# PCJ at SemEval-2023 Task 10: A Ensemble Model Based on Pre-trained Model for Sexism Detection and Classification in English

**Chujun Pu**
Yunnan University
1448592081@qq.com

**Xiaobing Zhou***
Yunnan University
zhouxb@ynu.edu.cn

## Abstract

This paper describes the system and the resulting model submitted by our team "PCJ" to the SemEval-2023 Task 10 sub-task A contest. In this task, we need to test the English text content in the posts to determine whether there is sexism, which involves emotional text classification. Our submission system utilizes methods based on RoBERTa, SimCSE-RoBERTa pre-training models, and model ensemble to classify and train on datasets provided by the organizers. In the final assessment, our submission achieved a macro average F1 score of 0.8449, ranking 28th out of 84 teams in Task A.

## 1 Introduction

Sexism is a very serious problem on the Internet nowadays, where unsuspecting people attack women by making statements against them, exposing them to unfair treatment, and eventually causing harmful effects on society. Therefore, tools that can automatically detect sexism are widely deployed (Bordia and Bowman, 2019), which is an important task in the field of NLP (Mai et al.).

Sexism detection is a text classification problem. Text classification is a classical problem in the field of natural language processing, and the main methods can be divided into two categories: machine learning-based methods and deep learning-based methods. For the machine learning-based methods (Agarwal and Mittal), the main idea is artificial feature engineering plus shallow classification model (Li et al., 2020). Text feature engineering is divided into three parts: text preprocessing, feature extraction, and text representation, with the ultimate goal of converting text into a computer-understandable format and encapsulating enough information for classification, and then the classifier uses statistical classification methods, such as SVM, plain Bayesian classification algorithms. This approach overall relies heavily on discrete manual features. For deep learning-based text classification (Minaee et al., 2021), the main idea is that the text to be classified is represented as a word vector, and then deep learning networks and other variants are used for automatic feature extraction, represented by models such as Transformer (Vaswani et al., 2017).

However, there are still some deficiencies in current sexism detection, which need to be improved, such as the abuse of language detection models, and the misestimation of sexism (Park et al., 2018). Task 10 of SemEval-2023 is the explainable detection of online sexism (Kirk et al., 2023), which aims to address the issue of automated detection of large-scale online sexism. By proposing a hierarchical taxonomy with three tasks for detecting sexist content, it ultimately aims to improve the accuracy and explainability of automated detection tools, and contribute to reducing the harmful effects of online sexism.

In this paper, we describe the details of our research in Subtask A and how it performed on the evaluation data. The specific task is described as performing sexism detection, so that the system must predict whether the post is sexist or not, making a binary classification. The paper is organized as follows. Section 2 briefly describes the relevant background regarding text classification and text preprocessing. Section 3 describes the system architecture used in the experiments to explain the model approach. In Section 4, the experiments and results are presented, and Section 5 presents a summary of the whole paper and areas for future improvements.

## 2 Related Work

The emergence of pre-trained models (PTMs) has brought the field of natural language processing into a whole new era. Pre-training has been an effective strategy for learning deep neural network parameters, and fine-tuning the downstream tasks based on pre-training models is beneficial for NLP



*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 433–438

tasks. The development of PTMs goes through two stages from Non-Contextual Embeddings to Contextual Embeddings (Qiu et al.). For Non-Contextual Embeddings, the representative one is Word2Vec (Mikolov et al., 2013). This type of word embedding usually takes shallow networks for training, and when applied to downstream tasks, the rest of the whole model still needs to be learned from scratch. For Contextual Embedding, the main purpose is to solve the problem of multiple meanings of a word through a pre-trained encoder that can output contextually relevant word vectors, representative ones are OpenAI GPT (Radford et al., 2018), BERT (Devlin et al., 2018).In pre-training models, RoBERTa is an improved version of BERT, which uses a larger training dataset, longer training time, and some other optimization strategies. RoBERTa has achieved better results than BERT in multiple natural language processing tasks (Liu et al., 2019), including text classification tasks. Therefore, it is suitable to be used as the basic pre-training model for this task.

In conclusion, in this competition, we adopt a deep learning-based approach for text classification, using the existing dataset to fine-tune the pre-trained models RoBERTa and SimCSE-RoBERTa to obtain the embedding vectors, combining BiLSTM (Wang et al., 2021) and BiGRU (Yu et al., 2021) neural network structures for task classification. Finally, in order to improve the task classification, the model ensemble method is used to fuse the fine-tuned models together (Malla and Alphonse, 2021).

## 3 System overview

This section describes the models used. We use RoBERTa and SimCSE-RoBERTa pre-trained models as the coding layer for modification, and on top of Roberta-large and sup-Simcse-Roberta-large base models, we then down-join BiLSTM and BiGRU structures to let the models capture more information in combination to contextual semantics. Then, we construct 4 types of text classification models. Finally, the trained models are fused and the final classification results are selected using the voting method. The system architecture used is described below.

### 3.1 Model blocks

RoBERTa is a more finely tuned version of BERT, proposed by Facebook AI (Liu et al., 2019), which

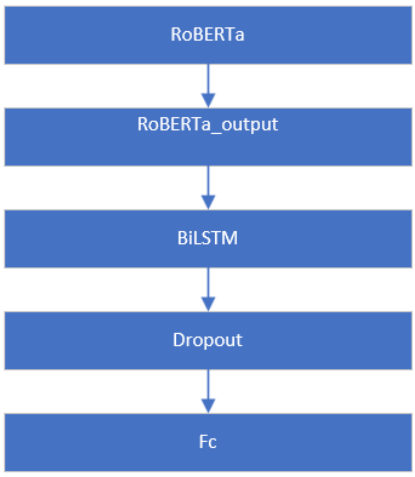


Figure 1: RoBERTa-BiLSTM

improves BERT in three aspects. It has two versions, RoBERTa-large and RoBERTa-base. Compared with the latter, the large version has 24-layer Transforms and is more capable of extracting features for this paper, so we choose RoBERTa-large as the base pre-training model. The structural modification based on the RoBERTa model is as follows.

Model 1: The last hidden layer of the RoBERTa model is plugged into the BiLSTM under the final layer, then classified after the Dropout layer, as in Figure 1.

Model 2: First, the output of the RoBERTa model is connected to BiLSTM and then to BiGRU, and then the average pooling and maximum pooling are performed on the output results, and finally, the results are spliced with the pooler-out of RoBERTa model, and then classified after the Dropout layer, as in Figure 2.

Model 3: Each layer of BERT's structure understands the text differently (Jawahar et al., 2019), and the same is true in RoBERTa. Therefore, based on the original RoBERTa model, only the features of the last four layers are used, and the <CLS> of the last four hidden layers are stitched together, and then after maximum pooling, the dimensionality is reduced and the classification is performed again after the Dropout layer.

SimCSE (Gao et al., 2021) is a simple contrast learning framework that can greatly improve sentence embedding. It can get better embedding of sentences in the task of doing sentence representation. Sup-Simcse-RoBERTa-large model is a RoBERTa model trained with supervised SimCSE.

Model 4: The structure modification method

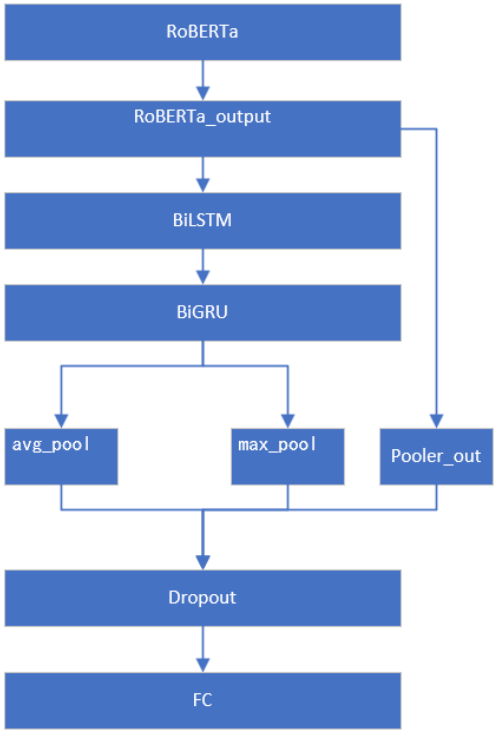


Figure 2: RoBERTa-BiLSTM-BiGRU

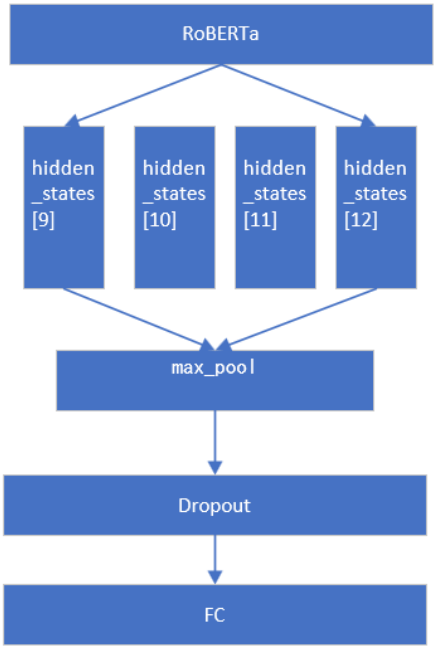


Figure 3: RoBERTa

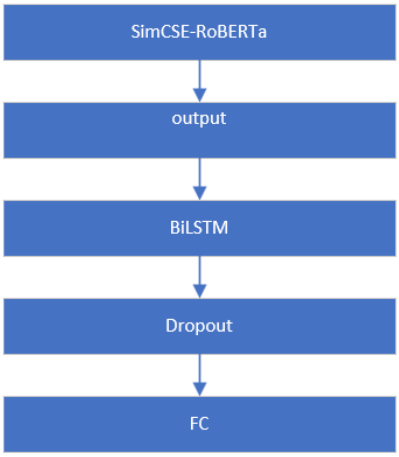


Figure 4: RoBERTa-BiLSTM

based on the SimCSE-RoBERTa model is as follows: the result of the last hidden layer obtained from the SimCSE-RoBERTa model is used as input to access the bidirectional LSTM, and then classified after the Dropout layer.

### 3.2 Model ensemble

In order to improve the classification effect of sexism detection, we use the voting method to obtain the final output from the output of the four models mentioned above. The voting method is an ensemble model approach that follows the principle of majority rule. It can combine each model's characteristics thus improving the model's robustness and generalization ability. This is achieved by first training each of the aforementioned models to obtain the best training results, and then having each model predict the same data. The final output is selected based on the majority vote.

## 4 Experiments and Results

### 4.1 Dataset and Processing

The dataset we use is the training data from the English dataset provided by the organizers (Kirk et al., 2023), which comes from Gab and Reddit and is divided into two categories sexist, and not sexist. Its training data consists of 14,000 entries, 3398 of which are classified as sexist. The category classification of the data is highly unbalanced, which greatly impacts the training of the model and the accuracy description.

In the experiment, the original dataset is divided into training set, validation set, and test set in an

| Name | Value |
|---|---|
| The learningrate of RoBERTa | 2e-5 |
| The learningrate of other models | 1e-5 |
| Warmup rate | 0.01 |
| Dropout rate | 0.2 |
| Weightdecay | 0.01 |
| Padsize | 64 |
| Batchsize | 32 |

Table 1: Final hyperparameter configuration

8:1:1 ratio. During this process, the stratify parameter is set for stratified sampling to ensure that the proportions of different class samples in each subset are the same as in the original dataset. Then the training and validation sets are preprocessed by using preprocessing first. In order to get a better understanding, or to build a better algorithm, a pair of data sets is needed to remove the noise that may affect the results, including emoji and '[xxx]'. On the other hand, the test set is not preprocessed, aiming to maintain similarity to the files tested in the final competition, allowing for a more realistic resulting score.

### 4.2 Implementation Details

In the specific training process, we used the BertAdam (Zhang et al., 2020) optimizer to optimize our model and adjusted the hyperparameters based on the results of the test set. The hyperparameter table using the final model is shown in Table 1.

### 4.3 Results

For the classification results, the Macro F1 score is used for evaluation, which calculates the precision and recall for each class, averages them and then calculates them according to the F1 score formula, as follows.

$$Precision_{macro} = \frac{\sum Precision_i}{n} \quad (1)$$

$$Recall_{macro} = \frac{\sum Recall_i}{n} \quad (2)$$

$$MacroF1 = \frac{2 * Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}} \quad (3)$$

Where,

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

| Model name | F1 score |
|---|---|
| RoBERTa | 0.8327 |
| RoBERTa-BiLTSM | 0.8311 |
| RoBERTa-BiLTSM-BiGRU | 0.8467 |
| SimCSE-RoBERTa-BiLSTM | 0.8359 |
| Model ensemble | **0.8480** |

Table 2: Results from different models

| Model name | F1 score |
|---|---|
| RoBERTa | 0.8164 |
| RoBERTa-BiLTSM | 0.8460 |
| RoBERTa-BiLTSM-BiGRU | 0.8375 |
| SimCSE-RoBERTa-BiLSTM | 0.8467 |
| Model ensemble | **0.8489** |

Table 3: Results of different models based on development phase data

The final results of the experiments are shown in Table 2. The table shows the Marco F1 scores for each model on the divided test set.The results show that the RoBERTa model can already model text well on the test set. Just adding a BiLSTM structure after RoBERTa cannot improve the classification performance and may even decrease it. Compared with the model of SimCSE-RoBERTa with a BiLSTM structure, SimCSE-RoBERTa can better utilize semantic information to model text through contrastive learning, so a BiLSTM structure can achieve better performance in this task than RoBERTa. Based on RoBERTa-BiLSTM, we added a BiGRU structure and found that it can significantly improve the performance. This is because the BiGRU structure can better capture long-distance dependencies in the text sequence and further extract semantic information, thereby enhancing the model's classification ability. For the ensemble model, it works better than any single model to improve the classification ability of sex discrimination detection, and it generalizes better. This is reflected in the gap between the results submitted in the development phase of the competition (Kirk et al., 2023) and our own results on the test set.

Table 3 shows the results of the scores based on the validation data provided during the development phase. It can be seen that the scores of the ensemble model do not differ much from the Macro F1 scores on their own divided training set, while the other models differ significantly. The model

| Model name | Recall&Precision(Not Sexist) | Recall&Precision(Sexist) |
|---|---|---|
| RoBERTa | 0.9292/0.9129 | 0.7235/0.7664 |
| RoBERTa-BiLTSM | 0.9302/0.9113 | 0.7176/0.7673 |
| RoBERTa-BiLTSM-BiGRU | 0.9425/0.9082 | 0.7029/0.7967 |
| SimCSE-RoBERTa-BiLSTM | 0.9387/0.9179 | 0.7382/0.7943 |

Table 4: Precision and recall of different models vary across different categories

scores of RoBERTa alone do not perform very well, while all others have improved, and the ensemble model remains the most effective model.

Table 4 shows the precision and recall of each model on different categories. We can see that there is a significant gap between the precision and recall of all single models in the two categories. In the non-sexism category, both values are above 0.9, indicating that the model can detect this category very well. However, in the sexism category, the values are only around 0.7, indicating that the model's ability to detect this category is weaker. The reason for this is due to the imbalanced sample distribution, which results in insufficient feature learning for the sexism category. This is also a key focus for future work.

## 5 Conclusion

This paper describes our binary sexism detection system for subtask A of SemEval-2023 Task 10, including system design, implementation, and evaluation. The final experimental results demonstrate the effectiveness of our pre-trained language ensemble model for sexism detection and classification. Due to the limited time of this event, there are still some improvement strategies that we have not tried. For future improvements, we can start from these aspects: (1) mitigating the impact of imbalanced sample distribution through sampling or loss function adjustments, such as using over-sampling, class weight adjustment; (2) using unlabeled data to perform semi-supervised training so that the system can learn more valid information; (3) using more effective model ensemble strategies, such as XGBoost; (4) using multi-task deep pre-training (Zhang and Yang, 2021).

## References


Basant Agarwal and Namita Mittal. Text classification using machine learning methods-a survey. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, pages 701–709.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

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*.

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the 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.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

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*.

Hanjie Mai, Xiaobing Zhou, and Liqing Wang. A multi-task learning model for fine-grain dialogue social bias measurement. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part II*.

SreeJagadeesh Malla and PJA Alphonse. 2021. Covid-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. *Applied Soft Computing*, 107:107495.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bingkun Wang, Donghong Shan, Aiwan Fan, Lei Liu, and Jingli Gao. 2021. A sentiment classification method of web social media based on multidimensional and multilevel modeling. *IEEE Transactions on Industrial Informatics*, 18(2):1240–1249.

Qing Yu, Ziyin Wang, and Kaiwen Jiang. 2021. Research on text classification based on bert-bigru model. In *Journal of Physics: Conference Series*, volume 1746, page 012019. IOP Publishing.

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

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.