EXIST 2024: sEXism Identification in Social neTworks and Memes

Laura Plaza[1], Jorge Carrillo-de-Albornoz[1], Enrique Amigó[1], Julio Gonzalo[1], Roser Morante[1], Paolo Rosso[2,3], Damiano Spina[4], Berta Chulvi[2], Alba Maeso[2], Víctor Ruiz[1]

[1] Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
lplaza,jcalbornoz,enrique,julio,rmorant,victor.ruiz@lsi.uned.es
[2] Universitat Politècnica de València (UPV), 46022 Valencia, Spain
prosso@dsic.upv.es,berta.chulvi@upv.es,amaeolm@inf.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 2024 lab on Sexism identification in social networks, that is expected to take place at the CLEF 2024 conference and represents the fourth edition of the EXIST challenge. The lab comprises five tasks in two languages, English and Spanish, with the initial three tasks building upon those from EXIST 2023 (*sexism identification in tweets*, *source intention detection in tweets*, and *sexism categorization in tweets*). In this edition, two new tasks have been introduced: *sexism detection in memes* and *sexism categorization in memes*. Similar to the prior edition, this one will adopt the Learning With Disagreement paradigm. The dataset for the various tasks will provide all annotations from multiple annotators, enabling models to learn from a range of training data, which may sometimes present contradictory opinions or labels. This approach facilitates the model's ability to handle and navigate diverse perspectives. Data bias will be handled both in the sampling and in the labeling processes: seed, topic, temporal and user bias will be taken into account when gathering data; in the annotation process, bias will be reduced by involving annotators from different social and demographic backgrounds.

Keywords: sexism identification · sexism categorization · learning with disagreement · memes · data bias

1 Introduction

Sexism is rooted in beliefs concerning the inherent nature of women and men, dictating the roles they are expected to fulfill in society. These beliefs often materialize as gender stereotypes, potentially elevating one gender above the other in a hierarchical manner. This hierarchical mindset can manifest either consciously and overtly, or subtly and unconsciously, often appearing as unconscious bias. The impact of sexism is widespread, affecting individuals across the gender spectrum, with women bearing a significant brunt of its consequences.

Inequality and discrimination against women persist within society and are increasingly replicated online [1]. The internet has the power to perpetuate and normalize gender differences and sexist attitudes [2], especially among teenagers. Previous studies have shown how sexism towards women is perpetuated on social networks and their influence on the teenagers' health [3]. Although platforms like Twitter continually strive to identify and eradicate hateful content, they grapple with the overwhelming volume of user-generated data [4]. In this context, leveraging automatic tools becomes crucial not only for detecting and alerting against sexist behaviors and discourses, but also for estimating the prevalence of sexist and abusive situations on social media platforms, identifying common forms of sexism, and understanding how sexism is expressed through these channels.

EXIST 2024[5] will be the fourth 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 last edition of the EXIST shared task was held as a lab in CLEF 2023 [5, 6], while the first two editions were held in the IberLEF Spanish evaluation forum [7, 8]. The EXIST 2024 Lab will be one of the participating labs of the new MonsterCLEF initiative that focuses on the versatility of Large Language Models.[6]

In past editions, more than 75 teams participated from research institutions and companies from all around the world. While the three previous editions focused solely on detecting and classifying sexist textual messages, this new edition incorporates new tasks that center around images, particularly memes. Memes are images, typically humorous in nature, that are spread rapidly by social networks and Internet users. With this addition, we aim to encompass a broader spectrum of sexist manifestations in social networks, especially those disguised as humor. As indicated in [9], the consumption of memes and their increasing visual complexity is on the rise each day. Additionally, as indicated by [10], memes can propagate globally through repackaging in different languages and they can also be multilingual. Consequently, it becomes imperative to develop automated multimodal tools capable of detecting sexism in both text and images.

Similar to the approach in the 2023 edition, this edition will also embrace the Learning With Disagreement (LeWiDi) paradigm for both the development of the dataset and the evaluation of the systems. Our prior research demonstrated that perceiving sexism is significantly influenced by the demographic and cultural background of the individual. Consequently, when identifying sexist attitudes and expressions, and even when categorizing them into distinct sexist categories, disagreements among individuals are common. The LeWiDi paradigm doesn't rely on a single "correct" label for each example. Instead, the model is trained to handle and learn from conflicting or diverse annotations. This enables the

[5] https://nlp.uned.es/exist2024

[6] https://monsterclef.dei.unipd.it

system to consider various annotators' perspectives, biases, or interpretations, resulting in a fairer learning process.

In the following sections, we provide comprehensive information about the tasks, the dataset and the evaluation methodology that will be adopted in the EXIST 2024 challenge at CLEF.

2 EXIST 2024 Tasks

The last edition of EXIST focused on detecting sexist messages in Twitter as well as on categorizing these messages according to the type of sexist behavior they enclose and the intention of the source (Tasks 1 to 3). For the 2024 edition, we will address also multimedia content. In particular, we will aim to detect and categorize sexism in memes. Therefore, five tasks are proposed which are described below.

2.1 Sexism Identification

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

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 will allow us 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

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:

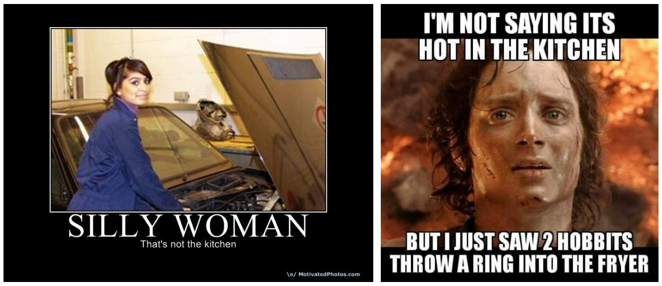
- **Ideological and inequality**: this category includes all tweets that discredit the feminist movement in order to devalue, belittle and defame the struggle of women in any aspect of their lives. It also includes messages that reject inequality between men and women, or present men as victims of gender-based oppression. Some examples of this type of sexism discourse can be found in the following tweets:
  (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. Figure 1 shows examples of sexist and non sexist memes, respectively.

2.5 Task 5: Sexism Categorization in Memes

This task aims to classify sexist memes according to the categorization provided for Task 3: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence.


SILLY WOMAN
That's not the kitchen
\o/ MotivatedPhotos.com
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 2024 dataset

3 The EXIST 2024 Dataset

In recent years, a number of proposals have advocated for considering disagreement as informative content that can enhance task performance [11]. Evidence suggests that gold labels are an idealization and that irreconcilable disagreement is common, especially in tasks involving highly subjective judgments, such as sexism detection.

Following this trend, for the development of the dataset we will follow the learning with disagreement paradigm, so that rather than aggregating the (potentially) contradictory labels provided by the annotators, all of them will be kept and provided to the participating systems so that they may exploit disagreement in their systems. The LeWiDi paradigm may also help to mitigate bias and produce equitable NLP systems.

For the development of the EXIST 2024 dataset we will collect memes from different sources such as Google image search, Twitter, special forums such as Reddit or Forocoche. The downloaded images will be processed to remove duplicate items, as well as analyzed with an OCR software to extract the text following the methodology used in [12]. For the crawling and filtering, we will use the same methodology as in EXIST 2023.

The selection of annotators for the development of the EXIST 2024 dataset will take into account the heterogeneity necessary to avoid gender and age biases. Annotation will be performed by crowd-workers using for instance Prolific [13]. The Prolific crowdsourcing platform was specifically selected because of the features it provides to define participant criteria in the recruiting process – in our case, gender, age, and fluency in the different languages. Each tweet/meme will be annotated by six workers that will be selected according to their different demographic characteristics in order to minimize the label bias.

4 Evaluation Methodology and Metrics

As in EXIST 2023, we will carry out a **hard evaluation** and a **soft evaluation**. In the hard evaluation, both the system and the gold standard consist of one or more labels assigned to each instance in the dataset. In contrast, the soft evaluation is intended to measure the ability of the model to capture disagreement, by considering the probability distribution of labels in the output as a soft label and comparing it with the probability distribution of the annotations.

The LewiDi paradigm can be considered in both sides of the evaluation process:

- **The ground truth.** In a hard setting, the variability in the human annotations is reduced by selecting one and only one gold category per instance, the hard label. In a soft setting, the gold standard label for one instance is the set of all the human annotations existing for that instance. Therefore, the evaluation metric incorporates the proportion of human annotators that have selected each category (soft labels).
- **The system output.** In a hard traditional setting, the system predicts one or more categories for each instance. In a soft setting, the system predicts a probability for each category, for each instance. The evaluation score is maximized when the probabilities predicted match the actual probabilities in a soft ground truth.

Two different types of evaluation will be accomplished:

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 [14]), ICM-Soft (see details in [5]), 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.
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 [14]. We will also report a normalized version of ICM (ICM Norm) and F1.

Acknowledgments This work has been financed by the European Union (Next Generation EU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Economic Affairs and Digital Transformation and by the UNED University. It has also been financed by the Spanish Ministry of Science and Innovation (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 (DE200100064 and 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. Burgos, A., Donoso-Vázquez, T., López, E.,Cabañó, E., Martínez, Y., Martín, A., Prado-Soto, N., Rubio, M., Velasco-Martínez, A., Vila, R: Violencias de Género 2.0, pp. 13–27 (2014)
3. 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.
4. Twitter's Famous Racist Problem. The Atlantic, https://bit.ly/38EnFPw . Last accessed 17 Oct 2023.
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. 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.
6. 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.
7. 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)
8. 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.
9. Valensise, C. M., Serra, A., Galeazzi, A., Etta, G., Cinelli, M., Quattrociocchi, W.: Entropy and Complexity Unveil the Landscape of Memes Evolution. Scientific Reports, 11(1), 1-9. 2021.
10. Sharma, S., Alam, F., Akhtar, M. S., Dimitrov, D., Martino, G. D. S., Firooz, H., Chakraborty, T.: Detecting and Understanding Harmful Memes: A Survey. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pages 5597–5606, 2022.
11. 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.
12. Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., Lees, A., and Sorensen, J.: SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 533–549. Association for Computational Linguistics. 2022.
13. Prolific, https://www.prolific.com/. Last accessed 18 Oct 2023.
14. 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.