

Leveraging GPT-2 for Automated Classification of Online Sexist Content

Notebook for the Exist 2023 Lab at CLEF 2023

Advaitha Vetagiri^{1,*}, Prottay Kumar Adhikary¹, Partha Pakray¹ and Amitava Das^{2,3}

¹National Institute of Technology Silchar, Assam, 788010, India

²Artificial Intelligence Institute of UofSC (AIISC) 1112 Greene St. Columbia, South Carolina, USA

³Wipro AI Lab, Bangalore, Karnataka, India

Abstract

In today's digital culture, sexism and misogyny on online platforms have grown to be serious issues. To solve these problems, efficient automated sexist content detection and classification techniques must be created. In this study, we investigate the application of the GPT-2 model, a cutting-edge pre-trained language model, to the shared Exist 2023 job of sexism categorization. On the Exist 2023 dataset, we fine-tuned the GPT-2 model by adding adjustments like a classification head and weighted cross-entropy loss to tackle class imbalance. Our experimental findings show the GPT-2 model's potential for precisely recognizing and classifying instances of sexism. Using the official assessment measure ICM (Information Contrast Measure), we assess our strategy while taking into account various evaluation modes, such as hard-hard, hard-soft, and soft-soft. The results show how well the GPT-2 model handles the problem of sexism categorization, assisting in the creation of automated techniques for fostering a safer and more welcoming online environment.

Keywords

Sexism Classification, GPT-2, Exist 2023, ICM

1. Introduction

Sexism, encompassing various forms of oppression and prejudice against women due to gender, remains a pervasive issue today. It manifests in numerous ways, including stereotyping [1], ideological biases [2], and even explicit acts of sexual violence [3]. As a result of the realization that online forums significantly influence public debate, identifying and combatting sexism in social networks has emerged as a crucial subject of research. As scholars and practitioners strive to understand better and address this complex problem, the EXIST 2023 shared task [4] emerges as a pivotal platform for advancing state of the art in sexism detection and classification.


Numerous studies have shown how seriously sexism affects both individuals and society as a whole. It perpetuates gender inequalities, restricts opportunities, and reinforces harmful stereotypes, ultimately hindering progress towards gender equality [5]. Previous studies have


CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

✉ advaitha21_rs@cse.nits.ac.in (A. Vetagiri); prottay_ug@cse.nits.ac.in (P. K. Adhikary); partha@cse.nits.ac.in (P. Pakray); AMITAVA@mailbox.sc.edu (A. Das)

ORCID 0000-0002-0651-4171 (A. Vetagiri); 0000-0003-3834-5154 (P. Pakray)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

focused on explicit misogyny and violence against women [6]. However, the EXIST campaigns seek to broaden the scope of the investigation, encompassing a comprehensive range of sexist phenomena, including overt and covert manifestations. By doing so, EXIST aims to shed light on the nuanced nature of sexism and provide insights into its prevalence, dynamics, and potential countermeasures.

In order to address the ubiquitous problem of sexism in social networks, academics, professionals, and practitioners from diverse disciplines have joined forces to create the EXIST 2023 shared task at CLEF (Conference and Labs of the Evaluation Forum). This collaborative job offers a chance to expand the field's knowledge and capacity to confront sexism in online platforms with an emphasis on creating automated tools for sexism recognition [7], source intention analysis, and sexism categorization. The EXIST 2023 shared task offers us invaluable tools to develop and assess our tactics, encouraging creativity and cooperation in the fight against sexism. It does this by utilizing a vast dataset that includes English and Spanish tweets. To effectively eradicate sexism on social networks, the EXIST 2023 shared task is essential. This shared task aims to advance the development of effective automated systems for identifying and categorising sexist content by addressing the diverse forms of sexism and incorporating the learning with disagreements paradigm.

Generative Pre-trained Transformer 2 (GPT-2) [8] is an advanced language model developed by OpenAI [9]. It is part of the GPT series, which stands for "Generative Pre-trained Transformer," and represents a breakthrough in natural language processing (NLP) research. GPT-2 utilises a deep learning architecture known as a transformer, designed to understand and generate human-like text based on vast training data. GPT-2 is pre-trained on a massive corpus of text from the internet, allowing it to learn the statistical patterns and structures of language. The model's architecture enables it to capture contextual information, making it adept at understanding and generating coherent text based on the input. One of the remarkable features of GPT-2 is its ability to generate highly realistic and contextually appropriate text, making it suitable for a wide range of NLP tasks, such as language translation [10], text completion, and text classification [11]. The model has been trained on diverse text types, allowing it to exhibit a broad understanding of various topics and writing styles.

GPT-2 can be a potent automated analytic technique for categorizing sexism. GPT-2 can learn the underlying patterns and characteristics that are indicative of sexism by training the model on a labelled dataset that differentiates between sexist and non-sexist text. The trained GPT-2 model uses its linguistic knowledge to assess the possibility that a text fragment, such as a tweet, contains sexist material throughout the classification phase. GPT-2's ability to classify texts for the purpose of identifying sexism is based on its capacity to extract semantic and contextual information from the input text. GPT-2 can recognize the subtle subtleties of language that may imply sexism, such as the use of pejorative phrases, stereotypes, or statements that are degrading to women, by learning from a variety of samples. Through its extensive training and exposure to large-scale language data, GPT-2 can contribute to automated systems that aid in identifying and combating sexism in social networks.

2. Literature Survey

As shown in prior work by [12], one method for overcoming the difficulty of sexism identification is to use conventional machine learning approaches, such as n-grams. Creating a dataset for recognizing and categorizing sexist language on Twitter in both Spanish and English was the main goal of this study. Similarly, [13] used two datasets in a similar manner to identify online hate speech directed towards women.

However, recent developments in the field have investigated the use of sophisticated deep-learning approaches to produce cutting-edge outcomes in sexism detection. For instance, [14] used modified LSTMs with attention mechanisms [15] and GloVe embeddings [16] to automatically recognize sexist remarks often heard in the workplace. These studies show how deep learning techniques may be used to overcome the difficulty of identifying sexism and misogyny in natural language writing.

A strong model for a variety of natural language processing (NLP) tasks, such as text classification [11], sentiment analysis [17], and language modelling [10], is GPT-2 (Generative Pre-trained Transformer 2) [8]. [18] used a GPT-2 model that has already been trained to perform binary classification on a dataset of news items, determining whether or not each article had sexist material. The outcomes showed that the GPT-2 model performed better in terms of accuracy than a number of other machine learning models.

Similar to this, [19] achieved cutting-edge performance in sentiment analysis tasks using a pre-trained GPT-2 model on a dataset of customer reviews. This research shows the possibility of using pre-trained language models like GPT-2 for diverse natural language processing tasks, even though they did not explicitly focus on sexism categorization. Due to their performance, comparable methods leveraging GPT-2 may also be successful in solving the challenge of classifying sexism [20]. Notably, Encoder-Decoder models have been the mainstay of prior methods for categorizing sexism in English and Spanish.

3. Data

The dataset used in the context of the EXIST 2023 shared task on sexism identification in social networks plays a crucial role in training and evaluating models for automated classification. The dataset is carefully constructed to encompass a wide range of expressions and terms commonly used to undermine the role of women in society, both in English and Spanish. With over 400 expressions included, the dataset aims to capture various forms of sexism, from explicit misogyny to more subtle and implicit sexist behaviours. It covers different dimensions of sexism, such as stereotyping, ideological issues, and sexual violence which is represented as a flow in Figure 3. Including descriptive or reported assertions further expands the scope of the dataset, allowing for the analysis of tweets where the sexist message is a report or description of sexist behaviour.

3.1. Data Sampling and Annotation

A crawling process was conducted to create the dataset, resulting in the collection of more than 8,000,000 tweets in English and Spanish. The crawling period spanned from September

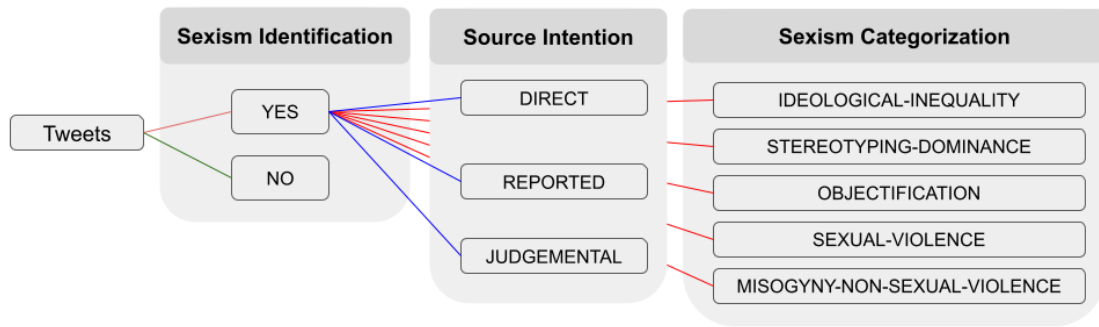


Figure 1: Dataset Classes

2021 to September 2022, ensuring a diverse and up-to-date set of tweets. To maintain balance among the seeds, those with fewer than 60 tweets were removed. The dataset is carefully labelled through an annotation process involving crowdsourcing annotators selected through the Prolific app. To mitigate label bias, gender and age parameters are taken into account during the annotation process. Six annotators annotate each tweet, and their diverse views and annotations are captured rather than relying on a single aggregated label.

3.2. Learning with Disagreements

The premise that natural language expressions possess a solitary and unequivocal interpretation within a given context is a convenient abstraction, yet it deviates significantly from reality, particularly in subjective tasks such as the identification of sexism. The learning with disagreements approach endeavours to address this predicament by enabling systems to acquire knowledge from datasets where definitive gold annotations are absent but instead encompass information pertaining to annotations from all annotators, encompassing many perspectives. In line with methodologies proposed for training directly from discordant data, as opposed to relying on aggregated labels, we will furnish all annotations per instance across the six distinct strata of annotators. This learning with disagreements paradigm acknowledges the subjective nature of sexism identification and aims to capture the diversity of perspectives.

3.3. Development, Training, and Test Data

Table 1
Dataset Split Statistics

	Dev	Train	Test
Spanish	549	3660	1098
English	489	3260	978
Total	1038	6920	2076

The dataset is partitioned into development, training, and test sets with specific temporal distributions to mitigate temporal bias. The training set consists of 3,660 tweets in Spanish and 3,260 tweets in English, while the development consists of 549 tweets in Spanish and 489 tweets in English, and the test sets consist of 1,098 tweets in Spanish and 978 tweets in English, respectively. Additionally, tweets containing less than five words are removed to ensure the inclusion of meaningful content. The EXIST dataset, with its comprehensive coverage of sexist expressions and behaviours, enables researchers and participants in the EXIST 2023 shared task to develop and evaluate models for sexism identification. Its carefully designed labelling process and inclusion of diverse perspectives contribute to a robust analysis of sexism in social networks.

4. System Overview

The GPT-2 (Generative Pre-trained Transformer 2) model is a highly advanced language model that can be leveraged for sexism classification in the Exist 2023 shared task context. The task consists of three specific subtasks: Task 1 focuses on sexism identification (binary classification), Task 2 involves source intention classification (a multiclass hierarchical classification), and Task 3 deals with sexism categorisation (multiclass hierarchical multi-label classification).

For Task 1, the GPT-2 model can classify text instances as sexist (YES) or not sexist (NO). The model can learn the patterns and linguistic cues indicative of sexism by fine-tuning the pre-trained GPT-2 model on a dataset specifically annotated for sexism identification. During the fine-tuning process, the model's parameters are adjusted to optimise its performance in distinguishing between YES and NO text. The output of the model for Task 1 will be a binary classification label, indicating whether the text is sexist (YES) or not sexist (NO).

In Task 2, the GPT-2 model can be utilised to classify the source intention of the sexist text. The classification is hierarchical, with the first level distinguishing between sexist and non-sexist text and the second level categorising the sexist text into three mutually exclusive subcategories: Direct, Reported, and Judgmental. The model can learn the specific linguistic cues associated with each subcategory by training the GPT-2 model on a dataset that includes source intention annotations. The model output for Task 2 will provide the source intention classification label for each text instance.

Task 3 involves sexism categorisation, a multiclass hierarchical multi-label classification problem. Like Task 2, the GPT-2 model can be fine-tuned on a dataset with annotations for sexism categorisation. The first classification level distinguishes between sexist and non-sexist text. In contrast, the second level includes subcategories such as Ideological-Inequality, Stereotyping-Dominance, Objectification, Sexual-Violence, and Misogyny-Non-Sexual-Violence. As Task 3 allows multiple subcategories to be assigned to a single text instance, the GPT-2 model must generate multi-label predictions. The model output for Task 3 will provide the probabilities or confidence scores for each subcategory, indicating the extent to which the text instance belongs to each category.

By leveraging the contextual understanding and language modelling capabilities of GPT-2, the model can effectively capture the nuanced linguistic patterns associated with sexism. It can consider the relationships between words, phrases, and sentences to make informed predictions

for each task. The GPT-2 model can be optimised through the fine-tuning process to achieve high performance in sexism classification, addressing the specific requirements of Task 1, Task 2, and Task 3 in the Exist 2023 shared task.

5. Experimental Setup

The first step in the experimental setup for using the GPT-2 model for sexism classification was to fine-tune the pre-trained GPT-2 model on the Exist 2023 dataset. This was done using the PyTorch framework and the Hugging Face Transformers library, which provided access to the pre-trained GPT-2 model.

During the fine-tuning process, the GPT-2 model was trained for five epochs with a learning rate (LR) of $1e-5$, a commonly chosen value for fine-tuning pre-trained language models. The GPT-2 model used in this experiment had a hidden size of 768 and a maximum sequence length of 128 tokens. To adapt the model for the sexism classification task, a classification head with two output classes (sexist and not sexist) was added to the model architecture.

To train the GPT-2 model on the Exist 2023 dataset, the model was first initialised with the weights from the pre-trained GPT-2 model. Then, the model was trained on the dataset using cross-entropy loss as the objective function and the Adam optimiser. The training was performed with a batch size of 8, meaning that the model processed eight instances simultaneously during each training iteration. It's worth noting that only the weights of the classification head were updated during training, while the weights of the pre-trained model were frozen.

A regularisation technique called dropout was employed to mitigate the risk of overfitting. Dropout randomly sets a fraction (in this case, 0.1) of the input units to zero during each training step. This helps prevent the model from relying too heavily on specific features and improves its generalisation capability.

Furthermore, the weighted cross-entropy loss was utilised to handle the imbalanced distribution of classes in the Exist 2023 dataset. This approach assigned higher weights to the minority class (e.g., YES) and lower weights to the majority class (e.g., NO). Doing so encouraged the model to pay more attention to the less frequent class during training, thus addressing the class imbalance issue.

6. Results & Discussion

In this section, we present the outcomes of our methodologies applied to Tasks 1, 2, and 3 using the test dataset, as well as the final results provided by the organizers of the Shared Task.

6.1. Test Result

The test results demonstrate the effectiveness of our approach, which validates the efficacy of our methodologies in solving the given tasks.

6.1.1. Task 01

Classification report on Tabel 2 provides a summary of the model's performance for each label in the classification task. For the "NO" label, the model achieved a precision of 0.76, a recall of 0.56, and an f1-score of 0.65. This suggests that the model accurately identified 76% of the instances as "NO," correctly recalling 56% of the actual "NO" instances. On the other hand, for the "YES" label, the precision was 0.62, a recall was 0.81, and the f1-score was 0.7. This indicates that the model accurately classified 62% of the instances as "YES," correctly recalling 81% of the actual "YES" instances. Overall, the accuracy of the model was reported as 0.68, meaning that it correctly classified 68% of the instances. These results suggest that the model performs reasonably well in distinguishing between "YES" and "NO" classes, but there is room for improvement, particularly in increasing the precision for the "YES" label and the recall for the "NO" label.

Table 2

Task 01 Testing Classification Report

Label	precision	recall	f1-score
NO	0.76	0.56	0.65
YES	0.62	0.81	0.70
accuracy			0.68

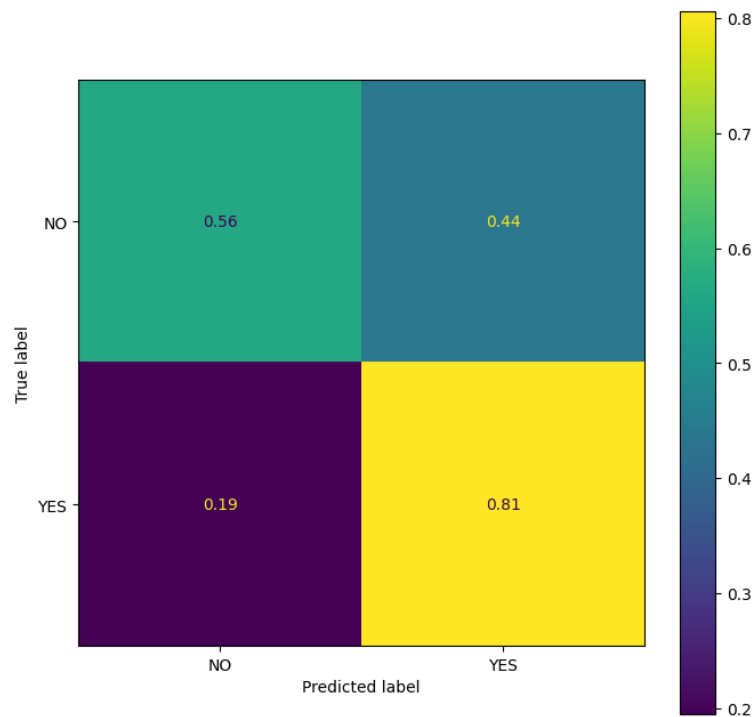


Figure 2: Task 01 Testing confusion Matrix

6.1.2. Task 02

The classification report in Table 3 provides an evaluation of the model's performance across different labels. For the "NO" label, the model achieved a precision of 0.46, a recall of 0.6, and an f1-score of 0.52. This suggests that the model accurately identified 46% of the instances as "NO" and correctly recalled 60% of the actual "NO" instances. However, for the "DIRECT" label, the precision was 0.4, the recall was 0.33, and f1-score was 0.25, indicating that the model struggled to classify instances as "DIRECT" accurately. The "REPORTED" label had precision, recall, and f1-score values of 0, indicating that the model failed to identify any instances as reported. On the other hand, the model performed well in classifying instances as "JUDGEMENTAL," achieving a precision of 0.72, recall of 0.85, and an f1-score of 0.78. The overall accuracy of the model was reported as 0.64, meaning that it correctly classified 64% of the instances. These results highlight the model's strengths in identifying "JUDGEMENTAL" instances but its limitations in accurately distinguishing between "DIRECT" and "REPORTED" classes.

Table 3
Task 02 Testing Classification Report

Label	precision	recall	f1-score
NO	0.46	0.60	0.52
DIRECT	0.40	0.33	0.25
REPORTED	0.00	0.00	0.00
JUDGEMENTAL	0.72	0.85	0.78
accuracy			0.64

6.1.3. Task 03

The classification report in Table 4 provides an overview of the model's performance for each label in the classification task. The precision, recall and f1-score values indicate how well the model performed for each label. For the "NO" label, the model achieved a precision of 0.71, a recall of 0.72, and an f1-score of 0.71. This suggests that the model accurately identified 71% of the instances as "NO" and correctly recalled 72% of the actual "NO" instances.

However, for the other labels such as "IDEOLOGICAL-INEQUALITY," "STEREOTYPING-DOMINANCE," "OBJECTIFICATION," "SEXUAL-VIOLENCE," and "MISOGYNY-NON-SEXUAL-VIOLENCE," the model's performance was relatively lower. The precision, recall, and f1-score values for these labels indicate that the model struggled to accurately classify instances in these categories, with scores ranging from 0.15 to 0.35 for precision, 0.08 to 0.39 for recall, and 0.11 to 0.37 for f1-score.

The overall accuracy of the model was reported as 0.51, meaning that it correctly classified 51% of the instances. These results suggest that the model had relatively better performance in identifying instances as "NO" but faced challenges in accurately classifying the other labels.

6.2. Final Result

This section presents the evaluation methodology and metrics utilized for each task in the EXIST 2023 competition. Three types of evaluations are performed for each task, and the

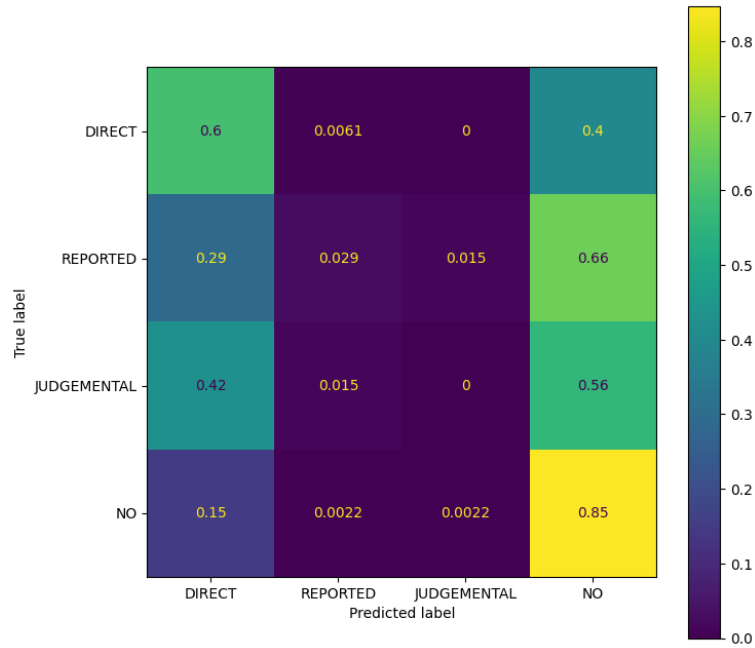


Figure 3: Task 02 Confusion Matrix

Table 4

Task 03 Testing Classification Report

Label	precision	recall	f1-score
NO	0.71	0.72	0.71
IDEOLOGICAL-INEQUALITY	0.35	0.39	0.37
STEREOTYPING-DOMINANCE	0.23	0.29	0.26
OBJECTIFICATION	0.23	0.18	0.20
SEXUAL-VIOLENCE	0.23	0.23	0.23
MISOGYNY-NON-SEXUAL-VIOLENCE	0.15	0.08	0.11
accuracy			0.51

official metric used across all evaluation contexts is the Information Contrast Measure (ICM). Additionally, details about the evaluation package, including the Python script and the contents of the evaluation folder, are provided. Different evaluation metrics are employed for the three tasks in EXIST 2023 based on the classification problems' nature and the hierarchical structure of the categories involved. It also presents a comprehensive analysis of the results obtained from various runs and variants, highlighting the performance based on the ICM-Soft and ICM-Hard scores.

Task 1: Sexism Identification Task 1 requires binary classification to identify sexism. The evaluation metric for this task is mono-label classification. To determine the ground truth labels, a "hard" setting is adopted, where the majority vote from human annotators is used. In this setting, the class annotated by more than three annotators is selected as the ground truth label.

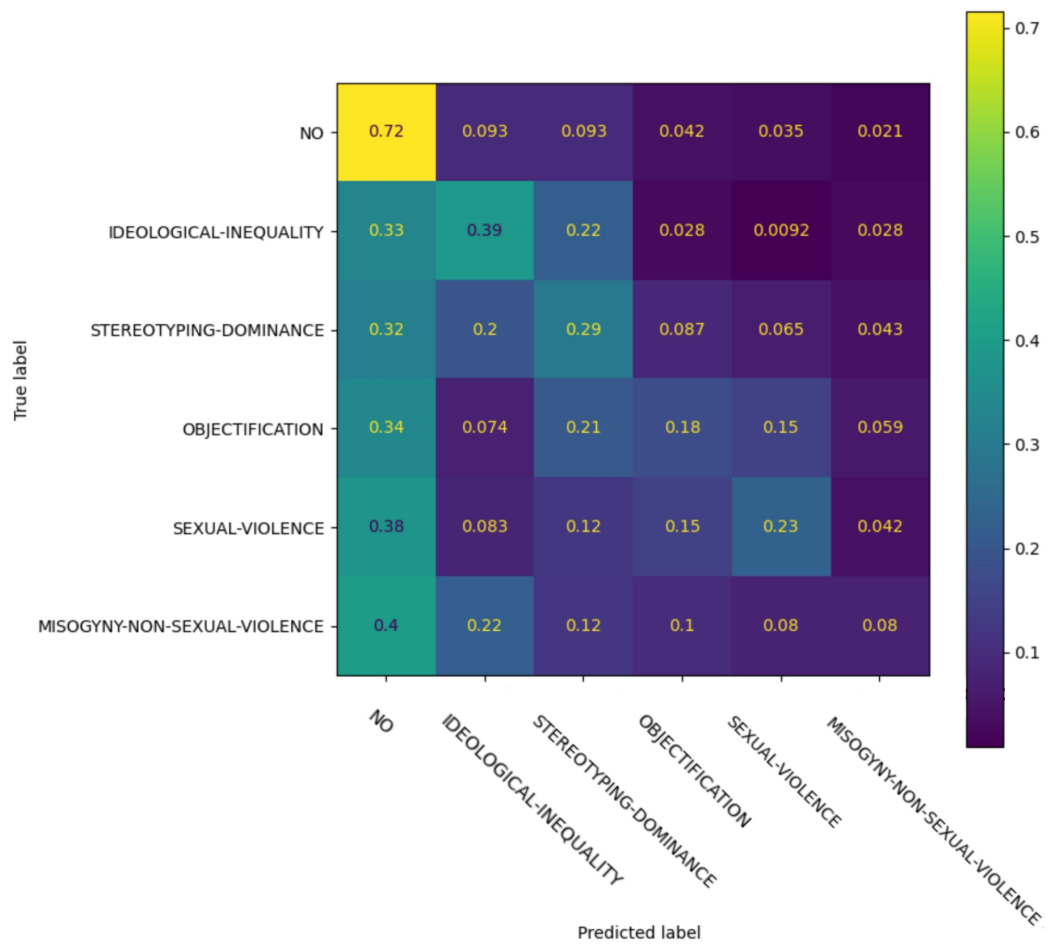


Figure 4: Task 03 Confusion Matrix

The evaluation is performed in both "hard-hard" and "hard-soft" contexts. The ICM serves as the official metric for Task 1.

Task 2: Source Intention Task 2 focuses on multiclass hierarchical classification, specifically categorizing the source intention as either sexist or not sexist, with further subcategorization into direct, reported, and judgmental. The evaluation metric for this task considers the severity of confusion between different categories. In the "hard" setting, the class annotated by more than two annotators is chosen as the ground truth label. The evaluation is conducted in all three contexts: "hard-hard," "hard-soft," and "soft-soft." The ICM is the official metric for Task 2.

Task 3: Sexism Categorization Task 3 involves multiclass hierarchical classification with

multi-label assignments, where a tweet may belong to multiple subcategories simultaneously. Similar to Task 2, the evaluation metric for this task considers the hierarchical structure and the possibility of multiple labels. The ground truth labels are determined using a "hard" setting, selecting the labels assigned by more than one annotator. The evaluation is performed in all three contexts: "hard-hard," "hard-soft," and "soft-soft." The official metric for Task 3 is the ICM, which is extended to ICM-soft to accommodate soft system outputs and ground truth assignments.

Evaluation Variants for Each Task The evaluation is conducted in three different modes for each task: "hard-hard," "hard-soft," and "soft-soft." In the "hard-hard" evaluation, systems that provide a conventional hard output are evaluated using hard ground truth labels. The official metric used to measure the system's performance is the ICM. Additionally, F1 scores are calculated and reported for comparison purposes, considering task-specific considerations.

For systems that provide a hard output, the "hard-soft" evaluation is performed. In this variant, the categories assigned by the system are compared with the probabilities assigned to each category in the ground truth. The official evaluation metric used in this context is the ICM-soft. Probabilities for each class are calculated based on the distribution of labels and the number of annotators for each instance.

The "soft-soft" evaluation is conducted for systems that provide probabilities for each category. In this context, the system's probabilities are compared with the probabilities assigned by the human annotators. Similar to the "hard-soft" evaluation, the ICM-soft metric is used. Additional metrics may be reported in the final evaluation report.

The use of ICM and ICM-soft metrics in the evaluation process ensures the consideration of the hierarchical structure of categories and the possibility of multiple labels, providing a superior analytical evaluation framework compared to alternatives in the current state of the art.

6.2.1. Task 01

In Task 01, Table 5 represents the results obtained from different runs and variants of a submission, comparing them based on the ICM-Soft and ICM-Hard scores. In the Soft-Soft (All) variant, the gold data achieved a score of 3.1182 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. However, the CNLP-NITS-PP submission obtained a score of -0.4237 for ICM-Soft Norm, ranking at 36. Shifting to the Hard-Hard (All) variant, the gold data did not provide a score for ICM-Soft, but it achieved a score of 0.9948 for ICM-Hard. On the other hand, the CNLP-NITS-PP submission obtained a score of 0.1093 for ICM-Hard Norm, ranking at 56. Regarding the Hard-Soft (All) variant, the gold data achieved the same scores as the Soft-Soft (All) variant, while the CNLP-NITS-PP submission obtained a score of -0.9509 for ICM-Soft Norm, ranking at 57.

Moving on to the Soft-Soft (ES) variant, the gold data achieved a score of 3.1177 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. However, the CNLP-NITS-PP submission obtained a score of -0.7839 for ICM-Soft Norm, ranking at 43. For the Hard-Hard (ES) variant, the gold data did not provide a score for ICM-Soft, but it achieved a score of 0.9999 for ICM-Hard. Conversely, the CNLP-NITS-PP submission obtained a score of -0.0382 for ICM-Hard Norm, ranking at 58.

In the case of the Hard-Soft (ES) variant, the gold data achieved the same scores as the Soft-Soft (ES) variant, while the CNLP-NITS-PP submission had a score of -1.1877 for ICM-Soft Norm, ranking at 59.

Now considering the Soft-Soft (EN) variant, the gold data achieved a score of 3.1141 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. Unfortunately, the scores for the CNLP-NITS-PP submission are not available for this variant. For the Hard-Hard (EN) variant, the gold data did not provide a score for ICM-Soft, but it achieved a score of 0.9798 for ICM-Hard. Conversely, the CNLP-NITS-PP submission obtained a score of 0.2508 for ICM-Hard Norm, ranking at 53. Lastly, for the Hard-Soft (EN) variant, the gold data achieved the same scores as the Soft-Soft (EN) variant, while the CNLP-NITS-PP submission had a score of -0.7779 for ICM-Soft Norm, ranking at 54.

Table 5
Task 01 Evaluation Result

Variants	Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm
Soft-Soft (All)	A	0	3.1182	-	1	-
	C	41	-0.5775	-	0.4029	-
Hard-Hard (All)	B	0	-	0.9948	-	1
	C	20	-	0.1093	-	0.4356
Hard-Soft (All)	A	0	3.1182	-	1	-
	C	8	-0.9509	-	0.3426	-
Soft-Soft (ES)	A	0	3.1177	-	1	-
	C	43	-0.7839	-	0.3145	-
Hard-Hard (ES)	B	0	-	0.9999	-	1
	C	23	-	-0.0382	-	0.3127
Hard-Soft (ES)	A	0	3.1177	-	1	-
	C	11	-1.1877	-	0.2436	-
Soft-Soft (EN)	A	0	3.1141	-	1	-
	C	36	-0.4237	-	0.4895	-
Hard-Hard (EN)	B	0	-	0.9798	-	1
	C	19	-	0.2508	-	0.5567
Hard-Soft (EN)	A	0	3.1141	-	1	-
	C	7	-0.7779	-	0.4384	-
EXIST2023_test_gold_soft = A; EXIST2023_test_gold_hard = B; CNLP-NITS-PP = C						

6.2.2. Task 02

For Task 02, Table 6 In terms of the Soft-Soft (All) variant, the gold data achieved a score of 6.2057 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. The CNLP-NITS-PP submission scores are not provided for this variant. Moving on to the Hard-Hard (All) variant, the gold data scored 1.5378 for ICM-Hard, while the CNLP-NITS-PP submission obtained a score of -0.3601 for ICM-Hard Norm, ranking at 24. For the Hard-Soft (All) variant, the gold data achieved the same scores as the Soft-Soft (All) variant, whereas the CNLP-NITS-PP submission obtained a negative score of -7.2467 for ICM-Hard and ICM-Hard Norm, ranking at 20.

Next, looking at the Soft-Soft (ES) variant results, the gold data achieved a score of 6.2431 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. However, the CNLP-NITS-PP submission scores are not available for this variant. For the Hard-Hard (ES) variant, the gold data scored

1.6007 for ICM-Hard, while the CNLP-NITS-PP submission obtained a score of -0.5075 for ICM-Hard Norm, ranking at 24. In the case of the Hard-Soft (ES) variant, the gold data achieved the same scores as the Soft-Soft (ES) variant, but the CNLP-NITS-PP submission had a score of -7.0396 for ICM-Hard, ranking at 23.

Moving on to the Soft-Soft (EN) variant, the gold data obtained a score of 6.1178 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. The CNLP-NITS-PP submission scores are not provided for this variant. For the Hard-Hard (EN) variant, the gold data scored 1.4449 for ICM-Hard, while the CNLP-NITS-PP submission obtained a score of -0.1945 for ICM-Hard Norm, ranking at 23. Lastly, for the Hard-Soft (EN) variant, the gold data achieved the same scores as the Soft-Soft (EN) variant, whereas the CNLP-NITS-PP submission had a score of -7.9564 for ICM-Hard, ranking at 13.

Table 6
Task 02 Evaluation Result

Variants	Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm
Soft-Soft (All)	A	0	6.2057	-	1	-
	C	-	-	-	-	-
Hard-Hard (All)	B	0	-	1.5378	-	1
	C	24	-	-0.3601	-	0.5955
Hard-Soft (All)	A	0	6.2057	-	1	-
	C	20	-7.2467	-	-7.2467	-
Soft-Soft (ES)	A	0	6.2431	-	1	-
	C	-	-	-	-	-
Hard-Hard (ES)	B	0	-	1.6007	-	1.6007
	C	24	-	-0.5075	-	0.5356
Hard-Soft (ES)	A	0	6.2431	-	1	-
	C	23	-7.0396	-	0.62	-
Soft-Soft (EN)	A	0	6.1178	-	1	-
	C	-	-	-	-	-
Hard-Hard (EN)	B	0	-	1.4449	-	1
	C	23	-	-0.1945	-	0.6666
Hard-Soft (EN)	A	0	6.1178	-	1	-
	C	13	-7.9564	-	0.6914	-
EXIST2023_test_gold_soft = A; EXIST2023_test_gold_hard = B; CNLP-NITS-PP = C						

6.2.3. Task 03

In task 03, Table 7 presents the results from different runs and variants of a submission. In terms of the Soft-Soft (All) variant, the gold data achieved an impressive score of 9.4686 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. However, the scores for the CNLP-NITS-PP submission are not available for this variant. Shifting to the Hard-Hard (All) variant, the gold data did not provide a score for ICM-Soft, but it scored 2.1533 for ICM-Hard. On the other hand, the CNLP-NITS-PP submission obtained a score of -0.8412 for ICM-Hard Norm, ranking at 20. Regarding the Hard-Soft (All) variant, the gold data achieved the same scores as the Soft-Soft (All) variant, while the CNLP-NITS-PP submission obtained a score of -11.3206 for ICM-Hard, ranking at 8. Moving on to the Soft-Soft (ES) variant, the gold data achieved a score of 9.6071 for ICM-Soft and a perfect score of 1 for ICM-Soft Norm. However, the CNLP-NITS-PP submission

scores are not provided for this variant. For the Hard-Hard (ES) variant, the gold data did not provide a score for ICM-Soft, but it obtained a score of 2.2393 for ICM-Hard. On the other hand, the CNLP-NITS-PP submission obtained a score of -1.115 for ICM-Hard Norm, ranking at 23. In the case of the Hard-Soft (ES) variant, the gold data achieved the same scores as the Soft-Soft (ES) variant, but the CNLP-NITS-PP submission had a score of -10.9613 for both ICM-Hard and ICM-Hard Norm, ranking at 11.

Table 7

Task 03 Evaluation Result

Variants	Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm
Soft-Soft (All)	A	0	9.4686	-	1	-
	C	-	-	-	-	-
Hard-Hard (All)	B	0	-	2.1533	-	1
	C	20	-	-0.8412	-	0.4332
Hard-Soft (All)	A	0	9.4686	-	1	-
	C	8	-11.3206	-	0.6259	-
Soft-Soft (ES)	A	0	9.6071	-	1	-
	C	-	-	-	-	-
Hard-Hard (ES)	B	0	-	2.2393	-	1
	C	23	-	-1.115	-	0.3966
Hard-Soft (ES)	A	0	9.6071	-	1	-
	C	11	-10.9613	-	-10.9613	-
Soft-Soft (EN)	A	0	9.1255	-	1	-
	C	-	-	-	-	-
Hard-Hard (EN)	B	0	-	2.0402	-	1
	C	19	-	-0.5318	-	-0.5318
Hard-Soft (EN)	A	0	9.1255	-	1	-
	C	7	-11.8244	-	-11.8244	-
EXIST2023_test_gold_soft = A; EXIST2023_test_gold_hard = B; CNLP-NITS-PP = C						

7. Conclusion

Sexism is a pervasive issue that continues to affect individuals and society as a whole. The EXIST 2023 shared task has provided a valuable platform for advancing the understanding and detection of sexism in social networks. Significant progress has been made in identifying and combating sexism online through the application of automated methods, such as the Generative Pre-trained Transformer 2 (GPT-2) language model. GPT-2's ability to capture semantic and contextual information from text has proven to be a powerful tool in classifying sexist content. However, the results and discussions from the tasks indicate that there is still room for improvement in the accuracy and precision of the models.

The EXIST 2023 shared task results indicate that the models' performance varied across different tasks and labels. In Task 1, the models showed promising performance in distinguishing between sexist and non-sexist content, but further improvements are needed to enhance precision and recall for both categories. In Task 2, the models struggled with accurately classifying instances into the "DIRECT" and "REPORTED" labels, indicating the need for more refined approaches to handle these categories effectively. Similarly, in Task 3, the models faced

challenges in categorizing instances into specific subcategories of sexism, such as ideological inequality, stereotyping dominance, objectification, sexual violence, misogyny and non-sexual violence. These findings highlight the complexity and nuances of detecting and categorizing sexism in social networks.

Future work should focus on refining and developing models that can address the specific challenges identified in the shared task. Improving the accuracy and precision of the models in distinguishing between different forms of sexism will be crucial in advancing the field. Additionally, efforts should be made to expand the datasets and incorporate more diverse examples to improve the models' generalization capabilities. Collaboration between researchers, experts, and practitioners will be essential in advancing the field and developing more effective automated systems for combating sexism in social networks. Ultimately, the EXIST 2023 shared task has provided valuable insights and a foundation for further research, bringing us closer to addressing the pervasive issue of sexism in online platforms and promoting a more inclusive and equal society.

Acknowledgments

We would like to express our gratitude to the National Institute of Technology Silchar for allowing us to conduct our research and experimentation. We are thankful for the resources and research atmosphere provided by the CNLP & AI Lab, NIT Silchar.

References

- [1] D. H. Felmlee, C. Julien, S. C. Francisco, Debating stereotypes: Online reactions to the vice-presidential debate of 2020, *PloS one* 18 (2023) e0280828.
- [2] L. Jussim, J. T. Crawford, S. M. Anglin, S. T. Stevens, Ideological bias in social psychological research, *Social psychology and politics* (2015) 107–126.
- [3] C. J. Burns, L. Sinko, Restorative justice for survivors of sexual violence experienced in adulthood: A scoping review, *Trauma, Violence, & Abuse* 24 (2023) 340–354.
- [4] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023*, pp. 593–599.
- [5] A. Di Vaio, R. Hassan, R. Palladino, Blockchain technology and gender equality: A systematic literature review, *International Journal of Information Management* 68 (2023) 102517.
- [6] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing., *Neurocomputing* (2023) 126232.
- [7] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, *IEEE Access* 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.
- [8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI blog* 1 (2019) 9.

- [9] OpenAI, Gpt-4 technical report, 2023. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [10] P. Budzianowski, I. Vulić, Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems, [arXiv preprint arXiv:1907.05774](https://arxiv.org/abs/1907.05774) (2019).
- [11] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do not have enough data? deep learning to the rescue!, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7383–7390.
- [12] M. E. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: International Conference on Applications of Natural Language to Data Bases, 2018.
- [13] S. Frenda, B. Ghanem, M. M. y Gómez, P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, *J. Intell. Fuzzy Syst.* 36 (2019) 4743–4752.
- [14] D. Grosz, P. C. Céspedes, Automatic detection of sexist statements commonly used at the workplace, [ArXiv abs/2007.04181](https://arxiv.org/abs/2007.04181) (2020).
- [15] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, [arXiv](https://arxiv.org/abs/1412.1967) (2014).
- [16] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>. doi:10.3115/v1/D14-1162.
- [17] G. Alexandridis, I. Varlamis, K. Korovesis, G. Caridakis, P. Tsantilas, A survey on sentiment analysis and opinion mining in greek social media, *Information* 12 (2021) 331.
- [18] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, [arXiv preprint arXiv:2205.01068](https://arxiv.org/abs/2205.01068) (2022).
- [19] L. Mathew, V. Bindu, A review of natural language processing techniques for sentiment analysis using pre-trained models, in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2020, pp. 340–345.
- [20] A. Vaca-Serrano, Detecting and classifying sexism by ensembling transformers models, *language* 2 (2022) 1.