Emotion Detection for Spanish with Data Augmentation and Transformer-Based Models

Hongxin Luo

School of Information Science and Engineering Yunnan University, Yunnan, P.R. China
1104792873@qq.com

Abstract. In this paper we describe the participation of Yeti team in IberLEF EmoEvalEs task, which is based on the Spanish Semantic Analysis in TASS 2020 version, and proposes as separate task for 2021 in IberLEF. We introduce the methods we used in the emotion detection task and the results obtained. First, we used back-translation data augmentation technology to solve the problems of data scarcity and data imbalance. Our method is based on transfer learning using the BETO language model for sentiment classification in Spanish. This system showed excellent performance and finally achieved an accuracy score of 0.7125. We won third place in the final result, which is only 0.0151 points away from the best result.

Keywords: Natural Language Processing, Transformers, Data Augmentation, Sentiment Analysis

1 Introduction

Sentiment analysis in tweets is a challenging task because a lot of subjective information is generated every day. It is very difficult to deal with these messages with potential language phenomena [6], and these subjective languages can be used to express private states beyond opinions [1]. People have been looking for efficient sentiment analysis algorithms based on tweets [15]. In the past few years, most of the work on sentiment analysis has combined neural network models and word embedding techniques to achieve better results [4][11]. This work is to promote the development of a Twitter sentiment classification system in Spanish.

Iberian Languages Evaluation Forum (IberLEF) is a comparative evaluation campaign for Natural Language Processing Systems in Spanish and other Iberian languages [12]. The main content of EmoEvalEs task [13] includes classifying the emotion expressed in a tweet as one of six Ekman's basic emotions [5] that best represents the mental state of the Twitter sender:*Anger, Disgust, Fear, Joy,*

IberLEF 2021, September 2021, Málaga, Spain.

Sadness, Surprise or Others. In the task, the data set is divided into training set, development set, and test set.

This article mainly summarizes our participation in Emotion Detection and Evaluation tasks [13]. According to the results of TASS 2020 [6], we can find that the performance of the BERT-based model [3] on the task is very competitive [6]. We considered a number of state-of-the-art neural network models, and finally our approach is to adaptively fine-tune the Transformer architecture based on multi-language pre-trained. We used ALBERT as the baseline for comparison.

The rest of this paper is organized as follows. Chapter 2 describes the task and the corpus. Chapter 3 introduces our system in detail. Chapter 4 introduces the experimental setup. Chapter 5 outlines the evaluation process, and the conclusions are in Chapter 6.

2 Corpus description

The organizer proposed a sentiment detection task, which is a single-label multi-classification task, which divides the sentiment labels corresponding to tweets into seven different sentiments. The seven sentiments are *Anger, Disgust, Fear, Joy, Sadness, Surprise* and *Others.* The data set is mainly collected from related events in different domains in April 2019, including entertainment, catastrophe, politics, global commemorations and global strikes [14]. The corpus is divided into three parts: training, development, and testing, with a total of 8223 items. The data in the training set and the development set have a total of five attributes, namely *id, event, tweet, offensive,* and *emotion.* The test set does not contain the emotion label. In order to prevent the classifier from relying on hashtags to classify sentiment with tweets, the organizer replaced the hashtag in the dataset with the keyword "HASHTAG" [6]. The challenges we have to face are as follows [14]:

- Lack of context: Tweets are short (up to 240 characters).
- Informal language: Misspellings, emojis and onomatopoeias are common.
- Multiclass classification: The dataset is labeled with seven different classes.

Table 1 shows the number distribution characteristics of various labels in this data set. It can be seen from the table that the distribution of data is extremely unbalanced. The largest number is the *Others* category, and the smallest number is the *Fear* category. The difference between the two is close to 43 times.

3 Materials and methods

3.1 Pre-processing

Data preprocessing is particularly important for reducing the noise information in tweets. High-quality input data can improve the output performance of the model [8]. Before conducting our experiments, we performed the following

Table 1. Distribution of sentiment labels in the corpus.

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Others	Total
Train	1227	693	589	238	111	65	2800	5723
Dev	181	104	85	35	16	9	414	844
Test	-	-	-	-	-	-	-	1656
Total	1408	797	674	273	127	74	3214	8223

preprocessing on the data respectively. First of all, we delete the URLs and punctuation marks from the text content. In order to remove unnecessary semantic information, we removed the stop words through the NLTK toolkit and converted the content of the tweets to lowercase. Finally, we also used the emoji library to convert the emojis in the tweets into text content. At the same time, we also kept the original version of the data set. In the experiment, we compared the results of various pre-processing in the experiment.

3.2 Data augmentation

Due to the extremely unbalanced distribution of the data set, the model tends to be over-fitting, predicting the most frequent category. We decided to use data augmentation technology to solve this problem. A simple and effective method is back-translation [16]. Back-translation is to translate sentences into other languages (such as Spanish to English), and then translate English back to Spanish. Check whether the newly generated sentence is different from the original sentence. If it is not the same, use the newly generated sentence as the data augmentation version of the original text. Run back-translation in multiple different languages at the same time to generate more variants. This augmentation technique helps to introduce changes in vocabulary and syntax in tweets, most of the time it can maintain the original meaning [10]. We used two representative languages (Chinese and English) to run back-translation to expand the training data, because we found during the experiments that using more languages to run back-translation does not significantly improve the experimental results. In order to obtain translation results, we used Baidu Translation API service.*

3.3 ALBERT

We used the ALBERT model as the baseline of our work, because ALBERT [9] is a newly released model that has excellent performance in various Natural Language Processing (NLP) tasks. ALBERT solves the problem of memory and training speed by designing a Lite BERT architecture, which has fewer parameters than the traditional BERT architecture [9]. The structure of ALBERT is basically the same as BERT, and there are three specific improvements, including embedding layer parameter factorization, cross-layer parameter sharing,

* Baidu Translation API available at `https://api.fanyi.baidu.com/`

Next Sentence Prediction (NSP) task is changed to Sentence Order Prediction (SOP) task. The hyperparameter settings of the model are as follows (the settings found to perform well in several fine-tunings; the parameters not mentioned keep the default values):

- **albert_model :** albert_base_v2
- **max_seq_length :** 128
- **optimizer :** AdamW
- **warmup_step :** 200
- **learning_rate :** 3e-5
- **train_step :** 800
- **train_batch_size :** 64

3.4 BETO

Inspired by the results of the TASS 2020 seminar and our emotion classification task, we decided to use the BETO model to complete this challenging Emotion Detection and Evaluation for Spanish task. BETO is a BERT model trained on a large Spanish corpus. BETO model combines the pre-training model with the downstream task model, which means that the BETO model is still used when doing downstream tasks, and it naturally supports text classification tasks, and there is no need to modify the model when doing text classification tasks. BETO is trained on a large Spanish corpus, which can more accurately represent the text features of Spanish, and can solve the problem of the dependence of the task on the Spanish language. It has BETO-uncased and BETO-cased. We used BETO-cased as our Language Model (LM). The size of BETO is similar to BERT-base, according to the guidelines presented by Cañete et al. [2], BETO has received Whole Word Masking (WWM) training, and both use about 31k Byte Pair Encoder (BPE) subwords constructed by Sentence Piece Vocabulary list, and have been trained in 2M steps [2]. In the training process, the dynamic mask technology is introduced, which is to use 10 different masks for the same sentence in the corpus. When using WWM to mask a specific token, if the token corresponds to a subword in a sentence, all consecutive tokens that make up the same word will be masked. We use the ADAM optimizer [7] for optimization. We hope to use BETO as a basic initial LM to construct a robust method to complete this challenge and achieve excellent performance in the final result. The settings of the optimal hyper-parameters in the experiment are as follows (the settings that performed well in several fine-tunings; the parameters not mentioned keep the default values):

- **beto_model :** BETO-cased
- **max_seq_length :** 128
- **train_batch_size :** 32
- **learning_rate :** 2e-5
- **num_train_epochs :** 3.0

4 Experimental Setup

In this section we introduce our experimental procedure. In order to compare the results of this experiment, we will use ALBERT as the baseline. We compare with the BETO model with the best result obtained by fine-tuning during the experiment (the hyperparameter settings are shown in the introduction in section 3.3), both are pre-trained deep models. The IberLEF organization released three corpora for training, development, and testing. The label of each tweet corresponds to one of the 6 emotions.

In the exhaustive search process with the BETO model as the main research object, we determined the model configuration parameters (as shown in section 3.4). Inspired by TASS 2020 Task 1, through observing the data, we found that the tweet also corresponds to an *Offensive* label, so based on the nearly determined model configuration parameters, we tried to input the *Offensive + Tweet* content into the model for prediction and compare with the result of input only *Tweet*.

Then we processed the unbalanced data in the corpus by using back-translation data augmentation technology, mainly using Chinese and English as intermediate languages for back translation. We mainly enhance the two very few categories (Fear and Disgust) to expand the data volume and prevent the model from overfitting and predicting a large number of categories.

Finally, we also tried to convert the emojis in data into corresponding content texts to explore better model performance. We input the processed data into the two models for comparison. All experiments are performed on a machine equipped with Nvidia GPU (Tesla V100).

5 Results

The results of our model on the validation set of the Emotion Detection and Evaluation for Spanish task are shown in Table 2. Our final submission results and rankings on the official test set are shown in Table 5. The final result of our system is quite competitive. In the final submitted result, it got the fourth place overall with a score of 0.7125, which is only 0.0151 behind the best result.

The results given in Table 4 show that the results obtained by using the BETO model are better than the baseline, and at the same time far better than the BERT model using the multi-language model. This fully shows that the result of using the specific languages pre-trained model is higher than that of the multi-language pre-trained model. Through Table 2 we also observe that our data preprocessing does not promote the performance of the model. Compared with the unprocessed raw data, the effect is reduced. After our discussion, we concluded that we believe that the reason for the drop in results may be related to the pre-trained of the model. The pre-trained of the original BERT model is to ensure that the context and semantic connections can be learned, and the input data set is a raw material that has not undergone any pre-processing raw data, and I added preprocessing when fine-tuning downstream tasks, which

Table 2. The results of BETO model on the development set. Pre-1 means data preprocessing for deleting URLs, punctuation, and stop words; pre-2 means data preprocessing for deleting URLs and punctuation; Pre-3 means data preprocessing for deleting only URLs; Input means that the input data only contains one column of *Tweet* or contains two columns of *Offensive + Tweet*.

Back-translation	Input	pre-processing	ACC
No	Offensive + Tweet	No	**0.7215**
		Pre-1	0.6919
		Pre-2	0.6990
		Pre-3	0.6919
	Tweet	No	**0.7156**
		Pre-1	0.7097
		Pre-2	0.6954
		Pre-3	0.7014
Yes	Offensive + Tweet	No	**0.7322**
		Pre-1	0.6931
		Pre-2	0.6978
		Pre-3	0.7298
	Tweet	No	**0.7132**
		Pre-1	0.6966
		Pre-2	0.7002
		Pre-3	0.7061

Table 3. Comparison of results before and after data augmentation.

Model	Data	Acc
BETO	raw	0.7049
	back-translation	**0.7322**
	back-translation + emojis-to-text	0.7315

may destroy the contextual text relationship, which will result in poor results. Finally, it can be seen from Table 3 that the back-translation data augmentation technology we used is helpful to the improvement of model performance, and the conversion of emojis into text content also slightly improves the effect. Our final submission results and rankings on the official test set are shown in Table 5.

6 Conclusions

We propose a BETO-cased sentiment classification system for IberLEF 2021 EmoEvalEs task. This method is based on BETO transfer learning. It is mainly applied to the sentiment analysis of Spanish tweets, which includes an additional data augmentation step, and has achieved good results in the Spanish task. We are very satisfied with the results of our first participation in the IberLEF workshop. Although the method is relatively simple, it is important that we

Table 4. Comparison of the results of our model on the validation set.

Model	Data augmentation	Pre-process	Acc
Baseline	Yes	No	0.6866
BERT	Yes	No	0.6646
BETO	Yes	No	**0.7322**

Table 5. Our final submitted results and rankings.

Team Name	Acc	MP	MR	MF_1	Rank
GSI-UPM	0.7276	0.7094	0.7276	0.7170	1
Yeti	**0.7125**	**0.7044**	**0.7125**	**0.7054**	**3**
qu	0.4498	0.6188	0.4498	0.4469	15

have achieved very good results in the task by exploring the hyperparameters of the model and configuring our model reasonably in the task. Careful selection of language models and data augmentation techniques play an important role in sentiment analysis of small sample data set. However, there are still huge challenges in sentiment analysis regarding the content of tweets, and our system still has a lot of room for improvement. In the future work, I hope to use more powerful data augmentation technology to solve the problem of data scarcity. We look forward to further exploring more advanced techniques to solve the sentiment analysis of Spanish tweets.

Acknowledgments

First of all, thank the organizer for the valuable opportunity provided to us. Then I would also like to thank the teacher for supporting my research work and the patience of future reviewers.

References

1. Algeo, J.: A comprehensive grammar of the english language. by randolph quirk, sidney greenbaum, geoffrey leech, and jan svartvik. london: Longman. 1985. x+ 1779. Journal of English Linguistics **20**(1), 122–136 (1987)
2. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Díaz Galiano, M.C., Martínez Cámara, E., García Cumbreras, M.Á., García Vega, M., Villena Román, J.: The democratization of deep learning in tass 2017 (2018)
5. Ekman, P.: Are there basic emotions? (1992)

6. García-Vega, M., Díaz-Galiano, M.C., García-Cumbreras, M.Á., del Arco, F.M.P., Montejo-Ráez, A., Jiménez-Zafra, S.M., Cámara, E.M., Aguilar, C.A., Antonio, M., Cabezudo, S., et al.: Overview of tass 2020: introducing emotion detection (2020)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Data preprocessing for supervised leaning. world academy of science, engineering and technology. International Journal of Computer, Electrical, Automation, Control and Information Engineering **1**, 4104–4109 (2007)
9. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
10. Luque, F.M.: Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. arXiv preprint arXiv:1909.11241 (2019)
11. Martínez Cámara, E., Almeida-Cruz, Y., Díaz Galiano, M.C., Estévez-Velarde, S., García Cumbreras, M.Á., García Vega, M., Gutiérrez, Y., Montejo Ráez, A., Montoyo, A., Munoz, R., et al.: Overview of tass 2018: Opinions, health and emotions (2018)
12. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez Carmona, M., 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. 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)
14. 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
15. Villena Román, J., Lana Serrano, S., Martínez Cámara, E., González Cristóbal, J.C.: Tass-workshop on sentiment analysis at sepln (2013)
16. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)