HHS at SemEval-2023 Task 10: A Comparative Analysis of Sexism Detection Based on the RoBERTa Model

Yao Zhang
Yunnan University
18314275485@163.com

Liqing Wang*
Yunnan University
wlq@ynu.edu.cn

Abstract

This paper describes the methods and models applied by our team HHS in SubTask-A of SemEval-2023 Task 10 about sexism detection. In this task, we trained with the officially released data and analyzed the performance of five models, TextCNN, BERT, RoBERTa, XLNet, and Sup-SimCSE-RoBERTa. The experiments show that most of the models can achieve good results. Then, we tried data augmentation, model ensemble, dropout, and other operations on several of these models, and compared the results for analysis. In the end, the most effective approach that yielded the best results on the test set involved the following steps: enhancing the sexist data using dropout, feeding it as input to the Sup-SimCSE-RoBERTa model, and providing the raw data as input to the XLNet model. Then, combining the outputs of the two methods led to even better results. This method yielded a Macro-F1 score of 0.823 in the final evaluation phase of the SubTask-A of the competition.

1 Introduction

Sexism has long existed in society, causing various asymmetries and inequities, and harming some people. Given the enormity of data and the fast-paced flow of information on the internet, addressing sexism still poses a significant challenge (Kirk et al., 2023). Web technologies are rapidly evolving and automated detection is widely used for sexism, except that most tools only give generic categories without further explanation of what sexism is and do not improve the interpretability, trust and understanding of the tools.

In the past, computers were frequently used to conduct application tests for sexism by many people. Multiple social media platforms have been used to analyze cyberbullying detection using various topics based on deep learning models and transfer learning (Agrawal and Awekar, 2018). A method has been proposed that can detect and analyze three characteristics of Twitter tweets: patriarchal behavior, misogyny and sexism (Frenda et al., 2019). (Farrell et al., 2019) demonstrated an increasing pattern of misogynistic content and users as well as violent attitudes. Fine-grained, multi-label sexism account classification has been explored, and the first semi-supervised multi-label sexism account classification has been proposed (Abburi et al., 2021). (Guest et al., 2021) proposed a new hierarchical taxonomy for online misogyny and produced a related labeled dataset. Sexism takes many forms on the web, some are fixed words of sexism, some are new words, and many forms of sexism require contextual, temporal and geographical understanding. Thus much of this sexism detection is subjective and relative. This also means that sexism data is hard to find, requiring a lot of manual annotation effort. Therefore, we participated in the SemEval-2023 shared Task 10 (Kirk et al., 2023), which aims to identify whether utterances contain sexism against women.

In this task, we carried out several works as follows.

a) Firstly, we tried 5 different models on the dataset to compare the experimental results of several models

b) Secondly, we also tried to enhance the data using methods such as dropout and pseudo-labeling to improve the Macro-F1 scores

c) Finally, we introduced the Focal Loss function to analyze its effect

In SubTask-A, we got the best Macro-F1 score of 0.84 for the development phase and a Macro-F1 score of 0.823 for the final testing phase, obtaining a final rank of 47. The paper is organized as follows: the model, experimental methods and data are presented in Section 2; the experimental details are presented in Section 3; the analysis of the results is performed in Section 4; and the summary is presented in Section 5.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 963–968
July 13-14, 2023 ©2023 Association for Computational Linguistics

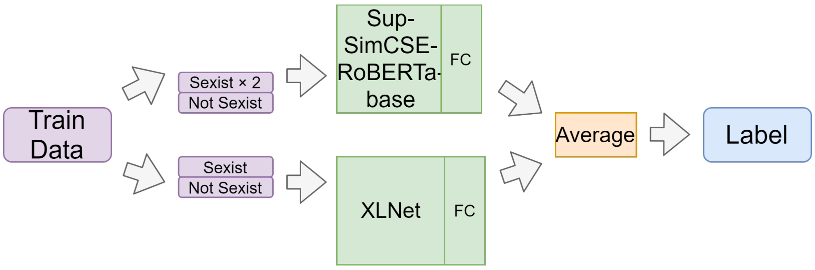

Figure 1: Model overview (FC is fully connected layer)

2 System Overview

We use Sup-SimCSE-RoBERTa-base (Gao et al., 2021) as the first model and use dropout augmented sexist data as this model input to get the first classification output through the fully connected layer. Meanwhile, the classification output of the second model is obtained by using the data without enhancement as the XLNet model input. The final result is obtained by calculating the average of the two outputs. The models used here are all sourced from HuggingFace[1]. The composition is shown in Figure 1.

2.1 Models Comparison

SubTask-A can be viewed as a simple text classification task. We initially chose five models to take for the experiment: TextCNN, BERT, RoBERTa, Sup-SimCSE-RoBERTa, and XLNet.

In 2014, Yoon Kim et al. made some changes to the CNN and proposed the text classification model TextCNN (Zhang and Wallace, 2015). The main idea of TextCNN is to pass several different convolutional kernels to obtain the features of the n-gram of the text. These features are pooled and spliced to get the features of the text to perform operations such as classification. BERT is a pre-trained model proposed in 2018 (Devlin et al., 2018). BERT is an important milestone in the history of NLP development, and it can be used as a basic model to compare with other pre-trained models, so we choose it as one of the comparison models. RoBERTa is an optimized version of BERT, which has a larger number of parameters, more training data, and some improvements in training methods. Its performance is even better than BERT in some NLP tasks (Liu et al., 2019). Sup-SimCSE-RoBERTa uses the idea of contrastive learning to better distinguish between data with different labels, and may be able to achieve good results in this task (Gao et al., 2021). XLNet is a model that is also worthy of consideration for comparison. Its application of autoregressive language modeling, instead of autoencoding language modeling, overcomes the limitations of masking and provides a unique perspective on the model's influence on the task (Yang et al., 2019).

2.2 Methods

After observing the experimental results of the model, we found that the Macro-F1 scores of the prediction results of the sexism sample in the experiment is low. Analysis of the officially given data clearly shows that the samples are unbalanced, and only 24% of the data are sexist samples. For this situation, we tried several methods to solve this problem.

2.2.1 Pseudo Label

The problem of sample imbalance can also be solved by the pseudo label. The pseudo label technique involves utilizing a model trained on labeled data to make predictions on unlabeled data, filtering the samples based on the prediction outcomes, and re-entering them into the model for further training (Lee et al., 2013).

2.2.2 Data Augmentation

Text data can be enhanced in a variety of ways, with methods such as EDA, text generation (Peng et al., 2020), back translation, etc. In 2019 Protago Labs published a paper on EDA data enhancement (Wei and Zou, 2019) that improved their experimental scores.

2.2.3 Dropout

Dropout has been used in many applications, including preventing the overfitting of neural networks and improving the generalization ability of models (Srivastava et al., 2014). It has also been used for data augmentation, and Gao et al. obtained similar data pairs by this method (Gao et al., 2021).

[1] https://huggingface.co/princeton-nlp/sup-simcse-roberta-base

2.2.4 R-drop

Through previous experiments, we have confirmed that dropout does work well in some aspects. Therefore, we tried another dropout method with a relatively mature theory, R-drop (Wu et al., 2021). R-drop mitigates the inconsistency between training and testing by calculating the KL scatter between the samples generated by dropout.

2.2.5 Model Ensemble

The model ensemble has played a good role in many competitions, which can effectively stabilize the model and improve the generalization ability (Nai et al., 2022).

2.3 Dataset

In this task, all data were obtained from Kirk et al (Kirk et al., 2023). The dataset of markers published by the task organizer consists of 20,000 entries, 10,000 from Gab and 10,000 from Reddit. The training data consisted of 14,000 entries (70%), of which 3,398 were sexist. The composition is shown in Figure 2. where the labeled data were divided into three levels of labels, of which SubTask-A used only the first level of labels [2]. In our experiments, we divide the original 14,000 training data into a training set, validation set and test set according to the ratio of 8:1:1. The training set is used to train the model, the validation set is used to evaluate the effect of the model during the training process, and the test set is used to calculate the final model results. The task organizers also provided two million unlabeled entries, half from Gab and half from Reddit.

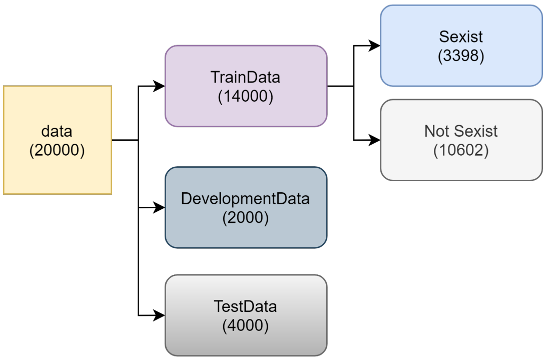

Figure 2: Dataset composition

3 Experiment

In this section, we will present some experimental details and parameters.

3.1 Model

For the TextCNN model, an index is first created for the words that appear in the training data to build a corpus. The learning rate is set to 1e-5, and 5 epochs are trained. And for the BERT, RoBERTa, and Sup-SimCSE-RoBERTa models, the basic parameters of these three models can be set the same, pad size can be set to 256, learning rate to 1e-5, training 5 epochs, and batch size to 32. Use the BertAdam (Zhang et al., 2020) optimizer to optimize the model. For the XLNet model, train 3 epochs, the batch size is 32, and the learning rate is 1e-5. The experimental results are shown in Table 1.

Table 1: Table of Macro-F1 scores for each model on the local test set

Model	Macro-F1 score
TextCNN	0.4915
BERT	0.8153
XLNet	0.8224
RoBERTa	0.8272
Sup-SimCSE-RoBERTa	0.8390

3.2 Method

This section presents some experimental details of the method, and the experimental results are shown in Table 2.

3.2.1 Dropout

In this experiment, we can use the model's dropout parameter to achieve a similar effect. That is, each sample is loaded twice, and the dropout is used to make the input samples have similar but not identical representations. We have tried dropout for all labels, and we have also tried multiple dropouts for the sexist label data alone and finally evaluated that twice dropout for the sexist label data alone is better.

3.2.2 R-dorp

In this experiment, we tried to use an R-drop for the sexist sample alone, but this made the difference in loss between different labeled samples and the final result was not very good, so the R-drop operation was applied to the whole training set.

[2] https://codalab.lisn.upsaclay.fr/competitions/7124

Table 2: Table of Macro-F1 scores for each method on the local test set

Methods/Models	BERT	RoBERTa	SimCSE-RoBERTa
Normal	0.8153	0.8272	0.8370
R-drop	0.7745	0.7777	0.7933
Data Augmentation	0.8038	0.7988	0.8116
Pseudo label	0.8084	0.8281	0.8343
Dropout	0.8278	0.8396	0.8425
Ensemble(SRX)	-	-	0.8441

Table 3: Comparison of Macro-F1 scores for cross-entropy loss and Focal loss

Model	Cross entropy	Focal loss
BERT	0.8153	0.8278
RoBERTa	0.8272	0.8376
Sup-SimCSE-RoBERTa	0.8390	0.8304
Sup-SimCSE-RoBERTa+Pseudo label	0.8343	0.8303
Sup-SimCSE-RoBERTa+Dropout	0.8425	0.8405
SRX	0.8441	0.8401

3.2.3 Pseudo Label

We train a Sup-SimCSE-RoBERTa model that performs well on the task dataset and then use this model to predict unlabeled data. The sexist-labeled data with accuracy greater than 0.97 are filtered from the prediction results, and then a portion of them are selected and added to the training set to alleviate the data imbalance problem.

3.2.4 Data Augmentation

Each word in the experiment has a 40% probability of synonym substitution and a 30% probability of random insertion. We used this method to augment the sexist data 5 times. These augmented data were added to the dataset for training tests.

3.2.5 Model Ensemble

In our experiments, we tried to aggregate the results of several selected models, among which the Sup-SimCSE-RoBERTa and XLNet models have better aggregate results. Because dropout was found to improve certain scores during the training process, this training technique was also used. We abbreviate it as SRX.

3.3 Loss Function

We used the cross-entropy function (De Boer et al., 2005) in sklearn in our experiments, and after observing the problem of sample imbalance, we tried to introduce the Focal Loss (Lin et al., 2017) for comparison. Focal loss helps solve the sample imbalance problem, and we test its impact on the task. The experimental results are shown in Table 3.

4 Results and Analysis

In this section, we report the results of the models and methods presented in Sections 2 and 3 and make a basic analysis.

Table 1 shows the results of the Macro-F1 score performance of each model on our test set. Among them, Sup-SimCSE-RoBERTa has the best performance, while the score of TextCNN is very low.

From Table 2, of the several methods, only dropout has a significant improvement on the Macro-F1 score. In Table 3, the cross-entropy loss and Focal Loss have their advantages and disadvantages, the combination of model and method also improves the Macro-F1 score to some extent. Meanwhile, the pseudo label has very limited improvement on the experimental results, while data augmentation and R-drop are negatively affecting the model. Our initial inference is that the experimental methods were not adjusted to the task and that there were some objective errors in the process. For example, in the data augmentation section, we tried different probabilities of synonym replacement and random insertion, but never found a probability that would improve the results. Moreover, we believe that the effect of synonym replacement and random insertion is limited in our experiments, and perhaps it would be more effective to focus on some specific words, such as some iconic sexism words and symbols.

We finally settled on using the Sup-SimCSE-RoBERTa model and used dropout to input the sexist data twice into the model and finally ensemble the output of the XLNet model(SRX). We used this model in the final test phase of SubTask-A to get a Macro-F1 score of 0.823. Its performance was also the best in our experiments.

5 Conclusion

We discuss and compare several sexism detection models with different approaches to try to improve model capabilities and report relevant experimental results. The models still have considerable potential for improvement. Therefore, in the future plan, we will continue to try to improve the effectiveness of methods such as pseudo labels. And try some large models such as large RoBERTa to achieve better results.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2021. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Science and Engineering*, 6(4):359–379.

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 141–153. Springer.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67.

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*.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 87–96, New York, NY, USA. Association for Computing Machinery.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

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.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

Zihan Nai, Jin Wang, and Xuejie Zhang. 2022. YNU-HPCC at SemEval-2022 task 8: Transformer-based ensemble model for multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1215–1220, Seattle, United States. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.