KingsmanTrio at SemEval-2023 Task 10: Analyzing the Effectiveness of Transfer Learning Models for Explainable Online Sexism Detection

Fareen Tasneem, Tashin Hossain, and Jannatun Naim
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{fareen.tasneem, tashin.hossain.cu, and
jannatun.naim.cu}@gmail.com

Abstract

Online social platforms are now propagating sexist content endangering the involvement and inclusion of women on these platforms. Sexism refers to hostility, bigotry, or discrimination based on gender, typically against women. The proliferation of such notions deters women from engaging in social media spontaneously. Hence, detecting sexist content is critical to ensure a safe online platform where women can participate without the fear of being a target of sexism. This paper describes our participation in subtask A of SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS). This subtask requires classifying textual content as sexist or not sexist. We incorporate a RoBERTa-based architecture and further fine-tune the hyperparameters to entail better performance. The procured results depict the competitive performance of our approach among the other participants.

1 Introduction

The free flow of information is the most fascinating feature of social media that has enabled sharing of information, thoughts, and opinions across the world. But this very feature is now used as a tool to incite hatred and threats against individuals or communities. Our society has always exploited women with prejudice, stereotype, and discrimination. The creation and proliferation of social media have provided our society with a more extensive platform to induce such unjust treatment against women. People can post sexist content anonymously using fake and untraceable identities without the fear of accountability. Moreover, the hatred can be spread to the wider community impeding the integration of women into the social network. So it is essential to detect such sexism in social media content so that women can partake and collaborate in social media without being mistreated and threatened. With this intention, SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) is introduced. The aim of the task is to detect sexism in the textual content of social media. The task is divided into three hierarchical tasks. We participated in the subtask A - Binary Sexism Detection where a text must be classified as sexist or not-sexist. To explicate this task, two examples are shown in 1.

Sample texts	Class
#1: Im beginning to see why women werent allowed in politics or allowed to vote.	Sexist
#2: Kaine is so creepy He looks like a raging alcoholic wife beater Ick	Not-sexist

Table 1: Sexist and Non-sexist samples.

Here, in the first example, the capability of women in political activities is being doubted just based on their sex. Hence, the first text is considered sexist. On the other hand, the second example depicts violence toward women but doesn't encourage or support it. So the second text is not sexist.

Detecting sexism in social media has been of great concern for researchers. Numerous works have been conducted on detecting sexism in texts. For instance, (Parikh et al., 2019) employed LSTM and CNN-based architecture along with several word embeddings including ELMo, GloVe, and fastText, and also sentence embedding incorporating BERT, Universal Sentence Encoder (USE), and InferSent. Convolutional filters were explored by (Sharifirad et al., 2019). to extract the most important n-grams in a specific category of sexism (i.e., Indirect sexism, Sexual sexism, and Physical sexism). They filtered out the irrelevant n-grams and clustered the significant ones to enhance the

All of the authors have equal contributions.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1916–1920

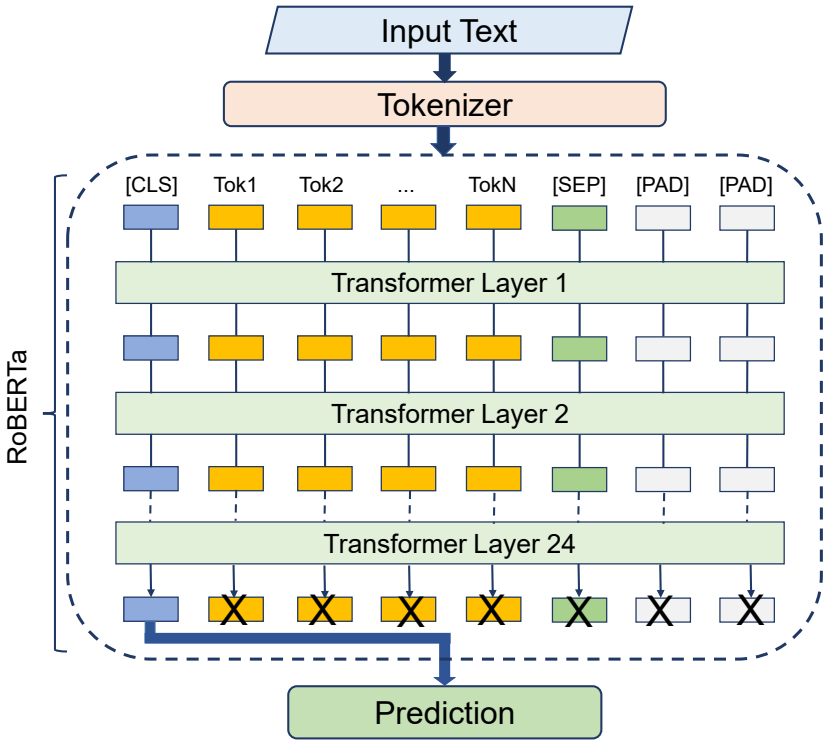

Figure 1: Proposed Framework.

performance. (Grosz and Conde-Cespedes, 2020) worked on a specialized Twitter dataset that focuses on workplace sexism. They employed GloVe and random word embeddings along with attention-based LSTMs. Moreover, (Samory et al., 2021) discuss how most of the previous works deal with explicit expressions of sexism. So when faced with implicit sexism and adversarial sample, most of the proposed models don't work well. Transformer models such as XLM-R and mBERT were also proposed (Schütz et al., 2021). (Chiril et al., 2021) leveraged the problem of detecting gender stereotypes to entail better outcomes in the detection of sexism. On the other hand, (Vaca-Serrano, 2022) employed an ensemble of finetuned DeBERTa, RoBERTa, and BERTweet for detecting sexism in English tweets.

In this paper, we elucidate our findings achieved from our investigation on this task. We propose a RoBERTa-based architecture and further hyperparameter tuning to achieve an enhanced result. The organization of this paper is as follows: we illustrate our proposed methodology in Section 2. Section 3 elaborates on the experimental details. We also incorporate our study on the performance of various approaches and error analysis in Section 4. Finally, we culminate this paper with some future aspects in Section 5.

2 Proposed Framework

For the aim of detecting explainable online sexism, we described our tested models in this part. Various transfer learning and deep learning models were explored in this aspect.

RoBERTa-Sexism-Classifier: The RoBERTa-Sexism-Classifier is a well-liked and adaptable architecture that makes use of an enhanced bidirectional encoder mechanism. It can understand more complex intricate relationships between words and phrases because it was trained on a broader and more diverse corpus. We fine-tuned the RoBERTa pre-trained transformer-based architecture with the explainable online sexism dataset to achieve our goal. The overall fine-tuned process is depicted in Figure 1.

An input text is initially tokenized using the RoBERTaTokenizer (Wolf et al., 2019). All of the tokens were padded to the same sequence length. It is then fed to the first transformer layer, and the output of that layer is fed to the second. This procedure continues until the transformer layer 24's output is obtained. We used the RoBERTaForSequenceClassification (Liu et al., 2019) technique to obtain the final prediction.

3 Experiments and Evaluations

3.1 Dataset Description

We utilized the dataset proposed on the EDOS task (Kirk et al., 2023) in SemeEval-2023. The dataset has a total of 20000 entries gathered from Gab and Reddit (10000 each). Three annotators label the gathered entries, and the expert reviews any disagreements in entries. The task is broken down into three smaller tasks. In this paper, we only concentrated on Task A. They separated the 20000 entries into three subsets, with the training, validation, and testing sets consisting of 14000, 2000, and 4000 text entries, respectively. The dataset is used for the training model to determine whether or not a post is sexist. Moreover, we discovered that only 24.3 percent of the data in our dataset has sexist and not sexist labels after evaluating the distribution from Table 2, containing sexist and not sexist labels. The Macro F1-Score is the key assessment statistic utilized for this task.

Category	Train	Dev	Total
Sexist	3398	486	3884
Not Sexist	10602	1514	12116
Total	14000	2000	16000

Table 2: The quantitative properties of the dataset for Task A.

3.2 Experimental Setup

We proposed the RoBERTa-Sexism-Classifier as our final KingsmanTrio system since finetuning its pre-trained embedding layers helps predict better than our other evaluated approaches. Table 3 shows the experimental setup for our best-scoring system.

3.3 Results Analysis

We now evaluate our system's performance against that of our other participants' systems. Table 4 shows the performance of the 84 valid submissions with that of the top-performing teams.
It shows that our system stands out among the participants' systems in a competitive manner. It narrowly falls short of PingAnLifeInsurance, the top-performing team, by 3.8%.

System	Settings
RoBERTa-Sexism Classifier	*1. Tokenizer*: roberta-large *2. Model*: roberta-large *3. Optimizer*: AdamW *4. Learning_rate*: 2e-5 *5. Epsilon*: 1e-8 *6. Epochs*: 10 *7. Batch Size*: 32

Table 3: System settings.

Team_Name	Rank	Score
PingAnLifeInsurance	01	0.8746
PALI	04	0.8717
Zhegu	06	0.8674
Aston NLP	14	0.8583
KingsmanTrio	39	0.8366
msharma95	48	0.8230
tsingriver	57	0.8180
PadmaDS	68	0.7826
OPEN SESAME	81	0.6044

Table 4: Analysis of the participating systems' comparative performance.

4 Discussion

We explored numerous deep learning and transfer learning models in order to estimate and identify the most effective strategy. In Table 5, several model performances are displayed based on the development set.

Method	Macro F1-Score
KingsmanTrio	0.8500
Efficacy of other tried approaches	
Finetuned BERT+RoBERTa	0.8439
Finetuned BERT	0.8010
Bi-LSTM+GloVe	0.7501
Dense model	0.7130
LSTM	0.6900
CNN	0.6930
GRU	0.4370

Table 5: Performance of individual models on the development set.

In Table 5, the analysis of the model's performance shows that transfer learning models outper-

Text	Gold Label	Predicted Label
1. How the hell do you accidentally decapitate a baby? What that bitch did was stone-cold murder.	Sexist	Sexist
2. I wouldn't receive that treatment as a woman. You also couldn't talk shit about women the way they do, and get off scot-free. You'd be called a chauvinist asshole, at best.	Sexist	Not Sexist
3. There are only TWO genders & I suggest the only thing flowing will be 2 ltr bottles of coke down her throat.	Not Sexist	Sexist
4. Not common these days. They just don't have an interest in it, even my female housemate said that.	Not Sexist	Not Sexist

Table 6: Examples of misclassified text for sexism classification.

form deep learning models. Among the other tried approaches in Finetuned BERT+RoBERTa, we consider an ensemble of the labels to be not-sexist if it is predicted by both of the individual model. However, the ensemble does not bring a good outcome though.

Moreover, Models of transfer learning made use of prior knowledge to produce universal features and patterns. Transfer learning models can exceed deep learning models which are trained from scratch on smaller datasets by making use of this general knowledge. Yet, because transfer learning models begin with pre-trained weights, they can converge more quickly during training, enabling the model to pick up knowledge about a given task quickly. Also, we can observe that the RoBERTa model outperforms the BERT model. It might be a result of RoBERTa's training on a significantly larger corpus of text data than BERT, which likely contained more diverse and high-quality content. As a result, RoBERTa can develop more accurate representations of the language, which will help it perform better on downstream tasks.

We look into the cause of our suggested system's incorrect predicted labels further. Table 6 provides several instances in this respect.

Our proposed model was able to identify the fact that the first example used derogatory language against women and the use of the word "bitch" is a form of gender-based harassment towards women, which contributes to a culture of sexism and misogyny. However, in the second example, the text makes a generalization using a double standard meaning that women are not treated equally when it comes to speaking about men versus women. Here, the term "chauvinist asshole" further reinforces the sexist nature of the text. But, because our system was unable to comprehend the significance of this double standard, it made a false forecast. Even though the third example has the gold label "Not Sexist," our methodology predicts "Sexist" due to its rude and demeaning character towards a certain gender. The term "only TWO genders" can be seen as being dismissive and exclusionary since it suggests that non-binary people do not exist. Also violent and unpleasant is the idea that someone is forced to drink entire bottles of coke. To elaborate further, we can say that our method failed to detect 123 out of 486 sexist posts. This may be attributed to the fact that there is an insufficient number of sexist posts in the training dataset.

5 Conclusion and Future Notion

In this paper, we traversed several transformer-based architectures and deep learning models. We further conducted a performance analysis of these models and investigated the mispredictions of the top-performing model. In the future, we intend to incorporate domain-specific features along with other transformer architectures. Moreover, as social media content are not limited to text only, we aspire to detect sexism in other modalities too.

References

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844.

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2020 Workshops, DSFN, GII, BDM, LDRC and LBD, Singapore, May 11–14, 2020, Revised Selected Papers 24*, pages 104–115. Springer.

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*.

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*.

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.

Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepcevic, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2021. Automatic sexism detection with multilingual transformer models ait fhstp@ exist2021. In *IberLEF@ SEPLN*, pages 346–355.

Sima Sharifirad, Alon Jacovi, Israel Bar Ilan Univesity, and Stan Matwin. 2019. Learning and understanding different categories of sexism using convolutional neural network's filters. In *WNLP@ ACL*, pages 21–23.

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

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*.