ColBERT at HAHA 2021: Parallel Neural Networks for Rating Humor in Spanish Tweets

Issa Annamoradnejad[1] [0000−0003−3147−6389]
and Gohar Zoghi[2] [0000−0003−0298−4069]

[1] Department of Computer Engineering, Sharif University of Technology, Iran
i.moradnejad@gmail.com
[2] Golestan University of Medical Sciences, Iran
zoughi.g@goums.ac.ir

Abstract. Previously, we proposed ColBERT, a humor detection model based on the general linguistic structure of humor for formal English texts. ColBERT uses BERT model to produce embeddings for the text sentences, which will be put as inputs into a parallel neural network. In this paper, we utilized the proposed model on informal Spanish texts to detect humor and rate its level. The current task has three differences compared to the original humor detection task on the ColBERT dataset: (1) rating humor is a regression task rather than binary classification, (2) texts are informal, and (3) texts are in a different language. Using our general model and without any knowledge of the Spanish language, we participated in HAHA shared task at IberLEF 2021 Forum and achieved 2nd place for humor rating and 3rd place for binary humor detection. The results confirm robustness of our proposed model.

Keywords: humor rating · parallel neural networks · computational humor · informal texts · Spanish tweets

1 Introduction

Computational humor detection is a an ongoing research track that has several delicacies due to the linguistic features of humor and the various mechanisms that can be incorporated to bring laughter in humans.

In previous works, researchers mostly focused on the binary task of humor detection, where the goal is to separate humor from non-humor texts. However, it would also be beneficial to rate the level of humor existing in a humorous text. The new task would contribute in fixing the level of humor for chatbots based on the mood of user or query.

Computational humor has a long list of history, from using statistical and N-gram analysis [10], Regression Trees [7], Word2Vec combined with K-NN Human

IberLEF 2021, September 2021, Málaga, Spain.

Centric Features [12], Convolutional Neural Networks [3, 11], and pre-trained models based on transfer learning [1].

Previously, we proposed the ColBERT model [1] based on the linguistic features of humor for the binary task of humor detection in formal English texts. Our approach separates sentences and uses the English BERT-base-uncased pretrained model to encode them into sentence embeddings. They are separately fed into parallel hidden layers of neural network to extract mid-level features for each sentence (related to context, type of sentence, etc). The final layers combine the output of all previous lines of hidden layers in order to predict the final output. In theory, these final layers should determine the congruity of sentences and detect the transformation of reader's viewpoint after reading the punchline.

In this paper, we aim to test the robustness of the model by applying the model on a new context, which has three new challenges compared to the previous one:

1. The new task is on Spanish language, a language that we have no previous knowledge.
2. The input texts are informal texts of Twitter users.
3. The new task is about rating humor, thus it is a regression task.

We did not change the model structure, its hyper-parameters, or any of the pre-processing functions. However, in order to extract sentence embedding for Spanish texts, we changed the pre-trained model from the English BERT-base-uncased to a recent Spanish equivalent (BETO-uncased [2]).

To evaluate our performance, we participated in HAHA shared task at IberLEF 2021 Forum [4]. Thus the paper is also a description of the methods that we used for the competition track. The competition task is divided into four sub-tasks all of which target informal Spanish tweets. While the task include four sub-tasks (Table 1), we particularly focused on the first two sub-tasks that required the least amount of modifications on our original model.

Table 1. The four sub-tasks at the HAHA task 2021 [4]

Sub-task Title	Description
Humor Detection	A binary classification of humor.
Humor Rating	Predicting a score between 1 to 5 for a tweet assuming it as a joke.
Humor Mechanism Classification	A multi-class classification task that predicts the mechanism by which the tweet conveys humor, such as irony, wordplay or exaggeration.
Humor Target Classification	A multi-label classification task that predicts the target or context of the tweet, such as weight, racist, sexist, etc.

2 Proposed Method

In this section, we present the three conceptual steps that lead to the proposed method. First, we will explore the linguistic structure that we particularly focused to achieve this method. Then, we explain the overall structure of the ColBERT model, and finally, we give a description of the changes that we made in this paper.

2.1 Humor Structure

Many linguists theorized that humor arises from the sudden transformation of an expectation into nothing [6]. Therefore, the structure of a joke generally includes two or three stages of storytelling that ends with a punchline [5, 9]. Punchline as the last part of a joke brings laughter through its incongruity to the perceiver's previous expectations.

Based on Raskin's Semantic Script Theory of Humor (SSTH) [8], humor has the necessary condition of having two distinct related scripts opposite in nature, such as real/unreal, possible/impossible. This is compatible with the two-staged theory which ends with a punchline. While the punchline is related to previous sentences, it is included as an opposition to previous lines in order to transform the reader's expectation of the context.

2.2 ColBERT Model

Based on the linguistic theories on the structure of humor, if one reads each sentence of a joke separately, it will be perceived as normal and non-humorous texts. However, if we try to comprehend all sentences together in one context or in one line of story, the text becomes humorous. ColBERT model utilizes this linguistic characteristic of humor in its structure.

In short, it uses separate paths of hidden layers especially designed to extract latent features from each sentence. In addition, it has an additional path to extract latent features of the whole text. Hence, the neural network structure includes one parallel path to view text as a whole and several other paths to view each sentence separately. It is composed of a few steps (Figure 1)[1]:

1. Separate sentences and tokenize them individually.
2. Convert textual parts to proper numerical inputs for the neural network, using a pre-trained BERT based model. This step is performed individually on each sentence (left side in Figure 1) and also on the whole text (right side in Figure 1).
3. Feed embeddings into parallel hidden layers of neural network to extract mid-level latent features for each path (related to context, type of sentence, etc). The output size is 20 for each path.
4. Feed embeddings for the whole text in the last path, similar to previous step. The output size is 60. We do this as there may exist meaningful relationships, such as synonyms and antonyms.

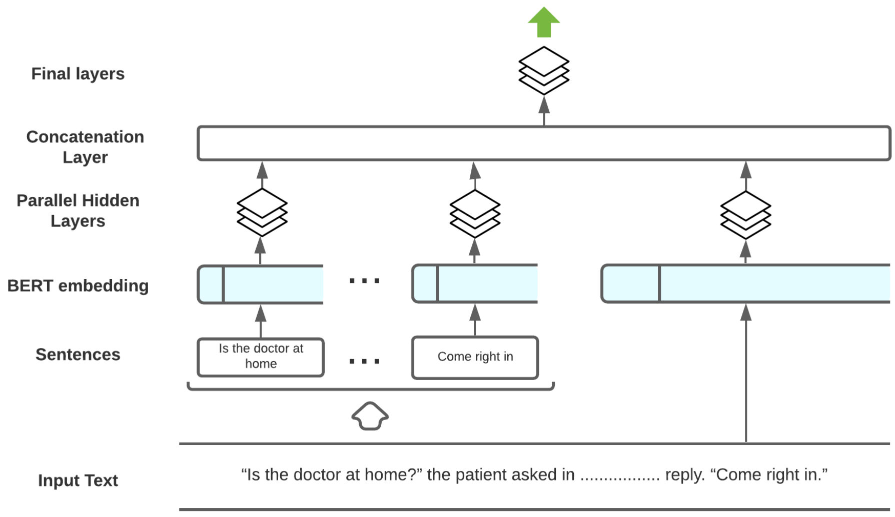

Fig. 1. Architecture of the ColBERT method. [1]

5. Three sequential layers combine the output of previous paths of hidden layers in order to predict the final output. In theory, these final layers should determine the congruity of sentences and detect the transformation of reader's viewpoint after reading the punchline.

2.3 Changes for the Spanish Language

In order to change the target context from formal English texts to informal Spanish tweets, we kept the model structure and hyper-parameters as before. However, in order to extract sentence embedding for Spanish texts, we changed the pre-trained model from the English BERT-base-uncased to a recent Spanish equivalent. This was a required and logical step to achieve meaningful embeddings. For this goal, we selected BETO model [2] from a long line of huggingface models.

BETO [2] is a BERT model trained on a big Spanish corpus that has a similar size to the BERT-base model and was trained with the Whole Word Masking technique. Compared to other BERT based models proposed for the Spanish language (including multilingual BERT models), it achieved higher accuracy in several selected tasks.

We tried a few pre-processing methods for separating sentences and cleaning the text that were previously proposed for the Spanish language, but they were unsuccessful in achieving higher accuracy in our evaluations. Therefore, we split the sentences as before, and did not perform any data cleaning.

2.4 The multi-class and multi-label sub-tasks

For the last two sub-tasks, we reduced multi-class and multi-label classification to multiple regression tasks using classical one-against-all (OAA) approach. In this approach, we kept the single-target model structure as before and added one post-processing step to accumulate all results and predict the final value.

For the multi-class classification task, we predicted the probability of each class separately and used the class with the maximum predicted value as our final prediction. For the multi-label classification sub-task, we used a threshold (0.5) to select all labels applicable for the given text.

3 Results and Discussions

17 teams participated in HAHA 2021 shared task. Based on the official results reported by the organizers, we managed to achieve the 2$^{\text{nd}}$ place for humor rating sub-task with a very close score to the first team (0.002 difference). As reported in Table 2, our model achieved 0.6246 score for the official test data based on Root-Mean-Squared-Error (RMSE) metric. The first three teams are separated by a huge gap from the rest of participants, even from the fourth team which also used BERT language model for their predictions.

Table 2. Performance of our model in rating humor compared to other teams (Evaluation by RMSE)

Rank	Model	Score (RMSE)
1	UMUTeam	0.6226
2	**ColBERT**	0.6246
3	Jocoso	0.6296
4	BERT4EVER	0.6587
...	...	...

As we mentioned earlier, HAHA 2021 also organized three more sub-tasks about computational analysis of humor, all of which are evaluated using F1-Score. While our focus was on the task of humor rating, we managed to use our model to participate in those sub-tasks. Table 3 compares our performance with other top-ranking teams.

For the binary humor detection sub-task, we did not have to perform any extra steps or modifications on our method. Our proposed method was able to achieve 3$^{\text{rd}}$ place among 17 teams with 0.8696 F1-score. Compared to our previous evaluations for humor detection on formal English texts, this is a 10 percent drop in F1-score, which can be attributed to the informality of texts, our lack of knowledge on Spanish language, and weaker cleaning and sentence embedding methods.

For the last two sub-tasks, we managed to achieve 7th and 5th places, accordingly (Table 3). This lower performance compared to the first two sub-tasks can largely be attributed to the simple taken approach (OAA) and the naive threshold for all classes (0.5).

Table 3. Comparison of our model with other high-scoring models for the rest of sub-tasks (all evaluated by F1-score)

Model	Humor Detection	Humor Mechanism	Humor Target
	(binary)	(multi-class)	(multi-label)
Jocoso	0.8850	0.2916	0.3578
icc	0.8716	0.2522	0.3110
kuiyongyi	0.8681	0.2187	0.2836
BERT4EVER	0.8645	0.3396	0.4228
UMUTeam	0.8544	0.2087	0.3225
Baseline	0.6619	0.1001	0.0527
ColBERT	0.8696 (3rd)	0.2060 (7th)	0.3099 (5th)

4 Conclusions

In this paper, we showed robustness of our proposed method for computational humor, through its performance evaluation on rating and detecting humor in informal Spanish tweets. The new context has three important challenges compared to the previous task of detecting humor in formal English texts.

For the new context, we did not change any part of our proposed method and only replaced the chosen pre-trained model that we used for generating sentence embeddings from an to a state-of-the-art Spanish model. For the multi-class and multi-label classification sub-tasks, we reduced them to multiple regression tasks and achieved the final output using one-against-all (OAA) approach.

We participated at HAHA 2021 competition at IberLEF 2021 Forum, and competed alongside several teams from all over the world. Based on the official results, our general method was able to achieve the second place for rating humor and third place for detecting humor in Spanish tweets.

5 Acknowledgements

We would like to thank the organizers of HAHA shared task for supporting research interest in the field of computational humor in the past few years.

References

1. Annamoradnejad, I., Zoghi, G.: Colbert: Using bert sentence embedding for humor detection. arXiv preprint arXiv:2004.12765 (2020)

2. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
3. Chen, P.Y., Soo, V.W.: Humor recognition using deep learning. 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. 113–117 (2018)
4. 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)
5. Eysenck, H.J.: The appreciation of humour: an experimental and theoretical study 1. British Journal of Psychology. General Section **32**(4), 295–309 (1942)
6. Kant, I.: Kritik der urteilskraft, vol. 39. Meiner (1913)
7. Purandare, A., Litman, D.: Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 208–215 (2006)
8. Raskin, V.: Semantic mechanisms of humor, vol. 24. Springer Science & Business Media (2012)
9. Suls, J.M.: A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. The psychology of humor: Theoretical perspectives and empirical issues **1**, 81–100 (1972)
10. Taylor, J.M., Mazlack, L.J.: Computationally recognizing wordplay in jokes. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 26 (2004)
11. Weller, O., Seppi, K.: Humor detection: A transformer gets the last laugh. arXiv preprint arXiv:1909.00252 (2019)
12. 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. pp. 2367–2376 (2015)