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

Rakib Hossain Rifat **Abanti Chakraborty Shruti**

Marufa Kamal **Farig Sadeque**

BRAC University, Dhaka, Bangladesh
rakib.hossain.rifat@g.bracu.ac.bd, abanti.chakraborty.shruti@g.bracu.ac.bd,
marufa.kamal1@g.bracu.ac.bd, farig.sadeque@bracu.ac.bd

Abstract

People are expressing their opinions online for a lot of years now. Although these opinions and comments provide people an opportunity of expressing their views, there is a lot of hate speech that can be found online. More specifically, sexist comments are very popular affecting and creating a negative impact on a lot of women and girls online. This paper describes the approaches of the SemEval-2023 Task 10 competition for Explainable Online Sexism Detection (EDOS). The task has been divided into 3 subtasks, introducing different classes of sexist comments. We have approached these tasks using the bert-cased and uncased models which are trained on the annotated dataset that has been provided in the competition. Task A provided the best F1 score of 80% on the test set, and tasks B and C provided 58% and 40% respectively.

1 Introduction

Sexism is prejudice or discrimination based on one's sex or gender. Sexism can affect anyone, but it primarily affects women and girls who face negative sentiment or abuse based on their gender combined with one or more other identity attributes (e.g. Black women, Muslim women, Trans women). It has been linked to stereotypes and gender roles and may include the belief that one sex or gender is intrinsically superior to another(Wikipedia contributors, 2023). Social media platforms have promoted the spread of hate speech through anonymization and accessibility, spurring greater study into developing automatic algorithms to recognize these sorts of writings. According to a study(Duggan, 2017) on online harassment, women experience harassment due to their gender twice as often as males do. Bullying online can lead to depression, moreover, a study(Fulper et al., 2014) related to rape cases found a connection between the number of sexist tweets and the number of rapes in the United States of America. An online poll conducted by Amnesty International across eight high-income countries in 2017 revealed that 23% of women had experienced some form of abuse or harassment on social media platforms(Center on Gender Equity and Health (GEH) – UC San Diego, 2023). It can cause harm to targeted women, render online places inhospitable and inaccessible, and maintain societal inequities and asymmetries. Social media text can be inconsistent and mixed with sarcasm often times which can make it challenging to detect sexist texts. The categorization of sexism differs from and may be complemented by the detection of hate speech. Although sexist comments may be regarded as hate speech and there is a lot of research conducted regarding hate speech detection, sexism sentences have a broader aspect, and studies on sexism detection have a huge scope. Despite using various automated tools in the digital space to identify and flag high-level sexist content, however, they fail to recognize potential content and explain the cause behind it. The capacity to identify sexist content and to explain why it is sexist enhances the interpretability, trust, and comprehension of the choices made by automated systems, giving users and moderators more control. It is crucial to be able to improve the automatic detection and classification of sexism. It may aid in the analysis of sexism in order to enhance sensitization campaigns and implement other countermeasures towards this oppression.

This research is based on the classification of sexist data and its different categories. Firstly, we target a binary classification of a sentence indicating whether it is sexist. Secondly, if it is classified as sexist, we further look into fine-grained sexism categories. We have dealt with both labeled (14k) and large amounts of unlabelled data(2M) and used the BERT transformer model to classify and categorize the texts.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 724–732
July 13-14, 2023 ©2023 Association for Computational Linguistics

2 Background

2.1 Task Description

This task supports the development of 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.(Kirk et al., 2023)

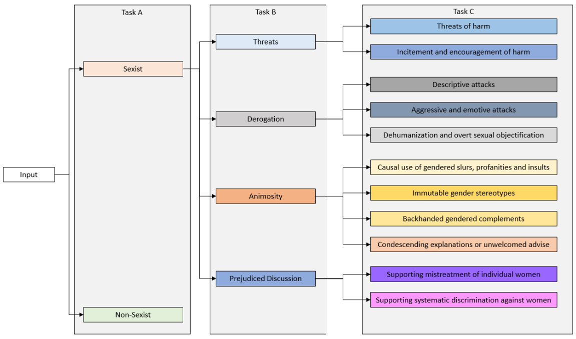

Task A
Task B
Task C
Input
Sexist
Non-Sexist
Threats
Derogation
Animosity
Prejudiced Discussion
Threats of harm
Incitement and encouragement of harm
Descriptive attacks
Aggressive and emotive attacks
Dehumanization and overt sexual objectification
Causal use of gendered slurs, profanities and insults
Immutable gender stereotypes
Backhanded gendered complements
Condescending explanations or unwelcomed advise
Supporting mistreatment of individual women
Supporting systematic discrimination against women

Figure 1: Task Details

The problem includes **three** hierarchical sub-tasks as shown on figure 1 -

- **TASK A - Binary Sexism Detection**: a two-class (or binary) classification where systems have to predict whether a post is sexist or not sexist.
- **TASK B - Category of Sexism**: for posts that are sexist, a four-class classification where systems have to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussion.
- **TASK C - Fine-grained Vector of Sexism**: for posts that are sexist, an 11-class classification where systems have to predict one of 11 fine-grained vectors.

The basic input and output of the system are shown below in the Table 1. The input of the system is basically textual data in this case a comment or tweet from a user and the output is the predicted class of the text.

2.2 Dataset

For this research, we have used the labeled and unlabeled datasets provided by SemEval 2023 (Kirk et al., 2023). The given labeled dataset consists of 14,000 entries. And the unlabeled dataset provided consists of 2M data of which,1,00,000 are sampled from Gab and 1,00,000 from Reddit. Both the labeled and unlabeled datasets are in the English language. All 14,000 labeled data are first labeled by three trained annotators and disagreements are adjudicated by one of two experts. All annotators and experts were self-identifying women.

The training data consists of 14,000 entries (70% split), of which 3,398 are sexist. There is one CSV file containing labels for Tasks A, B, and C.

The problem also includes **Auxiliary Unlabelled Data** of 2M where -

- **Gab**: 1 million entries collected from publicly available Gab posts.
- **Reddit**: 1 million entries collected from sub-reddits comments.

Moreover, The development data consists of 2,000 entries (10% split). And the test data consists of 4,000 entries (20% split). Table 2 shows the detailed dataset distribution.

2.3 Related Works

In recent years, a lot of work has been put into the identification of hate speech, including tasks but few works have dealt with sexism detection and, in particular, they have dealt with sexism as the cause of hate speech. Specifically, a noticeable amount of work for hate speech detection from social media is done. In this section, we briefly describe some of the related works done on sexism detection and hate speech detection.

In the year 2020, Rodríguez-Sánchez et al. released the first **Spanish** dataset containing sexist tweets and named it as MeTwo. (Rodríguez-Sánchez et al., 2020). They validated their dataset containing 3600 tweets with various machine learning and deep learning techniques including logistic regression, Random Forest, SVM, BiLSTM, and also transformer models like BERT. Their model achieved the highest F1 Score of 0.64 using BiLSTM and Bert models. However, their dataset only included three classes- sexist, nonsexist, and doubtful. Another interesting work - detecting sexist memes on the web has been done in the year of 2019 by Fersini et al.(Fersini et al., 2019) that can detect sexism from the combination of pictorial and textual components. Their released dataset includes 800 memes divided broadly into two classes- sexist and nonsexist and categorizing the sexist content into - aggressive/non-aggressive and ironic/non-ironic content. The work also includes both uni-

Task no	**Input**	**Output**
A	Damn, this writing was pretty chaotic.	*Non-sexist*
	Leg day is easy. Hot girls who wear miniskirts get asked out.	*Sexist*
B	Finger her with your penis.	*threats, plans to harm and incitement*
	I went into PC world to get a laptop for my wife. Bastards wouldn't do the deal, and said they preferred cash.	*derogation*
	little girl, youre really the boss–come back tomorrow and find out how	*animosity*
	[URL] Women want equality until equality comes knocking their ass out. #Feminism #DamnPatriarchy	*prejudiced discussions*
C	I'd grab her and pull it's ass into my house and literally beat the shit out of her	*threats of harm*
	wish i had a girlfriend named alexa so i can force her to do stuff,sexist	*incitement and encouragement of harm*
	Girls are weird they secretly take your pictures and do all kinds of weird things but only other girls know about it usually	*descriptive attack*
	[USER] #HillaryHatesAmericans And she calls US deplorable? Please. Hillary is a disgusting	*aggressive and emotive attacks*
	Fucking disgusting you can see the kid trying to hold on single motherhood is a plague avoid that shit AT ALL COST	*dehumanising attacks & overt sexual objectification*
	Bad because its society that has to pick up the tab if it turns sour; Slut walks and single mothers comes to mind.	*casual use of gendered slurs, profanities, and insults*
	I don't know. I think girls are smart now days and know guys don't have as many options as girls	*immutable gender differences and gender stereotypes*
	How great is this pic of Duterte with some of his troops? I usually am not in favor of female soldiers but I may make an exception here	*backhanded gendered compliments*
	That's what happens little girl when you mix cocaine and milkshakes together.	*condescending explanations or unwelcome advice*
	All females must take collective responsibility for false rape accusations!	*supporting mistreatment of individual women*
	Child-bearing White women are around 2% of the world population. Bring back the patriarchy.	*supporting systemic discrimination against women as a group*

Table 1: Input and Output of the System

Class Label			**Data Amount**
Non-Sexist			10602
Sexist	Threats	Threats of harm	56
		Incitement and encouragement of harm	254
	Derogation	Descriptive attacks	717
		Aggressive and emotive attacks	673
		Dehumanizing and overt sexual objectification	200
	Animosity	Casual use of gendered slurs, profanities and insults	637
		Immutable gender differences and gender stereotypes	417
		Backhanded gendered compliments	64
		Condescending explanations or unwelcome advice	47
	Prejudiced discussions	Supporting mistreatment of individual women	75
		Supporting systemic discrimination against women as a group	258

Table 2: Data Distribution of labeled dataset

modal and multimodal approaches to detect sexism from textual content or pictorial content or from a combination of both. SVM, Naive Bayes, Decision Tree, and 1-Nearest Neighbor are used as baseline models where for the unimodal model they achieved the highest F1 Score of 0.841, and for the multimodal model, 0.744 is achieved.

If we look into other languages and research based on them, the first **Chinese** Sexism dataset named SWSR has been proposed by Jiang et al. in the year 2021 (Jiang et al., 2022) and includes the classes - I) Sexist, nonsexist, ii) stereotype based on appearance, stereotype based on cultural background, microaggression, and sexual offense. iii) individual target, generic target. Their exploratory analysis achieved the highest F1 score of 0.780 using the RoBerta model. Sexism detection from French tweets is done in 2020 by Chiril et al in the work (Chiril et al., 2020). They released an annotated corpus for **French** tweets which contained 12000 tweets broadly classified into sexist [Divided into directed, reporting, and descriptive classes], nonsexist, and no decision classes. They also experimented using this dataset with various deep learning techniques like - SVM, CNN, BiLSTM, and BERT models among which BERT showed the best F1 score of 0.762 in the binary classification task.

Research has also been conducted on sexist comments in the **English** language. In a work by Parikh et al.(Parikh et al., 2019) focus has been put on the multicategory classification of sexist content which is comprised of 23 categories of sexism and a total of 13023 data that are in the English language. The authors also proposed a novel architecture that includes BERT, ElMo, and Glove embedding and also attention-based BiLSTM models among which they achieved the highest micro F1 score of 0.718 in multiclass classification. Another work of the year 2021 by Samory et al. (Samory et al., 2021) is grounding sexism in social media on a psychological scale. They categorized their dataset into 4 classes based on psychological aspects and experimented with different models including Logistic regression, CNN, and BERT Finetuned, and got a 72% F1 score using the BERT. Moreover, studies on sexist and harassment detection have been done previously over the years using different Recurrent neural models such as LSTM, CNN(Bugueño and Mendoza, 2020), (Basu et al., 2021), Transformer model BERT(Yan and Luo, 2021), (Butt et al., 2021) and proven to give prominent results.

All the summary of the related works can be found in the table 3. Different studies have been conducted and it can be observed that models such as LSTM, BiLSTM, and BERT are some common models that have showcased good results in the case of performance. Moreover, we can see that there are some recent works done in sexism detection but most of them are confined to only the broader category of Sexism detection. However, our task comprises a fine-grained classification of sexist comments dividing the Sexist class into 4 classes and then sub-classifying these 4 classes into more 11 classes.

3 System Overview

3.1 Data Preparation

During the model-building process, we experimented with different methods on the training set to get the best result on the model which included data cleaning and preprocessing. However, as the input is sequential and preprocessing causes valuable information loss and the model showed poor results. Thus, directing us towards using the raw data from the dataset in our models without any preprocessing.

In this study, the dataset is split into a 90:10 ratio for the training and test set. The max input length is 55 collected from the dataset. So, we set the max input length to 60 with a padding length of 5.

3.2 Embedding

The pre-trained **GloVe: Global Vectors for Word Representation**(Pennington et al., 2014) embedding of 6B tokens and 100 dimensions of features is used in these tasks. GloVe embeddings can aid in capturing both semantic and syntactic connections between words in a phrase, hence enhancing precision. Focusing on the meaning and context of the word can make it more efficient, providing valuable information while training the models.

3.3 Active Learning(AL)

The number of sexist data was comparatively low and we had 14000 labeled data and 2M unlabeled data for the tasks. A large amount of data when fed into the Natural Language Processing model, produces finer results. That's why we decided to opt for an active learning approach and experimented with it. Active learning is the subset of machine learning in which a learning algorithm can query a user interactively to label data with the desired outputs. We have trained our active learning model with 14000 labeled data and 2,00,000 unlabeled data from Reddit and Gab using the active learning method.

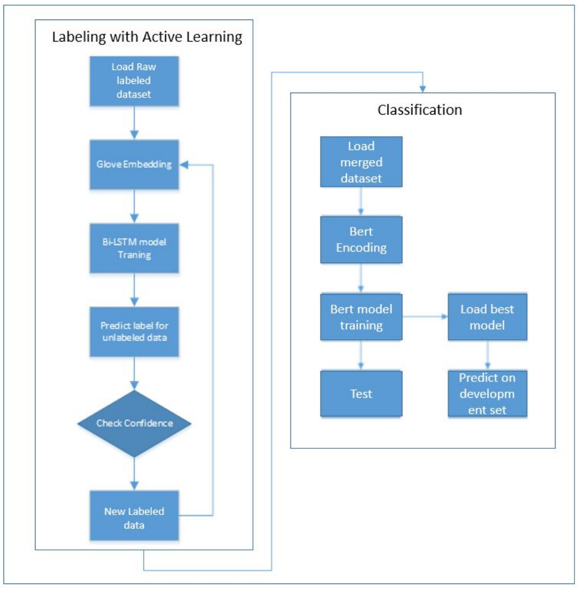


Figure 2: Active Learning Model Flowchart

For the active learning model, we used the BiLSTM method. First, a 0.3 dropout is added to the embedding layer then the BiLSTM method is used which had 128 output dimensions and a dropout of 0.2. Two dense layers were also added. After that, the max pool layer and normalization were used along with the Softmax activation function on the output layer. Adam optimizer with a learning rate of 0.001 is used in this model. Then the model is trained with batch size=32, epochs=20.

For unlabeled data, we used a confidence level of 0.85, 0.90, and 0.95 on the predicted data labels for Task A. This means if the confidence of a predicted label is equal or more than 90% then the predicted label will be taken and the unlabeled data will be labeled with the prediction. This newly labeled data is then merged with our labeled dataset for the main classification model. From this new dataset, we took only the sexist labeled data for Task B and used active learning on them again. For both tasks B and C, the confidence level of the predicted label is 65%. This process was again used for Task

Dataset	Language	No. of Classes	No. of Sub Classes of Sexist	Data Amount	Year	F1 Score	Ref.
MeTwo	Spanish	3	0	3600	2020	0.64	(Fersini et al., 2019)
MEME	English	2	4	800	2019	0.841	(Jiang et al., 2022)
SWSR	Chinese	2	4	8969	2021	0.780	(Chiril et al., 2020)
Chiril et. al	French	3	3	12000	2020	0.762	(Parikh et al., 2019)
Parikh et al.	English	2	23	13023	2019	0.718	(Samory et al., 2021)
Samory et al.	English	2	4	16000	2021	0.720	(edo)

Table 3: Related Works Summary

C. Table 4 gives an outlook of the result of active learning on the dataset. The total data represents the Active Learning labeled data in addition to the already given labeled dataset.

Task	AL Labeled data	Total Data
A	5592	19592
B	4901	8299
C	7660	11058

Table 4: Active Learning Results

Despite trying the active learning approach our experimental results proved that the raw dataset was better in terms of detecting the sexist data in comparison to the newly labeled dataset we prepared using active learning. Thus the experiment was continued using the raw dataset mentioned earlier.

3.4 Classification Model

Transformer works great with sequential data and in our case, we have used the pre-trained language model BERT(Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018) for the classification all of three tasks. For Task-A, Task-B, and Task-C, the Simple Transformers NLP library through the Transformers library by HuggingFace is used(Wolf et al., 2019). We have used a 'bert-large-uncased' pre-trained model consisting of 24-layer, 1024-hidden, 16-attention heads, totaling 336M parameters which were trained on raw labeled text. Our defined model contained a dropout rate of 0.2 with 2 dense layers, one with the ReLU activation function and the other with the softmax activation function. Adam optimizer was used with a learning rate of 0.0001. For training, the model used a batch size of 10 for all tasks and 10 epochs for the training of task A, task B, and task C. Prior to training the model, we used class weights in the loss function using sklearn's compute_class_weight method to combat the imbalance of data and avoid putting priority on any specific class. The training data set has been split by 10% testing and 90% training. We have used the early stopping technique if training accuracy reached 95% accuracy. After every epoch when the validation accuracy was increased, we saved the model. Next, we used the best-performing model to test each development test set for tasks A, B, and C. Google Colab was used to conduct all the experiments.

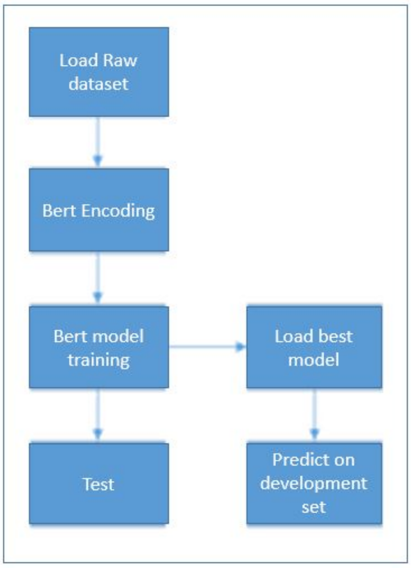


Figure 3: Final Model Flowchart

We have used this classification model for training on both actively learned labeled dataset and our raw labeled dataset. Then we validated our models with the development set provided. The active learning results can be found here in Table 2. As we saw that we achieved better performance in the raw labeled dataset, we finally selected the labeled dataset as our model dataset.

4 Results

If we compare the results based on the F1 score of Table 5 we can see the F1 score for task A is nearly the same for both the model using active learning labeled dataset and the model using raw data. However, we could see an improvement for Task B and Task C. As our model was performing

Task A

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Active Learning Labeled Data	Conf-85	0.87	0.82	0.83	**0.82**
	Conf-90	0.83	0.78	0.83	0.79
	Conf-95	0.87	0.83	0.81	**0.82**
Raw Data	BERT	0.87	0.82	0.82	**0.82**

Task B

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Active Learning Labeled Data	Conf-85	0.63	0.61	0.60	0.61
	Conf-90	0.59	0.58	0.61	0.58
	Conf-95	0.60	0.58	0.64	0.59
Raw Data	BERT	0.64	0.64	0.62	**0.63**

Task C

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Active Learning Labeled Data	Conf-85	0.48	0.45	0.41	**0.42**
	Conf-90	0.47	0.38	0.38	0.36
	Conf-95	0.49	0.41	0.38	0.39
Raw Data	BERT	0.54	0.44	0.42	**0.42**

Table 5: Result Comparison of Different Models on Development Set

poorly on task B and task C we decided to move forward with the model trained only on the raw labeled dataset.

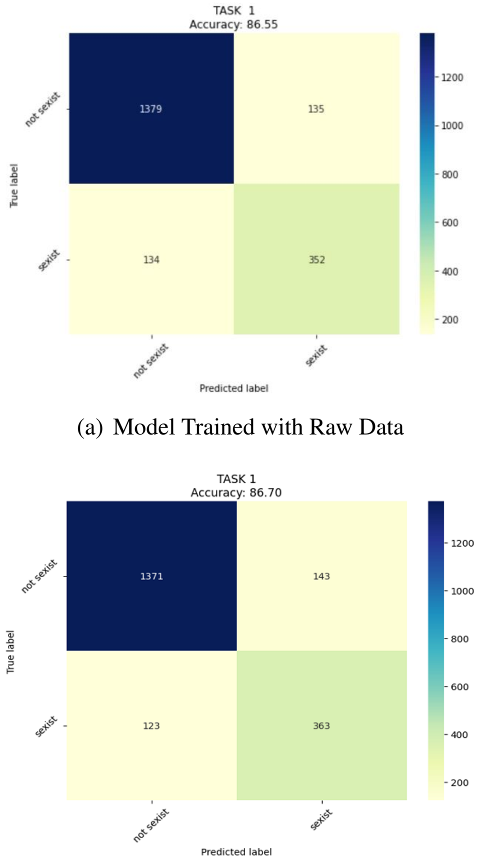
From the Table 5, we can see that the model gets an accuracy of 87% on the binary classification task of Task A where it gets an accuracy of 64% and 54% on multi-classification Task B and C respectively. The macro F1 Score for Task A and B also shows prominent results which are 82% and 63%. However, Task C where we performed fine-grained classification achieved a Macro F1 Score of 42% which is not satisfactory and can be improved.

We can see the detailed comparison of the result between the model trained on the raw dataset and the model trained on the active learning labeled dataset in the figures 4. Here, we compared the confusion matrix of the best-performing model among confidence scores 85, 90, and 95 using active learning labeled dataset and model using raw dataset. We could see from Table 5 that model with 85% confidence showed the best result among all of them. So we have compared the model trained on the raw dataset with it.

Figure 4 shows confusion matrices for Task A. Figure 4(a) shows that the model trained on the raw dataset predicts 1379 non-sexist and 352 sexist data correctly. However, figure 4(b) shows the model trained on the active learning labeled dataset predicted 1371 non-sexist and 363 sexist data correctly. Compared to the model trained with raw data it predicts the sexist data more correctly than the non-sexist data.

The results of Task B can be seen in figure 5 where in figure 5(a) the raw dataset trained model, the derogation class classifies 162 data correctly but misclassifies some into the animosity class. The animosity class also correctly predicted 94 data which is notable. The other two classes were mostly correctly classified. For active learning labeled data trained model in figure 5(b) shows that the derogation and animosity classes are mostly wrongly classified which are 43 and 61 respectively. Both of the models get confused between these two classes. However, the model using the active learning labeled dataset misclassifies every class item more than the model trained on the raw dataset.

From the confusion matrix shown in figure 6 for Task C in figure 6(a), we can see class-2, 3, 4, and 6 are somewhat classified correctly. Classes 1, 8, 9, and 11 are mostly misclassified. Significantly, the model fails to classify class 8, the "condescending explanations or unwelcome advice" class and classifies most of them to other classes. Moreover, class 9 'supporting mistreatment of individual women' classifies the data as another class instead of its own class label. We can see that in figure 6(b), the correct classification data number for the model using active learning labeled data is less than the



(a) Model Trained with Raw Data

(b) Model Trained with Active Learning Labeled Data

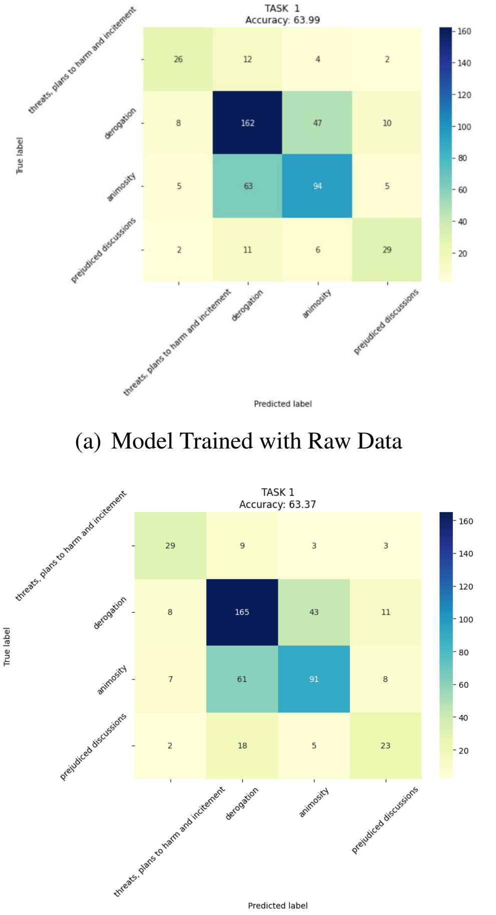
Figure 4: Task A Confusion Matrix

model using raw data. For example, for class 3 the raw data model predicts 65 correct data whereas the other predicts only 55 correct. It mostly misclassified them to class 7 and class 9. But the raw data model misclassified them into class 4 mostly. Therefore, it is seen that the model using the raw data performs well than the model using active learning labeled data in task C too. However, due to the fine-grained classification in Task C, our main model fails to distinguish them with better results leaving room for improvement.

Task	Macro F1 Score
A	0.8009
B	0.5849
C	0.4067

Table 6: Results on Test set

From the Table 6, we can see the current results on the Test set which shows better results on Task A and Task B than Task C which performs poorly.



(a) Model Trained with Raw Data

(b) Model Trained with Active Learning Labeled Data

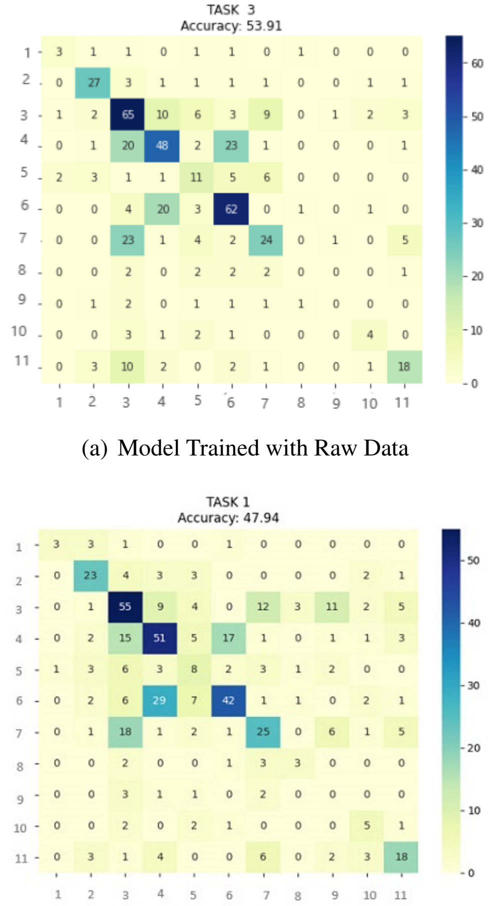
Figure 5: Task B Confusion Matrix

5 Conclusion

In conclusion, this paper describes our approach for the shared task on Explainable Detection of Online Sexism (EDOS) by SemEval 2023. A largely fine-grained classified labeled data imposed challenges that we tried to resolve using existing NLP models. Although there is room for improvement and further experiment, using a pre-trained multilingual-BERT model our approach obtained promising results. In the future, additional pre-processing steps might also improve the results along with further experimenting with different recurrent models and tuning the hyperparameters with different settings.

Limitations

The limitations of this paper include the need for further experimentation with different recurrent models and tuning hyperparameters with different settings. While the approach obtained promising results using a pre-trained multilingual-BERT model,



	1	2	3	4	5	6	7	8	9	10	11
1	3	1	1	0	1	1	0	1	0	0	0
2	0	27	3	1	1	1	1	0	0	1	1
3	1	2	65	10	6	3	9	0	1	2	3
4	0	1	20	48	2	23	1	0	0	0	1
5	2	3	1	1	11	5	6	0	0	0	0
6	0	0	4	20	3	62	0	1	0	1	0
7	0	0	23	1	4	2	24	0	1	0	5
8	0	0	2	0	2	2	2	0	0	0	1
9	0	1	2	0	1	1	1	1	0	0	0
10	0	0	3	1	2	1	0	0	0	4	0
11	0	3	10	2	0	2	1	0	0	1	18

(a) Model Trained with Raw Data

	1	2	3	4	5	6	7	8	9	10	11
1	3	3	1	0	0	1	0	0	0	0	0
2	0	23	4	3	3	0	0	0	0	2	1
3	0	1	55	9	4	0	12	3	11	2	5
4	0	2	15	51	5	17	1	0	1	1	3
5	1	3	6	3	8	2	3	1	2	0	0
6	0	2	6	29	7	42	1	1	0	2	1
7	0	1	18	1	2	1	25	0	6	1	5
8	0	0	2	0	0	1	3	3	0	0	0
9	0	0	3	1	1	0	2	0	0	0	0
10	0	0	2	0	2	1	0	0	0	5	1
11	0	3	1	4	0	0	6	0	2	3	18

(b) Model Trained with Active Learning Labeled Data

Figure 6: Task C Confusion Matrix

there is still room for improvement. Moreover, we have seen how our model does not produce promising results for Task B and specially for Task C. The fine-grained classified labeled data used in the shared task C may have restrictions that could affect how generalizable the results are to other datasets, however these limits are not discussed in the research.

Ethics Statement

We made use of an annotated dataset from the SemEval-2023(Kirk et al., 2023) Task 10 competition that was gathered in accordance with ethical standards. The remarks in the dataset were sexist, which is bad for the people who were being targeted. We made sure our classification algorithms weren't applied to expose or hurt those who were the targets of the sexist remarks in order to reduce the danger of harm. We used the BERT transformer model to classify and categorize the text, which is a widely accepted approach in the NLP community. We try to recognize the importance of addressing online sexism to create a more inclusive and equitable society. In order to improve awareness campaigns and put stop to this oppression, our study intends to help develop automated technologies that can help with the analysis of sexism. We envision our work as a modest first step toward achieving a more just and equal online environment for everyone.

References

Priyam Basu, Tiasa Singha Roy, Soham Tiwari, and Saksham Mehta. 2021. Cyberpolice: Classification of cyber sexual harassment. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 701–714. Springer.

Margarita Bugueño and Marcelo Mendoza. 2020. Learning to detect online harassment on twitter with the transformer. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 298–306. Springer.

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.

Center on Gender Equity and Health (GEH) – UC San Diego. 2023. When social media is sexist: A call to action against online gender-based violence. [February 21, 2023].

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in french tweets. In *Proceedings of the 12th language resources and evaluation conference*, pages 1397–1403.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maeve Duggan. 2017. Online harassment 2017.

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.

Rachael Fulper, Giovanni Luca Ciampaglia, Emilio Ferrara, Y Ahn, Alessandro Flammini, Filippo Menczer, Bryce Lewis, and Kehontas Rowe. 2014. Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, pages 57–64.

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

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

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*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584.

Wikipedia contributors. 2023. Sexism — Wikipedia, the free encyclopedia. [Online; accessed 21-February-2023].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Mingrui Yan and Xudong Luo. 2021. Bert-based detection of sexual harassment in dialogues. In *2021 5th International Conference on Computer Science and Artificial Intelligence*, pages 359–364.