HAHA at EmoEvalEs 2021: Sentiment Analysis in Spanish Tweets with Cross-lingual Model

Kun Li

School of Information Science and Engineering, Yunnan University, Yunnan, P.R. China
2106967047@qq.com

Abstract. This article describes HAHATeam participation in the IberLEF 2021 EmoEvalEs task: Emotion detection and Evaluation for Spanish. In this task, we use Masked Language Model-based data augmentation to enhance the data to increase the training data set and prevent overfitting. We improve the performance of the model through a series of data processing and data augmentation technologies. We use three different cross-language models, BERT, XLM, and XLM-RoBERTa for comparative experiments. Very competitive results have been achieved on both the development set and the test set. The best model achieved the third place in the development set and the fifth place in the test set.

Keywords: Sentiment Analysis, Data Augmentation, Machine Learning, Cross-lingual Model.

1 Introduction

Sentiment analysis (SA) is the process of analyzing, processing and categorizing subjective text. According to the emotional tendency of the text, the text can be classified. In addition to the traditionally recognized positive and negative emotions, there should also be anger, disgust, fear, joy, sadness, surprise and other emotion types. This article describes the emotion classification task (EmoEvalEs [11]) that we participated in in IberLEF 2021 [7], which requires the classification of Spanish tweets into various categories based on emotions, such as anger, disgust, fear, joy, sadness, surprise, or others. Due to the lack of voice regulation and facial expressions, understanding the emotions expressed by users on social media is a difficult task.

The earliest work on SA largely relied on feature engineering [15,3], and subsequent neural network-based methods [8,17,14,16] achieved higher accuracy. Recently, Ma [6] said: "Incorporate useful common sense knowledge into deep neural networks to further enhance the results of the model." Liu [5] optimized the memory network and applied it to its model to better capture the language structure.

IberLEF 2021, September 2021, Málaga, Spain.

Recently, pre-trained language models, such as ELMo [10], OpenAIGPT [13], and BERT [2], have shown their effectiveness in mitigating feature engineering efforts. In particular, BERT has achieved excellent results in Question Answering (QA) and Natural Language Inference (NLI). However, the SA task that directly uses the pre-trained BERT model does not improve much.

In order to solve this task, we proposed the use of pretrained classic deep learning models to apply to downstream task. Furthermore, we not only tested the data on a single model, but also compared three language models. We use the three mainstream cross-language models: bert-multiling-cased (BERT), xlm-mlm-100-1280 (XLM), xlm-roberta-base (XLM-RoBERTa). These three models all come from the transformers package, and the hugging face provided us with great help. The system will sort according to accuracy and the weighted-averaged versions of Precision, Recall, and F1.

The rest of the article is structured as follows. Section2 proposes the task to be solved and introduces the corpus of the task. It analyzes the data set and proposes some tricks. In the Section3, we introduce the methods and techniques we used. The first is to use some conventional operations to process the data set, the second is to introduce the data augmentation method we use, and the third is to introduce the language model we use. Section 4 presents the results and discuss the performance of the system. In Section 5, we give the conclusions and future work.

2 Task and Corpus description

2.1 Task description

The organizer only proposed one task: EmoEvalEs [11]: Emotion detection and Evaluation for Spanish. 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 "Emotion detection and Evaluation for Spanish" has been designed to encourage research in this area. The task consists of classifying the emotion expressed in a tweet as one of the following emotion classes: Anger, Disgust, Fear, Joy, Sadness, Surprise or Others.

2.2 Corpus description

The datasets [12] is based on events related to different fields that occurred in April 2019: entertainment, disasters, politics, global memorials, and global strikes. The corpus contains approximately 8.2k tweets for this task. Each tweet contains 5 features: id, event, tweet, offensive and emotion. The most important feature is tweet. At the same time, event and offensive may also have a positive impact on the final result. We will verify our conjecture in the next part of the experiment. For this task, the corpus is released as three data sets, namely training, development and testing. Table 1 shows the distribution of training set and development set.

Table 1. Distribution of tweets in the training and development set of the task.

	Joy	Sadness	Anger	Surprise	Disgust	Fear	Others	Total
training	1408	797	674	273	127	74	3214	5723
development	181	104	85	35	16	9	414	844

3 Method and technology

3.1 Data Preprocessing

To achieve good results, data preprocessing is essential. The first step in our model is to process these tweets. For how to preprocess the data, we refer to IberLEF 2019 [9], which gives us some ideas and and more in-depth analysis. We found that in addition to tweets, there are two other features that may affect that final result, so we process the datasets as follows:

- Combining tweet, event, and offensive together as a new text. The advantage of this is to make better use of the information given by the organizer.
- Removing URLs.
- Contracting white-space characters, such as space, tabs or punctuation.
- Removing stop word in Spanish.

We ignored spelling errors in text tweets. After data processing, the sentence length in the tweet is distributed within 60. At the same time, we also keep the version that not remove the stop word in Spanish, so that we can judge whether removing the stop word in Spanish is good for our model.

3.2 Data Augmentation

Data augmentation technology is widely used in computer vision, but it is rarely effectively used in Natural Language Processing (NLP). The essential reason is that some data augmentation methods in the image will not change the meaning of the image itself while augmenting the data. In this article, we try to use a data augmentation method: Masked Language Model-based data augmentation. Masked Language Modeling (MLM) can not only predict words, but also can be used for text data augmentation, and other methods, such as based on synonym substitution and word embedding. Compared with the replacement of, the text generated by data augmentation based on the Masked Language Model is more grammatically fluent, because the model considers the context information when making predictions. We can easily use HuggingFace's transformers library. One place to pay attention to in this method is how to determine which part of the Mask text is to be used. Do data augmentation for each processed tweet, and determine the number of masks according to the length of each sentence, details as follows:

- Fixed 1 [Mask], random position, repeatable.

– Fixed 2 [Mask], random position, repeatable.
– Fixed 3 [Mask], random position, repeatable.
– 15% of the length of the sentence [Mask], the position is random, not repeatable.
– 20% of the length of the sentence [Mask], the position is random, not repeatable.

Next is a concrete example: 'This is [Mask] cool'
'score': 0.515411913394928, 'sequence': 'This is [pretty] cool'
'score': 0.1166248694062233, 'sequence': 'This is [really] cool'
'score': 0.07387523353099823, 'sequence': 'This is [super] cool'
'score': 0.04272908344864845, 'sequence': 'This is [kinda] cool'

"score" indicates the degree of similarity between the word and [Mask], and "sequence" indicates the sequence after data augmentation. The sequence with the highest score is selected as the sentence augmented by our data.

3.3 Model

For SA Task, we combine some classic deep learning models. In the deep learning model, we mainly use the most popular cross-lingual models. These cross-lingual models can be better used to process Spanish text. These models include two parts: tokenizer and SequenceClassification. Among them, the tokenizer divides the input into words and processes it into the input required by the model. SequenceClassification is applied to specific downstream tasks.

– tokenizer: First of all, the tokenizer was splitting strings in sub-word token strings, converting tokens strings to ids and back, and encoding or decoding. Adding new tokens to the vocabulary in a way that is independent of the underlying structure (Byte Pair Encoding). Managig special tokens (like mask, begining-of-sentence, etc). The model needs language embedding during inference. We will use the tokenizer module to process the input before entering the model. The input required for each cross-language model is different, and the tokenizer module will generate the corresponding embedding according to the specific model. The tweet first undergoes data preprocessing, and then uses the tokenizer to vectorize the input. According to different models, Token embeddings, Segment Embeddings, and Position Embeddings are obtained.
– SequenceClassification: The model with a sequence classification/regression heads on top(a linear layer on top of the pooled output). We add a dropout layer and a linear layer to the last layer. The input feature of the linear layer we added is 768, the output feature is 384, and the p value of the dropout layer is set to 0.1. The input feature of the last layer is 384, and the output feature is 7, to satisfy our 7 classification task. The model does not require language embedding at inference time. They can identify the language used in the context and infer accordingly.

XLM-RoBERTa [1]: The XLM-RoBERTa model was proposed in Unsupervised Cross-lingual Representation Learning. It is a large multi-lingual language model, trained on 2.5TB of filtered Common Crawl data. XLM-RoBERTa is a multilingual model trained on 100 different languages. Unlike some XLM multilingual models, it does not require lang tensors to understand which language is used, and should be able to determine the correct language from the input-ids. XLM-RoBERTa was evaluated on classification tasks and they showed very good performance. During fine-tuning, multi-language annotation data is used based on the ability of the multi-language model to improve the performance of downstream tasks. This allows our model to obtain state-of-the-art results in cross-language benchmark tests, while surpassing the performance of the single-language BERT model in each language. The main body of the XLM-RoBERTa model is Transformer, and the training target is the multilingual MLM objective, which is basically the same as XLM. We sample the text from each language, and then predict the tokens that are masked. We use model Sentence Piece with unigram language model to perform Sentence Piece on the original text data.

XLM [4]: Although the BERT model can be pre-trained on hundreds of languages, the information between languages is not interoperable, and there is no shared knowledge between different language models. Facebook's XLM model overcomes the problem of non-interoperability of information, puts different languages together and uses new training targets for training, so that the model can master more cross-language information. A significant advantage of this cross-language model is that, for subsequent tasks after pre-training (such as text classification or translation tasks), languages with relatively rare training corpus can use the information learned on other corpora. We propose to use xlm-mlm-100-1280 as part of our experiment.

BERT: The bert-base-multilingual-cased model is based on BERT, which has obtained state-of-the-art performance on most NLP task [2]. The pre-trained BERT models are trained on a large corpus (Wikipedia). Trained on cased text in the top 104 languages with the largest Wikipedias. So bert-base-multilingual-cased model is better for Spanish text in Sentiment Analysis (SA). Because bert-base-multilingual-cased model can splits tokens in its vocabulary into sub-tokens, while will affect the result of the classification task. The bert-base-multilingual-cased model also fixes normalization issues in many languages, so it is recommended in languages with non-Latin alphabets (and is often better for most languages with Latin alphabets).

4 Experiments

We have fully experimented with the above three model. The process and result of the experiment will be described in detail below. We first experiment with different data augmentation methods to obtain better data augmentation methods. Then we use the best data augmentation method on the three cross-language models and compare them on the development set. Finally, we use the best model to experiment on the test set and get the final result.

4.1 Experimental Setup

During the development set experiment, we divided the training set into a new training set and a development set, where the division ratio was 0.9 for the new training set, and the development set was used as the test set.

We will do the same preprocessing before sending the data to each model to ensure the comparability of the experiment, that is the ablation experiment. Use ablation experiments to evaluate the quality of our different data processing methods. And here some hyper parameter sett of the model; the max length of a single sentence in each batch is adjusted according to the preprocessed results, the max length does not exceed 60; and add "[CLS]" and "[SEP]" to each input; and padding for sentences with insufficient length; We use a batch size of 32, a base learning rate of 1e-5; the activation function is AdamW; and the number of fine-tuning is set to 3. For every 10 batch-size training set data processed, model.eval will be executed to process the development set data. And save the model and print the accuracy rate.

4.2 Learning Rate Warm-up

Use learning rate warm-up. During warm-up, the learning rate linearly increases from 0 to the initial learning rate of the optimizer. After the warm-up phase, a schedule is created to linearly reduce the learning rate from the initial learning rate in the optimizer to 0. The number of steps for the warm up phase is set to 0, that is, the learning rate is warmed up from the beginning. The total number of training steps are set to N.

$$N = len(train\,dataloader) \times num\,epochs \tag{1}$$

4.3 Data Augmentation Experiments

We conduct data augmentation experiments on the XLM-RoBERTa model to get the best data augmentation method, and use the data augmentation method in all subsequent models. From Table 2, we can see that when using 15% of

Table 2. The results of different data augmentation strategies on the development data set on the XLM-RoBERTa model.

XLM-RoBERTa	Accuracy	Precision weighted	Recall weighted	F1 weighted
Fixed 1	0.697	0.671	0.697	0.680
Fixed 2	0.698	0.670	0.698	0.678
Fixed 3	0.687	0.657	0.687	0.666
15% random	**0.710**	0.676	**0.710**	**0.685**
20% random	0.703	**0.677**	0.703	0.681

the length of the sentence [Mask], XLM-RoBERTa is the best among the three

evaluation indicators Accuracy, Recall weighted and F1 weighted score. But when the XLM-RoBERTa model uses 20% of the sentence length [Mask], the precision of the model is the best. The choice of pre-training model has a great impact on the final performance, even though they are all based on transformers.

4.4 Finally Experiments

Next, we conduct ablation experiments on three cross-language models for data augmentation. From Table 3, we can see that the four evaluation indicators

Table 3. The results of different data augmentation strategies on the development data set on the cross-language model.

Model	Data Augmentation	Accuracy	Precision weighted	Recall weighted	F1 weighted
BERT	Yes	0.626	0.606	0.626	0.614
	No	0.666	0.618	0.666	0.637
XLM	Yes	0.659	0.654	0.659	0.656
	No	0.684	0.673	0.685	0.669
XLM-RoBERTa	Yes	**0.710**	0.676	**0.710**	**0.685**
	No	0.704	**0.702**	0.704	0.677

Table 4. The results of the XLM-RoBERTa model on the test set.

Model	Data Augmentation	Accuracy	Precision weighted	Recall weighted	F1 weighted
XLM-RoBerta	Yes	**0.698**	0.647	**0.698**	**0.670**
	No	0.692	**0.680**	0.692	0.664

of the BERT model and the XLM model during the data augmentation operation are less than those without the data augmentation operation. In the XLM-RoBERTa model, data augmentation can improve the performance of the model to a certain extent, except for Precision weighted scores. Furthermore, whether data augmentation is performed or not, the performance of the XLM-RoBERTa model is better than the other two models. The biggest advantage of XLM-RoBERTa is the significant increase in the amount of training data. And the effect is especially obvious in languages with fewer resources.

Finally, we use the best performing XLM-RoBERTa model on the test data set, with data augmentation and no data augmentation at the same time. From Table 4, we can see that data augmentation will effectively improve the performance of the model in test data set, especially in Accuracy, Recall weighted and F1 weighted scores. But it will also reduce the Precision weighted score of the model.

4.5 Result Analysis

We use three popular cross-language models BERT, XLM, and XLM-RoBERTa for comparative experiments. It can be seen from Figure 3 that the XLM-RoBERTa model surpasses BERT and XLM in all evaluation indicators. Because the XLM-RoBERTa model increases the number of languages and the number of training data sets, it uses more than 2TB of pre-processed CommonCrawl data sets. And during the fine-tuning of the XLM-RoBERTa model, based on the ability of the multi-language model to use multi-language annotation data to improve the performance of downstream tasks. Beyond this, although we use data augmentation technology to only double the amount of data, the performance improvement is still very obvious. Accuracy and Recall weighted have increased by 0.6%, and F1 weighted has increased by 1.2%. Moreover, the data processing method and Learning Rate Warm-up technology we use also play a certain role in the improvement of model performance and the speed of model inference.

Table 5. The results and ranking of our model on the official test sets (Ranked by Accuracy).

Team	Accuracy	Precision weighted	Recall weighted	F1 weighted	rank
GSI-UPM	0.728	0.709	0.728	0.717	1
HAHA	0.698	0.647	0.698	0.670	5
qu	0.450	0.619	0.450	0.447	15

5 Conclusions and further work

This article introduces the EmoEvalEs task: Emotion detection and Evaluation for Spanish in the IberLEF 2021. We use three cross-language models BERT, XLM and XLM-RoBERTa. Using the cross-language pre-training model can handle Spanish tweets well, and the pre-training model only needs to modify the final output layer to be well applied to downstream tasks, which greatly reduces the training time. We improve the performance of the model through a series of data processing and data augmentation technologies. As can be seen from Table 5, The results and ranking of our model on the official test sets (Ranked by Accuracy). Our final model got an Accuracy of 0.692, a Precision weighted score of 0.680, a Recall weighted score of 0.692, and an F1 weighted score of 0.664 on the official test set, ranking 5th, 6th, 5th, and 8th among all participating teams.

We are very satisfied with our first participation in IberLEF 2021 and our performance in the EmoEvalEs task, but some work still needs improvement. The future work is to explore more advanced technology, use better models, and achieve better results in the next competition. For example, use better data

augmentation technology to increase the amount of data set, select the best hyperparameter adjustment technology, and use other pre-training models such as Longformer, ELECTRA, ALBERT or even T5.

Acknowledgments

First of all, we would like to thank the organizers for the opportunity and organization, as well as the help of teachers and seniors. Finally, we would like to thank the school for supporting my research and the patient work of future reviewers.

References

1. 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. CoRR **abs/1911.02116** (2019), http://arxiv.org/abs/1911.02116
2. 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
3. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 437–442. Association for Computational Linguistics, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/S14-2076, https://www.aclweb.org/anthology/S14-2076
4. Lample, G., Conneau, A.: Cross-lingual language model pretraining. CoRR **abs/1901.07291** (2019), http://arxiv.org/abs/1901.07291
5. Liu, F., Cohn, T., Baldwin, T.: Recurrent entity networks with delayed memory update for targeted aspect-based sentiment analysis. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 278–283. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2045, https://www.aclweb.org/anthology/N18-2045
6. Ma, Y., Peng, H., Cambria, E.: Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16541
7. 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)
8. Nguyen, T.H., Shirai, K.: PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2509–2514. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1298, https://www.aclweb.org/anthology/D15-1298

9. 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 (2019)
10. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR **abs/1802.05365** (2018), http://arxiv.org/abs/1802.05365
11. 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)
12. Plaza-del-Arco, F., Strapparava, C., Ureña-Lopez, L.A., Martin-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
13. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
14. Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3298–3307. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/C16-1311
15. Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J., Tounsi, L.: DCU: Aspect-based polarity classification for SemEval task 4. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 223–229. Association for Computational Linguistics, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/S14-2036, https://www.aclweb.org/anthology/S14-2036
16. Wang, B., Liakata, M., Zubiaga, A., Procter, R.: TDParse: Multi-target-specific sentiment recognition on Twitter. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 483–493. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1046
17. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 606–615. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1058, https://www.aclweb.org/anthology/D16-1058