AU_NLP at SemEval-2023 Task 10: Explainable Detection of Online Sexism Using Fine-tuned RoBERTa

Amit Das[1], **Nilanjana Raychawdhary**[1], **Tathagata Bhattacharya**[2],
Gerry Dozier[1], **and Cheryl D. Seals**[1]

[1]Department of Computer Science & Software Engineering, Auburn University, AL, USA
[2]Department of Computer Science, Auburn University at Montgomery, AL, USA
[1]{azd0123,nzr0044,doziegv,sealscd}@auburn.edu
[2]{tbhatta1}@aum.edu

Abstract

Social media is a concept developed to link people and make the globe smaller. But it has recently developed into a center for sexist posts that target especially women. As a result, there are more events of hostile actions and harassing remarks present online. In this paper, we introduce our system for the task of Explainable Detection of Online Sexism (EDOS), a part of SemEval 2023 task 10. We introduce a fine-tuned RoBERTa model by encoding the initial representation of the data using Roberta and setting three distinct Multilayer Perceptrons (MLPs) corresponding to the three sub-tasks to address this specific problem. The effectiveness of the proposed strategy is demonstrated by the experimental results reported in this research.

1 Introduction

Discriminatory ideas and sentiments, particularly those directed at women, are frequently found online and on social media. This typically indicates the presence of additional hazardous content types like hate speech (Demus et al., 2022) or misinformation (Schütz et al., 2021). It might be difficult to identify such remarks because sexism and misogyny can take many different forms and vary across linguistic and cultural boundaries.

Sexism is one of the most critical difficulties because it may greatly harm users. As an illustration, objectification, a kind of sexism, has been linked to serious risks for eating disorders, unipolar depression, and sexual dysfunction (Fredrickson and Roberts, 1997). A study published by Fox et al. (Fox et al., 2015) requested participants to retweet or post tweets with sexist content before conducting a set of activities meant to uncover sexist behaviors. Additionally, since the anonymity of Twitter can encourage people to demonstrate an even greater sexist behavior, the authors came to the conclusion that the users who were protected by the anonymity of a social media profile were more likely than non-anonymous users to exhibit aggressive sexism.

Sexism is defined by the Oxford English Dictionary as *prejudice, stereotyping or discrimination, typically against women, on the basis of sex.*[1] Sexist attitudes and languages undervalue the contribution of women. Furthermore, considering that a sizeable portion of Internet users, particularly those who utilize social networks, are teenagers, the rise in online sexism necessitates urgent research and social debate that results in action (Rodríguez-Sánchez et al., 2020).

However, detecting online sexism may be difficult, as it may be expressed in very different forms. In this study, we focus mainly on using Natural Language Processing (NLP) methods and state-of-the-art models for two important tasks: (i) binary sexism detection, which aims to determine whether a given sentence contains any sexist content; and (ii) fine-grained sexism classification, which aims to further identify which class a sexist sentence belongs to.

In the specific context of this work, we seek to solve the issue of sexism identification and classification in social media posts using a fully autonomous, NLP technique. This is a pressing issue that has not been adequately addressed previously (Fortuna et al., 2021; de Paula et al., 2021; Rodríguez-Sánchez et al., 2022), and the suggested solution is scalable and does not rely on inputs from humans directly.

The structure of this work is as follows: Section 2 lists some significant research in the detection and classification of sexism; Section 3 describes the task; Section 4 describes the details of the dataset used here; Section 5 describes the fundamental architecture of transformer models and how they are used for identifying and categorizing sexism as well as the key steps of the methodology used; Sec-

[1] https://www.oed.com/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 707–717
July 13-14, 2023 ©2023 Association for Computational Linguistics

tion 6 contains the primary findings and analyses of this study and in Section 7 the work is concluded, providing suggestions for future works.

2 Related work

A lot of work has been done in recent years to identify hate speech, including tasks like identifying racist or xenophobic content, but very few of these studies have addressed sexism detection, and in particular, they have dealt with sexism as the identification of hate speech against women (Rodríguez-Sánchez et al., 2020). However, several concepts and methods from the identification of hate speech may be applicable to our issue (Rodríguez-Sánchez et al., 2020). So, in this Section, we briefly discuss related research in the realm of hate speech as well as earlier studies on the identification of sexism and misogyny.

There have been several approaches to classify hate speech spreaders online. Bag-of-words (BOW) techniques were the foundation for the first efforts on hate speech identification (Davidson et al., 2017; Waseem, 2016; Waseem and Hovy, 2016). One of the earliest studies, published in 2012, used machine learning-based classifiers (Xiang et al., 2012) rather than pattern-based techniques (Gianfortoni et al., 2011) for detecting abusive language. To identify hate speech, many traditional machine learning techniques have been used, including decision trees (Davidson et al., 2017), logistic regression (Waseem, 2016), and support vector machines. Recently Das et al. (Das et al., 2022) used BERT-TFIDF based approach for profiling irony and stereotype spreading authors. There are various methods that use specific types of sentiment data as features.

Peter Burnap and colleagues utilized a dictionary-based method to find cyberhatred on Twitter. They employed an N-gram feature engineering method to create the numeric vectors using a specified vocabulary of offensive words. Njagi Dennis et al. (Gitari et al., 2015) used an ML-based classifier to categorize hate speech in internet forums and blogs. The authors choose to build the master feature vector using a dictionary-based method. The use of emotive language as well as semantic and subjective components were influenced by the focus on hate speech (William et al., 2022). The resulting feature vector was then input to a rule-based classifier.

Misogyny typically connotes the display of anger and hatred toward women, albeit it is not always the same as sexism (Pamungkas et al., 2020). The term 'expressions of hate towards women' is misogyny (Ussher, 2016), although the term 'sexism' also refers to more subtly veiled implicit forms of abuse and prejudice that can nevertheless significantly affect women. When defining sexism, Glick and Fiske (Glick and Fiske, 2001) distinguish between two types: benevolent sexism and aggressive sexism. Benevolent sexism is more subtle with traits that appear to be beneficial, whereas aggressive sexism is characterized by an overtly negative attitude toward women. Sexism can take many different forms, such as direct, indirect, descriptive, or recorded behavior (such as stereotyping, ideological disagreements, sexual aggression, etc.) (Anzovino et al., 2018; Manne, 2017). Misogyny is thus just one example of sexism (Manne, 2017). The majority of earlier research has focused more on identifying hostile and explicit sexism while ignoring covert or implicit sexism (Waseem and Hovy, 2016; Frenda et al., 2019; Anzovino et al., 2018; Pamungkas et al., 2020). As a result, it is important to deal with the identification of sexism in different sexist attitudes and behaviors because these are the most prevalent and harmful to society (Hellinger and Pauwels, 2008).

Hate speech detection is related to recent studies on the detection of sexism. One of the first datasets was created to investigate the relationship between sexism and racism (Waseem, 2016; Waseem and Hovy, 2016). However, this dataset ignores other forms of sexism and only includes instances of hatred or aggressive sexism directed towards women. A classification of sexism by Sharifirad and Jacovi (Sharifirad et al., 2019) included direct, sexual, and physical forms of sexism. A more recent study by (Parikh et al., 2019) aims to classify sexism reports. Other jobs to protect women from hatred on the internet have emerged as a result of the increased interest in hate detection towards women. For instance, sexist MEME detection (Fersini et al., 2019) and sexist advertisement classification (Gasparini et al., 2018) are two examples.

The use of neural models to detect hate speech has drawn interest in recent years. These models frequently employ deep learning techniques like Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs), achieving outstanding results in a variety of natural language processing applications (Pitsilis et al., 2018;

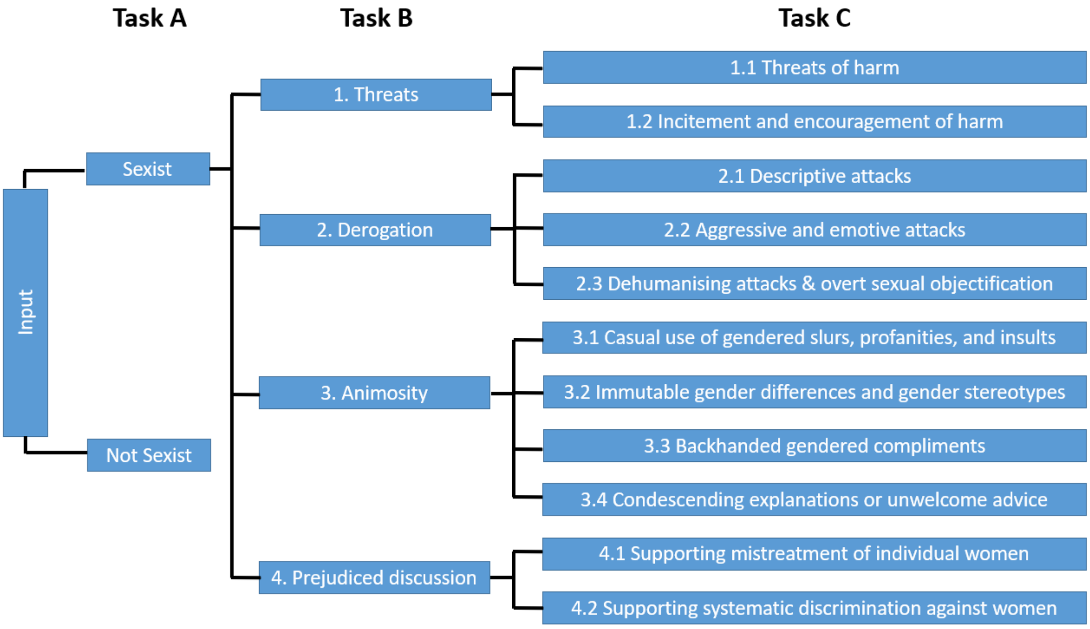

Task A
Task B
Task C
Input
Sexist
Not Sexist
1. Threats
2. Derogation
3. Animosity
4. Prejudiced discussion
1.1 Threats of harm
1.2 Incitement and encouragement of harm
2.1 Descriptive attacks
2.2 Aggressive and emotive attacks
2.3 Dehumanising attacks & overt sexual objectification
3.1 Casual use of gendered slurs, profanities, and insults
3.2 Immutable gender differences and gender stereotypes
3.3 Backhanded gendered compliments
3.4 Condescending explanations or unwelcome advice
4.1 Supporting mistreatment of individual women
4.2 Supporting systematic discrimination against women

Figure 1: Task Description

Badjatiya et al., 2017; Zimmerman et al., 2018; Park and Fung, 2017). Rahgouy et al. (Rahgouy et al., 2022) recently used Stylistically Fused Contextualized Representation and Deep Learning for sarcasm detection. In this study, we combine several of these approaches to address the issue.

3 Task

The task (Kirk et al., 2023) deals with identifying online sexism in English and offers a system to evaluate the effectiveness of the automated solutions. The task contains three subtasks: Task A, Task B and Task C, explained in Figure 1.

1. **TASK A - Binary Sexism Detection:** At the first level, the taxonomy divides content into two categories, sexist and non-sexist. Any maltreatment of women based on their gender or on the combination of their gender with one or more additional identification features (such as Black women or Muslim women), whether explicit or implicit, is referred to as sexist content.

2. **TASK B - Category of Sexism:** Sexist content is broken down into four conceptually and analytically separate groups at the second level of the taxonomy. Because the harm produced by content is idiosyncratic and speaker intent is difficult to determine, especially without larger context, the categories were purposefully designed to not distinguish categories by the purported effect on the recipient or the supposed motivation of the speaker. The four categories are:

   - 1. Threats
   - 2. Derogation
   - 3. Animosity
   - 4. Prejudiced discussion

3. **TASK C - Fine-grained Vector of Sexism:** Each category of sexism is further broken down into fine-grained sexism vectors at the third level of the taxonomy. Vectors that are mutually exclusive (each vector is separate) and collectively exhaustive (each vector contains all content that is sexist) are included. An 11-class classification for sexist posts requires the system to predict one of the following 11 fine-grained vectors:

   - 1.1 Threats of harm (Th)
   - 1.2 Incitement and encouragement of harm (Ieh))
   - 2.1 Descriptive attacks (Da)
   - 2.2 Aggressive and emotive attacks (Aea)
   - 2.3 Dehumanising attacks & overt sexual objectification (Doso)
   - 3.1 Casual use of gendered slurs, profanities, and insults (Cugspi)
   - 3.2 Immutable gender differences and gender stereotypes (Igdgs)

- 3.3 Backhanded gendered compliments (Bgc)
- 3.4 Condescending explanations or unwelcome advice (Ceua)
- 4.1 Supporting mistreatment of individual women (Smiw)
- 4.2 Supporting systematic discrimination against women (Ssdaw)

4 Dataset

The dataset with labels had 20,000 entries, 10,000 of them were drawn from Reddit and 10,000 from Gab (Kirk et al., 2023). For labeling them, three trained annotators first labeled each entry, and one of two experts then decided on any differences. The task contains three tasks: task A, task B and task C. When all of the annotators agreed on one label for Task A, that label is considered to be the gold label. One of the experts evaluated the entry and chose the gold label if there was any disagreement. When two or more annotators agreed on a label for Tasks B and C, the label was considered the gold label; however, if there was a three-way tie, one of the experts determined the gold label. The experts and annotators were all self-identified women.

Category	**Train**	**Dev**	**Test**
Not sexist	10602	1514	3030
Sexist	3398	486	970
Total	**14000**	**2000**	**4000**

Table 1: Dataset split for Subtask A

Category	**Train**	**Dev**	**Test**
1. Threats	310	44	89
2. Derogation	1590	227	454
3. Animosity	1165	167	333
4. Prejudiced discussion	333	48	94
Total	**3398**	**486**	**970**

Table 2: Dataset split for Subtask B

Task A is a classification task between ‘sexist’ and ‘non sexist’ posts. For task A, 14000 posts for training, 2000 for validation and 4000 for testing were used. The training dataset was imbalanced. Out of 14000 posts used for training, 3398 posts were labeled as ‘sexist’, remaining 10602 as ‘non sexist’. For task B, the ‘sexist’ posts were further classified into four sub categories and for task C, these four categories were further classified into 11 categories. For both tasks B and C, 486 posts for validation and 970 posts for testing were used. Tables 1, 2 and 3 show the dataset splits for tasks A, B and C respectively.

Category	**Train**	**Dev**	**Test**
1.1 Th	56	8	16
1.2 Ieh	254	36	73
2.1 Da	717	102	205
2.2 Aea	673	96	192
2.3 Doso	200	29	57
3.1 Cugspi	637	91	182
3.2 Igdgs	417	60	119
3.3 Bgc	64	9	18
3.4 Ceua	47	7	14
4.1 Smiw	75	11	21
4.2 Ssdaw	258	37	73
Total	**3398**	**486**	**970**

Table 3: Dataset split for Subtask C

5 System overview

Recently, Natural Language Processing has seen a rise in popularity of Pretrained Language Models (LMs). A pre-trained language model has the drawback of taking a long time during sentence pair regression processes like clustering and sentence similarity analysis (Seo et al., 2022). The sentence can be embedded to remedy the issue. Recently, many sentence embedding techniques with varied generating mechanisms have been proposed. In this study, we have implemented some models that are listed below.

5.1 BERT

Devlin et al. (Devlin et al., 2018) created the Bidirectional Encoder Representations from Transformers (BERT) model to enhance the predominantly unidirectional language model training. BERT requires two segments concatenated as its input (sequences of tokens). Usually, segments have more than one naturally occurring sentence. With unique tokens separating them, the two segments are provided to BERT as a single input sequence. A sizable unlabeled text corpus is used for the model’s pretrained training, and end-task labeled data is then used to refine it. BERT takes advantage of the design known as the transformer (Vaswani et al., 2017). Masked language modeling (MLM) and next sentence prediction are the two aims that BERT uses during pretraining. A cross-entropy loss on forecasting the masked tokens is the MLM aim. Next Sentence Prediction (NSP) is a binary classification loss for determining whether two segments in the original text follow one another. The NSP objective was created to enhance performance on tasks that come later, including Natural Language Inference (Bowman et al., 2015), which calls for deducing the links between phrase pairs. With Adam optimizer, BERT is optimized (Kingma and Ba, 2014).

Before the embeddings could be made, the sentence had to be tokenized. It should be noted that BERT can only accept sentences that are 512 tokens or shorter in length. Unless it is obvious that adopting a case-sensitive model will be advantageous to the task, the authors of BERT advocate utilizing the BERT Base Uncased model in the majority of situations. BERT is trained on and anticipates sentence pairings, using 1s and 0s to distinguish between the two sentences (McCormick and Ryan, 2019). In other words, each token in "tokenized text" must be indicated as to whether it belongs in sentence 0 (a string of 0s) or sentence 1. (a series of 1s). For each character in the input sentence, a vector of 1s was constructed since single-sentence inputs only need a string of 1 (McCormick and Ryan, 2019).

5.2 SBERT

SBERT is a variant of BERT (Devlin et al., 2018) that creates semantically significant sentence embeddings using siamese and triplet network structures (Reimers and Gurevych, 2019). In this instance, the text documents are divided into paragraphs before using the average of SBERT paragraph embeddings as text document representations. The text documents are then categorized based on the document's SBERT representations that have the highest cosine similarity (Schopf et al., 2022).

Using a variation of the masked language modeling aim used in the pre-training of the BERT model, the SBERT model is pre-trained on a vast volume of text data. The model can be fine-tuned on particular downstream goals like text classification, question answering, or semantic similarity tasks after pre-training. On various benchmark datasets for tasks like phrase categorization and semantic similarity, SBERT has demonstrated state-of-the-art performance, demonstrating its excellent efficacy in capturing sentence semantics. Due to its capacity to produce high-quality sentence embeddings, it is a preferred option for many NLP applications, including chatbots, recommendation systems, and hate speech detection.

5.3 BERT-TFIDF

The BERT representation combined with the well-known Term Frequency Inverse Document Frequency (TFIDF) weighting system to extract traditional features referred to as BERT-TFIDF was used previously for author profiling (Das et al., 2022). TFIDF is a combination of two different terms: Term Frequency (TF) and Inverse Document Frequency (IDF) (Qaiser and Ali, 2018). The algorithm evaluates all keywords similarly while calculating a document's term frequency, regardless of whether they are stop words or not, which is incorrect because the relevance of all key words are not same (Hakim et al., 2014). The inverse document frequency method gives less weight to more frequently occurring words and more weight to infrequently occurring terms (Hakim et al., 2014). Mathematically, term frequency (TF) and inverse document frequency (IDF) are multiplied to create TFIDF. In general, TFIDF's purpose is to decrease the impact of less informative tokens that appear frequently in a data corpus (Gaydhani et al., 2018). We used TfidfVectorizer features from scikit-learn to perform the TFIDF task (Pedregosa et al., 2013).

TFIDF is used to assess a word's relevance to a particular document in a group of papers (Das et al., 2022). The Bert model's performance can be enhanced by feeding it the TFIDF score (Das et al., 2022). We adopted this embedding strategy to generate a richer and more understanding quantitative representation of the data.

5.4 Fine-tuned RoBERTa

BERT was significantly undertrained for few specific tasks. In order to address this weakness, a new recipe for training BERT models termed RoBERTa was proposed (Liu et al., 2019). RoBERTa stands for Robustly optimized BERT approach. RoBERTa can match or outperform the performance of all post-BERT approaches. The changes in RoBERTa consist of the following: (1) training the model for a longer period of time with larger batches of data; (2) eliminating the objective of next sentence prediction; (3) training on longer sequences; and (4) dynamically altering the masking pattern used on the training data (Liu et al., 2019). In particular, dynamic masking, FULL-SENTENCES without

NSP loss, big mini-batches, and a bigger byte-level BPE are used to train RoBERTa (Liu et al., 2019).

RoBERTa is a multilingual model with Transformer as the primary structure, that was used for the representation. The model has 12 layers with 768 output dimensions. After obtaining the embedding vectors of the texts, RoBERTa was improved to make it better suited for the subsequent task of identifying hate speech. The 12 layers that makeup RoBERTa, each learn a distinct type of semantic data. In general, more word level semantic information is learned when the layers are thinner. More broad semantic knowledge is learned as one delves further into the levels. For binary classification tasks like subtask A, global semantic information is more beneficial. The RoBERTa model's hidden layer has 12 layers of Transformer and 768 dimensions. The 12th layer was removed as the dimension was (0,2) and was spliced it with a vector (Tc) where Tc's shape was [32,768] and the hidden vector of the 12th layer was [32,60,768]. The resulting data was then passed to the classifier and the results were sent to Softmax. By incorporating the transformed data into the model, both tasks were trained. Similar activities can provide useful information for multitask learning.

We used RoBERTa large for our research. 24 transformer layers, each with 16 self-attention heads and 1,024 hidden units, make up the RoBERTa large model. The RoBERTa model's transformer architecture is a critical component that enables the model to handle input text in a highly parallelized and effective manner while capturing contextual data and distant dependencies. Each transformer layer in RoBERTa is divided into two sub-layers: a feed-forward neural network and a self-attention mechanism. The model can weigh various input sequence components differently depending on how relevant they are to the present context thanks to the self-attention mechanism. This approach allows the model to concentrate on the most crucial components of the input and reject unimportant data, making it particularly helpful for processing lengthy sequences. The self-attention layer's output is given a non-linear activation function by the feed-forward sub-layer, which enables the model to recognize more intricate patterns and connections among the input's many components.

We utilized a multi-task learning approach to address all three tasks simultaneously with a single model. This was achieved by encoding the initial representation of a post using Roberta and setting three distinct Multilayer Perceptrons (MLPs) corresponding to the three sub-tasks. Our model can be viewed as a hard multi-task learning paradigm where all three heads share the 24 layers defined in Roberta large. The average loss obtained from each head is used to backpropagate and adjust the parameters of the model, resulting in a reduction of the final model size by one-third compared to the separated version without sacrificing performance.

6 Results & Discussion

6.1 Task A

We first implemented the sentence-BERT representation and used logistic regression model for classification on the validation dataset. We got a macro-F1 score of 67.26%. We then did the same experiment with BERT representation and got a score of 68.94%. We then further implemented the BERT representation combined with TFIDF (Das et al., 2022) and used multilayer perceptron model for classification on the validation dataset. The F1 score was increased to 71.93%. It is a point to be noted that when the TFIDF features were included with the BERT representation, the F1 score was increased. We then implemented the fine-tuned RoBERTa for classification and the F1 score was significantly improved to 83.64%. Table 4 shows the F1 scores of the models we used on validation data. Out of all the models listed here, fine tuned RoBERTa had significantly higher F1 score than the other models on the development data.

Representation	**Macro-F1**
SBERT	67.26
BERT	68.94
BERT-TFIDF	71.36
Fine tuned RoBERTa	83.64

Table 4: F1 scores of different models on the development dataset for Task A

After trying different models, we decided to implement the fine-tuned RoBERTa model for our task. Figure 2 shows the confusion matrix of our prediction with true label on task A development data. The number of correct predictions of 'not sexist' and 'sexist' comments are 1423 and 345 respectively which makes total 1768 correct predictions out of 2000 data.

The test dataset contained 4000 posts. The F1 score of our model on the test dataset for task A is

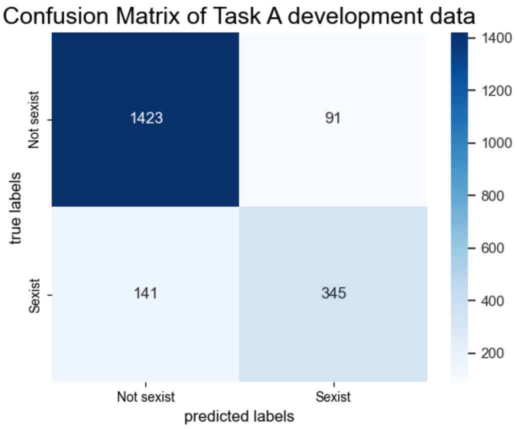


Figure 2: Confusion Matrix of Task A development data

shown in Table 5.

Representation	Macro-F1
Fine tuned RoBERTa	79.43

Table 5: F1 scores of our model on the test dataset for Task A

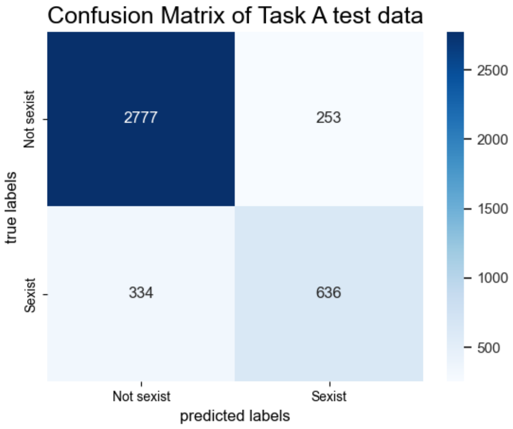


Figure 3: Confusion Matrix of Task A test data

Figure 3 shows the confusion matrix of our prediction with true label on task A test data. The number of correct predictions of 'not sexist' and 'sexist' comments are 2777 and 636 respectively which makes total 3413 correct predictions out of 4000 data.

6.2 Task B

For task B also we first implemented the sentence-BERT representation and used logistic regression model for classification on the validation dataset. We got a macro-F1 score of 53.74%. We then did the same experiment with BERT representation and got a score of 54.58%. The BERT representation was more efficient than SBERT representation for task B. We then further implemented the BERT representation combined with TFIDF (Das et al., 2022) and used multilayer perceptron model for classification on the validation dataset. The F1 score was decreased to 50.56%. Here is a point to be noted that when the TFIDF features were combined with the BERT representation, the overall F1 score for task B decreased than using only BERT representation. We then implemented the fine-tuned RoBERTa for classification and the score was significantly improved to 65.88%. It is interesting to see that not only binary classification, but our model was efficient for multi class classification task also. Table 6 shows the F1 score of the models we used on validation data.

Representation	Macro-F1
SBERT	53.74
BERT	54.58
BERT-TFIDF	50.56
Fine tuned RoBERTa	65.88

Table 6: F1 scores of different models on the development dataset for Task B

Figure 4 shows the confusion matrix of task B on the development dataset where number 1 refers to category 'Threats', number 2 refers to category 'Derogation', number 3 refers to category 'Animosity' and number 4 refers to category 'Prejudiced Discussion'. The number of correct predictions of the four categories are 31, 182, 84 and 32 respectively.

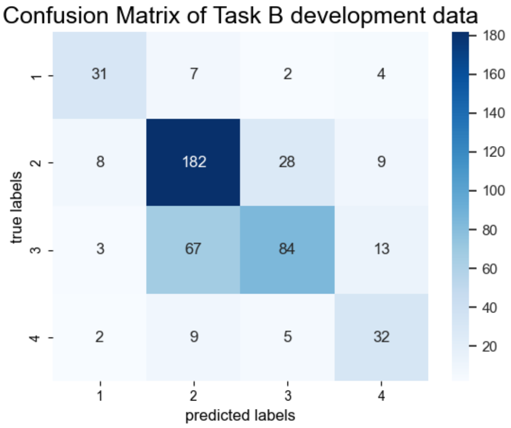


Figure 4: Confusion Matrix of Task B development data

The test dataset contained 970 posts. The F1 score of our model on the test dataset is shown in

Table 7.

Representation	Macro-F1
Fine tuned RoBERTa	61.91

Table 7: F1 scores of our model on the test dataset for Task B

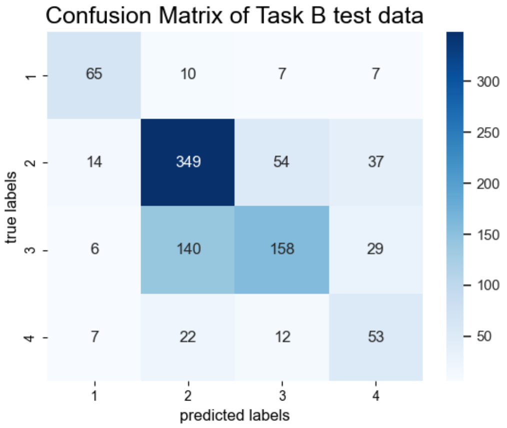


Figure 5: Confusion Matrix of Task B test data

Figure 5 shows the confusion matrix of task B on the test dataset. The number of correct predictions of the four categories are 65, 349, 158 and 53 respectively.

6.3 Task C

For task C also we first implemented the sentence-BERT representation and used logistic regression model for classification on the validation dataset. We got a macro-F1 score of 32.74%. We then did the same experiment with BERT representation and got a score of 32.14%. We then further implemented the BERT representation combined with TFIDF (Das et al., 2022) and used multilayer perceptron model for classification on the validation dataset. The F1 score was increased to 26.64%. Here is a point to be noted that when the TFIDF features are combined with the BERT representation, although the overall F1 score increased for task A, the F1 score decreased for both tasks B and C. A possibility is that the TFIDF features did not add any significant value to multiclass classifications task. We then implemented the fine-tuned RoBERTa for classification and the F1 score was significantly improved to 33.20%. Table 8 shows the F1 scores of the models we used on validation data.

Representation	Macro-F1
SBERT	32.74
BERT	32.14
BERT-TFIDF	26.64
Fine tuned RoBERTa	33.20

Table 8: F1 scores of different models on the validation dataset for Task C

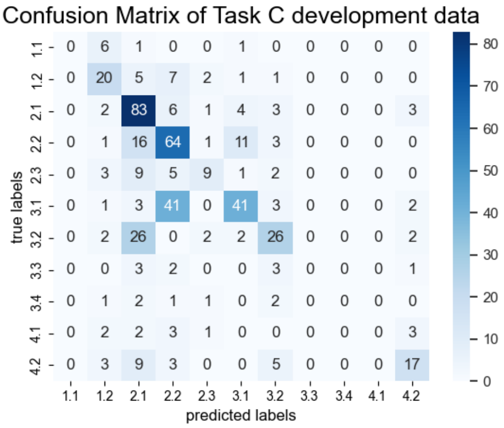


Figure 6: Confusion Matrix of Task C development data

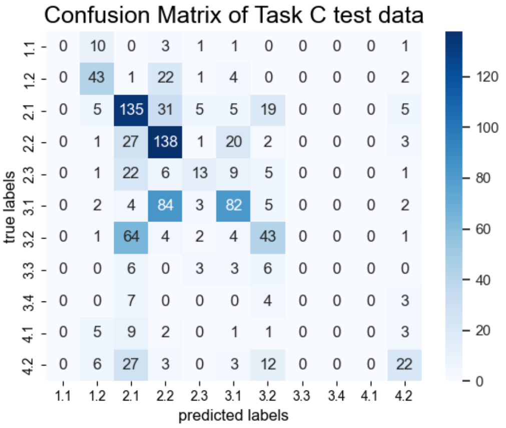
Figure 6 shows the confusion matrix of task C on the development dataset where number 1.1 refers to category ‘Threats of harm’, number 1.2 refers to category ‘Incitement and encouragement of harm’, number 2.1 refers to category ‘Descriptive attacks’, number 2.2 refers to category ‘Aggressive and emotive attacks’, number 2.3 refers to category ‘Dehumanisation and overt sexual objectification’, number 3.1 refers to category ‘Casual use of gender slurs, profanities and insults’, number 3.2 refers to category ‘Immutable gender stereotypes’, number 3.3 refers to category ‘Backhanded gendered compliments’, number 3.4 refers to category ‘Condescending explanations or unwelcome advice’, number 4.1 refers to category ‘Supporting mistreatment of individual women’ and number 4.2 refers to category ‘Supporting systemic discrimination against women’. The number of correct predictions of the eleven categories are 0, 20, 83, 64, 9, 41, 26, 0, 0, 0, and 17 respectively.

The test dataset contained 970 posts. The F1 score of our model on the test dataset is shown in Table 9.

Representation	Macro-F1
Fine tuned RoBERTa	29.9

Table 9: F1 score of our model on the test dataset for Task C

Figure 7 shows the confusion matrix of task C on the development dataset. The number of correct predictions of the eleven categories are 0, 43, 135, 138, 13, 82, 43, 0, 0, 0, and 22 respectively.



	1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4	4.1	4.2
1.1	0	10	0	3	1	1	0	0	0	0	1
1.2	0	43	1	22	1	4	0	0	0	0	2
2.1	0	5	135	31	5	5	19	0	0	0	5
2.2	0	1	27	138	1	20	2	0	0	0	3
2.3	0	1	22	6	13	9	5	0	0	0	1
3.1	0	2	4	84	3	82	5	0	0	0	2
3.2	0	1	64	4	2	4	43	0	0	0	1
3.3	0	0	6	0	3	3	6	0	0	0	0
3.4	0	0	7	0	0	0	4	0	0	0	3
4.1	0	5	9	2	0	1	1	0	0	0	3
4.2	0	6	27	3	0	3	12	0	0	0	22

Figure 7: Confusion Matrix of Task C test data

7 Conclusion

In this paper we presented our method to the SemEval 2023 Explainable Detection of Online Sexism task to address the sexism detection problem on Gab and Reddit data. The task contained three subtasks - one for binary sexism classification, one for category of sexism classification and one for fine-grained vector of sexism classification. To address the tasks, we implemented fine tuned RoBERTa model and compared it to several other models like BERT, SBERT, BERT combined with TFIDF etc. The results show that our model gave the best Macro-F1 score. To conclude, we have shown some a very useful technique for online sexism detection. How this model behaves to a different type of dataset, will be a future direction to explore.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Amit Das, Nilanjana Raychawdhary, Gerry Dozier, and Cheryl D Seals. 2022. Irony and stereotype spreading author profiling on twitter using machine learning: A bert-tfidf based approach.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Angel Felipe Magnossão de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv preprint arXiv:2111.04551*.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for german offensive language and conversation analysis. In *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022), Association for Computational Linguistics, Online*, pages 54–61.

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*.

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.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

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.

Barbara L Fredrickson and Tomi-Ann Roberts. 1997. Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of women quarterly*, 21(2):173–206.

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.

Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. 2018. Multimodal classification of sexist advertisements. In *ICETE (1)*, pages 565–572.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

Philip Gianfortoni, David Adamson, and Carolyn Rose. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 49–59.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Peter Glick and Susan T Fiske. 2001. Ambivalent sexism. In *Advances in experimental social psychology*, volume 33, pages 115–188. Elsevier.

Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, and Wahyu Muliady. 2014. Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach. In *2014 6th international conference on information technology and electrical engineering (ICITEE)*, pages 1–4. IEEE.

Marlis Hellinger and Anne Pauwels. 2008. 21. language and sexism. *Handbook of language and communication: Diversity and change*, pages 651–684.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*.

Kate Manne. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Chris McCormick and Nick Ryan. 2019. Bert word embeddings tutorial. *URL: https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial*.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

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*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

F Pedregosa et al. 2013. sklearn. feature_extraction. text. tfidfvectorizer.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

Mostafa Rahgouy, Hamed Babaei Giglou, Taher Rahgooy, and Cheryl Seals. 2022. Null at semeval-2022 task 6: Intended sarcasm detection using stylistically fused contextualized representation and deep learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 862–870.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Francisco Rodríguez-Sánchez, 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. *Procesamiento del Lenguaje Natural*, 69:229–240.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. *arXiv preprint arXiv:2211.16285*.

Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic fake news detection with pre-trained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*, pages 627–641. Springer.

Jaejin Seo, Sangwon Lee, Ling Liu, and Wonik Choi. 2022. Ta-sbert: Token attention sentence-bert for improving sentence representation. *IEEE Access*, 10:39119–39128.

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.

Jane M Ussher. 2016. Misogyny. *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, pages 1–3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

P William, Ritik Gade, Rup esh Chaudhari, AB Pawar, and MA Jawale. 2022. Machine learning based automatic hate speech recognition system. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 315–318. IEEE.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

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)*.