SSS at SemEval-2023 Task 10: Explainable Detection of Online Sexism using Majority Voted Fine-Tuned Transformers

Sriya Rallabandi   **Sanchit Singhal**   **Pratinav Seth**

Manipal Institute of Technology

Manipal Academy of Higher Education, Manipal, India

{sriyarallabandi,sanchitsinghal,seth.pratinav}@gmail.com

Abstract

This paper describes our submission to Task 10 at SemEval 2023-Explainable Detection of Online Sexism (EDOS) Kirk et al. (2023), divided into three subtasks. The recent rise in social media platforms has seen an increase in disproportionate levels of sexism experienced by women on social media platforms. This has made detecting and explaining online sexist content more important than ever to make social media safer and more accessible for women. Our approach consists of experimenting and finetuning BERT-based models and using a Majority Voting ensemble model that outperforms individual baseline model scores. Our system achieves a macro F1 score of 0.8392 for Task A, 0.6092 for Task B and 0.4319 for Task C and the code is available on GitHub.[1]

1 Introduction

Social media has become a free and powerful tool for communication, allowing people to share their ideas and connect with others worldwide. For women, social media has provided a platform for bringing visibility, empowering, and educating people on women's issues. However, with an increase in the number of users, there is a significant increase in sexism on these platforms has led to online harassment and disinformation, making social media less accessible and unwelcoming, continuing to uphold unequal power dynamics and unfair treatment within society. Social media's anonymity and ease of access have made it easier for individuals to spread sexist content online. This has led to a growing interest and needs to detect and identify online content as sexist and explain why that content is sexist. Natural Language Processing (NLP) techniques have shown promise in enabling the automated detection of sexist language and providing insights into its nature of it.

This paper describes our contributions to SemEval 2023 Task 10:Explainable Detection of Online Sexism Kirk et al. (2023), which proposed a new dataset of sexist content from Gab and Reddit to detect and explain online sexist content through three hierarchical classification subtasks.

Task 10 of SemEval 2023 includes three hierarchical subtasks: Task A, which is a binary classification task to classify whether a post is sexist or not sexist, Task B, a multi-class classification task for sexist posts which further classifies them into four categories and Task C, another multi-class classification task to classify the sexist posts into one of eleven fine grained vectors. We conducted detailed experiments on BERT-based models, such as RoBERTa Liu et al. (2019) and DeBERTaV3 V3 He et al. (2021), using different training methods and loss functions to reduce class imbalance along with hyperparameter tuning. We then ensembled these individual models using a majority voting ensemble and show how these methods affect the model's performance and how an ensemble of these models performs better than the individual baselines.

2 Background

2.1 Problem and Data Description

SemEval 2023 Task 10: Explainable Detection of Online Sexism aims to detect and flag any abuse or negative sentiment towards women based on their gender or based on their gender combined with one or more other identity attributes. The dataset consists of 20,000 labelled entries. Of these, 10,000 are sampled from Gab and 10,000 from Reddit. Initially, three skilled annotators assign labels to all entries, and any discrepancies are resolved by one of the two experts. In Task A, if all annotators unanimously agree on a label, it is considered the final label. However, if there is any disagreement, an expert will review the entry and determine the final label. For Task B and C, if at least two anno-

---

[1] https://github.com/sriya26/SemEval

tators agree on a label, that is considered the final label. However, an expert will determine the label if there is a three-way disagreement. The challenge is divided into three hierarchical subtasks:

Subtask A: Binary Sexism Classification- The task is a two-class (binary) classification that detects whether a post is sexist. The training set consists of 14000 (70% split) entries, of which 3398 are sexist, as shown in Table 1.

Post	Label
yeah but in order to keep the benefit i have to be good tommorow because i told her we could try	not sexist
As Roosh said having discussions with woman is a waste of time anyway	sexist

Table 1: Examples for Task A

Subtask B: Category of Sexism- The task is a multi-class classification where we have to predict the sexist class into four categories (1) threats, (2) derogation, (3) animosity, (4) prejudiced discus-

Post	Label
Shove PORK down her THROAT!	1.threats,plans to harm and incitement
Classic toxic white females	2.derogation
Rule one: No woman can be incel, only volcel.	3. animosity
English: robot know best person for job. Robot no choose woman. Robot choose man.	4. prejudiced discussions

Table 2: Examples for Task B categories

sions, shown above in Table 2.

Subtask C: Fine-grained Vector of Sexism- The task is a multi-class classification where the sexist posts are further subdivided into 11 fine-grained Vector of Sexism: (1.1) Threats of harm and Incitement and (1.2) encouragement of harm are included in the Category of threats ; (2.1) Descriptive attacks, (2.2) Aggressive and emotive attacks and (2.3) Dehumanising attacks and overt sexual objectification in Derogation ; (3.1) Casual use of gendered slurs, profanities and insults, (3.2) Immutable gender differences and gender stereotypes,(3.3) Backhanded gendered compliments and (3.4) Condescending explanations or unwelcome
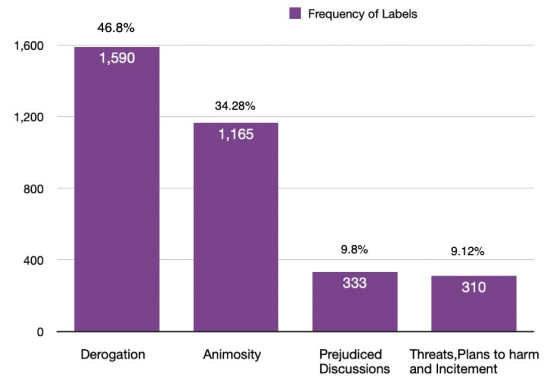

Figure 1: Frequencies of Task B labels in training set

advice in Animosity; (4.1) Supporting mistreatment of individual women and (4.2) Supporting systemic discrimination against women as a group in Prejudiced discussions, as shown in Table 3.
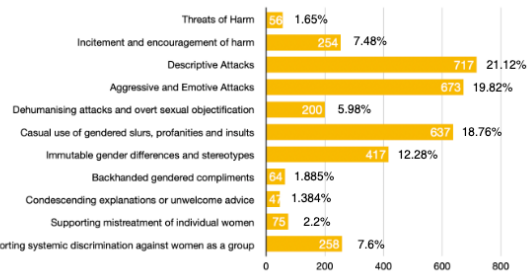

Figure 2: Frequencies of Task C labels in training set

2.2 Related Work

2.2.1 Onine Sexism

Online Sexism is considered a type of hate speech by many and can be found on several big platforms like Twitter and Reddit. This increases the importance of improving detection and classification accuracy for the same. Major work has been done in Harmful Speech Detection in previous SemEval tasks like Hate news detection Basile et al. (2019); Propaganda Detection Da San Martino et al. (2020); and Rumor Gorrell et al. (2019) Detection. Stanczak and Augenstein (2021) conducted a survey

Fersini et al. (2022) aimed to detect misogynous memes on the web by taking advantage of available texts and images, and Mina et al. (2021) and Rodríguez-Sánchez et al. (2020) aims to automatically identify sexism in social media content by applying machine learning methods.

Post	Label
Shove PORK down her THROAT!	1.2 incitement and encouragement of harm
Classic toxic white females	2.1 descriptive attacks
Rule one: No woman can be incel, only volcel.	3.2 immutable gender differences and gender stereotypes
English: robot know best person for job. Robot no choose woman. Robot choose man.	4.2 supporting systemic discrimination against women as a group

Table 3: Examples for Task C vectors

2.2.2 Methods for detecting Online Sexism

de Paula and da Silva (2022) explored the use of different transformers architectures for two tasks in English and Spanish: sexism detection and sexism classification. Parikh et al. (2019) introduced the first work on multi-label classification for sexism detection and also provided the largest dataset on sexism categorization. They built a BERT-based neural architecture with distributional and word-level embeddings to perform the classification task. Abburi et al. (2021) proposed a neural model for this sexism detection and classification that can combine representations obtained using RoBERTa model and linguistic features such as Empath, Hurtlex, and PerspectiveAPI by involving recurrent components. Agrawal and Mamidi (2022) and Rao (2022) performed binary and multi-class classification using transformer-based models and showed how ensembling them outperforms the baseline model performances.

3 System Overview

After conducting extensive experiments on pre-trained transformer-based models such as BERT Devlin et al. (2018), RoBERTa Liu et al. (2019), DeBERTa He et al. (2020),DeBERTaV3 He et al. (2021), AlBERTa Lan et al. (2019) and XL-Net Yang et al. (2019) and comparing their baseline per-

formances based on accuracy and Macro F1 score, we finally used DeBERTaV3 for Task A and both RoBERTa and DeBERTaV3 for Task B and C.

3.1 Ensemble

To increase the overall accuracy of the predictions and robustness of the predictive model, models were first individually tuned on the entire dataset. Then, their predictions were combined using ensembling methods. Final results were obtained using the Majority Voting Ensembling method, a hard voting method in which the class predicted by most models is considered the final output and weighted average ensemble, in which the weights are obtained by grid search or other optimization techniques on the validation dataset.

3.2 Task A: Binary Classification

This consisted of classifying texts into binary classes: sexist and not sexist. The data was trained on the pre-trained BERT-based models, of which DeBERTaV3 gave the best baseline score. We experimented with the DeBERTaV3 model with tuning parameters such as different learning rates and number of epochs to test their effect on the model's accuracy. We also used Focal Loss Lin et al. (2017) and Class Weights to deal with class imbalance in the dataset. We finally used the Majority Voting ensembling method on the top 3 model combination results to obtain the final labels.

3.2.1 Class Imbalance

We noticed a major class imbalance in the training dataset of Task A, with 10602, not sexist examples and only 3398 sexist examples. Hence to deal with this, we experimented with class imbalance handling measures such as Focal Loss and Class Weights assigned to each class to reduce class imbalance.

Focal Loss: Focal Loss is primarily used in classification tasks to deal with class imbalance.

$$L_{FL} = \alpha(1-p)^{\gamma}\log(p) \qquad (1)$$

where γ and α are the hyperparameters.

It is an improvement over the Cross-Entropy loss function.Zhang and Sabuncu (2018) and deals with class imbalance by assigning more weights to hard or easily misclassified examples and down weight easy examples, thereby emphasizing correcting misclassified ones.$\gamma>1$ will rescale the modulating factor such that the easy examples are down-weighted more than the hard ones, reducing

their impact on the loss function.

3.3 Task B and C: Multi-Label Classification

These tasks were a multi-label classification problem. We had to predict the category label in task B and the vector label in task C of the text already classified as sexist in task A. We experimented with parameters like learning rates and the number of epochs. We used Focal Loss and Class Weights for any class imbalance. Apart from these methods, we also trained the baseline DeBERTaV3 and RoBERTa models using a 5-fold cross-validation method and compared the effect of the training method on accuracy. The best-performing models were ensembled using the Majority Voting ensemble method to get the final labels.

4 Experimental Setup

The training data consists of 14,000 entries out of 20,000 total entries, and these are used for Task A. 3,398 of these 14,000 are labelled as sexist, which are further used in Tasks B and C. We strictly used the data provided by the organisers to train and test our models in the development and testing phases. We used 80% of the training dataset to train the models, while 20% remaining was used for cross-validation. Upon using both methods, we found that the majority voting ensemble method gives better results than the weighted average ensemble method. Using the majority voting ensemble reduced individual model errors, and the impact of outliers was also minimized, leading to a more accurate and robust system.

For Task A, we used our best three performing models for the ensemble, including the baseline DeBERTaV3 model, with focal loss and different hyperparameter values. The exact hypermeters used for these three models are mentioned in Table 4.

Parameters	Model 1	Model 2	Model 3
Learning Rate	2e-5	2e-5	4e-5
Epochs	6	6	4
Batch Size	6	6	6
Loss Function	None	Focal	Focal
Optimizer	None	Adam	Adam

Table 4: Hyperparameters for Task A DeBERTaV3 models

Model 1 is the baseline DeBERTaV3 model with a learning rate of 2e-5 and batch size and epochs of

6. Model 2 is the baseline as Model 1 but with Focal loss and Adam optimizer while Model 3 has a learning rate of 4e-5 and runs on 4 epochs instead. α and γ are the hyper-parameters of Focal Loss, where α is a weighting factor that gives more importance to the minority class, and γ is a tunable parameter that modulates the degree of downweighting. The values we take after experimenting are 1.0 for α and 2.0 for γ, the default value used in the original paper. We also made use of Class Weights along with Focal Loss, which are computed for each class using the formula,

$$\frac{n_{samples}}{n_{classes} \times np.bincount(y)} \quad (2)$$

However upon experimenting, we observed that the results obtained were lesser compared to baseline models, hence we did not include it in the final ensemble.

In Task B and C, after much experimenting, we noticed that an ensemble of RoBERTa and DeBERTaV3 models with 5-fold cross-validation gave a better result than the baseline models. The learning rate for both was taken as 2e-5, with 4 epochs in each fold.

Macro F1 score was the primary evaluation metric used. We used the pretrained models available in huggingface library and implemented the models using ktrain library Maiya (2020), a lightweight wrapper for deep learning libraries.

5 Results

Results for Task A, B and C are mentioned in Tables 6 and 5 as F1 scores. We have presented our results from the Development phase in Table 6 and the final results of the ensemble from the Testing phase in Table 5. We ranked 34th out of 84 for Task A, 36th out of 69 for Task B and 30th out of 63 for Task C.

Task	F1 Score
Task A	0.8392
Task B	0.6092
Task C	0.4319

Table 5: Results on Test Phase dataset with Ensemble

While using focal loss doesn't improve the F1 score compared to the baseline DeBERTaV3 model in Task A, ensembling both the models along with DeBERTaV3 with a learning rate of 4e-5 improves the overall performance of the model.

Task	Model	F1 Score
Task A	DeBERTaV3	0.8348
	DeBERTaV3 with focal loss	0.8346
	DeBERTaV3 with lr=4e-5	0.8345
	Ensemble	**0.8411**
Task B	DeBERTaV3	0.6381
	RoBERTa	0.6376
	DeBERTaV3 with 5-fold CV	0.6522
	RoBERTa with 5-fold CV	0.6490
	Ensemble	**0.6592**
Task C	DeBERTaV3	0.4285
	RoBERTa	0.4263
	DeBERTaV3 with 5-fold CV	0.4336
	RoBERTa with 5-fold CV	0.4327
	Ensemble	**0.4469**

Table 6: Results for Development phase

For Task B and C, Focal Loss models performed poorly compared to the baselines despite having some imbalanced classes. We also noticed that using a different training method, like 5-fold Cross Validation improved the performance over the baseline models for multi-class classification and ensembling them further improved the performance.

6 Conclusion and Future Works

In this work, we benchmarked various pre-trained Transformer based models such as BERT, RoBERTa and DeBERTaV3 and their majority voting Ensemble Models. This detection and classification of online sexism were done under Task 10 at SemEval 2023-Explainable Detection of Online Sexism (EDOS). We also used Focal loss to deal with class imbalance.DeBERTaV3 gave the best F1 score of 0.8348 on Task A, 0.6381 on Task B, and 0.6381 on Task C. RoBERTa gave results close to DeBERTaV3 on both Task B and C.

In future works, we would like to **(1)** compare the results of the models on different datasets and languages; **(2)** use unsupervised learning to improve the results of our models;**(3)** propose a novel custom architecture which provides more robust predictions.

References

Harika Abburi, Shradha Sehgal, Himanshu Maheshwari, and Vasudeva Varma. 2021. Knowledge-based neural framework for sexism detection and classification. In *IberLEF@ SEPLN*, pages 402–414.

Samyak Agrawal and Radhika Mamidi. 2022. LastResort at SemEval-2022 task 4: Towards patronizing and condescending language detection using pre-trained transformer based models ensembles. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 352–356, Seattle, United States. Association for Computational Linguistics.

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.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Angel Felipe Magnossão de Paula and Roberto Fray da Silva. 2022. Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *CoRR*, abs/1708.02002.

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.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Schütz Mina, Boeck Jaqueline, Liakhovets Daria, Slijepčević Djordje, Kirchknopf Armin, Hecht Manuel, Bogensperger Johannes, Schlarb Sven, Schindler Alexander, and Zeppelzauer Matthias. 2021. Automatic sexism detection with multilingual transformer models. *arXiv preprint arXiv:2106.04908*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Ailneni Rakshitha Rao. 2022. ASRtrans at SemEval-2022 task 4: Ensemble of tuned transformer-based models for PCL detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 344–351, Seattle, United States. Association for Computational Linguistics.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *CoRR*, abs/1805.07836.