HITSZQ at SemEval-2023 Task 10: Category-aware Sexism Detection Model with Self-training Strategy

Ziyi Yao[1], **Heyan Chai**[1], **Jinhao Cui**[1], **Siyu Tang**[1] and **Qing Liao**[1,2*]
[1] Harbin Institute of Technology, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
{yaoziyi,chaiheyan,cuijinhao,tangsiyu999}@stu.hit.edu.cn,
liaoqing@hit.edu.cn

Abstract

This paper describes our system used in the SemEval-2023 *Task 10 Explainable Detection of Online Sexism (EDOS)*. Specifically, we participated in subtask B: *a 4-class sexism classification task*, and subtask C: *a more fine-grained (11-class) sexism classification task*, where it is necessary to predict the category of sexism. We treat these two subtasks as one multi-label hierarchical text classification problem and propose an integrated sexism detection model for improving the performance of the sexism detection task. More concretely, we use the pretrained BERT model to encode the text and class label and a hierarchy-relevant structure encoder is employed to model the relationship between classes of subtasks B and C. Additionally, a self-training strategy is designed to alleviate the imbalanced problem of distribution classes. Extensive experiments on subtasks B and C demonstrate the effectiveness of our proposed approach.

1 Introduction

Sexism is especially rampant online due to the anonymity of the Internet, making the online community unfriendly towards women (Mosca et al., 2021; Abburi et al., 2020; García-Baena et al., 2022). Detecting sexism on social media is meaningful in building a more harmonious Internet world. Sexism detection (Ahuir et al., 2022; del Arco et al., 2022) can be considered as a particular sentiment classification task, applied to predict the categories (the detailed categories are shown in Figure 1) of prejudice or discrimination based on a person's gender in the whole sentence. For example, the sentence *"As Roosh said having discussions with woman is a waste of time anyway."* is a sexist comment, which belongs to a kind of *descriptive attack* expressing *derogation* against women.

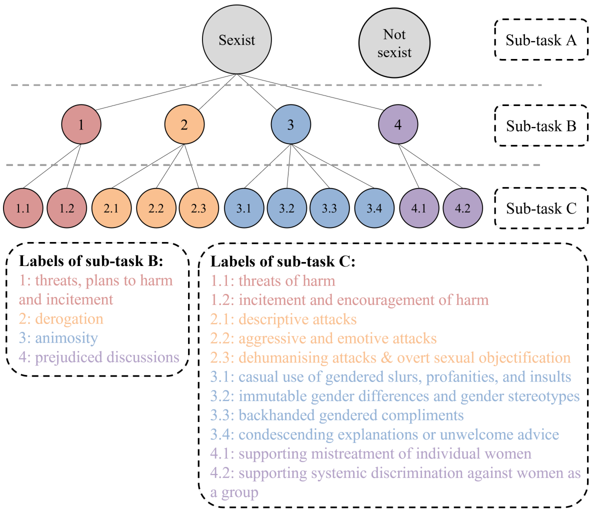

Sexist
Not sexist
Sub-task A
1
2
3
4
Sub-task B
1.1
1.2
2.1
2.2
2.3
3.1
3.2
3.3
3.4
4.1
4.2
Sub-task C
Labels of sub-task B:
1: threats, plans to harm and incitement
2: derogation
3: animosity
4: prejudiced discussions
Labels of sub-task C:
1.1: threats of harm
1.2: incitement and encouragement of harm
2.1: descriptive attacks
2.2: aggressive and emotive attacks
2.3: dehumanising attacks & overt sexual objectification
3.1: casual use of gendered slurs, profanities, and insults
3.2: immutable gender differences and gender stereotypes
3.3: backhanded gendered compliments
3.4: condescending explanations or unwelcome advice
4.1: supporting mistreatment of individual women
4.2: supporting systemic discrimination against women as a group

Figure 1: Taxonomic hierarchy structure of the three subtasks.

SemEval-2023 Task 10 (Kirk et al., 2023) (Explainable Detection of Online Sexism, EDOS), is an online sexism detection and classification task. The organizers provide a labeled dataset and two unlabeled datasets for training the three subtasks. The labeled dataset consists of 14,000 annotated samples of which 3398 are sexist. For the sexist entries, as shown in Figure 1, fine-grained labels of the 4 categories (subtask B) and 11 vectors (subtask C) are also provided.

Subtask A is a binary classification task aimed at predicting whether a post is sexist or not. Subtask B and subtask C classify the sexist posts into more fine-grained categories. There are four categories for subtask B, which are threats, derogation, animosity, and prejudiced discussions. For subtask C, there are eleven target vectors. The hierarchical structure of the three tasks is shown in Figure 1. One obvious characteristic of subtasks B and C is that categories in subtask C are the detailed classification of categories in subtask B, which means all the instances from the one category in subtask C are labeled the same in subtask B. Labels in subtask

[*] Corresponding Author

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 934–940
July 13-14, 2023 ©2023 Association for Computational Linguistics

B and C put together form a 15-class taxonomic hierarchy.

In this paper, we try to solve subtasks B and C simultaneously and treat these two tasks as one multi-label hierarchical text classification problem (Chen et al., 2020; Huang et al., 2021; Zhang et al., 2022). We investigate the EDOS dataset and summarize the two observations: Class label text holds important textual information regarding the class, and, the hierarchical structure of the classes preserves richer semantics dependency information between tasks. Based on it, we propose an approach based on the pre-trained BERT model (Devlin et al., 2019), Graph Convolutional Network (GCN)(Deng et al., 2021), and a retrieval-based attention mechanism (Vaswani et al., 2017) to obtain the final prediction. First, we use BERT model to obtain the sentence and label representation. Then, we propose a hierarchy-relevant structure encoder to capture the richer class representation. Moreover, a class-aware attention mechanism is proposed to capture the relationship between text and classes. In addition, a self-training strategy (Niu et al., 2020; Soleimani et al., 2022) is designed to alleviate the imbalanced problem of distribution classes, to get a better performance of the sexism detection model. Our system achieves the Macro F1 scores for subtasks B and C of 0.6030 and 0.3847, respectively.

2 System Overview

2.1 Task Definition

We define the combination of subtasks B and C of EDOS as a detailed hierarchical text classification problem. The hierarchy is predefined according to the class relationships shown in Figure 1.

Given a sentence consisting of n words $S = \{w_1, \cdots, w_n\}$, we construct a BERT-based pair sequence *"[CLS] S [SEP]"* and feed the sequence into pre-trained language model BERT (Devlin et al., 2019) to acquire the sentence representation H. As to class labels, 4 classes labeled $Y_B = \{y_1, y_2, y_3, y_4\}$ are provided for subtask B and 11 classes with label $Y_C = \{y_5, y_6, \cdots, y_{15}\}$ for subtask C. $child(i), i \in \{1, 2, 3, 4\}$ refers to the set of child nodes of label node i. As we consider the two tasks together, the 15 classes are labeled $Y = \{Y_B, Y_C\} = \{y_1, y_2, \cdots, y_{15}\}$. The goal of the task is to predict one label in Y_B and one label in Y_C for each input sentence.

2.2 Model Architecture

Figure 2 shows the overall architecture of our system. Our model mainly consists of four modules: 1) Pre-trained label encoder extracts contextualized class-label features through BERT. 2) Hierarchy-relevant structure encoder utilizes graph convolution neural network to produce class-structure feature. 3) Class-sentence attention module computes the attention score between the class feature, which is the concatenation of class-label and class-structure feature, and the input text feature to produce class-relevant text feature. 4) A self-training strategy is adopted to deal with unbalanced class distribution and an insufficient amount of labeled data.

2.2.1 Pre-trained Label Encoder

BERT (Devlin et al., 2019) encoder is utilized to extract class-label features. We input the class labels in the form $y_i = \{[CLS], w_{i_1}, w_{i_2}, \cdots, w_{i_n}, [SEP]\}, i \in \{1, 2, \cdots, 15\}$ and use the pooled output h_i^y from the encoder as the class-label feature $H^y = \{h_1^y, h_2^y, \cdots, h_{15}^y\}$.

2.2.2 Hierarchy-Relevant Structure Encoder

To capture the structure feature between classes, we construct a graph $G = \{A, H^s\}$, where vertices are the 15 classes in Y with randomly initialized representation (Glorot and Bengio, 2010). Following the study of HiGAM (Zhou et al., 2020), the prior probability of label dependencies is considered when building the adjacent matrix A. Suppose N_i is the number of entries of class label y_i, A is formulated as follows:

$$A_{ij} = \begin{cases} \frac{N_j}{N_i} & i \in Y, j \in child(i) \\ 1 & j \in Y, i \in child(j) \\ 1 & i = j \\ 0 & otherwise \end{cases} \tag{1}$$

We then use a graph convolutional neural network (GCN) (Kim et al., 2017) to obtain the relationship between classes. Formally,

$$H_l^s = \sigma(AW^s H_{l-1}^s + b^s) \tag{2}$$

where l is the layer of GCN. After graph convolution, H_l^s denotes the class structure feature of each class.

2.2.3 Class-Sentence Attention Module

We concatenate class-structure feature H_l^s and class-label feature H^y as the final class feature

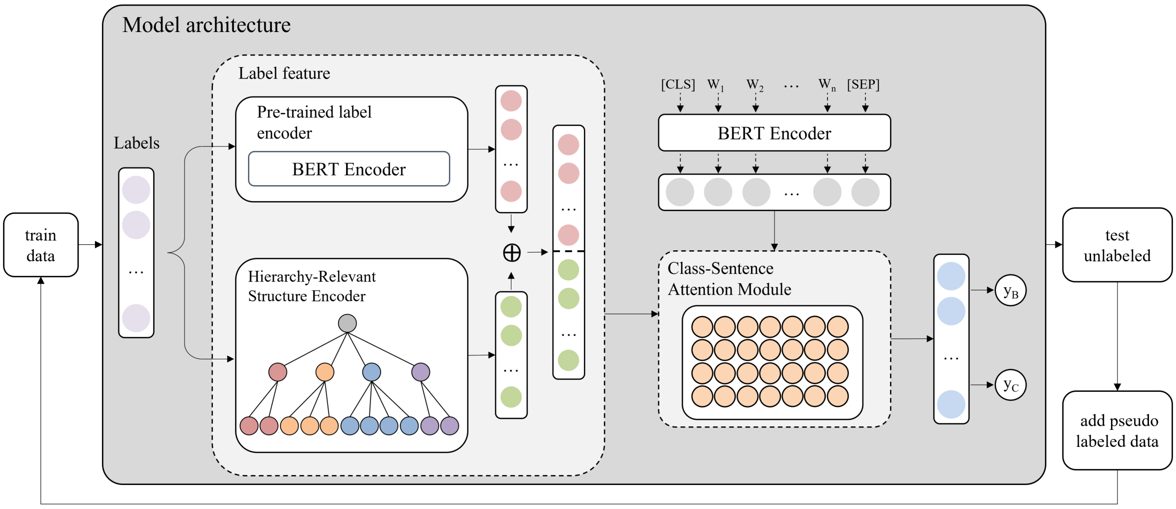


Figure 2: Model architecture with self-training.

representation $H^c = \{H_l^s; H^y\}$. We compute the attention score between sentence feature H^s and class feature H^c as:

$$H^{att} = softmax(\frac{HW^Q \times (H^c W^K)^T}{\sqrt{d}}) \quad (3)$$

where $H^{att} = \{h_1^{att}, h_2^{att}, \cdots, h_{15}^{att}\}$ is used for the final class-relevant sentence representation.

2.2.4 Classification

We use fully-connected layers to calculate the prediction probability distribution for subtasks B and C:

$$y_B = \sigma(W_B H_B^{att} + b_B) \quad (4)$$

$$y_C = \sigma(W_C H_C^{att} + b_C) \quad (5)$$

where $H_B^{att} = \{h_1^{att}, h_2^{att}, h_3^{att}, h_4^{att}\}$, $H_C^{att} = \{h_5^{att}, \cdots, h_{15}^{att}\}$. The final objective function of subtask B and C are defined by cross-entropy loss, respectively,

$$\mathcal{L}_B = -\frac{1}{N}(\sum_{i=1}^{N}\sum_{j=1}^{4} y_{ij}^{'} log(y_{Bij})) \quad (6)$$

$$\mathcal{L}_C = -\frac{1}{N}(\sum_{i=1}^{N}\sum_{j=5}^{15} y_{ij}^{'} log(y_{Cij})) \quad (7)$$

The final loss the model is:

$$\mathcal{L} = \alpha\mathcal{L}_B + \beta\mathcal{L}_C \quad (8)$$

where both α and β are set as 0.5 during training.

2.2.5 Self-training Strategy

The labeled dataset provided by SemEval-2023 Task 10 is very unbalanced. For example, class "*2.1 descriptive attacks*" has over 700 instances while class "*3.4 condescending explanations or unwelcome advice*" only has less than 50 instances. As a consequence, we utilize a self-training strategy to mitigate this problem based on CReST.

Specifically, we add certain pseudo-labeled data to the training set from the *Reddit* unlabeled dataset provided by the organizers. After each epoch, the unlabeled data is first tested on the newly attained model. Then, sentences with the prediction probability on both tasks higher than the corresponding terms in threshold list $T = \{T_1, T_2, \cdots, T_{15}\}$ are added to the training set. Suppose there are a total of N_{tot} instances in the labeled dataset, the threshold list is calculated as follows:

$$T_i = \begin{cases} 0.5 + 0.5 * \frac{N_i}{N_{tot}} & i \in \{1, 2, 3, 4\} \\ 0.2 + 0.8 * \frac{N_i}{N_{tot}} & i \in \{5, \cdots, 15\} \end{cases} \quad (9)$$

Higher thresholds are assigned to more frequent classes and lower thresholds to rare classes. According to CReST, classification models have a much higher accuracy on minority classes. To be more specific, models are usually reluctant to classify instances into minority classes, but once such a prediction is made, the accuracy is very high. Therefore, we make it easier to add less frequent class samples and harder to add the major classes.

Only sexist entries are needed for subtasks B and C, so we select the sexist entries from the original

B Label	B Size	C Label	C Size
1. threats	310	1.1 threats of harm	56
		1.2 incitement and encouragement of harm	254
2. derogation	1590	2.1 descriptive attacks	717
		2.2 aggressive and emotive attacks	673
		2.3 dehumanising attacks & overt sexual objectification	200
3. animosity	1165	3.1 casual use of gendered slurs, profanities, and insults	637
		3.2 immutable gender differences and gender stereotypes	417
		3.3 backhanded gendered compliments	64
		3.4 condescending explanations or unwelcome advice	47
4. prejudiced discussions	333	4.1 supporting mistreatment of individual women	75
		4.2 supporting systemic discrimination against women as a group	258

Table 1: Statistics of the labeled dataset

Models	Subtask B		Subtask C	
	$ACC.(\%)$	$F1.(\%)$	$ACC.(\%)$	$F1.(\%)$
BERT (Devlin et al., 2019)	54.64	47.01	39.67	23.09
RoBERTa (Liu et al., 2019)	55.09	50.96	40.79	26.03
HiAGM-LA (Zhou et al., 2020)	60.59	55.32	46.35	36.27
HTCInfoMax (Deng et al., 2021)	61.03	58.83	46.91	38.87
Our model + roberta-base	59.56	57.68	45.59	40.03
Our model + roberta-large	60.44	58.79	45.15	39.93
Our model + xlm-roberta-large	59.36	56.18	46.72	34.87
Our model + bert-base	60.88	59.19	47.21	41.61
Our model + bert-large	**62.65**	**60.31**	**54.12**	**44.34**
best model	-	73.26	-	56.06

Table 2: Experimental results of our proposed model compared to several comparison baselines.

Reddit unlabeled dataset through a BERT-based binary classification model. We train a binary classification model The sexist samples form the unlabeled dataset used in our model.

3 Experimental Results and Analysis

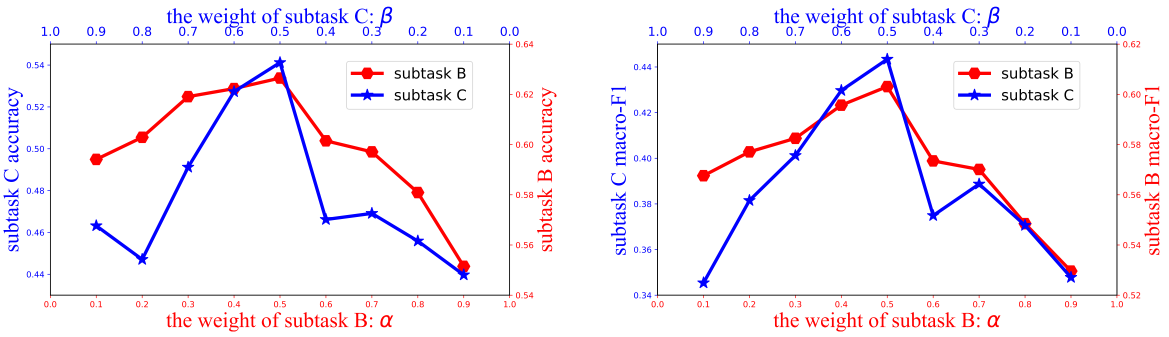
3.1 Experimental Setup

Dataset. We use the EDOS datasaet provided by SemEval-2023 Task 10 (Kirk et al., 2023). Detailed statistics of the labeled dataset are illustrated in Table 1. The labeled dataset is divided into a training dataset (80%) and a validation dataset (20%). We randomly sampled 80% of samples for each class in subtask C and used the rest 20% for validation.

Implementation Details. The pre-trained BERT encoder, which is used as the sentence encoder and class-label encoder, is set with a hidden dim of 1024 and a maximum length of token input of 100. The label embedding for class-structure feature H_0^s is initialized with Kaiming uniform (Wei et al., 2021). To avoid overfitting, dropout is employed at a rate of 0.5 for sentence embedding and 0.1 for the GCN and attention. We select Adam as the optimizer with a learning rate of 2×10^{-4}. The model is trained for 30 epochs with a batch size of 32. We assess the prediction results by calculating the accuracy and macro F1 of each subtask.

3.2 Experimental Results

To evaluate our model, we compare its performance on the development test dataset with several baselines , such as BERT, RoBERTa, HiAGM-LA, and HTCInfoMax. The comparison results are reported in Table 2. Our model with *bert-large* as the pre-trained BERT model outperforms all the



(a) Accuracy of subtask B and C.

(b) Macro F1 of subtask B and C.

Figure 3: The effect of task weight α and β.

other methods on both evaluation metrics in both tasks, which demonstrates that our model can solve the two sub-tasks effectively. Compared to bert-based models, our model adds additional hierarchical structure information to model the relationship between the labels of the two subtasks. Moreover, the hierarchical classification models only consider the structural relationship without making use of the label text. Our approach, which fully utilized the contextualized information from the label texts, is proven to be better. Among the variants of our model, using *bert-large* as the pre-trained encoder has the best performance surprisingly. We argue that the training dataset initially includes only 2718 labeled instances for a 15-class classification problem, making larger and more complicated models easily overfitted. This is the main reason why BERT as the encoder outperforms RoBERTa. The best model in the competition reports a Macro-F1 of 73.26 in subtask B and 56.06 in subtask C.

3.3 Ablation Study

To evaluate the effectiveness of components of our proposed model, we report the results of the ablation study in Table 3. Our model w/o H_y performs significantly worse on subtask C dropping nearly 10%, indicating that the class-label feature H_y is effective. The performance of our model w/o H_l^s also degrades, thus proving class-structure feature H_l^s improves the performance. The performance of our model w/o self-training is slightly worse on subtask B but drops substantially in terms of subtask C, which confirms the effectiveness of the self-training strategy.

3.4 Effect of task weight α and β

To explore the impact of the task weights on the performance of our proposed model, we vary the task weight from 0.1 to 0.9 and show the results in Figure 3. The accuracy and macro-F1 of both tasks tend to increase when the weight of subtask B α increases from 0.1 to 0.5 and drops when α continues to grow from 0.5 to 0.9. The same pattern is observed with the weight of subtask C β. The best performance is achieved when equal weights are assigned to the two losses, or $\alpha = \beta = 0.5$.

Models	Subtask B		Subtask C	
	$ACC.$	$F1.$	$ACC.$	$F1.$
Our model	**62.65**	**60.31**	**54.12**	**44.34**
- *w/o* H_l^s	59.56	58.76	47.50	43.40
- *w/o* H^y	57.98	56.71	45.88	31.01
- *w/o* Self-training	62.20	60.15	46.32	35.65

Table 3: Ablation study results.

4 Conclusion

In this paper, we propose a model to address subtasks B and C of SemEval-2023 Task 10 Explainable Detection of Online Sexism. Our goal is to identify sexist statements and explain why they are sexist. The model treats the two subtasks as one multi-label hierarchical classification problem to capture class features from two perspectives: label texts and class hierarchy, utilizing a pre-trained BERT model and a graph convolution neural network respectively. Our model is proven effective by extensive comparison experiments.

Online sexism detection is getting more and more attention as people begin to realize its importance to the Internet community. More efforts need to be put into this cause to make the Internet world better for everybody.

Acknowledgements

This work was supported by The Major Key Project of PCL(Grant No.PCL2022A03) and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005)

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5810–5820.

Vicent Ahuir, José-Ángel González, and Lluís-Felip Hurtado. 2022. Enhancing sexism identification and categorization in low-data situations. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022*.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5578–5593.

Flor Miriam Plaza del Arco, María Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. 2022. Exploring the use of different linguistic phenomena for sexism identification in social networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022.*

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip S. Yu. 2021. Htcinfomax: A global model for hierarchical text classification via information maximization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3259–3265.

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, NAACL*.

Daniel García-Baena, Miguel Ángel García Cumbreras, Salud María Jiménez Zafra, and Manuel García Vega. 2022. SINAI at EXIST 2022: Exploring data augmentation and machine translation for sexism identification. In *IberLEF@SEPLN*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 249–256.

Wei Huang, Chen Liu, Yihua Zhao, Xinyun Yang, Zhaoming Pan, Zhimin Zhang, and Guiquan Liu. 2021. Hierarchy-aware T5 with path-adaptive mask mechanism for hierarchical text classification. *CoRR*, abs/2109.08585.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured attention networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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. *CoRR*, abs/1907.11692.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, SocialNLP@NAACL 2021, Online, June 10, 2021*, pages 91–102.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A self-training method for machine reading comprehension with soft evidence extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3916–3927.

Amir Soleimani, Vassilina Nikoulina, Benoît Favre, and Salah Ait-Mokhtar. 2022. Zero-shot aspect-based scientific document summarization using self-supervised pre-training. In *Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 49–62.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan L. Yuille, and Fan Yang. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10857–10866.

Yupei Zhang, Md Shahedul Islam Khan, Yaya Zhou, Min Xiao, and Xuequn Shang. 2022. An effective chinese text classification method with contextualized weak supervision for review autograding. In *Intelligent Computing Methodologies - 18th International Conference, ICIC 2022, Xi'an, China, August 7-11, 2022, Proceedings, Part III*, pages 170–182.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1106–1117.