Applying Pre-trained Model and Fine-tune to Conduct Humor Analysis on Spanish Tweets

Yongyi Kui

Information Institute of Yunnan University, Yunnan,China
3964438@qq.com

Abstract. This paper describes in detail the four subtasks of HAHA@ IberLEF 2021 [7] : Humor Analysis based on Human Annotation. Subtask 2 is a regression problem, and the other three subtasks are all text classification problems. The data comes from the Twitter social platform, and the language is Spanish. The classification problem is mainly solved by integrating the Multilingual Bert model and the LSTM model, and the regression problem is solved by the GPT-2 Model. According to the official evaluation results, the result of the method proposed in this paper ranks fourth, eighth, fifth, and sixth on the four subtasks, respectively. For this task, I have uploaded the code to GitHub kuiyongyi for easy reference by others.

Keywords: Spanish Text Classification · Humor Analysis · Pre-trained Model · Fine-tuning.

1 Introduction

Humor is a very common phenomenon in human communication. It is relatively easy for humans to understand whether the content of a text is humorous, but the computer can learn the characteristic information in the text, only after learning a large amount of corpus and then detect whether the text is humorous [14].

Humor detection has been a relatively hot field for many years. Semeval-2015 Task 11 proposes the influence of figurative language such as metaphor and irony on sentiment analysis. Semeval-2017 Task 6 requires predicting the ranking of comedy shows based on humorous tweets of comedy shows. Both IberEVAL 2018 and IberLEF 2019 include two subtasks: humor detection and interest score prediction. Castro [6] established a corpus of annotated tweets, allowing annotators to judge subjectively which tweets are humorous, and then build a humor classifier for Spanish tweets based on the method of supervised learning. Barbieri and Saggion [2] proposed a machine learning method based on a set of language-driven features. Radev [15] described a method for humor

IberLEF 2021, September 2021, Málaga, Spain.

detection in cartoon subtitles. Yang [16] constructed different classifiers through feature sets to detect humor. However, many existing systems use a combination of Pre-trained Models and fine-tuning [9] to achieve humor detection.

The rest of the paper is organized as follows. In the second part of the paper, I give an overview of the task and the data. The third part describes the model used in this task. In the fourth part, the experiments of the four subtasks are described. The fifth part gives the test results of the model. Finally, the sixth part of the article is the conclusion.

2 Task and Data Description

2.1 Subtasks

HAHA@IberLEF2021 includes four sub-tasks, which aim to predict the humor classification or humor degree of the text. The four subtasks are defined as follows:

Subtask 1: Given the Twitter text in Spanish, the goal is to predict whether each tweet is humorous. This is a binary classification problem. This question is measured by F1 score.

Subtask 2: To measure the degree of humor on the text predicted to be humorous, this is a regression task. The evaluation standard of this subtask is the root mean square error.

Subtask 3: It is required to predict a humor mechanism category for tweets among the twelve humor mechanisms given. This is a multi-classification problem. The performance of this task will be measured using the Macro-F1 score.

Subtask 4: Given tweets and fifteen humorous target tags, the goal is to predict the corresponding humorous target tags for each tweets (at least zero and at most 15 tags), which is a text multi-label classification task. The evaluation standard for this question is also Macro-F1.

2.2 Dataset

The Dataset [8] provided by HAHA@IberLEF 2021 includes three Subdatasets: haha_2021_train (24,000 tweets), haha_2021_dev (6,000 tweets), and haha_2021_test (6,000 tweets).

Each record in haha_2021_train consists of twelve columns of data, which are respectively Id, the tweets corresponding to each Id (text includes punctuation and emoticons), non-humorous voting, the number of votes for each of the five levels, is humor, humor rating, humor mechanism, and humor_target.

2.3 Data cleaning

Since the text in the HAHA Dataset comes from Twitter social media, the length of the text varies, most of the text is generally short, and the content or format of the tweets is informal, (including spelling errors, character, and vocabulary

repetition). It is necessary to clean up the tweets in the Dataset, because in this way the Language Model can predict the content to be covered based on the context and semantics. The data cleaning methods involved in this paper are as follows:

- Replace the repeated characters or words (three times or more) in the tweet with a single character or word;
- Delete Delete the emoticons appearing in the text;
- Delete Delete the HTML tags that appear in the text;
- Delete Replace the newline character in the text with spaces.

3 System Description

In this section, I will give an overview of the system that implements these four subtasks. We can simply regard the task of Humor Analysis based on Human Annotation in IberLEF 2021 as text classification and regression problems, given input text, predict one or more labels, and predict the humor score value of the text.

One disadvantage of the LSTM model is that the input information is compressed into a vector with the same length as the number of LSTM memory units, so the LSTM model cannot remember long tweets. However, the Twitter text length of this task is very short, so I consider using LSTM to further extract text feature information after the output of the Pre-training Model.

3.1 Model

Since the text of the Dataset is Spanish, this paper uses three Cross-language [12] Pre-training Models of Multilingual Bert [1] , XLM, and XLM-RoBerta, as well as models such as XLNet, Albert, and GPT-2 to extract text feature information. Here is a brief introduction to each Pre-training Model.

The Bert model uses Masked Language Modeling [11] to train the bidirectional Transformers to generate deep bidirectional language representations. After the Pre-training phase is over, you just need to add an output layer for fine-tuning. This time, Bert-base-multilingual-uncased is used. This model uses a corpus of one hundred and two languages, including Spanish, during the Pre-training phase.

Albert uses Transformer and GELU activation function. Albert uses such as parameter sharing and matrix decomposition to reduce the number of model parameters. Albert uses Sentence Order Prediction Loss to replace Next Sentence Prediction Loss.

XLM is a Cross-language model, similar to Bert, it is also a Masked Language Model, and its Pre-training method is the next token prediction. In this task, XLM-mlm-tlm-xnli15-1024 is used. The model uses fifteen corpus including Spanish for Cross-language sentence training.

XLM-Roberta is a model trained on a corpus of one hundred different languages. Unlike XLM, it doesn't require lang tensors to understand which language is used and should be able to determine the correct language from the input ids. In this task, XLM-RoBerta-base is used.

XLNet mainly optimized the Bert model in three aspects. First, the Auto-encoding Model [4] is replaced with an Autoregressive [10] Model, and then the Transformer of the Bert model is improved with Transformer-xl, and XLNet uses a dual-stream [13] attention mechanism. In the Pre-training phase, the Next Sentence Prediction method of the Bert model is discarded.

The GPT-2 model is usually padding the inputs on the right and was trained using causal language modeling targets. The GPT-2 model and the Bert model are constructed through the decoder and encoder modules of the transformer, respectively.

3.2 Parameter setting

In subtask 1, the optimizer chooses AdamW, train for 50 epochs with a 3e-5 learning rate and a 32 batch size. And the weight_decay parameter is set to 1e-2, the maxlen parameter value is set to 64, and the loss function is selected as Crossentropy.

In subtask 2, the loss function used is mse_loss, the learning rate is set to 1e-5, the maxlen parameter value is set to 64, batch_size is set to 32, and after every sixty-four steps of training, a verification is carried out.

In subtask 3 and subtask 4, the learning rate and weight_decay parameters are both 5e-6 and 1e-2, the epoch is 50 and 100 respectively, the loss function uses CategoricalCrossentropy and BinaryCrossentropy respectively, both the drop_out parameter and the optimizer are set to 0.5 and Adam.

4 Experiments

4.1 Subtask 1

Divide haha_2021_train according to the ratio of 4/1 as the Training Dataset and Validation Dataset of the models in subtask 1. The actual Training Dataset, Validation Dataset, and Testing Dataset lengths of subtask 1 are 19200, 4800, and 6000 respectively. Next, the data is cleaned, word segmented, and coded before it is ready to be input into the model.

In subtask 1, First, I used three Pre-trained Models (Albert-base-v2, XLNet-base-cased and Bert-base-multilingual-uncased) for text binary classification. The results show that the Multilingual Bert model is used to obtain the highest score in the text classification problem of Spanish materials. The F1 score of this model on the Validation Dataset is 0.8712, and then add a 4-layer unidirectional LSTM network after the model to further extract text features, and finally input the LSTM results into the fully connected layer for classification. This combination has an F1 score of 0.8785 on the Validation Dataset, compared

Table 1. Performance of the four models on the Validation Dataset of subtask 1, performance is measured by F1 value

Model	F1
XLNet-base-cased	0.8611
Albert-base-v2	0.8653
Bert-base-multilingual-uncased	0.8712
Bert-base-multilingual-uncased+LSTM	**0.8785**

Table 2. The performance of the four pre-trained Language models on the Validation Dataset in subtask 2, measured using root-mean-squared error (RMSE)

Model	RMSE
Bert-base-multilingual-uncased	0.6833
XLM-RoBerta-base	0.6761
XLM-mlm-xnli15-1024	0.6719
GPT-2	**0.6683**

to a single pre-trained model, the score increased by 0.0073. Thus, the prediction result of this combined model is the answer I finally submitted on subtask 1. The performance of these models on the Validation Dataset is shown in Table 1.

4.2 Subtask 2

There are 9253 pieces of data available for subtask 2 in the haha_2021_train. Divide the data according to the method of subtask 1. The Training Dataset length of subtask 2 is 7402, the Validation Dataset length is 1851, and the Testing Dataset length is 6000.

In subtask 2, I compared the performance of three Cross-language Pre-training Models and the GPT-2 model in the regression problem. These four Pre-training Models are Bert-base-multilingual-uncased, XLM-mlm-xnli15-1024, XLM-RoBerta-base, and GPT-2. The experimental results show that the performance of the GPT-2 model in subtask2 is slightly better than the other three models, therefore, I use the prediction result of the GPT-2 model as the answer I finally submitted in subtask 2. The specific performance of these four models on the Validation Dataset is shown in Table 2.

4.3 Subtask 3 & 4

In haha_2021_train, humor_mechanism column and humour_target column have 4800 and 1629 non-empty data respectively. Divide the data according to the method of subtask 1. The Training Dataset lengths of subtask 3 and subtask 4

Table 3. The performance of our solutions on the Validation Datasets of subtask 3 and subtask 4 respectively, The evaluation standard of them are both Macro-F1.

Model	subtask 3 Macro-F1	subtask 4 Macro-F1
Bert-base-multilingual-uncased	0.2433	0.2927
Bert-base-multilingual-uncased+BiLSTM	0.2516	0.3044

are 3840 and 1303, respectively, the Validation Dataset lengths are 960 and 326, respectively, and, the length of the test Dataset for both is 6000.

After dividing the data, first, encode the labels of the Training Dataset and Validation Dataset in subtask 3 and subtask 4 into one_hot vectors. Next, the Dataset is processed in the same way as subtask 1. Then, use the sequential function to build a double-layer BiLSTM [5]. In the double-layer BiLSTM model, LSTM Cells are set to 192 and 64 respectively, and the return_sequences parameter is set to True and False respectively. Subtask 3 is a multi-class classification problem, while subtask 4 is a multi-label classification. The biggest difference between the two is only the activation function used in the last layer of the network. The activation function used in subtask 3 is Softmax, while subtask 4 uses Sigmoid.

The results of subtask 1 show that the Multilingual Bert model performs slightly better than Albert and XLNet in the Spanish Twitter text classification task. So in subtask 3 and subtask 4, I use the multilingual Bert model as the basis. First, input the processed data into the Multilingual Bert model; next, input the result of the Bert model into the two-layer BiLSTM network to further extract features; finally, input the result of the LSTM network into the fully connected layer (The number of neurons in the fully connected layer of subtasks 3 and 4 are twelve and fifteen respectively) for classification. Subtask 3 finally outputs a twelve-dimensional vector, taking the element with the largest value upwards to 1, and setting the remaining elements to 0 (The 1 and 0 in subtask 3 and 4 respectively indicate that it is predicted to be or not predicted to be a certain category). The output of subtask 4 is a fifteen-dimensional vector, and the threshold is set to 0.5. For each element in the output vector, if it is greater than the threshold, it is set to 1 upwards, otherwise, it is set to 0 downwards. I used the prediction results of the combined model of Multilingual Bert and BiLSTM as the answers I finally submitted on the last two subtasks. The performance of my solution on verification sets of subtask 3 and subtask 4 is shown in Table 3.

5 Results

Among all the teams participating in this task, the result of my solution ranked fourth, eighth, fifth, and sixth among the four subtasks. Table 4 lists the scores of the best performing teams in the four subtasks of the IberLEF 2021 HaHa competition, and the scores of my method on the official Testing Dataset.

Table 4. The final evaluation result of my solution in the Humor Analysis based on Human Annotation Forum challenge, the best results of each subtask and the baseline score

Method	subtask 1 F1	subtask 2 RMSE	subtask 3 Macro-F1	subtask 4 Macro-F1
Winning approach	0.8850	0.6226	0.3142	0.3787
Our proposal	**0.8681**	**0.6797**	**0.2187**	**0.2836**
Baseline	0.6619	0.6704	0.1001	0.0527

6 Conclusion

This paper describes the data processing, the comparison of Pre-trained Models, and the final model construction in the HAHA@ IberLEF2021 challenge. Although the solution I proposed has achieved good results, it is undeniable that there is still a lot of room for improvement. Due to time constraints, I can not do error analysis at present. In my future work, I will conduct a detailed error analysis to understand the limitations of the program.

In future work, first of all, we can try to extract the emoticon information in tweets instead of deleting it directly. If both the emoticons and the text information can be extracted, it will promote humor classification. Secondly, during the experiment, I found that the model had overfitting problems. For this reason, I will try to use transfer learning [3] to make improvements in the future. Finally, the distribution of data of each category in subtask 3 and subtask 4 is not balanced, especially in subtask 4. For this reason, I will consider dealing with unbalanced data by setting the weight of the loss function.

References

1. Azhar, A.N., Khodra, M.L.: Fine-tuning pretrained multilingual bert model for indonesian aspect-based sentiment analysis (2021)
2. Barbieri, F., Saggion, H.: Automatic detection of irony and humour in twitter. In: ICCC. pp. 155–162 (2014)
3. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML workshop on unsupervised and transfer learning. pp. 17–36. JMLR Workshop and Conference Proceedings (2012)
4. Bi, B., Li, C., Wu, C., Yan, M., Wang, W., Huang, S., Huang, F., Si, L.: Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation (2020)
5. Bohnet, B., Mcdonald, R., Simes, G., Andor, D., Maynez, J.: Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018)
6. Castro, S., Cubero, M., Garat, D., Moncecchi, G.: Is this a joke? detecting humor in spanish tweets. In: Springer International Publishing (2016)

7. Chiruzzo, L., Castro, S., Góngora, S., Rosá, A., Meaney, J.A., Mihalcea, R.: Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. Procesamiento del Lenguaje Natural **67**(0) (2021)
8. Chiruzzo, L., Castro, S., Rosá, A.: Haha 2019 dataset: A corpus for humor analysis in spanish. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 5106–5112 (2020)
9. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D.: Codebert: A pre-trained model for programming and natural languages (2020)
10. Gong, X.R., Jin, J.X., Zhang, T.: Sentiment analysis using autoregressive language modeling and broad learning system. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2019)
11. Goyal, N., Du, J., Ott, M., Anantharaman, G., Conneau, A.: Larger-scale transformers for multilingual masked language modeling (2021)
12. Li, J., He, R., Ye, H., Ng, H.T., Bing, L., Yan, R.: Unsupervised domain adaptation of a pretrained cross-lingual language model. arXiv preprint arXiv:2011.11499 (2020)
13. Li, R., Li, S.: Human behavior recognition based on attention mechanism. In: 2020 International Conference on Artificial Intelligence and Education (ICAIE). pp. 103–107. IEEE (2020)
14. Mihalcea, R., Strapparava, C.: Making computers laugh: investigations in automatic humor recognition. In: Conference on Human Language Technology Empirical Methods in Natural Language Processing (2005)
15. Radev, D., Stent, A., Tetreault, J., Pappu, A., Iliakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Jaimes, A., Jha, R., et al.: Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. arXiv preprint arXiv:1506.08126 (2015)
16. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015)