JUST_ONE at SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS)

Doaa Obeidat, Wala'a Shnaigat, Heba Nammas, Malak Abdullah
Jordan University of Science and Technology, Irbid, Jordan
dfobeidat20, washnaigat20, hrnammas20@cit.just.edu.jo, mabdullah@just.edu.jo

Abstract

The problem of online sexism, which refers to offensive content targeting women based on their gender or the intersection of their gender with one or more additional identity characteristics, such as race or religion, has become a widespread phenomenon on social media. This can include sexist comments and memes. To address this issue, the SemEval-2023 international workshop introduced the "Explainable Detection of Online Sexism Challenge", which aims to explain the classifications given by AI models for detecting sexism. In this paper, we present the contributions of our team, JUST_ONE, to all three sub-tasks of the challenge: subtask A, a binary classification task; subtask B, a four-class classification task; and subtask C, a fine-grained classification task. To accomplish this, we utilized pre-trained language models, specifically BERT and RoBERTa from Hugging Face, and a selective ensemble method in task 10 of the SemEval 2023 competition. As a result, our team achieved the following rankings and scores in different tasks: 19th out of 84 with a Macro-F1 score of 0.8538 in task A, 22nd out of 69 with a Macro-F1 score of 0.6417 in task B, and 14th out of 63 with a Macro-F1 score of 0.4774 in task C.

Keywords: Sexism, Muslim woman, Fine-grain, NLP, Transformer, Hugging face, Text classification

1 Introduction

Social media has facilitated the freedom of expression, but it has also enabled the spread of hate speech against specific communities, such as races, sexes, and ethnicities (Zhang and Luo, 2019; Habash et al., 2022; Faraj and Abdullah, 2021). Sexism is a particularly challenging form of hate speech to address because it can be difficult to identify and significantly impacts women's lives. Women are often subjected to abusive and discriminatory behaviors based on gender, religion, or other identity factors (Younus and Qureshi, 2022; Abburi et al., 2021; Frenda et al., 2019). Given the seriousness of the problem, researchers in natural language processing (NLP) have been inspired to develop novel methods for sexism in text detection and classification (Butt et al., 2021).

This paper discusses the JUST_ONE team's contribution to SemEval 2023-task10. The task aimed to detect and identify various types of sexist content using Gab and Reddit data (Kirk et al., 2023). We used the pre-trained language models: BERT (Devlin et al., 2018), and RoBERTa / unlabeled RoBERTa (Liu et al., 2019) from Hugging face, alongside the ensemble method (selective ensemble), and optimized the models' hyperparameters for each of the three subtasks (A, B, C). It is worth mentioning that our team ranked 19th out of 84 in task A, with a Macro-F1 score of 0.8538, 22nd out of 69 in task B with a Macro-F1 score of 0.6417, and 14th out of 63 in task C, with a Macro-F1 score of 0.4774.

The rest of the paper is structured as follows: A background of task setup and data description is covered in section 2. Section 3 provides insights into the related work. A system overview is provided in section 4. The experiments are described in section 5; section 6 provides our results. Finally, we conclude this work in section 7.

2 Background

2.1 Task Setup

In SemEval2023-task 10, there are three hierarchically arranged subtasks (Kirk et al., 2023). Below is the brief definition for each subtask:

- TASK A - Binary Sexism Classification: The system aims to determine if the input text is sexist, with two possible outcomes: sexist or not sexist. An example of input text is: "O come on, there's no way any men have attracted to her she a pig", which contains a

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 526–531

sexist statement, and the corresponding output is sexist.

- TASK B - Category of Sexism: The system aims to categorize sexist content into one of four categories: 1. threats, 2. derogation, 3. animosity, and 4. prejudiced discussions. An example of sexist input text is "O come on, there's no way any men are attracted to her she is a pig", and the output is class 2, indicating that the text falls into the derogation category because it involves insulting and belittling a woman based on her appearance.

- TASK C - Fine-grained Vector of Sexism: By using fine-grained vectors, the system can capture more detailed information about the category of sexism displayed in a given sexist post, detecting one of eleven Fine-grained Vectors. An example of sexist input text is Ö come on thereś no way any men are attracted to her shes a pig, and the output: 2.3 dehumanizing attacks & overt sexual objectification, which refers to a more detailed category derogation. The phrase s̈heś a pigïs dehumanizing, as it compares the woman to an animal, implying that she is dirty, unattractive, and unworthy of respect or consideration.

2.2 Data Description

The dataset used in this task consists of 20,000 posts labeled in English (Kirk et al., 2023), divided into training, development, and testing sets with proportions of 70%, 10%, and 20%, respectively. Table 1 provides detailed information on these sets, with training data applicable to all subtasks and separate data for each subtask in the development and testing sets. Using this dataset, we created a task-specific dataset, as illustrated in Figure 1. In addition, table 1 provides the definitions for each variable used in the dataset.

tabularx

3 Related Work

The field of text classification has seen significant advancements thanks to machine learning, with a particular focus on identifying instances of sexism in text. Researchers have explored various approaches in this area, including collecting sexist data and employing different techniques to detect the presence of a sexual agenda in a piece of

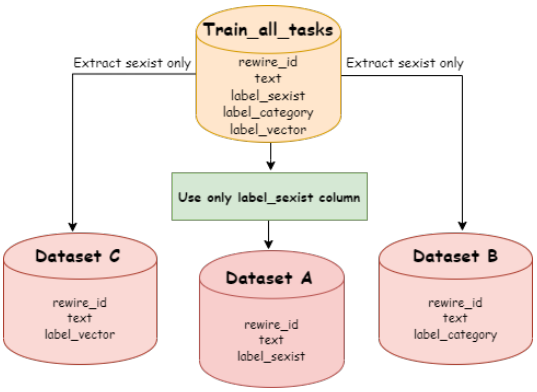

Figure 1: The Preparation of data

Table 1: Dataset Description

Variable name	Description
rewire_id	unique identifier for each entry
text	input text (post)
label_sexist	the label is sexist, or not
label_category	category of each sexist post
label_vector	sub-category of each sexist post

text. For example, in a recent study by the authors of (Liakhovets et al., 2022), three transformers were used to detect sexism and classify text into multiple categories in response to the 2022 IberLEF EXIST challenge. These included a monolingual T5 model for English, a multilingual BERT, and the XLM-RoBERTa model, which achieved a Macro F1-score of 0.7496 for binary classification and 0.4675 for multi-class classification. The researchers further improved performance by using unsupervised pre-training with additional data and data augmentation.

The authors of (Fersini et al., 2019) developed a unimodal and multimodal classifier for detecting instances of written and visual sexism in social media memes using various machine learning models. The researchers collected a benchmark dataset of 800 memes from various social media platforms and extracted visual and textual features. Using these features, they developed unimodal classifiers for detecting sexism in memes based on visual or textual features and a multimodal classifier created through early and late fusion. The best binary classification performance was achieved using a Support Vector Machine (SVM) classifier with a recall of 0.81 and 0.83 for the neural network classifier.

Shibly and colleagues employed multiple ma-

chine learning algorithms, including Decision Tree, Logistic Regression, Random Forest, Long Short-Term Memory, and a voting classifier, to automatically detect instances of hate speech targeting women, as detailed in (Shibly et al., 2023). They used the MeToo open-source dataset for their study and found that the voting classifier model achieved the highest F1-measure value of 0.9. In (El Ansari et al., 2020), two different strategies were suggested for creating a thematic training package that integrates manual and machine translations of Arabic materials related to gender inequality and violence against women.

In the SemiEVAL 2020 competition focused on identifying the offensive language in social media across multiple languages (OffensEval 2: Multilingual Offensive Language Identification in Social Media), Das and their team participated and explained their methodology in their publication cited as (Das et al., 2020). The team utilized a checkpoint ensemble-based transformer model composed of Ensembled GPT-2, Ensembled RoBERTa, and Ensembled DistilRoBERTa to address the competition's three subtasks. Additionally, the attention mask dropout technique was employed to mitigate the issue of poor composition in social media texts. The team's performance in the English edition of the competition was notable, with Macro F1 scores of 0.909, 0.551, and 0.616 achieved for the three subtasks across five languages. The focus of (Zimmerman et al., 2018) was on hate speech in social media and the concerns around censorship, with a particular emphasis on human rights. The authors proposed new methods to identify and address discrimination while preserving freedom of expression more effectively. They introduced a novel clustering method that utilizes a neural network approach to improve the classification of hate speech. To evaluate its performance, the method was tested on the Twitter hate speech dataset and the common sentiment dataset using a publicly available inclusion model. The results demonstrated an improvement of around 5 points on an F1 scale compared to the original work on a publicly available hate speech assessment dataset. The paper also acknowledges the challenges of reproducing deep learning methods and comparing the results of other studies.

4 System Overview

The study outlines the approach for designing the system, which is tailored for each sub-task based on factors such as the task's difficulty level and dataset size. Multiple experiments were conducted to ensure the design met the system's needs. This flexible approach allows for adaptation to the unique requirements of the system. The details of the work on each sub-task are presented separately as follows:

4.1 Task A

We fine-tuned the RoBERTa architecture with specific hyperparameters to identify sexist content through binary classification. The optimal hyperparameters for this model include the following: A learning rate of 2e-5. Four epochs with a batch size of 16. A dropout of 0.2. Two fully connected layers.

4.2 Task B

In this sub-task, we designed a simple selective ensemble approach that depends on the highest probability among two models. The technique combines the predictions of two pre-trained models (RoBERTa and BERT) and selects the one with the highest probability as the final prediction. This approach takes advantage of the strengths of both models, leading to improved performance and reduced overfitting. The selective ensemble was designed to improve the overall F1 score of the system and address the limitations of individual models. We achieved improved results by carefully selecting the models to be combined after diverse experiments. For example, if model 1 generates a prediction with a probability of 0.95 for class 2, and model 2 generates a prediction with a probability of 0.56 for class 1, the final prediction would be class 2 as it has the highest probability among the two models. The approach of the selective ensemble is shown in Figure 2.

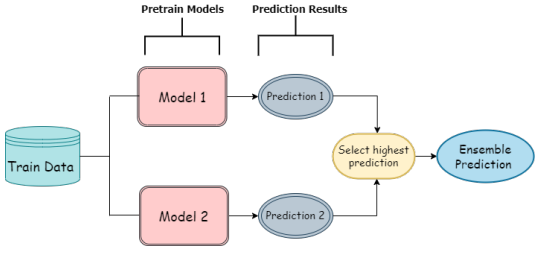

Figure 2: Selective Ensemble Method

4.3 Task C

For this task, we employed the same approach as in task B, which combined the predictions of two pre-trained models. However, for this task, we utilized one labeled model (RoBERTa) and one unlabeled model (RoBERTa) from Hugging Face to make predictions. Our experiments utilizing a selective ensemble approach demonstrated that choosing the prediction with the highest probability among the two predictions generated the most optimal outcomes. Additionally, to overcome the issue of class imbalance, we incorporated the widely utilized Focal loss in text classification for all sub-tasks.

5 Experimental Setup

The labeled dataset comprises 10,000 samples from GAB and 10,000 from Reddit, split into 14,000 entries for training, 2,000 for development, and 4,000 for testing. Additionally, 2 million unlabeled data entries augment the training stage. We evaluated multiple pre-trained models by training them on the labeled training set and testing their performance on a validation set with an 80:20 ratio. The models were trained for 3 or 4 epochs, depending on the size of the dataset, and we selected the best model checkpoint based on its Macro F1 score on the development set. We took steps to prevent overfitting during this process. Transfer learning was employed in our approach by reusing fine-tuned embedding parameters. This is done by saving the checkpoint of fine-tuned model embeddings from the earlier stage and utilizing it as the initialization embeddings for the next model stage. The illustration of transfer learning is shown in Figure 3. The methodology employed in this study involved running our code on the Kaggle environment using Python 3.9 and PyTorch. For all experiments, we utilized the implementation by Hugging Face to import pre-trained models. We also used the focal loss to deal with the imbalanced dataset. We performed multiple experiments to fine-tune various pre-trained models, determine which ones were best suited for our sub-tasks, and identify the most appropriate hyperparameters. Table 2 offers a detailed explanation of the particular terms utilized in the experiments. Moreover, Tables 3, 4, and 5 display the outcomes of our fine-tuning and experimentation across three distinct tasks on the development dataset.

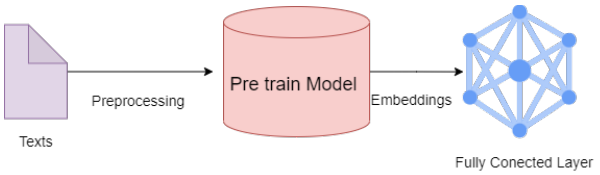

Figure 3: Transfer Learning Model

6 Results

Our team performed satisfactorily for the three subtasks of SemEval 2023 - Task 10: Explainable Detection of Online Sexism. The RoBERTa large model produced promising results for task A, achieving a Macro-F1 score of 0.8538 and ranking 19 out of 84 teams. For task B, the selective ensemble model resulted in a Macro-F1 score of 0.641 and an order of 22 out of 69, while for task C, it achieved a Macro-F1 score of 0.4774 and a ranking of 14 out of 63. Our participation highlights the importance and urgency of detecting online sexism. We used different approaches for each task to adapt to them. In Task A, we used fine-tuned pre-trained model to get the best performance we reached. While we realized that using the ensemble approach in tasks B and C gave better results than fine-tuning, this is because we assume that ensemble approaches are more effective when there is variance in the data and avoid overfitting when fine-tuning is used with more complex data. Table 6 shows the results we have obtained throughout our contribution.

7 Conclusion

This paper outlines our approach to participating in SemEval 2023 - Task 10: Explainable Detection of Online Sexism (EDOS), a crucial and relevant issue that negatively impacts women. The task attracted numerous teams this year and comprised three subtasks. We utilized several models based on pre-trained RoBERTa and BERT architecture and a selective ensemble approach. For the binary classification subtask (A), we fine-tuned the RoBERTa model, which resulted in a Macro F1 score of 0.8538. Our selective ensemble model achieved a Macro F1 score of 0.6417 for subtask B and 0.4774 for subtask C. In summary, our future work includes investigating alternative ensemble strategies like weighting average and majority voting to improve the performance of our EDOS system in multi-class subtasks (B) and (C). We also plan to explore other pre-training techniques, such

Table 2: Elucidation of Certain Terminology.

Terminology	Explaining
Learning Rate (LR)	A hyperparameter that determines the step size at which the optimizer makes updates to the model parameters during training
The number of epochs	Refers to the number of complete iterations of the training process over the entire training dataset
Pretrain Model	Is ML model trained on a large dataset and then fine-tuned to make predictions on new data without training.
Max sequence length (Max_seq)	Determines the maximum number of time steps or tokens to which a data sequence will be padded or truncated.
Train batch size	A hyperparameter that determines the number of training samples to use in each iteration of the model's parameters during training.
NN	Neural Network.
Dropout	Is a regularization technique in deep learning used to prevent overfitting.
F1 Score	Is a commonly used metric to evaluate the performance of a binary classification model.

Table 3: Fine-tuning of Hyperparameters during the Training Phase of our Model on task A, with dropout = 0.2

#	LR	Epochs	Pretrain model names	Max_ seq	Train batch size	NN	F1 Score
1	4e-5	4	Roberta Large	128	8	2 FCL	0.8411
2	2e-6	4	Roberta Large	128	8	2 FCL	0.8111
3	5e-5	4	Roberta Large	128	8	2 FCL	0.7943
4	2e-5	4	Roberta Large	128	16	3 FCL	0.7243
5	**2e-5**	**4**	**Roberta Large**	**128**	**16**	**2 FCL**	**0.8611**
6	2e-5	4	Roberta Large (Liu et al., 2019)	128	32	2 FCL	0.8233
7	2e-5	4	"unlabeled Roberta-270000sample"	128	16	2 FCL	0.8254
8	2e-5	4	"distilbert-base-uncased-finetuned-sst"	128	16	2 FCL	0.84165
9	2e-5	4	"cardiffnlp/twitter-roberta-base-sentiment"	128	16	2 FCL	0.84488
10	2e-5	4	"ProsusAI/finbert"	128	16	2 FCL	0.8245

Table 4: Fine-tuning of Hyper-parameters during the Training Phase of our Model on task B with dropout = 0.2

#	LR	Epochs	Pretrain model names	Max_seq	Train batch size	NN	F1 Score
1	2e-5	4	BERT- Large-uncased	128	16	2 FCL	0.6643
2	2e-5	3	Roberta Large	128	16	2 FCL	0.6737
3	**-**	**-**	**Selective-Ensemble**	**-**	**-**	**-**	**0.6869**

Table 5: Fine-tuning of Hyper-parameters during the Training Phase of our Model on task C with dropout =0.2

#	LR	Epochs	Pretrain model names	Max_ seq	Train batch size	NN	F1 Score
1	2e-5	3	Roberta Large	128	8	1 FCL	0.478
2	2e-5	3	" unlabeled Roberta-270000sample"	128	8	2 FCL	0.4819
5	**-**	**-**	**Selective ensemble**	**-**	**-**	**-**	**0.4843**

Table 6: Results of different subtasks in the testing phase

Models	Score
Task A (RoBERTa Large)	0.8538
Task B (Selective Ensemble of RoBERTa and BERT)	0.6417
Task C (Selective Ensemble of RoBERTa (Labeled and unlabeled)	0.477

as GPT-3, and to use semi-supervised and unsupervised approaches to address data scarcity by utilizing 2 million unlabeled data.

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.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F Gelbukh. 2021. Sexism identification using bert and data augmentation-exist2021. In *IberLEF@ SEPLN*, pages 381–389.

Kaushik Amar Das, Arup Baruah, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2020. Kafk at semeval-2020 task 12: Checkpoint ensemble of transformers for hate speech classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2023–2029.

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.

Oumayma El Ansari, Zahir Jihad, and Mousannif Hajar. 2020. A dataset to support sexist content detection in arabic text. In *International Conference on Image and Signal Processing*, pages 130–137. Springer.

Dalya Faraj and Malak Abdullah. 2021. Sarcasmdet at semeval-2021 task 7: detect humor and offensive based on demographic factors using roberta pretrained model. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 527–533.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Mohammad Habash, Yahya Daqour, Malak Abdullah, and Mahmoud Al-Ayyoub. 2022. Ymai at semeval-2022 task 5: Detecting misogyny in memes using visualbert and mmbt multimodal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784.

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.

Daria Liakhovets, Mina Schütz, Jaqueline Böck, Medina Andresel, Armin Kirchknopf, Andreas Babic, Djordje Slijepčević, Jasmin Lampert, Alexander Schindler, and Matthias Zeppelzauer. 2022. Transfer learning for automatic sexism detection with multilingual transformer models.

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.

FHA Shibly, Uzzal Sharma, and HMM Naleer. 2023. Automatic detection of online hate speech against women using voting classifier. In *International Conference on Innovative Computing and Communications*, pages 735–745. Springer.

Arjumand Younus and Muhammad Atif Qureshi. 2022. A framework for sexism detection on social media via byt5 and tabnet.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.