niceNLP at SemEval-2023 Task 10: Dual Model Alternate Pseudo-labeling Improves Your Predictions

Yu Chang[1] , Yuxi Chen[1] , Yanru Zhang[1,2]*
[1]University of Electronic Science and Technology of China, Chengdu
[2]Shenzhen Institute for Advanced Study, UESTC
{yuchang, yuxi.ch}@std.uestc.edu.cn,
yanruzhang@uestc.edu.cn

Abstract

Sexism is a growing online problem. It harms women who are targeted and makes online spaces inaccessible and unwelcoming. In this paper, we present our approach for Task A of SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS), which aims to perform binary sexism detection on textual content. To solve this task, we fine-tune the pre-trained model based on several popular natural language processing methods to improve the generalization ability in the face of different data. According to the experimental results, the effective combination of multiple methods enables our approach to achieve excellent performance gains.

1 Introduction

Gender discrimination is any mistreatment or negative feelings directed at women based on their gender, or based on their gender in combination with one or more other identity attributes (e.g., Black women, Muslim women, trans women). As the Internet continues to grow, sexism has become a growing online problem. It can be harmful to women who are discriminated against while also posing a threat to the healthy functioning of cyberspace. Large-scale pre-trained models have been widely used to automate the identification of discriminatory statements. However, these models are often capable of simple labeling and not interpretative descriptions. Improving the interpretability of sexism detection can help to thereby enhance the understanding and behavioral decision-making ability of users and moderators of automated tools. To address this issue, SemEval 2023 - Task 10 supports the development of English-language models for sexism detection that are more accurate as well as explainable, with fine-grained classifications for sexist content from Gab and Reddit. This task proposes a new English sexism detection dataset based on Gab and Reddit, which contains four fine-grained and eleven more fine-grained category labels. In Task A, each sentence was first asked to perform a dichotomous classification of the presence or absence of sexism. And tasks B and C require a finer-grained distinction between sentences with sex discrimination (Kirk et al., 2023).

In many past studies, language models called BERT have been used extensively for several tasks (Devlin et al., 2018). BERT models can be pre-trained with a self-supervised approach to generate word/tag or sentence representations that are rich in a priori knowledge (Jin et al., 2020). They can then be specifically fine-tuned for many downstream tasks, including text classification. Moreover, various model optimization methods have been shown to be effective in enhancing the fine-tuning of pre-trained models. Therefore, in this work, we employ the current state-of-the-art pre-training model DeBERTa-v3 to exploit the prior knowledge in its pre-trained resources (He et al., 2021). And to adapt it to the linguistic context in the task, we pre-trained the model again using a label-free corpus provided by the organizer. In addition, we improve the model classification based on various model tuning methods including adversarial training (Shafahi et al., 2019) and Stochastic Weight Averaging (SWA) (Yang et al., 2019). Finally, based on the characteristics of the binary classification task, we constructed a dual model alternate pseudo-labeling (Cascante-Bonilla et al., 2021) approach using multi-sample dropout (Inoue, 2019) to complete the final classification.

Our approach finally achieved the eleventh place in the test phase leaderboard of task A. The experimental results demonstrate the effectiveness of the multiple methods we used and proposed.

2 Related Work

In recent years, there has been an increasing interest in sex discrimination detection tasks. Such as by

*Corresponding author

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 307–311

using GloVe Embeddings (Pennington et al., 2014) and improved LSTMs to see sexism in the workplace, adding attention mechanisms to it (de Paula et al., 2021).

Masked Language Modeling (MLM) is one of the pre-training methods used by BERT. Although the DeBERTa model does not use this method, it can still improve the performance of the model from our experimental results.

Inspired by previous pre-training models such as BERT, RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019), DeBERTa-v3 employs a series of innovative techniques to improve the performance and efficiency of the model. Among the most notable features are: the use of a dynamic masking mechanism in the pre-training phase, which allows the model to better capture relationships in long texts; the use of a new cross-layer interaction mechanism that allows the integration of different semantic information at different levels; and the use of a global adaptive regularization method that reduces the overfitting problem.

Adversarial training is a training method that introduces noise and adds perturbations to the samples while trying not to change the distribution of the original samples, allowing the model to ignore such perturbations and thus improving the robustness of the model. The Fast Gradient Method (FGM) that we use is mainly useful in NLP tasks (Miyato et al., 2016). Unlike the direct perturbation method in the image domain on the input samples, FGM perturbs on the word vectors because a sentence that changes a word (e.g., synonym substitution) may cause the meaning of the whole sentence to change, and the strength of the perturbation is not easy to control.

SWA is a technique for optimizing neural networks by using models with random weights during training to smooth out the loss curve of the model. This technique improves the generalization ability of the model, prevents overfitting, and generally leads to better test set performance.

Multi-sample dropout is a neural network regularization method based on the dropout technique. Unlike the traditional dropout technique which is only used during network training, multi-sample dropout can be used both during training and testing to improve the generalization ability of the model.

3 Approach Overview

For task A, each sentence in the dataset was labeled as 0 if sex discrimination was present and 1 otherwise. We chose the large version of the DeBERTa-v3 model as the backbone. The DeBERTa-v3-large model comes with 24 layers and a hidden size of 1024. It has 304M backbone parameters with a vocabulary containing 128K tokens which introduce 131M parameters in the Embedding layer. This model was trained using the 160GB data as DeBERTa V2. Figure 1 shows the architecture of the whole approach.

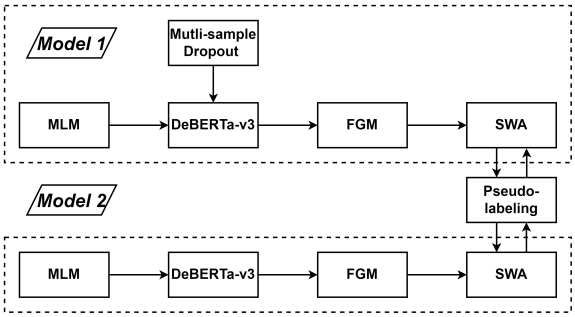

Figure 1: The architecture of the whole approach.

We first pre-trained the DeBERTa-v3 model using the MLM method on one million Gab texts and one million Reddit texts provided by the organizer and then used FGM and SWA to further improve the model's generalization and resistance to overfitting. In Model 1, multi-sample dropout is used to improve the neural network and form an alternative dual-model structure with Model 2. Finally, we use the classification results of each model in the test set as pseudo-labels and add them to the training set of another model. Several iterations of alternative training are performed until the classification effect no longer improves.

4 Experiments

We conducted experiments on the training and test sets provided by the task organizer, where data from the development phase were also added to the training set. The training set contains 16,000 samples, of which 12,116 are "not sexist" samples and 3,884 are "sexist" samples. We divide 25% of them into validation sets and ensure that the sample proportion is constant. The test set contained 4,000 samples, including 3,030 "not sexist" samples and 970 "sexist" samples. Table 1 shows the example sentences with two different labels.

Sentence	Label
"This is like the Metallica video where the poor mutilated bastard was saying ""Please kill me"" over and over again, only with emojis instead of Morse code."	no sexist
"I agree with that but at the same time I know myself well enough to say I can't love a woman. The minute she begins to hit the wall and some hotter, younger women enters the picture, it's time for impulse control because I'm going to want that."	sexist

Table 1: Examples of sentences with different labels.

4.1 Pre-trained Models

First, we compare the basic classification effects of several different pre-trained models that are currently popular. They are DeBERTa-v3-large, XLM-RoBERTa-large (Conneau et al., 2019) and covid-twitter-BERT-v2 (Müller et al., 2020). They have a maximum input length of 96, batch size of 16 and learning rate of 1e-06.

For their training results, we always take the one epoch with the smallest validation set loss as the final model for prediction. Their performance on the test set at this time is shown in Table 2.

Models	F1 Score
DeBERTa-v3-large	0.8433
XLM-RoBERTa-large	0.8251
covid-twitter-BERT-v2	0.8316

Table 2: The comparison of base performance of three pre-trained models.

We chose DeBERTa-v3-large as the backbone model for the subsequent experiments because of its obvious advantages over other models. We then use the MLM method to pre-train DeBERTa-v3-large on the unlabeled dataset. Instead of randomly initializing the weights, we inherit the model's original weights. This is because we believe that the prior knowledge contained in the actual weights of the model is important for this task. Figure 2 shows the variation in the loss of the model and the performance when applied to task A for different pre-training steps.

4.2 Trick Methods

In the subsequent experiments, we choose the model obtained by pre-training 225000 steps as the benchmark. It has an F1 score of 0.8496 on the test set. For the choice of adversarial training methods, we compared FGM, Projected Gradient Descent (PGD) (Madry et al., 2017) and Adversarial Weight Perturbation (AWP) (Dong et al., 2020). Their performance is compared in Table 3.

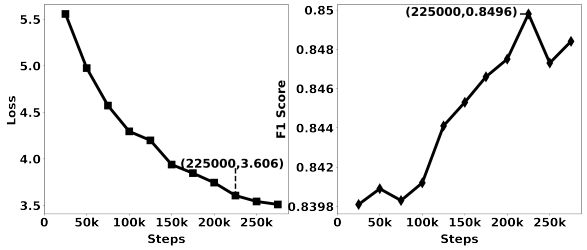

Figure 2: The variation in the loss and the performance for different pre-training steps.

Methods	FGM	PGD	AWP
F1 Score	0.8521	0.8504	0.8477

Table 3: The comparison of base performance of three adversarial training methods.

Then we performed comparison experiments using SWA and Exponential Moving Average (EMA) with FGM. Both EMA and SWA are methods for averaging the model weights. They assume that weight jitter in the last few steps of model training will impact the model effect. The performance degradation caused by jitter can be effectively reduced by weight averaging. The results of their comparison experiments are shown in Table 4.

Methods	F1 Score
Origin+FGM	0.8521
Origin+EMA	0.8487
Origin+SWA	0.8504
Origin+FGM+EMA	0.8524
Origin+FGM+SWA	0.8548

Table 4: The comparison of EMA and SWA with FGM.

4.3 Dual Model Alternate Pseudo-labeling

Further to solve the overfitting problem, we use multi-sample dropout. While traditional dropout selects a random set of samples from the input (called dropout samples) in each training round, multi-sample dropout creates multiple dropout samples and then averages the loss of all samples to

obtain the final loss. This approach simply replicates parts of the training network after the dropout layers and shares the weights between these duplicated fully connected layers without new operators. The network parameters are updated by combining the losses of M dropout samples so that the final loss is lower than the loss of any of the dropout samples. This has the effect of repeating the training M times for each input in a minibatch. Thus, it significantly reduces the number of training iterations. We experimentally determine the optimal M. Unfortunately, even with the best performance of M=5, the F1 score of the model is only 0.8544. Since this score is similar to the best score without this method, we tried to incorporate the pseudo-label corresponding to this score into the model training. The results were surprising. After we added the classification results of the two models with and without the multi-sample dropout to each other's training sets, the model performance improved substantially. The performance of the single-model self-looping pseudo labeling and two-model alternative pseudo labeling are compared in Table 5. Model 1 and Model 2 are the same as Figure 1.

Methods	Iter0	Iter1	Iter2	Iter3	Iter4
Model1 self	**0.8548**	0.8451	0.8476	0.8544	0.8518
Model2 self	0.8548	0.8564	0.8568	**0.8579**	0.8562
Model1 alternative	0.8548	0.8563	0.8584	0.8601	**0.8614**
Model2 alternative	0.8548	0.8568	0.8591	0.8609	**0.8624**

Table 5: The performance of our different models.

The F1 scores do not increase after the fifth iteration and beyond, so these data are not represented in the table above. To summarize the experimental results, it can be seen that pseudo labeling alone on Model 1 with multi-sample dropout degrades the performance. Pseudolabeling alone on Model 2 without multi-sample dropout results in a better improvement. In the model with interactive pseudo labeling, Model 2 performs better than Model 1. This phenomenon indicates first that the pseudo-labeling approach is effective for this task. This is because the task is simpler for the pre-trained model, and the base model classifies well. This leads to a higher quality of the pseudo-labeling and thus facilitates further model training. However, with the enhanced dropout capability of multi-sample dropout, the model can obtain better text information on the one hand and learn additional errors on the other. This two-sidedness is manifested as a negative impact of the mistake in the single model and as an enhanced generalization ability in the dual model.

5 Conclusion

This paper compares and summarizes the different methods used in solving task A. MLM, FGM, SWA, and multi-sample dropout form the backbone of the whole approach. The pseudo-labeling technique based on dual-model interaction substantially improves the final performance of the model. The role played by multi-sample dropouts in this technique deserves further discussion.

References

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Angel Felipe Magnossão de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv preprint arXiv:2111.04551*.

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*.

Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. 2020. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33:8270–8283.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification

and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Chris De Sa. 2019. Swalp: Stochastic weight averaging in low precision training. In *International Conference on Machine Learning*, pages 7015–7024. PMLR.