Chride at SemEval-2023 Task 10: Fine-tuned Deberta-V3 on Detection of Online Sexism with Hierarchical Loss

Letian Peng, and Bosung Kim
Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
{lepeng,bosungkim}@ucsd.edu

Abstract

Sexism is one of the most concerning problems in the internet society. By detecting sexist expressions, we can reduce the offense toward females and provide useful information to understand how sexism occurs. Our work focuses on a newly-published dataset, EDOS, which annotates English sexist expressions from Reddit and categorizes their specific types. Our method is to train a DeBERTa-V3 classifier with all three kinds of labels provided by the dataset, including sexist, category, and granular vectors. Our classifier predicts the probability distribution on vector labels and further applies it to represent category and sexist distributions. Our classifier uses its label and finer-grained labels for each classification to calculate the hierarchical loss for optimization. Our experiments and analyses show that using a combination of loss with finer-grained labels generally achieves better performance on sexism detection and categorization. Codes for our implementation can be found at https://github.com/KomeijiForce/SemEval2023_Task10.

1 Introduction

The advent of the internet has drastically changed the way we communicate and interact with one another. It has enabled people from different parts of the world to connect and share their thoughts and ideas on various platforms. However, with the increased usage of the internet, there has been a concerning rise in the prevalence of sexism in online communities and platforms. This issue has been especially rampant towards women and other marginalized groups, with the negative sentiment and abuse contributing to the deterioration of the friendly atmosphere that online groups strive to foster.

While various censorship systems have been implemented to filter out sexist content, they often fail to provide a deeper understanding of the underlying reasons for the sexism. This is where SemEval2023 Shared Task 10 (Kirk et al., 2023) comes in, aiming to categorize online comments into fine-grained vectors of sexism. Doing so, it takes the first step towards exploring the root causes of sexism and developing more effective strategies to combat it.

The labels used in SemEval2023 Shared Task 10 are annotated hierarchically as sexist, category, and vector, providing a nuanced understanding of the different forms of sexism present in online comments. The statistics of the labels are depicted in Figure 1, highlighting the prevalence and diversity of the issue.

In our approach, we recognize the hierarchical nature of the problem and propose a method to improve expression representations by training a classifier with all three hierarchies of labels. We obtain a probability distribution on vector labels from the classifier's output and use this to compute category and sexist probabilities. To optimize the model, we calculate the loss of probability distributions in different hierarchies and aggregate them to obtain the final loss. For encoding, we leverage the power of the DeBERTa-V3 model, which has shown exceptional performance in various natural language processing tasks. Our experiments demonstrate the superiority of our approach over other pre-trained language model encoders in learning sexist text representations. Moreover, we conduct a thorough analysis of the impact of different combinations of loss functions on training results, highlighting the benefits of using finer-grained label annotations.

2 Related Works

Online sexism detection has been actively studied in recent years to face the growing offensive content against women in online spaces. Many tasks and competitions have been introduced to expedite the study of sexism detection, such as SemEval-2019 Task 5 (Basile et al., 2019), IberLEF 2021 EXIST (Rodríguez-Sánchez et al., 2021),

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1670–1675

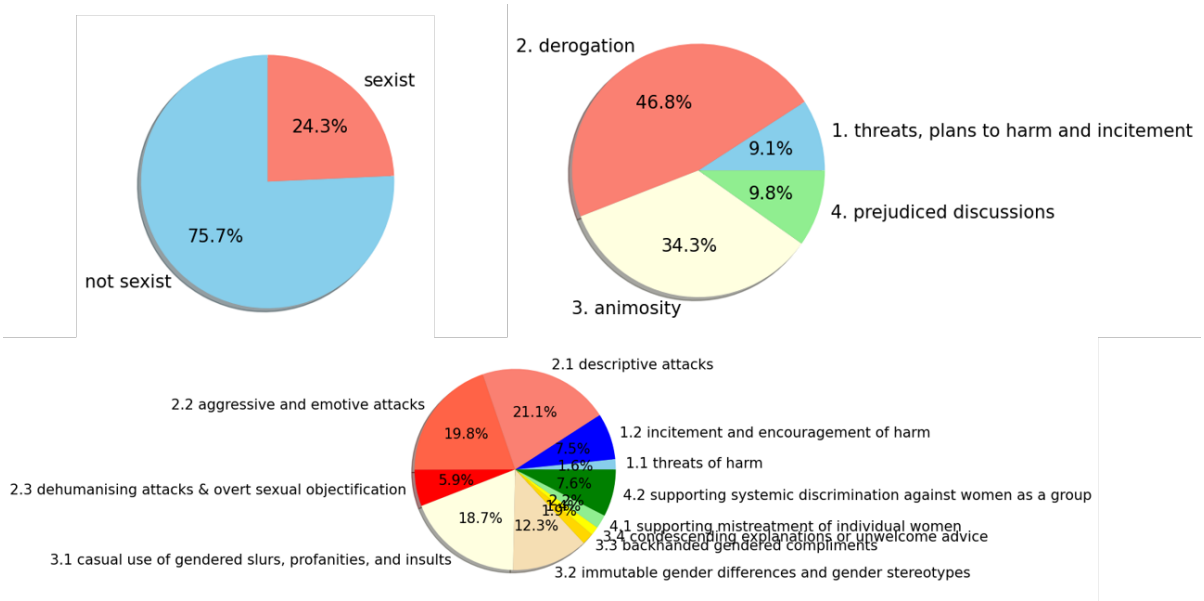

Figure 1: Distributions of labels for Task A, B and C in the EDOS dataset.

and SemEval-2022 Task 5 (Fersini et al., 2022). The datasets have also facilitated the development of automated methods for identifying online sexism in various languages, including English, French, Spanish, and Chinese (Chiril et al., 2020; Rodríguez-Sanchez et al., 2021, 2022; Jiang et al., 2022).

Most existing methods rely on the neural networks models, such as Convolutional Neural Networks (Boriola and Paetzold, 2020), transformer-based large language models (LMs) (Wiedemann et al., 2020; Wang et al., 2020; Davies et al., 2021), and Graph Neural Networks (Wilkens and Ognibene, 2021). Other approaches include data augmentation with back translation (Butt et al., 2021), knowledge distillation (Wang et al., 2020), leverage external resources (García-Baena et al., 2022) and multi-task learning (del Arco et al., 2021). In this work, we use DeBERTa-V3 (He et al., 2021a), which has shown superior performance over other transformer-based LMs. In experiments, we show that DeBERTa-V3 achieved competitive performance with a simple hierarchical loss function compared to the counterpart LMs.

be

3 Task and Dataset

We work on the EDOS sexism detection and explanation dataset[1] (Kirk et al., 2023). Identified by the task, Sexism is "Any abuse or negative sentiment that is directed towards women based on their gender, or based on their gender combined with one or more other identity attributes".[2] Our goal is to detect and explain the sexism in the expressions. Specifically, the dataset contains three subtasks:

- **Task-A: Binary Sexism Detection** is a binary classification task to determine whether an entry is sexist or not.
- **Task-B: Category of Sexism** is a 4-class classification task that categorizes the specific types of sexist. The 4 types are threats, derogation, animosity, and prejudiced discussions.
- **Task-C: Fine-grained Vector of Sexism** is an 11-class classification task that provide more fine-grained classification for sexism. Each of the sexist types in Task-B has subtypes of 2, 3, 4, and 2, respectively.

For the dataset $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$, the format of instances D is (S, L_A, L_B, L_C) where S refers to the sentence and L_N is the label of S for Task-N. L_B, L_C exists only when L_A is *sexism*. Also, L_C is fully dependent on L_B since it is a sublabel of L_B. Notice that Task-B and C only categorize sexist expressions, instances with L_A = *not sexism* will not be involved in training or evaluation.

[1] https://semeval.github.io/SemEval2023/tasks

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

Model	Task A		Task B		Task C	
	Dev	Test	Dev	Test	Dev	Test
$\text{BERT}_{\text{Large}}$ (Devlin et al., 2019)	82.74	82.63	61.16	57.48	36.65	36.17
$\text{RoBERTa}_{\text{Large}}$ (Liu et al., 2019)	84.29	83.58	61.09	57.88	42.28	37.57
$\text{BERTweet}_{\text{Large}}$ (Nguyen et al., 2020)	84.97	84.64	65.20	64.04	**46.90**	**39.22**
$\text{DeBERTa-V3}_{\text{Large}}$ (He et al., 2021a)	**85.48**	**85.67**	**70.39**	**66.36**	45.61	38.25

Table 1: The performances of different pre-trained language models.

4 Methodology

In this section, we present our method for training a classifier to predict labels from three different categories, using a hierarchical loss function. We acknowledge the reviewer's feedback on the need for additional information on the training setting and losses, and we have made revisions accordingly.

Given an input sentence S, our classifier outputs $P_C = P(L_C|S)$, a probability distribution over all categories in L_C and an additional *not-sexist* label. We then calculate the probability distributions for L_B and L_A as follows:

$$P(L_B = l_b|S) = \sum_{l_c \in l_b} P(L_C = l_c|S)$$

$$P(L_A = \textit{not sexist}|S) = P(L_B = \textit{not sexist}|S)$$

$$P(L_A = \textit{sexist}|S) = 1 - P(L_B = \textit{not sexist}|S)$$

Here, $l_c \in l_b$ denotes that l_c is a sub-category of l_b. We then use the cross-entropy loss for all three predictions:

$$\mathcal{L} = \sum_{h \in A,B,C} \text{CELoss}(P_h, L_h)$$

where $\text{CELoss}(P_h, L_h)$ is the cross-entropy loss between the predicted distribution P_h and the true labels L_h for each category $h \in A, B, C$.

For our backbone model, we fine-tune DeBERTa-V3 (He et al., 2021a) on the classification task. DeBERTa-V3 is a pre-trained language model that improves upon the original DeBERTa (He et al., 2021b) by replacing the masked language modeling objective with a replaced token detection task. This change results in state-of-the-art performance across various natural language processing tasks.

Configuration We train our classifier with an AdamW optimizer initialized by a learning rate of 10^{-5}. The batch size is set to 16, and the maximum number of epochs is set to 20. We save the best-performing model on the development set and evaluate its performance on the test set. To predict labels in a hierarchy, we only use the labels and finer-grained labels of interest (e.g., for Task B, we optimize $\mathcal{L}_B + \mathcal{L}_C$, excluding $\mathcal{L}_A$). We note that our reported results in Section 5 differ from those on the leaderboard as we merged the training and development sets and split them by a ratio of $19:1$ to train the model for the submitted version. Here, we report the results achieved with the standard splits of train and development sets. On the leaderboard, our submission includes all losses in lower hierarchies, i.e., $L_A + L_B + L_C$ for Task A, $L_B + L_C$ for Task B, and L_C for Task C.

5 Experiment and Result

5.1 Main Results

The results from our experiments are presented in Table 1. We compare the results of DeBERTa-V3 with other pre-trained LMs including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and BERTweet (Nguyen et al., 2020). BERTweet is trained following the RoBERTa's pre-training procedure, but it uses an additional large corpus, which is about 850M English Tweets. Even though DeBERTa-V3 is pre-trained with less size of datasets than RoBERTa and BERTweet (He et al., 2021a), DeBERTa-V3 achieved best performance in Task A and Task B, outperforming 1.03 and 2.32 points over BERTweet on the test set. However, in Task C, BERTweet shows the best performance over DeBERTa-V3 with a gap of 0.97 points. We provide a detailed analysis in Section 5.2.

5.2 Analysis

Table 2 shows examples and the prediction of each model. We observed that BERTweet performs well even if the input text includes slangs, abbreviations, and hashtags. Many users in social network services (SNS) use variations of curse words to avoid automated censorship (e.g., some SNS users

Example	BERT	RoBERTa	BERTweet	DeBERTa
When a girl gives you s**t test, scare the s**t out of her.	X	O	O	O
Maybe if one of their women were molested... But no, they're all too ugly. #Sweden	X	X	O	O
Can't hit girls though.....but.... a c@#t punt might be an idea..	X	X	O	X
I like how it says "battle for equality" even though it's an all-female coalition.	X	X	X	X

Table 2: Examples and model predictions in Task C. BERTweet performs well even if the input sentence contains slangs or variations of curse words. On the one hand, in the example in the last row, all models fail to correctly predict sarcasm expressions. X/O indicates incorrect/correct prediction.

Loss	Task A		Task B		Task C	
	Dev	Test	Dev	Test	Dev	Test
$\mathcal{L}_A$	85.20	84.18	9.72*	9.50*	3.66*	4.02*
$\mathcal{L}_B$	84.88	85.26	66.33	63.66	9.85*	8.93*
$\mathcal{L}_C$	84.67	85.41	65.46	59.87	**45.61**	**38.25**
$\mathcal{L}_A + \mathcal{L}_B$	**85.82**	**86.10**	47.97	43.87	11.14*	10.63*
$\mathcal{L}_B + \mathcal{L}_C$	84.04	84.12	**70.39**	**66.36**	41.49	35.95
$\mathcal{L}_A + \mathcal{L}_C$	84.45	85.18	46.44	42.44	34.70	31.25
$\mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_C$	85.48	85.67	47.25	45.37	34.46	32.49

Table 3: The performances of DeBERTa-V3$_{\text{Large}}$ with different combinations of loss hierarchies. * : This configuration does not contain the label annotation or any finer-grained annotation for this task.

replace the alphabet *a* with the special character @). BERTweet shows its strength in handling these types of variation since it is specialized on SNS texts. However, without fine-tuning on additional datasets, DeBERTa-V3 shows competitive performance overall. On the one hand, we also found that all models fail to correctly predict sarcasm patterns. In the example in the last row in Table 2, it is labeled as L_A=*sexist*, L_B=*prejudiced discussions*, and L_C=*supporting systemic discrimination against women as a group*. However, the sentence is indirectly expressing hostility even using "I like ~". In this case, most models still struggle to predict that it contains a negative intention.

5.3 Hierarchical Loss Contribution

Table 3 shows the performances of DeBERTa-V3 according to the combination of loss hierarchies. We observed that using the loss of similar tasks performs better rather than using whole loss terms. For example, in the results on Task A, the model trained with $\mathcal{L}_A$ and $\mathcal{L}_B$ slightly outperforms than the result with $\mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_C$. Similarly, using $\mathcal{L}_B$ and $\mathcal{L}_C$ shows significant improvements over other results in Task B. However, we found that $\mathcal{L}_A$ and $\mathcal{L}_C$ are not compatible and conjecture this is due to the gap of class granularity between tasks, thus the model hardly benefits from other tasks.

6 Conclusion

In this paper, we study the detection of online sexism and the categorization of the underlying reasons for sexism. We train a DeBERTa-V3 classifier to learn fine-grained representations of sexism in online expression. Results from our experiments show a combination of labels with finer-grained labels generally improves classification performance.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Aurélio Hermogenes Boriola and Gustavo Henrique Paetzold. 2020. UTFPR at SemEval 2020 task 12: Identifying offensive tweets with lightweight ensembles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2232–2236, Barcelona (online). International Committee for Computational Linguistics.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F. Gelbukh. 2021. Sexism identification using BERT and data augmentation - EXIST2021. In

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, pages 381–389. CEUR-WS.org.

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 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1397–1403. European Language Resources Association.

Lily Davies, Marta Baldracchi, Carlo Alessandro Borella, and Konstantinos Perifanos. 2021. Transformer ensembles for sexism detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 390–394. CEUR-WS.org.

Flor Miriam Plaza del Arco, M. Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín-Valdivia. 2021. Sexism identification in social networks using a multi-task learning system. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 491–499. CEUR-WS.org.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

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 *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*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

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 (SemEval-2023)*. 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.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Francisco J. Rodríguez-Sanchez, Jorge Carrillo-de-Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of EXIST 2021: sexism identification in social networks. *Proces. del Leng. Natural*, 67:195–207.

Francisco J. Rodríguez-Sanchez, Jorge Carrillo-de-Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of EXIST 2022: sexism identification in social networks. *Proces. del Leng. Natural*, 69:229–240.

Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, and Yu Sun. 2020. Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1448–1455, Barcelona (online). International Committee for Computational Linguistics.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. Uhh-lt at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. *arXiv preprint arXiv:2004.11493*.

Rodrigo Souza Wilkens and Dimitri Ognibene. 2021. Mb-courage @ EXIST: GCN classification for sexism identification in social networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 420–430. CEUR-WS.org.