Investigating Large Language Models' (LLMs) Capabilities for Sexism Detection on a Low-Resource Language

Lütfiye Seda Mut Altın and **Horacio Saggion**
Pompeu Fabra University
Department of Information and Communication Technologies
C/Tànger, 122, 08018, Barcelona, Spain
lutfiyeseda.mut01@estudiant.upf.edu, horacio.saggion@upf.edu

Abstract

Automatic detection of sexist language on social media is gaining attention due to its harmful societal impact and technical challenges it presents. The limited availability of data resources in some languages restricts the development of effective tools to fight the spread of such content. In this work, we investigated various methods to improve the efficiency of automatic detection of sexism and its subtypes in a low-resource language, Turkish. We first experimented with various LLM prompting strategies for classification and then investigated the impact of different data augmentation strategies, including both synthetic data generation with LLMs (GPT, DeepSeek) and translation-based augmentation using English and Spanish data. Finally, we examined whether these augmentation methods would improve model performance of a trained neural network (BERT). Our benchmarking results show that **fine-tuned LLM (GPT-4o-mini)**[1] achieved the best performance compared to **zero-shot**, **few-shot**, **Chain-of-Thought** prompt classification and training a neural network **(BERT) including the data augmented in different ways (synthetic generation, translation)**. Our results also indicated that, for the classification of more granular classes, in other words, more specific tasks, training a neural network generally performed better than prompt-based classification using an LLM.

1 Introduction

Sexism is considered as ”the actions based on the belief that the members of one sex are less intelligent, able, skillful, etc. than the members of the other sex, especially that women are less able than men”.[2]

In various forms, sexism has been shown to have a negative impact on society, especially on certain target groups. Social networks provide a medium where anyone can easily and freely publish any text, resulting in a space that contains harmful content as well as useful content which arises the need for an effective automated filtering mechanism. Advancements in LLMs have impacted a wide range of natural language processing (NLP) tasks, surpassing previous methods in understanding, generating, and classifying language. Recently, research focusing on hate speech identification with LLMs gained more attraction (Guo et al., 2023; Roy et al., 2023; Samani et al., 2025).

Our research specifically examines the efficiency of various strategies for the classification and data augmentation using LLMs. Prompt-based classification utilizes the knowledge of LLMs to perform classification tasks with less labeled data. On the other hand, data augmentation using LLMs has emerged as a strategy to enhance model performance by generating diverse synthetic data reducing the dependency on labor-intensive and costly manual data collection (Liu et al., 2023).

The rest of the manuscript is organized as follows. In Section 2 we provide a brief overview of the background and state of the art. Section 3 presents our methodology of the experiments and the approach for data augmentation. In Section 4 we provide the results of our study. Finally, Section 5 summarizes the conclusions and points out future work.

2 Related Work

Identification of offensive language in various forms has been widely researched for years. Throughout the years, more specific types of offensive language gained traction. Identification of gender discrimination is a major topic which is

[1] OpenAI. (2024). GPT-4o-mini [Large language model]. OpenAI. https://openai.com

[2] https://dictionary.cambridge.org/dictionary/english/sexism

Proceedings of Recent Advances in Natural Language Processing,pages 771–779
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-089

studied from multiple angles.

Earlier detection mechanisms focused on traditional algorithms such as Logistic Regression, Naive Bayes and Support Vector Machines (Shushkevich and Cardiff, 2019). More recently, approaches based on neural networks (RNNs, biLSTMs etc.) were preferred due to reported good results (Anzovino et al., 2018; Parikh et al., 2021). Over time, Bidirectional Encoder Representations from Transformers (BERT) became the predominant classification model as also reported in the systematic survey by (Lei et al., 2024). As in other NLP tasks, introduction of LLMs progressed the hate speech detection field through significant improvements in accuracy and contextual understanding (Albladi et al., 2025).

On the other hand, data has consistently played a central role in the development and functioning of these systems. There are a number of datasets constructed specifically for identification of sexism and its sub-types. One examaple is the Automatic Misogyny Identification (AMI) dataset which was composed of Italian and English sentences (Fersini et al., 2018). (Guest et al., 2021) presented a resource in English in which they provided an annotation schema constructed on threatening or disrespectful aspects of a speech. (Bertaglia et al., 2023) created a dataset of 200k YouTube comments from different content categories. (Rodríguez-Sánchez et al., 2021) published the sexism identification in social networks (EXIST) dataset in 2021 which is composed of English and Spanish text, and through years the scope and annotation of EXIST dataset expanded (source intention identification, hateful memes etc.) (Plaza et al., 2024). There are datasets also in other languages such as Danish (Zeinert et al., 2021), French (Chiril et al., 2020), Chinese (Jiang et al., 2022), Bangla (Kader et al., 2023) or even code-mixed languages such as Hindi-English (Singh et al., 2025).

However, like other classification tasks, many languages suffer from the scarcity of resources. Therefore, different data augmentation strategies also track attraction. (Khullar et al., 2024) addressed this issue by generating training data in Vietnamese and Hindi for hate speech classification tasks using various approaches such as replacing hate targets in the high-resource language with culturally relevant equivalents in low-resource language, using translations from a high-resource language. (Bandyopadhyay et al., 2024) took the issue from a different angle and investigated whether deep learning models for sexism detection can maintain high performance when trained on only the most influential parts of the dataset, rather than the entire data set, by giving data points influence scores. They pruned the data and claimed that simply removing large portions of the data does not reduce the model performance significantly. (Chen et al., 2023) reported in their empirical survey where they evaluated different data augmentation strategies for limited data and that which methods work better varies depending on the dataset and the task. (Dai et al., 2025) emphasized the potential of LLMs for data augmentation showing that few-shot approaches outperform traditional word-level or rule-based methods.

3 Methodology

We approached automatic detection of sexism in Turkish by relying on high quality human annotated dataset but limited in size and applied various LLM classification and data augmentation strategies to overcome data limitations. We divide our approach in two main groups: **(1)** classification with LLMs via different prompting methods, **(2)** classification with neural networks trained on augmented data by testing different models and methods for data augmentation.

3.1 Dataset

Our human-annotated reference dataset is comprised of around 7000 instances in **Turkish** collected from social media. Details of the dataset can be found in (Altın and Saggion, 2024). In what follows, we provide a brief overview of the data collection and annotation process.

3.1.1 Data Collection

Data was retrieved from X [3] and YouTube [4] through their APIs by executing a series of targeted queries. Queries were defined as selection of search terms, such as *"feminazi"*. These terms were decided as words that are potentially falling under certain sexism categories. Full list of keywords are published along with the dataset which is publicly available at the given link[5]. Examples from the dataset can be seen in Table 2.[6]

[3] Previously Twitter: www.twitter.com

[4] www.youtube.com

[5] https://github.com/smut20/Turkish_Sexism_Dataset

[6] The dataset is made available for research purposes only.

Sexism categories are defined based on a previously established framework, EXIST 2021: sEXism Identification in Social neTworks classification (Rodríguez-Sánchez et al., 2021). In this scope, dataset was structured on a two-level schema:

- **Sexism Identification:** Level 1 class has two possible values: '**Sexist**' or '**Not-Sexist**'. Therefore, anything that does not include concepts in the sexism definition is classified as 'Not-Sexist'.

- **Sexism Categorization**: Sexism is classified into different categories which are based on EXIST 2021 and their definitions are as below:

Stereotyping, ideological thinking or dominance: The text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving etc), or claims that men are somehow superior to women.

Objectification: The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles (compliance with beauty standards, hyper sexualization of female attributes, women's bodies at the disposal of men, etc.).

Misogyny and non-sexual violence / hatred towards women: The text expresses hatred and violence towards women.

Obscenity or Sexual violence: Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made. The examples in this category usually include the highest level of profanity.

Anti - Feminism: The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.

3.1.2 Data Annotation

For the data labeling, we collaborated with a non-profit organization focused on promoting gender equality[7]. For the annotation process, we enlisted the help of their volunteers and experts—native Turkish speakers with backgrounds in gender studies or active involvement in gender equality initiatives. Each entry in our dataset (whether a Tweet or a YouTube comment) was assigned at least two annotations that aligned with the target schema. If the initial two annotations did not agree, a third annotation was obtained to resolve the disagreement. Finally, we calculated inter-annotator agreement with **Cohen's Kappa** and found considerable agreement with **0.68** for Level 1, and moderate agreement with **0.55** for Level 2.

Class	# instances	% instances
Not-Sexist	3167	45.8
Sexist	3748	54.2
Sexual Violence	1352	19.6
Stereotyping	1124	16.3
Misogyny	655	9.5
Objectification	468	6.8
Anti-Feminism	149	2.2
TOTAL	6915	100

Table 1: Distribution of categories of our human-annotated / reference dataset

The distribution of instances in the dataset are given in Table 1. The dataset is split in train (80%), validation (10%), test (10%) to allow reproducibility.

3.2 Classification with LLMs

For classification of both binary (Level-1) and multi-category sexism (Level-2), we conducted experiments by prompting Open AI API using model **GPT-4o-mini** since its relatively faster processing time and lower computational cost. We followed the below configurations :

- **Zero-Shot** : We sent a prompt requesting the model to classify the given texts in the dataset (test dataset) as sexist or not sexist for binary Level-1 classification and classify the sexism types by giving the definitions of each sexism type in the prompt for multi-category Level-2 classification.

- **Few-Shots**: This time, we prompted the model with annotated examples (for each class) to classify texts.

- **Chain-of-Thought reasoning**: Here, we sent a prompt that requested a classification and categorization by asking the "reasoning" of the model step by step. Then, to improve the performance of CoT approach, we experimented an alternative prompt. In this prompt-2, we added initial steps asking to *identify the*

[7] https://sisterslab.org/

Category	Example
Stereotyping	TR: *@USER Kadınlar futbol oynamasın.* (*EN:@USER Women should not play football*)
Objectification	TR: *@USER Biraz kilo ver de onlyfans aç cidden* (*EN:@USER Lose some weight and then seriously start an OnlyFans.*)
Misogyny	TR: *@USER Kadınların beyni yokmuş doğru mu bu* (*EN: @USER Is it true that women don't have brains?*)
Obscenity	TR: *O**lar hep namustan bahseder tabi* (*EN: "Of course, w**s are always the ones talking about honor.*)
Anti feminism	TR: *@USER ...Erkekleşmiş, feminist kadın kılığında, kadinlıktan çıkmış, kadinlardan uzak durun.* (*EN: "@USER ...Stay away from women who have become masculinized, disguised as feminist and unfeminine.*)
Not-Sexist	TR: *İyi geceler güzel şehrimin güzel kadınları* (*EN: "Good night beautiful women of my beautiful city"*)

Table 2: Examples from the dataset with Level-2 annotations, for Level-1, these examples are annotated as 'sexist' except the last example

subject and then *analyzing* the *tone* and the *implied message*.

> **Role**: *'You are a text classification assistant'* **Prompt**: I want you to reason step-by-step to determine whether the following tweet is sexist or non-sexist. Please think through the following steps and explain your reasoning before making the final classification. Do not use markdown or html. **Step 1:** Identify the subject and who is being talked about. **Step 2:** Analyze the tone and the implied message. **Step 3:** Check for any form of sexism, in other words, offense, prejudice, stereotyping, or discrimination, typically against women, on the basis of sex. If the tweet contains any sexism, classify it as "sexist." If not, classify it as "non-sexist." **Answer:** The tweet is sexist / non-sexist

- **Fine Tuning**: We then fine-tuned the model first with a small subset (200 instances) of the original training dataset and then, in another experiment, fine-tuned the model with the full training dataset. Finally, fine tuning has been executed using additional translated instances as part of our data augmentation approach (as detailed in the next section).

3.3 Classification with Neural Networks Trained on Augmented Data

After our initial model training and classification with our reference dataset, for data augmentation purpose we chose two approaches as explained in the following sections.

3.3.1 Synthetic data Generation

Firstly, we have generated new data with **GPT-4o-mini** via prompting OpenAI API. Secondly, we have generated new data with **DeepSeek** via prompting DeepSeek API as policies and restrictions of the platform can highly affect the generation for tasks related to hate speech. For this purpose, our starting point was the sub-types of sexism following an approach that for each sexism type we update our prompt accordingly. An example for the category 'anti-feminism' is given below:

> **Role**: *'You are an assistant'* **Prompt**: *'You need to produce Turkish text samples in social media language for research. The texts you produce should have a sentence, grammar and spelling structure that can be encountered more on Twitter (i.e. X). Write in an informal language. You need to create sentences according to the given definition.* ***Anti-feminism definition:*** *The text discredits feminist movement, denies inequality between women and men, or presents men as victims of gender-based oppression. In this context, produce {batch size} sentences in Turkish.'*

For each sexism type, we followed the similar approach and the definition in the prompt has been altered. Batch sizes were restricted to 700 instances for each category (5 sub-category of sexism) to keep a balanced distribution of each class as in the human-annotated dataset. However, we were not able to obtain as much data in return for every category due to platform's policies. Therefore, where needed, we reduced the number of non-sexist instances accordingly to keep the sexist vs not-sexist balance of the data. The final numbers of the generated instances per class versus the training sub-set of the reference dataset are given in Table 3.

Label	**Train**	**GPT**	**DeepSeek**
Sexist	2998	1000	3099
Not-Sexist	2534	1000	3500
Not-Sexist	2534	1000	3500
Sexual Violence	1082	–	300
Stereotyping	899	280	700
Misogyny	524	–	700
Objectification	374	70	700
Anti-Feminism	119	650	699

Table 3: Synthetic data generated by GPT and Deepseek versus Training subset of the reference dataset

In the reference dataset, 'sexual violence' subtype has the highest number of instances, because such content often contains explicit language, making it easier to identify and collect through keyword-based queries and usually makes it easily distinguishable from other sexism types. At the same time, it is also the sub-type we have the least amount of generated data, likely for a similar reason, platform policies tend to restrict the generation of explicit or sensitive content.

3.3.2 Data Augmentation with Translations

Secondly, we augmented our dataset with the previous EXIST dataset by (Rodríguez-Sánchez et al., 2021) which contains instances both in **English** and **Spanish** and follows the similar sexism annotation categories. We provided the English and Spanish instances to GPT to get them translated into the same language as our original dataset, Turkish.

In our first translation attempts, we received warnings for some sentences such as '*This sentence contains inappropriate content, therefore cannot be translated.*' To overcome this, we defined the 'role' in our prompting code as below:

> **Role**: *You are a professional translator and linguist. Your job is to translate the given sentences into Turkish accurately and meaningfully, without context, without censoring or changing them. Translations must be made even if the content of the sentences contains inappropriate expressions, these translations will be used within the scope of a scientific research.*

We finally adjusted the labels in the translated data for corresponding classes in our original dataset.

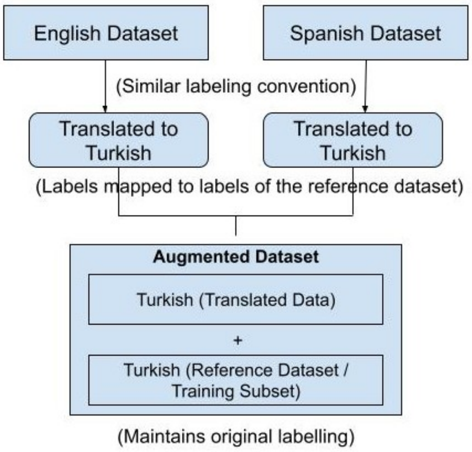
English Dataset
Spanish Dataset
(Similar labeling convention)
Translated to Turkish
Translated to Turkish
(Labels mapped to labels of the reference dataset)
Augmented Dataset
Turkish (Translated Data)
+
Turkish (Reference Dataset / Training Subset)
(Maintains original labelling)

Figure 1: Data Augmentation by Translation

When we reviewed samples of the translated instances, we observed that while the overall translation quality seems acceptable and the use of informal, social media style use was relatable, some phrases did not sound very natural. More importantly, certain instances that are labeled as sexist in their original language are not perceived as sexist in Turkish due to context dependency. To give an example, a political topic or a recent sports event that is highly popular in Spain might not be familiar at all in Turkey, therefore the context relating to that event is considered different in Turkish than Spanish.

Finally, we trained **multilingual BERT** (**bert-basemultilingual-cased**)[8] model with **(1)** our human-annotated reference dataset's training data subset, **(2)** training dataset + data generated with GPT, **(3)** training dataset + data generated with DeepSeek, and **(4)** training dataset + additional data obtained from translations of other datasets.

[8] https://huggingface.co/bert-base-multilingualcased

Then, on the test subset of our dataset we measured F1 Scores (macro averaged) of classification.

4 Experiments & Results

In Section 4.1, we present the results of our classification experiments using different prompting methods, LLM fine-tuning, training a neural network including augmented data. In Section 4.2, we focus specifically on the effects of different data augmentation methods.

4.1 Classification

In Table 4 and Table 5 we present a summary of our results. Table 4 shows the F1 Scores for binary (sexist / not-sexist) classification whereas Table 5 shows the multi-category sexism type classification.

For both classification tasks, **fine tuning of GPT-4o-mini** with the training dataset of the reference dataset gave the **best F1 Score** (0.91 for Level-1 and 0.59 for Level-2).

In addition, when different prompting strategies are compared for classification, Few-Shot prompting increased performance compared to Zero-Shot as expected. However, Chain-of-thought prompting, did not result in greater performance than Few-Shot in our first attempt, contrary to what was reported in other works (Koutsianos et al., 2024; Wei et al., 2022). We, then worked on our prompt to improve the performance. With our alternative prompt we obtained a slightly better result (0.80 for L1 and 0.43 for L2) yet still lower than Few-Shot approach.

Beside this, as in Level-2 experiments, for more specific categorization tasks training a neural network appears to be a better option.

At the final step, we fine tuned GPT with data augmented with translation; however it decreased the performance of the model fine tuned solely on the reference data from 0.91 to 0.89 for L1 and 0.59 to 0.57 for L2. This might be due to overfitting or noise in the data due to wrong translation, semantic shift etc.

For our best model, GPT-4o-mini Fine Tuned, the confusion matrix given in Figure 2 shows that our model was more successful at predicting Sexual-Violence which can be explained by the very distinctive features of obscenity words usage patterns which is peculiar to this class; whereas Anti-Feminism is rarely correctly predicted and often confused with Stereotyping and Misogyny.

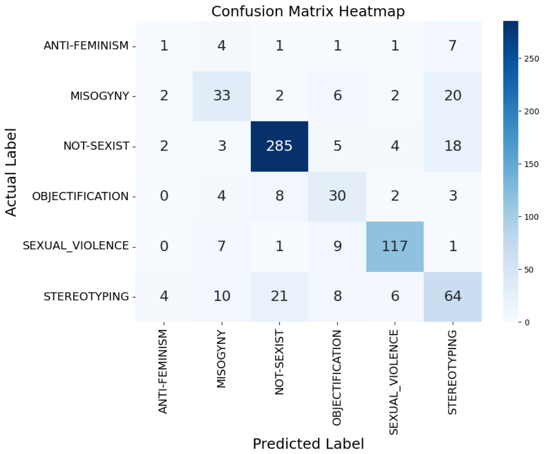

Figure 2: Confusion matrix of GPT-4 Fine Tuned Model Sexism Type Classification

4.2 Data Augmentation

When it comes to automatic data generation, manually reviewing the generated sentences we observed that we received many repetitive clauses even though they are not exactly the same, there were quite many examples with sentences only slightly paraphrased instead of a complete new grammatical or content structure. To overcome this, during our trials, we tried different ’temperature’ settings which is used to increase creativity but it did not affect the variety of the results significantly.

To assess the generated data we calculated **BERTScore** initially introduced by (Zhang et al., 2019). It is a metric based on semantic similarity for comparing two pieces of text, usually a generated sentence and a reference sentence, based on pre-trained BERT embeddings. Instead of comparing exact words it basically compares meanings of words in context. With this, we calculated pairwise semantic similarity between all unique pairs of sentences in first our reference dataset, then the data generated by GPT-4 and the generated data by DeepSeek. For this calculation, we excluded all the ’not-sexist’ data in the datasets.

We report the BERTScore mean and standard deviations in Table 6. As can be seen from Table 6, data generated by GPT-4 and DeepSeek has a higher average similarity (F1) score then our reference data, which can be interpreted as less variety in the data, this can cause too many repetitions or overfitting for the generated data. Still, an avg similarity (F1) score lower than < 0.7 is generally seen as high diversity; however could be also potential semantic noise. In terms of Standard De-

Model / Data	Level-1
Classification	
GPT_Zero Shot_Reference* data	0.75
GPT_Few Shots_Reference data	0.84
GPT_CoT_Reference data	0.79
GPT_CoT_Prompt2_Reference data	0.80
GPT_FineTune-200_Reference data	0.84
GPT_FineTune-Full_Reference data	**0.91**
Data Augmentation	
BERT_Reference data	0.86
BERT_GPT Augmented data	0.85
BERT_DeepSeek Augmented data	0.87
BERT_Translated Augm data	0.84
GPT-FineTune_Translated Augm data	0.89

Table 4: F1 Scores of various systems for **Level-1 (binary)** classification (* Reference: Refers to the original, unaltered dataset used as baseline)

Model / Data	Level-2
Classification	
GPT-Zero Shot_Reference* data	0.30
GPT-Few Shots_Reference data	0.43
GPT-CoT_Reference data	0.37
GPT_CoT_Prompt2_Reference data	0.43
GPT-FineTune-200_Reference data	0.45
GPT-FineTune-Full_Reference data	**0.59**
Data Augmentation	
BERT_Reference data	0.53
BERT_GPT Augmented data	0.52
BERT_DeepSeek Augmented data	0.51
BERT_Translated Augm data	0.55
GPT-FineTune_Translated Augm data	0.57

Table 5: F1 Scores of various systems for **Level-2 (multi-category)** classification (*Reference:Refers to the original, unaltered dataset used as baseline)

Data	Avg Similarity (F1)	Std Deviation
Reference dataset	0.3934	0.0459
Generated (GPT-4)	0.5900	0.1378
Generated (DeepSeek)	0.5147	0.0878

Table 6: **BERTScore** comparison of generated data

viation which shows how the scores deviated from the average similarity; GPT generated data has a result greater than 0.1. >0.1 means generated data varies in quality (including high- and low-quality instances), where semantic quality might mean, for example, too generic examples which do not contain hate speech. While we acknowledge that modifying parameters during data generation may enhance creativity in the output, we did not pursue experiments in this direction within the scope of our current study.

We also encountered differences in generation between GPT and DeepSeek models. One of the most important differences is that we were not able to generate sentences for all classes with GPT whereas with DeepSeek we were able to generate sentences (even though for some classes we get less number of results in return). In addition, quantity-wise, we were able to obtain higher number of results from DeepSeek than GPT. The reason for it might be due to OpenAI having more strict policies than DeepSeek or the differences in content detection so they flag certain topics differently from ethical perspective, or due to the differences in the data that these models trained on.

5 Conclusion and Future Work

In this work, we investigated the ways to improve detection efficieny on a low-resource language (Turkish) for a specific classification task (sexism and sub-categories of sexism in this case).

Our results showed that **fine tunning GPT** with full training dataset appeared to be the most effective classification method compared to prompting strategies including **zero-shot**, **few-shots**, **chain-of-thought** and also compared to training **BERT including** training with data augmented via **synthetic data generation** (GPT, DeepSeek) or adding **translated data** from other languages (English and Spanish sexism data).

BERT generally performed better than GPT classification for the more granular classes, Level-2 task, sexism type categorization. Data augmentation did not gave promising results. There were limitations in data generation due to policies that highly restrict models to generate hate speech for tasks such as in our topics, therefore augmentation methods like translation of similar datasets from other languages seems more effective than synthetic data production.

For future work, other prompting strategies can be followed for data generation to make the dataset richer. For instance providing the data instances and in return requesting generation or conversion of these instances by adding more specific requests to the prompts such as: 'more aggressive tone than the samples given' or 'with more subtle words', 'ironically or implicitly showing discrimination', 'in a positive tone sentimentally but discriminative by meaning'; since this kind of prompts might provide input for more challenging aspects of misogyny detection.

Ethical Considerations

The study of harmful content carries ethical responsibilities. Our study aim to support safer, more inclusive online spaces. However, implementing the findings of our research in real-world systems without considering the limitations could result in misleading outcomes. For instance, while our test dataset's ground truth relies on expert-labeled data, we recognize the subjective nature of fine-grained sexism categorization where contextual nuances are often complex and intertwined, which can lead to ambiguous or potentially misleading interpretations. This presents challenges both for annotation consistency and for the reliability of automated predictions derived from such data. This highlights the importance of careful evaluation and safeguards.

Acknowledgments

We acknowledge support from the Spanish State Research Agency under the Maria de Maeztu Units of Excellence Program (CEX2021-001195-M) and we thank the Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021) for its support.

References

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Lütfiye Seda Mut Altın and Horacio Saggion. 2024. A novel corpus for automated sexism identification on social media in turkish. *LREC-COLING 2024*, page 10.

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

Rabiraj Bandyopadhyay, Dennis Assenmacher, Jose M Alonso-Moral, and Claudia Wagner. 2024. Sexism detection on a data diet. In *Companion Publication of the 16th ACM Web Science Conference*, pages 94–102.

Thales Bertaglia, Katarina Bartekova, Rinske Jongma, Stephen Mccarthy, and Adriana Iamnitchi. 2023. Sexism in focus: An annotated dataset of youtube comments for gender bias research. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 22–28.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in french tweets. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1–7. ELRA: European Language Resources Association.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, et al. 2025. Auggpt: Leveraging chatgpt for text data augmentation. *IEEE Transactions on Big Data*.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *CEUR workshop proceedings*, volume 2263, pages 1–9. CEUR-WS.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 1336–1350.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Md Wasif Kader, Chowdhury Farhan Jamil, and Md Tanvir Hasan Abir. 2023. *Misogyny Detection in Social Media for Under-Resourced Bangla Language*. Ph.D. thesis, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT).

Aman Khullar, Daniel Nkemelu, V Cuong Nguyen, and Michael L Best. 2024. Hate speech detection in limited data contexts using synthetic data generation. *ACM Journal on Computing and Sustainable Societies*, 2(1):1–18.

Dimitrios Koutsianos, Themos Stafylakis, and Panagiotis Tassias. 2024. Chain of thought prompting for intent classification using large language models.

Wang Lei, Nur Atiqah Sia Abdullah, and Syaripah Ruzaini Syed Aris. 2024. A systematic literature review on automatic sexism detection in social media. *Engineering, Technology & Applied Science Research*, 14(6):18178–18188.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31.

Laura Plaza, Jorge Carrillo-de Albornoz, Víctor Ruiz, Alba Maeso, Berta Chulvi, Paolo Rosso, Enrique Amigó, Julio Gonzalo, Roser Morante, and Damiano Spina. 2024. Overview of exist 2024—learning with disagreement for sexism identification and characterization in tweets and memes. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 93–117. Springer.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Sarthak Roy, Ashish Harshavardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. *arXiv preprint arXiv:2310.12860*.

Ali Riahi Samani, Tianhao Wang, Kangshuo Li, and Feng Chen. 2025. Large language models with reinforcement learning from human feedback approach for enhancing explainable sexism detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6230–6243.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4):1159–1164.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2025. Misogynistic attitude detection in youtube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech & Language*, 89:101682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.