ZYJ at SemEval-2021 Task 7: HaHackathon: Detecting and Rating Humor and Offense with ALBERT-Based Model

Yingjia Zhao
Yunnan University / Yunnan, P.R. China
zyj1309700118@gmail.com

Xin Tao
Yunnan University / Yunnan, P.R. China
taoxinwy@126.com

Abstract

Although humorous language can bring joy to people, it is also easy to cause offense. Therefore, in order to effectively identify whether a sentence is humorous or offensive, the system needs to obtain abundant semantic information. This article introduces the submission of subtask 1 and subtask 2 that we participate in SemEval-2021 Task 7: HaHackathon: Detecting and Rating Humor and Offense, we use a model based on ALBERT that uses ALBERT as the module for extracting text features. We modify the upper layer structure by adding specific networks to better summarize the semantic information. Finally, our system achieves an F-Score of 0.9348 in subtask 1a, RMSE of 0.7214 in subtask 1b, F-Score of 0.4603 in subtask 1c, and RMSE of 0.5204 in subtask 2.

1 Introduction

A sense of humor is a positive psychological quality that can make people feel better about themselves, relieve stress, and gain a sense of connection with others. Humor can help bring about happiness experiences and optimism. Humor, by its nature, stirs up emotions. The speaker presents the audience with an unexpected conflict. The audience feels nervous and anticipatory, and at the same time feels pleased and released. Humor is also a highly subjective phenomenon, with age, gender, and socioeconomic status is known to influence the perception of jokes. That's why some things make people laugh and others don't. When the brain lacks the cognitive resources to accurately understand the context in which a joke takes place, it generalizes it into an everyday behavior, and the benign offense becomes hostile aggression. Like most metaphorical languages, humor emphasizes multiple word meanings, cultural knowledge, and pragmatic capabilities, making it a challenging task to detect humor and offense in a sentence.

SemEval-2021 Shared Task 7: HaHackathon: Detecting and Rating Humor and Offense (Meaney et al., 2021) shared task has two subtasks. Subtask 1 emulates previous humor detection tasks in which all ratings were averaged to provide mean classification and rating scores. Subtask 1a: predict if the text would be considered humorous (for an average user). This is a binary task. Subtask 1b: if the text is classed as humorous, predict how humorous it is (for an average user). The values vary between 0 and 5. Subtask 1c: if the text is classed as humorous, predict if the humor rating would be considered controversial, i.e. the variance of the rating between annotators is higher than the median. This is a binary task. Subtask 2 aims to predict how offensive a text would be (for an average user) with values between 0 and 5. This score was calculated regardless of whether the text is classed as humorous or offensive overall.

In this paper, we use ALBERT: A Lite BERT for self-supervised Learning of Language (Lan et al., 2019) as the module for extracting sentence semantic information. The features extracted by ALBERT are then further processed through a specific structure. Besides, for subtasks 1a and 1c, since it is a binary classification problem, we use k-fold stratified sampling to reinforce the training process. The rest of the paper is organized as follows. Part 2 gives a brief introduction to the relevant work. Part 3 describes the dataset and our approach. Part 4 describes the hyperparameters of the study method used and our results. Finally, the fifth part summarizes our work.

2 Related Work

Computational research in the field of humor detection has been going on for some time, and the diversity of tasks allows for different analyses according to the type and expression of humor. In

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1175–1178
Bangkok, Thailand (online), August 5–6, 2021.

previous shared tasks, T3 team (Vanroy et al., 2020) used the pre-trained language model Roberta in the SemEval-2020 shared task7 on Assessing the Funninness of Edited News Headlines (Hossain et al., 2020) to learn the latent features in news headlines and predict how funny each headline is. UniTuebingenCL team (Ammer and Grüner, 2020) used a Ridge Regression model using Elmo and Glove embeddings as well as Truncated Singular Value Decomposition at SemEval-2020 Task 7: Humor Detection in News Headlines. Humor and emotions such as offense and hatred are not only different but also related to one another. It is often challenging to classify them accurately. Badlani et al. (2019) proposed a composite two-step model. In the first step, features related to irony, humor, hate speech, and emotion are extracted, and in the second step, these features are combined to classify emotions. This multi-step method is better than a single step. Models that predict sentiment have better empirical performance in sentiment classification.

Sentiment analysis of humor data requires a deep semantic understanding of the text, and signals of nuances in the language may enhance or completely change the sentiment of the sentence. Morales and Zhai (2017) proposed a generative language model based on incongruity theory to model humorous text, using background text sources, such as Wikipedia entry descriptions, and being able to construct multiple features to identify humorous comments. Besides, Deep Learning (DL) (Goodfellow et al., 2016) methods of multi-layer Neural Networks (NN) (Mikolov et al., 2011) stacked was also a common method. Ortega-Bueno et al. (2018) used a recurrent neural network(RNN) (Mikolov et al., 2010) that combines language features and attention-based to classify Spanish tweets as humorous or not and predict how funny they are. The attention (Vaswani et al., 2017) layer helps calculate the contribution of each term to the target humor category. In recent years, various pre-training models based on Transformer have shown outstanding performance, and some researchers have also applied them in the field of humor detection. Weller and Seppi (2019) used a Transformer framework to evaluate whether a joke was humorous, and performed well on the short joke and pun datasets.

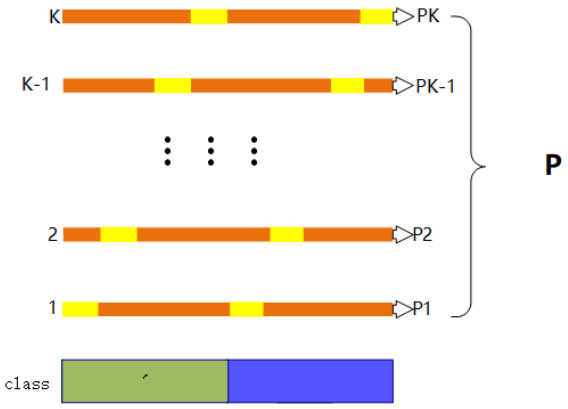

Figure 1: K-fold stratified sampling to the training set

3 Methodology and Data

3.1 Data description

The organizer of Semeval-2021 Shared Task 7 provided a complete training data set for our system, including 8000 pieces of data for the training set, 1000 for the validation set, and 1000 for the test set. Each piece of data has four tags.

In our experiment, for subtasks 1a and 1c, we use a Stratified-K-fold technique to randomly segment all combined training datasets. As shown in Figure 1, we use the Stratified-K-fold cross-validation instead of the ordinary K-fold cross-validation. Stratified-K-fold can ensure that the proportion of each class in the generated training set and validation set is consistent with the original training set, thus avoiding the generated data distribution disorder. In this experiment, we set the value of K as 5.

3.2 Description of the system

Our model is one based on ALBERT, which is shown in Figure 2. BERT (Devlin et al., 2018) has good performance, but too many parameters and long training time are also its disadvantages. ALBERT is designed by Google mainly to solve the problem of BERT, which is a simplified model based on BERT. ALBERT solves these problems by using two parameter reduction techniques, one of them is cross-layer parameter sharing, in order to avoid quantity increases along with the network depth; factorized embedding parameterization is another approach, by putting a big word embedded matrix is decomposed into two small matrix makes the relationship between the size of the hidden layer and dictionary apart, thus, when the size of the hidden layer is increased, the parameter size of

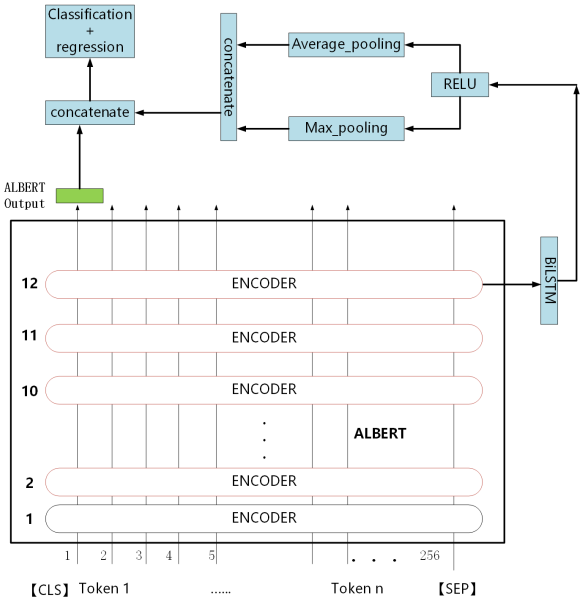

Figure 2: Schematic overview of the architecture of our model

the vocabulary embedding will not be significantly increased.

We input the preprocessed text into the model through the input layer, and then carry out vector representation. Position vector representation, text vector representation, and word vector representation make up the vector representation. Then the ALBERT model takes the sum of the input embedding, position encoding, and token type embedding as input, which is further processed by the Transformer Encoder module. After that, we input the output of the last ALBERT hidden layers into BiLSTM. Then, through the Relu function, the vector is nonlinearly mapped to the lower dimension. After average pooling and maximum pooling are achieved to obtain the feature vector, we concatenate the average pooling and maximum pooling output. Finally, the feature vector will concatenate with the original output of ALBERT, the classification or regression task is then performed.

4 Experiment and results

4.1 Experiment setting

In this experiment, albert_base_v2 is used. After adding a new module based on ALBERT, the whole model is fine-tuned. The main hyper-parameters we adjust are the maximum sequence length, the learning rate, the gradient accumulation steps, and batch size. As is shown in Table 1.

maximum sequence	**length learning rate**
128	2e-5
gradient accumulation	**steps batch size**
4	4

Table 1: Details of the hyper-parameters.

Team Name: ZYJ	
Task 1a Humor Detection	
F-Score	Rank
0.9348	41
Task 1b Average Humor Score	
RMSE	Rank
0.7214	43
Task 1c Humor Controversy	
F-Score	Rank
0.4603	33
Task 2 Average Offensiveness Score	
RMSE	Rank
0.5204	31

Table 2: The results of our methods.

4.2 Results

According to the leaderboard provided by the organizer, our team's F-score is 0.9348, ranking 41st place in subtask 1a Humor Detection. RMSE is 0.7214, ranking 43rd place in subtask 1b Average Humor Score. F-Score is 0.4603, ranking 33rd place in subtask 1c Humor Controversy. RMSE is 0.5204, ranking 31st place in subtask 2 Average Offensiveness Score. As shown in Table 2.

5 Conclusion

In this task, we detect and rate humor and offense using a deep-learning-based model. In the construction of the model, we use ALBERT as a module of the model and add a custom network structure to further process the extracted feature vector. As for the classification task, we perform Stratified-K-Fold cross-validation based on the model and get the optimal value through the voting mechanism. Although our system has fewer parameters and is easier to train, the performance needs to be improved. In future work, we will further optimize the system structure, so that the model can obtain rich semantic information characteristics.

References

Charlotte Ammer and Lea Grüner. 2020. Unituebingencl at SemEval-2020 task 7: Humor detection in news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1060–1065.

Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. Disambiguating Sentiment: An Ensemble of Humour, Sarcasm, and Hate Speech Features for Sentiment Classification. *W-NUT 2019*, page 337.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. *arXiv preprint arXiv:2008.00304*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky. 2011. Rnnlm-recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, pages 196–201.

Alex Morales and ChengXiang Zhai. 2017. Identifying humor in reviews using background text sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. 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 (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 204–213.

Bram Vanroy, Sofie Labat, Olha Kaminska, Els Lefever, and Véronique Hoste. 2020. Lt3 at SemEval-2020 task 7: Comparing feature-based and transformer-based approaches to detect funny headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1033–1040. International Committee for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. *arXiv preprint arXiv:1909.00252*.