

# A BERT-based Approach for Automatic Humor Detection and Scoring

Jihang Mao<sup>1</sup>, Wanli Liu<sup>2</sup>

<sup>1</sup> Montgomery Blair High School,  
51 University Blvd E, Silver Spring, MD 20901, USA

<sup>2</sup> TAJ Technologies, Inc.,  
7910 Woodmont Ave #1214, Bethesda, MD 20814, USA

jim-blair@hotmail.com

**Abstract.** In this paper we report our participation in the 2019 HAHA task where a corpus of crowd-annotated tweets is provided and required to tell if a tweet is a joke or not and predict a funniness score value for a tweet. Our approach utilizes BERT, a multi-layer bidirectional transformer encoder which can help learn deep bi-directional representations, and the pretrained model is fine-tuned on training data for HAHA task. The representation of a tweet is fed into an output layer for classification. To predict the funniness score, we apply another output layer to generate scores by using float labels and train it with the mean squad error between the prediction scores and the labels. Our best F-Score on the test set for Task 1 is 0.784 and RMSE for Task 2 is 0.910. We find that our approach is competitive and applicable to multilingual text classification tasks.

**Keywords:** BERT; Text Classification; Score Prediction; Multilingual Model; Evaluation.

## 1 Introduction

The exponential growth of user-generated-content demands an effective way to detect specific information that is relevant to an automatic text classification task. Automatic humor detection has drawn attention within this domain, as one of the challenging research field in Natural Language Processing [1]. It is because there are many types of humor such as irony, wordplay, metaphor and sarcasm [2], but few formal taxonomies of humor characteristics, making the task of automated humor recognition difficult. In addition, humor plays a role that can be viewed as a sociological phenomenon and function, while the extent to which a person may sense humor depends on her/his personal background. There exist some previous works [3-5], but a characterization of humor that allows its automatic recognition and generation is far from being specified.

Humor is an example of figurative language and has relevance in the field of sentiment analysis and opinion mining [6, 7]. There are some previous researches regarding this

topic. Semeval-2015 Task 11 proposed to work on figurative language, such as metaphors and irony, but focused on Sentiment Analysis [8]. Semeval-2017 Task 6 presented a similar task to this one as well. Majority of the researches on social media texts is focused on English. However, Schroeder’s work [9] shows that a high percentage of these texts are in non-English languages. HAHA - Humor Analysis based on Human Annotation, is a task to classify tweets in Spanish as humorous or not, and to determine how funny they are [10]. The 2019 edition of HAHA is the 2nd year round of the shared task. The aim of this task is to gain better insight in what is humorous and what causes laughter.

Based on tweets written in Spanish, the HAHA task comprises two subtasks: Humor Detection (Task 1) - telling if a tweet is a joke or not (intended humor by the author or not) and Humor Scoring (Task 2) - predicting a funniness score value (average stars) for a tweet in a 5-star ranking. We participated in both sub-task this year. A corpus of crowd-annotated tweets based on [11] is provided and divided in 80% for training and 20% tweets for testing. Participating teams should include a row for each one of the 6000 tweets in the test corpus. All tweets are classified as humorous or not humorous by the “is\_humor” column. For Task 2, all the rows have a “funniness\_average” as a predicted score.

A brief description of our method for HAHA task is presented in Section 2. In Section 3 we show the results of our method on the official HAHA test datasets. In section 4 we present a discussion of the results and conclusions of our participation in this challenge.

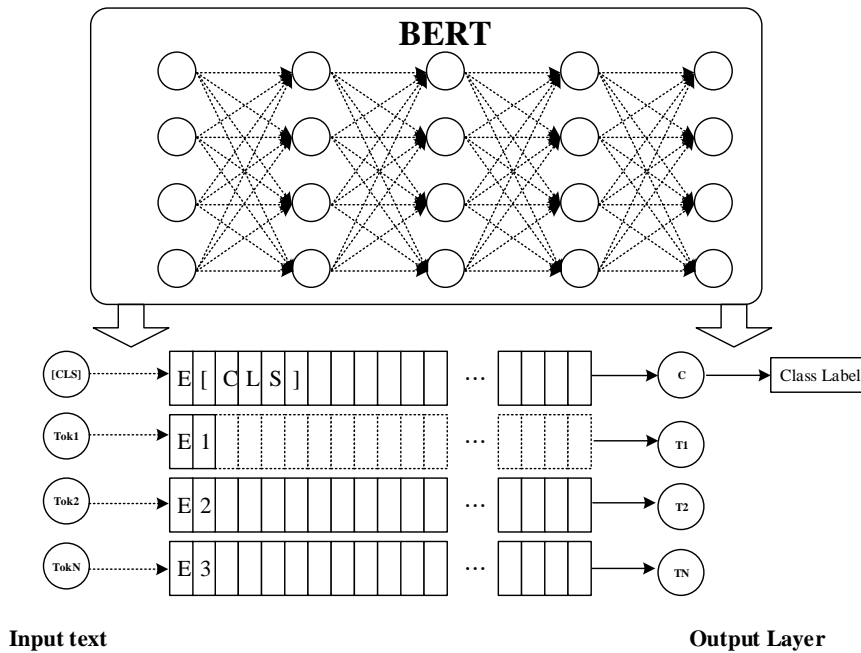
## 2 Methods

For HAHA Task, our approach builds on BERT, which has obtained state-of-the-art performance on most NLP tasks [12]. More specifically, given a tweet, our method first obtains its token representation from the pre-trained BERT model using a case-preserving WordPiece model, including the maximal document context provided by the data. Next, we formulate this as a single-sentence classification task by feeding the representation into an output layer, a binary classifier over the class labels. Finally, we apply another output layer to generate scores by using float labels and train it with the mean squad error between the prediction scores and the labels.

BERT utilizes a multi-layer bidirectional transformer encoder which can learn deep bi-directional representations and can be later fine-tuned for a variety of tasks such as NER. Before BERT, deep learning models, such as Bi-directional Long Short-Term Memory (Bi-LSTM) and convolutional neural network (CNN) have greatly improved the performance in text classification over the last few years [13]. More recently, ULMFiT [14] has proved the ability of transfer learning for any NLP task and achieved great results.

The pre-trained BERT models are trained on a large corpus (Wikipedia + BookCorpus). There are several pre-trained models releases. In HABA, we chose BERT-Base, multilingual cased model for the following reasons: First, multilingual model is better for Spanish documents in HABA because the English-only model splits tokens not available in its vocabulary into sub-tokens, which will affect the accuracy of the classification task. Second, although BERT-Large generally outperforms BERT-Base in English NLP tasks, BERT-Large versions of multilingual models haven't been released. Third, the multilingual cased model fixes normalization issues in many languages, so it is recommended in languages with non-Latin alphabets (and is often better for most languages with Latin alphabets). In Task 1, we use the final hidden state corresponding to a special token ([CLS]) as the aggregate sequence representation, then feed it into an output layer for classification (Figure 1).

**Fig. 1.** Architecture of our model for sentence classification. Similar to [11], we denote input embedding as  $E$ , the final hidden vector of the special [CLS] token, and the final hidden vector for the  $i$ th input token as  $T_i$



For Task 2, we make several changes to the above output layer. First, we change the label type to a float instead of an int. Second, we change the measure for training to mean squad error, Pearson and Spearman correlation instead of accuracy. Finally, we change the output to a scorer instead of a classifier. Since high recall (0.825) and low precision (0.724) were observed while submitting results, we utilize the scores in Task

2 to optimize the F1 score in Task 1, i.e. changing the tweet classified as “Is humorous” in Task 1 to “Not humor” if it has received a low score in Task 2.

### 3 Results & Discussion

The HAHA corpus has been divided into two subsets: the train and the test set. The training set contains 24000 tweets, and the test set 6000 tweets.

In Task 1, the results are measured using accuracy and F-measure for the humorous category. F-measure is the main measure for this task. In Task 2, The results are measured using root-mean-squared error. Here we present the results on the test set. In our best submission, the model was fine-tuned using the hyperparameter values suggested in [11]: learning rate (Adam)= $2e-5$ , number of epochs=3, max sequence length=256, and batch size=16. When fine-tuning the model for HAHA task, we randomly sample the training set into two subsets: 18000 tweets for training, and 6000 tweets for development. We use a new set of random seeds each time to prevent over-fitting. In addition, tweets classified as “Is humorous” with prediction score  $< 0.2$  are reclassified as “Not humor”, while tweets classified as “Not humor” with prediction score  $> 1.7$  are reclassified as “Is humorous” in the final submission. Table 1 shows the improvement obtained using the scores of Task 2 to optimize Task 1.

As shown in Table 2 and Table 3, our best submission significantly outperformed the Baseline “hahaPLN” in both F-measure for Task1 and root mean squared error for Task 2, while the F1 score of our submission is near to the highest score for Task 1 (-0.037). We are in the first third of all participants in both Task 1 and Task 2, which demonstrates a good performance of our system in detecting humor tweets in Spanish and predicting the funniness scores.

**Table 1:** The improvement obtained using the scores of Task 2 to optimize Task 1

Systems	F1	Precision	Recall	Accuracy
Original	0.771	0.724	0.825	0.809
After optimizing	0.784	0.745	0.827	0.822

**Table 2:** Official results for Task 1 of HAHA

Systems	F1	Precision	Recall	Accuracy
Our proposal	0.784	0.745	0.827	0.822
Baseline	0.440	0.394	0.497	0.505
Winning approach	0.821	0.791	0.852	0.855

**Table 3:** Official results for Task 2 of HAHA, measured using root-mean-squared error (RMSE)

Method	RMSE
Our proposal	0.910
Baseline	2.455
Winning approach	0.736

## 4 Conclusion

We described our BERT-based approach that participated in the HAHA - Humor Analysis based on Human Annotation task in IberLEF 2019. Compared to previous methods, our approach has several significant differences from system architecture to the processing flow. It is a general and robust framework and showed competitive performance during the HAHA evaluations. With more and more training corpora available, we plan to explore the application of it in other text classification task in future work.

## Acknowledgements

The authors would like to thank Dr. Yutao Zhang for providing Jihang Mao the intern opportunity at George Mason University and valuable suggestions and comments on the manuscript. The authors would also like to thank the HAHA task organizers for providing the data of the task.

## References

1. Attardo, S. Linguistic theories of humor, volume 1. Walter de Gruyter (1994).
2. Raz, Y. Automatic humor classification on twitter. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 66–70. Association for Computational Linguistics. (2012).
3. Mihalcea, R., & Strapparava, C. Making Computers Laugh: Investigations in Automatic Humor Recognition. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05, (pp. 531–538). Association for Computational Linguistics, Vancouver, British Columbia, Canada. (2005).
4. Sjöbergh, J., and Araki, K. Recognizing Humor Without Recognizing Meaning. In WILF, (pp. 469–476). Springer, (2007).
5. Castro, S., Cubero, M., Garat, D., & Moncecchi, G. Is This a Joke? Detecting Humor in Spanish Tweets. In Ibero-American Conference on Artificial Intelligence (pp. 139-150). Springer International Publishing. (2016).

6. Reyes, A., Rosso, P., and Veale, T. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 1–30. (2013).
7. Pang, B., and Lee, L. Opinion Mining and Sentiment Analysis. *Found. Trends Information Retrieval* 2(1-2):1–135. (2008).
8. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 470-478). (2015).
9. Schroeder, S. Half of messages on twitter aren't in english[stats]. (2010).
10. Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J. & Rosá, A. Overview of Haha at IberLEF 2019: Humor Analysis based on Human Annotation. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR Workshop Proceedings, Bilbao, Spain, (2019)*.
11. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., & Moncecchi, G. A crowd-annotated spanish corpus for humor analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (pp. 7-11). (2018).
12. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Minneapolis, Minnesota, USA, (2019).
13. Zhang, T., Huang, M., & Zhao, L. Learning structured representation for text classification via reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*. (2018).
14. Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics. (2018).