PoSh at SemEval-2023 Task 10: Explainable Detection of Online Sexism

Shruti Sriram, Padma Pooja Chandran and Shrijith M R
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
shruti2010197@ssn.edu.in, padmapooja2010291@ssn.edu.in, shrijith2010542@ssn.edu.in

Abstract

The growing popularity of online platforms and social media has intensified the problem of sexism, making it more prevalent and widespread. Online platforms have provided an anonymous space where individuals can freely express their misogynistic beliefs with discriminatory and abusive language, creating a culture of online harassment and hate speech, mainly targeting women. The dataset used in this task was provided by SemEval-2023 Task 10: Explainable Detection of Online Sexism. To precisely identify the different forms of online sexism, we utilize several sentence transformer models such as ALBERT, BERT, RoBERTa, DistilBERT, and XLNet. By combining the predictions from these models, we can generate a more comprehensive and improved result. Each transformer model is trained after pre-processing the data from the training dataset. Our team obtained macro f1 scores of 0.7937 for subtask A, 0.5284 for subtask B and 0.2674 for subtask C.

1 Introduction

In today's technology-driven world, it has become increasingly common for users to utilize social media platforms to spread hatred and engage in sexist behaviour. Therefore, addressing online sexism is a critical way to create a safe online space for everyone, where people can express their opinions without fear of discrimination or harassment. Online sexism is a form of gender-based discrimination or harassment that occurs in online spaces, including social media platforms, online forums, chat rooms, and messaging apps. It can take many forms, from overly sexist comments, hashtags (Fox et al., 2015) and slurs in the forms of discrimination, such as dehumanizing language, gender-based stereotypes, and microaggressions. Further, it can have a significant impact on the mental and emotional status of individuals who experience it. Also, it can lead to feelings of shame, anxiety, and depression, as well as physical symptoms such as headaches and insomnia. Online sexism can be directed at individuals or groups of people based on their gender, sexual orientation, race, religion, or other personal characteristics. Women are particularly vulnerable to online sexism as they are more likely to be targeted with derogatory comments and threats of violence, and overall experience higher levels of online harassment. Given the rising issue of online sexism, it is essential to develop effective strategies for detecting and addressing the same. One promising approach is to use natural language processing (NLP) transformers to analyze social media data and identify instances of online sexism.

We have participated and submitted our solutions for all three subtasks. The goal of these tasks is to enhance the accuracy and explainability of English-language models for detecting online sexism. The main focus is to create fine-grained classifications for sexist content on Gab and Reddit (based on training data). By developing more accurate and explainable models, we can improve our ability to identify, address and prohibit online sexism, creating safer and more inclusive online exposure. Our main strategy for detecting online sexism was to utilize various NLP sentence transformers, including ALBERT, BERT, RoBERTa, DistilBERT, and XLNet. The first step of our approach involved cleaning the dataset of all unwanted characters. After filtering out the data, it was used as input for each of the transformer models. To generate comprehensive and accurate results, we combined each of the predictions from all the transformer models. This enabled us to identify instances of online sexism with greater accuracy and provide a more nuanced understanding of its different forms.

Overall, participating in this task highlights the need for continued research and development of

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1276–1281
July 13-14, 2023 ©2023 Association for Computational Linguistics

automated tools to effectively detect the slightest forms of online sexism. Our paper is structured as follows: In section 2, we provide detailed Background work on the problem of online sexism and existing research on the topic. In section 3, we describe the NLP sentence transformer models we utilized in our approach. Section 4 presents the experimental setup we used to evaluate the performance of our models, also providing a comparative analysis of our results and discussing the strengths and limitations of our approach. Finally, in section 5, we conclude our work by summarizing our findings and highlighting the implications of our work.

2 Background

2.1 Task Description

The task (Kirk et al., 2023) consists of three hierarchical subtasks. Task A is a binary classification problem where the goal is to predict whether a post is sexist or not sexist. Task B is a four-class classification problem where the system has to further classify sexist posts as (1) threats, (2) derogation, (3) animosity, or (4) prejudiced discussions. The goal of task C is to classify sexist posts as one of the 11-fine grained vectors. A common training dataset contains rewire_id, text, and label_sexist for task A, label_category for task B and label_vector for task C. The training dataset contains 14,000 rows of data. The development and test dataset for each task has rewire_id and text. For task A there were 2,000 rows of development and 4,000 rows of test data and for tasks B and C, 486 rows of data were for development and 970 for testing. All the data was in English. The final submission file contains rewire_id and label_pred which is the label predicted by the model for the corresponding text.

2.2 Related Work

Istaiteh et al. (Istaiteh et al., 2020) aim to provide a survey of sexist and racist hate speech detection approaches focusing on three different aspects; namely, available datasets, features exploited, and machine learning models. Firstly, an overview of the available datasets that researchers can utilize to determine whether to use a pre-labelled dataset or to create their own dataset through labelling. Secondly, discuss the significance of selecting appropriate features for classification. Proper feature representation is a crucial aspect of all NLP tasks. Finally, selecting the appropriate machine learning algorithm is equally important and must be carefully considered alongside the above-mentioned aspects.

Senn et al. (Senn et al., 2022) explore the effectiveness of different BERT models and ensembles in classifying depression from transcripts of clinical interviews. AudiBERT was used for depression classification from audio recordings, but the ablation study shows that BERT was the most influential component. RoBERTa has demonstrated better performance than BERT, and ensembles of models have produced more robust classifications. DistilBERT was utilized in combination with various ensemble techniques. The study aims to show the hypothesis that ensembling multiple BERT variants improve performance.

Mina et al. (Mina et al., 2021) proposed an approach for the task of sexism detection as both a coarse (binary) classification problem and a fine-grained classification task that involves identifying various types of sexist content such as dominance, stereotyping, and objectification. To solve the task (of detecting sexist content), two multilingual transformer models were utilized - one based on multilingual BERT and the other on XLM-R. The approach employed two different strategies to adapt the transformers for detecting sexist content. The first strategy was unsupervised pre-training with additional data, and the second was supervised fine-tuning with augmented and additional data.

Liakhovets et al. (Liakhovets et al., 2022) utilized two multilingual transformer models and a monolingual T5 model to solve two related tasks(similar to that of the previous paper). One transformer model was based on a multilingual BERT architecture, while the other was based on an XLM-RoBERTa architecture. The transformers were adapted to detect sexist content using two strategies: unsupervised pre-training with additional data and supervised fine-tuning with additional and augmented data. The XLM-RoBERTa model, which applied a combination of both strategies, outperformed the other two models for both tasks.

De Paula et al. (de Paula and da Silva, 2022) proposed a methodology for identifying and classifying sexist content in social media posts using various transformer architectures. Detecting and identifying sexist content is a challenging problem due to characteristics such as sarcasm and multiple forms of sexism. The authors evaluated single-language and multilingual versions of BERT, RoBERTa, and single-language versions of Electra and GPT2 architectures for two tasks, namely sexism detection and classification in both English and Spanish languages.

3 System Overview

3.1 BERT

BERT (Devlin et al., 2018) is a neural network architecture composed of multiple layers of transformers. The encoder is a stack of transformer blocks that processes text in a bidirectional manner and generates contextualized word representations. There are two sub-layers in each transformer block. The sub-layers are followed by a residual connection and a layer normalization step.

3.2 ALBERT base v1

A Lite BERT (ALBERT) (Lan et al., 2019) architecture has significantly fewer parameters as compared to the traditional BERT architecture. Additionally, in order to overcome the challenges of scaling pre-trained models in NLP, ALBERT contains two-parameter reduction approaches. In a nutshell, ALBERT configurations have fewer parameters compared to BERT-large but achieve significantly better performance.

3.3 RoBERTa-base

RoBERTa (Robustly Optimized BERT Approach) (Liu et al., 2019) is a BERT model variant. It is pre-trained on a much larger amount of data than BERT, with a larger batch and context size and a longer training time. To prevent over-fitting, it also employs dynamic masking during pre-training to randomly mask tokens during each training epoch. RoBERTa-Base has 12 transformer layers, each with 12 attention heads and a hidden size of 768 and 125 million parameters.

3.4 DistilBERT-base-cased

DistilBERT (Sanh et al., 2019) is a transformer-based language model with the same architecture as BERT but fewer layers and parameters. The model is trained on both uppercase and lowercase text versions. It has six layers and a total of 110 million parameters. Distillation is used to transfer the knowledge learned by a larger model, followed by pruning to remove the less important parameters, to reduce the size and improve performance.

3.5 XLNet-base-cased

XLNet-base-cased (Yang et al., 2019) is a transformer-based language model with 116 million parameters that was pre-trained using an auto-regressive method. There are 768 hidden units and 12 attention heads in each encoder layer. In addition, XLNet employs permutation language modelling to forecast the probability distribution of various input sequence permutations, followed by a task-specific output layer.

4 Experimental Setup

4.1 Dataset Analysis

The labelled dataset provided by the organizers consists of 20,000 entries in total. Of these, 10,000 are sampled from Gab and 10,000 from Reddit. This dataset is further divided into training, development, and test set. The training data consists of 14,000 entries (70% split), of which 3,398 are labelled 'sexist'. The development data consists of 2,000 entries (10% split) and the test data consists of 4,000 entries (20% split).

4.2 Data Pre-processing

The given dataset of online posts is highly noisy and contains excessive use of punctuation, URLs, symbols, and misspelt words, thus pre-processing the text is essential. The NLTK toolkit, abbreviated as the Natural Language Toolkit, is used to process the data. It offers a variety of text-processing libraries for tokenization, parsing, classification, semantic reasoning, etc. The data is altered in the ways listed below:

- Stopwords are removed.
- Hashtags, HTML tags, mentions, and URLs are removed.
- Emoticons and other symbols are removed.
- The contractions are expanded and the text is lemmatized.
- Special characters are removed.

- Extra white spaces are reduced.

The Regexp() module and the RegexpTokenizer() method are used to extract the tokens from the string (tokenization). Tokenizing is a crucial step when it comes to cleaning the text. It is employed to split the text into words or sentences, dividing it into more manageable pieces that may be understood independently of the remainder of the text. To analyze the text, we must tokenize each word and each sentence separately. Unstructured data is converted in this way to structured data, which is simpler to examine.

4.3 Building the model

We implemented ALBERT transformer model for task A. The model was trained for 3 epochs with a batch size of 16 and a learning rate of 1e-4.
We explored other NLP sentence transformers like BERT, RoBERTa, DistilBERT and XLNet for task B. But the performance significantly improved when the collective output of the above-mentioned transformer models was used. Subsequently, BERT, RoBERTa, DistilBERT, and XLNet were trained for 3 epochs with a batch size of 16 and a learning rate of 1e-4. Then majority voting was used to determine the final predicted class for that text.

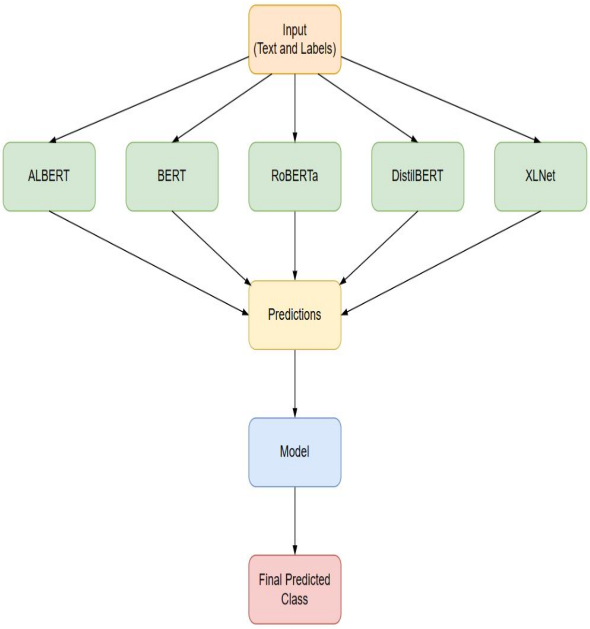

Figure 1: Ensemble model for task C

For task C, BERT, ALBERT, RoBERTa, DistilBERT, and XLNet were trained for 3 epochs with a batch size of 16 and a learning rate of 1e-4. Then the mode of the predictions of all the models for a particular text was calculated and compared with ALBERT. The ensemble gave a better Macro f1 score.

5 Results

This section presents the analysis of the results and submitted official results.

Model	Macro f1
ALBERT	0.754762

Table 1: Development results of task A

For task A, our final submission was the predictions made by ALBERT for the test dataset.

Model	Macro f1
ALBERT	0.508084
BERT	0.578587
RoBERTa	0.549079
DistilBERT	0.559483
XLNet	0.537929
Ensemble model	0.580871

Table 2: Development results of task B

We explored several transformer models individually for task B. Yet the performance increased when the combined output of the different transformer models was used. Therefore our final submission for task B was the collective output of a few transformer models like BERT, RoBERTa, DistilBERT, and XLNet.
Similarly for task C, the collective output of the models ALBERT, BERT, RoBERTa, DistilBERT, and XLNet was submitted.

Model	Macro f1
ALBERT	0.279955
BERT	0.297759
RoBERTa	0.308759
DistilBERT	0.301552
XLNet	0.290285
Ensemble model	0.312036

Table 3: Development results of task C

Finally, the confusion matrices were plotted to visualize and summarise the performance of our model.

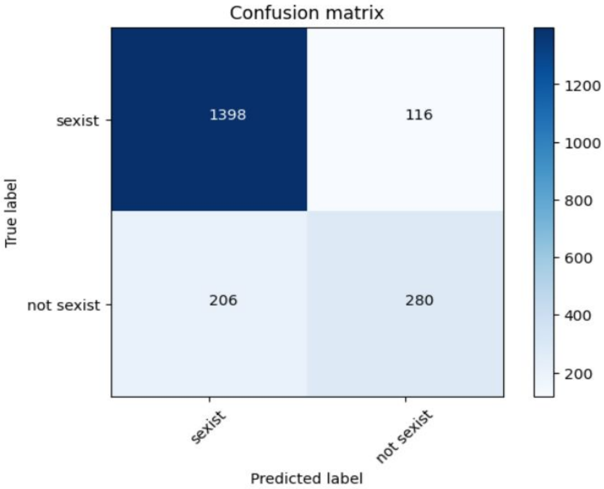

Figure 2: Confusion matrix for task A

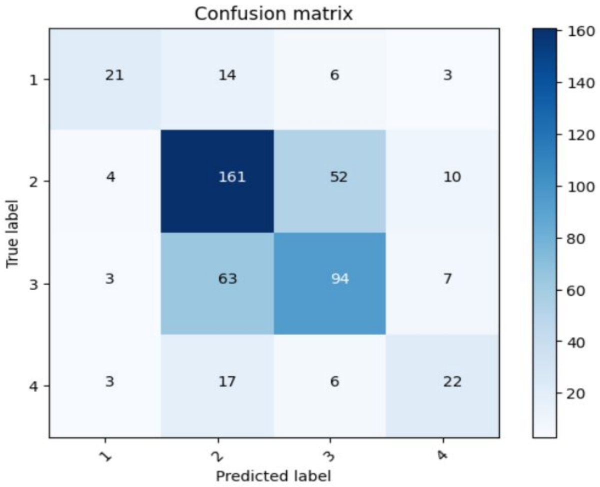

Figure 3: Confusion matrix for task B

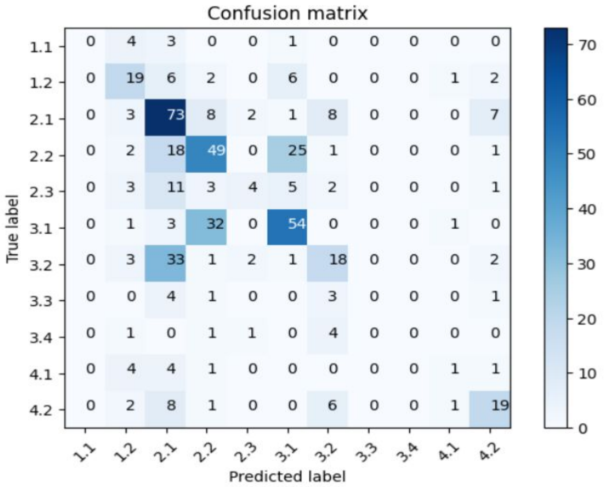

Figure 4: Confusion matrix for task C

Further, we tried a different ensemble and compared the results with the previous approach that employed majority voting. In the new system, we obtained the training accuracies of the models used in the ensemble, and for each text, we calculated the sum of the accuracies of models that predicted the given class. The final predicted class was the one with the maximum sum. It was observed that the second approach produced a comparable f1 score for task B (0.573080) and a better f1 score for task C (0.445315).

6 Conclusion

This paper presents the submitted runs for the task 'Explainable Detection of Online Sexism (EDOS)' in SemEval 2023. We experimented with different approaches such as a BERT model, ALBERT, RoBERTa base, DistilBERT, and XLNet and Ensemble models. Our team's submission had macro f1 scores of 0.7937 for subtask A, 0.5284 for subtask B and 0.2674 for subtask C. Our team placed 66th on the leaderboard for task A, 57th for task B and 53rd for task C. For future work, we will explore other ensembling techniques like averaging the probabilities of the models. Additionally, we will experiment with external resources and employ different data augmentation techniques to enhance the performance of our model.

Acknowledgements

We thank the CSE department of SSN College of Engineering for helping us implement this task.

References

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. *arXiv preprint arXiv:1810.04805*.

Jesse Fox, Carlos Cruz, and Ji Young Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in human behavior*, 52:436–442.

Othman Istaiteh, Razan Al-Omoush, and Sara Tedmori. 2020. Racist and sexist hate speech detection: Literature review. In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 95–99.

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. *arXiv preprint arXiv:1909.11942*.

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. *arXiv preprint arXiv:1907.11692*.

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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Saskia Senn, ML Tlachac, Ricardo Flores, and Elke Rundensteiner. 2022. Ensembles of bert for depression classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 4691–4694.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.