# EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos

Laura Plaza[1], Jorge Carrillo-de-Albornoz[1], Iván Arcos[2], Paolo Rosso[2,3], Damiano Spina[4], Enrique Amigó[1], Julio Gonzalo[1], Roser Morante[1]

[1] Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
lplaza,jcalbornoz,enrique,julio,rmorant@lsi.uned.es
[2] Universitat Politècnica de València (UPV), 46022 Valencia, Spain
prosso@dsic.upv.es,iarcgab@etsinf.upv.es
[3] ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence, 46022 Valencia, Spain
[4] RMIT University, 3000 Melbourne, Australia
damiano.spina@rmit.edu.au

**Abstract.** The paper describes the EXIST 2025 lab on Sexism identification in social networks, that is expected to take place at the CLEF 2025 conference and represents the fifth edition of the EXIST challenge. The lab comprises nine tasks in two languages, English and Spanish, which are the same three tasks (*sexism identification*, *source intention detection*, and *sexism categorization*) applied to three different types of data: text (tweets), image (memes) and video (TikToks). This multimedia approach will help identify trends and patterns in sexism across media formats and user interactions, contributing to a deeper understanding of the social dynamics at play. As in EXIST 2023 and 2024, this edition will use the "Learning With Disagreement" approach. The datasets for the nine tasks will include annotations from multiple annotators, showing different or even conflicting opinions. This helps models learn from diverse perspectives, making them better at understanding a range of human viewpoints, and contributing towards effective human-centric development of solutions.

**Keywords:** sexism identification · sexism categorization · learning with disagreement · tweets · memes · TikTok videos · human-centric AI

## 1 Introduction

Sexism refers to prejudice or discrimination based on a person's sex or gender, often manifesting in the belief that one sex is superior to another. It can take various forms, including overt actions, societal norms, and institutional practices that disadvantage individuals based on their gender. The effects of sexism are far-reaching, influencing individuals of all genders, although women often experience the most significant consequences.

Sexism manifests and proliferates on the Internet, particularly through social media platforms like Twitter and TikTok, where the rapid sharing of content can amplify harmful messages [1, 2]. On Twitter, sexism often takes the form of derogatory comments, harassment, and trolling, especially targeted at women and marginalized groups. The platform's character limit encourages succinct, sometimes aggressive expressions of bias, enabling users to engage in hostile exchanges that can quickly go viral. Additionally, misogynistic hashtags and trends can create echo chambers that normalize and perpetuate sexist attitudes, allowing harmful stereotypes to spread unchecked. On TikTok, sexism is expressed in various ways. The platform's algorithm can promote videos that garner high engagement, which sometimes includes content that trivializes or mocks gender issues. Users, particularly young women, may face pressure to conform to specific beauty standards or behaviors, leading to the reinforcement of sexist ideals. As a result, both platforms play significant roles in shaping cultural narratives around gender, making it essential to address the ways sexism is communicated and shared in these digital spaces.

EXIST 2025[5] will be the fifth edition of the sEXism Identification in Social neTworks challenge. EXIST is a series of scientific events and shared tasks that aim to capture sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviors. The two last editions of the EXIST shared task were held as labs in CLEF [3, 4], while the first two editions were held in the IberLEF Spanish evaluation forum [6, 7]. Along the four years, more than 100 teams from research institutions and companies have participated in the EXIST challenge. While the three first editions focused on detecting and classifying sexism in textual messages, the fourth edition incorporates new tasks that center around images, particularly memes. The fifth edition of EXIST will extend to a new type of data: TikTok videos.

TikTok is a popular social media platform that allows users to create, share, and discover short-form videos, typically ranging from 15 seconds to 3 minutes. TikTok has grown rapidly, boasting over 1 billion active users globally, with a significant portion of its user base being younger people (around 60% of TikTok users are aged 16-24, making Gen Z the platform's largest demographic). Studies have shown that TikTok's algorithm can reinforce sexism by progressively exposing users to sexist content based on their prior interactions, creating filter bubbles that normalize misogynistic attitudes [10]. Additionally, trends of hypersexualization and gender stereotypes in videos, along with disparities in content moderation, contribute to the perpetuation of traditional gender roles and the objectification of women [11]. This is particularly concerning for adolescents, as repeated exposure to such content can negatively impact their self-esteem and gender perceptions [12].

Building on the success of the previous two editions, the 2025 edition of EXIST will also adopt the Learning With Disagreement (LeWiDi) paradigm for both dataset creation and system evaluation. The LeWiDi paradigm can significantly contribute to achieving human-centric AI by prioritizing diverse human

[5] https://nlp.uned.es/exist2025

perspectives and incorporating disagreements into the decision-making process [8]. Instead of relying on a single "correct" label, it teaches AI systems to handle and learn from conflicting annotations, reflecting the complexity and subjectivity of human reasoning. This approach helps reduce bias, ensures more inclusive, fair and equitable decision-making, and builds transparency. By incorporating multiple viewpoints, LeWiDi enables AI to better align with human values and promote fairness in real-world applications.

In the following sections, we provide comprehensive information about the tasks, the dataset and the evaluation methodology that will be adopted in the EXIST 2025 challenge at CLEF.

## 2 EXIST 2025 Tasks

The last edition of EXIST focused on detecting and categorizing both sexist tweets and memes. For the 2025 edition, we will revisit these tasks. The motivation behind this is that the results for tasks related to the categorization of sexism and intention identification remain relatively low, suggesting there is still room for improvement. Additionally, previous approaches have not fully leveraged the sociodemographic information of the annotators. Our aim is to promote more human-centered approaches. As a novelty, in the 2025 edition we will also aim to detect and categorize sexism in TikTok videos. As a result, nine tasks are proposed which are described below.

### 2.1 Sexism Identification in Tweets

The first task is a binary classification task where systems must decide whether or not a given tweet is sexist. The following statements show examples of sexist and not sexist messages, respectively.

(1) **Sexist:** *Woman driving, be careful!.*
(2) **Non sexist:** *Just saw a woman wearing a mask outside spank her very tightly leashed dog and I gotta say I love learning absolutely everything about a stranger in a single instant.*

### 2.2 Source Intention Detection in Tweets

This task aims to categorize the message according to the intention of the author. We propose a ternary classification task: (i) direct sexist message, (ii) reported sexist message and (iii) judgmental message. This distinction allows to differentiate sexism that is actually taking place in online platforms from sexism which is being suffered by women in other situations but that is being reported in social networks with the aim of complaining and fighting against sexism. The following categories are defined:

- **Direct** sexist message: the intention was to write a message that is sexist by itself or incites to be sexist, as in:

(3) *A woman needs love, to fill the fridge, if a man can give this to her in return for her services (housework, cooking, etc), I don't see what else she needs.*

- **Reported** sexist message: the intention is to report and share a sexist situation suffered by a woman or women in first or third person, as in:

  (4) *Today, one of my year 1 class pupils could not believe he'd lost a race against a girl.*
- **Judgemental** message: the intention is judgmental, since the tweet describes sexist situations or behaviors with the aim of condemning them.

  (5) *21st century and we are still earning 25% less than men #Idonotrenounce.*

### 2.3 Sexism Categorization in Tweets

Many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. In this task, each sexist tweet must be categorized in one or more of the following categories:

- **Ideological and inequality**: this category includes tweets that discredit the feminist movement in order to devalue and defame the struggle of women in any aspect of their lives. It also includes tweets that reject inequality between men and women, or present men as victims of gender-based oppression.

  (6) *Think the whole equality thing is getting out of hand. We are different, that's how were made!*
- **Stereotyping and dominance**: this category includes messages that express false ideas about women that suggest they are more suitable or inappropriate for certain tasks. It also includes any claim that implies that men are somehow superior to women.

  (7) *Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.*
- **Objectification**: Objectification and physical stereotyping includes messages where women are presented as objects apart from their dignity and personal aspects. We also include messages that assume or describe certain physical qualities that women must have in order to fulfill traditional gender roles, for example, ideas that suggest that women should maintain a standard and ideal of beauty or attacks on a woman's body.

  (8) *No offense but I've never seen an attractive african american hooker. Not a single one.*
- **Sexual violence**: this category includes messages where sexual suggestions, requests or harassment of a sexual nature (rape or sexual assault) are made.

  (9) *I wanna touch your tits..you can't imagine what I can do on your body.*
- **Misogyny and non sexual violence**: this category includes expressions of hatred and violence towards women.

  (10) *Domestic abuse is never okay... Unless your wife is a bitch.*

### 2.4 Task 4: Sexism Identification in Memes

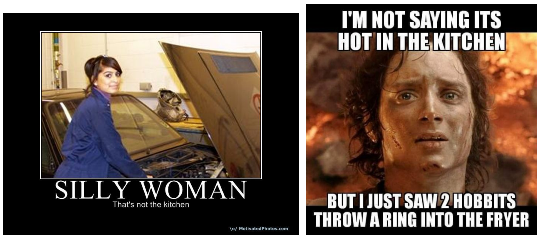
This is a binary classification task consisting on deciding whether or not a given meme is sexist (see Figure 1).

### 2.5 Task 5: Source Intention in Memes

Similar to Task 2, this task aims to categorize memes according to the author's intention. However, in this task, systems should only classify memes as either direct or judgmental, as reported memes are not frequent. Examples of direct and judmental sexist memes can be found in [4].

### 2.6 Task 6: Sexism Categorization in Memes

This task aims to classify sexist memes according to the categorization provided for Task 3. Examples of memes from different categories can be found in [4].


SILLY WOMAN
That's not the kitchen
I'M NOT SAYING ITS HOT IN THE KITCHEN
BUT I JUST SAW 2 HOBBITS THROW A RING INTO THE FRYER

(a) Sexist meme (b) Non sexist meme

Fig. 1: Examples of memes from the EXIST dataset

### 2.7 Task 7: Sexism Identification in TikToks

As in Tasks 1 and 4, systems must determine whether a given short video shared on TikTok is sexist.

### 2.8 Task 8: Source Intention in TikToks

As in Tasks 2 and 5, this task aims to categorize the short video according to the intention of the author, as direct or judgemental.

### 2.9 Task 9: Sexism Categorization in TikToks

As in Tasks 3 and 6, this task aims to categorize short videos according to the categorization provided for Task 3.

## 3 The EXIST 2025 Dataset

The EXIST 2025 Dataset is made up of three distinct subsets: the EXIST Tweets Dataset, the EXIST Memes Dataset and the EXIST TikTok Dataset.

### 3.1 EXIST 2023 Tweets Dataset

For Tasks 1 to 3, we will use the EXIST 2023 dataset as the test set was kept private.[6] The EXIST 2023 dataset consists of 10,034 annotated tweets from a large and diverse group of 1,065 annotators, selected by their gender (male/female) and age (18–22 y.o./23–45 y.o./+46 y.o) [3].

### 3.2 EXIST 2024 Memes Dataset

For Tasks 4 to 6, we will use the EXIST 2024 dataset as the test set was kept private. It consists of 3,000 memes annotated by 6 different persons, selected by their gender and age (see [4] for a description of the dataset).

### 3.3 EXIST 2025 TikTok Dataset

The TikTok dataset has been specifically developed for the 2025 edition of EXIST. A detailed description of the dataset may be found in [13]. We have collected more than 3,500 videos in English and Spanish from different TikTok accounts, following the same methodology as in EXIST 2023 and 2024 for downloading the data, using the same seeds and filtering process. However, labeling is done by trained annotators rather than crowd workers. The learning with disagreement paradigm has been adopted, as in EXIST 2023 and 2024.

The data was collected using the Apify's TikTok Hashtag Scraper tool,[7] using 185 Spanish hashtags and 61 English hashtags associated with potentially sexist content. The annotation was conducted using Servipoli's service[8] with eight students organized in pairs. Each pair, consisting of one male and one female student to avoid biases, was tasked with annotating 1,000 TikTok videos in either Spanish or English. The Spanish TikTok dataset consists of a total of 1,969 TikToks, with a cumulative duration of 13.86 hours. The English TikTok dataset comprises a total of 1,773 TikToks, with a cumulative duration of 11.83 hours.

## 4 Evaluation Methodology and Metrics

As in EXIST 2023 and 2024, we will carry out two types of evaluations:

1. **Soft-soft evaluation**. This evaluation is intended for systems that provide probabilities for each category, rather than a single label. We will use a modification of the ICM metric (Information Contrast Measure [9]), ICM-Soft (see details in [3]), as the official evaluation metric in this variant and we will also provide results for the normalized version of ICM-Soft (ICM-Soft Norm). We will also provide results for Cross Entropy.

[6] The test sets will remain private to prevent contamination of generative models, like ChatGPT, which continuously scrape the web and could produce skewed results.

[7] https://apify.com/clockworks/tiktok-hashtag-scraper

[8] https://www.servipoli.es/

2. **Hard-hard evaluation**. This evaluation is intended for systems that provide a hard, conventional output. To derive the hard labels in the ground truth from the different annotators' labels, we will use a probabilistic threshold computed for each task. The official metric for this task will be the original ICM, as defined by Amigó and Delgado [9]. We will also report a normalized version of ICM (ICM Norm) and F1 Score.

**Acknowledgments** This work has been financed by the Spanish Ministry of Science, Innovation and Universities (project FairTransNLP (PID2021-124361OB-C31 and PID2021-124361OB-C32)) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe, and by the Australian Research Council (CE200100005).

## References

1. Social Media and the Silencing Effect: Why Misogyny Online is a Human Rights Issue. NewStatesman, https://bit.ly/3n3ox68. Last accessed 18 Oct 2023.
2. Gil Bermejo J.L., Martos Sánchez C., Vázquez Aguado O., García-Navarro E.B.: Adolescents, Ambivalent Sexism and Social Networks, a Conditioning Factor in the Healthcare of Women. Healthcare (Basel). 2021 Jun 12;9(6):721.
3. Plaza, L., Carrillo-de-Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P.: Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. Experimental IR Meets Multilinguality, Multimodality, and Interaction. In Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro, Eds. September 2023, Thessaloniki, Greece.
4. Plaza, L., Carrillo-de-Albornoz, J., Ruiz, V., Maeso, A., Chulvi, B., Rosso, P., Amigó, E., Gonzalo, J., Morante, R., Spina, D.: Overview of EXIST 2024 — Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes. Experimental IR Meets Multilinguality, Multimodality, and Interaction. In Proceedings of the 15th International Conference of the CLEF Association (CLEF 2025). Lorraine Goeuriot, Philippe Mulhem, Georges Quénot · Didier Schwab, Giorgio Maria Di Nunzio, Laure Soulier, Petra Galušcáková, Alba García Seco de Herrera, Guglielmo Faggioli and Nicola Ferro, Eds. September 2024, Grenoble, France.
5. Plaza, L., Carrillo-de-Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P.: Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview). Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum. Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro and Michalis Vlachos, Eds.
6. Rodríguez-Sánchez, F.,Carrillo-de-Albornoz, J.,Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of EXIST 2021: Sexism Identification in Social Networks. Procesamiento del Lenguaje Natural,**67**, 195–207 (2021)
7. Rodríguez-Sánchez, F.,Carrillo-de-Albornoz, J.,Plaza, L., Mendieta-Aragón, A., Marco-Remón,G., Makeienko, M., Plaza, M., Spina, D., Gonzalo, J., Rosso, P.: Overview of EXIST 2022: Sexism Identification in Social Networks. Procesamiento del Lenguaje Natural,**69**, 229–240. 2022.

8. Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A.: We Need to Consider Disagreement in Evaluation. In Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 15–21, Online. Association for Computational Linguistics. 2021.
9. Amigó, E., Delgado, A.: Evaluating Extreme Hierarchical Multi-label Classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pages 5809–5819. 2022.
10. Morales Rodríguez, G., Lopez-Figueroa, J. The Portrayal of Women in Media. Journal of Student Research, 13(2). 2024.
11. Davis, S. E. Objectification, Sexualization, and Misrepresentation: Social Media and the College Experience. Social Media + Society, 4(3). 2018.
12. Harriger, J.A., Thompson, J.K., Tiggemann, M. TikTok, TikTok, the time is now: Future directions in social media and body image. Body Image, 44: 222-226. 2023.
13. Arcos, I., Rosso, P. Sexism Identification on TikTok: A Multimodal AI Approach with Text, Audio, and Video. In: Goeuriot, L., et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. Lecture Notes in Computer Science, vol 14958. 2024.