FiRC at SemEval-2023 Task 10: Fine-grained Classification of Online Sexism Content using DeBERTa

Fadi Hassan† **Abdessalam Bouchekif†** **Walid Aransa†**

Huawei Technologies Oy (Finland) Co. Ltd., Finland

firstname.lastname@huawei.com

Abstract

SemEval 2023 shared task 10 "Explainable Detection of Online Sexism" focuses on detecting and identifying comments and tweets containing sexist expressions and also explaining why it is sexist. This paper describes the system that we used to participate in this shared task. Our model is an ensemble of different variants of fine-tuned DeBERTa models that employs a k-fold cross-validation. We have participated in the three tasks A, B and C. Our model ranked 2^{nd} position in tasks A, 7^{th} in task B and 4^{th} in task C.

Index Terms : Hate Speech, DeBERTa, Ensemble Models, Cross-validation.

1 Introduction

Social media platforms have been facing an increase in the number of hate speech targeting an individual or group, most often on the ground of religion, race, sexual orientation and gender, which makes these platforms hostile and toxic. Sexism is a form of discrimination that can be directed toward women. Sexism identification has been gaining attention recently among NLP community, mainly through shared tasks.

The EXIST 2021 challenge (Rodríguez-Sanchez et al., 2022) is the first shared task aimed to attract and motivate research efforts in building sexism detection models for English and Spanish languages. The proposed systems in EXIST 2021 challenge were mainly based on transformers (monolingual or multilingual). Hence, the best model in sub-task 1 (for Spanish language) (Butt et al., 2021) is obtained by applying specific pre-processing techniques and by using BERT model with data augmentation. The new training data was created by translating the comments into German language and then back-translated into Spanish and English. In (del Arco et al., 2021), the authors explore how transferred knowledge from tasks related to toxicity language may help in sexism tasks. Therefore, the proposed system follows a multitask learning approach where multiple tasks related to toxic comment identification are learned in parallel while using a shared representation. The model achieves convincing performance in both subtasks, 1^{st} place in sub-task 1 English and 2^{nd} place in sub-task 2 Spanish.

The second edition of EXIST was organized in 2022 (Rodríguez-Sanchez et al., 2022). Similar to the first edition, EXIST 2022 focused on "Sexism Identification and Sexism Categorization", both in Spanish and English. EXIST 2021 data sets are used as training data and a new data set consisting of 1058 tweets is used as a test set. Unsurprisingly, all participating teams used some sort of transformer architecture : Bert (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021) and BETO [*] (Cañete et al., 2020). (Vaca-Serrano, 2022) obtained powerful models for both tasks using different transformer-based language models with RoBERTa-large, BERTweet-large (Nguyen et al., 2020) and DeBERTa v3-large for English and RoBERTuito [**] (Pérez et al., 2021) for Spanish. The training was done in three steps. First, the hyper-parameters were optimized for every single model. Second, these optimized hyper-parameters were used to train models with more data. Thirdly, a simple ensembling strategy for combining models is applied. In (Villa-Cueva et al., 2022), 20 models are trained, 10 RoBERTuito and 10 BERT models. Each one of them is trained individually using different seeds. Finally, voting strategy is used to perform the final decision.

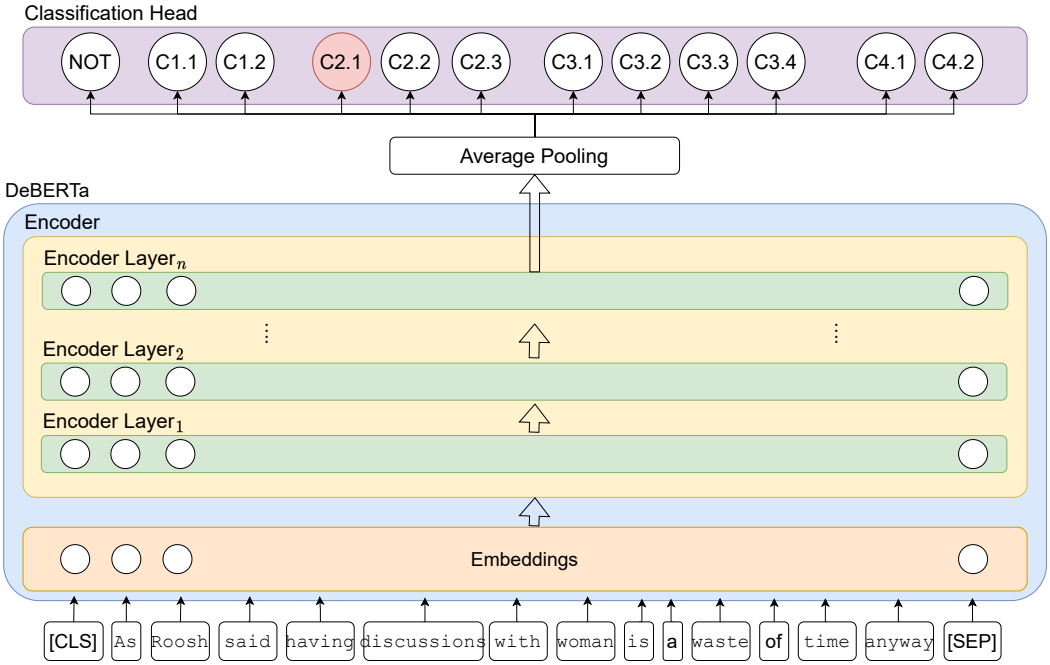
†. These authors contributed equally to this work

*. BETO is BERT model fine-tuned on Spanish data

**. RoBERTuito is a RoBERTa fine-tuned with 500M tweets in Spanish.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1824–1832

FIGURE 1 – Multi-class model for task A, B and C



The current task, SemEval 2023 Task 10 (Kirk et al., 2023), focuses on developing English-language models for sexism detection that are more accurate as well as explainable, with fine-grained classifications for sexist content from Gab and Reddit. The organizers propose three hierarchical subtasks :

— Task A (Binary Sexism Detection) : the post should be classified as sexist or not sexist.

— Task B (Category of Sexism) : The aim is to assign one of four sub-categories to the post that has been identified as sexist.

— Task C (Fine-grained Vector of Sexism) : The aim is to assign one of 11 sub-categories to the post that has been identified as sexist.

In Table 1, we described the list of classes of tasks B and C and a sample from each class training data. We also added our assigned class IDs so we can refer to these classes easily through this paper. In this paper, based on our related work on hate-speech detection (Ripoll et al., 2022), we proposed a system that consists of an ensemble of transformers and a k-fold cross-validation. In this work, we only used the shared task data.

The rest of the paper is organized as follows : Section 2 describes our proposed system. The experimental setup is described in Section 3, including the description of the data. Results and their analysis are given in Section 4. Finally, a conclusion is given with a discussion of future work in section 5.

2 System Description

One of the main challenges in this work was to apply a state-of-the-art model to a small dataset without getting an over-fitted model. First, we used DeBERTa (Decoding-enhanced BERT with disentangled attention) model. DeBERTa was proposed by Microsoft (He et al., 2021), and it improves upon the BERT and RoBERTa models using two novel techniques. The first technique is the disentangled attention mechanism, where each word is represented by two vectors that encode its semantic and positional information, respectively, and the attention scores among words are computed by using separate matrices for their semantic and relative positional features. The second technique is the enhanced mask decoder, which incorporates absolute positional embeddings in the decoding layer to predict the masked tokens during model pre-training.

To overcome the data scarcity problem. We addressed this issue by using k-fold cross-validation. We divided the train data into k different subsets. Therefore, k DeBERTa models were individually fine-tuned using nine subsets and one subset was reserved for validation. The final system was a combination of 10 models. To aggregate the output of the ten models, we used a soft voting approach

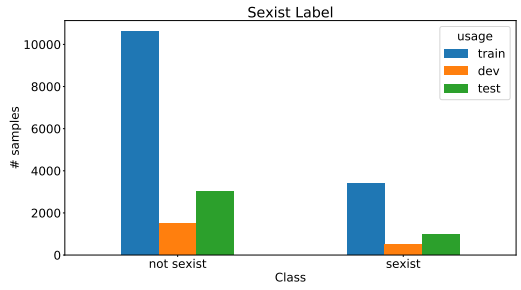
TABLE 1 – Examples of text for each task

Task A (**Sexism**)		**Text**
Task B (**Category**)	Task C (**Vector**)	
C1 Threats, plans to harm and incitement	C1.1 Threats of harm	I personally would have hit that bitch and she'd never have done it again.
	C1.2 Incitement and encouragement of harm.	Husbands. Kill your piece of shit commie wives
C2 Derogation	C2.1 Descriptive attacks	Classic toxic white females
	C2.2 Aggressive and emotive attacks	shes ugly af in both tho. looks like a tranny on the right
	C2.3 Dehumanising attacks & overt sexual objectification	Dude.. She is a prostitute. Fuck her, do not date her.
C3 Animosity	C3.1 Casual use of gendered slurs	Men can be whores also.
	C3.2 Immutable gender differences and gender stereotypes	Nobody cares about your emotions unless you're female
	C3.3 Backhanded gendered compliments	Femininity is more support/passive. So i guess it is true
	C3.4 Condescending explanations or unwelcome advice	All women benefit from the actions of violent men.
C4 Prejudiced discussions	C4.1 Supporting mistreatment of individual women	one thing i learned is to tell the judge she stayed the night. no woman who was raped would willing stay in a bed with a rapist
	C4.2 Supporting systemic discrimination against women as a group	There is no way to do so without enslaving yourself to a woman and the state.

where we averaged the prediction probability of all models to form the final score.

The proposed system serves multiple purposes and can be applied to all three tasks. Firstly, in task A, the system functions as a one-vs-all solution, distinguishing the negative class "not sexist" from the positive class "sexist", which could be any of the 11 sexist classes. Secondly, in task C, the proposed system can be utilized as a standalone model capable of predicting both the negative and positive classes or as a secondary system by suppressing the score of the negative class, thereby applying it exclusively to positive samples extracted by another binary classifier. Lastly, the proposed system can tackle task B in the same manner as task C, but an additional post-processing step is necessary to deduce one of the four sexist classes from the 11 fine-grained sexist classes.

FIGURE 2 – Task A data sets



3 Experimental Setup

3.1 Data

We had two approaches for our system in this competition. The first one was a top-down approach, where we would build a system consisting of three sub-models : 1) Task A model, a binary classifier to detect sexist samples, 2) Task B model, a multi-class classifier to assign one of four categories to the extracted sexist samples, and 3) a multi-class classifier to assign one of 11 subcategories to the extracted sexist samples. The second one was a bottom-up approach, where we would build a single model that could predict all 12 classes, including the 'not Sexist' class. In this approach, we aimed to solve all three tasks simultaneously using a single model.

We used the second approach, which is less complex than the first option and to avoid error propagation among the models. The architecture of our system is illustrated in Figure 1. It consists of a pre-trained DeBERTa model and a multi-class classification head. The output of the DeBERTa model is passed to an average pooling layer that aggregates all the generated embedding for each token and computes their mean to obtain a single vector representation. The vector obtained from the average pooling layer is fed into the classification head, where a softmax layer is applied to produce the predicted class probabilities. Finally, The predicted class is determined by selecting the class with the highest probability. The organizers of SemEval 2023 Task 10 (Kirk et al., 2023) provided two types of data : labeled and unlabeled data. Our models were trained solely on the labeled data, which was available only in English. Due to time constraints, we were unable to explore the potential of the unlabeled data. The training data For task A consists of 14K samples, the development set is 2k samples and the test set is 4k samples as shown in Figure 2. For this task, the "sexist" label count is 3398 samples (i.e. 24.27%) and the "not sexist" label count is 10602 samples (i.e. 75.72%). This shows that the class distribution is imbalanced, which can impact the model performance trained on this data as it may be biased towards the "not sexist" class and have lower accuracy on the "sexist" class. We had the same observation looking into task B and C training data.

FIGURE 3 – Task B data sets

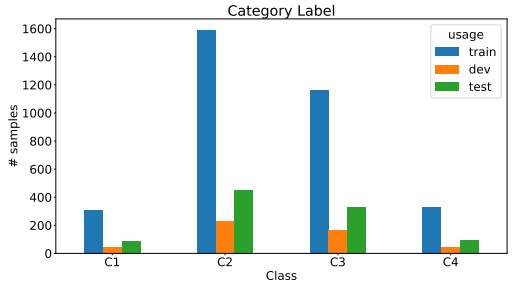
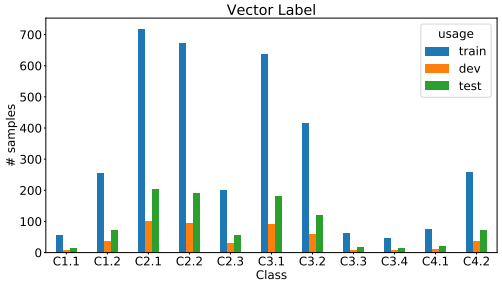


FIGURE 4 – Task C data sets



Since the training data of tasks B and C (see Figure 3 and Figure 4) is much less as they concern only the positive samples labeled "sexist" (*i.e.* 3398 samples), we decided to use the whole data including task A data to train a multi-task model. This model was trained on the whole $14k$ labeled samples for task C (including non-sexist samples). Task B result was extracted from task C results. Task B labels were not used for training but only during the extraction of task B categories from task C vectors. Table 1 provides a sample for each class as an illustration.

3.1.1 Lexicon Analysis

We conducted a lexicon analysis using the frequency of the words to find the most frequently used words in each class in task C. For example, the word 'smack' has a higher probability compared to other words in C1.1 "threats of harm". For C1.2 "descriptive attacks", the word 'treason' has a higher probability. We listed the words that have a higher probability in each class in Table 2. This lexicon analysis allowed us to find common words associated with each category. For example, samples in task B C1 "threats, plans to harm and incitement" are associated with attack and threats like 'smack','punch', 'kick', 'hang' and 'punishment'.

Also, in task C, C4.2 "supporting systemic discrimination against women as a group" is associated with related words like 'jobs', 'laws', 'government' and 'workplace'. This lexicon analysis gave us more understanding of the differences between classes on keywords level.

3.1.2 Text Pre-processing

We tried different text pre-processing techniques to clean the data, but we did not get any significant improvement on the development data set. Our final model used for the submission does not use any text pre-processing.

3.2 Training

Before constructing the final system, we tested multiple pre-trained models by fine-tuning them on the training set. As all tasks contained only English language data, monolingual pre-trained models consistently outperformed cross-lingual models. Table 3, 4 and 5 show that DeBERTa demonstrated the best performance for all tasks. All models are 10-fold cross-validation except the first and the second row, and the final output was the average of the 10-fold models. To determine the optimal number of epochs, we use the early stopping mechanism for each fold with macro F1-score as the evaluation metric.

4 Results and Analysis

We initiated our model development process with the creation of a strong baseline model that served as a benchmark for evaluating the performance of our subsequent models. Tables 3, 4, and 5 present the results of our developed models compared to the baseline for tasks A, B, and C, respectively. Precision, recall, and F1-score were used to evaluate the models' performance on both the dev and test sets, with F1-score serving as the official evaluation metric for the shared task. Our submitted model achieved an F1-score of 87.4%, 70.58%, and 54.04% on the test sets for tasks A, B, and C, respectively.

In the subsequent sections, we delve into the various experiments we conducted to select the optimal pre-trained model or improve the model's performance on the target task.

TABLE 2 – Lexicon analysis : top 10 related words in task C fine-grained vectors

Task C vector	# samples	Words
C1.1 threats of harm	56	smack, rofl, drinking, punch, nazi, ive, pull, knees, push, love
C1.2 incitement and encouragement of harm	254	treason, hang, thrown, asap, rope, four, punishment, kick, traitor, tried
C2.1 descriptive attacks	717	movement, sexism, brains, earned, messed, highest, becomes, spending, random, desperate
C2.2 aggressive and emotive attacks	673	stormy, roastie, wondering, cats, witch, nudes, opened, died, lying, slag
C2.3 dehumanising attacks & overt sexual objectification	200	foids, dolls, femoids, videos, besides, foid, dream, knowing, legs, option
C3.1 casual use of gendered slurs, profanities, and insults	637	bitching, panties, solid, happiness, delete, karma, faggots, faggot, thots, speech
C3.2 immutable gender differences and gender stereotypes	417	figures, traditional, blonde, emotions, romance, partners, chase, catch, smv, depends
C3.3 backhanded gendered compliments	64	space, travel, whale, hotter, acts, successful, speakfreely, safe, tits, balls
C3.4 condescending explanations or unwelcome advice	47	misogyny, definition, skin, step, treat, air, color, lesbian, therefore, pink
C4.1 supporting mistreatment of individual women	75	weinstein, entirely, traitors, twice, assholes, victim, raping, changed, metoo, choices
C4.2 supporting systemic discrimination against women as a group	258	quotas, spaces, supremacy, norms, accused, jobs, laws, privilege, government, workplace

TABLE 3 – Task A - The impact of various model choices on the system performance

Model	Dev set			Test set		
	Precision	Recall	F1-score	Precision	Recall	F1-score
XLM-RoBERTa$_{large}$: (Baseline)	85.72	82.72	84.06	84.73	82.6	83.58
DeBERTa$_{large}$: (1-fold)	86.14	83.72	84.82	86.81	85.46	86.1
DeBERTa$_{large}$	87.04	85.98	86.49	86.26	86.91	86.58
DeBERTa$_{xlarge}$	87.33	85.67	86.45	87.09	86.83	86.96
$^{(*)}$**DeBERTa**$_{large+xlarge}$	87.37	86.36	86.84	87.54	**87.25**	**87.4**
$^{(**)}$**DeBERTa**$_{xxlarge}$	**87.86**	**86.45**	**87.12**	**87.56**	86.54	87.03
$^{(***)}$**DeBERTa**$_{large+xlarge+xxlarge}$	87.19	86.29	86.73	87.24	86.93	87.08

* Submitted system - ensemble of 20 models ** Post-submission system - ensemble of 10 models *** Post-submission system - ensemble of 30 models

TABLE 4 – Task B - The impact of various model choices on the system performance

Model	Dev set			Test set		
	Precision	Recall	F1-score	Precision	Recall	F1-score
XLM-RoBERTa$_{large}$: (Baseline)	65.64	67.26	64.78	60.74	64.46	60.74
DeBERTa$_{large}$: (1-fold)	65.06	63.18	63.69	65.89	64.78	65.08
DeBERTa$_{large}$	72.91	72.34	72.41	71.28	70.57	70.72
DeBERTa$_{xlarge}$	74.42	**76.66**	75.42	70.14	70.34	70.09
$^{(*)}$**DeBERTa**$_{large+xlarge}$	76.21	76.1	75.91	70.72	70.72	70.58
$^{(**)}$**DeBERTa**$_{xxlarge}$	**77.41**	76.05	76.22	**73.32**	69.5	70.87
$^{(***)}$**DeBERTa**$_{large+xlarge+xxlarge}$	76.96	76.59	**76.46**	71.92	**71.55**	**71.57**

* Submitted system - ensemble of 20 models ** Post-submission system - ensemble of 10 models *** Post-submission system - ensemble of 30 models

TABLE 5 – Task C - The impact of various model choices on the system performance

Model	Dev set			Test set		
	Precision	Recall	F1-score	Precision	Recall	F1-score
XLM-RoBERTa$_{large}$: (Baseline)	50.07	44.12	42.56	40.31	41.39	38.45
DeBERTa$_{large}$: (1-fold)	42.45	40.17	40.41	44.38	40.9	41.31
DeBERTa$_{large}$	53.95	55.87	54.01	49.4	49.58	48.8
DeBERTa$_{xlarge}$	62.98	58.17	59.33	56.7	51.7	53.15
$^{(*)}$**DeBERTa**$_{large+xlarge}$	65.05	**60.48**	**61.29**	56.37	53.18	54.04
$^{(**)}$**DeBERTa**$_{xxlarge}$	66.5	57.15	57.91	**59.32**	51.4	53.2
$^{(***)}$**DeBERTa**$_{large+xlarge+xxlarge}$	**69.58**	59.88	61.05	57.8	**53.32**	**54.61**

* Submitted system - ensemble of 20 models ** Post-submission system - ensemble of 10 models *** Post-submission system - ensemble of 30 models

4.1 Monolingual or Multilingual Pre-trained Model ?

We experiment using monolingual and multilingual pre-trained models. As all tasks contained only English language data, monolingual pre-trained models consistently outperformed cross-lingual models. Additionally, DeBERTa demonstrated the best performance for all tasks vs. other transformer models like BERT/RoBERTa-large.

4.2 Does Model Size Matter ?

We experiment with various DeBERTa models with different sizes. We showed in our results in Tables 3, 4, and 5, the impact of using different model sizes (i.e. large, xlarge and xxlarge) on the task performance. The bigger the model, the better performance. Additionally, ensembling different sizes of models performs better than using a single model. Our best-submitted and post-submission models use an ensemble of various models with different sizes.

4.3 Effectiveness of K-fold Cross Validation

K-fold models consistently outperform the 1-fold models, as evidenced by the significantly higher performance scores observed in the third row compared with the second row of Tables 3, 4, and 5. These results further support our hypothesis that the number of available samples is insufficient and that employing k-fold models can effectively enhance the robustness of our model.

4.4 Result Error Analysis

We conducted a thorough analysis of our models on the test set. For each task, we utilized the test set confusion matrix to identify areas of confusion, identify classification errors, and identify potential challenges. In the following sections, we present our detailed analysis along with possible solutions that may help enhance the performance of our models.

4.4.1 Task A Error Analysis

The test set confusion matrix for this task is shown in Figure 6. You can see that our model has high precision given the high TN (i.e. "not sexist" class). In contrast, the model has lower recall given the lower TP (*i.e.* "sexist" class). This was expected since the training data is imbalanced and it has significantly more samples from "not sexist" class (*i.e.* 75.73%) compared to "sexist" class (*i.e.* 24.27%). There are several approaches to improve the model to overcome the imbalanced training set like adding more positive samples to the training data or using data augmentation techniques.

4.4.2 Task B Error Analysis

The test set confusion matrix for this task is shown in Figure 7. We found that the accuracy of classes C1/C3 are better than the other two classes C2/C4. Clearly, our model is confusing the latter two classes.

Here, we try to use the explainable detection objective of this shared task to shed some light on the reason the model was confused C4 with "C3 and C4" since task B C4 has two fine-grained vectors (i.e. sub-classes) in task C : C4.1, C4.2. Task B C2 has three fine-grained vectors in task C : C2.1, C2.2 and C2.3. Task B C3 has four fine-grained vectors in task C : C3.1, C3.2, C3.3 and C3.4. We will use the test set confusion matrix for task C shown in Figure 5 to investigate

FIGURE 5 – Task C test set confusion matrix

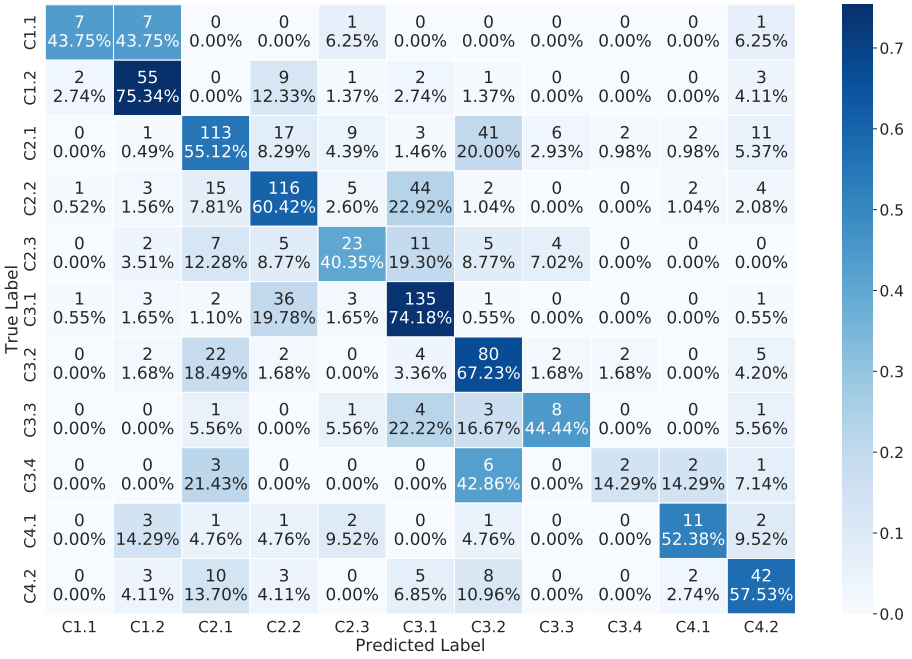


FIGURE 6 – Task A test set confusion matrix

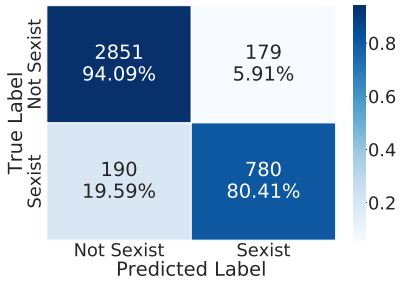
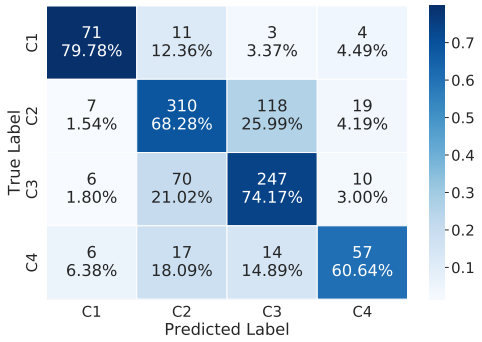


FIGURE 7 – Task B test set confusion matrix



the root cause of the confusion. We found that there are two main confusions as follows :

— Between C4.2 and C2.1 (i.e. 10 samples)
We found that C4.2, which is "supporting systemic discrimination against women as a group" and C2.1 "descriptive attacks" are both very close, which explains why the model was struggling to learn to discriminate between these two classes.

— C4.2 and C3.2 (i.e. 8 samples).
We observed that C4.2 and C3.2 "immutable gender differences and gender stereotypes" are both very close, which explains why the model was struggling to learn to discriminate between these two classes.

If the model is confused between "C4.2 and C2.1" and "C4.2 and C3.2", this means that there should be confusion between "C2.1 and C3.2" as well.

In order to confirm our findings, we had to check the test set confusion matrix of task C for

any confusion between C2.1 and C3.2 and we found that there is high confusion for C2.1 as C3.2 (*i.e.* 41 samples). On the other direction, the model is also confusing C3.2 as C2.1 (*i.e.* 22 samples), which confirms our findings.

4.4.3 Task C Error Analysis

The test set confusion matrix for this task is shown in Figure 5. We found that our model confused C1.1 with C1.2. This can be explained by the assumption that there is a high similarity between these two fine-grained vectors (*i.e.* Threats of harm vs. Incitement and encouragement of harm). However, this assumption failed to explain why the model is not confusing C1.2 with C1.1 (*i.e.* only two samples). By checking the training data, we found that the number of training samples for C1.1 is 56 samples which is 22% of the training samples for C1.2. This makes the model more biased towards C1.2, which has a higher number of training samples leading to a high number of C1.1 false negatives or false positives. There are several possible solutions for this problem, for example, adding more training samples for C1.1, using data augmentation or re-sampling to balance the training data.

4.4.4 Explainablility using Bottom-up Classification

Deep neural networks are often considered opaque or "black box" systems, with their decision-making processes being difficult to interpret. The design of this shared task is of significant importance to us as it provides an opportunity to evaluate a data-driven approach to enhancing the explainability of decisions made by neural networks.

We choose to use this shared task to emphasize this data-driven approach for explainability since the annotated data that has a sexist class of Task A has been categorized to one of 4 sub-categories of Task B, which is subsequently categorized to one of the fine-grained vectors of Task C.

When binary classification models are used for content moderation in social media (*e.g.,* 'not sexist'/'sexist' binary classifier), the flagged content may be hidden or removed. However, the author may request republishing or further justification for the removal. In such cases, a human reviewer must spend additional time reviewing the content to provide an explanation for the model's decision at a high cost.

In contrast, our model can provide the author with detailed information on the sexist violation, including the category and fine-grain vector. This approach offers a data-driven, affordable, automatic and rapid level of explainability. While this strategy may require more effort in the data annotation phase to carefully annotate data at the level of fine-grained sub-classes, we believe it will ultimately provide long-term benefits in the production phase and allow for greater explainability of the model's decision.

5 Conclusion

In this paper, we described our system for SemEval 2023 shared task 10, "Explainable Detection of Online Sexism", which ranked 2^{nd}, 7^{th} and 4^{th} in task A, B and C, respectively. Our system is an ensemble of different variations of fine-tuned DeBERTa models that employs the k-fold cross-validation.

Since this task supports the development of English-language models for sexism detection, We showed that the choice of the monolingual pre-trained model has better performance than the multi-lingual pre-trained model. We also showed that the k-fold cross-validation allowed our model to get the best from the training data, especially for classes with a low number of training samples.

We also presented the impact of the size of the pre-trained model. We think that all these techniques made our model more robust and well generalized with the good results on the test set that our system achieved. For future work, we would improve our model by using monolingual data and data augmentation techniques to overcome the limited training data for some classes.

References

Flor Miriam Plaza del Arco, M. Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. 2021. SINAI at iberlef-2021 DETOXIS task : Exploring features as tasks in a multi-task learning approach to detecting toxic comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*.

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

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020(2020) :1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kris-

tina Toutanova. 2018. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

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

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. *arXiv preprint arXiv :1907.11692*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet : A pre-trained language model for english tweets. *arXiv preprint arXiv :2005.10200*.

Juan Manuel Pérez, Damián A Furman, Laura Alonso Alemany, and Franco Luque. 2021. Robertuito : a pre-trained language model for social media text in spanish. *arXiv preprint arXiv :2111.09453*.

María Luisa Ripoll, Fadi Hassan, Joseph Attieh, Guillen Collell, and Abdessalam Bouchekif. 2022. Multi-lingual contextual hate speech detection using transformer-based ensembles. In *Forum for Information Retrieval Evaluation*.

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.

Alejandro Vaca-Serrano. 2022. Detecting and classifying sexism by ensembling transformers models. *language*, 2 :1.

Emilio Villa-Cueva, Fernando Sanchez-Vega, and Adrián Pastor López-Monroy. 2022. Bi-ensembles of transformer for online bilingual sexism detection.