# Cracking Down on Digital Misogyny with MULTILATE a MULTImodaL hATE Detection System

Notebook for the Exist 2024 Lab at CLEF 2024

Advaitha Vetagiri[1,*], Prateek Mogha[2] and Partha Pakray[1]

[1]Department of Computer Science and Engineering, National Institute of Technology Silchar, Assam, 788010, India
[2]Department of Electrical Engineering, National Institute of Technology Silchar, Assam, 788010, India

## Abstract

Sexism in social networks manifests in various forms, from blatant misogyny to subtle, implicit biases, presenting a significant societal challenge that necessitates effective detection and mitigation strategies. Addressing this issue involves participation in the EXIST 2024 tasks, a competition designed to advance the identification of sexist content in social media. This year's contest includes both traditional text-based data from tweets and an innovative meme dataset, incorporating both images and text. The approach leverages sophisticated models to analyze these multimodal inputs. For textual modalities, a Convolutional Neural Network - Bidirectional Long Short-Term Memory model is employed to discern sexist language and tweet behaviours. For image modalities, a combination of Residual Network 50 and text-based analysis is utilized to detect and interpret sexist elements within memes. Both models undergo hyperparameter tuning and k-fold cross-validation to ensure robustness and accuracy. Preliminary results indicate that integrating these methods enhances the precision and effectiveness of sexism detection, providing a comprehensive tool for identifying and addressing sexist content in diverse social media formats.

## Keywords

EXIST 2024, Convolutional Neural Networks - Bidirectional Long Short-Term Memory, Residual Network 50, Sexism Detection, Sexist Content

## 1. Introduction

The phenomenon of sexism remains a crucial problem in the contemporary world at large; in turn, sexism encompasses stereotyped perceptions [1], ideological prejudice, and even acts of bigotry in relation to both male and female sex [2]. As the internet continues to assert its presence in people's lives, especially within the context of social discourse through social networks, it becomes important to track and tackle sexism in these platforms. The Conference and Labs of the Evaluation Forum (CLEF) sponsored shared task for sEXism Identification in Social neTworks (EXIST) 2024 [3] explores the potential of developing and improving the methodology and tools to effectively detect and classify sexism in language, bringing together researchers from various disciplines.

Sexism is one of the most multifaceted and consequential concepts because, in fact, it affects not only individuals but also entire societies, preserving gender inequalities, limiting personal and social opportunities, and strengthening discourses and representations of gendered divisions. This ultimately leads to the marginalization of women and the enhancement of inequalities, hence delaying the achievement of gender equity. Categorical and severe sex-related hatred has long been the main object of concern in the field of research [4], but recently, there has been an increased understanding and recognition of the need for a constant investigation of a range of sexist manifestations [5]. The goals set by the EXIST campaigns are to include both blatant and subtle instances of sexism to embrace the full range of its expression and the ways that it could be encountered daily.

*CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France*

[*]Corresponding author.

advaitha21_rs@cse.nits.ac.in (A. Vetagiri); prateek_ug@ee.nits.ac.in (P. Mogha); partha@cse.nits.ac.in (P. Pakray)
http://www.parthapakray.com/ (P. Pakray)
0000-0002-0651-4171 (A. Vetagiri); 0000-0003-3834-5154 (P. Pakray)

The chief aim of EXIST [6] is to encompass a wide array of sexist expressions, ranging from blatant misogyny to more nuanced and implicit behaviours. This initiative has steadily evolved since its inception, with the 2024 edition introducing fresh challenges and broadening its scope to include multimodal data. The fourth iteration, hosted at the University of Grenoble Alpes, France, from September 9-12, 2024, builds on the groundwork established in previous years while incorporating novel elements to further enhance detection capabilities.

A significant addition to this year's challenge is incorporating a meme dataset alongside the traditional tweet dataset. Memes, which blend images and text to convey humour or commentary, present distinct challenges due to their multimodal nature and the subtleties involved in their interpretation. The EXIST 2024 [7] task aspires to develop robust models adept at identifying sexist content within both textual and visual contexts.

In order to address these issues, this paper proposes a multifaceted approach that includes CNN and BiLSTM models [8]. CNNs, especially those designed for image processing, are proficient at identifying localized representations and characteristics of text as well as images, which makes them appropriate for categorizing the subtle forms of expression identified in the form of memes [9]. Long-range dependencies and context information in textual content make BiLSTM networks outperform on the other hand. This Duplex architecture aligns itself with the best aspects of both architectures, providing a complete answer to the task as nuanced as sexism detection. The structure provided by the CNN-BiLSTMs [10] of the hybrid model allows for capturing fine-grained features and contextual information, which is critical in the identification and categorization of instances of sexism. This [11] present work contributes to improving current approaches by proposing a new tool that enables automatic identification and categorization of stereotypically sexual comments in social media compared to previous work using a Generative Pre-trained Transformer 2 (GPT-2) [12].

OpenAI [13] owns GPT-2 [14], a language model that marks a severe improvement in the Natural Language Processing (NLP) area. Being a member of the 12 GPT generation learning models, GPT-2 utilizes a unique neural network design and copious amounts of training data to produce human-like text. The actual upbringing of GPT-2 is done on a vast dataset of the language available on the internet, which lays down the statistical probabilities and features of language in its structure. This complex structure allows it to understand contextual details, thus making it capable of creating coherent and contextually appropriate text from the input given to it. In particular, GPT-2 has been used in the past as an effective means for the automatic identification and classification of subdivisions containing sexism. This is achieved by training GPT-2 on a dataset that is marked to contain sexist and non-sexist [15] text, hence making it acquire prior knowledge on the typical characteristics of text samples that might contain sexist tendencies.

Recently, the CNN-BiLSTM model was found to outperform the previous GPT-2 model to be used in the EXIST 2023 task. CNN-BiLSTMs, which are specifically designed for text classification tasks, appear more effective at identifying sexist language and behaviours. The improvement of this model in the identification of contextual and semantic features of a text also contributes towards better identification of sexism. Moreover, ResNet50's combination with the textual model called *"MULTIHATE"* [16] made a substantial enhancement in analyzing sexism embedded in memes. They have both been hyperoptimized and tested using k-fold cross-validation (CV), thereby presenting them as very reliable models. Altogether, these have made sexism detection more accurate and efficient than it was with the help of the previous GPT-2 model.

## 2. Literature Survey

Sexism, a pervasive issue in society, is defined as discrimination based on sex or gender, especially against women and girls. Sexism can be a belief that one sex is superior to another sex. It imposes limits on what men should do and what women should do. Sexism in a society is most commonly applied against women and girls due to patriarchy or male domination. The problem of sexism extends beyond individual discrimination to systemic inequalities that affect various aspects of the life of a woman, including employment, education, and social interactions. Researchers have identified multiple categories and impacts of sexism in society, leading to a broad body of literature exploring its various dimensions and proposing methods for its identification and mitigation, especially in the field of deep learning

## 2.1. Negative Effects of Sexism

The consequences of sexism are far-reaching and detrimental. Sexism contributes to gender inequality in the workplace, limiting opportunities for women in terms of promotions, salaries, and job roles. This inequality is often perpetuated through both discrimination and more subtle biases that affect hiring practices and workplace culture. Educational disparities also emerge, with sexist attitudes influencing the subjects that individuals are encouraged to pursue, often steering women away from STEM fields. Furthermore, sexism can lead to unequal access to resources and support systems, worsening the challenges faced by women.

Moreover, studies showing that persistent exposure to sexist attitudes and behaviours can contribute to sexism can lead to depression, post-traumatic stress disorder, lower self-esteem, and a heightened risk of mental health issues [17]. This pervasive issue affects not only individuals but also the broader society by perpetuating gender inequalities and limiting the potential contributions of all its members.

## 2.2. Categories of Sexism in EXIST 2024

Ideological and Inequality: Ideological sexism often manifests in cultural norms and legal systems reinforcing gender disparities. For example, some historical and cultural narratives portray females as inferior to males in terms of their abilities.

Stereotyping and Dominance: A gender stereotype is a generalized view or perception about attributes or characteristics that should be possessed by women, or that should be performed by men and women; stereotypes can be both positive and negative, such as ‘women are nurturing’ and ‘women are weak’. Stereotyping stops women from moving up by creating unfair doubts about their skills and leadership. This blocks their chances for promotion and equal recognition at work. [18] Dominance, on the other hand, manifests in power dynamics, where men are considered to be dominant over women.

Objectification: Objectification means viewing or treating individuals as objects, reducing them to their physical appearance. This form of sexism is more common in media representations and advertising [19], where women are frequently depicted in ways that only focus on their physical attributes over their skills or personalities. Objectification can lead to dehumanization, where women are valued less for their personality or work and more for their appearance.

Sexual Violence: This severe form of sexism is when an individual is forced or manipulated into unwanted sexual activity without their consent. This includes sexual assault or rape, harassment, exploitation, public flashing and watching someone in a private act without their knowledge or permission. Sexual violence can have a permanent effect on a woman’s life, which can lead to depression [20].

Misogyny and Non-Sexual Violence: Misogyny refers to hatred towards women; it is a form of sexism that can keep women at a lower social status than men [21]. Misogyny has taken various forms, such as discrimination, objectification, belittlement, or violence, and often stems from deeply rooted societal attitudes and stereotypes about gender roles and power dynamics. Non-sexual violence includes behaviours such as verbal abuse, threats, and other forms of intimidation that are not explicitly sexual but are driven by gender bias [22].

## 2.3. Identification of Sexism

With advancements in technology, particularly in the domain of machine learning (ML) and deep learning (DL), Many methods have been developed to identify and analyze instances of sexism on the digital platform.

Machine learning techniques, such as support vector machines (SVM) and random forests, have been applied to classify and detect sexist content in text data. These methods rely on feature extraction and supervised learning to differentiate sexist remarks from non-sexist ones. By training models on labelled datasets, researchers can develop systems that automatically identify and categorize sexist language [23] in both Spanish and English. These models use various textual features, such as word frequencies, n-grams, and syntactic structures, to distinguish between different types of sexist content. Similarly,[24] used two datasets in a similar manner to identify online hate speech directed towards women.

Deep learning approaches have further enhanced the ability to identify sexism by automatically learning hierarchical representations of text data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are commonly employed in this domain. CNNs effectively capture local patterns in text, while RNNs, particularly Long Short-Term Memory (LSTM) networks, excel at modelling sequential dependencies and context [25]. These methods have been shown to outperform traditional ML techniques in various natural language processing tasks, including sentiment analysis and hate speech detection .[26]

To tackle the issue of sexism in memes, researchers have applied CNN-BiLSTM models for image classification. CNNs are employed to extract features from the image components of memes, while BiLSTM networks process the textual content, capturing both spatial and contextual information effectively [27]. This combination allows for a comprehensive analysis of memes, considering both visual and textual cues to accurately identify sexist content. Integrating these models enables the detection of subtle and complex forms of sexism that may not be apparent through text or image analysis alone.

ResNet50 is a deep residual network that has shown superior performance in image recognition [28] and can be used for identifying sexist content in memes. It handles the complexities of image data, making it well-suited for detecting subtle visual cues that indicate sexism. Hence, it can be used to differentiate between hateful and not-hateful memes when combined with other models such as LSTM [29].

## 3. Dataset

Since 2021, the primary aim of the EXIST campaigns has been the detection of sexism in tweets [30, 31, 32]. Over the years, three distinct corpora of annotated tweets have been amassed for various EXIST tasks. In line with this tradition, the focus of EXIST 2024 remains on identifying sexism in textual content, utilizing the EXIST 2023 [32] dataset, and expanding to encompass memes. Memes, which are images typically adorned with text captions, often carry humour and circulate widely on social media, forums, and other digital platforms. These memes can serve as vehicles for misinformation, perpetuate stereotypes, or degrade individuals. For EXIST 2024, a comprehensive lexicon of terms and expressions indicative of sexist memes has been meticulously curated, drawing from expressions that have proven effective in identifying sexism in previous EXIST editions. This lexicon includes a diverse array of topics, incorporating terms used in both sexist and non-sexist contexts, all centred around women. The final compilation includes 250 terms, with 112 in English and 138 in Spanish.

**Table 1**
Dataset Split Statistics for Tweets [32]

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

**Table 2**
Dataset Split Statistics for Memes [6]

| | Train | Test |
|---|---|---|
| Spanish | 2046 | 540 |
| English | 2010 | 513 |
| Total | 4056 | 1053 |

### 3.1. Crawling

These terms were employed as search queries on Google Images to retrieve the top 100 images. Through rigorous manual curation, efforts were made to define memes accurately and eliminate noise, such as

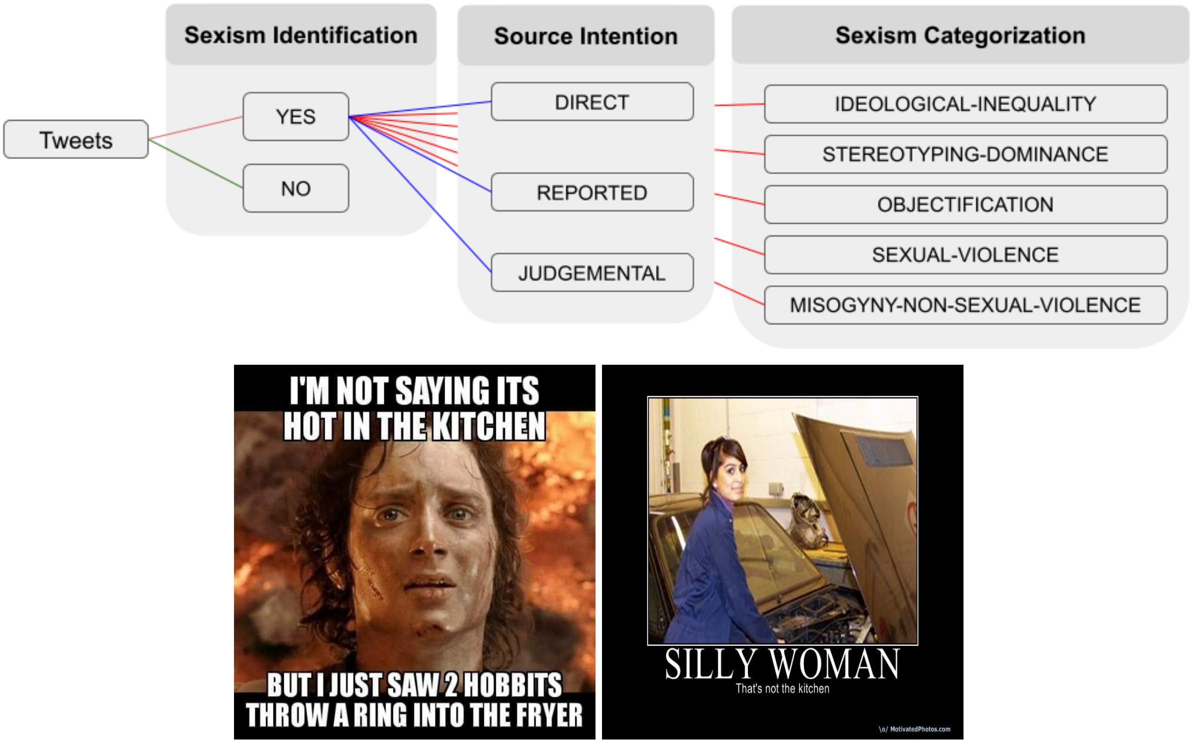


**Figure 1:** Comprehensive Visualization of Dataset Classes: An Overview of Various Categories Including Non-Sexist and Sexist Examples.

images without text, text-only images, advertisements, and duplicates. The final collection comprises over 3,000 memes per language. Given the heterogeneous proportion of memes per term, the most unbalanced seeds were discarded to ensure that each seed had at least five memes. Furthermore, the final dataset was curated to achieve the most equitable distribution of memes per seed. To avoid selection bias, memes were randomly chosen, adhering to the appropriate distribution per seed. Consequently, the training set contains more than 2,000 memes per language, while the test set includes over 500 memes per language.

### 3.2. Labeling Process

As with the previous edition, potential sources of *"label bias"* have been carefully considered. Label bias can stem from the socio-demographic differences of the individuals involved in the annotation process, as well as from scenarios where multiple correct labels exist or where labelling decisions are highly subjective, as shown in Figure 1. To mitigate label bias, two different social and demographic parameters were considered: gender (MALE/FEMALE) and age (18-22 years/23-45 years/46+ years). Each meme was annotated by 6 crowdsourcing annotators selected via the Prolific app, following guidelines established by two gender issues experts. As an added feature in the datasets, both for 2023 and 2024, three additional demographic characteristics of each annotator have been included: level of education, ethnicity, and country of residence.

### 3.3. Learning with Disagreements

The notion that natural language expressions have a singular and clearly identifiable interpretation in any given context is an oversimplification, particularly in the realm of highly subjective tasks like sexism detection. The learning with disagreements paradigm addresses this by enabling systems to learn from datasets without gold-standard annotations, instead providing information about all annotator responses to capture the diversity of perspectives. In line with methods proposed for training directly from data with disagreements, all annotations per instance from the 6 different strata of annotators will be provided rather than using an aggregated label.

## 4. System Overview

In the EXIST 2024 shared task, the CNN-BiLSTM and ResNet50 models are employed for the detection and classification of sexism across various tasks. The competition encompasses six distinct tasks: Task 1 focuses on identifying sexism (binary classification), Task 2 involves classifying the source intention (a multiclass hierarchical classification), Task 3 deals with categorizing sexism (multiclass hierarchical multi-label classification), Task 4 targets the identification of sexism in memes (binary classification), Task 5 categorizes the source intention in memes (multiclass classification), and Task 6 involves the categorization of sexism in memes (multiclass hierarchical multi-label classification).

For Tasks 1, 2, and 3, which concentrate on textual data, the CNN-BiLSTM model is utilized. This architecture merges Convolutional Neural Networks (CNN) for feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) networks for capturing sequential dependencies and contextual nuances. The CNN-BiLSTM model undergoes fine-tuning on annotated datasets to optimize its efficacy for each specific task. In Task 1, the CNN-BiLSTM model classifies text instances as either sexist (YES) or not sexist (NO). By training on a dataset labelled for sexism identification, the model discerns patterns and linguistic signals indicative of sexism. The output for Task 1 is a binary classification label that denotes whether the text is sexist or not. Task 2 employs the CNN-BiLSTM model for source intention classification. This hierarchical classification task first differentiates between sexist and non-sexist texts and subsequently categorizes the sexist texts into Direct, Reported, and Judgmental subcategories. The model learns to identify linguistic cues associated with each subcategory by training on a dataset annotated for source intention. The output for Task 2 provides the source intention classification label for each text instance. Task 3 involves sexism categorization, a multiclass hierarchical multi-label classification problem. The CNN-BiLSTM model undergoes fine-tuning on a dataset with sexism categorization annotations. The first classification level distinguishes between sexist and non-sexist text, while the second level includes subcategories such as Ideological-Inequality, Stereotyping-Dominance, Objectification, Sexual-Violence, and Misogyny-Non-Sexual-Violence. As Task 3 permits multiple subcategories for a single text instance, the model generates multi-label predictions, outputting probabilities or confidence scores for each subcategory.

For Tasks 4, 5, and 6, which concentrate on image and text data, the ResNet50 model for image and CNN-BiLSTM for text data as above is employed. ResNet50, a deep convolutional neural network with 50 layers, is utilized for its robust image feature extraction capabilities. The model undergoes fine-tuning on a curated dataset of memes to enhance its performance for each task. In Task 4, the ResNet50 model performs binary classification to determine whether a given meme is sexist. By training on labelled data, the model learns to recognize visual and textual elements indicative of sexism. The output for Task 4 is a binary classification label indicating whether the meme is sexist or not. Task 5 involves categorizing the source intention in memes and distinguishing between Direct and Judgmental intentions. The ResNet50 model learns to identify visual and textual cues associated with each subcategory by training on annotated meme datasets. The output for Task 5 provides the source intention classification label for each meme. Task 6 requires the ResNet50 model to categorize sexism in memes, a multiclass hierarchical multi-label classification problem. The model undergoes fine-tuning on a dataset annotated for sexism categorization in memes, learning to identify subcategories such as Ideological-Inequality, Stereotyping-Dominance, Objectification, Sexual-Violence, and Misogyny-Non-Sexual-Violence. The output for Task 6 provides probabilities or confidence scores for each subcategory, indicating the extent to which the meme belongs to each category.

The models effectively capture the intricate patterns associated with sexism by harnessing the combined strengths of CNN-BiLSTM for textual data and ResNet50 for image data. These models are meticulously fine-tuned through rigorous training and validation processes, achieving high performance in the classification and categorization tasks of EXIST 2024.

## 5. Experimental Setup

The experimental framework for the EXIST 2024 tasks entails a meticulous approach to model training, validation, and assessment. This section delineates the data preprocessing protocols, model configurations, hyperparameter optimization, and evaluation metrics employed to achieve peak performance across the six tasks.

In preparing the textual data for Tasks 1, 2, and 3, several preprocessing steps are undertaken. Initially, texts are tokenized into words utilizing the Global Vectors (GloVe) [33], ensuring consistent treatment of punctuation marks and special symbols. Following this, all text is converted to lowercase to maintain uniformity. Common stopwords are excised to minimize noise within the dataset. Sequences are then padded to a standardized length to facilitate batch processing, with excessively long texts truncated to manageable sizes.

For the image data in Tasks 4, 5, and 6, a different set of preprocessing procedures is followed. Images are resized to 224x224 pixels to match the input dimensions required by the ResNet50 model. Pixel values are normalized to fall within the range of [0, 1], standardizing the input data. To enhance model robustness, data augmentation techniques such as random rotation, flipping, and colour jitter are applied, augmenting the training dataset's variability.

The CNN-BiLSTM model, used for the textual tasks, is configured with a convolutional layer employing a filter size of 300d, followed by a ReLU activation function and max pooling. The bidirectional LSTM layer comprises 128 units, adept at capturing contextual information in both forward and backward directions. Subsequently, two fully connected layers with 64 and 32 units, respectively, are incorporated, with a dropout rate of 0.5 to mitigate overfitting. The output layer is tailored to the specific task: a single sigmoid neuron for Task 1 and softmax layers corresponding to the number of classes for Tasks 2 and 3.

For the image tasks, the ResNet50 model, as well as CNN-BiLSTM for text, makes them a multimodal approach. The ResNet50 model is initialized with pre-trained weights from ImageNet. A custom classification head replaces the original, incorporating a global average pooling layer, followed by dense layers with 128 and 64 units, respectively, and a dropout rate of 0.5. The output layer is similarly tailored, with a single sigmoid neuron for Task 4 and softmax layers corresponding to the number of classes for Tasks 5 and 6.

Hyperparameter tuning is conducted via grid search and k-fold cross-validation to pinpoint the optimal parameters for each model. Key hyperparameters, such as learning rate, batch size, and the number of epochs, are systematically varied. Learning rates are explored within the range of 1e-5 to 1e-3, batch sizes are tested at 32, and the number of epochs is determined based on early stopping criteria, with patience set to five epochs. The models' performance is evaluated using a variety of metrics. Accuracy, defined as the proportion of correct predictions, is used for both binary and multiclass classification tasks. Precision, recall, and F1-score are calculated to provide a nuanced understanding of the balance between false positives and false negatives. The training and validation processes utilize the training and development datasets, respectively, with final evaluations conducted on the test datasets. This experimental setup ensures a thorough and comprehensive approach to model development, striving to attain superior performance across all tasks in the EXIST 2024 competition.

## 6. Results & Discussion

In this section, the outcomes of the methodologies applied for Tasks 1-6 in EXIST 2024 using the training dataset, as well as the final results provided by the organizers of the Shared Task on the test dataset.

### 6.1. Training Results

The training outcomes provide a comprehensive evaluation of the employed methodologies, underscoring their efficacy in addressing the specified tasks.

### 6.1.1. Task 1, 2 & 3

For Task 1, which involves binary classification for sexism identification, the average classification report across five folds shows a precision of 0.72, a recall of 0.72, and an F1-score of 0.71. These metrics suggest a well-balanced model performance, where the precision and recall are in harmony, indicating consistent identification of sexist and non-sexist instances. This balance between precision and recall results in a robust F1-score, demonstrating the model's reliability in distinguishing between the two classes. The detailed performance metrics for Task 1 are summarized in Table 3, and the model's accuracy and loss curves are illustrated in Figure 2, while the confusion matrix is shown in Figure 3.

In Task 2, which addresses the hierarchical classification of source intention, the model achieved an average precision of 0.60, a recall of 0.65, and an F1-score of 0.61. The results reflect a reasonable performance, with recall slightly outperforming precision. This indicates the model's slightly better ability to identify all relevant instances of each class than its precision. However, the moderate precision points to some challenges in avoiding false positives, suggesting that further fine-tuning could enhance the model's specificity. The detailed performance metrics for Task 1 are summarized in Table 4, and the model's accuracy and loss curves are illustrated in Figure 4, while the confusion matrix is shown in Figure 5.

Task 3, which involves multiclass hierarchical multi-label classification for sexism categorization, exhibited average precision and recall scores of 0.58, with an F1-score also at 0.58 across five folds. These results point to the complexity of the task, where the model faces difficulties in accurately predicting multiple labels simultaneously. The consistent scores across precision, recall, and F1-score indicate that while the model is competent, there is room for improvement, particularly in refining its ability to handle multiple overlapping categories.5, and the model's accuracy and loss curves are illustrated in Figure 6, while the confusion matrix is shown in Figure 7.

**Table 3**
Task 01 Testing Classification Report

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| NO | 0.72 | 0.78 | 0.75 |
| YES | 0.74 | 0.67 | 0.70 |
| | **Accuracy** | | **0.73** |

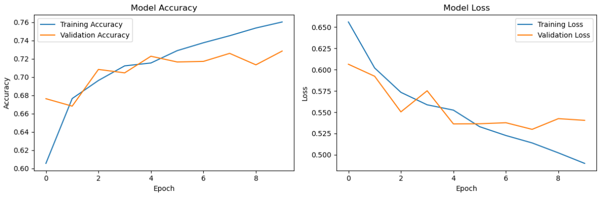


**Figure 2:** Task 01 Accuracy and Loss Curves

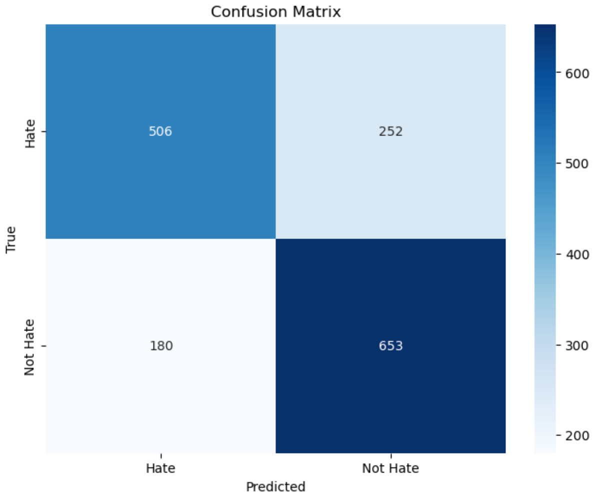


**Figure 3:** Task 01 Confusion Matrix

**Table 4**
Task 02 Testing Classification Report

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| NO | 0.16 | 0.04 | 0.07 |
| DIRECT | 0.41 | 0.16 | 0.23 |
| REPORTED | 0.46 | 0.57 | 0.51 |
| JUDGEMENTAL | 0.77 | 0.86 | 0.81 |
| **Accuracy** | | | **0.67** |

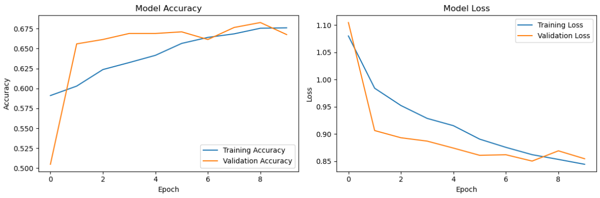


**Figure 4:** Task 02 Accuracy and Loss Curves

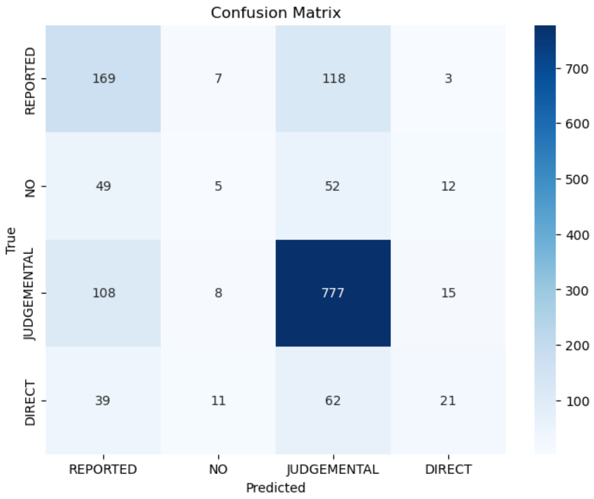


**Figure 5:** Task 02 Confusion Matrix

**Table 5**
Task 03 Testing Classification Report

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| NO | 0.67 | 0.04 | 0.08 |
| IDEOLOGICAL-INEQUALITY | 0.31 | 0.19 | 0.24 |
| STEREOTYPING-DOMINANCE | 0.71 | 0.94 | 0.81 |
| OBJECTIFICATION | 0.46 | 0.38 | 0.41 |
| SEXUAL-VIOLENCE | 0.30 | 0.11 | 0.17 |
| MISOGYNY-NON-SEXUAL-VIOLENCE | 0.56 | 0.06 | 0.11 |
| **Accuracy** | | | **0.66** |

### 6.1.2. Task 4, 5 & 6

Task 4 focused on image-based classification using the ResNet50 model, and the average classification report across five folds yielded a precision and recall of 0.63 and an F1-score of 0.62. These results highlight the model's consistent performance in classifying images accurately. The close alignment of precision and recall suggests that the model maintains a good balance between correctly identifying positive instances and minimizing false positives. The detailed performance metrics for Task 1 are summarized in Table 6, and the model's accuracy and loss curves are illustrated in Figure 8, while the confusion matrix is shown in Figure 9.

Task 5 presented a more challenging scenario, reflected in the lower average precision of 0.48, recall of 0.53, and an F1-score of 0.50 across five folds. These figures indicate that the model struggled with this task, likely due to the finer granularity required for distinguishing between closely related categories. The disparity between precision and recall suggests that the model identified many relevant instances

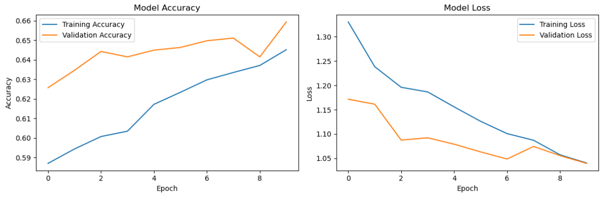
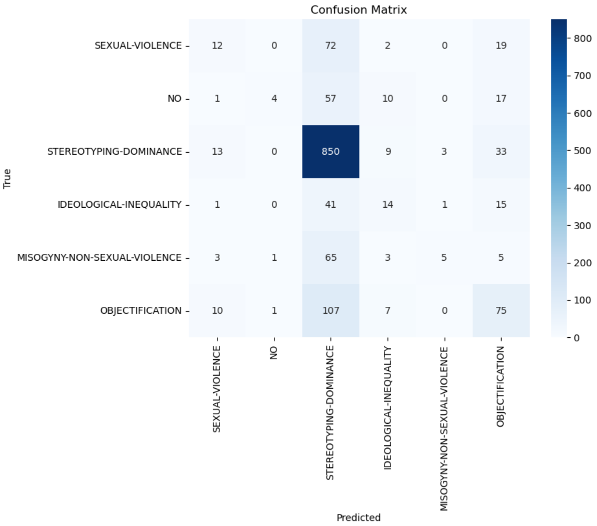


**Figure 6:** Task 03 Accuracy and Loss Curves



| | SEXUAL-VIOLENCE | NO | STEREOTYPING-DOMINANCE | IDEOLOGICAL-INEQUALITY | MISOGYNY-NON-SEXUAL-VIOLENCE | OBJECTIFICATION |
|---|---|---|---|---|---|---|
| SEXUAL-VIOLENCE | 12 | 0 | 72 | 2 | 0 | 19 |
| NO | 1 | 4 | 57 | 10 | 0 | 17 |
| STEREOTYPING-DOMINANCE | 13 | 0 | 850 | 9 | 3 | 33 |
| IDEOLOGICAL-INEQUALITY | 1 | 0 | 41 | 14 | 1 | 15 |
| MISOGYNY-NON-SEXUAL-VIOLENCE | 3 | 1 | 65 | 3 | 5 | 5 |
| OBJECTIFICATION | 10 | 1 | 107 | 7 | 0 | 75 |

**Figure 7:** Task 03 Confusion Matrix

but also produced a higher rate of false positives. The detailed performance metrics for Task 1 are summarized in Table 7, and the model's accuracy and loss curves are illustrated in Figure 10, while the confusion matrix is shown in Figure 11.

In Task 6, which also dealt with hierarchical multi-label classification using images, the model achieved an average precision and recall of 0.51, with an F1-score of 0.52. These results mirror the challenges seen in Task 3, underscoring the inherent difficulty of multi-label classification tasks. The equal precision and recall values reflect a balanced performance yet also highlight the need for further improvements to enhance the model's accuracy in handling complex, multifaceted data inputs. The detailed performance metrics for Task 1 are summarized in Table 8, and the model's accuracy and loss curves are illustrated in Figure 12, while the confusion matrix is shown in Figure 13.

**Table 6**
Task 04 Training Classification Report

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| NO | 0.67 | 0.69 | 0.68 |
| YES | 0.65 | 0.62 | 0.64 |
| | **Accuracy** | | **0.66** |

**Table 7**
Task 05 Training Classification Report

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| NO | 0.60 | 0.68 | 0.64 |
| DIRECT | 0.47 | 0.53 | 0.50 |
| JUDGEMENTAL | 0.32 | 0.06 | 0.10 |
| | **Accuracy** | | **0.54** |

The results across these tasks demonstrate the varying degrees of model effectiveness, with strengths in binary and simpler hierarchical classifications, but also revealed significant challenges in more complex multi-label tasks. These insights are crucial for directing future enhancements and refinements to improve overall performance.

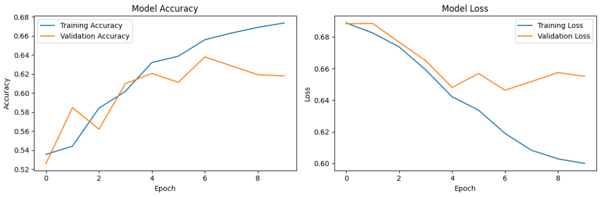


**Figure 8:** Task 04 Accuracy and Loss Curves

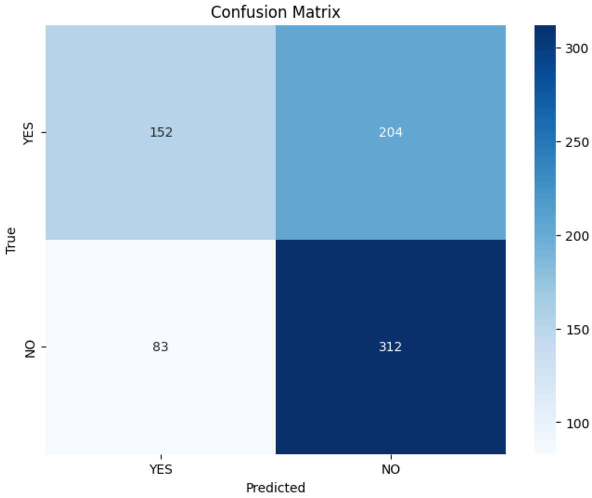


**Figure 9:** Task 04 Confusion Matrix

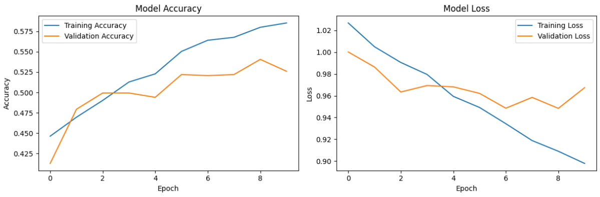


**Figure 10:** Task 05 Accuracy and Loss Curves

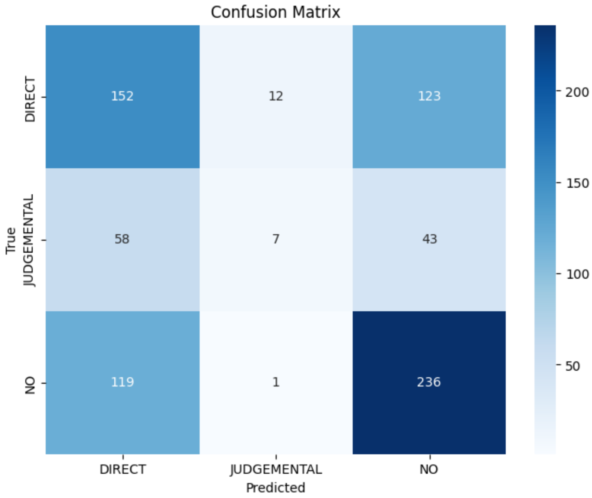


**Figure 11:** Task 05 Confusion Matrix

**Table 8**
Task 06 Training Classification Report

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| NO | 0.50 | 0.03 | 0.06 |
| IDEOLOGICAL-INEQUALITY | 0.28 | 0.15 | 0.20 |
| STEREOTYPING-DOMINANCE | 0.60 | 0.80 | 0.69 |
| OBJECTIFICATION | 0.40 | 0.32 | 0.36 |
| SEXUAL-VIOLENCE | 0.25 | 0.10 | 0.14 |
| MISOGYNY-NON-SEXUAL-VIOLENCE | 0.50 | 0.05 | 0.09 |
| **Accuracy** | | | **0.52** |

## 6.2. Final Result

This section presents the evaluation methodology and metrics utilized for each task in the EXIST 2024 competition. The primary evaluation metric used across all tasks is the Information Contrast Measure (ICM) [32]. 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 based on the classification problems' nature and the hierarchical structure of the

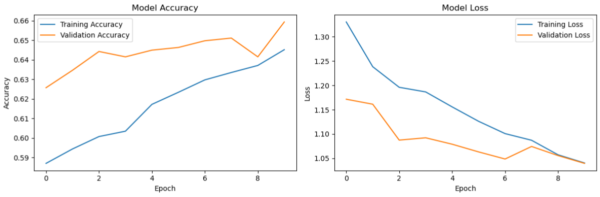


**Figure 12:** Task 06 Accuracy and Loss Curves

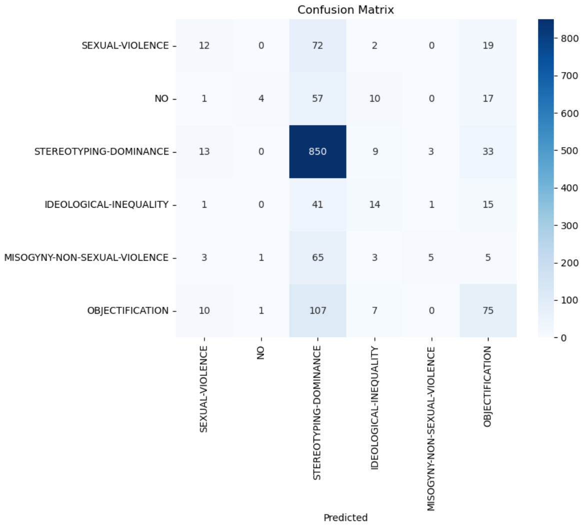


**Figure 13:** Task 06 Confusion Matrix

categories involved. It also presents a comprehensive analysis of the results of various runs and variants, highlighting the performance based on the ICM scores.

**Tasks 1 & 4: Sexism Identification** Tasks 1 & 4 require 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. The ICM serves as the official metric for Task 1.

**Tasks 2 & 5: Source Intention** Tasks 2 & 5 focus 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 ICM is the official metric for Task 2.

**Tasks 3 & 6: Sexism Categorization** Tasks 3 & 6 involve 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 multiple annotators. 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 two different modes for each task: "hard-hard" 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.

The "soft-soft" evaluation is conducted for systems that provide probabilities for each category. In this context, the system's probabilities are compared with those assigned by the human annotators. The ICM-soft metric accounts for the probabilistic nature of both system outputs and ground truth labels.

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.

**Table 9**
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 | 25 | -0.2086 | - | 0.4666 | - |
| Hard-Hard (All) | B | 0 | - | 0.9948 | - | 1 |
| | C | 54 | - | 0.1977 | - | 0.5994 |
| Soft-Soft (ES) | A | 0 | 3.1177 | - | 1 | - |
| | C | 31 | -0.3845 | - | 0.4383 | - |
| Hard-Hard (ES) | B | 0 | - | 0.9999 | - | 1 |
| | C | 55 | - | 0.0672 | - | 0.5336 |
| Soft-Soft (EN) | A | 0 | 3.1141 | - | 1 | - |
| | C | 23 | -0.0220 | - | 0.4965 | - |
| Hard-Hard (EN) | B | 0 | - | 0.9798 | - | 1 |
| | C | 50 | | 0.3320 | - | 0.6694 |
| EXIST2024_test_gold_soft = A; EXIST2024_test_gold_hard = B; CNLP-NITS-PP = C | | | | | | |

**Table 10**
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 | 15 | -2.4732 | - | 0.3007 | - |
| Hard-Hard (All) | B | 0 | - | 1.5378 | - | 1 |
| | C | 33 | - | -0.2694 | - | 0.4124 |
| Soft-Soft (ES) | A | 0 | 6.2431 | - | 1 | - |
| | C | 17 | -2.7097 | - | 0.2830 | - |
| Hard-Hard (ES) | B | 0 | - | 1.6007 | - | 1.6007 |
| | C | 33 | - | -0.3778 | - | 0.3820 |
| Soft-Soft (EN) | A | 0 | 6.1178 | - | 1 | - |
| | C | 9 | -2.2452 | - | 0.3165 | - |
| Hard-Hard (EN) | B | 0 | - | 1.4449 | - | 1 |
| | C | 31 | | -0.1572 | - | 0.4456 |
| EXIST2024_test_gold_soft = A; EXIST2024_test_gold_hard = B; CNLP-NITS-PP = C | | | | | | |

**Table 11**
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 | 17 | -5.7385 | - | 0.1970 | - |
| Hard-Hard (All) | B | 0 | - | 2.1533 | - | 1 |
| | C | 25 | - | -0.9571 | - | 0.2778 |
| Soft-Soft (ES) | A | 0 | 9.6071 | - | 1 | - |
| | C | 17 | -6.2485 | - | 0.1748 | - |
| Hard-Hard (ES) | B | 0 | - | 2.2393 | - | 1 |
| | C | 27 | - | -1.0686 | - | 0.2614 |
| Soft-Soft (EN) | A | 0 | 9.1255 | - | 1 | - |
| | C | 14 | -4.9948 | - | 0.2263 | - |
| Hard-Hard (EN) | B | 0 | - | 2.0402 | - | 1 |
| | C | 22 | | -0.8331 | - | 0.2958 |
| EXIST2024_test_gold_soft = A; EXIST2024_test_gold_hard = B; CNLP-NITS-PP = C | | | | | | |

## 7. Conclusion

The case of detecting and countering sexism in social networks has remained a vital and prevalent topic of future research through the assessments drawn by the research and results of the shared task of EXIST 2024. The task has now used more advanced models, among which are the CNN-BiLSTM model for text data and the ResNet50-CNN-BiLSTM model for meme data, to help understand and detect the presence of sexist content in images and text. Due to their ability to exploit the semantics and context of

**Table 12**
Task 04 Evaluation Result

| Variants | Run | Rank | ICM-Soft | ICM-Hard | ICM-Soft Norm | ICM-Hard Norm |
|---|---|---|---|---|---|---|
| Soft-Soft (All) | A | 0 | 3.1107 | - | 1 | - |
| | C | 27 | -1.2354 | - | 0.3014 | - |
| Hard-Hard (All) | B | 0 | - | 0.9832 | - | 1 |
| | C | 27 | - | -0.1234 | - | 0.4372 |
| Soft-Soft (ES) | A | 0 | 3.1360 | - | 1 | - |
| | C | 28 | -1.2557 | - | 0.2998 | - |
| Hard-Hard (ES) | B | 0 | - | 0.9815 | - | 1 |
| | C | 35 | - | -0.2781 | - | 0.3584 |
| Soft-Soft (EN) | A | 0 | 3.0794 | - | 1 | - |
| | C | 28 | -1.2140 | - | 0.3029 | - |
| Hard-Hard (EN) | B | 0 | - | 0.9848 | - | 1 |
| | C | 23 | | 0.0289 | - | 0.5147 |
| EXIST2024_test_gold_soft = A; EXIST2024_test_gold_hard = B; CNLP-NITS-PP = C | | | | | | |

**Table 13**
Task 05 Evaluation Result

| Variants | Run | Rank | ICM-Soft | ICM-Hard | ICM-Soft Norm | ICM-Hard Norm |
|---|---|---|---|---|---|---|
| Soft-Soft (All) | A | 0 | 4.7018 | - | 1 | - |
| | C | 5 | -1.5907 | - | 0.3308 | - |
| Hard-Hard (All) | B | 0 | - | 1.4383 | - | 1 |
| | C | 9 | - | -0.3370 | - | 0.3829 |
| Soft-Soft (ES) | A | 0 | 4.8140 | - | 1 | - |
| | C | 5 | -1.8008 | - | 0.3130 | - |
| Hard-Hard (ES) | B | 0 | - | 1.4356 | - | 1 |
| | C | 8 | - | -0.3809 | - | 0.3674 |
| Soft-Soft (EN) | A | 0 | 4.5834 | - | 1 | - |
| | C | 5 | -1.4400 | - | 0.3429 | - |
| Hard-Hard (EN) | B | 0 | - | 1.4409 | - | 1 |
| | C | 5 | | -0.2944 | - | 0.3978 |
| EXIST2024_test_gold_soft = A; EXIST2024_test_gold_hard = B; CNLP-NITS-PP = C | | | | | | |

**Table 14**
Task 06 Evaluation Result

| Variants | Run | Rank | ICM-Soft | ICM-Hard | ICM-Soft Norm | ICM-Hard Norm |
|---|---|---|---|---|---|---|
| Soft-Soft (All) | A | 0 | 9.4343 | - | 1 | - |
| | C | 8 | -6.6782 | - | 0.1461 | - |
| Hard-Hard (All) | B | 0 | - | 2.4100 | - | 1 |
| | C | 14 | - | -1.7920 | - | 0.1282 |
| Soft-Soft (ES) | A | 0 | 9.6290 | - | 1 | - |
| | C | 11 | -6.9019 | - | 0.1416 | - |
| Hard-Hard (ES) | B | 0 | - | 2.4432 | - | 1 |
| | C | 16 | - | -1.8559 | - | 0.1202 |
| Soft-Soft (EN) | A | 0 | 9.2546 | - | 1 | - |
| | C | 8 | -6.4165 | - | 0.1533 | - |
| Hard-Hard (EN) | B | 0 | - | 2.3532 | - | 1 |
| | C | 13 | | -1.6954 | - | 0.1398 |
| EXIST2024_test_gold_soft = A; EXIST2024_test_gold_hard = B; CNLP-NITS-PP = C | | | | | | |

texts and recognise visual prompts that characterize memes, these models have become very useful in categorizing Sexist posts. While attention-modulated models demonstrate state-of-the-art performance for a number of tasks, the variation in results across tasks and labels suggests that there is more work to be done in refining the methods.

The outcomes from the experiment of the EXIST 2024 shared task declare that models offer reliable performance in the detection of differentiating between sexist and non-sexist content; however, accuracy and reproducibility must be improved. Potential problems of this approach, such as significant distinctions between ‘DIRECT’ and ‘REPORTED’ categories or between several kinds of sexism that include ideological inequality, stereotyping, objectification, sexual violence, misogyny, and non-sexual violence, explain why this question is disputable.

There is also a paramount need to increase the models’ reliability with regard to identifying different types of sexism. These are some of the areas that would require expansion of datasets and diversification of the datasets that are fed to the system. Therefore, effective collaboration with the researcher, experts, and practitioners will be advisable to enhance the development of a better framework of automated systems to reduce sexism in social networks.

The EXIST 2024 shared task has offered a wealth of information and a clear starting point for future speculations, pushing machine learning forward in determining and eradicating the existing sexism on the Internet. These moves are expected to contribute towards the achievement of improving society, especially as it embraces the use of the digital platform.

## 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] 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.

[3] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, Exist 2024: sexism identification innbsp;social networks andnbsp;memes, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, p. 498–504. URL: https://doi.org/10.1007/978-3-031-56069-9_68. doi:10.1007/978-3-031-56069-9_68.

[4] A. Di Vaio, R. Hassan, R. Palladino, Blockchain technology and gender equality: A systematic literature review, International Journal of Information Management 68 (2023) 102517.

[5] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing., Neurocomputing (2023) 126232.

[6] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[7] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes (extended overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[8] A. Vetagiri, G. Kalita, E. Halder, C. Taparia, P. Pakray, R. Manna, Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces, arXiv preprint arXiv:2404.02013 (2024).

[9] N. R. Barman, K. Sharma, Y. Poddar, A. Vetagiri, P. Pakray, Addressing hate speech: Atlantis for efficient hate span detection (2023).

[10] A. Vetagiri, P. Pakray, A. Das, A deep dive into automated sexism detection using fine-tuned deep learning and large language models, Available at SSRN 4791798 (2024).

[11] G. Kalita, E. Halder, C. Taparia, A. Vetagiri, D. Pakray, Examining hate speech detection across multiple indo-aryan languages in tasks 1 & 4, Working Notes of FIRE (2023).

[12] A. Vetagiri, P. K. Adhikary, P. Pakray, A. Das, Leveraging gpt-2 for automated classification of online sexist content, Working Notes of CLEF (2023) 1107–1122.

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[14] 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.

[15] A. Vetagiri, P. Adhikary, P. Pakray, A. Das, CNLP-NITS at SemEval-2023 task 10: Online sexism prediction, PREDHATE!, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 815–822. URL: https://aclanthology.org/2023.semeval-1.113. doi:10.18653/v1/2023.semeval-1.113.

[16] A. Vetagiri, E. Halder, A. Das Majumder, P. Pakray, A. Das, "multilate": A synthetic dataset for multimodal hate speech detection, Available at SSRN 4733628 (2024).

[17] S. N. Vigod, P. A. Rochon, The impact of gender discrimination on a woman's mental health, EClinicalMedicine 20 (2020).

[18] M. E. Heilman, Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder, Journal of social issues 57 (2001) 657–674.

[19] A. Zimmerman, J. Dahlberg, The sexual objectification of women in advertising: A contemporary cultural perspective, Journal of advertising research 48 (2008) 71–79.

[20] R. Jina, L. S. Thomas, Health consequences of sexual violence against women, Best practice & research Clinical obstetrics & gynaecology 27 (2013) 15–26.

[21] J. Holland, A brief history of misogyny: The world's oldest prejudice, Hachette UK, 2012.

[22] L. Kelly, J. Radford, 'nothing really happened': the invalidation of women's experiences of sexual violence, Critical Social Policy 10 (1990) 39–53.

[23] 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.

[24] 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. URL: https://api.semanticscholar.org/CorpusID:156056029.

[25] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[26] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, Semantic Web 10 (2019) 925–945.

[27] A. Kumar, S. Kumar, K. Passi, A. Mahanti, A hybrid deep bilstm-cnn for hate speech detection in multi-social media, ACM Transactions on Asian and Low-Resource Language Information Processing (2024).

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[29] M. R. Ahmed, N. Bhadani, I. Chakraborty, Hateful meme prediction model using multimodal deep learning, in: 2021 International Conference on Computing, Communication and Green Engineering (CCGE), IEEE, 2021, pp. 1–5.

[30] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[31] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[32] 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.

[33] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.