Humor Analysis in Spanish Tweets with Multiple Strategies

Lianxi Wang[1,2], Xiaotian Lin[1], Nankai Lin[1(✉)], Yingwen Fu[1], Kaiying Wu[1] and Jiajun Wu[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
neakail@outlook.com

Abstract. In this article, we report the solution of the team BERT 4EVER for the Humor Analysis based on Human Annotation task in IberLeF 2021, which aims to identify humorous articles from a computational perspective. We propose the BERT model to tackle the problem. In addition, we leverage various strategies including pseudo-label technology, Task-Adaptive Pre-training and ensemble learning to improve the generalization capability. Experimental results as well as the leading position our team on the task leaderboard demonstrate the effectiveness of our method with the first place in two subtasks

Keywords: Humor Analysis, Multiple Strategies, BERT.

1 Introduction

Humor, a complex phenomenon in human communication that results in amusement or laughter, not only serves to interchange information or share implicit meaning, but also engages a relationship between those exposed to the funny message. However, while humor has been historically studied from a psychological, cognitive and linguistic standpoint, there have been only few attempts to create computational models for humor recognition or generation. Besides, the existing research mainly focuses on high-resource languages such as Chinese and English and a characterization of humor that allows its automatic recognition and generation is far from being specified.

Luckily, HAHA @IberLEF 2021 propose the task "Humor Analysis based on Human Annotation" [1], which aims to gain a better insight into what is humorous and what causes laughter, and propose to go further in the direction of analyzing humor structure and content. During the task, four subtasks are proposed: (1) Humor

IberLEF 2021, September 2021, Málaga, Spain.

Detection: determining if a tweet is a joke or not (intended humor by the author or not). (2) Funniness Score Prediction: predicting a Funniness Score value for a tweet in a 5-star ranking, assuming it is a joke. (3) Humor Mechanism Classification: for a humorous tweet, predict the mechanism by which the tweet conveys humor from a set of classes such as irony, wordplay, hyperbole, or shock. (4) Humor Content Classification : for a humorous tweet, predict the content of the joke based on its target (what it is making fun of) from a set of classes such as racist jokes, sexist jokes, dark humor, dirty jokes, etc. Our team, BERT 4EVER, also participated in this task and achieved good results with the first place in two subtasks. In this report, we will review our solution to this task, namely, the BERT model aided by pseudo-label technology, task-adaptive pre-training and teacher-student network with MSE loss function.

2 Related Work

To our best knowledge, the existing researches on humor detection are mainly focus on identifying whether the text is humor or not and humor rating.

Previous research for humor recognition is mainly based on taking the problem into account as a classification problem. Barbieri et al. [2] proposed to train classification procedures with a rich set of features and representation though casting it as a classification problem. Chen et al. [3] presented a Convolutional Neural Network (CNN) for humor recognition concentrating on lexical cues. Furtherly, Zhang et al. [4] designed several simple but effective features to capture the emotionality and subjectivity in humorous texts, which enables the model to make full use of the contextual knowledge.

Although humor recognition has commonly been regarded as a binary classification task, recent works have further toward humor detection as a relative ranking task. Semeval-2017 Task 6 [5] asked competitors to predict the ranking gave by the comedy program's audience and producers using the humorous tweets submitted to a comedy program. To better identify funnier captions, Shahaf et al. [6] proposed to analyze the caption pairs. Besides, they further find significant differences between the funnier and less-funny captions.

As regards to Spanish language, Castro et al. [7] construct a tweet corpus labeled as humor/no humor and a funniness score from 1 to 5 for Spanish humor recognition tasks containing 27000 tweets. Ortega-Bueno et al. [8] proposed to combine both linguistic features and an Attention-based Recurrent Neural Network, where the attention layer helps to calculate the contribution of each term towards targeted humorous classes.

3 Method

3.1 Humor Detection

BERT. BERT [9], a language model based on bidirectional encoder characterization, is designed to pretrain deep bidirectional representations via two unsupervised subtasks (namely Mask Language Model and Next Sentence Prediction) from unlabeled text by

jointly conditioning on both left and right context in all layer, meaning it can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as text classification, without substantial tasks-specific architecture modifications. Based on these, we choose BERT as our language model shown in Figure 1 to conduct our own various strategies on it.

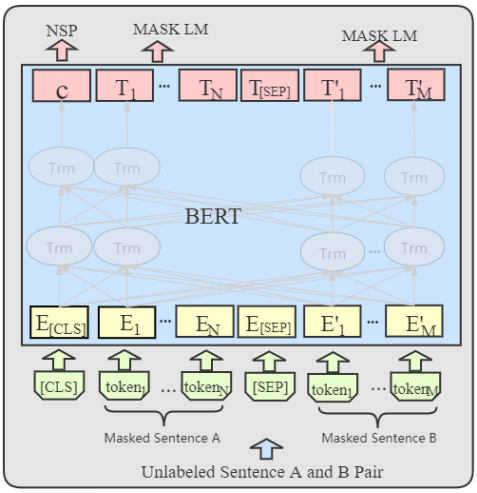

Fig. 1. BERT Model.

Task-Adaptive Pre-training. Following Gururangan et al. [10], Our approach to task-adaptive pretraining is straightforward—we continue pretraining BERT on the unlabeled training set provided by HAHA @IberLEF 2021. Specially, we select the data whose labels are marked as 1 from the training set in HAHA@IberLEF 2021. Compared to the BERT without task-adaptive pre-training or using all the training data, it uses a smaller pretraining corpus but one that is much more task-relevant to further improve the performance of the task.

MSE Strategy. According to Hinton et al. [11], they pointed out that crudely using one-hot encoding may lose additional information for different labels for classification tasks and proposed to utilize the probability outputted by the teacher network to instruct the student network for re-training. Inspired by their research, we train a teacher model with the training data and use its probabilities to the loss function MSE and Cross Entropy to train the student model with the same training data.

Five-fold Cross-validation Models Fusion. In our conducted experiment, in order to fairly increase the robustness of the model, 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 1 part is for verification. Afterwards we leverage the average results of 5 cross models as an estimation of the effectiveness of the strategy.

3.2 Funniness Score Prediction

During this task, our method consists of three models with five-fold cross-validation: XGBoost with word frequency matrix, LightGBM with TFIDF and BERT.

XGBoost. XGBoost [12], a highly effective and widely used machine learning method, achieves state-of-the-art results on many machine learning challenges and Regression tasks. Based on its scalability in all scenarios and algorithmic optimizations, we choose this model with word frequency matrix to select text features.

LightGBM. Being a variant of XGBoost, it outperforms the XGboost model in terms of accuracy and speed. For the shortcomings of the XGboost model, the LightGBM [13] proposed two effective methods called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to tackle them. During this paper, we leverage this model and select the text features with TFIDF to achieve the task of Funniness Score Prediction.

Multiple Models Fusion. To our best knowledge, different models usually focus on different information about the same task, which will cause the difference performance on regression tasks. Based on this, in order to further increase the robustness of the model, we merge these three models with five-fold Cross-validation during the task of Funniness Score Prediction.

3.3 Humor Mechanism Classification

Although manually annotating dataset is expensive, it is relatively easy to collect massive unlabeled data in the target domain. Hence, it becomes desired to improve the generalization of the model by leveraging unlabeled data D_U and limited labeled data D_L [14][15]. In this competition, only 4800 pieces of data were annotated with the humor mechanism tag, which were regarded as the limited labeled data D_L. The rest of the data is considered as the unlabeled data D_U.

In this task, we also use the task-adaptive pretraining model in the humor detection task. We further use data D_L to train a humor mechanism classification model M_1. We use model M_1 to predict the label of the unlabeled data D_U, and keep samples with label probability greater than 0.8. Through this strategy, we obtained a total of 1940 pseudo-labeled data D_P. We merge the labeled data D_L and the pseudo-labeled data D_P to train a new model M_2 which has the stronger generalization capability. In the final evaluation phase, we use M_2 to predict the test data.

3.4 Humor Target Classification

Similar to the humor mechanism classification task, we use pseudo-label technology to solve the task. However, we did not use the task-adaptive pretraining model in the humor detection task. At the same time, because humor target classification is a multi-label classification task, that is, a sample may have zero or more labels for humor target, we added a “None” label on the basis of the original label set. In addition, since many of the 4800 annotated samples have no labels, when generating the pseudo-labeled data set, we only selected the data whose prediction results contained one or more labels. The pseudo-labeled data set contains a total of 774 samples.

4 Experiment

4.1 Experiment Settings

We use Transformers[2] library using Pytorch[3] as backend to construct BERT-based models and scikit-learn[4] to construct machine learning models. What's more, we leveraged BETO[5] [16] as our base model. The hyperparameters are shown in Table 2.

Table 1. Hyperparameters.

Parameter	Value
Learning Rate	5e-5
Batch Size	16
Epoch	15
Optimizer	Adam
Task-Adaptive Pre-training Epoch	3

4.2 Results

Table 2. Results for Humor Detection Task.

Method	F1-validation	F1-test
BERT	83.06%	-
BERT with pretrained	85.35%	85.00%
Teacher-student network with pretrained and MSE[6]	86.32%	86.45%

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.

The experimental results of humor detection task are shown in Table 2. The experimental results show the effectiveness of the task-adaptive pre-training strategy and the MSE strategy. Among them, the model that uses two strategies at the same time has the best performance, and the F value is 86.45% on the final test data set.

Table 3. Results for Funniness Score Prediction Task.

Method	MMR-validation	MMR-test
BERT	0.6471	-
BERT with pretrained	0.6470	0.6673
XGBoost	0.6470	0.6615

[2] https://github.com/huggingface/transformers
[3] https://github.com/pytorch/pytorch
[4] https://github.com/scikit-learn/scikit-learn
[5] https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
[6] We only used the four folds fusion as the final submission result since the fifth fold did not perform well.

LightGBM	0.6532	0.6643
Merge	-	0.6587

The results in Table 3 show that on the funniness score prediction task, among the three base models, XGBoost has the best performance. The results on the validation set and test set are 0.6470 and 0.6615, respectively. In the end, the multiple models fusion strategy brought further improvement to the task. The result on the test set was 0.6587, ranking fourth place among all teams.

Table 4. Results for Humor Mechanism Classification Task.

Method	F1-validation	F1-test
BERT	26.50%	29.99%
BERT with pretrained	28.50%	32.27%
BERT with pretrained and pseudo-label	33.67%	33.96%

Table 5. Results for Humor Target Classification Task

Method	F1-validation	F1-test
BERT	26.50%	37.20%
BERT with pretrained	23.47%	29.42%
BERT with pseudo-label	29.31%	42.28%

We achieved the first place in the leaderboard in both the humor mechanism classification task and the humor target classification task. Whether on validation and test sets, we can see that pseudo-label technology has brought significant improvements to the model.

5 Conclusion

Aiming at humor analysis task for Spanish tweets in HAHA@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, for different tasks, we leverage different strategies to enhance the classic fine-tuned model. Experimental results demonstrate the effectiveness of our method. In the future, we will further try more strategies to achieve better results on the humor 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. Luis, C., Santiago, C., Santiago, G., Aiala R., Meaney, J. A. and Rada, M.: Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. Procesamiento del Lenguaje Natural 67. (2021).
2. Barbieri, F., Saggion, H.: Automatic Detection of Irony and Hu-mour in Twitter. In: International Conference on Computer and Communications, pp. 155-162. (2014).
3. Mihalcea, R., Strapparava, C.: Making computers laugh: Investi-gations in automatic humor recognition. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 531-538. (2005).
4. Zhang, D., Song, W., Liu, L., Du, C., et al.:(2017, December). Investigations in automatic humor recognition. In: 2017 10th International Symposium on Computational Intelligence and Design. pp. 272-275. IEEE (2017).
5. Potash, P., Romanov, A., Rumshisky, A.: SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 49–57. Association for Computational Linguistics, Vancouver, Canada (2017).
6. Shahaf, D., Horvitz, E., Mankoff, R.: Inside jokes: Identifying hu-morous cartoon captions. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1065-1074. (2015).
7. Castro, S., Chiruzzo, L., Rosá, A.et al.: A crowd-annotated Spanish corpus for humor analysis. CORR. (2017).
8. Ortega-Bueno, R., Muniz-Cuza, C. E., Pagola, J. E. M.: UO UPV: Deep linguistic humor detection in Spanish social media. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, pp. 204-213. (2018).
9. Devlin, J., Chang, M. W., Lee K., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding, In: Proceedings of NAACLHLT 2019, pp. 4171-4186. (2019).
10. Gururangan, S., Marasović, A., Swayamdipta, S., et al.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of ACL, pp. 8342—8360. (2020).
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CORR. (2015).
12. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. (2016).
13. Ke, G., Meng, Q., Finley, T., et al.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 3146-3154(2017).
14. Lee, D.: Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In: ICML 2013 Workshop: Challenges in Representation Learning. pp. 1-6. (2013).
15. Shi, W., Gong, Y., Ding, C., et al.: Transductive semi-supervised deep learning sing min-max features. In Proceedings of the European Conference on Computer Vision. pp. 299-315. (2018).

16. 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).