iREL at SemEval-2023 Task 10: Multi-level Training for Explainable Detection of Online Sexism

Nirmal Manoj, Sagar Joshi, Ankita Maity, Vasudeva Varma
IIIT Hyderabad, India
{nirmal.manoj, sagar.joshi, ankita.maity}@research.iiit.ac.in
vv@iiit.ac.in

Abstract

This paper describes our approach for SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS). The task deals with identification and categorization of sexist content into fine-grained categories for explainability in sexism classification. The explainable categorization is proposed through a set of three hierarchical tasks that constitute a taxonomy of sexist content, each task being more granular than the former for categorization of the content. Our team (iREL) participated in all three hierarchical subtasks. Considering the inter-connected task structure, we study multilevel training to study the transfer learning from coarser to finer tasks. Our experiments based on pretrained transformer architectures also make use of additional strategies such as domain-adaptive pretraining to adapt our models to the nature of the content dealt with, and use of the focal loss objective for handling class imbalances. Our best-performing systems on the three tasks achieve macro-F1 scores of 85.93, 69.96 and 54.62 on their respective validation sets.

1 Introduction

The proliferation of online sexism targeted towards women not only makes online spaces inhospitable but also perpetuates existing social inequities. This has lead to the use of online systems for the identification of such content through automated technologies based on NLP methods. However, a blatant categorization into general, high-level categories as is often performed by such technologies leads to lack of explainability and trust in these systems. The nature of sexist content being varied based on the motivation of the writer, a discrete identification of categories as different as threats and prejudiced expressions is beneficial for differential treatment. A granular study also helps in identifying the weaknesses of the model in identifying certain specific categories of sexist content.

In our work, we work on the dataset for Explainable Detection of Online Sexism (introduced as a part of the SemEval-2023 Task 10) (Kirk et al., 2023) based on English language posts from the social media platforms Reddit[1] and Gab[2]. The dataset, discussed in a greater detail in Section 3, introduces three inter-connected subtasks (labeled A, B, C) for the problem, each dealing with classification of sexist comments in a more fine-grained sense. Task A consists of a simple binary classification of posts into sexist and non-sexist categories. For posts which are detected as sexist, Task B deals with their categorization into 4 distinct categories of sexism. Finally, in Task C looks at a more granular classification of the sexist posts into 11 classes, or "fine-grained vectors of sexism". We worked on all the three subtasks, and report our experimental findings on the same.

To benefit from the inter-connected, hierarchical nature of the three subtasks forming a taxonomy of sexist content labeling, we primarily study the technique of multi-level training in this paper. This technique, explained in Section 4.4, serves to transfer learn on a task at the lower level from the learnings on a higher level task. We experiment with five different pretrained transformer (Vaswani et al., 2017) architectures for our approaches. Besides cross-entropy based vanilla finetuning (Section 4.1), we employ domain-adaptive pretraining (refer to Section 4.3) (Gururangan et al., 2020) as a precursor to multi-level training for adaptation of the models to the nature of the data in this domain. Considering the class imbalances in Tasks B and C, we also make use of the focal loss (Lin et al., 2017) objective, elaborated in Section 4.2. We specify the experimental settings for these methods in Section 5 and provide the results in Section 6. We show the suitability of domain-adaptive pretraining on Task A, multi-level training on Task B and focal loss on

[1] https://www.reddit.com/
[2] https://gab.com/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1691–1696

Task C as the best performing approaches for the tasks. The best results (on the basis of macro-F1) were obtained on the pretrained RoBERTa-large (Zhuang et al., 2021) model, the results for which we show on the validation split of the provided data. We open source our code for replicability of our methods[3].

2 Related Work

Abburi et al. (2021) employ self-training to augment the set of labelled instances by selectively adding unlabeled samples for multi-label sexism classification. In their work, the focus is on self-reported accounts of sexism rather than the classification of sexist content per se. The paper delves into a comprehensive and fine-grained categorization of sexism, encompassing a total of 23 categories in a multi-label setting. This classification scheme is organized into a 3-level hierarchical taxonomy, wherein fine-grained labels are systematically grouped together. The taxonomy comprises 8 categories at level 1, 15 categories at level 2, and 23 categories at level 3. This structured approach facilitates a more nuanced understanding of sexism and its various manifestations, enabling more effective detection and analysis of sexist content.

ElSherief et al. (2017) show increased engagement with gender based violence related tweets in comparison to other non gender based violence tweets.

Zhang and Luo (2018) propose deep neural network structures serving as feature extractors that are particularly effective for capturing the semantics of hate speech including racism and sexism. Frenda et al. (2018) present an approach for detecting sexism and misogyny from Spanish and English tweets. Schrading et al. (2015) design classifiers to detect text discussing domestic abuse on Reddit.

Karlekar and Bansal (2018) investigate an RNN-CNN model for categorizing personal experiences of sexual harassment into one or more of three classes. In Anzovino et al. (2018), tweets identified as misogynist are classified as discredit, stereotype and objectification, dominance, or derailing, sexual harassment, and threats of violence using features involving n-grams, part of speech (POS) tags and text embedding.

These prior works lack the study from a taxonomical perspective for categories on online sexist content. In our work, we fill the gap by proposing methods for dealing with hierarchical categorization of sexism.

3 Data

In addition to the primary labeled dataset for the task, the organizers also provided two additional unlabeled datasets, each containing 2M comments from Reddit and Gab. The primary dataset consists of 14k / 2k / 4k labeled instances each in the train / dev / test splits. Of the 14k instances in the train split, ∼3.4k comments are labeled under the "sexist" category for task A, hence limiting the amount of usable data for tasks B and C to this subset, which are further categorized into 4 and 11 classes respectively. Although the classes for task A do not have a significant imbalance (3:1 distribution for "non-sexist" versus "sexist"), the classes for tasks B and C have a significant imbalances.

We make use of the labeled dataset as well as the unlabeled Reddit and Gab datasets for our experiments. No other additional data was made use of. As a part of data exploration, we measured the average lengths of the comments in each dataset on the basis of white space-separated tokens. Average length of comments in the given labeled dataset is 23.42 tokens, slightly above that for Gab (16.17 tokens) and Reddit (17.89 tokens) unlabeled datasets. We also noticed that the length distributions of the comments from all the three data sources to conform when plotted on a graph. This observation was useful in merging the datasets for applying domain-adaptive pretraining, discussed in Section 4.3.

4 Methodology

In this section, we describe the core components of our modeling we experimented with to tackle the three subtasks. We start with basic finetuning in Section 4.1, explain the focal loss objective in Section 4.2, domain-adaptive pretraining in Section 4.3, and multi-level training in Section 4.4.

4.1 Basic Finetuning

The first component of our approach is Basic Fine Tuning, which is a widely used technique for adapting pretrained transformer language models to specific tasks. This approach is based on the idea that the pretraining of the transformer model on large amounts of unlabeled data provides a good generalization of natural language, which can be

[3] https://github.com/NirmalManoj/EDOS-iREL-MultiLevel-Training

further fine-tuned on the specific task of interest. Fine-tuning the pretrained model on a specific task allows the model to adapt to the specific characteristics of the task, such as the vocabulary, syntax, and semantics.

4.2 Focal Loss

In Task B and C, we observed a class imbalance in the fine-grained classification tasks. Specifically, some classes had significantly fewer examples than others, which can lead to bias in the model's predictions towards the majority class. To address this issue, we employed the Focal Loss as a loss function for some experiments.

The Focal Loss is a modified version of the cross-entropy loss that downweights easy examples and focuses on hard examples. It was introduced by Lin et al. (2017) for object detection tasks in computer vision, but has since been applied to various natural language processing tasks, including sentiment analysis and named entity recognition.

By using the Focal Loss, we aim to give more weight to hard examples, which are examples that the model is currently misclassifying, while downweighting easy examples that the model is already accurately classifying. This helps to address the problem of class imbalance by making the model focus more on the minority classes, which are typically the harder examples.

4.3 Domain-Adaptive Pretraining

The task of detecting and classifying sexism in online comments requires a nuanced understanding of language and context. To better adapt our models to this domain, we employed Domain-Adaptive Pretraining (Gururangan et al., 2020), a technique that involves pretraining our models on data from a similar domain or task before fine-tuning on the target task. By doing so, we aimed to improve the performance of our models, especially on the fine-grained classification tasks (Task B and C), which require a more nuanced understanding of sexist language.

In particular, we utilized the 2 million unlabelled data points collected from Reddit and Gab, which were provided by the task organizers. We used this data to pretrain several transformer language models. Our motivation for using this approach is based on the observation that the language patterns and structures used in social media or online forums are different from those used in more formal contexts such as news articles or books, which are part of the corpus used to create the pretrained LMs. Further Pre Training on data from social media or online forums can help our models learn to identify the specific linguistic cues that are indicative of sexism in this domain, such as informal language, sarcasm, and emoticons.

By leveraging Domain Adaptive Pretraining, we aimed to improve the generalizability of our models to the task and enhance their ability to detect and classify sexism in online comments.

4.4 Multi-level Training

Each subtask requires a progressively more fine-grained hierarchical classification of sexist content. We found that the previous tasks could serve as a good starting point for the current task, as they provide the model with valuable knowledge and representations that can be leveraged to improve its performance on the current task.

To this end, we employed Multi-Level Training, a technique that uses the best checkpoint of a previous task as the starting point for the current task. By using this technique, we aim to capitalize on the knowledge and representations learned from previous tasks and transfer them to the current task. For example, we use the best checkpoint from Task A to initialize the training for Task B, and the best checkpoint from Task B to initialize the training for Task C.

We believe that Multi-Level Training is particularly useful for Tasks B and C since each subtask builds upon the previous one, with the final subtask requiring the most fine-grained classification. By initializing the current task with the best checkpoint from the previous task, our model can leverage the knowledge learned from previous tasks and improve its performance on the current task, especially on the more challenging subtasks.

5 Experimental Details

5.1 Modeling

All our experimental strategies are tested on a set of 5 pretrained transformer encoders. These include a base, uncased version of the BERT (Devlin et al., 2019) model (BBU), base and large versions of the RoBERTa (Zhuang et al., 2021) model (RB, RL), two versions of the DeBERTa (He et al., 2021) base model: the original base model (DB) and the v3 version (DBV3) (He et al., 2023), which was an improved version of the same. Modeling the classification task for each experiment involved ad-

dition of a linear layer with the output size equal to the number of classes for that task on the pooled embeddings of the last hidden layer of each model. For the pretraining experiment, a linear layer was added for the Masked Language Modeling (MLM) objective as in (Devlin et al., 2019). 15% of the tokens were masked randomly for MLM pretraining.

Transferring the model from one task to the other involved replacement of the final linear layer with a new layer suitable for that task. All the layers of the model were then trained, with the model encoder parameters adapting to the new task from their previous checkpoint, and the linear layer being trained on the task-specific objective from scratch.

5.2 Training Hyperparameters

For training, we employed a fine-tuning approach using the Slanted Triangular Learning Rates (STLR) technique with a 10% warmup. We varied the values of the maximum learning rate in this scheduling to $2e-6, 5e-6$ and found the value of $5e-6$ to work well across experiments. AdamW optimizer (Loshchilov and Hutter, 2017) was used with a weight decay of 0.01. In the experiments using focal loss, the gamma value of the hyperparameter was set to 2.0. The training was carried out over a course of 10 epochs using a batch size of 4 and accumulated gradient steps of 2. We relied on the Huggingface (Wolf et al., 2020) library for pretrained encoder models and training support.

The aforementioned settings were used across all classification experiments and found them to work well. In domain-adaptive pre-training, we trained the model on 2 epochs over the complete data of $\sim$2M samples. Rest of the hyperparameters were kept the same as in the classification experiments.

Best checkpoint was determined on the basis of the performance on the validation (dev) dataset using the macro-F1 score, the primary evaluation metric for the task. Besides macro-F1, we also measure the precision and recall values calculated in macro mode along with accuracy.

6 Results

The experimental results for our approaches for Task A, B and C can be seen in Table 1, Table 2 and Table 3 respectively. The values of F1, Precision and Recall metrics are noted in terms of their macro versions, as noted in Section 5.2.

In Task A, we observed that all the models, except DB, performed better after the domain-adaptive pretraining step. Among all the experiments, RL with pretraining gave the best F1 score of 85.93.

In Task B, we explored the usage of focal loss in basic finetuning, which showed improved performance in the best performing model, RL. However, we observed that the improvement from focal loss to not be consistent when compared to cross entropy. Pretraining followed by fine-tuning with focal loss showed improvement only for BBU. Multi-level training with pretraining gave the best model, which again was RL with a macro-F1 score of 69.96. Interestingly, BBU showed the highest performance improvement with continued pretraining and multi-level training where the score increased by 5.42 points from the second-best performing setup.

In Task C, we observed that the performance improvement from focal loss was more pronounced due to the high class imbalance of 11 categories. In basic finetuning, changing the objective from cross entropy to focal loss improved the performance of each model by at least 3 points in the macro-F1 score. The best performing model in basic fine-tuning was RL, which showed an improvement of more than 4.27 points in macro-F1. The model that saw the highest improvement was DBV3 with a difference of 6.25 points in macro-F1. We also noted that continued pretraining gave improved performance for all models except RL. Multi-level training with Task A -> Task B -> finetuning with focal loss gave the best performance for BBU and RB. However, overall the best performing model remained RL in the basic finetuning with focal loss setup, giving a macro-F1 score of 54.62.

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning	BBU	81.62	81.97	82.47	86.80
	RB	83.22	83.94	82.58	87.90
	DB	83.69	83.69	83.69	88.00
	DBV3	82.51	82.69	82.32	87.20
	RL	85.73	**87.11**	84.56	89.85
Pretraining → finetuning	BBU	81.93	83.02	81.00	87.10
	RB	83.86	84.56	83.22	88.35
	DB	82.40	85.02	80.52	87.85
	DBV3	83.93	84.15	83.72	88.25
	RL	**85.93**	86.90	**85.09**	**89.90**

Table 1: Performance of our models on the validation set of Task A.

7 Conclusion

In our work, we present a study of our approaches to tackle the problem of explainable detection of online sexism as a set of 3 inter-connected, hierar-

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning: CE	BBU	58.91	62.39	56.68	62.14
	RB	63.10	64.84	63.85	65.23
	DB	68.33	67.07	69.90	69.34
	DBV3	61.43	61.71	62.13	63.79
	RL	69.41	69.19	69.71	70.99
Basic finetuning: FOCAL	BBU	58.60	60.43	57.21	62.14
	RB	64.02	65.36	62.90	64.81
	DB	67.79	67.53	68.21	69.14
	DBV3	62.72	62.03	63.86	62.55
	RL	69.67	69.11	71.47	**71.60**
Pretraining → finetuning: FOCAL	BBU	60.91	59.97	62.00	61.32
	RB	57.30	57.17	57.49	57.61
	DB	62.13	60.33	64.80	62.55
	DBV3	56.98	56.16	58.69	56.17
	RL	69.57	67.24	**73.39**	69.55
Pretraining → Task A → finetuning: FOCAL	BBU	66.33	65.94	67.28	67.28
	RB	62.87	64.92	61.36	62.96
	DB	67.05	65.60	68.87	68.93
	DBV3	62.53	64.19	61.21	64.20
	RL	**69.96**	**69.24**	70.98	70.99

Table 2: Performance of our models on the validation set of Task B.

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning: CE	BBU	33.07	33.51	34.43	54.12
	RB	38.48	37.91	39.57	55.76
	DB	44.90	54.68	43.10	56.58
	DBV3	34.97	35.68	36.09	55.14
	RL	50.35	52.34	49.31	61.32
Basic finetuning: FOCAL	BBU	36.22	34.97	39.27	47.33
	RB	43.05	42.48	44.34	50.21
	DB	49.79	51.25	49.17	55.35
	DBV3	41.23	40.28	43.95	49.59
	RL	**54.62**	**56.33**	**53.79**	**61.32**
Pretraining → finetuning: FOCAL	BBU	37.32	41.34	38.19	47.33
	RB	39.83	38.87	42.07	47.33
	DB	51.16	51.90	52.19	58.23
	DBV3	41.38	40.21	43.67	50.62
	RL	51.67	55.95	50.97	60.08
Pretraining → Task A → Task B → finetuning: FOCAL	BBU	40.68	48.61	41.04	51.44
	RB	43.15	44.22	43.00	53.50
	DB	49.97	51.04	49.77	57.41
	DBV3	39.68	42.80	39.64	51.23
	RL	51.15	51.98	50.80	59.67

Table 3: Performance of our models on the validation set of Task C.

chical classification tasks. Our approaches involve domain-adaptive pretraining for adapting the classification models to the downstream data, implementation of the focal loss objective for handling the class imbalances and a multi-level training strategy for studying the benefit of learning from a coarser to finer classification task. On all our experiments involving 5 different pretrained transformer models, we observe the benefit of more trainable parameters and their suitability to small datasets with the noticeably high performance of the RoBERTa-large model over the rest across different training settings and tasks. We show a general improvement from domain-adaptive pretraining due to a better alignment of the pretrained models towards the downstream task and also showcase the benefit of the focal loss for very large imbalances in the data, such as in Task C. The multi-level training turns out to be the best performing approach for Task B, benefiting from the learnings gained from Task A.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2021. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Science and Engineering*, 6(4):359–379.

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

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.

Mai ElSherief, Elizabeth Belding, and Dana Nguyen. 2017. notokay: Understanding gender-based violence in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):52–61.

Simona Frenda, Bilal Ghanem, and Manuel Montes y Gómez. 2018. Exploration of misogyny in spanish and english tweets. In *IberEval@SEPLN*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

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

Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, pages 2805–2811, Brussels, Belgium. 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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, Lisbon, Portugal. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, Accepted.

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.