UMUTeam at EmoEvalEs 2021: Emotion Analysis for Spanish based on Explainable Linguistic Features and Transformers

José Antonio García-Díaz^{1[0000-0002-3651-2660]}, Ricardo Colomo-Palacios^{2[0000-0002-1555-9726]}, and Rafael Valencia-García^{1[0000-0003-2457-1791]}

 Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain {joseantonio.garcia8,valencia}@um.es
Faculty of Computer Sciences, Østfold University College, Halden, Norway ricardo.colomo-palacios@hiof.no

Abstract. Emotion Analysis extends the idea of Sentiment Analysis by shifting from plain positive or negative sentiments to a rich variety of emotions to get better understanding of the users' thoughts and appraisals. The move from Sentiment Analysis to Emotion Analysis requires, however, better feature engineering techniques when it comes to capturing complex language phenomena, which have to do with figurative language and the way of expressing oneself. In this manuscript we detail the participation of the UMUTeam in EmoEvalEs'2021 shared task from IberLEF, concerning the identification of emotions in Spanish. Our proposal is grounded on the combination of explainable linguistic features and state-of-the-art transformers based on the Spanish version of BERT. We achieved the 6th position in the official leader board with an accuracy of 68.5990%, only 4.1667% below the best result. In addition, we apply model agnostic techniques for explainable artificial intelligence to achieve insights from the linguistic features. We observed a correlation between psycho-linguistic processes and perceptual feel with the emotions evaluated and, specifically, with documents labelled as sadness.

Keywords: Emotion Analysis \cdot Feature Engineering \cdot Natural Language Processing.

1 Introduction

Emotion Analysis (EA) is a Natural Language Processing (NLP) task related to Sentiment Analysis (SA), Document Classification (DC) and Information Retrieval (IR), whose objective is the identification of emotions from a piece of

IberLEF 2021, September 2021, Málaga, Spain.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

text [21]. Standard SA, on the other hand, is focused on determining whether a document is positive, neutral, or negative. EA insights, therefore, are useful for creating better recommender systems that adapt better to the mood of the users [21]. Moreover, the oversimplification of SA could be misleading in some scenarios. For example, while analysing online reviews of movies, EA might identify as sadness the emotions that arouse in people from the film *La vita è bella*; however, these reviews can be wrongly classified as negative from conventional SA approaches because sadness and negative feelings are related in some way [9].

In this manuscript we describe the participation of the UMUTeam in the shared task EmoEvalEs 2021 [15] proposed at Iberian Languages Evaluation Forum (IberLEF) [11]. This task is focused on the classification of emotions in micro-blogging posts, which is challenging due to the absence of contextual clues such as voice modulation or facial expressions. Specifically, this task aims to distinguish among the following emotions: Anger, Disgust, Fear, Joy, Sadness, Surprise and Others.

One of our objectives for participating in this task is the evaluation of a set of linguistic characteristics extracted with the tool UMUTextStats [4, 5] of which it is part a doctoral thesis from a team member. It is worth mentioning that we participated with an previous version of this tool on TASS 2020 shared task [6], in which a similar EA subtask was proposed. However, for this task we present a major revision of the linguistic features and new forms of combining them with state-of-the-art transformers.

This manuscript is organised as follows: First, Section 2 provides background information regarding EA. Next, Section 3 describes briefly the corpus that was made available by the organisers of the shared task. The methodology is depicted in Section 4. Next, Section 5 contains the results achieved by our team and the comparison with the rest of the participants. In addition, an interpretation of the features is presented. Finally, the conclusions and promising future research directions are shown in Section 6.

2 Background information

Emotion categorisation is a challenging task. On the one hand, there are several emotion classifications [7], such as six Ekman's basic emotions [3], Plutchik's Wheel of Emotions [17] or Russel's Circumplex Model [20]. On the other, the detection of emotions is subtle to distinguish as several emotions can be present at the same time. Also, there are not too many studies and resources in Spanish focused on this task. Nevertheless, recent shared tasks are focusing on Spanish EA as TASS 2020 [24], which includes a subtask based on six Ekman's basic emotions in Spanish tweets. One of the approaches to address the lack of datasets in Spanish for EA was carried out in [16], in which the authors presented a dataset of tweets compiled in April 2019 annotated based on the six Ekman's basic emotions plus an extra emotion for neutral and others. Another recent work is [1], in which the authors apply EA to social media by incorporating to their pipeline affective lexical resources such as SEL [22], iSOL [10], and EmoLex [8]. The experiments performed in this work indicate that the usage of linguistic features and sentiment lexicons are advantageous for conducting EA. In the same line, the usage of linguistic features have proven effective in other related tasks such as satire identification [14], in which the authors employ Linguistic Inquiry Word Count (LIWC) [23] for distinguish among satiric and non-satiric texts from European Spanish and Mexican Spanish tweets.

3 Corpus

According to the organisers of the task, the EmoEvalEs' dataset consisted in tweets from April 2019 based on different events. The tweets were pre-processed to replace hashtags and mentions with some tokens to hinder the automatic classification task. The dataset was distributed in three splits: train, development, and testing. Table 1 depicts the distribution of the corpus. As we can observe, many of the tweets could not be labelled with one of the sentiments and they were rated as *others*. This fact gives an idea of the difficulty of the task, even for human annotators. The emotions with more number of instances are *joy*, and *sadness* that are, from our point of view, the most generic and polarised emotions. On contrast, *fear* and *disgust* emotions are underrepresented in the dataset. It may be that these emotions are difficult to categorise, or that people do not express those emotions on public social networks.

Label	Total T	raining	Val	Test
others	4028	2800	414	814
joy	1762	1227	181	354
sadness	996	693	104	199
anger	842	589	85	168
surprise	340	238	35	67
disgust	160	111	16	33
fear	95	65	9	21
Total	8223	5723	844	1656

Table 1. Corpus distribution by label and split ordered by number of instances

4 Methodology

This section describes the feature sets employed for solving this task, the neural networks evaluated, and the hyperparameter optimisation stage carried out.

Regarding the features employed, our proposal is grounded on linguistic features in combination with state-of-the-art transformers [25]. During our experimentation, we also evaluated word and sentence embeddings from pre-trained Spanish models. For the linguistic features (LF) we use UMUTextStats [4, 5]. This tool is inspired in LIWC [23] but designed from scratch the Spanish language. UMUTextStats takes into account more than 350 linguistic features categorised as follows: (1) phonetics, which handles techniques such as word elongation; (2) morphosyntax, that includes a fine-grained Part-of-Speech tags extracted from Stanza [18] and custom lexicons; (3) correction and style, that captures different stylistic and correction patterns used during writing; (4) semantics, that captures linguistic phenomena such as onomatopoeia, euphemism, dynamism, or synecdoc; (5) pragmatics, that includes figurative language phenomena [13], discourse markers and courtesy forms; (6) stylometry, including several corpora statistics such as Type-token ratio (TTR) and punctuation symbols; (7) lexical, that includes a wide variety of topics, including locations, organisations, animals, weapons, food, religion, or health among others; (8) psycholinguistic processes, that includes positive and negative expressions; (9) register, that includes the usage of informal speech, colloquialisms, or SMS language; and (10) social media, that captures jargon used in social networks. For the transformers we use the Spanish version of BERT, also known as BETO [2]. To obtain these vectors, we evaluated two methods, that we called *BE* and *BF* respectively, both extracting the [CLS] token in a similar way as is detailed at [19] and using HuggingFace (v4.4.2). The key difference is that for BE we obtained the vectors from BETO directly, whereas for BF we first fine-tuned BETO with the Emo-EvalEs dataset. Both BE and BF are fixed-vectors of 768 items per document. In addition to the transformers, we also evaluated neural networks with word and sentences embeddings from fastText, word2vec, and gloVe. We refer to these feature set as WE for the word embeddings and SE for the sentence embeddings.

Each feature set (LF, SE, BF, and WE) was trained separately and in combination using the functional API of Keras. For the fixed-sentence vectors we rely on multi-layer perceptrons but for WE we also evaluated a convolutional and two bidirectional recurrent neural networks, based on Long-Short Term Memory (BiLSTM) and Gated Recurrent Unit (BiGRU), that have provided good results in the past for conducting SA tasks [12].

The next step in our pipeline consisted in a hyperparameter optimisation. For this, we evaluate a total of 110 neural models per feature set (in isolation or combined). The best model was selected using the weighted F1 score. Most of the neural networks evaluated consisted in shallow multilayer perceptrons (MLP) with one or two hidden layers and with both hidden layers having the same number of neurons (8, 16, 48, 64, 128, 256). We also evaluated deep neural networks with a number of hidden layers between 3 and 8, with a different number of neurons per hidden layer organised in different shapes. For the rest of the hyper-parameters, we evaluated different dropout rates, several activation functions, and different learning rates. The source code is available at https://github.com/Smolky/emoevales-2021.

https://huggingface.co/sentence-transformers/bert-base-nli-cls-token https://huggingface.co/

https://github.com/dccuchile/spanish-word-embeddings

Table 2 depicts the results of the hyperparameter optimisation stage for each feature set separately and in combination. For the sake of simplicity, we have included only the combinations with LF. Regarding the feature sets separately, we can observe that the best results are obtained with shallow neural networks. with 2 hidden layers (except for SE) with brick shape. The number of neurons is always less than the number of parameters, resulting in 256 neurons for LF, 128 for SE, and 512 for BE and BF. All neural networks achieved their best results with dropout for the features in isolation. The learning rate varies from 0.001 for LF and BE to 0.01 for SE and BF. Out of the activation functions, relu achieves better results for LF, SE, and BE whereas tanh achieves better results for BF. When we observe the features combined in pairs, only the combination of LF with BE requires a complex deep neural network to achieve their best result, with 4 hidden layers and 512 neurons stacked in a diamond shape. However, when combining LF with BF, the best result is achieved with a simpler model composed by two hidden layers of 128 neurons each. When combined in groups of three, the combination of LF, SE, and BE requires also a deep neural network composed of four hidden layers (as the combination of LF with BE) but with 1024 neurons. However, the combination of LF, SE, and BF resulted in a simpler model of two hidden layers with 48 neurons each. A similar architecture can be found when combining LF, SE, BE, and BF. In this case, the network also results in a very simpler model with only one hidden layer of 48 neurons, a dropout of 0.2, and a learning rate of 0.01 with a sigmoid as activation function. The simplicity of the networks in which BF is present can be explained because the weights of BF have been trained with the EmoEvalEs dataset, so the embeddings have already been grouped based on the emotions within the latent space.

Feature set	shape hidde	n layers	neurons	dropout	learning rate	activation
LF	brick	2	256	0.3	0.001	relu
SE	brick	1	128	0.3	0.01	relu
BE	brick	2	512	0.1	0.001	relu
BF	brick	2	512	0.2	0.01	\tanh
LF-SE	brick	1	512	0.3	0.01	relu
LF-BE	diamond	4	512	-	0.001	elu
LF-BF	brick	2	128	0.3	0.01	sigmoid
LF-SE-BE	brick	4	1024	-	0.01	sigmoid
LF-SE-BF	brick	2	48	0.2	0.001	sigmoid
LF-SE-BE-BF	brick	1	48	0.2	0.01	sigmoid

Table 2. The best hyperparameters by separate feature sets and combinations in thatLF appears

5 Results

Participants were required to submit a maximum of three runs that are ranked by macro average F1-score but also by accuracy and the macro-averaged versions of Precision and Recall. The organisers of the task allowed the participants to send their runs in two separated time slots: A development phase, in which the participants could evaluate their results with the development dataset, and the official one, against the test split. Due to lack of time, we were able to send only a run during the development phase that achieved an accuracy of 70.8531% and a macro averaged F1-Score of 69.9542%, reaching the second position in a total of six participants.

For the official competition our first run consisted in an ensemble of the best model for each feature set: LF, SE, BE, and BF. We exclude WE because it requires a large amount of time for training and the results does not outperform the models based on fixed-length vectors. This ensemble model decides the final output with an averaged version of the mode. For that, we store the results of each model with the validation set in order to decide its weight for the final decision. We achieve an 68.599% accuracy with this run. The macro F1-Score is 66.8407%, the precision is 67.2546%, and the recall is 68.5990%. For our second run, we evaluate another form of ensemble based on the softmax layer of each neural network. We use the probabilities of each neural network to train an extra ensemble. This run achieves worst result than the previous ensemble with an accuracy of 68.2971%. Our last submission consisted in a MLP perceptron trained with two inputs LF, and BF, as we want to compare the results of methods non based on ensembles. We achieve an accuracy of 66.7874%.

The official results are depicted in Table 3. We achieve the 6th position in the official leader board with an accuracy of 68.5990%, a macro average precision of 67.2546%, a macro recall of 68.5990%, and a macro F1-score of 66.8407%. The best result is achieved by *fyinh*, with an macro F1-score of 71.7028%, followed by *fyinh* with an macro F1-score of 71.1373%. We can observe that all runs and participants achieve competitive results. On the one hand, the major accuracy difference is only of 10.9903% between the best and worst result. On the other hand, the relation regarding the macro precision and macro recall is similar among all the participants. It is not noting that we set the main metric for the hyper-parameter optimisation to the weighted f1-score but finally, macro f1-score was the official score. It is possible, therefore, that we could achieve better results with a better strategy.

We include the normalised confusion matrix of the best model, a ensemble that combines LF, SE, BE, and BF using the weighted mode, with the validation set (see Figure 1). We can observe that *anger* is predicted correctly most of the times, and the wrong classifications are about the *others* class. Emotions of *disgust* are classified wrongly as *anger*, followed by *others* and *fear*. Only a 12% of documents labelled as *disgust* are correctly classified. Documents labelled as *fear* by the annotators are correctly classified the 67%, but sometimes they are wrongly classified as *anger*, *disgust*, and *others*. It is worth noting that both classes, *fear* and *disgust* are the labels with less instances in the corpus and that

Rank	Team/User	Accuracy	M-precision	M-recall	M-F1 score
1	daveni	72.7657	70.9411	72.7657	71.7028
2	fyinh	72.2222	70.4695	72.2222	71.1373
3	HongxinLuo	71.2560	70.4496	71.2560	70.5432
4	JorgeFlores	70.2899	69.2397	70.2899	69.6675
5	hahalk	69.2029	67.9620	69.2029	66.3740
6	UMUTeam	68.5990	67.2546	68.5990	66.8407
7	ffm	68.4179	68.2765	68.4179	68.2487
8	fazlfrs	68.2367	66.4868	68.2367	66.8757
9	luischir	67.8140	65.8314	67.8140	65.7367
10	vitiugin	67.5725	65.7681	67.5725	66.1427
11	job80	66.8478	65.2840	66.8478	64.6085
12	aridearriba	65.2778	60.0479	65.2778	62.2223
13	Timen	61.7754	59.7877	61.7754	60.0217
14	QuSwe1d0n	53.6836	65.3707	53.6836	55.7007
15	qu	44.9879	61.8833	44.9879	44.6947

Table 3. Official results of the task, ranked by accuracy

our proposal is especially confused with the *disgust* class. For the class *joy*, our system classifies it correctly the 61%, labelling as *others* a 32%. The majority class *others* is classified correctly the 80%, but the 12% is wrongly classified as *joy*. Note that these were the classes with large number of instances. *Sadness* is correctly classified the 67%. Finally, for the documents labelled as *surprise*, our system is able to classify the 37%, but a 43% of the times are classified as *others*, an 11% as *anger*, and a 9% as *joy*. The strong points of our proposal are that there are not so many wrongly classifications as opposite emotions, as it could be labelling *sadness* as *joy* or vice-versa. However, our proposal is confused between *anger* and *disgust* and it achieves a low recall on the class *surprise*.

In order to provide some understanding of the linguistic features, we obtain the top ten discriminatory linguistic features per class (see Figure 2) and we generate a polar chart for each linguistic category and emotion (see Figure 3). Note that in both charts we exclude intentionally those tweets labelled as *others*.

As it was expected, the features related to sad emotions are strong discriminatory for the *sadness* label, but also has an strong impact on *disgust*. In a similar manner, *anger* label is also related with the psycho-linguistic process anger, but also with *disgust*. As a personal opinion, *anger* and *disgusts* are the emotions in which it is more difficult to differentiate. Another correlation is perceptual feel, which has a strong correlation with *sadness*. In the same line, negative process is also related to different emotions such as *anger*, *disgust*, *fear*, and *sadness*, but also is relevant for documents labelled as *surprise*. It draws our attention that the token $^{\circ}$ has a strong correlation for documents labelled as *sadness* and *surprise*. We manually checked which tweets contains that sign and the majority are related to the sports events, such as *La Liga* and *ChampionsLeague*. They appear to discuss about results by means of ordinal numbers. It can also be



Fig. 1. Confusion matrix with the validation split with an ensemble based on the weighted mode of LF, SE, BE, and BF



observed that tweets with fewer words correspond mostly to tweets labelled as *fear* and *joy*.

Fig. 2. Mutual information by label.

Regarding each linguistic feature category (see Figure 3), the major difference among emotions appears in the semantics category, that it is the one that includes positive, and negative emotions. Regarding phonetics, that include features such as word elongation to add emphasis, *sadness* is the emotion that makes less use of this linguistic device. Regarding correction and style, *fear* is the emotion in which most stylistic errors are detected. Regarding lexical and topics, there is a wide heterogeneity among the emotions, ordered from major to minor use of topics by *surprise*, *joy*, *fear*, *disgust*, *anger*, and *sadness*. This fact suggest that people describes the cause of their emotions to explain which causes its surprise or joy, but they are less likely to explain why they are sad or angry.

6 Conclusions

Here we have described the participation of the UMUTeam at EmoEvalEs 2021 shared task regarding EA in Spanish. As commented earlier, this task has been



Fig. 3. Polar chart of emotions.

an opportunity for us to evaluate our methods in real scenarios and we considered that we achieved competitive results but with room for improvement. From the point of view of explainable artificial intelligence, we have shown the potential of the linguistic features to provide model agnostic methods for explainability.

As promising research directions we suggest to continue with the interpretability of the neural network models and features. In this sense, we propose to find the correlations between the linguistic features and embeddings in order to determine in which cases they are complementary and in which not. Another promising direction is to provide contextual features to EA, in order to track how sentiments and emotions are changing on online conversations such as threads on Twitter.

Acknowledgments

This work was supported by the Spanish National Research Agency (AEI) through project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme.

References

- Plaza-del Arco, F.M., Martín-Valdivia, M.T., Ureña-López, L.A., Mitkov, R.: Improved emotion recognition in spanish social media through incorporation of lexical knowledge. Future Generation Computer Systems 110, 1000–1008 (2020)
- 2. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
- Ekman, P.: Lie catching and microexpressions. The philosophy of deception 1(2), 5 (2009)
- García-Díaz, J.A., Cánovas-García, M., Valencia-García, R.: Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. Future Generation Computer Systems 112, 614–657 (2020). https://doi.org/10.1016/j.future.2020.06.019
- García-Díaz, J.A., Cánovas-García, M., Colomo-Palacios, R., Valencia-García, R.: Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. Future Generation Computer Systems 114, 506 – 518 (2021). https://doi.org/10.1016/j.future.2020.08.032, http://www.sciencedirect.com/science/article/pii/S0167739X20301928
- García-Díaz, J.A., Almela, A., Valencia-García, R.: Umuteam at tass 2020: Combining linguistic features and machine-learning models for sentiment classification. Proceedings of TASS (2020)
- Kim, E., Klinger, R.: A survey on sentiment and emotion analysis for computational literary studies. CoRR abs/1808.03137 (2018), http://arxiv.org/abs/1808.03137
- Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. pp. 26–34 (2010)

- Mokryn, O., Bodoff, D., Bader, N., Albo, Y., Lanir, J.: Sharing emotions: determining films' evoked emotional experience from their online reviews. Information Retrieval Journal 23, 475–501 (2020)
- Molina-González, M.D., Martínez-Cámara, E., Martín-Valdivia, M.T., Perea-Ortega, J.M.: Semantic orientation for polarity classification in spanish reviews. Expert Systems with Applications 40(18), 7250–7257 (2013)
- Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.Á., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez Zafra, S.M., Lima, S., Plaza-de Arco, F.M., Taulé, M.: Proceedings of the iberian languages evaluation forum (iberlef 2021). In: CEUR workshop (2021)
- Paredes-Valverde, M.A., Colomo-Palacios, R., Salas-Zárate, M.d.P., Valencia-García, R.: Sentiment analysis in spanish for improvement of products and services: a deep learning approach. Scientific Programming **2017** (2017)
- del Pilar Salas-Zárate, M., Alor-Hernández, G., Sánchez-Cervantes, J.L., Paredes-Valverde, M.A., García-Alcaraz, J.L., Valencia-García, R.: Review of english literature on figurative language applied to social networks. Knowl. Inf. Syst. 62(6), 2105–2137 (2020). https://doi.org/10.1007/s10115-019-01425-3, https://doi.org/10.1007/s10115-019-01425-3
- del Pilar Salas-Zárate, M., Paredes-Valverde, M.A., Rodríguez-García, M.Á., Valencia-García, R., Alor-Hernández, G.: Automatic detection of satire in twitter: A psycholinguistic-based approach. Knowl. Based Syst. 128, 20–33 (2017). https://doi.org/10.1016/j.knosys.2017.04.009, https://doi.org/10.1016/j.knosys.2017.04.009
- Plaza-del-Arco, F.M., Jiménez-Zafra, S.M., Montejo-Ráez, A., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. Procesamiento del Lenguaje Natural 67(0) (2021)
- Plaza-del-Arco, F., Strapparava, C., Ureña-López, L.A., Martín-Valdivia, M.T.: EmoEvent: A Multilingual Emotion Corpus based on different Events. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1492– 1498. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.186
- 17. Plutchik, R., Kellerman, H.: Theories of emotion, vol. 1. Academic Press (2013)
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
- Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bertnetworks. CoRR abs/1908.10084 (2019), http://arxiv.org/abs/1908.10084
- Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology 39(6), 1161 (1980)
- Sailunaz, K., Dhaliwal, M., Rokne, J., Alhajj, R.: Emotion detection from text and speech: a survey. Social Network Analysis and Mining 8(1), 1–26 (2018)
- 22. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Trevino, A., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: Mexican international conference on Artificial intelligence. pp. 1–14. Springer (2012)
- Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29(1), 24–54 (2010)

- Vega, M.G., Díaz-Galiano, M.C., Cumbreras, M.Á.G., del Arco, F.M.P., Montejo-Ráez, A., Zafra, S.M.J., Cámara, E.M., Aguilar, C.A., Cabezudo, M.A.S., Chiruzzo, L., Moctezuma, D.: Overview of TASS 2020: Introducing emotion detection. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Zafra, S.M.J., Zambrano, J.A.O., Miranda, A., Zamorano, J.P., Gutiérrez, Y., Rosá, A., Montes-y-Gómez, M., Vega, M.G. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 163– 170. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2664/tass_overview.pdf
- 25. Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45 (2020)