Tsingriver at SemEval-2023 Task 10: Labeled Data Augmentation In Consistency Training

Yehui Xu
Yunnan University
yehuixu24@163.com

Haiyan Ding*
Yunnan University
teidhy@163.com

Abstract

Semi-supervised learning has promising performance in deep learning, one of the approaches is consistency training on a large amount of unlabeled data to constrain model predictions to be invariant to input noise. However, The degree of correlation between unlabeled data and task objective directly affects model prediction performance. This paper describes our system designed for SemEval-2023 Task 10: Explainable Detection of Online Sexism. We utilize a consistency training framework and data augmentation as the main strategy to train a model. The score obtained by our method is 0.8180 in subtask A, ranking 57 in all the teams.

1 Introduction

Sexism is an increasingly serious network problem. It may harm targeted women, make cyberspace inaccessible and unpopular, and perpetuate social asymmetry and injustice. Explainable Detection of Online Sexism task targets to predict whether a sentence is sexism (Jane, 2014).

A small amount of labeled data limits the improvement of supervised learning model performance. In contrast, semi-supervised and self-supervised learning methods can learn from massive unlabeled data (Chen et al., 2020), even better than supervised learning. Bidirectional Encoder Representation from Transformers(BERT) and Robustly optimized BERT approach(RoBERTa) (Zhuang et al., 2021) are classic models that are trained by unlabeled data in unsupervised learning tasks. In this work, we aim to apply a semi-supervised learning method to accomplish the task, using labeled data as the train data of supervised training and augmented labeled data as the train data of unsupervised training(consistency training), respectively. Due to the difficulty of selecting domain-relevance unlabeled data in consistency learning and the fact that the labeled data is domain-relevance data that matches with task objective, we decide to augment labeled data instead of unlabeled data that mismatch the distributions of labeled data.

Improving model performance is as important as ensuring model stability (Gunel et al., 2020), which is why we choose to use consistency learning. Our method guarantees the same prediction result for the same sentence expressed with different words in a certain probability. In other words, it improves the model's performance in detecting synonymous sentences. To not lose BERT's original performance as much as possible, we combine two strategies: consistency training and synonym replacement method based on TF-IDF to fine-tune BERT, instead of directly retraining BERT (Ghalandari et al., 2022). Compared with the supervised learning method training our baseline model in this paper, our main training strategies have significantly improved model performance from 0.8004 to 0.8180. We have released the code at https://github.com/vincent-hyx/task_10.

2 Background

There are two main directions of research work, one is consistency training (Tarvainen and Valpola, 2017), and the other is unsupervised data augmentation(UDA) (Xie et al., 2020). The consistency training method relies on unmarked data training to overcome the weakness of supervised learning that requires a large number of labeled data (Tian et al., 2022). It evaluates the consistency between corresponding sentences that contain a pair of sentences with similar semantics. Data augmentation has achieved outstanding results (Baek et al., 2022). For instance, EDA(Easy Data Augmentation) (Wei and Zou, 2019)applies to most NLP tasks limited by a small amount of labeled data. Besides, to solve the problem of limited annotation data, computer vision tasks utilize auto-augmentation (Cubuk et al., 2019) to generate additional data. It is worth noting

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 782–786

that the proposed UDA method provides a way to obtain high-quality, unlabeled and supplementary data for consistency learning.

The general consistency training framework uses KL divergence as an evaluation of the consistency between the augmented data and the original data, and simultaneously combines it with a supervised learning approach. UDA method Using consistency training in the research of text classification proposes two data augmentation methods, back-translation and TF-IDF word replacement (Xie et al., 2020), to generate the data required for consistency training.

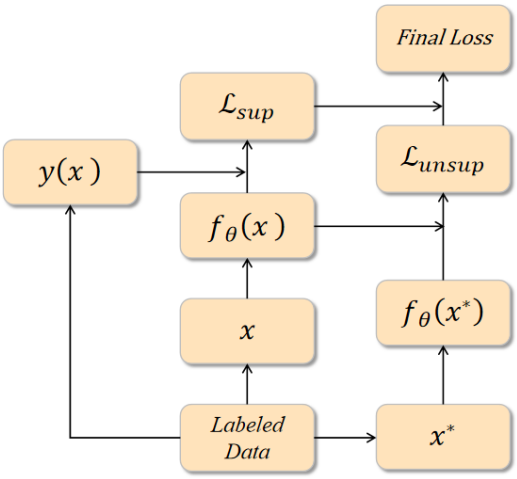

Figure 1: Model Framework

3 System Overview

3.1 Model Structure

In our model, we follow the semi-supervised learning framework(Figure 1) for consistency training, and use different loss functions in the supervised learning part of the framework to improve the model's performance. Formally, the full objective can be written as follows:

$$\min_{\theta} \mathcal{L}(\theta) = \mathcal{L}_{sup} + \mathcal{L}_{unsup} \quad (1)$$

$$\mathcal{L}_{sup1} = CE(y(x), f_\theta(x)) \quad (2)$$

$$\mathcal{L}_{sup2} = FL(y(x), f_\theta(x)) \quad (3)$$

$$\mathcal{L}_{unsup} = KL(f_\theta(x^*), f_\theta(x)) \quad (4)$$

where CE and KL denote cross entropy and KL-divergence respectively, $y(x)$ is the correct label vector with respect to x, $f_\theta(x)$ is the label vector predicted by the model with θ as weight parameters, and x^* is augmented data with respect to labeled data x. FL here refers to Focal Loss (Lin et al., 2020), we find that Focal Loss replacing the cross entropy of the supervised learning part improves the results of our method.

In the model design, we choose the structure of the pre-training language model BERT-base (Devlin et al., 2019) with three linear layers.

3.2 Data Augmentation Method

According to the consistency training framework, the unlabeled data and the augmented data must be domain-relevance data, which means the distribution of unlabeled data must match the distribution of labeled data. Considering this problem is difficult to deal with, we decided to use the labeled data and the synonym replacement method based on TF-IDF to generate augmented data instead of using unlabeled data to generate augmented data. It can be written as follows:

$$x^* = DA(x) \quad (5)$$

In addition to using the officially provided training data, we have generated additional augmented data through the use of the synonym replacement method based on TF-IDF. This method first performs word frequency statistics on the document of labeled data then calculates TF-IDF scores (Abdelminaam et al., 2021) for a word in the sentences. We replace the words below the threshold in the sentence with their synonyms to obtain the required augmented data. Since the algorithm for calculating the TF-IDF weight of words based on documents has been widely used, it will not be repeated here. The specific algorithm steps are are shown in Algorithm 1.

It is worth emphasizing that, in this algorithm, when searching for synonyms of a word, we should find out all its synonyms, then randomly select and replace them, which is equivalent to increasing the richness of the vocabulary in the sentence, without causing the word frequency of a word to be so high that misleads model thinks that this word is an important feature.

Algorithm 1: synonym replacement method based on TF-IDF

Input: $all\_sentence, tfidf, threshold$
Output: $new\_all\_sentence$

```
begin
    for setence ∈ all_sentence do
        for word ∈ sentence do
            if tfidf[sentence][word] < threshold then
                synonyms ← findSynonyms(word)
                word ← randomSelect(synonyms)
            end
        end
    end
    return new_all_sentence
end
```

3.3 Analysis

The consistency training procedure essentially enforces the model to be insensitive to the noise and hence smoother with respect to changes in the input space (Shao et al., 2021), thus improving the model's generalization ability. However, the out-of-domain data can easily result in performance loss (Oliver et al., 2018), which has been proven when we use the given unlabeled data from Gab and Reddit for training. This is the reason why we only select the labeled data to generate the augmented data required for consistency training. The results section will also show the impact of comparing the two kinds of data on the model.

Our data augmentation method makes the augmented data still belong to the domain-relevance data while introducing noise. Suppose that the label data x belongs to an unknown distribution. When we apply the previously mentioned data enhancement method to generate x^*, x^* still belongs to the distribution, because the replaced word has a low TF-IDF weight value, which means that the semantics will not be affected after the synonym is replaced.

Based on the previously mentioned, the selection of the superparameter threshold has an important impact on the effect of data enhancement. If the threshold is too large, it cannot increase the richness of words in the sentence and introduce the required noise, while if the threshold is too small, it will lead to the loss of important feature words in the sentence. Therefore, we get the model's performance under different thresholds through experiments, which are shown in the results section.

4 Experimental setup

In this work, our model accepts input data derived from given train data, while the model's output is the probability of binary classification(sexist or not sexist). The training data consists of 14,000 entries in English, of which 3,398 are sexist (Kirk et al., 2023).

In this task, we divide all tagged data into three parts, one accounting for 90% for training(train), one accounting for 5% for verification(dev), and one accounting for 5% for testing(test). In addition, the data used in the unsupervised learning setting is derived from the dataset for training augmented by the previously mentioned data augmentation approach, which has 14,000 examples.

In the parameters setting, it is appropriate to set the learning rate at 1e-5, and we use a batch size of 20 for the supervised loss and use a batch size of 40 for the unsupervised loss. Another superparameter about the temperature in softmax function is set to 0.8 for our experiment. If the focal loss is used as supervised loss, we recommend setting the parameter gamma at 2.

Because the original training data contains some unnecessary characters and emoji expressions, we performed a general data cleaning operation before using the data for training, and all the related code has been included in the project file uploaded to GitHub.

5 Results

In contrast to our method, we trained a baseline model which only utilizes the supervised learning method with cross entropy as a loss function, but it uses the same model structure as that used in our method.According to official requirements, we adopt macro F1(Ave F1 called in Figure 2) as evaluation metric to present the results.

In this task, we compare the results with and without our method, and compare the effect of focal

loss and cross entropy loss on our method. Our method scored 0.8180 in task A, ranking 57 in all the teams. All results are shown in Table 1.

Table 1: results for main strategy

	score for CE	score for FL
baseline	0.8004	×
our	0.8108	0.8180

In addition to showing the macro F1 score obtained through our method, we also want to show the impact of different threshold selection on the model performance, and the degradation of performance caused by using unlabeled data from Gab and Reddit for the model trained by consistency training approach.

Table 2: results of using labeled data and using unlabeled data from gab and reddit

	score for CE	score for FL
labeled data	0.8104	0.8180
unlabeled data	0.7855	0.7933
baseline	0.8004	×

The macro F1 score of the baseline in Table 2 is identical to the one in Table 1, and we put it here to show that the results of using unlabeled data from Gab and Reddit are not only inferior to those of using labeled data, but even worse than the baseline model.

We tested some different thresholds, crucial for the data augmentation strategy, and found that 0.8 is the optimal threshold corresponding macro F1 score is 0.818. It is shown as follows:

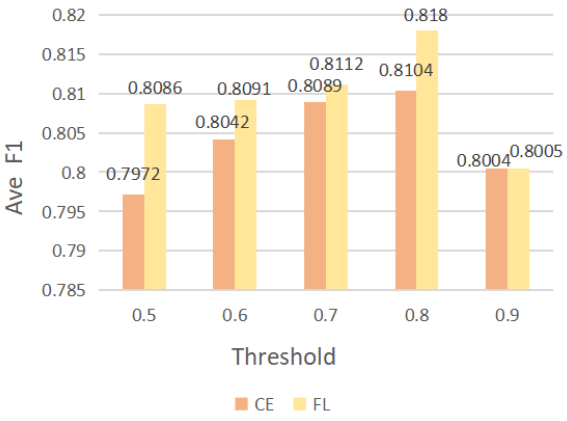

Figure 2: macro F1 score corresponding to different threshold

6 Conclusion

Compared with the baseline model, we use the semi-supervised consistency training framework and the data-augmented method of synonym replacement based on TF-IDF, which has significantly improved model performance from 0.8004 to 0.8180. In the future, we will explore different data augmentation strategies to provide more high-quality samples for consistency training, and research how to design better loss functions for consistency training and more training skills.

References

D. S. Abdelminaam, N. Neggaz, Iae Gomaa, F. H. Ismail, and A. Elsawy. 2021. Arabicdialects: An efficient framework for arabic dialects opinion mining on twitter using optimized deep neural networks. *IEEE Access*, PP(99):1–1.

Francis Baek, Daeho Kim, Somin Park, Hyoungkwan Kim, and Sanghyun Lee. 2022. Conditional generative adversarial networks with adversarial attack and defense for generative data augmentation. *Journal of Computing in Civil Engineering*, 36(3). Publisher Copyright: © 2022 American Society of Civil Engineers.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc.

Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2019. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123.

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)*, pages 4171–4186. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, and Georgiana Ifrim. 2022. Efficient unsupervised sentence compression by fine-tuning transformers with reinforcement learning. *ArXiv*, abs/2205.08221.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *ArXiv*, abs/2011.01403.

Emma A. Jane. 2014. "your a ugly, whorish, slut". *Feminist Media Studies*, 14(4):531–546.

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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318 – 327.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Nian Shao, Erfan Loweimi, and Xiaofei Li. 2021. Rct: Random consistency training for semi-supervised sound event detection. In *Interspeech*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jinchuan Tian, Jianwei Yu, Chao Weng, Shi-Xiong Zhang, Dan Su, Dong Yu, and Yuexian Zou. 2022. Consistent training and decoding for end-to-end speech recognition using lattice-free mmi. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7782–7786.

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)*, pages 6382–6388. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.