RETUYT-InCo at EmoEvalEs 2021: Multiclass Emotion Classification in Spanish

Luis Chiruzzo and Aiala Rosá

Universidad de la República
Montevideo, Uruguay
{luischir,aialar}@fing.edu.uy

Abstract. This paper presents the results for the team RETUYT-InCo of the participation in the EmoEvalEs 2021 challenge. We trained several systems using classical ML techniques and neural networks, and using a diverse set of features including word embeddings and features from Spanish BERT. Our best system achieved 0.6573 macro weighted average F1 score (position 10 in the ranking) and 0.6781 accuracy (position 9) over the test set. The most difficult classes to classify were surprise, disgust and fear, which are also the classes with fewer examples in the corpus.

Keywords: Emotion classification · Spanish · LSTM · BERT · word embeddings.

1 Introduction

Within the area of subjectivity analysis in texts, emotion analysis presents greater challenges and has been less studied than the more traditional task of classifying texts according to their polarity. It is necessary to define the set of categories and to have larger datasets than for polarity classification, where the different categories are sufficiently represented. This implies a more complex annotation process due to greater subtlety in choosing the category for each example, making it more difficult to assess inter-annotator agreement.

An important antecedent on emotion annotation is the corpus created by [11], used at SemEval-2018 Task 1: Affect in Tweets [10]. In this task, a subtask on emotion classification was proposed for three languages: English, Arabic and Spanish. The corpus was annotated according to a set of eleven categories: *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, suprise, trust*, and a *neutral or no emotion* extra class.

This year, for the second time in a row, the IberLEF workshop includes a task addressing this problem for Spanish texts. In IberLEF 2020, an emotion

IberLEF 2021, September 2021, Málaga, Spain.

classification subtask was part of the TASS 2020 task [6], which traditionally addressed tweets polarity classification. In IberLEF 2021 [12], a task only for emotion classification, EmoEvalEs [14], was proposed. For both editions a corpus [15] with 8,409 tweets written in Spanish was used, classified according to Ekman's categories: *anger, disgust, fear, joy, sadness*, and *surprise*. A neutral or no emotion class (with the label *others*) is also included.

In this paper, we describe the participation of the RETUYT-InCo team in the EmoEvalEs@IberLEF task. Based on the previous experience of the team in sentiment analysis tasks [16], [4], [13], we experimented with the different approaches that are described in the next section. In sections 3 and 4 we analyze the results and present some conclusions.

2 Experiments

We trained a set of classifiers from different families and using different sets of features. Some classifiers belong to the classic set of ML methods: Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR). On the other hand, we tried two architectures of neural network classifiers: Multilayer Perceptrons (MLP) and Long Short-Term Memory networks (LSTM).

The different classifiers were trained using a variety of features:

- **word2vec:** Word embeddings trained using word2vec over a 6 billion word Spanish corpus [2]. The embeddings collections contains 1.4 million vectors of size 300. These embeddding vectors were used as centroids in the fixed length input methods (SVM, RF, LR and MLP) and as separate vectors in LSTM, as the architecture supports inputs of variable length.
- **BERT:** BERT features from the cased Spanish BERT model (BETO) [3]. Only the 768 units vector corresponding to the whole sequence was used for training, not the individual vectors used for each token. When used together with the LSTM models, the BERT features are a separate track of features that is concatenated with the output of the LSTM.
- **Emoji:** We used the Python `emoji`[1] library for recognizing the use of emojis in tweets. We took the 50 most frequently used emojis in the training corpus and created a binary feature for each one of them indicating if the tweet uses the emoji or not. When combined with the LSTM we used the `emoji` library in a different way, applying the `demojize` method for transforming the emojis into descriptive strings in Spanish, and letting the description be part of the token sequence instead of the emoji.
- **Parser:** The parser features where calculated using a Spanish HPSG parser [5]. In the experiments using this features, we split each tweet into a set of coordinated elements found by the parser (the tweet might contain a set of coordinated sentences, or even a sentence could be the coordination of several statements), and for each element we created separated features that represent the verbal head, the subject, the complements and the modifiers.

[1] https://pypi.org/project/emoji/

As the original tweets could be split in several statements, and the model might predict a different emotion for each statement, we take the emotion that has more votes for all the statements of a tweet.
- **k-best** Top k word features found by `sklearn` using the ANOVA F-value method. We calculated the lists of k-best tokens of sizes 10, 20, 30 and 50 and trained variants of the experiments with each one of the lists.

Our methods did not use the event and offense features from the data, as we wanted to create a system that was based entirely on the text content of the tweet.

The classic ML methods and the MLP networks were developed using the `sklearn`[2] library, while the LSTM networks were developed on `keras`[3] with `tensorflow` [1].

For all neural network approaches we created several versions of the experiments varying the number of layers (dense or LSTMs) and the number of units in each layer (generally between 64 and 1024 units. In all cases we used adam [9] optimization and early stopping.

3 Results

Table 1 shows the results for the development set. Notice that the best results in this table in terms of accuracy and weighted F1 were using the MLP method with word2vec centroids, the LSTM method with plain word2vec features, and the LSTM enriched with BERT and k-best features. These were the models we submitted for the evaluation phase of the competition.

The MLP with word2vec centroid model uses two layers of size 256 and 64 with relu activation. The LSTM model with plain word2vec features uses only a single layer of bi-directional LSTM of size 96, followed by a dense layer of size 64, both with tanh activation. The best LSTM model enriched with BERT and k-best features is similar, using a single LSTM layer of size 96 and a dense layer of size 64 (with tanh activation), the output of the LSTM is concatenated with the BERT features and features for the 30 best words.

In general, we noticed the systems could more easily predict tweets belonging to the most numerous classes (`joy`, `sadness`, `anger` and `others`), while generally struggled to find tweets in the remaining categories (`surprise`, `disgust` and `fear`). Many of the classifiers were not even able to take a shot at classifying any tweet with one of those three categories. We also tried some strategies for training with a more balanced corpus by removing some tweets (as seen in [4]) or augmenting the least numerous classes with artificial examples using transformations of the parse trees, but none of these techniques yielded significant improvements in the results.

Table 2 shows the results for the test set. In the three cases, the results over the test set were between three and four points below the development results,

[2] https://scikit-learn.org/
[3] https://keras.io/

	Features						
Classifier	word2vec	BERT	Emoji	Parser	k-best	Acc	MwF1
SVM	X					0.6670	0.6342
	X			X		0.5308	0.4239
RF	X					0.5959	0.5353
	X			X		0.4940	0.3861
LR	X					0.6196	0.6113
	X			X		0.4443	0.4430
MLP	X					0.6729	0.6584
	X		X			0.6729	0.6547
	X		X	X		0.6342	0.6187
	X	X				0.6623	0.6440
	X	X	X			0.6623	0.6413
		X				0.6540	0.6299
		X		X		0.4579	0.3168
LSTM	X					0.6860	0.6550
	X		X			0.6590	0.6273
	X	X			X	0.7026	0.6815

Table 1. Results over the development set.

both for accuracy and weighted F1. The best results were obtained by the LSTM with word2vec features enriched by the BERT and k-best features. The results of this model achieved position 10 according to weighted F1 and position 9 according to accuracy in the official results of the competition. Table 3 shows a comparison of the top results for different teams in the competition.

Model	Acc	MwF1
MLP with word2vec centroids	0.6358	0.6076
LSTM with word2vec embeddings	0.6437	0.6116
LSTM with word2vec + BERT + k-best	0.6781	0.6573

Table 2. Results over the test set.

4 Conclusions

The classification of emotions, as it usually happens when working on automatic subjectivity analysis, is a highly challenging task. However, the results of previous campaigns have been far outperformed by the EmoEvalEs teams, reaching a Macro F1 score of 0.717. In SemEval 2018 the highest Macro F1 reached on the Spanish corpus was 0.440 [7] (it is not the same corpus than the one used in EmoEvalEs). In the subtask on emotions of the TASS 2020 task, only two teams participated and the highest Macro F1 score was 0.447 [8], evaluated on the same corpus used in EmoEvalEs.

Team	Acc	MwF1
GSI-UPM	0.7276	0.7170
BERT4EVER	0.7222	0.7113
Yeti	0.7125	0.7054
RETUYT-InCo (ours)	0.6781	0.6573
UMSNH	0.6684	0.6460
UPC Team	0.6527	0.6222
autoBOT	0.6177	0.6002

Table 3. Comparison of results for different teams in the competition. We show the top three results, our top result, and the bottom three results.

Our top system for this competition achieved 0.6573 macro weighted averaged F1 (position 10) and 0.6781 accuracy (position 9) over the test corpus, which is also higher than the performance obtained in previous years. However, there is still a lot of room for improvement in these systems, and we noticed that the most difficult categories to classify are (as expected) the ones with the fewest examples: *disgust*, *fear* and *surprise*. More research is needed to understand if it is only the number of examples what makes these categories particularly challenging. In order to analyze this hypothesis we plan to retrain our models using an extended corpus, merging the SemEval 2018 and EmoEval datasets, keeping only the common categories (the six used in EmoEval), conducting experiments on a larger and/or more balanced dataset. We are also working on collecting and generating emotion lexicons, with the goal of conducting new experiments using this information.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015), http://tensorflow.org/, software available from tensorflow.org
2. Azzinnari, A., Martínez, A.: Representación de Palabras en Espacios de Vectores. Proyecto de grado, Universidad de la República, Uruguay (2016)
3. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
4. Chiruzzo, L., Rosá, A.: RETUYT-InCo at TASS 2018: Sentiment Analysis in Spanish Variants using Neural Networks and SVM. In: TASS@SEPLN. pp. 57–63 (2018)
5. Chiruzzo, L., Wonsever, D.: Statistical Deep Parsing for Spanish using Neural Networks. In: Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies. pp. 132–144 (2020)

6. García-Vega, M., Díaz-Galiano, M.C., García-Cumbreras, M.A., del Arco, F.M.P., Montejo-Ráez, A., Jiménez-Zafra, S.M., Martínez Cámara, E., Aguilar, C.A., Cabezudo, M.A.S., Chiruzzo, L., Moctezuma, D.: Overview of TASS 2020: Introducing Emotion Detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). pp. 163–170 (2020)
7. González, J.Á., Hurtado, L.F., Pla, F.: ELiRF-UPV at SemEval-2018 tasks 1 and 3: Affect and irony detection in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 565–569. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1092, https://www.aclweb.org/anthology/S18-1092
8. Ángel González, J., Moncho, J.A., Hurtado, L.F., Pla, F.: ELiRF-UPV at TASS 2020: TWilBERT for Sentiment Analysis and Emotion Detection in Spanish Tweets. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). pp. 179–186 (2020)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: SemEval-2018 task 1: Affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 1–17. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1001, https://www.aclweb.org/anthology/S18-1001
11. Mohammad, S., Kiritchenko, S.: Understanding emotions: A dataset of tweets to study interactions between affect categories. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1030
12. 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. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
13. Pastorini, M., Pereira, M., Zeballos, N., Chiruzzo, L., Rosá, A., Etcheverry, M.: RETUYT-InCo at TASS 2019: Sentiment Analysis in Spanish Tweets. In: IberLEF@ SEPLN. pp. 605–610 (2019)
14. 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)
15. 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
16. Rosá, A., Chiruzzo, L., Etcheverry, M., Castro, S.: RETUYT en TASS 2017: Análisis de sentimientos de tweets en español utilizando SVM y CNN. Proceedings of TASS (2017)