spanish*. It consists of a fine-grained emotion classification of tweets from the EmoEvalEs corpus in one of these seven classes: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, or *others*. In this edition, 70 teams registered, 15 submitted results and 11 presented papers describing their systems. Most teams experimented with neural networks, being Transformers the most widely used model. It should be noted that few of them also considered the features of offensiveness and event that were provided in the corpus apart from the tweet texts.
Keywords: EmoEvalEs, emotion detection, natural language processing.

Resumen: Este artículo presenta la tarea EmoEvalEs, organizada en IberLEF 2021, en el marco del de la 37 edición de la Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. El objetivo de esta tarea es promover la *Detección y Evaluación de Emociones en Español*. Consiste en la clasificación de grano fino de los tweets del corpus EmoEvent en una de las siguientes siete clases: *ira*, *asco*, *miedo*, *alegría*, *tristeza*, *sorpresa* u *otros*. En esta edición, se registraron 70 equipos, 15 enviaron resultados y 11 presentaron artículos describiendo sus sistemas. La mayoría de los equipos experimentaron con redes neuronales, siendo *Transformers* el modelo más utilizado. Cabe destacar que pocos equipos consideraron también las características de ofensividad y evento que se proporcionaron en el corpus aparte de los textos de los tweets.
Palabras clave: EmoEvalEs, detección de emociones, procesamiento del lenguaje natural.

1 Introduction

Emotion detection from text is a research task in Natural Language Processing (NLP) aimed at classifying words, phrases, or documents into predefined emotion categories. This task can be considered an extension of the polarity classification task due to the presence of fine-grained categories based on fundamental emotion theories, with the Ekman (Ekman, 1992) and Plutchik (Plutchik, 2001) models the most commonly cited ones.

In the last years, great efforts have been made to address one of the most well-known Sentiment Analysis tasks, polarity classification. In fact, a number of shared tasks have been accomplished (Rosenthal, Farra, and Nakov, 2017; Díaz-Galiano et al., 2019; Martínez-Cámara et al., 2018). However, emotion classification is still considered a challenge for the NLP systems and its significance has increased in recent years.

With the aim of promoting research in the emotion analysis area in Spanish, the "Emotion Detection" task was introduced for the first time in the TASS workshop (Vega et al., 2020) at IberLEF 2020. This year we continued with the "Emotion detection and Evaluation for Spanish Shared Task" (EmoEvalEs) at IberLEF 2021 (Montes et al., 2021). Although only two teams participated in TASS

ISSN 1135-5948. DOI 10.26342/2021-67-13

2020, this edition has attracted 70 team registrations and received 11 system description papers, which demonstrates the interest of the research community in this topic.

With the "Emotion detection and Evaluation for Spanish Shared Task" (EmoEvalEs), participants had the challenge of classifying the emotion expressed in tweets related to certain events in seven categories: *anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise* and *others*. Unlike the previous edition, this year, two new features in dataset have been provided to participants: the event corresponding to the domain associated with the tweet and whether the tweet expresses offensiveness. The competition was organized through CodaLab and is accessible at the following link: `https://competitions.codalab.org/competitions/28682`.

The remainder of this paper is organized as follows: Section 2 describes the EmoEvalEs shared task. Section 3 presents the dataset that the participants used in the competition. Sections 4 and 5 present the approaches and results of the participants. Finally, Section 6 concludes and suggests some possible directions for future work.

2 Task description

Understanding the emotions expressed by users on social media is a hard task due to the absence of voice modulations and facial expressions. Our shared task, *EmoEvalEs: "Emotion detection and Evaluation for Spanish"*, has been designed to encourage research in this area. The task consists in classifying the emotion expressed in a tweet as one among the following emotion classes:

- *anger* (also includes annoyance and rage)
- *disgust* (also includes disinterest, dislike, and loathing)
- *fear* (also includes apprehension, anxiety, concern, and terror)
- *joy* (also includes serenity and ecstasy)
- *sadness* (also includes pensiveness and grief)
- *surprise* (also includes distraction and amazement)
- *others*: the emotion expressed in a tweet is neutral or there is no emotion

The challenges faced in this task are:

1. Lack of context: tweets are short (up to 240 characters)
2. Informal language: misspellings, emojis and onomatopoeias are common
3. Multiclass classification: the dataset is labeled with seven different classes
4. Imbalance dataset: the number of tweets per emotion category does not follow the same distribution

For developing their approaches, participants received the development and training partitions of the dataset and, at a later stage, the test partition was provided for evaluation. Finally, the participant's submissions were evaluated against the gold standard annotations to test their methods and determine the winner of the challenge.

The metrics used to evaluated the task were accuracy and the macro-averaged versions of Precision, Recall, and F1, being accuracy the one selected for ranking the systems.

3 Dataset

In this section, we describe the dataset of the EmoEvalEs shared task. EmoEvent (Plaza-del-Arco et al., 2020) is a multilingual emotion corpus of tweets based on events that took place in April 2019. They are related to different domains such as entertainment, catastrophe, political, global commemoration, and global strike. Each instance in the dataset is labeled with the main emotion expressed in the tweet by three annotators according to the following categories: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, "*neutral or no emotion*". The final emotion label of the tweet is the majority emotion labeled by the annotators, but in case the three annotators labeled the tweet with different emotions, the final label is "*neutral or no emotion*". In particular, for this task we used the Spanish version of EmoEvent.

This year, compared to the first edition in TASS 2020 (Vega et al., 2020), two new features from the EmoEvent dataset have been released to the participants: *offensive* (*OFF*: the tweet contains offensive language, *NO*: the tweet does not contain offensiveness) and *event* corresponding to

Emotion	Training	Dev	Test
Joy	1,227	181	354
Sadness	693	104	199
Anger	589	85	168
Surprise	238	35	67
Disgust	111	16	33
Fear	65	9	21
Others	2,800	414	814
Total	5,723	844	1,656

Table 1: Distribution of emotions by subset (Training, Dev, Test) in EmoEvalEs 2021.

the domain associated to the tweet (*WorldBookDay*, *GretaThunberg*, *Venezuela*, *GameOfThrones*, *LaLiga*, or *SpainElection*).

Before providing the dataset to the participants, we decided to replace the query hashtags by the keyword *HASHTAG* in order to prevent the automatic classifier from relying on hashtags to categorize the emotion associated with a tweet. Moreover, we replaced the user mentions by *@USER* to anonymize mentions to users.

Finally, different sets have been released for the competition. During the pre-evaluation phase, training and development (Dev) sets were provided to the participants; for the evaluation phase, the test set was released. Table 1 shows the number and percentage of tweets by emotion corresponding to each subset. It can be noticed that classes are highly imbalanced in the dataset.

4 Participant approaches

In this edition, 70 teams registered on the task, 15 submitted results and 11 presented working notes describing their systems. The following is a brief summary of the final proposals submitted.

- **GSI-UPM team - 1st (Vera, Araque, and Iglesias, 2021)**. This group has studied the combination of different features (like TF-IDF, n-grams, sentiment scores and the provided *event* and *offensiveness* columns) with encodings of a fine-tuned XLM-RoBERTa model. Although the best submission in the competition was their fine-tuned version of the multilingual RoBERTa model (XLM-RoBERTa), the reported scores on development set are not much higher than those obtained with Logistic Regression over text representations based on provided event and offensive categories along with sentiment analysis scores.

- **BERT4EVER team - 2nd (Fu et al., 2021)**. The authors adopt the monolingual Spanish BERT to tackle the task (BETO). In addition, they leveraged two augmented strategies to deal with the imbalanced emotion categories in the dataset, namely continual pre-training and data augmentation. The best result was obtained with the pre-training of BETO on the training set provided by the EmoEvalEs organizers, by ignoring the labels and performing back translation on the three categories with lower proportion: *disgust*, *fear* and *surprise*.

- **Yeti team - 3rd (Luo, 2021)**. This author used back-translation data augmentation technology to solve the problem of data scarcity and data imbalance. Chinese and English were used as intermediate languages for back translation. He mainly enhances the fear and disgust categories. The best result was by entering the offensive labels plus tweets text into the BETO-cased model.

- **URJC team - 4th (Sánchez, Herranz, and Unanue, 2021)**. The approach to the emotion detection problem proposed by this team was a fine-tuned version of BETO. They tried both, cased and uncased models, and a third tuning reducing, by a 30%, the number of samples within the *others* category. The best result was obtained with a voting system over the three trained models.

- **haha team - 5th (Li, 2021)**. The tweet, the event and offensive features are combined as a new text. Then, URLs, white-space characters and stop words are removed. The author adopts a masked language model technique for data augmentation in order to increase the training set and avoid over-fitting. Experiments were conducted with three cross-language models: BERT, XLM, and XLM-RoBERTa. The best performance was obtained with the XLM-RoBERTa model. The author shows that the technique used for data augmentation increased the generalization of the model.

- **UMU team - 6th (García-Díaz, Colomo-Palacios, and Valencia-García, 2021)**. The authors explore the combination of explainable linguistic features and state-of-the-art transformers. On the one hand, they used the UMUTextStats tool (García-Díaz, Cánovas-García, and Valencia-García, 2020; García-Díaz et al., 2021) to extract the linguistic features. On the other hand, they used sentence embeddings and word embeddings from fastText, GloVe and word2vec (although no details on how to compute sentence embeddings were provided) and sentence embeddings from BETO (pre-trained model) and from a fine-tuned BETO version on the EmoEvalEs dataset. Finally, they combine the features using an ensemble model based on the mode. In their analysis, they show the potential of the linguistic features to provide model agnostic methods for explainability.

- **RETUYT-InCo team - 9th (Chiruzzo and Rosá, 2021)**. They incorporate a diverse set of features to classical machine learning algorithms and traditional neural networks. The final model is LSTM where authors incorporate features from word2vec and BERT embeddings along with a the k word feature selection by a variance (ANOVA F-value) method. They mentioned that the most difficult emotion categories to classify by the model were *disgust*, *fear* and *surprise*.

- **WSSC team - 10th (Vitiugin and Barnabò, 2021)**. The authors propose a complex architecture which combines BiLSTM encodings with provided offensiveness and event features. Each of these three sets of features are the input of one or more feed forward networks, although not clear details are provided. Despite this complexity, results are not better than attention based mechanisms reported by other participants.

- **UPC team - 12th (de Arriba, Oriol, and Franch, 2021)**. The authors propose an approach based on a fine-tuned BETO model on pre-processed tweets. They pre-processed the tweets as follows: i) they removed URLs, hashtags, and numbers; ii) they replaced emojis, emoticons, abbreviations and laughs; iii) they removed punctuation marks, repeated characters, stopwords and blank spaces; and iv) they lemmatized the text. They concluded that the submitted system is less accurate for detecting the emotion categories with a small number of samples in the dataset: *fear*, *disgust* and *surprise*.

- **Dong team - 14th (Qu, Yang, and Que, 2021)**. It presents a combination of different neural networks (XLM-RoBERTa, Transformer enconding layer, TextCNN and a final linear one). In first place, they pre-processed the tweets by removing punctuation marks, emojis, empty characters and other special symbols. Then, they passedthe input data to XML-RoBERTa model to obtain word vectors with global features of sentences. Then they input the word vector into a Transformer Encoder for secondary feature extraction, and then pass the result into a TextCNN network. Finally, they passed the model output to a fully connected layer for classification.

- **Qu team - 15th (Qu, Jia, and Zhang,)**. The authors use the XLM-RoBERTa model to extract the features from training samples and then input the acquired word features into the Bi-GRU model to get the emotional features of comments. Finally, they classify the sentiment tendency by the softmax activation function.

5 Results and discussion

Table 2 shows the main results of the EmoEvalEs Shared Task. We received submissions through CodaLab from 15 participants. 11 teams provided their working notes explaining the systems that were developed for the competition. From all submissions, the best scoring system was that by GSI-UPM team, followed by BERT4EVER and Yeti. Between the first two participants, it can be observed that the difference in terms of macro-F1 and accuracy is minimal. The best team, GSI-UPM, achieved a macro-F1 score of 0.717028, exploring the combination of different features with a fine-tuned XLM-RoBERTa model. The team ranked in second

Team	Accuracy	Macro-P	Macro-R	Macro-F1
GSI-UPM	(1) 0.727657	(1) 0.709411	(1) 0.727657	(1) 0.717028
BERT4EVER	(2) 0.722222	(2) 0.704695	(2) 0.722222	(2) 0.711373
Yeti	(3) 0.712560	(3) 0.704496	(3) 0.712560	(3) 0.705432
URJC-TEAM	(4) 0.702899	(4) 0.692397	(4) 0.702899	(4) 0.696675
haha	(5) 0.692029	(6) 0.679620	(5) 0.692029	(8) 0.663740
UMUTeam	(6) 0.685990	(7) 0.672546	(6) 0.685990	(7) 0.668407
ffm	(7) 0.684179	(5) 0.682765	(7) 0.684179	(5) 0.682487
fazlfrs	(8) 0.682367	(8) 0.664868	(8) 0.682367	(6) 0.668757
RETUYT-InCo	(9) 0.678140	(9) 0.658314	(9) 0.678140	(10) 0.657367
WSSC	(10) 0.675725	(10) 0.657681	(10) 0.675725	(9) 0.661427
job80	(11) 0.668478	(12) 0.652840	(11) 0.668478	(11) 0.646085
UPCTeam	(12) 0.652778	(14) 0.600479	(12) 0.652778	(12) 0.622223
Timen	(13) 0.617754	(15) 0.597877	(13) 0.617754	(13) 0.600217
Dong	(14) 0.536836	(11) 0.653707	(14) 0.536836	(14) 0.557007
qu	(15) 0.449879	(13) 0.618833	(15) 0.449879	(15) 0.446947

Table 2: EmoEvalEs official ranking by accuracy (ranking position per metric is shown in parenthesis).

position was BERT4EVER ,with a macro-F1 score of 0.711373. It used BETO along with two augmented strategies to enhance the classic fine-tuned model, namely continual pre-training and data augmentation. The third team, Yeti, achieved a macro-F1 score of 0.705432, using back translation data augmentation technology to solve the problem of data scarcity and data imbalance, and tried to input the offensive labels and the text of the tweet into the BETO-cased model.

Most of the teams used neural network solutions to address the task. In particular, Transformers are the most widely used model by the participants in two ways: (1) as encoders to obtain contextualized sentence embeddings features from text, and (2) fine-tuning the pre-trained model on the task of emotion classification. As tweets from the dataset were in Spanish, most of the teams adopted two available pre-trained language models on Spanish corpora, the multilingual XLM-RoBERTa model and the monolingual BETO model.

Only three teams considered offensive and event information in their approaches (GSI-UPM, haha and WSSC). In most cases, this led to a slight improvement of the system, so both categories seem to retain certain semantic information related to emotions. In general, the enrichment of the learning process with additional data (by means of data augmentation techniques) or with additional features beyond neural network encodings, provide some insight on the relevance of hybrid methods for determining subjective information from texts. Although end-to-end solutions like BERT based models are clearly beneficial, additional characteristics are worth exploring, promoting ensemble based designs.

Focusing on the classification by emotion categories, some participants mention the challenge of classifying those emotions whose presence in the dataset is lower. In particular, these emotions are *fear*, *disgust* and *surprise*. Data augmentation by back translation was applied by two of the teams to address class imbalance. Also, some teams indicated that the systems faced the challenge of distinguishing complementary emotions, for example, the pairs *disgust* and *anger*, *fear* and *sadness* were often confused, a fact that is reflected by their close locations in the two-dimensional models of emotions.

6 Conclusions

Emotion classification is still considered a challenge in NLP systems and its significance has increased in recent years. With this task "Emotion detection and Evaluation for Spanish Shared Task" (EmoEvalEs) at IberLEF 2021, we want to promote research in the area of emotion analysis in Spanish, using the Spanish version of EmoEvent dataset.

As organizers, we are very satisfied with participation volume, as it was high. In this edition of EmoEvalEs, 70 participants registered, 15 of them submitted valid predictions and 11 contributed with a description

of their systems. As expected, deep learning approaches constitute the trend in this text classification task. In addition, the combination of linguistic information confirms the benefits of opting for hybrid solutions. Certainly, some of the most interesting challenges that participants faced were class imbalance and how to combine additional features with deep neuronal networks encondings.

Although different partitions from those of past editions have been provided this year, there is a clear improvement in performance. Best result reported in macro-F1 in 2020 was of 0.447. Compared to the best macro-F1 score in this year edition (0.717). Clearly, teams have gained skills in applying deep neural networks and adapting them to specific tasks. Besides, the participation has raised from 2 to 15 teams. We believe that moving the competition to CodaLab had the additional effect of more visibility, as can be noticed by the fact that five teams are located in China (four of them from Yunnan University), which represents one third of total participants.

As future work, we plan to include the English version of the EmoEvent dataset in the competition in order to promote multilingual emotion analysis research and explore how emotions are expressed for each event based on the cultural differences between English and Spanish speakers.

Acknowledgements

This work has been partially supported by a grant from Fondo Social Europeo, Administration of the Junta de Andalucía (DOC_01073 and P20_00956-PAIDI 2020), Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21) and the Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

References

Chiruzzo, L. and A. Rosá. 2021. RETUYT-InCo at EmoEvalEs 2021: Multiclass Emotion Classification in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

de Arriba, A., M. Oriol, and X. Franch. 2021. Applying Sentiment Analysis on Spanish Tweets Using BETO. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

Díaz-Galiano, M. C., M. Vega, E. Casasola, L. Chiruzzo, M. Á. G. Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. M. Ráez, M. A. S. Cabezudo, E. S. Tellez, M. Graff, and S. Miranda-Jiménez. 2019. Overview of tass 2019: One more further for the global spanish sentiment analysis corpus. In *IberLEF@SEPLN.*

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Fu, Y., Z. Yang, N. Lin, L. Wang, and F. Chen. 2021. Sentiment Analysis for Spanish Tweets based on Continual Pre-training and Data Augmentation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

García-Díaz, J. A., M. Cánovas-García, and R. Valencia-García. 2020. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Generation Computer Systems*, 112:614–657.

García-Díaz, J. A., R. Colomo-Palacios, and R. Valencia-García. 2021. UMUTeam at EmoEvalEs 2021: Emotion Analysis for Spanish based on Explainable Linguistic Features and Transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

García-Díaz, J. A., M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506 – 518.

Li, K. 2021. HAHA at EmoEvalEs 2021: Sentiment Analysis in Spanish Tweets with Cross-lingual Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

Luo, H. 2021. Emotion Detection for Spanish with Data Augmentation and Transformer-Based Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

Martínez-Cámara, E., Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. Á. G. Cumbreras, M. Vega, Y. G. Vázquez, A. M. Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, and J. Villena-Román. 2018. Overview of tass 2018: Opinions, health and emotions. In *TASS@SEPLN.*

Montes, M., P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Álvarez Carmona, E. Álvarez Mellado, J. Carrillo-de Albornoz, L. Chiruzzo, L. Freitas, H. Gómez Adorno, Y. Gutiérrez, S. M. Jiménez-Zafra, S. Lima, F. M. Plaza-del-Arco, and M. Taulé, editors. 2021. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021).*

Plaza-del-Arco, F., C. Strapparava, L. A. Ureña-Lopez, and M. T. Martin-Valdivia. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France, May. European Language Resources Association.

Plutchik, R. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Qu, S., Y. Yang, and Q. Que. 2021. Emotion Classification for Spanish with XLM-RoBERTa and TextCNN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

Qu, Y., S. Jia, and Y. Zhang. Sentiment Analysis in Spanish Tweets: The Model based on XLM-RoBERTa and Bi-GRU.

Rosenthal, S., N. Farra, and P. Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Sánchez, J. A. F., S. M. Herranz, and R. M. Unanue. 2021. URJC-Team at EmoEvalEs 2021: BERT for Emotion Classification in Spanish Tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

Vega, M., M. C. Díaz-Galiano, M. Á. G. Cumbreras, F. M. Plaza-del-Arco, A. M. Ráez, S. Jiménez-Zafra, E. Martínez-Cámara, C. Aguilar, M. A. S. Cabezudo, L. Chiruzzo, and D. Moctezuma. 2020. Overview of tass 2020: Introducing emotion detection. In *IberLEF@SEPLN 2020.*

Vera, D., O. Araque, and C. A. Iglesias. 2021. GSI-UPM at IberLEF2021: Emotion Analysis of Spanish Tweets by Fine-tuning the XLM-RoBERTa Language Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

Vitiugin, F. and G. Barnabò. 2021. Emotion Detection for Spanish by Combining LASER Embeddings, Topic Information, and Offense Features. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*