Sentiment Analysis for Spanish Tweets based on Continual Pre-training and Data Augmentation

Yingwen Fu[1], Ziyu Yang[1], Nankai Lin[1], Lianxi Wang[1, 2(✉)] and Feng Chen[1]

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, China
[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou
wanglianxi@gdufs.edu.cn

Abstract. In this paper, we report the solution of the team BERT4EVER for the sentiment analysis task for Spanish tweets in EmoEvalEs@IberLEF 2021, which aims to classify Spanish tweets into one of the following emotional categories: Anger, Disgust, Fear, Joy, Sadness, Surprise or Others. We adopt the monolingual Spanish BERT model to tackle the problem. In addition, we leverage two augmented strategies to enhance the classic fine-tuned model, namely continual pre-training and data augmentation to improve the generalization capability. Experimental results demonstrate the effectiveness of the BERT model and two augmented strategies.

Keywords: Sentiment Analysis, BERT, Continual Pre-training, Back Translation, Mix up.

1 Introduction

Sentiment analysis is an important task in the field of natural language processing (NLP). It is often used to determine which type of emotion a text belongs to [1]. However, due to the lack of voice modulations and facial expressions, understanding the emotions expressed by users on social media such as Twitter is a difficult task [2]. Researchers are constantly pursuing efficient algorithms to achieve better classification results. [3, 4]

Therefore, in EmoEvalEs@IberLEF 2021 [14], a sentiment analysis task is proposed [15], requiring participants to perform sentiment analysis and evaluation of tweets in Spanish and classify them into one of the following emotional categories: Anger, Disgust, Fear, Joy, Sadness, Surprise or Others. This track provides Spanish tweets and the

IberLEF 2021, September 2021, Málaga, Spain.

corresponding categories for participants to conduct sentiment classification experiment. However, there are two main challenges for this task:

1) The dataset size is relatively small, which is far from the amount of data required for training of commonly used classification models such as BERT [5] and Bi-LSTM [6].

2) The proportion of categories is extremely imbalanced, in the provided dataset, the proportion of Fear and Disgust is much smaller than that of Others and Joy.

In tackle to the issues above, we, the BERT4EVER team, have leveraged two strategies to boost the classification performance: Continual Pre-training and Data Augmentation. These two strategies can effectively compensate for the two problems of small data size and imbalanced category proportions, so that the trained model has yielded better performance.

The remaining structure of the article is as follows. In Section 2 we will describe the task and data set given by the organizer in detail. Then in Section 3 our specific implementation is given. The final experimental results and conclusions are shown in the Section 4 and Section 5 respectively.

2 Task Description

The aim of the task is to classify the sentiment conveyed in a Spanish tweet. The task is tough because it lacks the facial expression and intonation and the sentiment can be divided into the following sentiment classes: Anger, Disgust, Fear, Joy, Sadness, Surprise or Others (the sentiment conveyed in a tweet as 'neural' or no sentiment).

The datasets [7] involved in this task were provided by the organizer of the Codalab. There are about 18,000 training datasets. In addition to the tweet, the labels of the dataset also include whether the tweet is offensive and what event the tweet is about. Some statistics about the training set are shown in Table 1.

Table 1. Statistics of the dataset.

Class	Num. of Training Instances
Happy	4908
Fear	260
Anger	2356
Surprise	952
Sad	2772
Disgust	89
Others	2356

In our conducted experiment, in order to fairly explore the effectiveness of different strategies, we leveraged 5-fold cross-validation in which we divided all the datasets into 5 parts to obtain an ensemble model with a better generalization performance. 4 parts of them are for training and the remaining part is for verification. Afterwards we leverage the average results of 5 cross models as an estimation of the effectiveness of the strategy.

3 Method

3.1 Base Model

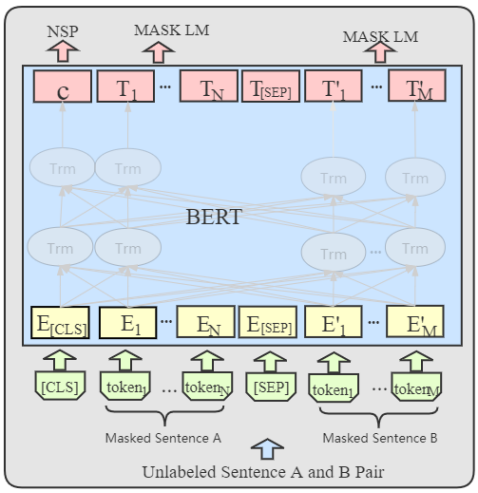

Fig. 1. BERT Model.

BERT (Bidirectional Encoder Representations from Transformers) model [5] is a pre-trained language model (PLM) which shows excellent performance on multiple downstream NLP tasks. The model architecture is shown in Fig. 1. It reads the input sequence at once and learns via two strategies, i.e., masked language modeling (MLM) and next sentence prediction (NSP). MLM is mean to randomly mask 15 percent of input words and replace them to other tokens, then predict those masked words. NSP refers to predict whether the input two sentences are consequent in the text or not to enhance the relationship between the sentences.

In this paper, **we leverage BETO** [13] **as our base model.** BETO is a BERT model trained on a big Spanish corpus Zenodo. BETO is of size similar to a BERT-Base and was trained with the Whole Word Masking technique. It uses a vocabulary of about 31k BPE [8] subwords constructed using SentencePiece and were trained for 2M steps.

However, since our data set is based on Spanish tweet, a general pre-trained model directly applied to this data set may be limited by insufficient domain knowledge. At the same time, the problem of category imbalance (as discussed in Introduction) is also a problem we need to solve. Therefore, we proposed two strategies, Continuous pre-training and Data augmentation, to alleviate the above problems.

3.2 Continual Pre-training

Inspired by [11], our continual pre-training approach to domain adaption is straightforward—we continue pretraining BETO on a large corpus of unlabeled domain-specific

text. Specifically, we try two domain corpora: (1) **Training set in EmoEvalEs@IberLEF 2021**: we ignore the labels in the training set and only use the raw text for continual pre-training. (2) **General Spanish tweet corpus + Training set in EmoEvalEs@IberLEF 2021:** in addition to the unlabeled training data in this track, we also leverage a large general Spanish tweet corpus [12] for domain-adaptive pretraining.

3.3 Data Augmentation

Data augmentation is to solve over-fitting from the data level and improve the generalization of the model. By increasing the diversity of training samples, the model can learn more essential features of the data and enhance the model's adaptability to subtle changes in samples.

Back Translation. In order to generate more training data, we use back translation generate paraphrases of an unlabeled sentence x_u in constructing x'_u. The paraphrase x'_u, generated via translating x_u to an intermediate language and then translating it back, describes the same content as x_u and should be close to x'_u semantically. In terms of the generated label, x_u and the corresponding back translation sample x'_u share the same labels. We leverage English as intermediate language in back translation.

By observing the Spanish dataset, we find that the three types of categories, Disgust, Fear, and Surprise, account for the lowest proportions. Therefore, we only perform back translation in these three categories. Increase the proportion of low-proportion categories, which not only enriches the amount of training data but also reduces the model's misjudgment rate for these three low-proportion labels.

Mix Up. Mix up [9] is a simple and quick data augmentation method. Its implementation method is to randomly extract two samples from the training sample to perform a simple random weighted summation. At the same time, the label of the sample corresponds to the weighted summation, and then the predicted result and the weighted summation loss is calculated for the subsequent tags, and finally the parameters are updated through backpropagation.

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \tilde{y} = \lambda y_i + (1-\lambda)y_j \tag{1}$$

where x_i, x_j are raw input vectors and y_i, y_j are one-hot label encodings. In this task, we simply set λ as 0.5 and get more stable predict results.

4 Experiment

4.1 Experiment Settings

We use Transformers library using Pytorch as backend to construct BERT-based models and ski-learn to construct machine learning models. The hyperparameters are shown in Table 2. As for evaluation, we leverage macro weighted averaged F1 score as our evaluation metric.

Table 2. Hyperparameters.

Parameter	Value
Learning Rate	1e-5
Batch Size	16
Epoch	15
Optimizer	Adam
Device	Nvidia 1080i

4.2 Experiment Results

We firstly report the offline results about some machine learning methods such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) and so on and latest neural methods such as fine-tuned XLM [10], fine-tuned BETO as well as some augmented strategies including continual pre-training and back translation. The results are shown in Table 3 and Table 4.

Table 3. Correspondence between model and ID.

ID	Model
1	LR
2	SVM
3	RF
4	Fine-tuned BETO
5	Fine-tuned XLM
6	ID 4 + Training set pre-training
7	ID 4 + General corpus pre-training
8	ID 6 + Whole data back translation
9	ID 6 + Low proportion data back translation
10	ID 6 + Mix up

Table 4. Offline Performance.

ID	Accuracy					
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
1	0.5163	0.5113	0.5305	0.5236	0.5236	0.5205
2	0.5351	0.5598	0.5704	0.5612	0.5797	0.5612
3	0.5346	0.5461	0.531	0.5461	0.5216	0.5367
4	0.708	0.7005	0.7133	0.7019	0.7044	0.7056
5	0.5969	0.6132	0.6158	0.6088	0.6108	0.6091
6	0.7036	0.7126	0.7197	0.7119	0.7126	0.7121
7	0.7121	0.7068	0.7112	0.7106	0.7042	0.709
8	0.7161	0.7098	0.7167	0.7172	0.7162	0.7172

9	0.7235	0.7197	0.7177	0.7294	0.7251	0.7231
10	**0.7255**	**0.7204**	**0.7233**	**0.7325**	**0.7311**	**0.7266**

From the table above, we can see that the SVM method works best in machine learning methods, outperforming LR and RF with 0.0407 and 0.0245. In addition, neural methods are far superior to machine learning methods, indicating the superiority of neural methods especially BERT-based methods. As for BERT-based method itself, we can see that the monolingual BETO achieves better performance than multilingual XLM with an improvement of almost 0.1, demonstrate the effectiveness of monolingual BETO for this task. Besides, two augmented strategies leveraged in this paper have made certain improvements to the base model, among which Mix up augmentation achieves the best effect, reaching an average accuracy of 0.7266. In addition, continual pre-training with training set and low proportion data back translation respectively outperforms continual pre-training with general corpus and whole data back translation.

Based on the offline results, we use the models (soft voting with 5 cross models) of ID 9, ID 10 and the combination of ID 9 and ID 10 (in Table 3) as our final submissions. The online results are shown in Table 5. We achieve **the second place** in the competition.

Table 5. Online Performance.

Model	Accuracy	Precision	Recall	F1-Score
ID 9	**0.7222**	0.7047	**0.7222**	**0.7114**
ID 10	0.7047	0.6927	0.7047	0.6942
Combination of ID 9 and ID 10	0.7204	**0.7082**	0.7204	0.7098

It can be seen from Table 5 that Fine-tuned BETO + Training set pre-training + Low proportion data back translation achieves the best result of 0.7222 in accuracy. It is worthwhile to note that the offline performance of Fine-tuned BETO + Training set pre-training + Mix up is excellent, but the online performance of it is not so good. That is also why the performance of the combination of the two models is not as good as that of the single model. We hold the opinion that the model training is over-fitting, resulting in poor generalization performance of the model, and thus the effect is impaired when tested on the test set.

5 Conclusion

Aiming at sentiment analysis task for Spanish tweets in EmoEvalEs@IberLEF 2021, we adopt a monolingual pre-trained Spanish BERT model as our base model and fine-tune it with the labeled tweets. In addition, focusing on two problems of small data size and class imbalance in the original training set, we leverage two augmented strategies to enhance the classic fine-tuned model, namely continual pre-training and data augmentation. Specifically, we try two data augmentation methods: back translation and

mix up. Experimental results demonstrate the effectiveness of two augmented strategies. In the future, we will further try more data augmentation methods to achieve better results on the sentiment analysis task for Spanish tweets.

Acknowledgements

This work was supported by the National Social Science Foundation of China (No. 17CTQ045), the Soft Science Research Project of Guangdong Province (No.2019A101002108), the Science and Technology Program of Guangzhou (No.202002030227), the National Natural Science Foundation of China (No. 61572145) and the Key Field Project for Universities of Guangdong Province (No. 2019KZDZX1016). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5(1), 1–167 (2012).
2. Rosenthal, S., Farra, N. and Nakov, P.: SemEval-2017 task 4: Sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 502–518. Vancouver, Canada (2017).
3. Cliche, M.: BB twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 573–580. Vancouver, Canada (2017).
4. Arasteh, S. T., Monajem, M., Christlein, V., Heinrich, P. and Evert, S.: How Will Your Tweet Be Received? Predicting the Sentiment Polarity of Tweet Replies. In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (2021).
5. Devlin, J., Chang, M. W., Lee K., and Toutanova K.: BERT: Pre-training of deep bidirectional transformers for language understanding, In: Proceedings of NAACLHLT 2019, pp. 4171-4186. (2019).
6. Hochreiter, S. and Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997).
7. Plaza-del-Arco, F. M., Strapparava, C., Urena Lopez, L. A. and 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 (2020).
8. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.:Google's neural machine translation system: Bridging the gap between human and machine translation. In: CoRR. (2016).
9. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D.: mixup: BEYOND EMPIRICAL RISK MINIMIZATION. In: Preceedings of ICLR 2018. (2018).
10. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave E., Ott M., Zettlemoyer L., and Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL 2020, pp. 8440-8451 (2020).

11. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. and Smith, N. A.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of ACL, pp. 8342—8360. Online (2020).
12. González, J. Á., Hurtado, L. F. and Pla, F.: TWilBert: Pre-trained Deep Bidirectional Transformers for Spanish Twitter. Neurocomputing 426, 58-69 (2021).
13. Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H. and Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: Preceedings of ICLR 2020. (2020).
14. Montes, M., Rooso, P., Gonzalo, J., et al.: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings. (2021).
15. 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).