# Explainable Detection of Online Sexism (EDOS)

**Rudraksh Sugandhi**
sugar01@pfw.edu

**Sharihar Bulbul**
bulbm01@pfw.edu

**Avani Dubey**
dubea02@pfw.edu

## Abstract

In the current digital era, when online platforms have developed into a fertile ground for sexist conduct, the identification of online sexism is an essential responsibility. Traditional machine learning techniques, however, frequently are unable to offer an explanation for the sexism that is observed, which might impede the creation of efficient mitigation solutions. This study suggests an explainable detection strategy for online sexism that combines cutting-edge explainable models with established machine learning techniques. The system is able to spot instances of sexism in online material while giving justification for its choices, improving the process's transparency and interpret ability.

## 1 Problem Statement

The goal of Explainable Detection of Online Sexism is to create a system that can properly identify instances of sexist speech or conduct in online settings while also explaining why a certain piece of information was labeled as sexist. The aim is to develop a tool that can assist online communities, social media platforms, and individuals in identifying and addressing instances of sexism and prejudice while also encouraging openness and responsibility in the process. The difficulty lies in creating a model that can properly identify instances of sexism by analyzing natural language and contextual data, and by presenting concise and accessible justifications for those conclusions, which can help people comprehend and address the problem.

## 2 Motivation

1. Taking action against the issue of online sexism: Online sexism is a ubiquitous and dangerous issue that may have detrimental repercussions on both people and society as a whole. We can address the issue and take action to lessen its effects by developing a system that can spot instances of online sexism.

2. Enhancing justice and equity: Certain groups of people might be disproportionately affected by online misogyny, notably women and non-binary persons. We can make sure that the system is just and equitable and that it doesn't feed or maintain preexisting prejudices and preconceptions by creating an explainable detection mechanism.

3. Increasing accountability and transparency: One of the main advantages of an explainable detection system is that it may contribute to greater accountability and transparency. We can guarantee that the method is applied fairly and appropriately by clearly stating the reasons why specific content is marked as sexist.

4. Supporting research and knowledge: Creating an understandable approach for detecting online sexism can help to enhance the understanding and investigation of the issue. Researchers may learn more about the type and scope of online sexism by examining the system's data, which can then guide attempts to combat the issue.

## 3 Objective

**Objective 1** - Binary Sexism Detection: This is a two-class (or binary) classification in which systems must predict whether or not a post is sexist.
**Objective 2** - Sexism Category: for sexist posts, a four-class classification in which systems must forecast one of four categories:
(1) threats (2) derogation (3) anger (4) biased discussion.

## 4 Models

There are various models that can be used to solve the problem of explainable detection of online sexism. Here are some examples:

**Supervised Learning Models**: These models use a labeled dataset to train a machine learning algorithm to identify and classify sexist text. Examples of supervised learning algorithms that could be used for this task include logistic regression.

**Deep Learning Models**: These models use neural networks to learn complex representations of text data and make predictions about whether a piece of text contains sexist language or not. Examples of deep learning models that could be used for this task include recurrent neural and transformer-based models like BERT or GPT.

**Hybrid Models**: These models combine different techniques from the above models to improve the performance and interpretability of the model. For example, a hybrid model could use a rule-based system to identify obvious instances of sexism, and a deep learning model to capture more subtle and nuanced forms of sexist language.

**BERT (Bidirectional Encoder:** Representations from Transformers) is a pre-trained language model. Using novel fine-tuning approaches based on transfer learning, we can use BERT to capture hateful contexts inside social media posts. We employ this model to datasets annotated for racism, sexism, hatred, or objectionable content to assess our suggested approach.

## 5 Data

The Dataset was provided by **SemEval 2023 Task 10 - Explainable Detection of Online Sexism (EDOS)**. Our labeled dataset has 20,000 entries. 10,000 are drawn from Gab and 10,000 from Reddit. All entries are labeled initially by three professional annotators, and any differences are resolved by one of two experts. All of the annotators and specialists were women. The training data consists of 14,000 entries. There are 2,000 entries in the development data The test data consists of 4,000 entries. One CSV file contains labels for Tasks A and B. The columns are as follows: rewire id: a unique identifier for each entry text: the input text label sexist: Task A label category: Task B label.

## 6 Measurement of Accuracy

Accuracy: Accuracy measures the percentage of correctly classified instances out of the total number of instances in the dataset. This metric is commonly used for binary classification tasks where there are only two possible labels (positive or negative).

**F1- Score:** The F1 score is a weighted average of precision and recall that provides a single value that describes the model's performance. It is especially beneficial in situations when precision and recall are critical, such as spam detection.

**AUC-ROC:** The Area Under the Receiver Operating Characteristic (AUC-ROC) curve is a measure of a binary classifier's performance at various classification thresholds. It compares the true positive rate (sensitivity) to the false positive rate (1 - specificity) and returns a single number that highlights the model's overall performance.sub

## 7 Possible Analysis

**Model evaluation analysis**: Way of measurement for specific models that tells how good/bad a model's performance is. It is the process of understanding how good or bad a model is. Generally, sample data (Training data) is used to build the model, and evaluation is done on out-of-sample data (Testing data) to see the performance.

**Bias assessment** It would be important to assess the potential biases in the model, especially given the sensitivity of the topic. This could involve analyzing the impact of different demographic factors such as gender, race, or nationality on the model's predictions

**Error Analysis:** The errors are then analyzed to identify patterns or common mistakes. This can help to improve the model by identifying areas where it is weak. The causes of the errors are investigated by looking at unclassified data or examining the model's predictions.

**Feature Analysis:** For the model, a variety of features can be used and here are some examples of features that could be used for EDOS: 1. Some features would involve the analysis of the words and phrases used in the text. For example, the presence of gendered pronouns (e.g., he, she, him, her), gendered insults (e.g., slut, whore), or gendered stereotypes (e.g., men are strong, women are weak) can be indicators of sexist content.

2.Some features would involve the analysis of the meaning of the text. For example, the presence of negative sentiment (e.g., hate, anger, disgust) towards a particular gender can be an indicator of sexist content. Also, the presence of dehumanizing language (e.g., referring to women as animals) can be a strong indicator of sexism.

3.Some features would involve the analysis of the frequency and co-occurrence of words and phrases in the text. For example, the presence of specific combinations of words or phrases (e.g., "boys will be boys," "she asked for it") can be strong indicators of sexist content.