DUTIR at SemEval-2023 Task 10: Semi-supervised Learning for Sexism Detection in English

Bingjie Yu, Zewen Bai, Haoran Ji, Shiyi Li, Hao Zhang, Hongfei Lin*
School of Computer Science and Technology, Dalian University of Technology, China
(yubingjie,986427968,18920392258,lishiyieee,zh373911345)@mail.dlut.edu.cn
hflin@dlut.edu.cn

Abstract

Sexism is an injustice afflicting women and has become a common form of oppression in social media. In recent years, the automatic detection of sexist instances has been utilized to combat this oppression. The Subtask A of SemEval-2023 Task 10, Explainable Detection of Online Sexism, aims to detect whether an English-language post is sexist. In this paper, we describe our system for the competition. The structure of the classification model is based on RoBERTa, and we further pre-train it on the domain corpus. For fine-tuning, we adopt Unsupervised Data Augmentation (UDA), a semi-supervised learning approach, to improve the robustness of the system. Specifically, we employ Easy Data Augmentation (EDA) method as the noising operation for consistency training. We train multiple models based on different hyperparameter settings and adopt the majority voting method to predict the labels of test entries. Our proposed system achieves a Macro-F1 score of 0.8352 and a ranking of 41/84 on the leaderboard of Subtask A.

1 Introduction

With the evolution of the internet, people are free to express their opinions on social media. This may lead to the mass dissemination of hateful or abusive messages (Chiril et al., 2020), such as sexist expressions that people use intentionally or unintentionally. Sexism is a complex phenomenon, broadly defined as "prejudice, stereotyping, or discrimination, typically against women, on the basis of sex."[1] Sexism takes many forms, including blatant, covert, and subtle sexism (Swim et al., 2004), causing sexism detection to be a challenge for the filtering mechanism of platforms.

The SemEval-2023 Task 10 (Kirk et al., 2023) is an explainable detection task of online sexism, and we participate in Subtask A, the binary classification task, to detect whether an English-language post is sexist or not. The competition organizers released datasets collected from Gab and Reddit, containing 20,000 labeled entries and 2 million unlabeled entries.

For the sexism detection task, we build a system based on the pre-trained language model RoBERTa. Firstly, we further pre-train RoBERTa on the unlabeled corpus to obtain our domain-specific encoder. Furthermore, we reuse the abundant unlabelled data by Unsupervised Data Augmentation (UDA) (Xie et al., 2020) approach for consistency training, which is a common practice in semi-supervised learning. As for the noising operation, we apply the Easy Data Augmentation (EDA) (Wei and Zou, 2019) method to transform the unlabeled examples into their noised versions. The consistency training can propagate label information from labeled to unlabeled examples to enhance the generalization of the model. To further improve the robustness of our system, we train models based on different hyperparameter settings and perform majority voting on the predicted labels to obtain the final prediction results.

The structure of this paper is as follows: We first take a brief overview of related work in section 2, and then describe our proposed system in section 3. We introduce the details of our experiments in section 4, including the experimental settings, results, and discussions. Finally, a brief conclusion is presented in section 5.

2 Related work

As more and more people suffer or witness sexism on social media, automatically detecting sexist posts can help combat this phenomenon(Abburi et al., 2020). Many studies have been proposed to identify gender-based violence in texts. For misogyny, Anzovino et al. (2018) build a corpus of misogynous tweets labeled from different perspectives and conduct exploratory investigations. Aiming at multi-label fine-grained sexism classification,

[1] Oxford English Dictionary

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 892–896

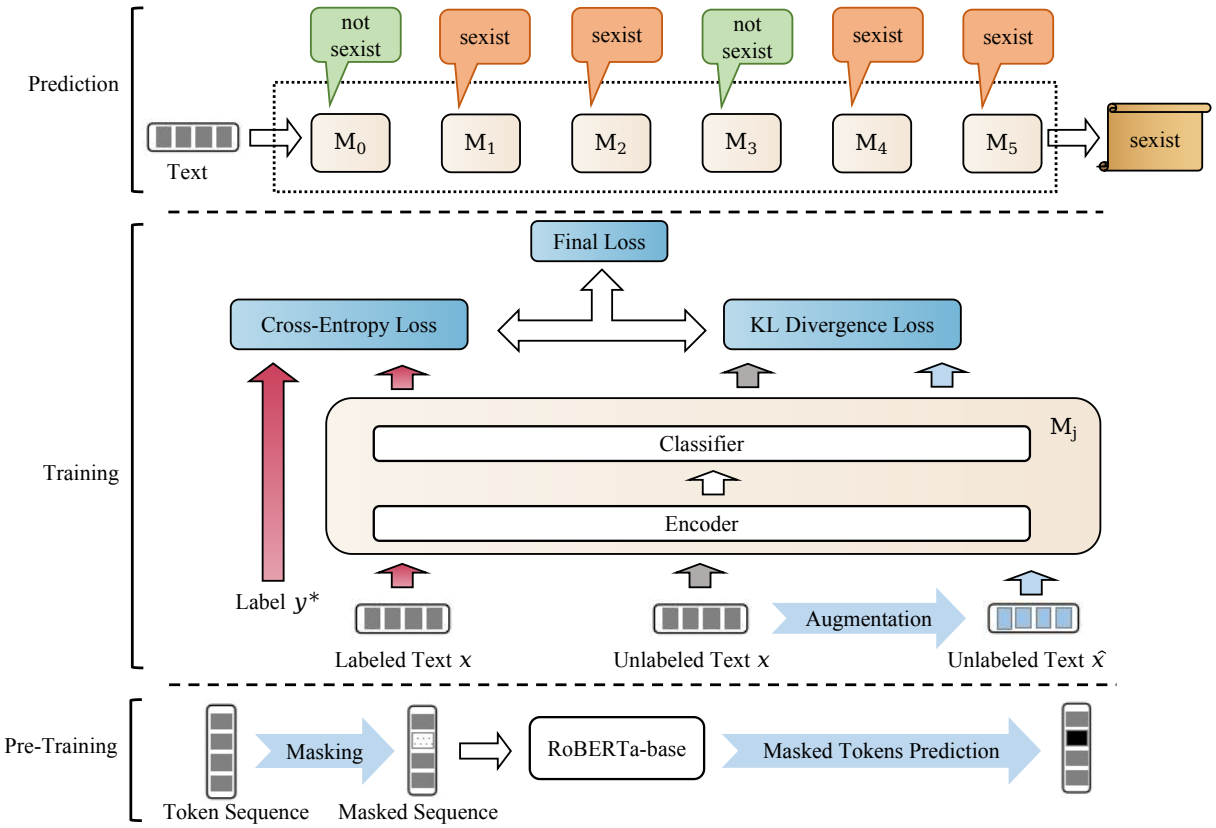

Figure 1: Model architecture of our method.

Abburi et al. (2020) introduce a semi-supervised multi-task neural approach. Inspired by speech acts theory and discourse analysis studies, Chiril et al. (2020) propose a new characterization of sexist content. Besides, some works (Badjatiya et al., 2017; Davidson et al., 2017; Espinosa Anke et al., 2019) regard sexism as a category of hate and detect it using hate speech classification methods.

3 System Description

Given a sentence $x = \{x_1, \cdots, x_n\}$ consisting of n tokens, the sentence is embedded as a vector by an encoder. We use a sentence-level embedding z to represent the entire sentence and feed it into a linear binary classifier to get the predicted label y. The goal of our system can be formulated as training a model $p_\theta(y|x)$, where the network parameters are denoted as θ, to predict the ground-truth label y^* for a given x.

As shown in Figure 1, the general implementation of our system can be described as 3 phases: pre-training, training, and prediction. In pre-training phase, we pre-train the language model on a relevant corpus so that the language model can learn more domain knowledge. Then we adopt the pre-trained language model as our domain-specific encoder. In training phase, we attach a linear classifier to the encoder and fine-tune the model through semi-supervised training. In prediction phase, we use the majority voting method to predict the final classification labels.

3.1 Pre-Training

As mentioned, a large-scale unlabeled corpus is available to participants, which we denote as $\mathcal{D}$. Since the labeled entries are sampled from this corpus, we consider $\mathcal{D}$ to have the same distribution as the labeled dataset D_l. Therefore, the unlabeled data can be used for self-supervised training of the pre-trained model, so that the generated embedding vectors will better match the domain-specific features. In our system, we adopt RoBERTa (Liu et al., 2019) as the pre-trained model, and pre-train it by Masked Language Model (MLM), which enables the model to learn the bidirectional contextual information of the text (Devlin et al., 2019). We then obtain a domain-specific language model denoted as RoBERTa$_A$ and serve it as the encoder for our system. Instead of using the entire $\mathcal{D}$, due to hardware and time constraints, we take a tenth of the entries from it and denote the subset as D_u.

We merge D_u with all the labeled entries as our pre-trained corpus.

3.2 Training

The sentence x is embedded by the domain-specific encoder as $s \in R^{n \times d}$, where d denotes the output hidden dimension size of the encoder. We adopt the last hidden state of $[CLS]$ token as the sentence-level embedding z and feed it into the linear binary classifier to get the output distribution $p_\theta(y|x)$.

In this phase, we fine-tune the model by minimizing the final loss $\mathcal{L}$ and optimizing the parameters θ based on the hyper-parameter setting β for getting the trained model M.

Supervised Learning. The supervised learning is conducted given the labeled data $(x, y^*) \in D_l$ with the aim of minimizing the divergence metric between ground-truth label y^* and the predicted label $y \sim p_\theta(y|x)$. The divergence metric is calculated by Cross-Entropy and denoted as $CE(y^*, y)$. With supervised learning, θ is optimized such that $p_\theta(y|x)$ approaches y^* for given (x, y^*).

Semi-supervised Learning. To enforce the robustness and smoothness of the model, we adopt the Unsupervised Data Augmentation (UDA) approach (Xie et al., 2020) to perform consistency training. To be concrete, given an unlabeled data $x \in D_u$ and its augmented version $\hat{x}$, we input the x and $\hat{x}$ into the model and then obtain their output distributions $p_\theta(y|x)$ and $p_\theta(y|\hat{x})$, respectively. With the minimization of the divergence metric between these two distributions, the model can constrain predictions to be invariant to input noise. We observe that consistency training is performed unsupervised so that the information of a large amount of unlabeled data can be learned. We use the Kullback-Leibler (KL) divergence to measure the above distribution divergence and define it as follows:

$$KL(x, \hat{x}) = D_{KL}(p_\theta(y|x)||p_\theta(y|\hat{x})) \quad (1)$$

Therefore, we set the final loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \sum^{N} CE(y^*, x_1) + \sum^{\alpha N} KL(x_2, \hat{x_2}) \quad (2)$$

where $(x_1, y^*) \in D_l$, $x_2 \in D_u$, $\hat{x_2}$ is the augmented version of x_2, N is the batch size of supervised data and α is a hyperparameter to control the ratio of unsupervised data in a batch.

In addition, we employ additional training strategies mentioned in Xie et al. (2020), including confidence-based masking and sharpening predictions, to solve some common training problems.

Data Augmentation. We employ Easy Data Augmentation (EDA), a simple but powerful data augmentation method, transforming unlabeled data x into $\hat{x}$. Details of the EDA augmentation operations are shown in Table 1. Here we list an example of a text and its augmented version:[2]

- **Original text**: *Is he really being an asshole? He's just stating what he likes.*
- **Augmented version**: *what is really an being asshole? he's just stating likes.*

Operation	Description	γ
SR	Synonym Replacement	0.3
RI	Random Insertion	0.3
RS	Random Swap	0.5
RD	Random Deletion	0.3

Table 1: The augmentation operations that we utilized. The γ indicates the proportion of word changes in a text.

3.3 Prediction

In this phase, we employ the models fine-tuned during training phase to predict whether the entry to be identified is sexist by majority voting, a simple ensemble method. Specifically, we optimize the model parameters based on the hyperparameter setting β_i, where $i \in \{0, \cdots, 5\}$, and get the corresponding fine-tuned model M_i. Given the entry x, we feed x into these fine-tuned models to obtain the classification result, i.e., *sexist* or *not sexist*, and count the number of *sexist* in these predicted labels. If the number exceeds 2, we mark x as *sexist*, and vice versa as *not sexist*. The details of hyperparameter setting are explained in 4.

4 Experiments

4.1 Experimental Settings

Datasets. The statistics of the labeled datasets are shown in Table 2. The unlabeled dataset D_u consists of 200,000 entries sampled from the unlabeled corpus with 2 million, and we adopt it for pre-training and semi-supervised learning. For all the data, we do not perform any other preprocessing because we consider that the competition organizers have already completed the initial clean-up and

[2]The texts contain offensive speech, and we oppose any use of this kind of speech act.

Dataset	Label		Total
	Sexist	Not sexist	
Training	3398	10602	14000
Dev	486	1514	2000
Test	970	3030	4000

Table 2: Statistics of the labeled datasets.

preparation of the data, and additional processing may lead to loss of information.

Evaluation and Model Saving. We evaluate the official metric, macro-F1 score of the model on the development set once every 100 training iteration steps and save the model with the highest current score. The iteration is terminated when macro-F1 stops growing in 20 consecutive evaluations.

Hyperparameters. The fixed hyperparameter settings of our system are described below. We adopt the RoBERTa-base as the pre-trained language model and AdamW as the optimizer, both provided by huggingface[3]. The pre-training for RoBERTa-base is conducted for 700 epochs with a batch size of 16. We use BCEWithLogitsLoss and KLDivLoss in Pytorch to compute the supervised and unsupervised losses. Since the labels of D_l are imbalanced, the parameter *pos_weight* for BCEWithLogitsLoss is set to 4. For confidence-based masking and sharpening predictions, we set the threshold and Softmax temperature to 0.45 and 0.85, respectively. In addition to the fixed hyperparameters, for majority voting, we set the learning rate from 1e-5 to 5e-5 in increments of 1e-5, the batch size of supervised data to 8, 16, 32, or 64, and the ratio α of unsupervised data to from 1 to 6 in increments of 1.

Dataset	Model	Macro-F1
Dev	RoBERTa-base	0.8205
	RoBERTa$_A$	0.8386
	RoBERTa$_A$+UDA	0.8545
Test	RoBERTa-base	0.8208
	RoBERTa$_A$	0.8289
	RoBERTa$_A$+UDA	0.8348
	RoBERTa$_A$+UDA+Vote	0.8352

Table 3: The main experimental results for Subtask A. RoBERTa$_A$ is the further pre-trained RoBERTa-base. Vote denotes the majority voting.

[3] https://huggingface.co/

4.2 Results and Analysis

Table 3 lists the experimental results of our system on the development set and test set. All models are trained using the same random seed, except for the model with the Vote on the test set, which integrates several models trained based on different hyperparameter settings. We can observe from the results that:

- Our RoBERTa$_A$ outperforms RoBERTa-base in both the development and test sets, proving the benefit of domain-specific **pre-training**. As stated earlier, pre-training the language model on the specific distribution is considered to make the output hidden states better match the domain features.

- The **UDA** approach improves the performance of the model, especially in the Development set, showing the effectiveness of data augmentation and semi-supervised learning for this task. This motivates us to continue exploring how to make better use of unlabeled data information to enhance the generalization of models in future research.

- The **majority voting** method has slight improvement on the models on the test set, probably because of the considerable homogeneity among the base models involved in voting (only the hyperparameter settings differ, not the model structure). We should examine models with various structures to make the majority voting more useful.

5 Conclusion

This paper describes our approach in SemEval-2023 Task 10 Subtask A to detect sexism in English-language social media posts. Our system employs self-supervised and semi-supervised learning to pre-train and fine-tune the pre-trained language model RoBERTa-base respectively. We demonstrate the effectiveness of our system. There are still considerable areas for improvement in our system, and how to better utilize the unlabeled data is an essential issue for us to explore in the future.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2020. Semi-supervised multi-task

learning for multi-label fine-grained sexism classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 5810–5820, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, page 57–64, Cham. Springer International Publishing.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in french tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, page 1397–1403, Marseille, France. European Language Resources Association.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luis Espinosa Anke, Thierry Declerck, Dagmar Gromann, Ziqi Zhang, Lei Luo, Dagmar Gromann, Luis Espinosa Anke, and Thierry Declerck. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semant. Web*, 10(5):925–945.

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.

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*, abs/1907.11692.

Janet K. Swim, Robyn Mallett, and Charles Stangor. 2004. Understanding subtle sexism: Detection and use of sexist language. *Sex Roles*, 51(3):117–128.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.