
**NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS**

**SCHOOL OF SCIENCES**

**DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**

**INTERDEPARTMENTAL PROGRAM OF POSTGRADUATE STUDIES IN LANGUAGE TECHNOLOGY**


**MASTER'S THESIS**

# A Greek Dataset for the Detection of Online Sexism Against Women


**Pinelopi A. Mantziou**

**Supervisor** **Dimitrios Galanis,** Researcher B' (ILSP)



**ATHENS**

**SEPTEMBER 2025**

**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗ ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Ένα ελληνικό σύνολο δεδομένων για την ανίχνευση διαδικτυακού σεξισμού κατά των γυναικών

**Πηνελόπη Α. Μάντζιου**

**Επιβλέπων** **Δημήτριος Γαλάνης,** Ερευνητής Β’ (ΙΕΛ)

**ΑΘΗΝΑ**

**ΣΕΠΤΕΜΒΡΙΟΣ 2025**

**MASTER'S THESIS**

# A Greek Dataset for the Detection of Online Sexism Against Women

**Pinelopi A. Mantziou**
**A.M.:** 7115182100019

**Supervisor** **Dimitrios Galanis,** Researcher B' (ILSP)

**EXAMINATION COMITTEE:**
**Dimitrios Galanis,** Researcher B' (ILSP)
**Georgios Paraskevopoulos,** Research Associate (ILSP)
**Sokratis Sofianopoulos,** Research Associate (ILSP)

October 2025

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Ένα ελληνικό σύνολο δεδομένων για την ανίχνευση διαδικτυακού σεξισμού κατά των γυναικών

**Πηνελόπη Α. Μάντζιου**
**Α.Μ.:** 7115182100019

**ΕΠΙΒΛΕΠΩΝ** **Δημήτριος Γαλάνης,** Ερευνητής Β' (ΙΕΛ)

**ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ**
**Δημήτριος Γαλάνης,** Ερευνητής Β' (ΙΕΛ)
**Γεώργιος Παρασκευόπουλος,** Επιστημονικός Συνεργάτης (ΙΕΛ)
**Σωκράτης Σοφιανόπουλος,** Επιστημονικός Συνεργάτης (ΙΕΛ)

Οκτώβριος 2025

# ΠΕΡΙΛΗΨΗ

Η παρούσα μελέτη παρουσιάζει, το πρώτο, από όσο γνωρίζουμε, μεγάλης κλίμακας (~14000 παραδείγματα), χειρωνακτικά επισημειωμένο σύνολο δεδομένων για την ανίχνευση σεξισμού στην ελληνική γλώσσα, με στόχο τις γυναίκες. Το σύνολο δεδομένων συλλέχθηκε από διάφορες πλατφόρμες κοινωνικής δικτύωσης, όπως το Twitter (X), το Reddit και το YouTube, και επισημειώθηκε τόσο σε δυαδικό επίπεδο (σεξιστικό, μη σεξιστικό) όσο και σε πιο λεπτομερές επίπεδο, καλύπτοντας οκτώ διαφορετικούς τύπους σεξισμού αλλά και τη διάκριση μεταξύ άμεσων και έμμεσων εκφάνσεων σεξισμού. Πραγματοποιήσαμε μια σειρά πειραμάτων χρησιμοποιώντας τα μοντέλα ανοιχτών βαρών Greek BERT, Meltemi και Kri-Kri, εφαρμόζοντας τόσο τεχνικές fine-tuning όσο και τεχνικές βασισμένες σε prompting. Prompting εφαρμόστηκε επίσης με τα μοντέλα κλειστών βαρών GPT-3.5 και GPT-4o μέσω των αντίστοιχων APIs. Όλα τα πειράματα διεξήχθησαν αποκλειστικά στο υποσύνολο των δεδομένων που προέρχονται από το Twitter. Τα αποτελέσματα δείχνουν ότι το GPT-4o, παρότι δεν χρησιμοποιεί fine-tuning, επιτυγχάνει υψηλή επίδοση και ξεπερνά όλα τα υπόλοιπα μοντέλα. Μεταξύ των μοντέλων που δοκιμάστηκαν, εκείνα που εκπαιδεύτηκαν (fine-tuned) απέδωσαν καλύτερα από τις αντίστοιχες prompt-based εκδοχές τους. Αξιοσημείωτο είναι ότι το Kri-Kri, το πιο πρόσφατο και μεγαλύτερο σε αριθμό παραμέτρων ελληνικό μοντέλο ξεπέρασε τα υπόλοιπα fine-tuned μοντέλα, επιδεικνύοντας ιδιαίτερη αποτελεσματικότητα στην ανίχνευση σεξισμού στα ελληνικά. Ωστόσο, όλα τα μοντέλα παρουσίασαν δυσκολία στην αναγνώριση έμμεσου σεξισμού, γεγονός που αναδεικνύει την εγγενή πρόκληση εντοπισμού λεπτών και έμμεσων μορφών σεξιστικής ρητορικής κατά των γυναικών.

**ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ**: Επεξεργασία Φυσικής Γλώσσας

**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ**: Ανίχνευση Σεξισμού, Επισημειωμένο ελληνικό Σύνολο Δεδομένων, Fine-tuning, Μάθηση με Prompts

# ABSTRACT

This study introduces, to the best of our knowledge, the first large-scale, (~14000 instances) manually annotated dataset for the detection of online sexism in the Greek language, against women. The dataset was collected from multiple social media platforms, including Twitter (X), Reddit, and YouTube, and annotated at both a binary level (sexist vs. non-sexist) and a fine-grained level, capturing eight distinct types of sexism as well as the distinction between direct and indirect expressions. We conducted a series of experiments using open-weight models such as Greek BERT, Meltemi, and Kri-Kri, applying both fine-tuning and prompt-based techniques. Prompt-based evaluation was also performed using the closed-weight GPT-3.5 and GPT-4o relying on the respective APIs. All experiments were conducted exclusively on the Twitter dataset. Results show that GPT-4o, even without fine-tuning, achieves strong performance and outperforms the other models overall. Among the models, those that were fine-tuned on the dataset performed better than their prompt-based counterparts. Notably, Krikri outperformed the other fine-tuned models, demonstrating strong effectiveness in the Greek sexism detection task. However, all models showed noticeable difficulty in detecting indirect sexism, highlighting the inherent challenge of identifying subtle and implicit expressions of sexism against women.



**SUBJECT AREA**: Natural Language Processing

**KEYWORDS**: Sexism Detection, Annotated Greek Dataset, Fine-tuning, Prompt-based Learning

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dimitris Galanis, for his invaluable guidance and support throughout the course of this thesis. His insights, encouragement, and constructive feedback were crucial at every stage of my research, and this work would not have been possible without his dedicated mentorship. I am also sincerely thankful to my fellow students, Manthoula and Andriana, for their voluntary and generous contribution in data annotation. Their effort and collaboration greatly enriched this project. Finally, I owe special thanks to my family and friends for their constant encouragement, understanding, and support during my studies.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# 1. INTRODUCTION

Online social media platforms have become central to public discourse, enabling instant communication and the exchange of ideas across diverse communities. However, this openness has also intensified the spread of harmful content, including hate speech, harassment, and gender-based discrimination. Sexism is generally defined as prejudiced attitudes, discriminatory language, or behaviors based on a person's gender, often reflecting and reinforcing stereotypes or power imbalances [1]. It can take many forms, ranging from direct sexism, which includes explicit insults or derogatory remarks, to indirect sexism, which is more subtle and often manifests through humor, stereotypes, or implicit bias. In this study, we specifically address sexism directed towards women, as this form of online abuse is not only prevalent but also deeply intertwined with cultural and societal norms. Automatic detection of such content is essential for promoting safer and more inclusive online environments.

Although significant progress has been made in Natural Language Processing (NLP) for hate speech detection, [2], [3] sexism detection remains a challenging task due to its subjective and context-dependent nature [4]. In addition, indirect sexism often relies on cultural nuances, sarcasm, or implicit bias, making it more difficult to annotate and classify correctly. A variety of initiatives, such as the EXIST (Sexism Identification in Social Networks) [5] shared tasks, have advanced research on sexism detection, offering multilingual datasets and performance benchmarks. A more detailed review of related datasets, annotation schemes, and model performance is provided in Section 2.

In addition, while English and other high-resource languages have benefited from the creation of specialized datasets and benchmarks, [6], [7], [8] to the best of our knowledge, there is currently no large-scale, publicly available annotated dataset for sexism detection in Greek. This limits the opportunities for development of models that can effectively recognize sexist content while accounting for linguistic and cultural nuances.

The goal of this study is to contribute to sexism detection in the Greek language by building a reliable annotated dataset and testing the effectiveness of different transformer-based models in identifying both direct and indirect forms of sexist content.

This study introduces a new dataset for sexism detection in Greek, collected from multiple social media platforms (Twitter, Reddit, and YouTube) to ensure a variety of linguistic registers and contexts. The dataset is manually annotated at both a binary level (sexist vs. non-sexist) and a fine-grained level covering various sexism subtypes such as objectification, stereotypes, and implicit bias. We evaluate several state-of-the-art models, including Greek BERT, Meltemi, Krikri, GPT-3.5 and GPT-4o, exploring both fine-tuning and few-shot prompting approaches. We also experimented with data augmentation techniques, including SMOTE-generated synthetic samples, to examine their effect on model performance. Our analysis also underscores the difficulty of detecting indirect sexism and the limitations of models when used without fine-tuning.

The remainder of this thesis is organized as follows. Section 2 reviews related work and provides background research on sexism detection datasets and models. Section 3 describes the dataset creation process, including data collection and annotation methodology. Section 4 outlines the experimental setup, covering both fine-tuning and prompt-based experiments, and presents the results along with an error analysis and a discussion of the challenges encountered. Finally, Section 5 concludes the study and suggests future research directions.

# 2. BACKGROUND

## 2.1 Introduction

The detection of sexism has been a subject of study in recent years in the field of Natural Language Processing and is examined either as a subcategory of toxic speech [2], [3] or as a separate research area. Classification approaches typically fall into two main categories: binary classification (sexist vs. not sexist) [5] and multiclass classification, where data are categorized into more fine-grained categories representing different types of sexism, such as slurs, stereotypes, objectification etc. [9]. Recent research has emphasized on the subjectivity of these labels and has taken into account annotator disagreement. For example, both the EXIST 2024 [10] and GermEval 2024 [4] shared tasks used the Learning with Disagreement (LwD) method, which recognizes that detecting sexism is subjective and different annotators may not always agree. So, instead of relying on one 'correct' label for evaluation (known as a hard label), LwD uses soft labels that reflect the probability distribution of the annotations of multiple annotators. More specifically for systems that provide probabilities for each class, the probabilities assigned by the system are compared with the probabilities assigned by a set of human annotators [10]. This approach allows models to learn from disagreement and better capture nuanced content.

While early research focused primarily on text [11], recent studies have expanded into multimodal datasets/tasks, including videos and images (e.g., memes) with the aim to identify sexist content that incorporates verbal and visual elements [12]. Most available sexism datasets are in English [6], but research has been conducted in several other languages too such as Spanish [7], German [8], Arabic [13], Chinese [14]; more details will be provided in the next paragraphs. Regarding Greek, while datasets have been published to detect toxic speech [15], [16] to the best of our knowledge there is currently no available dataset focusing exclusively on sexism.

The earliest approaches for sexism detection were based on classifiers such as Support Vector Machines (SVM), combined with features calculated using pre-trained word embeddings such as Word2Vec [17]. However, these approaches struggle to capture the nuances of natural language and the context dependent nature of sexism [18]. As research progressed, neural architectures, such as CNNs and RNNs along with variants like GRU, LSTM, and Bi-LSTM, were introduced, demonstrating strong performance in modeling sequential and contextual aspects of language [6]. LSTMs in particular, as reported in [18], proved effective in detecting a specific form of sexism; i.e. sexist statements used at the workplace. This effectiveness is because LSTMs are able to capture contextual information more effectively than non-deep learning models [19].

In the last few years, encoder-only Transformer-based models such as BERT [20], DistilBERT [21], RoBERTa [22], DeBERTa [23], have been proved particularly suitable for tasks such as text classification and sentiment analysis. They have been used extensively in sexism detection and have obtained better results than other machine learning methods in classifying and detecting sexism [24]. For example, in several comparative studies [14], [24], BERT-based models have shown superior performance to other approaches such as SVMs, CNNs, and BiLSTMs (see previous paragraph), achieving higher accuracy for two different languages, French [24] and Chinese [14]. Despite their success BERT-based models have shortcomings, particularly in categorizing subjective content from social media, which often includes implicit forms of sexism that make accurate classification difficult [25]. In the case of Greek, the release

of GreekBERT in 2020 [26] has enabled more effective language modeling for Greek, and it has also been used in studies on toxic speech detection, among others [27].

Encoder-based models such as BERT are primarily designed for text understanding and have been proven very competitive in various classification tasks. However, recent advancements in Natural Language Processing have led to the development of more versatile architectures known as Large Language Models (LLMs). These models incorporate decoder components, enabling them not only to interpret but also to generate fluent, human-like text [28]. This generative capability makes them suitable for applications such as dialogue systems, summarization, translation, and content generation [29]. LLMs are typically characterized by their massive scale, often comprising tens to hundreds of billions of parameters, and are pre-trained on vast and diverse text corpora. Compared to traditional pre-trained language models, LLMs not only exhibit significantly larger model sizes but also demonstrate better performance in both language understanding and generation, along with new abilities; e.g., few-shot in-context learning, as reported in [30]. LLMs have been explored for a variety of tasks, including toxic and sexism detection, thanks to their strong contextual understanding and generative capabilities. Below we will present tasks for which LLMs are used for sexism detection and we make the required comparisons with other types of models in various scenarios; e.g. fine-tuning, prompting etc.

### 2.2 GerMS-Detect task

The GerMS-Detect [4] shared task focused on detecting sexist and misogynistic language in German-language online news comments, with an emphasis on subjective annotation to capture the subtle and often unclear nature of sexist content. The task featured two subtasks: Subtask 1 on classifying with prediction of annotator disagreement, and Subtask 2 on predicting the full distribution of annotator labels. In terms of results, German-specific models performed best. More specifically, in Subtask 1, the goal was to predict a label for each comment. These labels were based on how multiple human annotators rated each comment for sexism. The task applied different strategies to combine the annotators' ratings, such as whether at least one, all, or the majority marked a comment as sexist, identifying the most frequently assigned label, or detecting disagreement among annotators. System performance was evaluated using the F1 macro score for each of the five label types: bin_maj, bin_one, bin_all, multi_maj, and disagree_bin. The final score was the average of those five scores. For Subtask 2, system performance was evaluated using the Jensen Shannon (JS) divergence, which measures the distance between the predicted and the gold label distributions. The final score was the average of the JS divergence computed for both the binary and the multiclass distributions. Both subtasks were organized into two tracks: a closed and an open track. In the closed track, participants were restricted to use only the provided training data and models that had not been pre-trained or fine-tuned on sexism-related tasks. In contrast, the open track allowed the use of external data sources, pre-trained language models, including large language models like GPT-3.5, and any proprietary tools or resources.

In Subtask 1, the top-performing systems were developed by the teams THAugs(F1 macro = 0.642) and ficode (F1 macro = 0.641). Both employed ensemble methods based on gbert-large, a German BERT-based model. Quabynar77 team, which used gbert-base in the closed track (F1 0.611), also competed in the open track using few-shot prompting via GPT-3.5 Turbo. GPT-3.5 is a decoder-only LLM (see previous paragraph) trained on vast amounts of text. However, this approach performed worse, achieving a lower F1 score of 0.45 compared to their closed-track result. In Subtask 2

as already mentioned, the goal was to predict the distribution of labels that annotators would assign to each comment. Each comment had multiple annotations on a 0–4 scale (representing the severity of sexism), and the task required predicting the percentage of annotators who would choose each label. Two types of distributions had to be predicted: a binary distribution, showing the proportion of annotators labeling the comment as sexist (labels 1–4) versus not sexist (label 0), and a multiclass distribution, showing the proportion of annotators selecting each severity level (0, 1, 2, 3, or 4). The evaluation metric for this task was Jensen–Shannon divergence, which measures how close the predicted distribution is to the actual one. Quabynar77 achieved the best results in Subtask 2 using a fine-tuned bert-base-german-cased model, while other top teams had a good performance by combining gbert-large-paraphrase embeddings with SVM classifiers. Across both subtasks, GPT-3.5-based few-shot learning consistently is inferior to encoder-only models. Their results suggest that fine-tuned, language specific BERT models are more effective than zero-shot or few-shot prompting with GPT-3.5 for the sexism detection task.

### 2.3 EXIST task

Similarly, in the multilingual EXIST 2024 [18] challenge that includes English and Spanish social media posts, several LLMs, including the encoder-only RoBERTa-Large, DeBERTa-V3-Large, and decoder-only Mistral-7B, were evaluated for sexism detection. On the combined development set (English + Spanish), Mistral-7B was fine-tuned and evaluated both as a standalone model and within an ensemble, but it performed worse than the encoder-only models. Specifically, it achieved an F1 score of 0.859, slightly lower than RoBERTa's 0.864 and DeBERTa's 0.866. The Dual-Transformer Fusion Network (DTFN), which combines the outputs of RoBERTa and DeBERTa, outperformed all three with an F1 score of 0.868. This indicates that although Mistral-7B has more parameters than the BERT-based models, it may be less effective at identifying sexist content in this task. In the test set results, the Multimodel Fusion Ensemble (MFE) (which uses a majority voting mechanism) achieved the 1st place out of 68 systems in the English track, with an F1 score of 0.7610, outperforming the Dual-Transformer Fusion Network (DTFN), which ranked 2nd (F1=0.7491). In the Spanish track, performance was lower overall: MFE ranked 12th (F1=0.7898), while DTFN ranked 25th (F1=0.7710). Finally, in the combined English + Spanish track, MFE obtained an F1 score of 0.7775 and ranked 4th out of 70, compared to DTFN's 13th place (F1=0.7614). These results highlight that MFE consistently outperforms DTFN, with particularly strong performance in English, while Spanish and cross-lingual scenarios remain more challenging.

Another study [1] that used EXIST 2024 data examined sexism detection in English and Spanish tweets, focusing on two key objectives: identifying whether a tweet is sexist (binary classification) and detecting the underlying source or intent behind sexist content, categorized as Direct, Reported, or Judgmental. The authors compare two approaches: fine-tuning the multilingual XLM-RoBERTa model and using few-shot prompting with GPT-3.5. The results reveal that XLM-RoBERTa consistently outperformed GPT-3.5 in both tasks, possibly because of the fine-tuning data. For Task 1 (sexism detection), XLM-RoBERTa achieved an F1 score of 0.78, while GPT-3.5 scored 0.71. In Task 2 (source/intention classification), XLM-RoBERTa again led with an F1 score of 0.48 compared to GPT-3.5's 0.43. Although GPT-3.5 demonstrated solid multilingual competence, it underperformed in terms of label consistency and overall predictive accuracy. These findings confirm that fine-tuned multilingual encoder-only models like XLM-RoBERTa remain superior to strong decoder-only non fine-tuned LLMs (like GPT3.5) for targeted classification tasks. While GPT-3.5 offers flexibility and

direct use in low resource scenarios, it requires more sophisticated example selection and prompt design to match the performance of dedicated, fine-tuned models.

## 2.4 EDOS task

The authors of SemEval-2023 Task 10, Explainable Detection of Online Sexism (EDOS) [9], created a new dataset of 20,000 English social media comments, sourced equally from Reddit and Gab, each carefully annotated using a three level taxonomy: (1) a binary label (sexist or not sexist), (2) one of four high level sexism categories, threats, derogation, animosity, and prejudiced discussion, and (3) one of eleven fine grained vectors (e.g., dehumanizing attacks, backhanded compliments). This labeling approach allows not just detection but also explanation of sexist content, enabling models to justify their predictions.

A very recent study [31] introduced a novel framework for explainable sexism detection based on Large Language Models (LLMs) enhanced by Reinforcement Learning from Human Feedback (RLHF). The research focuses on applying two recent open-source decoder-only LLMs, Mistral-7B and LLaMA-3-8B, to the Explainable Detection of Online Sexism (EDOS) dataset. The framework incorporates several techniques, including Supervised Fine-tuning (SFT) using QLoRA for efficient parameter adaptation, and RLHF using Direct Preference Optimization (DPO) to align model outputs with human preferences. Additionally, the approach utilizes prompt engineering to guide task-specific instructions. For Task A, Mistral-7B achieved an F1-macro score of 0.489 in the zero-shot setting, which improved to 0.822 with supervised finetuning (SFT) and further to 0.836 when SFT combined with RLHF. LLaMA-3-8B scored 0.508 in zero shot, 0.815 with SFT, and reached 0.860 with the addition of RLHF. The study highlights the strong impact of finetuning and reinforcement learning from human feedback (RLHF) on improving the ability of large language models (LLMs) to detect different forms of sexism, helping to build more explainable systems. The results also show that small sized LLMs (< 7B) perform poorly in zero-shot settings, meaning they struggle to obtain satisfactory results without task specific fine-tuning. Adding RLHF further improves these results by helping the models give clearer and more stable explanations when identifying sexist content.

Another paper [32] compared various fine-tuned and prompting based models on EDOS [9]. Encoder-only and decoder-only models were tested using fine-tuning, while encoder-decoder models were also evaluated using zero-shot and few-shot prompting. The results showed that, for detecting online sexism, the fine-tuning approach outperformed context based (few-shot) learning methods which confirms the conclusions of other studies; e.g. see previous paragraph. Specifically, among decoder-only models, Zephyr achieved the highest macro F1 score of 86.811, surpassing the encoder-only DeBERTa (83.913), while Mistral followed closely with a score of 86.041. In the prompting based experiments, the best performance came from the 5-shot configuration using Zephyr, which reached a macro F1 score of 70.936%. The fine-tuned model performed well in identifying both sexist and non-sexist content, showing no strong bias toward a particular label. In contrast, the context learning (few-shot) model struggled more with detecting sexist content and tended to overpredict non-sexist classifications. Further analysis revealed that the context learning approach often misclassified other forms of hate speech, such as racism, as sexism, leading to false positives. Additionally, both the fine-tuned and context learning models had difficulty recognizing figurative or subtle sexist language, which often does not explicitly mention gender.

Several other studies [2], [3], [31], [32], [33], [34] have explored the use of LLMs for automated sexism detection, applying fine-tuning, prompt engineering, and multilingual analysis across benchmark datasets. Despite variations in used approach, the findings consistently highlight the limitations of zero shot methods, the effectiveness of supervised fine-tuning, and the persistent challenges in detecting subtle or figurative sexist language.

# 3. DATASET

## 3.1 Data Collection

Previous research on toxic and sexist text has primarily relied on data from Twitter (currently X). Researchers also use data from other social media platforms to better reflect the diversity of online conversations [9], [35]. For this reason, we created three separate datasets from Twitter(X), Reddit, and YouTube.

Following the methodology described in [14], [15], tweets were collected covering the period from 2020 to 2024 based on a lexicon of terms/keywords primarily used to refer to women, with either offensive or neutral content. In particular, the lexicon includes slurs that specifically targeting women (e.g., *πουτάνα "whore*"), derogatory terms used to describe women (e.g*., μαντάμ "madam", γυναικάκι "little woman"*), as well as neutral terms that refer to women but can be used in either offensive or demeaning ways, either on their own or within phrases (e.g., *κορίτσι "girl", γυναίκα "woman", κουζίνα "kitchen", φεμινίστρια "feminist"*). In this way, we collected both positive and negative examples. We also included words in the lexicon that indicate important incidents related to sexism, such as *γυναικοκτονία "femicide"* and *ανδροκτονία "androcide".* Finally, we also added ten female names, 8 politicians who were recently in the news and 2 sexual abuse victims who shared their experiences publicly. The lexicon included 45 common nouns and 10 proper names. A total of 81,143 tweets were collected, of which 67,554 contained one of the 45 common nouns, and 13,589 included one of the 9 female names.

It is important to note that some slurs (e.g., *πουτάνα "whore*") may clearly refer to and target women in a sexist way. However, in many cases, such words appear within broader idiomatic expressions where they are not directed at a woman, but are rather used as generalized expletives (e.g., έγινε της *πουτάνας, "things went completely nuts"*). The presence of such terms does not necessarily indicate sexist content, which poses a challenge for classification models. For this reason, we intentionally included such cases in the dataset, in order to train and test the model with hard positive and negative examples and improve its ability to distinguish between them.

Reddit comments were collected using the Reddit API. Initially, a keyword-based search was performed using the same terms/lexicon as in the Twitter dataset. However, results were returned for only some of the keywords, likely due to Reddit not being widely used in the Greek language, as most users communicate in English. To augment our dataset, we downloaded content from Greek-language subreddits, 18 in total, resulting in a dataset of 13,570 comments.

For the YouTube data collection we didn't use the lexicon, automated scraping was performed on selected recent videos from podcasts featuring a female host or guest. Comments were collected from 12 videos, leading to the collection of 10,064 comments.

## 3.2 Data Cleaning and unbiasing

To ensure the quality and consistency of the data used in this study, a cleaning and anonymization process was applied. Specifically, we removed duplicates, empty entries, non-Greek text, comments containing only emojis or links, as well as those with fewer than three characters. Then, female proper names were replaced with the placeholder [WOMAN NAME] to prevent the model from becoming biased toward any specific name. The Twitter data were originally collected in .json format and later converted to

.csv files for easier preprocessing and annotation. In contrast, the Reddit and YouTube data were exported directly in csv format. After the data cleaning/unbiasing process, the final datasets were as follows: Twitter (X): 49,042 tweets, Reddit: 9,057 comments, YouTube: 8,897 comments. Next, the data were shuffled and a random subset was selected for manual annotation. Specifically, 10,090 tweets were randomly selected from the Twitter dataset, and 2,000 comments each from the Reddit and YouTube datasets.

### 3.3 Annotation Schema

Content Warning: The following section includes indicative examples of sexist language, used for the purpose of classification and analysis.

Before classifying comments as sexist or non-sexist, an initial first level distinction is made between toxic and non-toxic content (see Figure 2), to allow potential future research on hate speech. Toxic comments include offensive, aggressive, or demeaning language that creates a hostile environment, causes emotional harm, or undermines mental well-being [36]. This may involve insults, threats, or language that encourages hate or violence. At this stage, the focus is not on whether the content is sexist, but solely on its toxicity. However, all sexist comments are considered toxic, but not all toxic comments are sexist, since not all toxic comments contain sexist content.

The second level of annotation involves a binary classification of comments as sexist or non-sexist. Sexist content is defined as any form of abuse, explicit or implicit, targeted at women based on their gender or the intersection of gender with other identity traits (e.g., race, religion) [37]. Only comments targeting women (biologically or socially) are labeled as sexist. Offensive comments aimed at men, or at women without gender-based bias, are not labeled as sexist. If it's unclear whether a comment is sexist or directed at a woman, it is labeled as non-sexist by default. For instance, the phrase *"Γκόμενες και μαλακίες.. Καμία σας δεν αξίζει.. Μπέσα τώρα!"- Chicks and bullshit... None of you are worth anything... Seriously now!* is labeled as sexist and toxic, as it contains a gender-specific reference (*Γκόμενες*) and expresses generalized hostility toward women. On the other hand, comments like *"Σούργελο του κερατά"* may be labeled as toxic but not sexist, because although they include offensive or demeaning language, they do not clearly indicate a gender-specific target. If the comment does not clearly refer to a woman or use gendered language, it is not annotated as sexist.


**Figure 1: Final annotation scheme**

Next level of annotation (see Figure 1) introduces a fine-grained taxonomy of sexist comments, distinguishing *eight distinct subcategories* of sexist comments. While this classification is inspired by the hierarchical structure proposed in SemEval-2023 Task 10 [9], it has been carefully adapted to enhance annotation clarity and better capture subtle and culturally embedded forms of sexism. The SemEval-2023 Task 10 taxonomy for online sexism detection is structured into four main categories, each further divided into fine-grained vectors, resulting in a total of 11 distinct sexism types. The first category, (1) "Threats, plans to harm and incitement", includes vectors that involve either *direct threats of harm* or *incitement* and *encouragement of harm toward women*. The second category, (2) "Derogation," comprises three dimensions: *descriptive attacks*, *aggressive and emotive attacks*, and *dehumanizing or sexually objectifying comments*. The third category, (3) "Animosity," captures more implicit forms of sexism and includes the casual use of gendered *slurs and insults*, *immutable gender stereotypes*, *backhanded compliments*, and *condescending explanations or unwelcome advice*. The fourth category, (4) "Prejudiced Discussion," covers two dimensions: *support for mistreatment of individual women* and *endorsement of systemic discrimination against women as a group*. We didn't follow the structure with the four main categories but instead organized our scheme around the 11 fine-grained vectors, adjusting and restructuring them.

Specifically, the two subcategories *Threats of harm* and *Incitement and encouragement of harm*—were merged into our Category S0: “Incitement to violence (physical, sexual, or other invasive behavior)”. This decision was made because these types of content are semantically similar and appeared infrequently in our dataset. Similarly, the subcategories *Supporting mistreatment of individual women* and *Supporting systemic discrimination against women as a group* were merged into our broader Category S5: “Sexism through Denial, Language that Denies Discrimination, Justifies Gender Inequality, and Undermines Gender Issues”, as these types of expressions were also rare.

Several SemEval subcategories have clear counterparts in our taxonomy. *Dehumanising attacks and overt sexual objectification* correspond to our S2: “Objectification and sexualization”, while *Backhanded gendered compliments match* our S7: “Benevolent sexism: seemingly positive language or compliments”. Additionally, the “Aggressive and emotive attacks” subcategory corresponds to our S1: “Insults, offensive language, and slurs”, as both involve openly derogatory or hostile expressions. We also merged the subcategories *Descriptive attacks* and *Immutable gender differences and gender stereotypes* into a single category, S3 “Stereotypes”, because both involve generalizations about women’s roles, abilities, and traits.

The subcategory *Condescending explanations or unwelcome advice* maps partially to our S4: “Diminishing or patronizing comments”, which we expanded to include a broader range of subtle derogatory remarks, such as those using diminutives that belittles women without explicit profanity or aggression. This addition was made because a significant number of such comments appeared in our dataset.

Finally, we placed particular emphasis on sexist jokes, sarcastic remarks, and ironic expressions. While these often fall under category *Immutable gender differences and stereotypes*, which covers stereotypical assumptions and includes sexist jokes, it does not explicitly address irony or sarcasm. Given the frequency and subtlety of such forms in our data, we chose to foreground them more clearly in our annotation schema to better capture masked expressions of sexism.

Next, we present the eight subcategories that make up this fine-grained taxonomy of sexist comments, each capturing a distinct aspect of sexist discourse:

**0. Incitement to violence (physical, sexual, or other invasive behavior)**

Category S0 refers to incitement to violence, whether physical, sexual, or any other invasive behavior against women. It includes any comment or action that encourages, promotes, approves of, or suggests the use of violence or extreme behavior toward women. This type of language is particularly dangerous, as it can normalize violence and foster aggressive or abusive behavior. Specifically, it includes threats, insults, or any form of speech that justifies or legitimizes violence against women. Such statements or expressions, when directed at women, are identified as incitement to violence.

*Example: “Να σε βιάσουν να ηρεμήσεις!” (You need to be raped to calm down!)*

**1. Insults, offensive language, and slurs**

Category S1 refers to sexist comments that include direct insults, slurs, or abusive language. This type of language is used to demean or humiliate women through strongly derogatory or degrading expressions. Comments in this category represent open and explicit verbal attacks that reflect contempt toward women. Many of these statements also exhibit elements of misogyny.

*Example: "Ντύνονται σαν πόρνες πλέον οι γυναίκες" (Women dress like whores these days)*

**2.Objectification and sexualization**

Category S2 includes comments that objectify and sexualize women, disregarding their personality, abilities, or emotions. The key features of this category are a focus on women's bodies or appearance, references to women solely as objects of sexual gratification, and sexually suggestive remarks, often, though not always, expressed using vulgar language

*Example: "Μουναρά! Εμπαινα χωρίς δεύτερη σκέψη" (Total sexpot! I'd do her without a second thought)*

**3. Gender stereotypes**

Category S3 includes comments that reinforce, reproduce, or refer to gender stereotypes regarding women's roles, abilities, characteristics, or behavior. These comments are based on biases or traditional beliefs that limit women to specific roles or assign their value according to socially constructed expectations. The stereotypes in this category may be expressed explicitly or implied indirectly. Regardless of the speaker's intention, such comments contribute to the perpetuation of gender inequality. Stereotypes in this category may concern roles and abilities, behavior and traits, physical appearance, gendered expectations, and more [38].

*Example: "Τις δουλειές του σπιτιού τις κάνουν καλύτερα οι γυναίκες" (Women are better at doing housekeeping chores)*

**4. Diminishing or patronizing comments (e.g., use of diminutives)**

Category S4 includes comments that contain diminishing expressions or insults directed at women, often through the use of diminutives or subtle implications. Unlike S1, which involves explicit and direct insults, S4 reflects a more indirect form of disparagement that may appear "well-intentioned" or "advisory" on the surface. These types of comments reinforce gender inequalities and undermine women without necessarily using vulgar language. In addition to demeaning language and diminutives, this category may also include patronizing remarks that attempt to guide or advise women in a condescending tone. While S1 involves overt and harsh insults, S4 conveys sexism more subtly and without explicit obscenity.

*Example: "Το τουήτερ κοριτσάκι μου υπάρχει εδώ και δέκα χρόνια. Δεν ξεκίνησε τώρα." (Twitter, my little girl, has been around for ten years. It didn't just start now.)*

**5. Sexism through Denial, Language that Denies Discrimination, Justifies Gender Inequality, and Undermines Gender Issues**

Category S5 includes comments that dismiss or undermine the significance of gender-related issues and the discrimination experienced by women. These comments often attempt to justify gender inequality by questioning the existence or severity of systemic oppression, discrimination, or abuse. The language in this category may appear neutral on the surface, but in reality, it downplays gender-based problems, reinforcing sexism through denial.

*Example: "Και γιατί γυναικοκτονία και όχι ανδροκτονία ;" (Why femicide and not androcide?)*

**6. Sexism Hidden Behind Humor and Irony**

Category S6 includes comments that express sexist views in an indirect manner, often through irony, "humor," or sarcasm. In such comments, sexism may not be overt or

explicitly aggressive, but it becomes evident through context, implied attitudes, or a "humorous" tone. The language used in this category reinforces gender stereotypes and inequalities while often attempting to avoid being labeled as openly sexist.

*Example: "Γιατί οι γυναίκες πάντα έχουν τις καλύτερες ιδέες; Ποιος ξέρει, γιατί τις ακούει κανείς;" (Why do women always have the best ideas? Who knows, no one listens to them anyway.)*

**7. Benevolent sexism: seemingly positive language or compliments**

Category S7 includes comments that appear well-intentioned or positive but actually reinforce gender stereotypes and confine women to traditional roles or traits. Although these comments may be perceived as compliments, they are rooted in gender bias and represent a more subtle form of sexism. Benevolent sexism, while less overtly aggressive, contributes to the perpetuation of inequality by creating expectations about how women should behave or what their value is, based on stereotypical assumptions.

*Example: "Από τη γυναίκα πηγάζουν τα καλύτερα, τα ευγενέστερα γιατί μόνο αυτή ξέρει να αγαπά, να θυσιάζεται, να δίνει χωρίς να ζητά." (From woman springs the best and the noblest, for only she knows how to love, to sacrifice, and to give without expecting anything in return.)*

The above categories are organized in a scale from most severe, direct, and explicit (Category 0) to least severe and more implicit forms of sexism (Category 7). This gradation supports the decision rules used to assign the most appropriate label in ambiguous cases. Categories 0–2 correspond to direct sexism, while categories 3–7 reflect indirect or subtle sexism. This distinction between direct and indirect sexism is not made manually by the annotator. Instead, it is applied automatically through an if function in Excel, once the annotator has selected the specific category label (S0–S7). For instance, if an annotator labels a sexist comment as belonging to one of the categories 0–2 (Incitement to violence, Insults/slurs, Objectification) the system automatically assigns an additional label indicating direct sexism. Conversely, if a comment is labeled under categories 3–7 (Stereotypes, Diminishing comments, Sexism through denial, Humor and irony, Benevolent sexism), it is classified as indirect sexism.

The distinction between direct and indirect sexism is based on the intensity and clarity of the sexist language or behavior. Direct sexism includes overtly aggressive or offensive expressions that incite violence, abuse, or sexual degradation of women, such as threats, slurs and insults, or objectification and sexualization. These comments are explicit and clearly convey negative attitudes or behaviors toward women. On the other hand, indirect sexism consists of comments or actions that, while not openly hostile, perpetuate stereotypes, diminish women, or justify gender inequality in more subtle ways. These may include stereotypical statements, disrespectful or patronizing remarks, denial of gender-based discrimination, or irony and humor with underlying sexist messages. Benevolent sexism, although expressed through seemingly positive comments about women, often confines them to traditional gender roles and supports a sexist worldview.[39] Indirect sexism is more difficult to identify, but it remains equally harmful in reinforcing gender inequalities. A detailed set of annotation guidelines, including numerous examples, challenging cases, edge dilemmas with resolution strategies, and step-by-step instructions for annotating directly in Excel, was created to support the annotation process. The full version is provided in Appendix I and is written in Greek.

To facilitate annotation, the tweets were organized into a spreadsheet with six columns, as shown below. This format enabled annotators to apply both binary and fine-grained

labels, capture comments or uncertainties, and distinguish between direct and indirect forms of sexism. An excerpt of the annotation interface is presented in the figure.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Tweet | POLARITY | SEXISM | CATEGORY | COMMENTS | TYPE |
| 2 | Άντε γαμησου μωρή πουτανα | Toxic | Sexism | S1 | | Direct |
| 3 | Γκομενες και μαλακιες.. Καμια σας δεν αξίζει.. Μπέσα τώρα! | Toxic | Sexism | S1 | | Direct |
| 4 | Έχω σκοντάψει στο δρόμο για να σου απαντήσω στο μήνυμα μωρή ποι | Toxic | Sexism | S1 | | Direct |
| 5 | Χαχαχαχαχαχαχαχα. Ασχοληθείτε ρε λιγο με τις γυναικες σας και αφήσ | Non Toxic | Not Sexism | | | |
| 6 | Μυστικά ομορφιάς από μία μουνάρα | Toxic | Sexism | S2 | | Direct |
| 7 | Στα 47 λέω ότι, οι φοβερότερες και οι μόνες Γυναίκες που αξίζουν στην | Toxic | Sexism | S7 | | Indirect |
| 8 | γυναικες μπατσοι ηταν μλκ | Toxic | Not Sexism | | | |
| 9 | Η νδ καταργεί εργασιακά δικαιώματα και ο βόθρος του ΠΘ ανεβάζει αι | Toxic | Not Sexism | | | |
| 10 | το πιάτο σου δεν είχε ένταση-το είχα στο mute | Non Toxic | Not Sexism | | | |
| 11 | Καλημέρα στα κορίτσια τα όμορφα αν δείτε καμμία με λιλά μαλλί αγκα | Toxic | Sexism | S6 | | Indirect |
| 12 | Ντροπη Μωρη πατσαβουρα. | Toxic | Sexism | S1 | | Direct |
| 13 | Ώρα να γράψω κάτι σοβαρό φίλοι μου. ΜΗΝ ανεχεστε από κανέναν ΜΠ | Toxic | Not Sexism | | | |
| 14 | Οι καλύτεροι σύμμαχοι ενάντια στο κρυολόγημα, κρύβονται στην κουζί | Non Toxic | Not Sexism | | | |
| 15 | Μελέτη για την ανάπτυξη της Κοιλάδας των Τεμπών έγινε; | Non Toxic | Not Sexism | | | |
| 16 | Ρε παπάρα θα γίνεις άνθρωπος; Και πούτσα και φυλακή σε κάθε Γεωργ | Toxic | Not Sexism | | | |
| 17 | Στις 28.5.1952 Οι γυναίκες αποκτούν δικαίωμα εκλέγειν και εκλέγεσθα | Non Toxic | Not Sexism | | | |
| 18 | Το θέλω να φάω για σενα λιμιτ είναι καλή ατάκα για πέσιμο | Non Toxic | Not Sexism | | | |
| 19 | Η μαμά της Στέλλας αύριο το πρωί τηλεφωνική στην Κατερίνα Καινούρ | Non Toxic | Not Sexism | | | |
| 20 | Μαγειρεύω στη κουζίνα και με το που ειμαι έτοιμος να κάνω το πέταχτ | Non Toxic | Not Sexism | | | |
| 21 | Οι γριές οι πουτάνες έχουν το ζουμί | Toxic | Sexism | S1 | | Direct |

**Figure 2: Interface Used for Manual Annotation**

### 3.4 Annotation Agreement

The annotation guidelines were given to three female annotators for review along with a set of 50 tweets selected for a pilot annotation round. Each annotator labeled the tweets individually, without any communication with the others. Once the task was completed, the results were discussed collectively and the final set of guidelines was formed.

Subsequently, 1,000 tweets were selected from the Twitter dataset and assigned to the three annotators for annotation. During this phase, annotators were not allowed to communicate with each other but could flag comments they wished to discuss after the annotation process.

The annotation results were compiled into two separate files: one file containing the results of the binary classification (sexist / non-sexist) and another file containing the results of the multiclass classification (8 sexism subcategories).

Cohen's Kappa was calculated in both cases to evaluate inter-annotator agreement. The measurement was conducted separately for the binary and the multiclass classifications, offering valuable insights into the reliability of the annotations.

Cohen's Kappa is a statistical measure used to evaluate the level of agreement between two annotators when categorizing data into predefined categories. Its advantage is that it accounts for agreement that may occur by chance, thereby correcting for random coincidence. It assesses how much better the annotators agree than what would be expected by chance alone. The value of Cohen's Kappa ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating agreement equivalent to what would be expected by random chance. The formula for Cohen's Kappa is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:

- Po: The observed agreement between annotators.
- Pe: The expected agreement by chance.

This measure corrects for chance agreement and is widely used to assess inter-annotator reliability in categorical data. Kappa values below 0 indicate no agreement, values between 0.41 and 0.60 reflect moderate agreement, 0.61 to 0.80 indicate good agreement (aka substantial agreement), and values above 0.80 represent almost perfect agreement.

For the binary classification task (Sexist/Non-Sexist), Cohen's Kappa indicated a high level of agreement among annotators, suggesting strong consistency in the categorization of tweets into these two classes. Any disagreements were reviewed after the annotation process, and clarifications were added to the guidelines to improve the interpretation of cases that led to disagreement.

**Table 1: Cohen's Kappa Scores Between Annotators for Binary Classification**

| | |
|---|---|
| Annotator 1 & 2 | 0.912 |
| Annotator 1 & 3 | 0.940 |
| Annotator 2 & 3 | 0.863 |

For the multiclass classification task (S0 to S7), Cohen's Kappa scores were lower compared to the binary classification, as expected due to the larger number of categories and the complexity of the classification process. More disagreements were observed in categories that were more subjective or had conceptual ambiguity; these categories as expected belong to the broader class of indirect sexism which is more implicit. In particular, 'S3: Stereotypes' and 'S6: Irony/Humor' had the highest disagreement rates. Despite this, the dataset was considered sufficiently reliable, as the overall Kappa scores remained high, i.e. above 0.8 indicating strong inter-annotator agreement.

**Table 2: Cohen's Kappa Scores Between Annotators for Multiclass Classification**

| | |
|---|---|
| Annotator 1 & 2 | 0.861 |
| Annotator 1 & 3 | 0.898 |
| Annotator 2 & 3 | 0.828 |

To determine the final label for both the binary and multiclass classification tasks, a structured approach was followed to resolve disagreements. In cases of unanimous agreement among annotators, the shared label was directly accepted as final. When at least two annotators agreed on a label, that majority label was adopted. However, in instances of complete disagreement, where all three annotators selected different labels, an expert was consulted to assign the final label. Additionally, revisions were made to the annotation guidelines to reduce ambiguity and improve annotator consistency.

The 1,000 tweets that were annotated with the highest possible consistency were used as a reference for guiding the annotation of the remaining datasets, which were labeled by a single annotator.

### 3.5 Results of Annotation

The Twitter dataset contained 49,042 tweets, of which 10,090 were manually annotated. After removing instances labeled as “other” (not belonging to the predefined set of categories) or identified as errors, a final set of 10,000 tweets was kept; see Table 3 below. The Reddit dataset included 9,057 comments, with 2,000 manually annotated; following the removal of “other” labels and error entries, 1,973 comments remained. The YouTube dataset comprised 8,897 comments, 2,000 of which were manually annotated, resulting in a final set of 1,977 comments after similar filtering.

**Table 3: Statistics on the gathered data**

| Category | **Twitter** | **Reddit** | **YouTube** |
|---|---|---|---|
| Total Comments | 49,042 | 9,057 | 8,897 |
| Manually Annotated | 10,090 | 2,000 | 2,000 |
| Valid Annotated-Final Dataset | 10,000 | 1,973 | 1,977 |

The comparative results for Twitter, Reddit, and YouTube, highlight meaningful similarities and differences in the frequency and expression of sexist content across the three social media platforms. First all datasets were imbalanced in terms of sexist vs. non-sexist content with non-sexist being the prevalent category(Table 5). This reflects real-world conditions and aligns with the distribution found in other datasets, for example, for EDOS it is reported a ~24% of sexist comments [4]. Twitter, which is the largest dataset in this study, exhibited the highest percentage (see Table 4) of toxic (27.72%) and sexist comments (21.87%), while Reddit showed the lowest proportions in both categories; 14.05 % toxic,13.63 % sexist (See Table 3).

**Table 4: Distribution of Toxic/Nontoxic Across Platforms**

| | **Twitter** | **Reddit** | **YouTube** |
|---|---|---|---|
| Nontoxic | 7,203 (71.38%) | 1,683 (84.15%) | 1,546 (77.3%) |
| Toxic | 2,797 (27.72%) | 290(14.05%) | 431(21,55%) |

**Table 5: Distribution of Sexist/Non-Sexist Across Platforms**

| | **Twitter** | **Reddit** | **YouTube** |
|---|---|---|---|
| Non-Sexist | 7,813 (78.13%) | 1,704 (86.36%) | 1,663 (84.11%) |
| Sexist | 2187 (21.87%) | 269 (13.63%) | 314 (15.88%) |

Across all three platforms (Table 6), both benevolent sexism and incitement to violence were the least frequently observed categories. On Twitter, the majority of sexist comments fell into categories S1 (insults and slurs), S2 (objectification/sexualization), S3 (stereotypes), and S6 (humor and irony). YouTube showed a similar pattern, with the highest frequencies in categories S1, S2, S3, and S4 (diminishing or patronizing comments), indicating a slightly different distribution of sexist expression. Reddit presents a different distribution pattern, with around 50% of sexist comments falling under gender stereotypes (S3). Additionally, a significant portion of comments are categorized as sexism through denial (S5), suggesting that more implicit and indirect forms of sexist discourse are prevalent on this platform.

**Table 6: Distribution per category Across Platforms**

| | **Twitter** | **Reddit** | **YouTube** |
|---|---|---|---|
| S_0: Incitement to violence | 29 (1.32%) | 2 (0.74%) | 1 (0.31%) |
| S_1: Insults and slurs | 494 (22.58%) | 27 (10.03%) | 97 (30.89%) |
| S_2: Objectification sexualization | 450 (20.57%) | 26 (9.66%) | 56 (17.83%) |
| S_3: Stereotypes | 480 (21.94%) | 135 (50.18%) | 63 (20.06%) |
| S_4: Diminishing comments | 281 (12.84%) | 15 (5.57%) | 66 (21.09%) |
| S_5: Sexism through denial | 104 (4.75%) | 52 (19.33%) | 19 (6.05%) |
| S_6: Humor and irony | 311 (14.22%) | 9 (3.34%) | 11 (3.50%) |
| S_7: Benevolent sexism | 38 (1.73%) | 3 (1.11%) | 1 (0.31%) |

This pattern is also reflected in the distribution of direct versus indirect sexism (Table 7). While indirect sexism is more frequent across all three platforms, the difference is relatively small on Twitter and YouTube. In contrast, on Reddit, approximately 80% of the sexist comments are categorized as indirect sexism. This is likely due to the high presence of stereotype-related content (S3). YouTube, although its overall percentage of sexist comments (15.88%) was lower than Twitter, it had the highest proportion of direct sexism (49.04%) among the three. Insults and slurs (30.89%) and objectification (17.83%) were particularly prevalent.

**Table 7: Distribution of Direct/Indirect Comments Across Platforms**

| | **Twitter** | **Reddit** | **YouTube** |
|---|---|---|---|
| Direct Sexism | 973 (44.49%) | 55 (20.44%) | 154 (49.04%) |
| Indirect Sexism | 1214 (55.50%) | 214 (79.55%) | 160 (50.95%) |

Overall, the results show that the type and intensity of sexist language differ across platforms. Twitter includes a wide range of sexist comments, Reddit has more subtle and indirect sexism, while YouTube shows more direct objectification and offensive language. These differences highlight the need for platform-specific methods in detecting sexism, making sure to capture both clear insults and more hidden or disguised forms of bias. In the table below, all the actual counts and percentages are presented for each dataset and for each category individually.

### 3.6 Contribution of the Dataset

To the best of our knowledge, this dataset represents the first large-scale, manually annotated dataset for sexism in Greek, collected from three different platforms, Twitter, Reddit, and YouTube, to ensure linguistic and contextual diversity. A multi-level annotation schema was adopted, combining binary and fine-grained multiclass labels, including innovative categories such as benevolent sexism and sexism through denial. The annotation was based on detailed guidelines and evaluated through inter-annotator agreement. Three distinct annotated datasets were created: a large Twitter dataset containing 10,000 annotated tweets, and two smaller datasets from Reddit and YouTube, each with 2,000 annotated comments. The resulting dataset aims to support future research in Greek NLP and the automatic detection and analysis of sexist language, capturing a wide range of expressions and nuances of sexist online discourse.

# 4. EXPERIMENTS

## 4.1 Experimental Setup

All experiments were conducted using Google Colab Pro with access to GPU acceleration, providing sufficient computational resources for both training and inference. The experiments were conducted exclusively on the Twitter dataset, focusing only on the binary classification task; sexist vs. non-sexist. The final dataset consisted of 10,000 annotated tweets, split into 80% for training, 10% for validation, and 10% for testing, specifically, 8,000 training examples (6,250 non-sexist and 1,750 sexist), 1,000 for validation (781 non-sexist and 219 sexist), and 1,000 for testing (781 non-sexist and 219 sexist). The split was performed using stratified sampling to preserve the original label distribution across all subsets.

## 4.2 Models used in our experiments

### 4.2.1 Greek BERT

One of the models used in this study is Greek BERT [26], an open-source model (nlpaueb/bert-base-greek-uncased-v1), developed by the Athens University of Economics and Business (AUEB) NLP group. It follows the BERT-Base uncased architecture, which is an encoder-only transformer model consisting of 12 transformer encoder layers, 12 attention heads, and approximately 110 million parameters. The model was pre-trained on a 29 GB Greek corpus that includes (a) the Greek part of Wikipedia, (b) The Greek part of European Parliament Proceedings Parallel Corpus, and (c) the Greek section of OSCAR, a cleaned version of Common Crawl. The vocabulary of the model consists of 35,000 subword units and it was created using the SentencePiece tokenizer with byte pair encoding (BPE). The model is publicly available through the Hugging Face Model Hub[1].

### 4.2.2 Meltemi

Another model evaluated in this study is Meltemi-7B-Instruct, (ilsp/Meltemi-7B-Instruct-v1.5) which is developed by the Institute for Language and Speech Processing (ILSP) part of the Athena Research & Innovation Center.[2] It is based on the Mistral-7B architecture and features a context window of 8192 tokens. The model has 7 billion parameters and follows a decoder-only architecture, making it suitable for generation tasks such as instruction-following and conversational responses. Its foundational version, Meltemi-7B-v1, was pre-trained on approximately 40 billion tokens, including 28.5 billion in Greek, 10.5 billion in English, and 0.6 billion from Greek-English parallel corpora. Instruction tuning for the Instruct version was performed using around 100,000 Greek instruction-response pairs, compiled from translated open datasets (Open-Platypus, Evol-Instruct, Capybara) and manually curated multi-turn examples. The model's tokenizer is an extended version of the original Mistral tokenizer, enriched

[1] https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1

[2] https://medium.com/institute-for-language-and-speech-processing/meltemi-a-large-language-model-for-greek-9f5ef1d4a10f

with Greek vocabulary. Meltemi-7B-Instruct is open source, licensed under Apache 2.0, and available on the Hugging Face Model Hub[3].

### 4.2.3 Krikri

Llama-Krikri-8B-Instruct (ilsp/Llama-Krikri-8B-Instruct), also used in this study, is a decoder-only transformer model developed again by ILSP/Athena RIC, built upon Meta's LLaMA 3.1-8B architecture. It features approximately 8 billion parameters, supports a 128,000-token context window, and is open-source via the Hugging Face Model Hub[4]. The model was pre-trained on an extensive 91 billion-token multilingual corpus, including 56.7 billion Greek tokens, 21 billion English tokens, 5.5 billion parallel text pairs, and 7.8 billion code/math tokens. Subsequently, it underwent multi-stage instruction-tuning and alignment using over 1.4 million instruction-response pairs and preference data to support robust conversation and reasoning capabilities The tokenizer includes an extended Greek vocabulary (as Meltemi), ensuring efficient tokenization and speed without breaking Greek words into characters. Evaluation across several benchmarks indicates that Krikri outperforms larger commercial LLMs in Greek understanding, generation, and code tasks [40].

### 4.2.4 GPT-3.5-turbo

The evaluation also included GPT-3.5-turbo, a closed-source, decoder-only language model developed by OpenAI, accessible via the OpenAI API. While specific architectural details and parameter counts have not been officially released, estimates suggest the model contains between 20 and 40 billion parameters. It is optimized for instruction following, prompt-based inference, and multi-turn dialogue. As a proprietary model, it does not allow local deployment or training, and its training data and architecture remain undisclosed. Although GPT-3.5 demonstrates improvements in instruction following and generation quality, its performance does not consistently surpass that of the GPT-3 series across all tasks, particularly in areas requiring complex reasoning; e.g., Machine Reading Comprehension [41].

### 4.2.5 GPT-4o

Finally, this study employed GPT-4o, the latest publicly available multimodal, decoder only, large language model of OpenAI which was released in May 2024. As a closed-source model, it is accessible only via a web API, and no free fine-tuning or local deployment is currently supported. OpenAI has not disclosed the number of parameters or full training details. The suffix "o" in GPT-4o stands for "omni," indicating its ability to natively process text, vision, and audio inputs. GPT-4o is reported to be significantly faster, and more affordable than GPT-4, while offering comparable or improved performance across many benchmarks[5].

---

[3] https://huggingface.co/ilsp/Meltemi-7B-Instruct-v1.5

[4] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct

[5] https://openai.com/index/hello-gpt-4o/?utm_source=chatgpt.com

## 4.3 Approaches

### 4.3.1 Fine-tuning

The first approach we used was fine-tuning. Fine-tuning refers to the process of adapting a pre-trained language model to a specific downstream task by continuing training on labeled, task-specific data. In this study, we fine-tuned both encoder-only and decoder-based models on a binary classification task aimed at detecting sexist versus non-sexist content in Greek tweets. This approach was applied to three freely available models described above: Greek BERT, Meltemi-7B-Instruct, and Llama-Krikri-8B-Instruct. All of them were fine-tuned on the same annotated Twitter dataset to ensure fair comparison. As already mentioned, we used stratified data splitting to preserve label distribution across training, validation, and test sets. To ensure reproducibility, we used a fixed random state value of 42 for dataset splitting and other randomized operations (e.g., shuffling), so that data partitions remained identical across runs.

#### 4.3.1.1 Greek BERT fine tuning

We experimented with four variations of fine-tuning the Greek BERT model to evaluate its effectiveness in detecting sexist content:

- **BERT fine-tuning**: It was fine-tuned on the Twitter dataset using the standard cross-entropy loss function, without incorporating any weighting to compensate for class imbalance. The BERT encoder is followed by a dense classification layer with a softmax activation. The model was trained using AdamW with a batch size of 8, a learning rate of 1e-5, and the weight decay set to 0.01. Training ran for a maximum of 10 epochs, with early stopping enabled to reduce the risk of overfitting. Tokenization was performed using the Greek BERT tokenizer with a maximum sequence length of 256 tokens. This setup served as a baseline model for comparison.
- **BERT fine-tuning with weighted loss**: To address the imbalance between sexist and non-sexist tweets, we applied class weights to the loss function. Class weights were computed based on the inverse frequency of each class in the training set, assigning higher importance to the underrepresented class (sexist). The model architecture remained the same, consisting of the BERT encoder followed by a dense classification layer with a softmax activation. Training was conducted with a batch size of 8, learning rate of 1e-5, and the AdamW optimizer with weight decay set to 0.01. Again the model was trained for up to 10 epochs, with early stopping enabled to avoid overfitting. The tokenizer from the Greek BERT model was used, with a maximum sequence length of 256 tokens.
- **BERT fine-tuning with additional synthetic data**. To manage class imbalance at the data level, we also applied SMOTE [42] (Synthetic Minority Over-sampling Technique), a classic oversampling algorithm that generates synthetic instances of the minority class. Unlike naive oversampling, which simply duplicates existing examples and may lead to overfitting, SMOTE creates new samples by interpolating between existing minority-class examples in the feature space. Specifically, for each minority-class instance, the algorithm selects one or more of its *k*-nearest neighbors in the feature space and generates a new synthetic example by interpolating between the original instance and a neighbor — that is,

by creating a new point somewhere between them in the feature space. To ensure the relevance and appropriateness of the synthetic data, we applied an automatic filtering step using the same lexicon that had been employed during the initial data collection process. Each synthetic tweet was retained only if it contained at least one keyword from this lexicon. After this automated filtering and subsequent manual inspection for quality control, a total of 2,936 synthetic sexist tweets were retained for use. We explored the effect of different levels of augmentation on model performance by creating three experimental setups: (a) in the first setup, we added one-third of the SMOTE-generated examples to the training set, (b) on the second, we added two-thirds of the synthetic data and (3) in the third, we included all 2,936 synthetic examples. Each version was evaluated to observe whether increasing synthetic data would improve the model's ability to detect sexist content or lead to overfitting or noise-induced degradation. In all cases, the base model architecture and training configuration remained the same; i.e., standard cross-entropy loss function was used without class weighting.

- **BERT partial fine-tuning (frozen BERT layers)**: We experimented with freezing all or part of the pretrained BERT layers, allowing only the classification head to be trained, class weighting was not applied [43]. This approach reduces computational cost and tests whether the pretrained representations are sufficient without full fine-tuning. In this variation, we explored the effect of freezing the pretrained BERT encoder layers, allowing only the final classification head to be updated during training. Freezing was implemented by disabling gradient updates for all BERT parameters, keeping only the classification layer on top trainable. The rest of the training setup remained consistent with previous experiments; i.e., no class weighting. This experiment aimed to evaluate whether lightweight fine-tuning could still achieve competitive performance, especially in resource constrained settings where full model training is impractical.

#### 4.3.1.2 Meltemi fine-tuning

To fine-tune Meltemi-7B-Instruct, the Twitter dataset was first reformatted into a chat-style structure, where each example was represented as a sequence of system, user, and assistant messages. The assistant's response corresponded to the target label ("σεξιστικό" or "όχι σεξιστικό" i.e. sexist/non-sexist). Fine-tuning Meltemi required an efficient procedure due to the model's large number of parameters and the limited hardware resources of Google Colab. The issue was addressed using parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation, a.k.a LoRA [44]. Specifically, the model was loaded with 4-bit quantization using BitsAndBytes (nf4 quantization and float16 computation), and LoRA was applied to the q_proj and v_proj layers with parameters r=8 and lora_alpha=16. The training pipeline used the Hugging Face transformers library and trl.SFTTrainer. Tokenization was handled using the model's AutoTokenizer, setting the padding token equal to the end-of-sequence (EOS) token. Fine-tuning was done on the training part of our dataset, we used the AdamW optimizer, a maximum sequence length of 256 tokens, and lasted for 5 epochs with small batch sizes and gradient accumulation due to memory constraints. Predictions on the validation and test sets were generated using the model's generate() function with deterministic decoding (do_sample=False). The outputs were then post-processed and were turned into binary labels for evaluation. Training via LoRA and quantization significantly reduced both memory usage and runtime, making the fine-tuning of Meltemi feasible in a low-resource environment.

#### 4.3.1.3 KriKri fine-tuning

To fine-tune the Llama-Krikri-8B-Instruct model, the Twitter dataset was also reformatted into an instruction-based chat format aligned with the model's conversational architecture. Each example was structured as a sequence of system, user, and assistant messages, where the assistant's reply represented the ground-truth label in Greek (e.g., "Ναι, το tweet είναι σεξιστικό." or "Όχι, το tweet δεν είναι σεξιστικό."). As in the case of Meltemi parameter-efficient fine-tuning (PEFT) was performed using LoRA, with low-rank adapters (r=8, lora_alpha=16) inserted into the q_proj and v_proj layers. The model was quantized to 4-bit precision using BitsAndBytes (nf4 quantization with float16 computation), enabling training on limited hardware such as Google Colab. The tokenizer from the original model was used with the padding token set to the end-of-sequence (EOS) token, and the maximum input length was capped at 256 tokens. Fine-tuning was conducted using the SFTTrainer class from the trl library, with 5 training epochs and the AdamW optimizer. For inference, predictions on the validation and test sets were generated using the model's generate() function with greedy decoding, i.e., do_sample=False. The responses generated were post-processed into binary labels and evaluated using standard classification metrics.

### 4.3.2 In-Context learning

The second approach explored in this study was In-Context Learning (ICL), also referred to as prompt-based learning. This method relies on the use of prompts to guide pre-trained language models toward producing task-relevant outputs without any fine-tuning. Specifically, we conducted a series of zero-shot and few-shot experiments, using prompts that varied in size: 0, 2, 4, 6, and 8 labeled examples (shots). These prompts were designed to demonstrate the desired task behavior (i.e., distinguishing sexist from non-sexist content in Greek) and were formatted consistently across all models tested. The complete set of prompts used is provided in Appendix III. The same prompts were applied to all models to ensure comparability of results. The models evaluated using this method were: Meltemi-7B-Instruct, Llama-Krikri-8B-Instruct, GPT-3.5-turbo, and GPT-4o. Since no fine-tuning was performed, the used model relied entirely on the quality of the prompts and its ability to interpret them. In the zero-shot setting, where only a task description was provided without examples, models were generally able to correctly identify explicit forms of sexism, such as overt abuse or direct slurs, as preliminary experiments revealed. However, they frequently misclassified two particular types of inputs: (1) non-sexist comments containing offensive or aggressive language, which were wrongly labeled as sexist, and (2) indirectly sexist comments lacking abusive vocabulary, such as those expressing gender stereotypes, or sexist humor. To address these weaknesses, representative examples of both types were intentionally included in the few-shot prompts: (a) non-sexist utterances with toxic or insulting tone, and (b) sexist comments expressed in subtle or implicit ways. These additions were made through iterative experimentation using tweets from the training set, selected to be diverse and illustrative of the task's nuanced nature. Evaluation was initially performed on the 1,000-example validation set, followed by testing on the held-out test set, also consisting of 1,000 tweets.

#### 4.3.2.1 Meltemi and Kri-Kri Prompting

For the in-context learning experiments, both the Meltemi-7B-Instruct-v1.5 and Llama-KriKri-8B-Instruct models were loaded using 4-bit quantization via the BitsAndBytes

library to ensure memory efficiency, particularly within the Colab environment. In both cases, the classification task was framed as a structured dialogue in Greek, using a system prompt that defined the model's role (as a sexism detector) and presented each tweet for evaluation. A short definition of sexism was provided, and the prompt constrained the model's response to one of two valid options: "Ναι, το tweet είναι σεξιστικό." or "Όχι, το tweet δεν είναι σεξιστικό." Prompt formatting and tokenization were handled using Hugging Face's chat_template and apply_chat_template methods. Inference was performed using the generate() function with greedy decoding (do_sample=False), and the model's responses were post-processed into binary labels ("σεξιστικό" or "όχι σεξιστικό") before evaluation against ground truth annotations using standard classification metrics.

#### 4.3.2.2 GPT-3.5 and GPT-4o Prompting

For the in-context learning experiments with GPT-3.5-Turbo and GPT-4o, we used the OpenAI API to perform both zero-shot and few-shot classification without any task-specific fine-tuning. In both setups, each tweet was embedded within a carefully crafted Greek-language prompt that defined sexism and explicitly constrained the model's response to one of two fixed outputs: "Ναι, το tweet είναι σεξιστικό." or "Όχι, το tweet δεν είναι σεξιστικό."; the same prompt was used for Meltemi and Krikri. Prompts were structured as conversations, with a system message assigning the model the role of a sexism detector and a user message presenting the tweet for classification. Predictions were generated using greedy decoding (temperature=0), ensuring consistent outputs for identical inputs. Model responses were post-processed and mapped to binary labels ("σεξιστικό" or "όχι σεξιστικό"), which were then evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

### 4.4 Evaluation Metrics

To assess model performance on the binary classification task (sexist vs. non-sexist), we used four standard evaluation metrics: Accuracy and, Precision, Recall, and F1-score per category. These metrics are derived from the basic components of a confusion matrix:

- True Positives (TP): the number of tweets correctly classified as sexist.
- False Positives (FP): the number of tweets incorrectly classified as sexist.
- False Negatives (FN): the number of tweets incorrectly classified as non-sexist.
- True Negatives (TN): the number of tweets correctly classified as non-sexist.

Based on these, we calculate the metrics as follows:

- Accuracy: It represents the overall proportion of correctly classified samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + \text{FP} + \text{FN}}$$

- Precision: It measures the proportion of correctly predicted sexist tweets among all tweets labeled as sexist by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: It quantifies the model's ability to correctly identify actual sexist tweets.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-score: The F1-score provides a harmonic mean between Precision and Recall, offering a single metric that balances both false positives and false negatives.

$$\text{F1} - \text{score} = \frac{2 \times Precison \times Recall}{Precison + Recall}$$

While accuracy gives a general sense of correctness, it can be misleading in imbalanced datasets, such as the one used in this study, where the number of non-sexist tweets is significantly larger than sexist tweets. In such cases, a model might achieve high accuracy simply by predicting the majority class. For this reason, F1-score is particularly important, as it takes into account both types of classification errors (FP and FN) and provides a more reliable indicator of a model's ability in detecting the minority class (sexist content).

Finally, we report macro-averaged metrics: macro-precision, macro-recall, and macro-F1, which compute the metric independently for each class and then average the results, treating both classes equally regardless of size. This provides a more balanced evaluation by reflecting performance on both the majority and minority classes. The macro-F1 score is particularly important in our context, as it captures the trade-off between precision and recall for each class and highlights the model's ability to correctly identify sexist content, which is underrepresented in the dataset.

### 4.5 Experimental Results on validation set

All models were evaluated first on the validation set. The detailed classification metrics per class (i.e., for sexist and non-sexist tweets), along with their macro-averages, are presented in Table 9.

#### 4.5.1 Fine-tuning BERT

Among all fine-tuning variations of Greek BERT that were tested, the baseline model without class weights and additional synthetic data achieved the best overall performance, with an accuracy of 0.821 and a macro-F1 score of 0.759; this model was used as a baseline for comparison with LLMs (see Table 9). These results indicate that, despite the class imbalance, the model was able to achieve competitive results effectively without incorporating weighting.

The weighted-loss version yielded nearly identical outcomes, with an accuracy of 0.820 and macro-F1 of 0.757, suggesting that the inclusion of class weights had minimal impact on overall performance. In contrast, the introduction of synthetic data through SMOTE had a negative impact on model performance. As the proportion of synthetic examples increased, from one-third, to two-thirds, to full augmentation, the model's effectiveness progressively declined; see Table 8. This trend suggests that the generated examples may have introduced noise or failed to reflect the nuanced characteristics of authentic sexist language. Specifically, when one-third of the SMOTE-generated data was used, the model achieved an accuracy of 0.808 and a macro-F1 score of 0.729. With two-thirds, performance dropped to 0.784 accuracy and 0.725 macro-F1, while full augmentation led to a further decline, reaching 0.750 accuracy and 0.693 macro-F1; see Table 8.

**Table 8: Results on Val Set for BERT fine-tuned on SMOTE data**

| Val set | Class 0 (non-sexist) | Class 1 (sexist) | Macro average of 0, 1 | Accuracy |
|---|---|---|---|---|
| GreekBert + 1/3 SMOTE data | 0.891/0.860/0.875 | 0.550/0.620/0.583 | 0.721/0.740/0.729 | 0.808 |
| GreekBert + 2/3 SMOTE data | 0.916/0.798/0.583 | 0.502/0.736/0.597 | 0.709/0.767/0.725 | 0.784 |
| GreekBert + 3/3 SMOTE data | 0.897/0.764/0.825 | 0.468/0.703/0.562 | 0.683/0.733/0.693 | 0.750 |

The worst-performing BERT-based configuration by a significant margin was the partial fine-tuning setup, in which all BERT encoder layers were frozen and only the final classification head was trained. This outcome suggests that the fixed, pretrained representations from Greek BERT were insufficient for effectively capturing the nuances of sexist language in Greek tweets. Full model fine-tuning was essential to adapt the model to the specific characteristics of the task. The macro F1-score reached only 0.454, while the class-specific F1-score for the sexist category dropped sharply to 0.034, indicating severe difficulties in detecting the minority class. Despite an accuracy of 0.778, this result highlights the misleading nature of accuracy in imbalanced classification settings.

Based on the above results, as already mentioned only the baseline Greek BERT model without class weights and additional synthetic data is included in the final comparison table for the validation set; see Table 9, as it outperformed the other configurations.

### 4.5.2 Fine-tuning Meltemi and KriKri

Despite being loaded in 4-bit precision using quantization, a technique that often leads to performance degradation, fine-tuned Meltemi-7B-Instruct performed well. Specifically, when parameter-efficient fine-tuning via LoRA was used, Meltemi-7B-Instruct achieved an accuracy of 0.841 and a macro-F1 score of 0.794 on the validation set (Table 9); e.g., in terms of accuracy it surpassed the best BERT model. On the other hand Krikri-

8B-Instruct (4-bit quantization) delivered the strongest performance among all fine-tuned models, achieving an accuracy of 0.876 and a macro-F1 score of 0.804 on the validation set; winning by a significant margin Meltemi-7B-Instruct.

### 4.5.3 Prompting Meltemi and KriKri

We evaluated the performance of the Meltemi-7B-Instruct model under various few-shot prompting configurations, ranging from 0-shot to 8-shot. Performance improved significantly as the number of in-context examples increased. While the 0-shot configuration produced modest results (macro-F1: 0.634, accuracy: 0.661), adding examples led to marked gains. The best performance was observed with 6-shot prompting, achieving 0.667 macro-F1 and 0.728 accuracy, followed closely by the 8-shot (0.662 macro-F1, 0.723 accuracy) and 4-shot setups (0.664 macro-F1, 0.725 accuracy). This demonstrates Meltemi's strong capacity to benefit from in-context examples, despite being loaded with 4-bit quantization.

We also assessed the performance of the Krikri-8B-Instruct model across multiple few-shot prompting configurations (0–8 shots). Overall, the model demonstrated strong and stable performance, with only minor fluctuations across configurations. The highest macro-F1 score (0.714) and accuracy (0.756) were achieved using the 6-shot setup, closely followed by the 2-shot configuration (macro-F1: 0.701, accuracy: 0.759) and the 0-shot configuration (macro-F1: 0.699, accuracy: 0.758). Also it was proven better than Meltemi, the other Greek model. This is somewhat expected, e.g. because it is a larger model (+1B more parameters).

### 4.5.4 Prompting GPT-3.5 turbo and GPT-4o

The GPT-3.5-turbo model was evaluated across several prompting configurations, from 0-shot to 8-shot. Performance generally improved as more in-context examples were added. In the 0-shot setting, the model achieved an accuracy of 0.706 and a macro-F1 score of 0.665. With 2-shot prompting, results improved slightly to 0.716 accuracy and 0.675 macro-F1. The 4-shot setup reached 0.733 accuracy and 0.689 macro-F1. The best performance was achieved with 6-shot prompting (accuracy: 0.778, macro-F1: 0.713). The 8-shot configuration yielded 0.710 accuracy and 0.669 macro-F1, showing a slight drop compared to 6-shot.

The more recent OpenAI model, GPT-4o demonstrated the strongest overall performance among all evaluated models. In the few-shot prompting experiments, the 6-shot configuration yielded the highest accuracy (0.877) and macro-F1 score (0.829), outperforming all other models and setups (e.g. 2/4/8 shot). Even the other GPT-4o prompt configurations (2/4/8 shot) showed consistently high results, confirming the model's robustness across different numbers of in-context examples. These results also surpassed those achieved by the fine-tuned models, establishing GPT-4o as the best-performing model in this study.

**Table 9: Results on Val Set**

| Validation set | Class 0 (non-sexist) | Class 1 (sexist) | Macro average of 0,1 | Accuracy |
|---|---|---|---|---|
| Fine-tune Greek BERT - no class weights - no synthetic data | 0.914/0.850/0.881 | 0.573/0.716/0.636 | 0.743/0.783/0.759 | 0.821 |
| Fine Tune Meltemi-7B-Instruct | 0.947/0.844/0.892 | 0.5990/0.831/0.696 | 0.773/0.837/0.794 | 0.841 |
| Fine Tune Krikri-8B-Instruct | 0.898/0.948/0.922 | 0.771/0.616/0.685 | 0.834/0.782/0.804 | 0.876 |
| Meltemi-7B-Instruct 0_shot | 0.949/0.598/0.734 | 0.382/0.886/0.534 | 0.666/0.742/0.634 | 0.661 |
| Meltemi-7B-Instruct 2_shot | 0.939/0.607/0.737 | 0.380/0.858/0.527 | 0.656/0.733/0.632 | 0.662 |
| Meltemi-7B-Instruct 4_shot | 0.893/0736./0.807 | 0.421/0.685/0.522 | 0.657/0.711/0.664 | 0.725 |
| Meltemi-7B-Instruct 6_shot | 0.893/0.740/0.810 | 0.425/0.685/0.524 | 0.659/0.713/0.667 | 0.728 |
| Meltemi-7B-Instruct 8_shot | 0.891/0.735/0.806 | 0.419/0.680/0.518 | 0.655/0.708/0.662 | 0.723 |
| Krikri-8B-Instruct_0 shot | 0.908/0.768/0.832 | 0.466/0.721/0.566 | 0.687/0.745/0.699 | 0.758 |
| Krikri-8B-Instruct_2 shot | 0.909/0.768/0.833 | 0.468/0.726/0.569 | 0.688/0.747/0.701 | 0.759 |

| | | | | |
|---|---|---|---|---|
| Krikri-8B-Instruct 4_shot | 0.959/0.713/0.818 | 0.465/0.890/0.611 | 0.712/0.802/0.715 | 0.752 |
| Krikri-8B-Instruct 6_shot | 0.947/0.729/0.823 | 0.469/0.854/0.605 | 0.708/0.791/0.714 | 0.756 |
| Krikri-8B-Instruct 8_shot | 0.955/0.713/0.817 | 0.463/0.881/0.607 | 0.709/0.797/0.712 | 0.750 |
| GPT-3.5_0shot | 0.929/0.674/0.781 | 0.413/0.0817/0.549 | 0.671/0.746/0.665 | 0.706 |
| GPT-3.5 _2shot | 0.935/0.683/0.789 | 0.424/0.831/0.561 | 0.679/0.757/0.675 | 0.716 |
| GPT-3.5 _4shot | 0.931/0.711/0.806 | 0.441/0.812/0.572 | 0.686/0.762/0.689 | 0.733 |
| GPT-3.5 _6shot | 0.902/0.804/0.850 | 0.495/0.688/0.575 | 0.698/0.746/0.713 | 0.778 |
| GPT-3.5 _8shot | 0.929/0.679/0.785 | 0.417/0.817/0.552 | 0.673/0.748/0.669 | 0.710 |
| GPT 4o_0shot | 0.961/0.792/0.868 | 0.544/0.885/0.674 | 0.753/0.839/0.771 | 0.813 |
| GPT 4o _2shot | 0.938/0.902/0.920 | 0.694/0.790/0.739 | 0.816/0.846/0.829 | 0.878 |
| GPT 4o_4shot | 0.944/0.864/0.902 | 0.628/0.817/0.710 | 0.786/0.840/0.806 | 0.854 |
| GPT 4o_6shot | 0.939/0.900/0.919 | 0.690/0.794/0.738 | 0.815/0.847/0.829 | 0.877 |
| GPT 4o_8shot | 0.942/0.893/0.917 | 0.679/0.803/0.736 | 0.810/0.847/0.826 | 0.874 |

### 4.6 Experimental Results on test set

The test set was reserved for the final evaluation of the best-performing models as well as for thorough error analysis (see Sec 4.7) Accuracy, macro-average scores along with detailed classification metrics for each class (i.e., sexist and non-sexist tweets) are presented in Table 10.

It is observed that the ranking of the models in the validation set remains the same in the test set, which showcases the consistency of the results. Specifically, prompt-based GPT-4o (6-shot) remains the best-performing model overall (see Table 10), maintaining top-level performance without any fine-tuning. The prompt-based Krikri (6-shot) is less effective than both the fine-tuned Meltemi and the Greek BERT model, while the fine-tuned KriKri still stands out as the most effective model among the fine-tuned ones. This stability in model ranking across both sets indicates that the models' performance is generalizable and not overly dependent on the specific characteristics of a specific set of tweets.

Overall, the results show consistent behavior between the validation and test sets, with only minor differences. The fine-tuned models exhibit slight decreases in macro average and accuracy on the test set, without significant performance degradation. KriKri remains the most reliable among them, with minimal variation and consistently high performance. In contrast, the prompt-based Krikri shows a more noticeable drop, particularly in macro average, while GPT-4o shows nearly identical performance in both evaluation sets; which is somewhat expected due to its larger number of parameters. In conclusion, GPT-4o in the prompt-based setting stands out, outperforming even the best fine-tuned models, confirming the effectiveness of few-shot learning with large language models.

**Table 10: Results on Test Set**

| Test Set | Class 0 (non-sexist) | Class 1 (sexist) | Macro average of 0, 1 | Accuracy |
|---|---|---|---|---|
| Fine-tune GreekBERT | 0.916/0.845/0.879 | 0.567/0.726/0.637 | 0.742/0.785/0.758 | 0.819 |
| Fine Tune Meltemi-7B-Instruct | 0.935/0.845/0.888 | 0.588/0.790/0.674 | 0.762/0.818/0.781 | 0.833 |
| Fine Tune Krikri-8B-Instruct | 0.902/0.930/0.916 | 0.721/0.639/0.678 | 0.811/0.785/0.797 | 0.867 |
| Krikri-8B-Instruct 6 shot | 0.966/0.736/0.836 | 0.491/0.909/0.638 | 0.729/0.822/0.737 | 0.774 |
| GPT 4o – 6 shot | 0.948/0.891/0.918 | 0.680/0.826/0.746 | 0.814/0.858/0.832 | 0.877 |

## 4.7 Error Analysis

As already mentioned, we have based our error analysis on the test set, which was reserved for the final evaluation of the best-performing models. The identified errors reflect the remaining challenges encountered in our dataset. For this analysis, we also leveraged the fine-grained annotations (i.e., 8 categories and sexism type) created during the dataset's construction, which allowed us to identify patterns and examine in which specific categories each model performs better or worse.

### 4.7.1 Fine-tuning BERT

For the fine-tuned Greek BERT model, we observed 60 misclassified (False Negatives) out of 219 sexist tweets; see Fig. 3. Table 11 shows detailed misclassification statistics per sexism type. Among the in total 60 misclassified sexist tweets, the majority i.e., 42 were instances of indirect sexism, and 18 were direct sexism. This difference is due to the fact that indirect sexist tweets are more than indirect (120 vs 99) but also more difficult to be identified. The latter is revealed from the misclassification rate which was significantly higher (almost double) for indirect sexism; i.e., 35% (42/120) for indirect and 18.2% (18/99) for direct. These findings highlight that the model struggles more with detecting indirect sexism compared to direct expressions. Also it is worth to be noted that 18/60 (30%) FN belong to the stereotypes category and no other category has more than 15%. This indicates the BERT struggles with the specific cases.

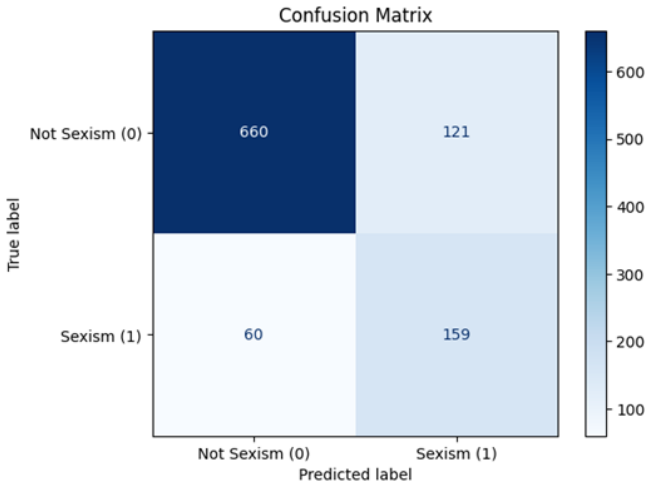



**Figure 3: Confusion Matrix of the Fine-Tuned Greek BERT Model on the Test Set**

**Table 11: False Negatives of the Fine-Tuned Greek BERT Model on the Test Set.**

| | FN | Sexism Type | FN per Sexism Type |
|---|---|---|---|
| *Incitement to violence* | 0 | direct | 18 |
| *Insults and slurs* | 9 | direct | |
| *Objectification sexualization* | 9 | direct | |
| *Stereotypes* | 18 | indirect | 42 |
| *Diminishing comments* | 4 | indirect | |
| *Sexism through denial* | 3 | indirect | |
| *Humor and irony* | 4. | indirect | |
| *Benevolent sexism* | 4 | indirect | |
| **TOTAL** | 60 | indirect | |

Additionally, the model produced 121 misclassified non-sexist tweets, often caused by overgeneralizing from surface-level cues like toxic or aggressive language, even when the context was not sexist.

### 4.7.2 Fine-tuning Meltemi

For the fine-tuned Meltemi-7B-Instruct model, we found 46 false negatives; see Fig. 4. Among them, 17 were related to direct sexism, and 29 to indirect sexism. The misclassification rate was higher for indirect sexism (24.16%) compared to direct sexism (17.1%). This suggests that the Meltemi model as in the case of BERT also faces challenges detecting indirect sexism. In general the model struggles with the sexism category if compared to BERT and Table 12 which presents misclassification statistics per sexism type shows that the errors are distributed to many types, e.g. “Stereotypes”, “Humor and irony”, “Insults and slurs” etc.

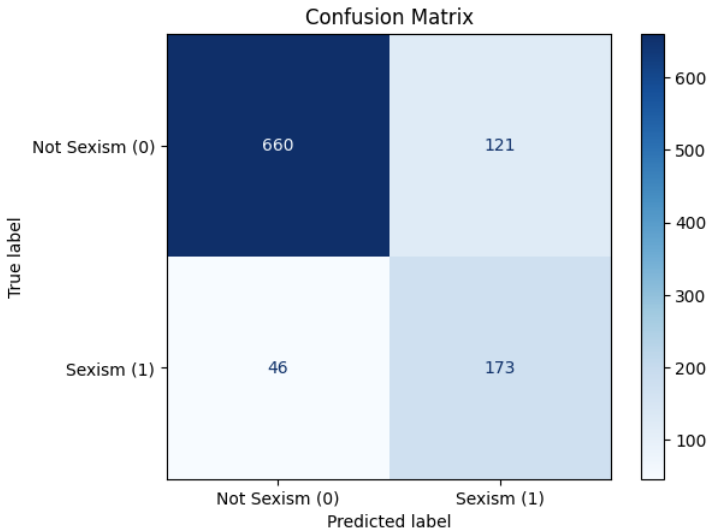



**Figure 4: Confusion Matrix of the Fine-Tuned Meltemi Model on the Test Set**

**Table 12: False Negatives of the Fine-Tuned Meltemi Model on the Test Set.**

| | FN | Sexism Type | FN per Sexism Type |
|---|---|---|---|
| *Incitement to violence* | 0 | direct | 17 |
| *Insults and slurs* | 11 | direct | |
| *Objectification sexualization* | 6 | direct | |
| *Stereotypes* | 10 | indirect | 29 |
| *Diminishing comments* | 6 | indirect | |
| *Sexism through denial* | 4 | indirect | |
| *Humor and irony* | 6 | indirect | |
| *Benevolent sexism* | 3 | indirect | |
| **TOTAL** | 46 | indirect | |

Additionally, 121 false positives were produced (Fig 4), often due to overgeneralizing surface-level cues, such as emotionally charged language or references to women, even in neutral contexts.

### 4.7.3 Fine-tuning Kri-Kri

The fine-tuned Kri-Kri model produced 79 false negatives, with 58 corresponding to indirect sexism and 21 to direct sexism. Again the misclassification rate for indirect sexism was notably higher (48.3%) than for direct sexism (21.21%). The model generated 54 false positives, often triggered by emotionally charged language that was not actually sexist.

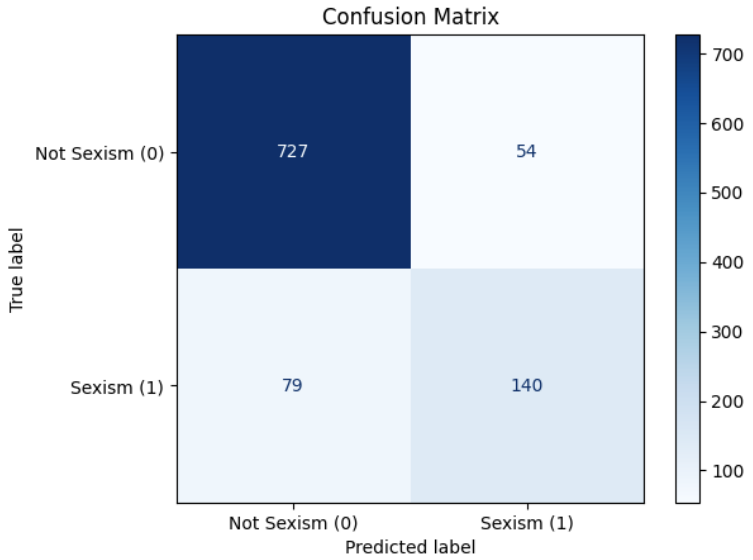



**Figure 5: Confusion Matrix of the Fine-Tuned Kri-Kri Model on the Test Set**

**Table 13: False Negatives of the Fine-Tuned Kri-Kri Model on the Test Set**

| | FN | Sexism Type | FN per Sexism Type |
|---|---|---|---|
| *Incitement to violence* | 0 | direct | 21 |
| *Insults and slurs* | 6 | direct | |
| *Objectification sexualization* | 15 | direct | |
| *Stereotypes* | 17 | indirect | 58 |
| *Diminishing comments* | 6 | indirect | |
| *Sexism through denial* | 4 | indirect | |
| *Humor and irony* | 26 | indirect | |
| *Benevolent sexism* | 5 | indirect | |
| **TOTAL** | 79 | indirect | |

### 4.7.4 Prompting Kri Kri

In the evaluation of the Kri-Kri prompts model, 20 false negatives were identified, 16 related to indirect sexism and only 4 to direct sexism. The misclassification rate for indirect sexism was higher (13.3%) compared to direct sexism (4%).

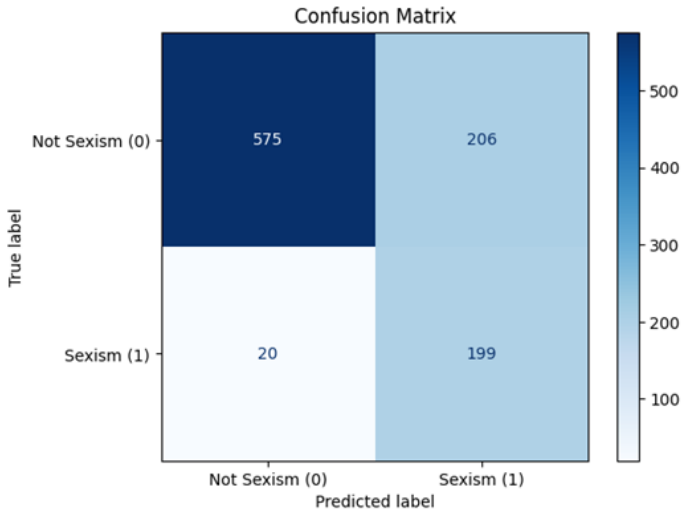



**Figure 6: Confusion Matrix for the prompt-based Kri-Kri Model Evaluated on the Test Set**

**Table 14: False Negatives of the prompt-based Kri-Kri Model on the Test Set**

| | FN | Sexism Type | FN per Sexism Type |
|---|---|---|---|
| *Incitement to violence* | 0 | direct | 4 |
| *Insults and slurs* | 2 | direct | |
| *Objectification sexualization* | 2 | direct | |
| *Stereotypes* | 5 | indirect | 16 |
| *Diminishing comments* | 3 | indirect | |
| *Sexism through denial* | 2 | indirect | |
| *Humor and irony* | 4 | indirect | |
| *Benevolent sexism* | 2 | indirect | |
| **TOTAL** | 20 | indirect | |

Notably, this model produced the fewest false negatives among all the models, but it also generated the highest number of false positives, with 206 instances.

### 4.7.5 Prompting GPT-4o

Finally, for the GPT-4o prompts model, we recorded 38 false negatives, 27 related to indirect sexism and 11 to direct sexism. In table 15 you can see the distribution of these errors; most belong to the “Humor and irony” category. Again the misclassification rate for indirect sexism (22.5%) was higher than for direct sexism (11.1%).

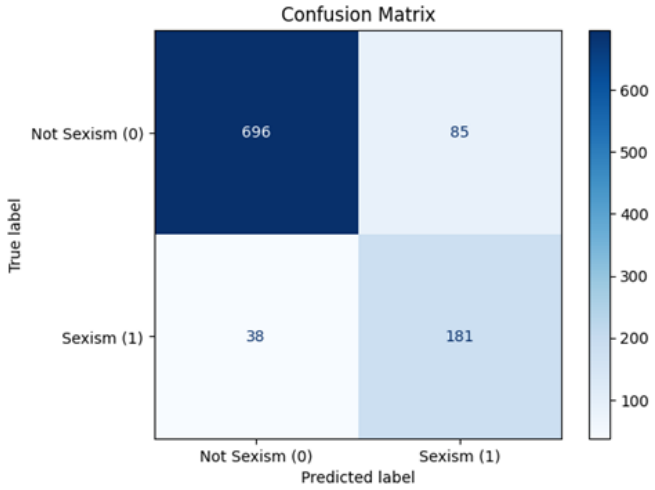



**Figure 7: Confusion Matrix for the prompt-based GPT4o Model Evaluated on the Test Set.**

**Table 15: False Negatives of the prompt-based GPT4o Model on the Test Set.**

| | FN | Sexism Type | FN per Sexism Type |
|---|---|---|---|
| *Incitement to violence* | 0 | direct | 11 |
| *Insults and slurs* | 6 | direct | |
| *Objectification sexualization* | 5 | direct | |
| *Stereotypes* | 6 | indirect | 27 |
| *Diminishing comments* | 5 | indirect | |
| *Sexism through denial* | 1 | indirect | |
| *Humor and irony* | 12 | indirect | |
| *Benevolent sexism* | 3 | indirect | |
| **TOTAL** | 38 | indirect | |

In addition, the model produced 85 false positives, again suggesting over-sensitivity to language features that resemble sexism but lack sexist intent.

# 5. CONCLUSIONS AND FUTURE WORK

This thesis presents, to the best of our knowledge, the first manually annotated Greek-language dataset specifically designed for the detection of online sexism against women. By collecting and annotating data from three major platforms, Twitter, Reddit, and YouTube, we aimed to reflect the diversity and complexity of sexist discourse in Greek digital spaces. Our detailed annotation scheme, which included both binary and fine-grained labels, allowed for the nuanced classification of sexist content, capturing a broad spectrum ranging from explicit abuse to more subtle, implicit expressions such as benevolent sexism and stereotypical assumptions.

Even with limited computational resources, our experimental evaluation demonstrated that fine-tuned models, particularly Llama-KriKri-8B-Instruct, outperformed both zero-shot and few-shot prompting approaches in most scenarios, especially in terms of macro F1-score. This finding aligns with previous literature [31], [32] for other languages, where it is similarly observed that performance without fine-tuning, relying solely on prompt-based methods, tends to be inferior.

However, it is worth noting that GPT-4o, even without fine-tuning, achieved the highest overall performance across evaluation metrics. This highlights the potential of very large state-of-the-art large language models when used through carefully crafted prompting strategies. This outcome is further supported by the fact that GPT-4o is a highly powerful model, with a significantly larger number of parameters compared to the other models evaluated. Additionally, our experiments with GPT-4o were conducted using the official OpenAI API that uses the full model, not a quantized version of it; it is known that quantization compromises performance in favor of efficiency.

Moreover, our error analysis revealed consistent challenges across all models in identifying indirect and nuanced forms of sexism, confirming the complexity of the task and the limitations of surface-level lexical cues. Nonetheless, the strong performance of fine-tuned instruction-tuned models, highlights the potential of language-specific LLMs for tackling sensitive sociolinguistic tasks in under-resourced languages like Greek.

There are several directions in which this research could be extended. First, the current evaluation focused exclusively on binary classification (sexist vs. non-sexist). Expanding this framework to include the fine-grained classification would allow for deeper insights into the nature of sexist discourse and enable models to distinguish between different types of harmful language, such as objectification, stereotyping, or benevolent sexism.

Although 3 datasets were created; Twitter (X), Reddit, and YouTube, the experiments were conducted exclusively on the Twitter (X) dataset. Future work could isolate and analyze platform-specific data to examine whether model performance varies across platforms with distinct discourse norms, community dynamics, and linguistic patterns.

Regarding annotation, while the process was informed by detailed guidelines and included inter-annotator agreement metrics, it remains inherently subjective, particularly for comments involving irony, ambiguity, or strong contextual dependencies. Although three annotators were involved overall, the majority of the data was labeled by a single individual, which may introduce annotator bias, reflecting their specific interpretations and value judgments. To address this limitation, future work could involve a larger and more balanced pool of annotators, as well as explore the use of soft labels, where disagreements are retained as probability distributions rather than being resolved into a single "correct" label. This would better reflect the uncertainty present in many examples and allow models to learn from ambiguity. Additionally, incorporating multiple

perspectives could reduce bias and enhance both fairness and robustness in model training.

Finally, integrating reasoning capabilities into model outputs could further improve the system's trustworthiness. Rather than simply producing a binary or categorical label, models could be designed to generate explanations or justifications for their decisions, helping users understand the rationale behind each prediction.

In conclusion, this work provides a valuable starting point for the computational study of online sexism in the Greek language, contributing both a dataset and a comparative evaluation of language models. It opens the path for more inclusive, accurate, and socially aware NLP systems tailored to the Greek digital sphere.

# ACRONYMS

| | |
|---|---|
| API | Application Programming Interface |
| AUEB | Athens University of Economics and Business |
| BERT | Bidirectional Encoder Representations from Transformers |
| Bi-LSTM | Bidirectional Long Short-Term Memory |
| BPE | Byte Pair Encoding |
| CNNs | Convolutional Neural Networks |
| DTFN | Dual-Transformer Fusion Network |
| DPO | Direct Preference Optimization |
| EDOS | Explainable Detection of Online Sexism |
| EOS | End of Sequence |
| EXIST | Sexism Identification in Social Networks |
| FN | False Negatives |
| FP | False Positives |
| GPU | Graphics Processing Unit |
| GRU | Gated Recurrent Unit |
| ICL | In-Context Learning |
| ILSP | Institute for Language and Speech Processing |
| LLMs | Large Language Models |
| LoRA | Low-Rank Adaptation |
| LSTM | Long Short-Term Memory |
| LwD | Learning with Disagreement |
| MFE | Multimodal Fusion Ensemble |
| NLP | Natural Language Processing |
| PEFT | Parameter-Efficient Fine-Tuning |
| QLoRA | Quantized Low-Rank Adaptation |
| RNNs | Recurrent Neural Networks |
| RLHF | Reinforcement Learning from Human Feedback |
| SFT | Supervised Fine-Tuning |
| SMOTE | Synthetic Minority Over-sampling Technique |
| SVM | Support Vector Machines |
| TN | True Negatives |
| TP | True Positives |

# APPENDIX I

## Lexicon Used for Data Collection:

| LEXICON | | |
|---|---|---|
| **Common nouns** | | **Number of Tweets** |
| 1 | Άγαμη | 425 |
| 2 | Αγάμητη | 38 |
| 3 | Ανδροκτονία | 37 |
| 4 | Ανοργασμικιά | 103 |
| 5 | Ανύπαντρη | 39 |
| 6 | Αξύριστη | 85 |
| 7 | Βίζιτα | 282 |
| 8 | Γκόμενα | 2,996 |
| 9 | Γκόμενες | 2,335 |
| 10 | Γκρίνια | 1,419 |
| 11 | Γυναίκα | 45 |
| 12 | Γυναίκες | 8,128 |
| 13 | Δεσποινίς | 4,319 |
| 14 | Δικαιώματα | 1811 |
| 15 | Κάμπια | 660 |
| 16 | Καριόλα | 770 |
| 17 | Κοπελια | 2,213 |
| 18 | Κοπελίτσα | 283 |
| 19 | Κορίτσι | 2,265 |
| 20 | Κορίτσια | 2,192 |
| 21 | Κότα | 6,150 |
| 22 | Κουζίνα | 2,863 |
| 23 | Κουτσομπόλα | 370 |
| 24 | Κυρία | 2,121 |
| 25 | Μαντάμ | 1,560 |
| 26 | Μουνάρα | 1,870 |
| 27 | Μωρή | 1,966 |
| 28 | Ξανθιά | 3,232 |
| 29 | Ομορφούλα | 131 |

| | | |
|---|---|---|
| 30 | Παρθένα | 1,171 |
| 31 | Παρθένες | 279 |
| 32 | Πιάτο | 2,244 |
| 33 | Πλαστική | 1,636 |
| 34 | Πουτάνα | 2,575 |
| 35 | Πουτάνες | 176 |
| 36 | Πούτσο | 1,284 |
| 37 | Σαύρα | 421 |
| 38 | Σέξι | 2,120 |
| 39 | Σούργελο | 1,161 |
| 40 | Τσούλα | 28 |
| 41 | Υστερία | 564 |
| 42 | Φεμινάζι | 92 |
| 43 | Φεμινιστικά | 1 |
| 44 | Φεμινίστρια | 475 |
| 45 | Φεμινίστριες | 828 |
| **TOTAL_1** | | **67,554** |
| | | |
| | **Proper Names** | **Number of Tweets** |
| 1 | Αχτσιόγλου | 1,470 |
| 2 | Δούρου | 1,380 |
| 3 | Καϊλή | 1,262 |
| 4 | Κωνσταντοπούλου | 2,084 |
| 5 | Μαρέβα | 1,762 |
| 6 | Μελέτη | 2,187 |
| 7 | Μενδώνη | 801 |
| 8 | Περιστέρα | 913 |
| 9 | Μπίκα | 409 |
| 10 | Μπεκατώρου | 1,321 |
| **TOTAL_2** | | **13,589** |

# APPENDIX II

Content Warning: Indicative examples of sexist language are presented to describe the classification and analyze the examples. The guidelines annotation is provided in Greek.

## GUIDELINES ANNOTATION - ΣΧΗΜΑ ΕΠΙΣΗΜΕΙΩΣΗΣ

Το βασικό σχήμα επισημείωσης, είναι binary:

***Σεξιστικό (S)***

***Μη σεξιστικό (N)***

### *ΠΡΩΤΟ ΕΠΙΠΕΔΟ ΕΠΙΣΗΜΕΙΩΣΗΣ*

Πριν το βασικό διαχωρισμό σε σεξιστικό και μη σεξιστικό σχόλιο, γίνεται μια γενικότερη διάκριση σε **toxic** και **non toxic** σχόλια, που προηγείται της διάκρισης σεξιστικού και μη σεξιστικού. Τα **τοξικά σχόλια** είναι όλα τα σχόλια που περιλαμβάνουν προσβλητική, επιθετική ή υποτιμητική γλώσσα, τα οποία δημιουργούν αρνητικό κλίμα, προκαλούν συναισθηματική βλάβη, ή υπονομεύουν τη συναισθηματική ευημερία και την ψυχική υγεία του ατόμου που τα δέχεται. Αυτά τα σχόλια μπορεί να περιλαμβάνουν βρισιές, απειλές, χλευαστικά σχόλια ή σχόλια που υποκινούν μίσος ή βία. Σε αυτό το πρωταρχικό επίπεδο δεν εξετάζουμε ακόμα εάν ένα σχόλιο είναι σεξιστικό ή όχι αλλά μόνο αν αποτελεί ρητορική μίσους.

Επίσης, προστέθηκε και η κατηγορία ***other (O)*** για tweets που δεν είναι γραμμένα με ελληνικούς χαρακτήρες και δεν θα έπρεπε να βρίσκονται στο dataset, για κενά tweets που δεν έχουν διαγραφεί, για ελλειπτικές προτάσεις που δεν μπορεί ο αναγνώστης να καταλάβει το περιεχόμενο αυτών κλπ. Τα σχόλια που έχουν επισημειωθεί ως other θα διαγραφούν από το dataset και δεν επισημειώνονται στα επόμενα επίπεδα.

### *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET POLARITY*

| **COMMENTS** | | **POLARITY** |
|---|---|---|
| 1. | *"Άντε γαμησου μωρή πουτανα"* | **Toxic** |
| 2. | *"τι λες βρε σουργελο που στην ευρωλιγκα επαιζες οταν η Φορτιτουτο ειχε μπασκετ στην Α1 Ιταλιας και η βιρτους λεγοταν Κιντερ βλακα"* | **Toxic** |
| 3. | *"Βρέθηκε πλαστική σακούλα στο πιο βαθύ σημείο των ωκεανών"* | **Non Toxic** |
| 4. | "Με δύο ακόμη επεισόδια για την αυλαία, αν δεν έρθει το τέλος του κόσμου και να γίνουν *όλα πουτάνα από κόσμο*, τότε..Κρίμα. #Banshee " | **Non Toxic** |
| 5. | "Έτσι...." | **Other** |
| 6. | "Οι 5 τύποι των Επιδαυρίων:<br>Κύρια Εικόνα:<br>Η κυρί..." | **Other** |

Παρατηρούμε ότι η **πρόταση_4** που περιέχει μια λέξη που ενδέχεται να θεωρηθεί προσβλητική ("*πουτάνα*"), δεν συνιστά αυτόματα προσβλητικό λόγο. Η διαφοροποίηση έγκειται στο πλαίσιο και τον τόνο του σχολίου. Στη συγκεκριμένη περίπτωση, η λέξη χρησιμοποιείται με τη σημασία της υπερβολής και του σχολιασμού για μια κατάσταση, χωρίς να απευθύνεται σε κάποιο άτομο ή ομάδα με σκοπό την υποτίμηση ή την προσβολή. Έτσι, ο λόγος δεν θεωρείται προσβλητικός.

### *ΔΕΥΤΕΡΟ ΕΠΙΠΕΔΟ ΕΠΙΣΗΜΕΙΩΣΗΣ*

Το βασικό επίπεδο της ταξινόμησής μας κάνει μια δυαδική διάκριση μεταξύ σεξιστικών και μη σεξιστικών σχολίων. Ορίζουμε το σεξιστικό περιεχόμενο ως **οποιαδήποτε κατάχρηση, σιωπηρή ή ρητή, που απευθύνεται προς τις γυναίκες με βάση το φύλο τους, ή με βάση το συνδυασμό του φύλου τους με ένα ή περισσότερα άλλα χαρακτηριστικά ταυτότητας (π.χ. μαύρες γυναίκες ή μουσουλμάνες γυναίκες).**

Επισημειώνονται ως σεξιστικά μόνο τα σχόλιο που έχουν ως **target τις γυναίκες** (βιολογικά και κοινωνικά) και όχι σεξιστικά σχόλια που μπορεί να απευθύνονται σε άνδρες, λόγω του σκοπού της παρούσας εργασίας. Επίσης, προσβλητικά σχόλια που απευθύνονται σε γυναίκες αλλά δεν είναι σεξιστικά, επίσης δεν επισημειώνονται ως σεξιστικά. Αν για κάποιο σχόλιο δεν υπάρχει βεβαιότητα αν είναι σεξιστικό ή όχι, ή αν απευθύνεται σε γυναίκα, επιλέγεται να επισημειωθεί ως **μη σεξιστικό**.

Θεωρείται αυτόματα **ότι όλα τα σεξιστικά σχόλια είναι και τοξικά**, καθώς περιέχουν αρνητική και προσβλητική γλώσσα προς τις γυναίκες, αλλά όλα τα σχόλια που εμπεριέχουν προσβλητικό λόγο δεν είναι απαραίτητα σεξιστικά. Ένα σχόλιο μπορεί να είναι τοξικό χωρίς να περιέχει σεξιστικό περιεχόμενο, όχι όμως το αντίθετο.

All Sexist comments → toxic speech

**BUT**

All Toxic comments **NOT** Sexist speech

### *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET_SEXISM*

| **COMMENTS** | | **POLARITY** | **SEXISM** |
|---|---|---|---|
| 1. | *«Άντε γαμησου μωρή πουτανα»* | **Toxic** | **Sexist** |
| 2. | *“τι λες βρε σουργελο που στην ευρωλιγκα επαιζες οταν η Φορτιτουτο ειχε μπασκετ στην Α1 Ιταλιας και η βιρτους λεγοταν Κιντερ βλακα”* | **Toxic** | **Not Sexist** |
| 3. | *«Σούργελο του κερατά!!!»* * | **Toxic** | **Not Sexist** |
| 4. | *«Πωωωω το πιάσατε αυτο που ανέβασε η Ορφανίδουν αρχίζουν και ανοίγουν στόματα..να δω μωρή χαμούρα Κοψιάλη που θα κρυφτείς παλιοξεπλένη #cancel_kopsialis»* ** | **Toxic** | **Not sexist** |

| | | | |
|---|---|---|---|
| 5. | *«Οι δηλώσεις της ανόητης Ζωής Κωνσταντοπούλου, όπως τις αναπαράγει η Ομάδα Αλήθειας της Νέας Δημοκρατίας» **** | **Toxic** | **Not sexist** |
| 6. | *Γκομενες και μαλακιες.. Καμια σας δεν αξίζει.. Μπέσα τώρα!* | **Toxic** | **Sexist** |

ΕΠΕΞΗΓΗΣΕΙΣ:

**«Σούργελο του κερατά!!!»*

→ θα επισημειωθεί ως **toxic** και **Nonsexist** (είναι τοξικός λόγος γιατί περιλαμβάνει προσβλητική λέξη αλλά δεν γνωρίζουμε καν αν απευθύνεται σε γυναίκα ή άνδρα ώστε να εξεταστεί αν πρόκειται για σεξιστικό)

***«Πωωωω το πιάσατε αυτο που ανέβασε η Ορφανίδουν αρχίζουν και ανοίγουν στόματα..να δω μωρή χαμούρα Κοψιάλη που θα κρυφτείς παλιοξεπλένη #cancel_kopsialis»*

→ θα επισημειωθεί ως **toxic** και **Nonsexist** είναι τοξικός λόγος γιατί περιλαμβάνει προσβλητική λέξη αλλά δεν είναι σεξιστικό δεν απευθύνεται σαν σε γυναίκα.

****«Οι δηλώσεις της ανόητης Ζωής Κωνσταντοπούλου, όπως τις αναπαράγει η Ομάδα Αλήθειας της Νέας Δημοκρατίας»*

→ θα επισημειωθεί ως **toxic** και **Nonsexist** είναι τοξικός λόγος γιατί περιλαμβάνει προσβλητική λέξη, απευθύνεται σε γυναίκα, ωστόσο δεν πρόκειται για καθαρή σεξιστική επίθεση που γίνεται εξαιτίας του φύλου.

### *ΤΡΙΤΟ ΕΠΙΠΕΔΟ ΕΠΙΣΗΜΕΙΩΣΗΣ*

Σε επόμενο επίπεδο διαχωρίζονται τα σεξιστικά σχόλια σε 8 υποκατηγορίες.

**0.Υποκίνηση για πρόκληση βίας, σωματική, σεξουαλική, παραβιαστική συμπεριφορά κλπ.**

*«Να σε βιάσουν να ηρεμήσεις!»*

**1. Προσβολή, βρισιές**

*«Ντύνονται σαν πόρνες πλέον οι γυναίκες»*

**2. Αντικειμενοποίηση, σεξουαλικοποίηση**

«*Μουναρά! Εμπαινα χωρίς δεύτερη σκέψη*»

**3.Στερεότυπα**

«*Τις δουλειές του σπιτιού τις κάνουν καλύτερα οι γυναίκες»*

**4.Υποτιμητικά σχόλια (π.χ. χρήση υποκοριστικών, πατροναριστικά σχόλια)**

*«Το τουήτερ κοριτσάκι μου υπάρχει εδώ και δέκα χρόνια. Δεν ξεκίνησε τώρα.»*

**5.Γλώσσα που αρνείται την ύπαρξη διακρίσεων , δικαιολογεί την ανισότητα των φύλων και υποτιμά εν γένει τα έμφυλα ζητήματα, σεξισμός μέσα από άρνηση**

*«Και γιατί γυναικοκτονία και όχι ανδροκτονία ; »*

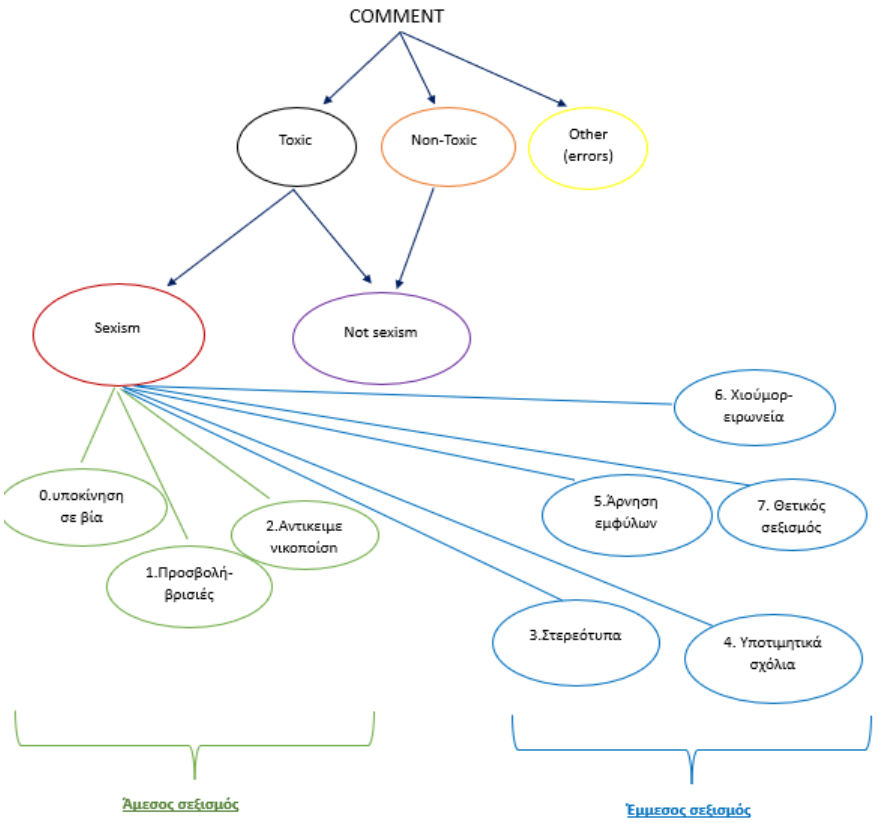
**6.Ειρωνεία, «χιούμορ» που υποβόσκει σεξισμό**

*«Γιατί οι γυναίκες πάντα έχουν τις καλύτερες ιδέες; Ποιος ξέρει, γιατί τις ακούει κανείς;»*

**7.Θετικός σεξισμός, λέξεις μη προσβλητικές, κομπλιμέντα**

*«Από τη γυναίκα πηγάζουν τα καλύτερα, τα ευγενέστερα γιατί μόνο αυτή ξέρει να αγαπά, να θυσιάζεται, να δίνει χωρίς να ζητά»*

*Οι παραπάνω κατηγορίες είναι διαβαθμισμένες 0 το περισσότερο σοβαρό, άμεσο και φανερό και 7 το λιγότερο σοβαρό. Η διαβάθμιση αυτή εξυπηρετεί την επιλογή του καταλληλότερου label σε διφορούμενα σχόλια. Οι κατηγορίες 0-2 αφορούν τον **άμεσο σεξισμό** ενώ οι κατηγορίες 3-7 τον **έμμεσο.** Αυτός ο διαχωρισμός σε έμμεσο και άμεσο σεξισμό δεν γίνεται από τον επισημειωτή, αλλά πραγματοποιείται αυτόματα μέσω συνάρτησης if στο excel, εφόσον έχει προηγηθεί η επιλογή του κατάλληλου label (S0-S7) από τον επισημειωτή.



## ΚΑΤΗΓΟΡΙΑ S0

***Υποκίνηση για πρόκληση βίας, σωματική, σεξουαλική, παραβιαστική συμπεριφορά κλπ.***

Η κατηγορία **S0** αφορά υποκίνηση για πρόκληση βίας, σωματικής, σεξουαλικής ή άλλης παραβιαστικής συμπεριφοράς εναντίον γυναικών και αναφέρεται σε κάθε σχόλιο ή ενέργεια που ενθαρρύνει, προτρέπει, επιδοκιμάζει, ή υποδεικνύει τη χρήση βίας ή άλλων ακραίων συμπεριφορών προς γυναίκες. Αυτού του είδους η γλώσσα είναι ιδιαίτερα επικίνδυνη, καθώς μπορεί να ενδυναμώσει την αποδοχή της βίας και να προάγει επιθετική ή καταχρηστική συμπεριφορά. Ειδικότερα, περιλαμβάνει απειλές, προσβολές ή οποιοδήποτε άλλο λόγο που δικαιολογεί ή νομιμοποιεί τη βία κατά γυναικών. Αυτές οι προτάσεις ή εκφράσεις, οι οποίες στοχεύουν σε γυναίκες, αναγνωρίζονται ως υποκίνηση βίας.

***«Να σε βιάσουν να ηρεμήσεις!»***

→ Ενθάρρυνση σεξουαλικής βίας εναντίον γυναικών.

***«Θέλει χαστούκια αυτή, και μάλιστα δυνατά!»***

→ Παρότρυνση για σωματική βία.

### *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S0*

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | *«θελω να μπω στο στούντιο να αστραψω ενα χαστούκι στην παπουτσακη -σαν αυτο της λιανη-κ να φύγω κυριος» ** | Toxic | Sexism | S0 |
| 2 | *«Καλά για κάποιες η γκρίνια είναι δεύτερη φύση τους... Δεν έχουν φάει απαραίτητα σφαλιάρες...» *** | Toxic | Sexism | S0 |
| 3 | *«Διακοπές σε κάμπινγκ; Καμιά γυμνή γκόμενά είδες κρυφά πίσω από τα δέντρα; :-)» **** | Toxic | Sexism | S0 |
| 4 | *«Σκάσε μωρή βρωμιαρα μην σε κάνω εγώ να σκάσεις» ***** | Toxic | Sexism | S0 |

1* απειλή για χρήση φυσικής βίας (χαστούκι) και ενδυνάμωση της αποδοχής τέτοιων πράξεων.

2** υποκινεί και κανονικοποιεί τη βία προς τις γυναίκες.

3*** υπονοεί σεξουαλική παρενόχληση.

4**** απειλή για βίαιη συμπεριφορά

## ΚΑΤΗΓΟΡΙΑ S1

***Προσβολή, βρισιές***

Η κατηγορία **S1** αφορά σεξιστικά σχόλια που περιέχουν άμεσες προσβολές με τη χρήση υβριστικών λέξεων ή φράσεων. Αυτού του είδους η γλώσσα χρησιμοποιείται για να απαξιώσει ή να ταπεινώσει τις γυναίκες με τη χρήση έντονα υποτιμητικών ή προσβλητικών χαρακτηρισμών. Τα σχόλια σε αυτή την κατηγορία αποτελούν μια

ανοιχτή, ωμή προσβολή που εκφράζουν απαξίωση προς τις γυναίκες. Σε αυτή την κατηγορία παρατηρούμε και πολλά μισογυνιστικά σχόλια.

***«Ντύνονται σαν πόρνες πλέον οι γυναίκες!»***

→ Άμεση προσβολή με βρισιά

***«Μωρή ηλίθια, πάλι μαλακίες κάνεις!»***

→ Ωμή και άμεση προσβολή με επιθετική γλώσσα.

***ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S1***

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | *«Άντε γαμησου μωρή πουτανα»* | Toxic | Sexism | S1 |
| 2 | *«Μωρη καριόλα άνοιξε κανένα βιβλίο όχι μόνο τα ποδάρια σου»* | Toxic | Sexism | S1 |
| 3 | *«Ειναι πουτανα όπως όλες.»* | Toxic | Sexism | S1 |
| 4 | *«Δεν χωνευω τις γυναικες ειναι τσουλες .Ολες τους. Τελος.»* | Toxic | Sexism | S1 |

**ΚΑΤΗΓΟΡΙΑ S2**

**Αντικειμενοποίηση, σεξουαλικοποίηση**

Η κατηγορία **S2** περιλαμβάνει σχόλια που αντικειμενοποιούν και σεξουαλικοποιούν τις γυναίκες, αγνοώντας την προσωπικότητα, τις ικανότητες ή τα συναισθήματά τους. Τα χαρακτηριστικά της κατηγορίας S3 είναι η εστίαση στο σώμα ή την εμφάνιση, σχόλια που αναφέρονται στις γυναίκες μόνο ως αντικείμενα σεξουαλικής ικανοποίησης, και σεξουαλικά υπονοούμενα συχνά –όχι πάντα- με χυδαία γλώσσα.

***«Μουναρά! Εμπαινα χωρίς δεύτερη σκέψη.»***

→ Χυδαία σεξουαλικοποίηση και αντικειμενοποίηση.

***«Οι ξανθιές είναι όλες για το κρεβάτι.»***

→ Γενίκευση που υποβιβάζει τις γυναίκες σε σεξουαλικά αντικείμενα.

***ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S2***

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | *«μουναρα!εμπαινα»* | Toxic | Sexism | S2 |
| 2 | *«Ο αντρας σε θελει κυρια στο TL και πουτανα στα DM.»* | Toxic | Sexism | S2 |
| 3 | *«Γιατί κορίτσι μου πάχυνες και χάλασες έτσι τον εαυτό σου;»* | Toxic | Sexism | S2 |
| 4 | *«Πάχυνε αυτή πάλι»* | Toxic | Sexism | S2 |

---

**ΚΑΤΗΓΟΡΙΑ S3**

**Στερεότυπα**

Η κατηγορία **S3** περιλαμβάνει σχόλια που ενισχύουν, αναπαράγουν ή αναφέρονται σε έμφυλα στερεότυπα, σχετικά με τους ρόλους, τις ικανότητες, τα χαρακτηριστικά ή τη

συμπεριφορά των γυναικών. Αυτά τα σχόλια βασίζονται σε προκαταλήψεις ή παραδοσιακές αντιλήψεις που περιορίζουν τις γυναίκες σε συγκεκριμένους ρόλους ή καθορίζουν την αξία τους με βάση κοινωνικά κατασκευασμένες προσδοκίες. Τα στερεότυπα που περιλαμβάνονται στην κατηγορία αυτή μπορεί εκφράζονται με τρόπο ξεκάθαρο ή μέσω υπονοημάτων. Ανεξάρτητα από την πρόθεση, αυτά τα σχόλια συμβάλλουν στη διατήρηση των έμφυλων ανισοτήτων. Τα στερεότυπα μπορεί να αφορούν ρόλους και ικανότητές , συμπεριφορά και χαρακτηριστικά , φυσική εμφάνιση, έμφυλες προσδοκίες κ.α.

***«Οι γυναίκες δεν είναι καλές οδηγοί.»***

→ Στερεότυπο για τη μειωμένη ικανότητα γυναικών.

***«Οι άντρες είναι καλύτεροι στη λήψη αποφάσεων από τις γυναίκες.»***

→ Στερεότυπο που παρουσιάζει τις γυναίκες ως λιγότερο ικανές.

***ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S3***

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | «σας παρακαλώ κυρα μου παντε πλύντε κανα πιάτο» | Toxic | Sexism | S3 |
| 2 | *«Στα κατηχητικά δεν υπάρχουν μικρά κορίτσια γιατί τα κορίτσια είναι μαρτυριάρικα....»* | Toxic | Sexism | S3 |
| 3 | *«σταμάτα τη γκρίνια, σα γκόμενα κάνεις. Έχει μπαλίτσα λέμε»* | Toxic | Sexism | S3 |
| 4 | *«RECPECT στα κοριτσια που συμπεριφεροντε σαν κυριες και οχι σαν πουτανες(!)»* | Toxic | Sexism | S3 |

## ΚΑΤΗΓΟΡΙΑ S4

### *Υποτιμητικά σχόλια*

Η κατηγορία **S4** αφορά σχόλια που περιέχουν υποτιμητικές εκφράσεις ή προσβολές προς τις γυναίκες, συχνά με τη χρήση υποκοριστικών, ή υπονοούμενων. Σε αντίθεση με την S1, όπου οι προσβολές είναι ωμές και ανοιχτές, στην S2 η υποτίμηση είναι πιο έμμεση και μπορεί να παρουσιάζεται ως σχόλιο "καλοπροαίρετο" ή "συμβουλευτικό". Αυτού του είδους τα σχόλια ενισχύουν τις έμφυλες ανισότητες και υποβαθμίζουν τις γυναίκες χωρίς να χρησιμοποιούν απαραίτητα χυδαία γλώσσα. Εκτός από τη χρήση υποτιμητικής γλώσσας και τη χρήση υποκοριστικών, μπορεί να έχουμε και πατροναριστικά σχόλια που προσπαθούν να καθοδηγήσουν ή να συμβουλέψουν τις γυναίκες με τόνο υποτιμητικό. Ενώ η S1 περιλαμβάνει ανοιχτές και ωμές προσβολές ή βρισιές, η S2 είναι πιο έμμεση και δεν χρησιμοποιεί χυδαία γλώσσα.

**«Αυτή η κοπελίτσα το μαλακομαγνήτη έχει, κρίμα.»**

→ Υποκοριστικό που υποτιμά τη γυναίκα.

**«Άκουσε με, καλή μου, και μην προσπαθείς να σκέφτεσαι μόνη σου.»**

→ Πατροναριστικός τόνος με υποτίμηση της νοημοσύνης.

## *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S4*

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | «κλεισαμε μια ωρα δουλεια κοπελια κουραστηκες κιολας αγαπουλα;» | Toxic | Sexism | S4 |
| 2 | «Το τουήτερ κοριτσάκι υπάρχει εδώ και δέκα χρόνια. Δεν ξεκίνησε τώρα.» | Toxic | Sexism | S4 |
| 3 | «Θίχτηκε η μικρή και αθώα κλωσσοπεριστέρα για τη "δήθεν ευνοϊκή ρύθμιση". Αν δεν θέλει να ασχολούνται μαζί της να... κάτσει στα αυγά της.» | Toxic | Sexism | S4 |
| 4 | «Η μανταμίτσα δεν ξέρει ούτε το πλέιμπλακ είμαι έξαλλη» | Toxic | Sexism | S4 |

## **ΚΑΤΗΓΟΡΙΑ S5**

**Σεξισμός μέσα από άρνηση, γλώσσα που αρνείται την ύπαρξη διακρίσεων , δικαιολογεί την ανισότητα των φύλων και υποτιμά εν γένει τα έμφυλα ζητήματα.**

Η κατηγορία **S5** περιλαμβάνει σχόλια που υποτιμούν ή αρνούνται τη σημασία των έμφυλων ζητημάτων και των διακρίσεων που βιώνουν οι γυναίκες. Αυτά τα σχόλια συχνά προσπαθούν να δικαιολογήσουν την ανισότητα των φύλων, αμφισβητώντας την ύπαρξη ή τη σοβαρότητα της συστημικής καταπίεσης, διάκρισης ή κακοποίησης. Η γλώσσα αυτής της κατηγορίας μπορεί να είναι φαινομενικά ουδέτερη, αλλά στην πραγματικότητα υποβαθμίζει τα έμφυλα προβλήματα, ενισχύοντας τον σεξισμό μέσω άρνησης.

***«Και γιατί γυναικοκτονία και όχι ανδροκτονία;»***

→ Άρνηση της έμφυλης βίας.

***«Δεν υπάρχει πια ανισότητα. Οι γυναίκες έχουν περισσότερα δικαιώματα από τους άντρες.»***

→ Υποτίμηση της συστημικής διάκρισης.

## *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S5*

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | «Τι σημαινει #γυναικοκτονια Υπάρχει και #ανδροκτονια Μετά πάμε και στα υπόλοιπα νέα φύλα LGTBQ κλπ κτονια» | Toxic | Sexism | S5 |
| 2 | «ο φεμινισμός είναι σαν τη μπρόκολο, αχρίαστος» | Toxic | Sexism | S5 |
| 3 | «οι γυναίκες είναι κομπλεξικές και βλέπουν σεξισμό παντου στο ισλαμ τι να πουν» | Toxic | Sexism | S5 |
| 4 | «Οι Ελληνες αντρες ειναι μια χαρα. να εκπαιδευσετε τις φεμινιστριες κορες σας να μην κανουν ψευδεις καταγγελιες κ να καταστρεφουν υποληψεις υπεργενικευοντας κιολας στο συνολο του αντρικου ελληνικου πληθυσμου. και ξεκιναμε απο το #metoo_στην_ψειρου» | Toxic | Sexism | S5 |

## ΚΑΤΗΓΟΡΙΑ S6

### Ειρωνεία, Χιούμορ που Υποβόσκει Σεξισμό

Η κατηγορία **S6** περιλαμβάνει σχόλια που εκφράζουν σεξιστικές απόψεις με έμμεσο τρόπο, συχνά χρησιμοποιώντας ειρωνεία, «χιούμορ» ή σαρκασμό. Σε αυτά τα σχόλια, ο σεξισμός μπορεί να μην είναι εμφανής ή άμεσα επιθετικός, αλλά διαφαίνεται μέσα από το πλαίσιο, τις υπονοούμενες απόψεις ή το «χιουμοριστικό» ύφος. Η γλώσσα αυτών των σχολίων ενισχύει έμφυλα στερεότυπα και ανισότητες, προσπαθώντας συχνά να αποφύγει τον άμεσο χαρακτηρισμό ως σεξιστική.

***«Τι είπε η ξανθιά όταν είδε τον πίνακα του Μονέ; 'Ω, τι ωραία χρώματα!'.»***

→«χιούμορ» που υπονοεί ότι οι ξανθιές γυναίκες δεν έχουν την ικανότητα να εκτιμήσουν την τέχνη και την κουλτούρα με τον ίδιο τρόπο που θα έκανε κάποιος άλλος.

***«Γιατί οι γυναίκες πάντα έχουν τις καλύτερες ιδέες; Ποιος ξέρει, γιατί τις ακούει κανείς;»***

→ έμμεσο σεξιστικό στοιχείο, αφού αναπαράγει την ιδέα ότι οι γυναίκες μιλάνε και δεν τις ακούει κανείς.

### *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S6*

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | «Όταν μπαίνει ωραία γκόμενα στη θάλασσα σηκώνονται και τα... κύματα» | Toxic | Sexism | S6 |
| 2 | «Ο εξάχρονος ανιψιός σου τη πέφτει σε όλα τα κορίτσια ακόμα και στα μεγαλύτερα στο σχολείο.Β1759<br>- Το ξέρω τον πήγαμε στο γιατρό.<br>- Και τι σας είπε;<br>- Ότι πάσχει απο πρόωρη γύπανση.» | Toxic | Sexism | S6 |
| 3 | «ζητιάνος πλησιάζει ξανθιά κυρία στο Κολωνάκι και λέει Δεν έχω φάει τίποτα τέσσερις ημέρες Κυρία μακάρι να είχα τη δύναμη να το κάνω και εγώ» | Toxic | Sexism | S6 |
| 4 | «Γυναίκα με πλαστική στήθους: φο βιζού.» | Toxic | Sexism | S6 |

## ΚΑΤΗΓΟΡΙΑ S7

### Θετικός σεξισμός

Η κατηγορία **S7** περιλαμβάνει σχόλια που φαίνονται καλοπροαίρετα ή θετικά, αλλά στην πραγματικότητα ενισχύουν έμφυλα στερεότυπα και περιορίζουν τις γυναίκες σε παραδοσιακούς ρόλους ή χαρακτηριστικά. Ενώ αυτά τα σχόλια μπορεί να εκλαμβάνονται ως κομπλιμέντα, έχουν τη βάση τους σε προκαταλήψεις για το φύλο και λειτουργούν ως μια πιο λεπτή μορφή σεξισμού. Ο θετικός σεξισμός, αν και λιγότερο επιθετικός, συμβάλλει στη διαιώνιση της ανισότητας, καθώς δημιουργεί προσδοκίες για το πώς πρέπει να συμπεριφέρονται οι γυναίκες ή ποια είναι η αξία τους, στηριζόμενος σε στερεοτυπικές αντιλήψεις.

***«Οι γυναίκες είναι πιο ευαίσθητες και τρυφερές από τους άντρες.»***

→ Καλοπροαίρετη δήλωση που ενισχύει στερεότυπα.

**«Τις όμορφες γυναίκες ξέρουμε να τις προσέχουμε!»**

→ Υποτιμητικός έπαινος που συνδέει την αξία με την εμφάνιση.

## *ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΠΟ ΤΟ DATASET S7*

| | COMMENT | POLARITY | SEXISM | CATEGORY |
|---|---|---|---|---|
| 1 | ***«Δεν παύουν οι γυναίκες να με εκπλήσσουν ποτέ! Τι πλάσματα!»*** | **Toxic** | **Sexism** | **S7** |
| 2 | ***«Οι εμφανίσεις των "καρυάτιδων" του Σύριζα, συγκεκριμένα της κυρίας Νοτοπούλου και της κυρίας Αχτσιόγλου απέτρεψαν με την κομψότητά τους τη μετατροπή της γιορτής της δημοκρατίας σε "γιορτάσι" "πλέριας" δημοκρατίας, όπως συμβαίνει συστηματικά τα τελευταία χρόνια. »*** | **Toxic** | **Sexism** | **S7** |
| 3 | ***«Πραγματικά ΚΥΡΙΑ η σημερινή Ούτε τσόκαρο, ούτε σαχλή, ούτε κακιά, ούτε σουργελο. Μπράβο της »*** | **Toxic** | **Sexism** | **S7** |
| 4 | ***«από τη γυναίκα πηγάζουν τα καλύτερα, τα ευγενέστερα»*** | **Toxic** | **Sexism** | **S7** |

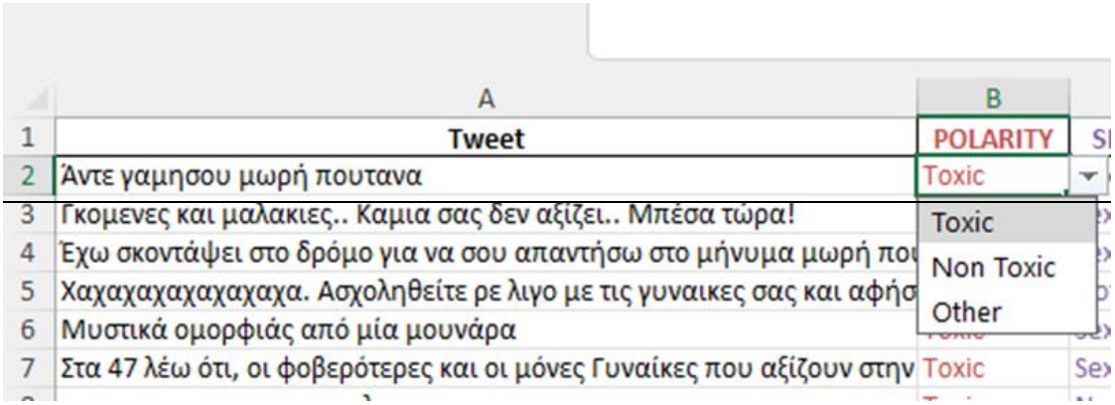
## ΕΠΙΣΗΜΕΊΩΣΗ ΣΤΟ EXCEL

Τα δεδομένα περάστηκαν στο excel για να πραγματοποιηθεί η επισημείωση και έχουν την εξής μορφή με 6 στήλες. Στην πρώτη βρίσκονται τα tweets που ο επισημειωτής θα κληθεί να επισημειώσει :

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Tweet | POLARITY | SEXISM | CATEGORY | COMMENTS | TYPE |
| 2 | Άντε γαμησου μωρή πουτανα | Toxic | Sexism | S1 | | Direct |
| 3 | Γκομενες και μαλακιες.. Καμια σας δεν αξίζει.. Μπέσα τώρα! | Toxic | Sexism | S1 | | Direct |
| 4 | Έχω σκοντάψει στο δρόμο για να σου απαντήσω στο μήνυμα μωρή ποι | Toxic | Sexism | S1 | | Direct |
| 5 | Χαχαχαχαχαχαχαχα. Ασχοληθείτε ρε λιγο με τις γυναικες σας και αφήσ | Non Toxic | Not Sexism | | | |
| 6 | Μυστικά ομορφιάς από μία μουνάρα | Toxic | Sexism | S2 | | Direct |
| 7 | Στα 47 λέω ότι, οι φοβερότερες και οι μόνες Γυναίκες που αξίζουν στην | Toxic | Sexism | S7 | | Indirect |
| 8 | γυναικες μπατσοι ηταν μλκ | Toxic | Not Sexism | | | |
| 9 | Η νδ καταργεί εργασιακά δικαιώματα και ο βόθρος του ΠΘ ανεβάζει αι | Toxic | Not Sexism | | | |
| 10 | το πιάτο σου δεν είχε ένταση-το είχα στο mute | Non Toxic | Not Sexism | | | |
| 11 | Καλημέρα στα κορίτσια τα όμορφα αν δείτε καμμία με λιλά μαλλί αγκα | Toxic | Sexism | S6 | | Indirect |
| 12 | Ντροπη Μωρη πατσαβουρα. | Toxic | Sexism | S1 | | Direct |
| 13 | Ώρα να γράψω κάτι σοβαρό φίλοι μου. ΜΗΝ ανεχεστε από κανέναν Μ | Toxic | Not Sexism | | | |
| 14 | Οι καλύτεροι σύμμαχοι ενάντια στο κρυολόγημα, κρύβονται στην κουζί | Non Toxic | Not Sexism | | | |
| 15 | Μελέτη για την ανάπτυξη της Κοιλάδας των Τεμπών έγινε; | Non Toxic | Not Sexism | | | |
| 16 | Ρε παπάρα θα γίνεις άνθρωπος; Και πούτσα και φυλακή σε κάθε Γεωρ | Toxic | Not Sexism | | | |
| 17 | Στις 28.5.1952 Οι γυναίκες αποκτούν δικαίωμα εκλέγειν και εκλέγεσθα | Non Toxic | Not Sexism | | | |
| 18 | Το θέλω να φάω για σενα λιμιτ είναι καλή ατάκα για πέσιμο | Non Toxic | Not Sexism | | | |
| 19 | Η μαμά της Στέλλας αύριο το πρωί τηλεφωνική στην Κατερίνα Καινούρ | Non Toxic | Not Sexism | | | |
| 20 | Μαγειρεύω στη κουζίνα και με το που ειμαι έτοιμος να κάνω το πέταχτ | Non Toxic | Not Sexism | | | |
| 21 | Οι γριές οι πουτάνες έχουν το ζουμί | Toxic | Sexism | S1 | | Direct |

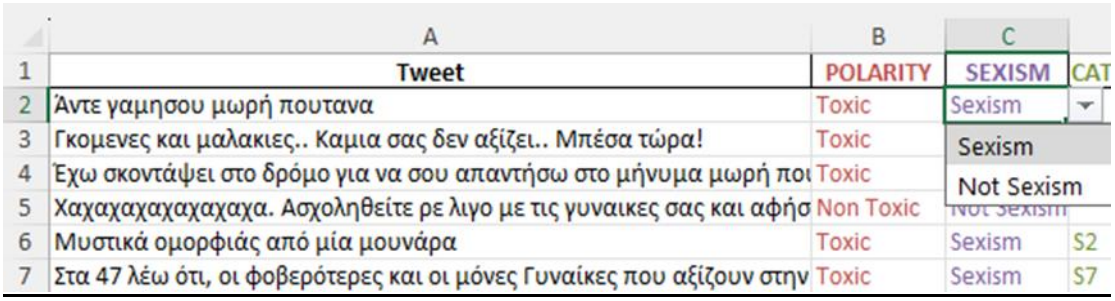
## ΒΗΜΑ 1ο, Polarity

Αρχικά ο annotator διαβάζει το tweet και επιλέγει αν το σχόλιο είναι τοξικό, μη τοξικό ή ανήκει στην κατηγορία «άλλο». Οι επιλογές είναι κλειστές και προκαθορισμένες για ευκολία και επιτάχυνση της διαδικασίας.



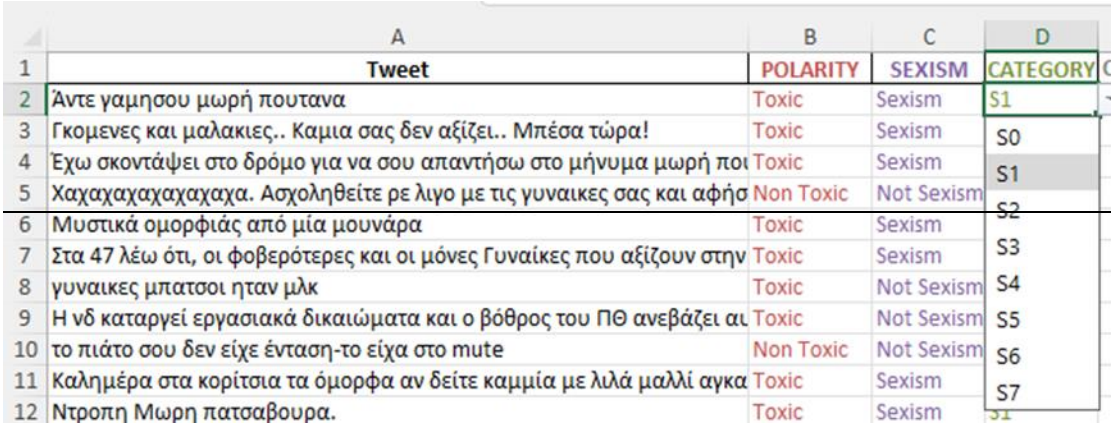

## ΒΗΜΑ 2ο, Sexism

Στη συνέχεια ο annotator θα κρίνει εάν το σχόλιο είναι σεξιστικό. Αυτό αφορά μόνο τα σχόλιο που έχουν polarity **toxic**. Όσα επισημειώθηκαν ως other στο προηγούμενο βήμα δεν μελετώνται στα επόμενα βήματα, αφού θα αφαιρεθούν στο τέλος από το dataset και όσα έχουν επισημειωθεί ως non toxic δεν είναι σεξιστικά.



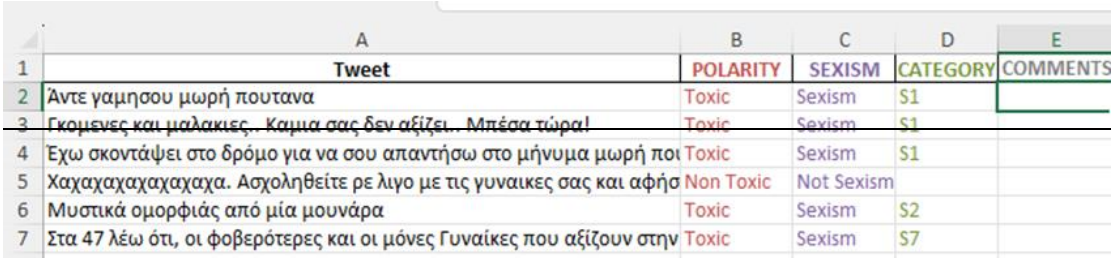

## ΒΗΜΑ 3ο, Sexism Category

Εφόσον το σχόλιο έχει κριθεί σεξιστικό, τότε ο επισημειώτης προχωράει στην κατηγοριοποίηση (S0-S7) του σύμφωνα με τις οδηγίες. Αν το σχόλιο έχει καταχωρηθεί ως non sexist δεν συμπληρώνεται τίποτα σε αυτό το πλαίσιο και πηγαίνουμε στο επόμενο.



## ΒΗΜΑ 4ο, Comments

Τα σχόλια είναι προαιρετικά, ο annotator μπορεί εδώ να γράψει ελευθέρα οτιδήποτε είναι προς συζήτηση με τους άλλους annotators.

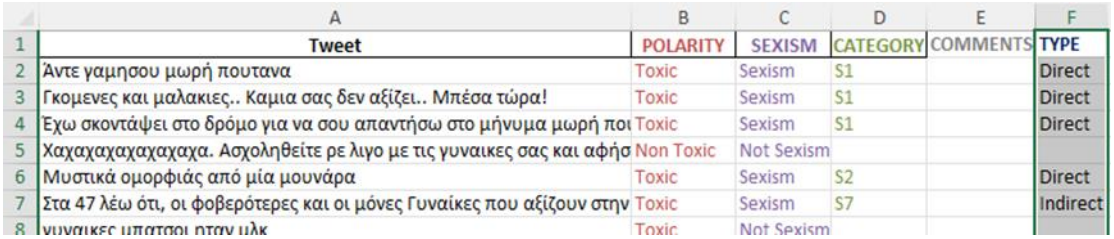


| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Tweet | POLARITY | SEXISM | CATEGORY | COMMENTS |
| 2 | Άντε γαμησου μωρή πουτανα | Toxic | Sexism | S1 | |
| 3 | Γκομενες και μαλακιες.. Καμια σας δεν αξίζει.. Μπέσα τώρα! | Toxic | Sexism | S1 | |
| 4 | Έχω σκοντάψει στο δρόμο για να σου απαντήσω στο μήνυμα μωρή ποι | Toxic | Sexism | S1 | |
| 5 | Χαχαχαχαχαχαχαχα. Ασχοληθείτε ρε λιγο με τις γυναικες σας και αφήσ | Non Toxic | Not Sexism | | |
| 6 | Μυστικά ομορφιάς από μία μουνάρα | Toxic | Sexism | S2 | |
| 7 | Στα 47 λέω ότι, οι φοβερότερες και οι μόνες Γυναίκες που αξίζουν στην | Toxic | Sexism | S7 | |

## ΒΗΜΑ 5ο, Sexism Type

Το βήμα αυτό **δεν αφορά** και **δεν συμπληρώνεται από τον annotator.** Συμπληρώνεται αυτόματα μέσω μιας συνάρτησης που έχει εισαχθεί στο excel.

Τα σχόλια που ανήκουν στις κατηγορίες S1,S1,S3 επισημειώνονται αυτόματα ως άμεσος σεξισμός (Direct) ενώ τα σχόλια που ανήκουν στις κατηγορίες S4,S5,S6, S7 επισημειώνονται αυτόματα ως έμμεσος σεξισμός (Indirect).



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Tweet | POLARITY | SEXISM | CATEGORY | COMMENTS | TYPE |
| 2 | Άντε γαμησου μωρή πουτανα | Toxic | Sexism | S1 | | Direct |
| 3 | Γκομενες και μαλακιες.. Καμια σας δεν αξίζει.. Μπέσα τώρα! | Toxic | Sexism | S1 | | Direct |
| 4 | Έχω σκοντάψει στο δρόμο για να σου απαντήσω στο μήνυμα μωρή ποι | Toxic | Sexism | S1 | | Direct |
| 5 | Χαχαχαχαχαχαχαχα. Ασχοληθείτε ρε λιγο με τις γυναικες σας και αφήσ | Non Toxic | Not Sexism | | | |
| 6 | Μυστικά ομορφιάς από μία μουνάρα | Toxic | Sexism | S2 | | Direct |
| 7 | Στα 47 λέω ότι, οι φοβερότερες και οι μόνες Γυναίκες που αξίζουν στην | Toxic | Sexism | S7 | | Indirect |
| 8 | γυναικες μπατσοι οταν μλκ | Toxic | Not Sexism | | | |

Ο διαχωρισμός ανάμεσα στον **άμεσο** και τον **έμμεσο σεξισμό** βασίζεται στην ένταση και τη σαφήνεια της σεξιστικής γλώσσας ή συμπεριφοράς. Ο **άμεσος σεξισμός** περιλαμβάνει φανερά επιθετικές ή προσβλητικές εκφράσεις που υποκινούν τη βία, την κακοποίηση ή τη σεξουαλική υποβάθμιση των γυναικών, όπως οι απειλές (S0), οι βρισιές (S1), ή η αντικειμενοποίηση και σεξουαλικοποίηση των γυναικών (S2). Αυτά τα σχόλια είναι ανοιχτά και εκφράζουν άμεσα αρνητικά συναισθήματα ή συμπεριφορές απέναντι στις γυναίκες. Από την άλλη, ο **έμμεσος σεξισμός** περιλαμβάνει σχόλια ή ενέργειες που, ενώ δεν είναι τόσο ανοιχτά επιθετικά, αναπαράγουν στερεότυπα, υποτιμούν ή δικαιολογούν τις ανισότητες φύλου με πιο συγκαλυμμένο τρόπο. Αυτά τα σχόλια μπορεί να περιλαμβάνουν στερεότυπα (S3), υποτιμητικά ή πατροναριστικά σχόλια (S4), σχόλια που αρνούνται την ύπαρξη φυλετικών διακρίσεων (S5) ή ακόμα και ειρωνεία και χιούμορ που υποβόσκει σεξισμό (S6). Ο θετικός σεξισμός (S7), αν και εκφράζει θετικά σχόλια για τις γυναίκες, μπορεί να περιορίζει τις γυναίκες σε παραδοσιακούς ρόλους, υποστηρίζοντας έναν σεξιστικό τρόπο σκέψης. Ο έμμεσος σεξισμός είναι πιο δύσκολο να αναγνωριστεί, αλλά παραμένει εξίσου βλαβερός με τον άμεσο σεξισμό στην ενίσχυση των ανισοτήτων φύλου.

## DECISION RULES

Ο επισημειωτής πρέπει να ακολουθεί σαφείς κατευθυντήριες γραμμές και να είναι προσεκτικός στις αποφάσεις του κατά την επισημείωση των σχολίων. Ακολουθούν γενικές οδηγίες συνοδευμένες από παραδείγματα για τις πιο περίπλοκες περιπτώσεις.

Οι παραπάνω κατηγορίες σεξισμού (0-7) είναι διαβαθμισμένες 0 το περισσότερο σοβαρό, άμεσο και φανερό και 7 το λιγότερο σοβαρό. Όταν ο annotator είναι ανάμεσα σε 2 ή και περισσότερες κατηγορίες αρχικά προτείνουμε να προσπαθήσει αναγνωρίσει το κύριο μήνυμα του σχολίου (πχ προσβολή, αστείο, κομπλιμέντο). Αυτό βοηθάει σε ένα πρώτο ξεκαθάρισμα. Ο διαχωρισμός σε άμεσο ή έμμεσο σεξισμό επίσης βοηθάει στην επιλογή, καθώς όταν υπάρχει δίλημμα για 2 κατηγορίες που η μία ανήκει στον άμεσο και η άλλη στον έμμεσο σεξισμό, η σκέψη αν το σχόλιο αποτελεί άμεσο ή έμμεσο σεξισμό βοηθάει.

Αν παρόλα αυτά 2 κατηγορίες βρίσκονται πολύ κοντά υπάρχει η γενική γραμμή να επιλεγεί ως label η πιο σοβαρή, σύμφωνα με την παραπάνω κατάταξη. Αν δηλαδή είμαι ανάμεσα σε 0 και 1 θα βάλω το 0. Δεν θέλουμε όμως υπερανάλυση από τους επισημειωτές καθώς αυτό που χρειαζόμαστε είναι κυρίως η διαίσθηση του φυσικού ομιλητή όποτε και κάθε απόφαση μπορεί να είναι διαφορετική ανάλογα με το άτομο. Μας ενδιαφέρει πολύ η συνέπειά στις απαντήσεις, αν επιλέξω ότι ένα σχόλιο συγκεκριμένου τύπου ανήκει στην χ κατηγορία, θα φροντίσω να το τηρήσω για όλο το σύνολο των δεδομένων.

**Συχνά διλήμματα και τρόπος επίλυσης:**

**S2 ή s7**: Πχ *«Μωρή Σμαράγδα Καρύδη πόσο μουνάρα είσαι! »*

Πολλά σχόλια παρουσιάζονται ως θετικά, δεν έχουν σκοπό να προσβάλλουν ωστόσο η συγκεκριμένη επιλογή λέξεων καταλήγει σε σεξουαλικοποίηση ή αντικειμενοποίηση της γυναίκας.

**S2:**

Το σχόλιο δεν έχει καλοπροαίρετο ή εξιδανικευτικό τόνο. Αντίθετα, είναι ωμό σεξουαλικό και αντικειμενοποιεί τη γυναίκα. Η χρήση της λέξης «μουνάρα» προσδίδει έναν χυδαίο χαρακτήρα που συνδέεται με τη σεξουαλικοποίηση και όχι με "θετικά" στερεότυπα. Αν και το σχόλιο φαίνεται «θετικό» με την έννοια του θαυμασμού, δεν έχει καλοπροαίρετο ή εξιδανικευτικό τόνο. Επικεντρώνεται ξεκάθαρα στη σεξουαλικότητα και τη φυσική εμφάνιση, χωρίς καμία αναφορά σε προσωπικά χαρακτηριστικά ή αξίες.

**S2 ή s7** Πχ «*"ΚΟΛΑΣΗ" ΤΟ ΚΟΡΙΤΣΙ!!!!! Η Candice Swanepoel είναι η πιο σέξι γυναίκα της χρονιάς»*

**S2:**

Αντίστοιχα, αν και χρησιμοποιείται θετικός τόνος και επικεντρώνεται στη σωματική εμφάνιση της γυναίκας, η χρήση της λέξης "ΚΟΛΑΣΗ" και η φράση "η πιο σέξι γυναίκα" επικεντρώνονται στο σεξουαλικό της χαρακτηριστικό, μειώνοντάς την σε αντικείμενο σεξουαλικής έλξης. Αυτό είναι ένα παράδειγμα αντικειμενοποίησης και σεξουαλικοποίησης (S2), αφού η γυναίκα παρουσιάζεται μόνο μέσω της φυσικής της εμφάνισης χωρίς να λαμβάνονται υπόψη οι άλλες πτυχές της προσωπικότητάς της. Αν και μπορεί να μην υπάρχει επιθετική γλώσσα, η αντικειμενοποίηση της γυναίκας σε αυτό το πλαίσιο είναι σαφής και ενισχύει τα στερεότυπα γύρω από την εμφάνιση και τον ρόλο των γυναικών.

**S1 ή S4** πχ «*Είσαι χαζή και άχρηστη*».

**S1:**

Το παραπάνω σχόλιο θα μπορούσε να θεωρηθεί S1 αλλά και S4 καθώς οι συγκεκριμένες λέξεις που έχουν επιλεγεί δεν έχουν πολύ μεγάλο αρνητικό σημασιολογικό βάρος. Ωστόσο, στόχος του μηνύματος είναι ξεκάθαρα η προσβολή και δεν αποτελεί έμμεση εκδήλωση σεξισμού.

**S6 ή S4** πχ «*Μπράβο, κοριτσάκι, πήρες δίπλωμα οδήγησης, ε; »*

**S4:**

Το σχόλιο αυτό ανήκει στην κατηγορία S4 γιατί έχουμε έμμεση υποτίμηση, μέσω σχολίων που μοιάζουν αθώα αλλά είναι προσβλητικά.

**S2 ή S4** πχ «*Επίσης... Εσείς οι γκόμενες! ΜΗΝ ΚΑΝΕΤΕ ΚΑΘΕ ΜΕΡΑ ΠΟΔΙΑ!»*

**S4:**

Ο όρος **"γκόμενες"** είναι υποτιμητικός, καθώς χρησιμοποιείται για να μειώσει τις γυναίκες και να τις αποκαλέσει με ένα μειωτικό όρο που φέρνει σε πρώτο πλάνο μόνο το φύλο τους και όχι την προσωπικότητα ή την αξία τους. Το σχόλιο αυτό θα μπορούσε να θεωρηθεί και αντικειμενοποίηση και πολλά σχόλια με τη λέξη «γκόμενα» ανήκουν όντως στην κατηγορία S2. Ωστόσο, στο συγκεκριμένο παράδειγμα δεν συνοδεύεται με κάποια άλλη λέξη που να ενισχύει τη σεξουαλικοποίηση ή την αντικειμενοποίηση, ενώ είναι γνωστό ότι η συγκεκριμένη λέξη δεν έχει μεγάλο αρνητικό σημασιολογικό βάρος, τουλάχιστον μόνη της, καθώς χρησιμοποιείται συχνά κυρίως στη γλώσσα των social media. Παράλληλα, το ύφος τους μηνύματος μπορεί να θεωρηθεί πατροναριστικό. Επομένως, η υποτίμηση της γυναίκας μέσω της γλώσσας και της χρήσης του όρου «γκόμενες» μπορεί να το κατατάξει στην κατηγορία **S4**, καθώς είναι υποτιμητικό και μειωτικό.

**S0 ή S6** πχ *«Τι μαλάκας ο πρίγκηπας στο παραμύθι της Χιονάτης... Εκεί που μπορούσε να έχει σεξ χωρίς γκρίνια, την ζωντάνεψε φιλώντας την στο στόμα. »*

**S0:**

Αν και χρησιμοποιείται μια μορφή χιούμορ, το παραπάνω σχόλιο θα κατηγοριοποιηθεί ως S0, καθώς πρόκειται για σχόλιο που κανονικοποιεί τη σεξουαλική παρενόχληση και προωθεί την κουλτούρα βιασμού. Λόγω αφενός σοβαρότητας και αφετέρου του παραπάνω κανόνα για διαβάθμιση σεξιστικών σχολίων θα επισημειωθεί ως S0.

S1 ή s2 Πχ *«Παντως κοριτσια το να ντυνεσαι σα φτηνη απομιμηση φτηνης πορνοσταρ ειναι ταλεντο απο μονο του.»*

**S2:**

Το σχόλιο προσβάλλει τις γυναίκες μέσω της χρήσης του όρου "φτηνή απόμιμηση φτηνής πορνοστάρ", αναφερόμενο στην εμφάνιση τους με υποτιμητικό τρόπο. Χρησιμοποιεί μια γλώσσα που υποβαθμίζει τη γυναίκα σε ένα σεξουαλικό αντικείμενο, βασισμένο στην εμφάνιση και τον τρόπο ντυσίματος. Η φράση αποσκοπεί να υπονοήσει ότι η γυναίκα είναι απλώς ένα σεξουαλικό αντικείμενο, χωρίς να λαμβάνει υπόψη την προσωπικότητα ή τις άλλες της ικανότητες. Το σχόλιο αυτό θα μπορούσε να θεωρηθεί και s1, καθώς περιλαμβάνει προσβλητική γλώσσα αλλά το στοιχείο της αντικειμενικοποίησης είναι πιο έντονο.

**S4 ή S3** πχ *«Τον έχει εύκολη λύση για κάποιο λόγο. Και στις αρχές τον ψήφιζε και έλεγε δεν κινδυνεύει... Μωρη τωρα είμαστε στους 10 δεν καις ψήφο μπηκε το ξανθο στον εγκεφαλο σου»*

**S3:**

Η χρήση της λέξης **"**Μωρή**"** είναι σαφώς υποτιμητική, καθώς χρησιμοποιείται με σκοπό να μειώσει το άτομο στο οποίο αναφέρεται. Ωστόσο, το στερεότυπο που περιλαμβάνει αυτή η πρόταση «*μπηκε το ξανθο στον εγκεφαλο σου»,* είναι πιο έντονο και για αυτό επιλέγεται η κατηγορία s3.

**S1 ή S3** *Πχ «ΣΚΑΣΕ ΜΩΡΗ ΚΟΤΑ ΓΑΜΩ ΤΟ ΣΠΙΤΙ ΣΟΥ. ΠΑΝΕ ΒΑΨΕ ΚΑΝΑ ΝΥΧΙ»*

**S1:**

Αν και η πρόταση περιέχει και υβριστικό λόγο και στερεότυπο, θα επιλέξουμε να την κατατάξουμε στην κατηγορία s1, καθώς είναι πιο κυρίαρχη.

**S1 ή S3** πχ «Οι Πουτανες...κερδιζουν την ΠΡΟΣΟΧΗ! Οι Σωστες κοπελες...τον ΣΕΒΑΣΜΟ!! #ainte_kouklitsa_mou»

**S3:**

Η πρόταση περιλαμβάνει και υβριστικό λεξιλόγιο και στερεότυπο που αφορά τις γυναίκες. Ωστόσο επειδή η προσβολή είναι πιο γενική και δεν στοχεύει συγκεκριμένα σε ένα πρόσωπο, αλλά χρησιμοποιείται για να ενισχύσει το στερεότυπο, θα επισημειωθεί ως s3.

**Επιπλέον επισημάνσεις**:

Σύμφωνα με τις γενικές οδηγίες που δόθηκαν παραπάνω, επισημειώνω ως σεξιστικό μόνο σχόλια που αναφέρονται σε γυναίκες, καθώς το αντικείμενο της παρούσας έρευνας είναι ο σεξισμός που αφορά γυναίκες. Έτσι, σεξιστικά σχόλια που απευθύνονται σε άνδρες, επισημειώνονται ως non sexist.

Ωστόσο, πολλές φορές ένα σχόλιο μπορεί να είναι σεξιστικό εμμέσως προς γυναίκες, παρότι απευθύνεται σε άνδρες.

Π.χ.

*1. «Αααα καλέ μεγάλη **κουτσομπολα** ο Κωστης.. ούτε μισή ώρα δ μπόρεσε να κρατήσει την πληροφορία για την πάρτη του .. και πως την είδε έτσι συμβουλατορας με όλους κ με ολα..#exapsi»*

2. «Αυτο το σούργελο ολκής ο Ντάφυ δακρύζει που τον θελουν να μπει στο #SurvivorGR και μας κάνει στορι τα μηνύματα. Ναι Τριαντάφυλλε μας έπεισες για το πόσο **ανισόρροπο attention whore** είσαι ΔΕΝ ΘΕΛΟΥΜΕ ΑΛΛΟ»

*3. «εγώ θα του έβγαζα τα εντερα θα τον τυλιγα κ θα τον εψηνα κοκορέτσι κ.ζαιο ...κάτι σαν κ εσένα κ τον τσιπρα έχουν καταντήσει την κοινωνία παράδεισο στα αποβρασματα*

*στους τρομοκράτες στους φονιάδες και σε κάθε είδος λαμογιο δυστυχώς...**μπράβο στη κοπελιά έχει καρύδια indeed»***

Και στις 2 πρώτες περιπτώσεις ενώ σκοπός του σχολίου είναι η υποτίμηση του συνομιλητή, ο οποίος δεν είναι γυναίκα, χρησιμοποιούνται λέξεις και εκφράσεις που χαρακτηρίζουν μία γυναίκα. Στην πρώτη περίπτωση επιλέγεται υποτιμητικά η λέξη κουτσομπόλα, όχι κουτσομπόλης, ενώ απευθύνεται σε άνδρα. Το σχόλιο αναπαράγει το στερεότυπο ότι οι γυναίκες είναι κουτσομπόλες και χρησιμοποιεί αυτό το χαρακτηριστικό για να προσβάλει τον άνδρα. Αυτό το σχόλιο ενισχύει το στερεότυπο ότι το κουτσομπολιό είναι κυρίως χαρακτηριστικό των γυναικών και το μεταφέρει με αρνητική χροιά στον άνδρα. Η χρήση του όρου "κουτσομπολα" και η σύνδεση της συμπεριφοράς του άνδρα με αυτό το χαρακτηριστικό γυναικών το καθιστά σαφώς ένα σχόλιο που ανήκει στην κατηγορία S3: Στερεότυπα.

Στη δεύτερη περίπτωση ενώ και πάλι το σχόλιο αναφέρεται σε έναν άνδρα, χρησιμοποιείται ο όρος "attention whore", ο οποίος συνήθως χαρακτηρίζει γυναίκες που επιδιώκουν υπερβολική προσοχή, συνήθως σε σχέση με την εμφάνισή τους ή τη συμπεριφορά τους. Αν και το σχόλιο αναφέρεται σε έναν άνδρα, χρησιμοποιεί έναν όρο που συνδέεται με στερεότυπα για τις γυναίκες, ιδιαίτερα για την ανάγκη τους για προσοχή. Αυτό το σχόλιο ενισχύει την αντίληψη ότι το "attention whore" είναι χαρακτηριστικό των γυναικών, κάτι που καθιστά το σχόλιο σεξιστικό μέσω του στερεοτύπου που αναπαράγει.

Στην 3η περίπτωση η φράση "*μπράβο στη κοπελιά έχει καρύδια indeed*" χρησιμοποιεί μια έκφραση που παραδοσιακά συνδέεται με ανδρική δύναμη και θάρρος. Η αναφορά στο να έχει κάποιος "καρύδια" είναι ένα στερεότυπο που συνήθως αποδίδεται στους άνδρες, καθώς η φράση αυτή χρησιμοποιείται για να δηλώσει θάρρος, γενναιότητα ή δύναμη, χαρακτηριστικά που παραδοσιακά θεωρούνται ανδρικά. Η χρήση της φράσης "μπράβο στη κοπελιά" και η ένσταση ότι "έχει καρύδια" είναι μια παραδοξότητα, γιατί ενώ αναφέρεται σε μια γυναίκα, της αποδίδεται ένα χαρακτηριστικό που συνήθως θεωρείται αρσενικό. Αυτή η αντίφαση αναπαράγει το στερεότυπο ότι η γενναιότητα ή η δύναμη είναι χαρακτηριστικά που ανήκουν περισσότερο στους άνδρες παρά στις γυναίκες. Αυτό το σχόλιο ενισχύει το στερεότυπο ότι οι γυναίκες δεν αναμένεται να έχουν τα ίδια χαρακτηριστικά με τους άνδρες, όπως η «γενναιότητα» ή η «δύναμη», και συνεπώς, όταν μια γυναίκα εκδηλώνει αυτές τις ιδιότητες, η κοινωνία την εξυμνεί για το γεγονός ότι "ξεφεύγει" από τα συνηθισμένα γυναικεία στερεότυπα. Επομένως, και αυτό το σχόλιο είναι σεξιστικό μέσω των στερεοτύπων, καθώς εστιάζει στην αντίθεση των φύλων και ενισχύει την αντίληψη ότι κάποια χαρακτηριστικά, όπως η γενναιότητα ή η δύναμη, ανήκουν περισσότερο στους άνδρες, ενώ οι γυναίκες που τα εκδηλώνουν φαίνονται ασυνήθιστες. Το σχόλιο αυτό ανήκει στην κατηγορία S3: Στερεότυπα.

# APPENDIX III

## Final Prompts used in Context Learning Experiments:

| PROMPTS | |
|---|---|
| 0 shot | {<br>"role": "system",<br>"content": (<br>"*Είσαι ένας ταξινομητής και ρόλος σου είναι να κρίνεις αν ένα tweet είναι σεξιστικό προς τις γυναίκες ή όχι. "*<br>*"Ένα tweet είναι σεξιστικό μόνο όταν μειώνει, προσβάλλει ή υποτιμά τις γυναίκες λόγω φύλου*. "<br>"Απάντα μόνο με μία από τις εξής φράσεις:\n"<br>"- Ναι, το tweet είναι σεξιστικό.\n"<br>"- Όχι, το tweet δεν είναι σεξιστικό."<br>)<br>}, |
| 2 shots | {"role": "user", "content": 'Το tweet είναι: "**Ούτε το αφεντικό να ήταν έλεος μωρη κότα!!!**"\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Όχι, το tweet δεν είναι σεξιστικό."},<br>{"role": "user", "content": 'Το tweet είναι: "**οι γυναίκες είναι πολύ καλές όταν ξέρουν τι θέλουν ή δεν ξέρουν ?θα σας γελάσω…Λες και ξέρουν ποτέ..**"\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Ναι, το tweet είναι σεξιστικό."}, |

| | |
|---|---|
| 4 shots | {"role": "user", "content": 'Το tweet είναι: "**Η στιγμή εκείνη που θέλεις να τα κάνεις πουτάνα όλα κ κάθεσαι τελικά στα αυγά σου..!!**"\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Όχι, το tweet δεν είναι σεξιστικό."},<br>{"role": "user", "content": 'Το tweet είναι: **"Η κόρη της «γυναίκας αράχνης» είναι η 17χρονη φίλη της που πήγαν μαζί στο πάρτι**"\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Ναι, το tweet είναι σεξιστικό."}, |
| 6 shots | {"role": "user", "content": 'Το tweet είναι: "**Μωρή σαφρακιασμενη ομοφοβικια κότα, με τι θα ήσουν ευχαριστημένος με το να τους θάψουν 2μετρα κάτω από τη γη?**"\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Όχι, το tweet δεν είναι σεξιστικό."},<br>{"role": "user", "content": 'Το tweet είναι: "**Το σωστό θα ήτανε [WOMAN NAME ] και [WOMAN NAME]την παράδοση-παραλαβή να την κάνανε με μαγιό για να συγκρίνουμε τα υπουργικά προσόντα.**"\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Ναι, το tweet είναι σεξιστικό."}, |

| | |
|---|---|
| 8 shots | {"role": "user", "content": 'Το tweet είναι: "**Τι λες βρε σουργελο της κοινωνίας!!!** "\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Όχι, το tweet δεν είναι σεξιστικό."},<br><br>{"role": "user", "content": 'Το tweet είναι: "**ξανθιά αλλά ξέρει τι είναι οφσάιντ** "\nΕίναι σεξιστικό;'},<br>{"role": "assistant", "content": "Ναι, το tweet είναι σεξιστικό."} |

# APPENDIX IV

## Results on Val Set for BERT fine-tuned with Class weights and partial fine-tuned BERT:

| Val Set | Class 0 | Class 1 | Macro average of 0, 1 | Accuracy |
|---|---|---|---|---|
| Fine-tuning GreekBERT with class weights | 0.912/0.843/0.876 | 0.561/0.712/0.627 | 0.736/0.778/0.752 | 0.815 |
| Partial Fine-tuning GreekBERT | 0.779/0.669/0.874 | 0.571/0.017/0.034 | 0.675/0.507/0.454 | 0.778 |

# REFERENCES

[1] A. Azadi, B. Ansari, S. Zamani, and S. Eetemadi, “Bilingual Sexism Classification: Fine-Tuned XLM-RoBERTa and GPT-3.5 Few-Shot Learning,” Jan. 05, 2025, *arXiv*: arXiv:2406.07287. doi: 10.48550/arXiv.2406.07287.

[2] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring Hate Speech Detection in Multimodal Publications,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA: IEEE, Mar. 2020, pp. 1459–1467. doi: 10.1109/WACV45572.2020.9093414.

[3] T. Kumarage, A. Bhattacharjee, and J. Garland, “Harnessing Artificial Intelligence to Combat Online Hate: Exploring the Challenges and Opportunities of Large Language Models in Hate Speech Detection,” Mar. 12, 2024, *arXiv*: arXiv:2403.08035. doi: 10.48550/arXiv.2403.08035.

[4] S. Gross, J. Petrak, L. Venhoff, and B. Krenn, “GermEval2024 Shared Task: GerMS-Detect – Sexism Detection in German Online News Fora”.

[5] M. Siino and I. Tinnirello, “Notebook for the EXIST Lab at CLEF 2024”.

[6] X. Luo *et al.*, “A Literature Survey on Multimodal and Multilingual Sexism Detection,” *IEEE Trans. Comput. Soc. Syst.*, pp. 1–19, 2025, doi: 10.1109/TCSS.2025.3561921.

[7] L. D. Grazia *et al.*, “MuSeD: A Multimodal Spanish Dataset for Sexism Detection in Social Media Videos,” Apr. 15, 2025, *arXiv*: arXiv:2504.11169. doi: 10.48550/arXiv.2504.11169.

[8] B. Krenn, J. Petrak, M. Kubina, and C. Burger, “GERMS-AT: A Sexism/Misogyny Dataset of Forum Comments from an Austrian Online Newspaper,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, Feb. 2024, pp. 7728–7739. Accessed: Jun. 14, 2025. [Online]. Available: https://aclanthology.org/2024.lrec-main.683/

[9] H. R. Kirk, W. Yin, B. Vidgen, and P. Röttger, “SemEval-2023 Task 10: Explainable Detection of Online Sexism,” May 08, 2023, *arXiv*: arXiv:2303.04222. doi: 10.48550/arXiv.2303.04222.

[10] L. Plaza *et al.*, “Overview of EXIST 2024 — Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, vol. 14959, L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco De Herrera, G. Faggioli, and N. Ferro, Eds., in Lecture Notes in Computer Science, vol. 14959. , Cham: Springer Nature Switzerland, 2024, pp. 93–117. doi: 10.1007/978-3-031-71908-0_5.

[11] W. Lei, N. A. S. Abdullah, and S. R. S. Aris, “A Systematic Literature Review on Automatic Sexism Detection in Social Media,” *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 6, pp. 18178–18188, Dec. 2024, doi: 10.48084/etasr.8881.

[12] M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, and N. Kumar, “A context-aware attention and graph neural network-based multimodal framework for misogyny

detection," *Inf. Process. Manag.*, vol. 62, no. 1, p. 103895, Jan. 2025, doi: 10.1016/j.ipm.2024.103895.

[13] D. Almanea and M. Poesio, "ArMIS - The Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, Mar. 2022, pp. 2282–2291. Accessed: Jun. 14, 2025. [Online]. Available: https://aclanthology.org/2022.lrec-1.244/

[14] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, "SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection," Aug. 06, 2021, *arXiv*: arXiv:2108.03070. doi: 10.48550/arXiv.2108.03070.

[15] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive Language Identification in Greek," Mar. 18, 2020, *arXiv*: arXiv:2003.07459. doi: 10.48550/arXiv.2003.07459.

[16] S. Markantonatou, V. Stamou, C. Christodoulou, G. Apostolopoulou, A. Balas, and G. Ioannakis, "The Corpus AIKIA: Using Ranking Annotation for Offensive Language Detection in Modern Greek," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 15861–15871.

[17] E. Fersini, D. Nozza, and P. Rosso, "Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)," in *EVALITA Evaluation of NLP and Speech Tools for Italian*, T. Caselli, N. Novielli, V. Patti, and P. Rosso, Eds., Torino: Accademia University Press, 2018, pp. 59–66. doi: 10.4000/books.aaccademia.4497.

[18] S. Khan, G. Pergola, and A. Jhumka, "Multilingual Sexism Identification via Fusion of Large Language Models".

[19] D. Grosz and P. Conde-Cespedes, "Automatic Detection of Sexist Statements Commonly Used at the Workplace," Jul. 08, 2020, *arXiv*: arXiv:2007.04181. doi: 10.48550/arXiv.2007.04181.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Mar. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Mar. 01, 2020, *arXiv*: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

[22] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.

[23] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," Oct. 06, 2021, *arXiv*: arXiv:2006.03654. doi: 10.48550/arXiv.2006.03654.

[24] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully, "An Annotated Corpus for Sexism Detection in French Tweets," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France: European

Language Resources Association, Feb. 2020, pp. 1397–1403. Accessed: Jun. 26, 2025. [Online]. Available: https://aclanthology.org/2020.lrec-1.175/

[25] A. Kalra and A. Zubiaga, “Sexism Identification in Tweets and Gabs using Deep Neural Networks,” Nov. 05, 2021, *arXiv*: arXiv:2111.03612. doi: 10.48550/arXiv.2111.03612.

[26] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos, “GREEK-BERT: The Greeks visiting Sesame Street,” in *11th Hellenic Conference on Artificial Intelligence*, in SETN 2020. New York, NY, USA: Association for Computing Machinery, Jun. 2020, pp. 110–117. doi: 10.1145/3411408.3411440.

[27] K. Perifanos and D. Goutsos, “Multimodal Hate Speech Detection in Greek Social Media,” *Multimodal Technol. Interact.*, vol. 5, no. 7, p. 34, Jun. 2021, doi: 10.3390/mti5070034.

[28] A. Matarazzo and R. Torlone, “A Survey on Large Language Models with some Insights on their Capabilities and Limitations,” Jan. 03, 2025, *arXiv*: arXiv:2501.04040. doi: 10.48550/arXiv.2501.04040.

[29] S. Minaee *et al.*, “Large Language Models: A Survey,” Feb. 20, 2024, *arXiv*: arXiv:2402.06196. doi: 10.48550/arXiv.2402.06196.

[30] W. X. Zhao *et al.*, “A Survey of Large Language Models,” 2023, *arXiv*. doi: 10.48550/ARXIV.2303.18223.

[31] A. Riahi Samani, T. Wang, K. Li, and F. Chen, “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*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds., Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 6230–6243. Accessed: Jun. 17, 2025. [Online]. Available: https://aclanthology.org/2025.coling-main.416/

[32] R. Pan, J. García-Díaz, and R. Valencia-García, “Comparing Fine-Tuning, Zero and Few-Shot Strategies with Large Language Models in Hate Speech Detection in English,” *Comput. Model. Eng. Sci.*, vol. 140, no. 3, pp. 2849–2868, 2024, doi: 10.32604/cmes.2024.049631.

[33] S. Sultana and M. Begum Kali, “Exploring ChatGPT for identifying sexism in the communication of software developers,” in *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, Crete Greece: ACM, Jun. 2024, pp. 400–403. doi: 10.1145/3652037.3663918.

[34] V. Basile *et al.*, “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Mar. 2019, pp. 54–63. doi: 10.18653/v1/S19-2007.

[35] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts, “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*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., Online: Association for Computational Linguistics, Dec. 2021, pp. 1336–1350. doi: 10.18653/v1/2021.eacl-main.114.

[36] T. Farrell, M. Fernandez, J. Novotny, and H. Alani, “Exploring Misogyny across the Manosphere in Reddit,” in *Proceedings of the 10th ACM Conference on Web Science*,

Boston Massachusetts USA: ACM, Jun. 2019, pp. 87–96. doi: 10.1145/3292522.3326045.

[37] H. Buie and A. Croft, “The Social Media Sexist Content (SMSC) Database: A Database of Content and Comments for Research Use,” *Collabra Psychol.*, vol. 9, no. 1, p. 71341, Mar. 2023, doi: 10.1525/collabra.71341.

[38] P. Chiril, F. Benamara, and V. Moriceau, “‘Be nice to your wife! The restaurants are closed’: Can Gender Stereotype Detection Improve Sexism Classification?,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Aug. 2021, pp. 2833–2844. doi: 10.18653/v1/2021.findings-emnlp.242.

[39] A. Jha and R. Mamidi, “When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data,” in *Proceedings of the Second Workshop on NLP and Computational Social Science*, D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O’Connor, O. Tsur, and A. S. Doğruöz, Eds., Vancouver, Canada: Association for Computational Linguistics, Dec. 2017, pp. 7–16. doi: 10.18653/v1/W17-2902.

[40] D. Roussis *et al.*, “Krikri: Advancing Open Large Language Models for Greek,” May 30, 2025, *arXiv*: arXiv:2505.13772. doi: 10.48550/arXiv.2505.13772.

[41] J. Ye *et al.*, “A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models,” Dec. 23, 2023, *arXiv*: arXiv:2303.10420. doi: 10.48550/arXiv.2303.10420.

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[43] D. Ingle, R. Tripathi, A. Kumar, K. Patel, and J. Vepa, “Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?,” in *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, J. Bastings, Y. Belinkov, Y. Elazar, D. Hupkes, N. Saphra, and S. Wiegreffe, Eds., Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Sep. 2022, pp. 238–248. doi: 10.18653/v1/2022.blackboxnlp-1.19.

[44] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” Oct. 16, 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685