cs_english@LT-EDI-EACL2021: Hope Speech Detection Based On Fine-tuning ALBERT Model

Shi Chen
School of Information Science and Engineering Yunnan University,
chen-s2020@139.com

Bing Kong
School of Information Science and Engineering Yunnan University,
kongbing@ynu.edu.cn

Abstract

This paper mainly introduces the relevant content of the task "Hope Speech Detection for Equality, Diversity, and Inclusion at LT-EDI 2021-EACL 2021". A total of three language datasets were provided, and we chose the English dataset to complete this task. The specific task objective is to classify the given speech into 'Hope speech', 'Not Hope speech', and 'Not in intended language'. In terms of method, we use fine-tuned ALBERT and K fold cross-validation to accomplish this task. In the end, we achieved a good result in the rank list of the task result, and the final F1 score was 0.93, tying for first place. However, we will continue to try to improve methods to get better results in future work.

1 Introduction

As we all know, we are currently facing an incompletely harmonious and secure network environment. Nowadays, the number of Internet users is very large, especially the proportion of minors is steadily increasing, which shows how important it is to create a hopeful social media environment.

Such an environment that embodies equality, tolerance, and diversity can help people who are in depression, confusion, lack of identity and other difficulties gain hope. At the same time, it also brings a better online social experience to the entire user community on social media.

Currently, the detection and classification of social media speech are mostly biased towards offensive and controversial speech detection, but this task is hope speech detection. The difference is that the former uses detection of negative speech to eliminate the impact of offensive speech on the creation of a healthy network environment. The hope speech detection task focuses on detecting hopeful words to create a good network environment, rather than detecting and deleting negative comments to encourage people and deprive individuals of their freedom of speech (Chakravarthi, 2020). The data source of this task comes from YouTube comments. What we have to do is to divide it into three categories: 'Hope speech', 'Not hope speech' and 'Not in intended language'. And we used K-fold cross-validation and ALBERT model to complete the detection of hope speech.

2 Related Work

At present, there are many studies on offensive and discriminatory speech detection, but there are few studies on hope speech detection. Therefore, we have looked for many related social media speech detection studies that can help us complete this task as a reference.

Marzieh Mozafari et al. used the regularization method to adjust the input text in the detection of hate and discriminatory speech and used a fine-tuned BERT model (Mozafari et al., 2020). Polychronis Charitidis et al. used the most advanced deep learning architecture to detect hate speech, trained and evaluated it using annotated data sets, and proposed an overall learning architecture that combines the predictive capabilities of each classifier (Charitidis et al., 2019).

Based on BERT's pre-trained language model, Marzieh Mozafari et al. made new fine-tuning of transfer learning and discussed the effect of using BERT to detect hate speech in a social media environment (Mozafari et al., 2019). Shriphani Palakodety et al. performed the qualitative and quantitative analysis of this embedding space for a variety of languages, they used Active Learning to train set construction in the analysis of peaceful speech (Palakodety et al., 2019). Zewdie Mossie et al. used Word2Vec word embedding technology for feature extraction when conducting Vulnerable community identification, and compared the effects

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 128–131

Figure 1: A word cloud image generated from the text marked with "Non hope speech" in the English training data set provided by the task organizer.

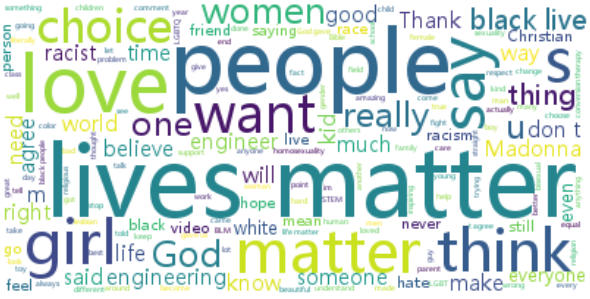

Figure 2: A word cloud image generated from the text marked with "Hope speech" in the English training data set provided by the task organizer.

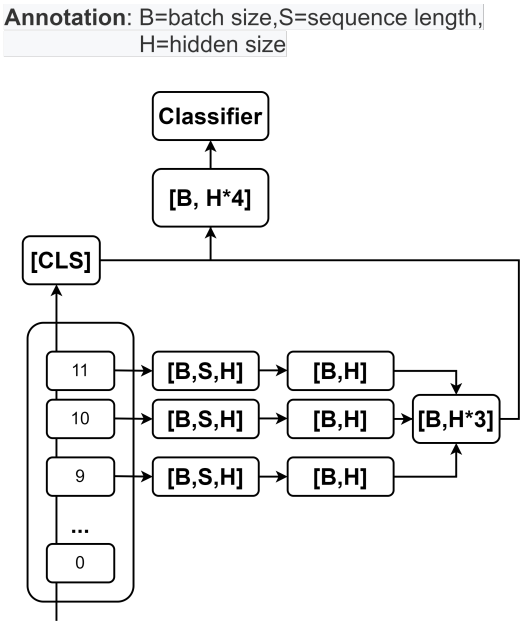

Figure 3: The fine-tuning strategies we use in English tasks. The output of layers 9-11 must be processed before being spliced with [CLS]. The processing method is to discard the first dimension data in the original data ([B, S, H] to [B, H]) to obtain a tensor output with the same shape as [CLS] (shape is [B, H]).

of classic classifiers with deep learning classifiers (Mossie and Wang, 2020). Shanita Biere et al. used natural language technology to detect hate speech on social media, using a convolutional neural network (CNN) (Biere et al., 2018).

Viktor Golem et al. tried different machine learning approaches, including traditional (shallow) machine learning models, deep learning models, and a combination of both for aggressive text detection(Golem et al., 2018). Jing MA et al. proposed a novel method that learns continuous representations of microblog events for identifying rumors. That model is based on recurrent neural networks (RNN) and that can capture the variation of contextual information of relevant posts over time (Ma et al., 2016). When discussing the detection of hate speech, Reynaldo Gil Pons et al. proposed an Attention-based Long Short-Term Memory Network, which is a model containing a bidirectional LSTM neural network (BILSTM)(De la Pena Sarracén et al., 2018).

3 Data

The data sets we can use are the training set and validation set in three languages(Tamil, Malayalam, and English) provided by the task organizer team(Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). We only participate in the English task, so here we only analyze the English data set.

We use the official data set for visual analysis. We can observe that some words such as "live", "people", "black", "matter" "racist" in the English data set appear very frequently. The appearance of these words is very suitable for the theme of the mission we participated in. However, these words mostly express some negative information. And these data mainly come from social media platforms, on the other hand, they also reflect the current status quo on social media platforms. We visualize the text data of two different labels of "Non hope speech" and "Hope speech". Figures 1 and 2 show the results of text visualization.

There are three types of category labels that appear in the English data set. They are "Hope speech", "Non hope speech", "Not English". Through further analysis of the data set, most of the mark types in the English data are "Non hope speech" (around 90%). There is an imbalance of category labels in the data sets of the English data set.

4 Methods

4.1 K-fold cross-validation

K-fold cross-validation reduces variance by averaging the results of k different group training.

The specific steps are as follows: The first step is to randomly divide the original data into k parts without repeated sampling; The second step is to extract 1 copy as the test set, and the remaining k-1 copies are applied to the training set for model training; In the third step, repeat the second step k times, so that a model is obtained after training on each training set and used to test on the corresponding test set, calculate and save the evaluation index of the mode; The fourth step is to calculate the average of k groups of test results as an estimate of model accuracy, and as the performance index of the model under current k-fold cross-validation.

4.2 ALBERT Fine-tuning

The reason we choose the ALBERT fine-tuning program is that ALBERT's parameters are shared. Compared with BERT, ALBERT has fewer parameters, and the corresponding model training requires less time and memory. The advantage is that when the model is large, parameter sharing will be a strong regularization method for the model, and it is not easy to overfit.

The related work and conclusions of Sun and others on the fine-tuning of BERT have inspired us (Sun et al., 2019). We try to apply these conclusions to the ALBERT model. We chose to combine a portion of the output of the ALBERT encoder layer with [CLS] (the output shape is [batch_size, hidden_size]). We take the output of the encoder of the ALBERT model (layers 9-11) (the output shape is [batch_size, sequence_length, hidden_size]). Then take out the data of the 0th and 2nd dimensions of the output tensor of each layer of 9-11 layers to obtain three Tensors of the same shape (shape is [batch_size, hidden_size]). Next, the three tensors obtained in the previous step and the [CLS] output result of the ALBERT model are spliced together (the shape is [batch_size, hidden_size*4]). Finally, input the results of the previous step into the classifier.

5 Experiment and Results

5.1 Data Preprocessing

We combine the English data training set and the English data validation set provided by the task organizer team in one file. The purpose is to integrate the validation set provided by the task organizer into the k-fold cross-validation data set. Then perform 5-fold cross-validation data processing on this merged file to split the data into five data sets. Each data set contains a new training set and a new validation set.

Data	Precision	Recall	F1 Score
Test set	0.93	0.94	0.93
Validation set	0.94	0.95	0.94

Table 1: The resulting score of our model method on the test set and validation set. The test set result score comes from the result score published by the task organization team on the leaderboard. The resulting score on the verification set is the result score obtained by us using the official evaluation index and the verification set.

5.2 Experimental setting

For the English tasks, we set model training parameters as follows: epoch, batch size, maximum sequence length, and learning rate of the task are 5, 32, 60, and 3e-5, respectively. We combined the training set and validation set provided by the task manager into a new data set. Then performed k-fold cross-validation on it to get the result of the k-fold cross-validation vote. The voting method in the reasoning process is to take the average value of the probability value of k outputs as the final predicted output logical value. We choose the training language models of ALBERT-V2 version[1].

5.3 Analysis of Results

Hope Speech Detection for Equality, Diversity, and Inclusion at LT-EDI 2021-EACL 2021 task evaluation indicators adopt F1 score weighted average. In the results announced by the task organizer team, our English scores were tied for first place on the list. This result is due to the pre-trained language model in the Transformer structure. The main reason is that the pre-trained language model has been pre-trained on a large number of English data sets. The ALBERT model is based on the Transformer model (Lan et al., 2019). The Transformer model has great advantages in capturing global information (Vaswani et al., 2017).

[1]https://huggingface.co/albert-base-v2

Label	Precision	Recall	F1 Score	Count
Hope speech	0.74	0.50	0.60	250
Non hope speech	0.95	0.98	0.97	2593
Not English	0.00	0.00	0.00	3

Table 2: On the labeled test set data published by the task organizer team, we compare the predicted results with the real results.

6 Conclusion

For the English task in this competition, we use the last three layers of the ALBERT model output to fuse with the [CLS] output. Use K-fold cross-validation on the integrated method to obtain better results. Our test results for Youtube's hopeful speech are not bad. In the final result, Precision was 0.93, Recall was 0.94, F1 Score was 0.93, and the ranking was tied for first place. But in future work, we will try to improve our method and choose to try other models, hoping to achieve better results on such tasks.

References

Shanita Biere, Sandjai Bhulai, and Master Business Analytics. 2018. Hate speech detection using natural language processing techniques. *Master Business AnalyticsDepartment of Mathematics Faculty of Science*.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2019. Towards countering hate speech and personal attack in social media. *arXiv preprint arXiv:1912.04106*.

Viktor Golem, Mladen Karan, and Jan Šnajder. 2018. Combining shallow and deep learning for aggressive text detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 188–198.

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*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2019. Kashmir: A computational analysis of the voice of peace. *CoRR, abs/1909.12940*.

Gretel Liz De la Pena Sarracén, Reynaldo Gil Pons, Carlos Enrique Muniz Cuza, and Paolo Rosso. 2018. Hate speech detection using attention-based lstm. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:235.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

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*.