GSI-UPM at IberLEF2021: Emotion Analysis of Spanish Tweets by Fine-tuning the XLM-RoBERTa Language Model

Daniel Vera[1 [0000−0001−8791−0274]], Oscar Araque[1 [0000−0003−3224−0001]], and Carlos A. Iglesias[1 [0000−0002−1755−2712]]

Universidad Politécnica de Madrid, Intelligent Systems Group, 28040 Madrid, Spain
d.vnieto@alumnos.upm.es
{o.araque,carlosangel.iglesias}@upm.es

Abstract. This work presents the participation of the Intelligent Systems Group (GSI) at Universidad Politécnica de Madrid (UPM) in the Emotion Analysis competition EmoEvalEs, part of IberLEF 2021 Conference. The addressed challenge proposes an emotion classification task of Spanish tweets, categorizing each message into seven emotions. We propose the design and development of a fine-tuned neural language model (XLM-RoBERTa) to tackle this challenge. We have obtained excellent results with this approach, obtaining the first place in the competition with a macro-averaged F1 score of 71.70%. Additionally, we also explore the application of several ensemble methods built over the neural language model.

Keywords: Emotion Analysis · Transformers · XLM-Roberta · Twitter

1 Introduction

Recent advances in machine learning research are rapidly pushing the sentiment analysis field forward. Works that use neural architectures to improve previous models are established, and current state-of-the-art models are heavily based on these techniques [1,21]. Although sentiment analysis still represents a challenging task and further study is needed, research in emotion analysis is also relevant. In this sense, estimating emotions from text is currently less studied and opens a new range of potential applications. Since sentiment and emotion analysis share many subproblems, the approaches that tackle these disciplines are frequently similar.

This paper presents our participation in IberLEF 2021 [12], describing our efforts towards EmoEvalEs, an emotion classification task [14]. The task presents an emotion classification challenge in the form of a multiclass classification task, where the emotions considered are *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and

IberLEF 2021, September 2021, Málaga, Spain.

other, category which represent emotions that are not included in the Ekman emotion model or the absence of any emotion. The data has been extracted from Twitter, and its contents address several domains: entertainment, catastrophe, political, global commemoration, and global strike. For more information on the dataset, please consult [15].

In the effort of addressing this task, we use the fine-tuned XLM-Twitter (XLM-T) language model [5] as the primary emotion estimator. Additionally, intending to improve the results obtained by XLM-T, we combine its predictions in an ensemble system that uses several types of features and models. The final result indicates that our efforts are oriented in the right direction since we have obtained the first place in the competition with a final macro-averaged F1 score that reaches 71.70%.

For replication reasons, we provide the source code used to generate the models and their respective submissions. It is available online at https://github.com/gsi-upm/emoevales-iberlef2021.

The rest of the paper is organized as follows. Section 2 provides the background of the methods used in this work. Next, the proposed approach and architecture for emotion classification are described in Section 3 and evaluated in Section 4. Finally, Section 5 states our conclusions and proposes future lines of work.

2 Background

Deep learning approaches are common in sentiment analysis and have proved helpful in emotion analysis [2]. Incorporating deep learning models was initiated with the popularization of word embedding models, such as word2vec [11] or GloVe [13]. Word embedding models have allowed researchers and practitioners to develop new deep learning models that use these distributed representations. One relevant example is Sentiment-Specific Word Embedding (SSWE) that computes sentiment-oriented word embeddings that can be later used to predict sentiment in texts [17]. Another model that makes efficient use of word embeddings is presented in [1]. This model uses a straightforward word vector aggregation method that extracts a unified document representation from a word embedding model. In this way, the aggregation of word vectors has proven effective in sentiment analysis, obtaining consistent performance in different data domains.

The use of word embeddings to elicit sentiment and emotion represents a large field. In this work, we use previous models to incorporate them into our ensemble model. One of those is the SIMilarity-based sentiment projectiON (SIMON) model, which computes the representation of a particular word in a document by considering its projection to a set of domain words [4]. Such projection is computed using the semantic similarity between words, as obtained from a word embedding model. Thus, a document word is represented by its similarity to the selection of domain words. However, as previously studied, this selection of words can be varied, and selecting the component words can highly affect the final prediction performance [3].

In recent years, an approach that has proved successful in most Natural Language Processing (NLP) tasks is using large pretrained language models based on a transformer architecture. Transformers [18] are an attention-based architecture that allows computing complex representations of information without using Recurrent Neural Networks, which have made it possible to parallelize the training of large language models efficiently. After the release of BERT [8] in 2018, the NLP community has created new improved language models. One of these language models is RoBERTa [10], an optimized BERT pretraining approach that achieves significantly better results than the previous BERT implementation. Furthermore, RoBERTa outperformed state-of-the-art results, becoming the baseline for many further works for different NLP tasks, such as cross-lingual language understanding (XLU).

In this domain, XLM-RoBERTa (XLM-R) [7] stands out as a model pretrained in 100 different languages, achieving state-of-the-art performance on cross-lingual classification, sequence labeling, and question answering. The lack of pretrained language models in languages different than English has geared researchers' interest towards multilingual models that have demonstrated that it is possible to have a single large model for all languages without sacrificing too much performance for each language. However, previous research shows that multilingual models tend to underperform monolingual models in language-specific tasks [16]. This context framed the pretrained language model we have used for this work, XLM-T, an XLM-R that achieves better results in the Twitter domain that its XLM-R baseline and has been pretrained on millions of tweets in over 30 different languages

3 Architecture

3.1 Fine-tuning XLM-T

We have fine-tuned the Twitter-specific pretrained language model in the downstream task of emotion classification following parameter-efficient transfer learning techniques [9]. In short, the language model parameters remain unchanged while the weights of a neural network classification head on top of the language model are trained.

We have implemented this architecture and run the training process using the modules and the Trainer API from the HuggingFace Transformer library [19], which is optimized and provides a wide range of training options and built-in features. We have tested three different approaches to solve the problem:

- **Multi-label classification problem**: We have trained the model to predict the class with a higher probability among the seven possibilities.
- **Binary classification problem**: Frame the multiclass problem as a one-vs-all problem where seven different models are trained. Ties are solved by selecting the the output from the model with the highest confidence score.
- **Additional Features**: We extend the classification head to use the additional features, *event* and *offensive*, available in the dataset as new inputs encoded as one-hot vectors.

The classification head consists of a dense layer at the output of the language model, followed by a dropout layer with the default dropout probability of the language model and a final projection layer with the number of labels. For the Additional Features model, we have added new inputs to the first dense layer.

We have followed the training process as described in Transformers documentation. For the sake of reproducibility, the source code is available at https://github.com/gsi-upm/emoevales-iberlef2021, and the fine-tuned model is available at the HuggingFace model hub.

The hyper-parameters used are a batch size of 16 per GPU, max length (tokenizer) of 200, and training for 5 epochs. The rest of the parameters are the default in the Trainer API. The trainer API also saves the checkpoint of the best epoch that usually occurs at epoch 3. We run the training process for 1 hour on two NVIDIA Titan X Pascal GPUs.

Before tokenizing, we have slightly preprocessed the tweets with the Twitter preprocessing module of the GSITK library [4]. We have found this helpful to achieve slightly better results.

3.2 Ensemble model

To improve the final prediction scores, we have developed an ensemble model that combines, at the prediction level, different models trained on varied data. In this sense, previous works [1,20] have successfully used ensemble models to boost the prediction performance. Furthermore, as outlined in previous works, an ensemble model improves its performance when using varied models trained with different feature types. We have used the following features in our ensemble model:

SIMON [4]. As mentioned, this model computes the representation of a document by measuring the similarity of the component words to those of a predefined domain word set. The main component of this model is the domain word set, from which the model adapts to the different language uses of a specific domain. Following previous work [3], we have extracted a custom word set from the dataset, selecting the words by their frequency of appearance in the dataset.

Word embedding combination (M_G). As described in Sect. 2, this model aggregates the component word vectors from a document, obtaining a fixed document representation. Previous works have found that this method is reliable across different domains. In this work, we aggregate the word vectors using the *average* aggregation operation.

Term Frequency–Inverse Document Frequency (TF-IDF). We use the TF-IDF as a simple text representation method. Our model instance considers both uni and bigrams.

N-grams. Similarly, as before, we use this feature to enhance the variety of the ensemble model. Moreover, as before, we consider uni and bigrams.

Additional features. The challenge dataset contains additional information that can be easily used as features. Concretely, we used the *event* and *offensive* data fields. The *event* category specifies the general event from which the message has been extracted. In contrast, the *offensive* category specifies whether the

message contains offensive language, which may aid in the task at hand. Please note that we do not train a learning model with these features solely. However, instead, we add them to other feature sets.

MeaningCloud. Since sentiment and emotion are intimately related we have incorporated a new feature, the sentiment estimation that MeaningCloud (https://www.meaningcloud.com/) offers. MeaningCloud offers a professional sentiment analysis service that can be accessed via a web API (https://www.meaningcloud.com/products/sentiment-analysis). We extract the sentiment estimations for all messages using this service. This information is included as an additional feature.

Considering the described feature types, we train different learning models that train on the mentioned features. We select a simple algorithm for the base learners, logistic regression, since all predictions are combined in an ensemble fashion. For feature combination, we concatenate the feature vectors. When using learning models to train for the ensemble, we have selected logistic regression and random forests.

4 Evaluation

4.1 XLM-T evaluation

This section describes the performance of the different approaches we have followed during the fine-tuning of the pretrained model. Table 1 shows the accuracy and weighted F1 scores of the different approaches on the development set, where the model fine-tuned in the multiclass classification problem has achieved the best results. Table 2 shows the same information for the test set, where the best model is the multiclass estimator again.

Table 1: Evaluation on dev set of the fine-tuned models

	Accuracy	**Weighted F1 score**
XLM-T Multi-label classification	73.10	71.10
XLM-T Binary one-vs-all classification	72.39	70.01
Additional Features	71.80	69.89

Table 2: Evaluation on test set of the fine-tuned models

	Accuracy	**Weighted F1 score**
XLM-T Multi-label classification	72.77	71.70
XLM-T Binary one-vs-all classification	71.43	68.93
Additional Features	71.67	69.66

These results show the superior performance of the multiclass classifier over the combination of various binary classifiers, although better combination strategies could improve the results of the latter. Moreover, including additional features without any preprocessing and at the same level as the output produced by the pretrained language model decreases the classifier performance.

Figure 1 depicts the confusion matrix produced by the XLM-T multi-label classifier on the test set. We observe the evident unbalancing of the dataset, where almost half of the records belong to the *others* class. Moreover, this class is usually confounded with the *joy* class. Additionally, this matrix shows the difficulty of distinguishing between emotions that share similar features, such as *anger* and *disgust*. Finally, the low number of records in some classes (*anger*, *disgust*, and *surprise*) is an additional challenge since the models tend to fail in those classes.

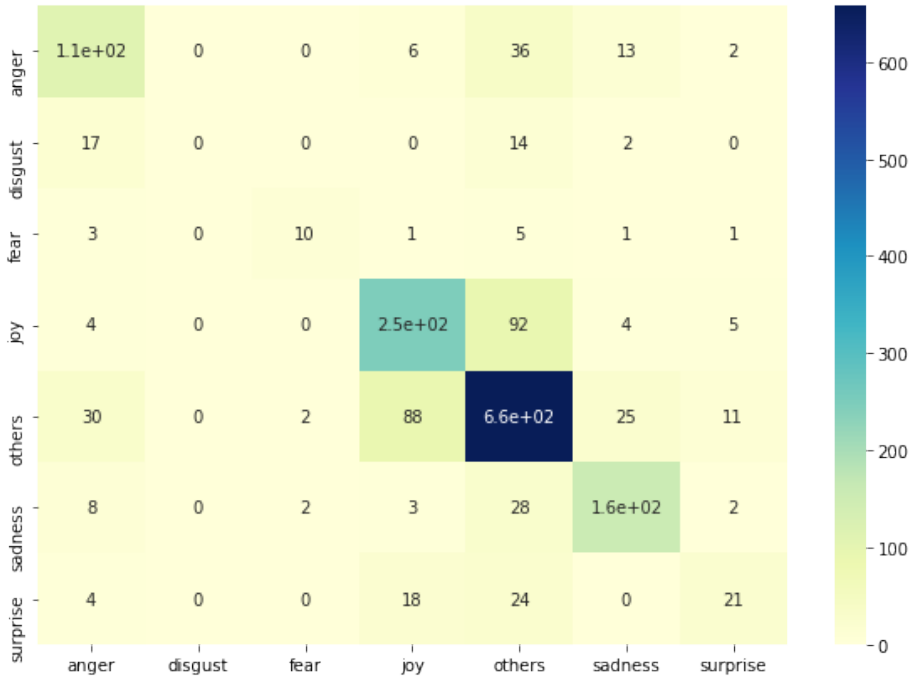

Fig. 1: Confusion matrix on test set produced by XLM-T multi-label classifier

4.2 Ensemble evaluation

As explained, we have designed an ensemble methodology to combine several base models and raise the final classification performance. Our primary model is the one we have obtained through the use of the XLM-T transformer fine-tuned for the multiclass classification problem, and thus we include this model into the ensemble. Table 3 details the models used in the ensemble, along with the features they use to train. To evaluate the different models, the accuracy and weighted averaged F1 score have been used.

Table 3: Models used for the ensemble with the used features.

Model name	**Features used**
XLM-T	XLM-T fine-tuned model (see Sect. 3.1).
word2vec	Word embedding combination (M_G): averaged word vectors.
n-gram	Uni and bi-gram representations.
TF-IDF	TF-IDF features, considering both uni and bigrams.
SIMON	SIMON features using an extracted word set.
n-gram + add. features	n-gram representations in combination with the dataset's additional features.
Ensemble LR	Ensemble that uses a logistic regression model to learn from base classifier predictions.
Ensemble RF	Ensemble that uses a random forest model to learn from base classifier predictions.
Ensemble LR + add. features	Ensemble LR combined with the dataset's additional features.
Ensemble RF + add. features	Ensemble RF combined with the dataset's additional features.
Ensemble LR + add. features + MeaningCloud	Ensemble LR + add. features, combined with the sentiment estimation obtained from MeaningCloud.

We have evaluated the models detailed in Table 3 on the development dataset [15]. The obtained metrics can be seen in Table 4. As described above, the XLM-T model obtains high accuracy and averaged F1 scores. As expected, the rest of the base models achieve lower metrics in comparison since they do not consider such complex relations in the analyzed text. Therefore, we can consider the average of word vectors (word2vec in Table 3), n-grams, and TF-IDF as baseline approaches in this task. Following, the SIMON model a higher score than the rest of the base methods, which can be explained by the increased complexity of the method in comparison. The SIMON model uses both a word embedding model and a selected word set to compute the text representation.

Following, we can observe that adding additional features (*event* and *offensive* categories) to the n-gram approach improves the regular n-gram features. This fact indicates that these additional features can be leveraged to improve classification performance.

Table 4 shows that combining all base models through an ensemble generally improves the classification performance when attending to the ensemble methods. Nonetheless, the metrics have not been improved over the XLM-T model, even though this model is included in the ensemble.

This situation changes when adding additional features and the MeaningCloud sentiment analysis to the ensemble. The ensemble using all features gets a lower accuracy, but the weighted F1 score is slightly higher in the development set, although this does not represent a relevant improvement. Please note that in this last case, the ensemble learner is trained with a combination of the predictions from the base classifiers, additional features, and MeaningCloud's sentiment analysis results.

Table 4: Development set results.

	Accuracy	**Weighted F1 score**
XLM-T	**73.10**	71.10
word2vec	61.02	58.81
n-gram	62.68	60.19
TF-IDF	56.99	59.63
SIMON	66.35	62.46
n-gram + add. features	64.22	61.95
Ensemble LR	71.56	68.93
Ensemble RF	69.31	67.56
Ensemble LR + add. features	70.97	70.76
Ensemble RF + add. features	68.01	66.43
Ensemble LR + add. features + MeaningCloud	71.09	**71.14**

When attending to the test set results, we have observed a different situation. The last ensemble, with additional features and sentiment analysis, does not improve the XLM-T final performance.

Table 5: Test set results.

	Accuracy	**Weigthed F1 score**
XLM-T	**72.77**	**71.70**
word2vec	59.78	57.25
n-gram	60.93	57.45
TF-IDF	56.04	57.34
SIMON	62.86	58.99
n-gram + add. features	62.44	59.30
Ensemble LR + add. features + MeaningCloud	70.89	70.78

5 Conclusions

This paper has described our participation in the EmoEvalEs competition framed in the IberLef 2021 Conference. Our proposal relies on using large pretrained language models, outperforming previous methods with little effort using the HuggingFace library, which provides a straightforward implementation of these pretrained language models. The pretrained model we have used is a RoBERTa transformer trained on a multilingual corpus of tweets, XLM-T. We have evaluated different strategies to approach the problem, finding that the fine-tuned model for a multiclass classification task obtains better results than the combination of various binary classifiers and the model with additional features. We have achieved first place in the EmoEvalEs competition with this model, obtaining a macro-averaged F1 score of 71.70%.

This work also presents an ensemble method that combines several base classifiers with the XLM-T model to improve the final performance by adding more knowledge to the system. Although we have found a slight improvement in the overall classification metrics in the development set, this enhancement has not continued in the test set. The obtained results suggest that combining such a transformer architecture with classical machine learning methods is a challenge, which must be done carefully.

We propose further lines of research to improve this work. Firstly, the effective combination of additional features that carry new information could enhance the classifier's overall performance. Secondly, we suggest using a weighted validation loss during the fine-tuning of the language model to deal with the unbalanced dataset problem. Moreover, using a monolingual pretrained model for the specific language of the task could improve the obtained results. In this sense, using the Spanish language model BETO [6] seems promising since it has demonstrated great results in similar tasks like sentiment analysis. Finally, this same method could be applied for emotion classification in other languages with the same pretrained language model, XLM-T.

6 Acknowledgements

The authors want to thank the help and support from the Cátedra Cabify (Cabify Chair) at the ETSI Telecomunicación of the Universidad Politécnica de Madrid. Moreover, the authors would also like to gratefully acknowledge MeaningCloud's support in facilitating this research work. Finally, the authors would like to acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used in this research.

References

1. Araque, O., Corcuera-Platas, I., Sánchez-Rada, J.F., Iglesias, C.A.: Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications **77**, 236–246 (2017). https://doi.org/https://doi.org/10.1016/j.eswa.2017.02.002, https://www.sciencedirect.com/science/article/pii/S0957417417300751
2. Araque, O., Gatti, L., Staiano, J., Guerini, M.: Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. IEEE Transactions on Affective Computing pp. 1–1 (2019). https://doi.org/10.1109/TAFFC.2019.2934444
3. Araque, O., Iglesias, C.A.: An approach for radicalization detection based on emotion signals and semantic similarity. IEEE Access **8**, 17877–17891 (2020). https://doi.org/10.1109/ACCESS.2020.2967219
4. Araque, O., Zhu, G., Iglesias, C.A.: A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowledge-Based Systems **165**, 346–359 (2019). https://doi.org/https://doi.org/10.1016/j.knosys.2018.12.005, https://www.sciencedirect.com/science/article/pii/S0950705118305926
5. Barbieri, F., Anke, L.E., Camacho-Collados, J.: XLM-T: A multilingual language model toolkit for twitter. CoRR **abs/2104.12250** (2021), https://arxiv.org/abs/2104.12250
6. 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)
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://www.aclweb.org/anthology/2020.acl-main.747
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
9. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. CoRR **abs/1902.00751** (2019), http://arxiv.org/abs/1902.00751
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

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. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
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. Rust, P., Pfeiffer, J., Vulic, I., Ruder, S., Gurevych, I.: How good is your tokenizer? on the monolingual performance of multilingual language models. CoRR **abs/2012.15613** (2020), https://arxiv.org/abs/2012.15613
17. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1555–1565 (2014)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. CoRR **abs/1910.03771** (2019), http://arxiv.org/abs/1910.03771
20. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Information sciences **181**(6), 1138–1152 (2011)
21. Yadav, A., Vishwakarma, D.K.: Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review **53**(6), 4335–4385 (2020)