Overview of EXIST 2023: sEXism Identification in Social neTworks

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

[1] Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
lplaza,jcalbornoz,rmorant,enrique,julio@lsi.uned.es
[2] RMIT University, 3000 Melbourne, Australia
damiano.spina@rmit.edu.au
[3] Universidad Politécnica de Valencia (UPV), 46022 Valencia, Spain
prosso@dsic.upv.es

Abstract. The paper describes the lab on Sexism identification in social networks (EXIST 2023) that will be hosted as a lab at the CLEF 2023 conference. The lab consists of three tasks, two of which are continuation of EXIST 2022 (*sexism detection* and *sexism categorization*) and a third and novel one on *source intention identification.* For this edition new test and training data will be provided and some novelties are introduced in order to tackle two central problems of Natural Language Processing (NLP): bias and fairness. Firstly, the sampling and data gathering process will take into account different sources of bias in data: seed, temporal and user bias. During the annotation process we will also consider some sources of “label bias” that come from the social and demographic characteristics of the annotators. Secondly, we will adopt the “learning with disagreements” paradigm by providing datasets containing also pre-aggregated annotations, so that systems can make use of this information to learn from different perspectives. The general goal of the EXIST shared tasks is to advance the state of the art in online sexism detection and categorization, as well as investigating to what extent bias can be characterized in data and whether systems may take fairness decisions when learning from multiple annotations.

Keywords: Sexism detection · Data bias · Learning with disagreements.

1 Introduction

Sexism[4] is pervasive, even in modern developed societies and among young generations who have been born in democratic societies. Recently, a video went viral where male students of a prestigious hall of residence in Spain were addressing extremely sexist and insulting messages to the girls of the neighbouring residence as a way of welcoming them for the new academic year.[5] Inequality and

[4] The Oxford English Dictionary defines sexism as “prejudice, stereotyping or discrimination, typically against women, on the basis of sex”.

[5] https://www.theguardian.com/world/2022/oct/06/spanish-pm-leads-outcry-over-students-filmed-chanting-abuse-at-womens-halls. Accessed 14 October 2022.

discrimination against women that remain embedded in society are increasingly being replicated online [1]. Internet perpetuates and even naturalizes gender differences and sexist attitudes [2]. Moreover, given that an important percentage of Internet users (especially social networks users) are teenagers, the increasing sexism on the Internet requires urgent study and social debate that leads to actions, especially from an educational point of view.

Social networks are the main platforms for social complaint, activism, etc., and movements like #MeTwoo, #8M or #Time'sUp have spread rapidly. Under this umbrella, many women all around the world have reported abuses, discrimination and sexist experiences suffered in real life. Social networks are also contributing to the transmission of sexism and other disrespectful and hateful behaviours. Even though social platforms such as Twitter are continuously creating new ways to identify and eradicate hateful content, they are facing many difficulties when dealing with the huge amount of data generated by users [3]. In this context, automatic tools not only may help to detect and alert against sexism behaviours and discourses, but also to estimate how often sexist and abusive situations are found in social media platforms, what forms of sexism are more frequent and how sexism is expressed in these media.

Up to date, most work dealing with sexism in online media is focused on detecting misogyny or hatred towards women [4–6]. However, sexism does not always imply misogyny. Sexism may sound "friendly" as in (1), which may seem positive, but is actually transmitting that women are weaker than men. Sexism may sound "funny", as reflected on sexist jokes or humour (2). Or sexism may sound "offensive" and "hateful", as in (3). However, even the most subtle forms of sexism can be as pernicious as the most violent ones and affect women in many facets of their lives, including domestic and parenting roles, career opportunities, sexual image and life expectations, to name a few.

(1) *Women must be loved and respected, always treat them like a fragile glass.*
(2) *You have to love women... just that... You will never understand them.*
(3) *Humiliate, expose and degrade yourself as the fucking bitch you are if you want a real man to give you attention.*

Until recently, no NLP shared tasks had addressed the detection of sexism in social media. To fill this research gap, in 2021 and 2022 the sEXism Identification in Social neTworks (EXIST) shared tasks were proposed at the IberLEF forum [7, 8]. The EXIST challenge was the first shared task on sexism detection in social networks whose aim was to identify and classify sexism in a broad sense, from explicit and/or hostile to other subtle or even benevolent expressions that involve implicit sexist behaviours. In the framework of these competitions, the first dataset of broad sexism was released. The 2021 and 2022 EXIST editions welcomed more than 50 teams from research institutions and companies from all around the world, demonstrating the importance of the problem, as well as the great interest of the research community around it. Given the success of the tasks, in 2023 a new edition of EXIST will take place as a lab at CLEF, which will continue with the tasks addressed in previous years, while facing yet a new challenge: the identification of the intention of the author of the sexist message. However, the main novelty will be the adoption of the "learning with disagreements" paradigm [10] for the development of the dataset and, optionally, for the evaluation of the systems. The adoption of this paradigm along with our effort to control bias in the annotations (see Section 3) will allow us to evaluate whether including the different views and sensibilities of the annotators contributes to the development of more accurate and fairer NLP systems.

In what follows we present the EXIST task for the 2023 edition.

2 EXIST Tasks

The two first editions of EXIST focused on detecting sexist messages in two social networks, Twitter and Gab footnotehttps://gab.com/. Accessed 14 Oct 2022. as well as on categorizing these messages according to the type of sexist behaviour they enclose. For the 2023 edition, we will focus on Twitter and we will address an additional task, namely "source intention classification". Therefore, three tasks are proposed which are described below.

2.1 Sexism Detection

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.

(4) **Sexist:** *Woman driving, be careful!.*
(5) **Not 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 Classification

This task aims to categorize the message according to the intention of the author, which provides insights in the role played by social networks on the emission and dissemination of sexist messages. In this task, we propose a ternary classification task: (i) direct sexist message, (ii) reported sexist message and (iii) judgemental 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:
  (6) *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.*
  (7) *Women shouldn't code...perhaps be influencer/creator instead...it's their natural strength.*
- **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:

(8) *I doze in the subway, I open my eyes feeling something weird: the hand of the man sat next to me on my leg #SquealOnYourPig.*

(9) *Today, one of my year 1 class pupils could not believe he'd lost a race against a girl.*

- **Judgemental** message: the intention was judgmental, since the tweet describes sexist situations or behaviours with the aim of condemning them.

  (10) *As usual, the woman was the one quitting her job for the family's welfare...*

  (11) *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. Automatically detecting which of these facets of women are being more frequently attacked in social networks will facilitate the development of policies to fight against sexism. According to this, 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:

  (12) *#Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity.*

  (13) *Think the whole equality thing is getting out of hand. We are different, that's how were made!*

- **Role 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.

  (14) *Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to.*

  (15) *I feel like everytime I flirt with a girl they start to imagine all the ways they can utilize me.*

- **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 physique.

  (16) *No offense but I've never seen an attractive african american hooker. Not a single one.*

  (17) *Most of the times I just want women for sex just because everytime I try to make a relationship out of it they always break my heart.*

- **Sexual violence**: this category includes messages where sexual suggestions, requests or harassment of a sexual nature (rape or sexual assault) are made:

  (18) *I wanna touch your tits..you can't imagine what I can do on your body.*

(19) *Fuck that cunt, I would with my fist.*

- **Misogyny and non sexual violence**: this category includes expressions of hatred and violence towards women.
  (20) *Domestic abuse is never okay.... Unless your wife is a bitch.*
  (21) *Some woman are so toxic they don't even know they are draining everyone around them in poison.*

3 The EXIST 2023 Dataset: Controlling Bias in Data and Annotation

An important problem in NLP that has gained attention in the recent years in parallel to the growing protagonism of big language models is bias, both in the data that are used to train and test systems, and in the way algorithms learn, mainly due to the bias in the data [15]. In EXIST 2023 we tackle one aspect of this problem, i.e., the data bias that may be introduced both during the data selection and during the labeling process.

Firstly, the sampling and data gathering process will take into account different sources of bias: seed, temporal and user bias. To mitigate the seed bias, we will use as seeds a wide range of terms that are employed in both sexist and non-sexist contexts. Temporal bias between training, validation and test data will be mitigated by selecting texts from different time spans, with a temporal gap between the sets. We will also check the temporal gap between tweets for each seed to ensure that data are spread all over a certain period. Finally, we will check the author of the messages to ensure an appropriate balance in the contribution of the different types of users. We will also avoid having tweets by the same user in the different subsets.

Secondly, during the annotation process we will consider some sources of "label bias" [11]. Label bias may be introduced by socio-demographic differences of the persons that participate in the annotation process, but also when more than one possible correct label exists or when the decision on the label depends on subjectivity. During the development of the datasets, we will devote special attention to avoid annotation bias. Some of the following socio-demographic sources of bias will be considered: gender, ethnicity, country, education and age. In order to avoid the bias derived from the individuality of the annotators, we will adopt the "learning with disagreements" paradigm (see Section 4).

The labelling process will be carried out by crowd-workers, selected according to their different social and demographic parameters in order to avoid the label bias. Each tweet will be annotated by enough crowd-workers to ensure diversity. At least 5,000 tweets will be crawled from Twitter in two different languages: English and Spanish. A train, a validation and a test set will be provided. To retrieve the tweets, more than 200 potentially sexist phrases will be used as seeds. These phrases have been extracted from different sources: (a) previous works in the area; (b) Twitter accounts (journalist, teenagers, etc.) or hashtags used to report sexist situations; (c) expressions extracted from the EveryDay-Sexism project[6]; d) a compendium of feminist dictionaries. We will also include

[6] Everyday sexism project: https://everydaysexism.com/. Accessed 14 October 2022.

other common hashtags and expressions less frequently used in sexist contexts to ensure a balanced distribution between sexist/not sexist expressions.

4 Learning with Disagreements

As stated in [10], the assumption that natural language expressions have a single and clearly identifiable interpretation in a given context is a convenient idealization, but far from reality. To deal with this, Uma et al. [10] have proposed the learning with disagreements paradigm (LeWiDi) , which consists mainly in letting systems learn from datasets where no gold annotations are provided but information about the annotations from all annotators, in an attempt to gather the diversity of views. In the case of sexism identification, this is particularly relevant, since the perception of a situation as sexist or not can be subjective and may depend on gender, age and cultural background of the person who is judging it. Following methods proposed for training directly from the data with disagreements, instead of using an aggregated label [12–14], we will provide multiple annotations per example, as was done in the SemEval-2021 shared task [10] on learning with disagreements that provided a unified testing framework for methods for learning from data containing multiple and possibly contradictory annotations. The LeWiDi paradigm is a first step towards more equitative and fairer algorithms that reflect multiple points of view.

The selection of annotators for the development of the EXIST 2023 dataset will take into account the heterogeneity necessary to avoid bias. Rather than eliminating disagreements by selecting the majority vote (as done in EXIST 2021 and 2022), EXIST 2023 will preserve the multiple labels assigned by an heterogeneous and representative group of annotators, so that participants may exploit disagreement in their systems.

5 Evaluation Methodology and Metrics

As in SemEval 2021, we will carry out a "hard evaluation" and a "soft evaluation". The hard evaluation will assume that a single label is provided by the systems for every example in the dataset. The soft evaluation is intended to measure the ability of the model to capture disagreements, by considering the distribution of labels in the output as a soft label and comparing it with the distribution of the annotations. Different metrics for both types of evaluations will be employed in order to facilitate comparison among participants and with previous works.

Acknowledgments This work is supported by the Spanish Ministry of Science and Innovation (project FairTransNLP (PID2021-124361OB-C31 and PID2021-124361OB-C32)) and the Spanish Ministry of Economy and Competitiveness (project pace for Observation of AI in Spanish (UNED and RED.ES, M.P., ref. C039/21- OT)).

References

1. Social Media and the Silencing Effect: Why Misogyny Online is a Human Rights Issue. NewStatesman, https://bit.ly/3n3ox68. Accessed 25 Sep 2022.
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. Twitter's Famous Racist Problem. The Atlantic, https://bit.ly/38EnFPw . Accessed 25 Sep 2022.
4. Anzovino, M., Fersini, E., Rosso, P.: Automatic Identification and Classification of Misogynistic Language on Twitter. In Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB '18), pp. 57–64 (2018)
5. Pamungkas, E., Basile, V., Patti, V.: Misogyny Detection in Twitter: A Multilingual and Cross-Domain Study. Information Processing & Management, **57**(6), p. 102360 (2020)
6. Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., Margetts, H.: An Expert Annotated Dataset for the Detection of Online Misogyny. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL '21), pp. 1336–1350 (2021)
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**, pp. 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**, pp. 229–240 (2022)
9. Amigó, E., Giner, F., Gonzalo, J., Verdejo, F.: On the Foundations of Similarity in Information Access. Information Retrieval Journal, **23**, pp. 216–254 (2019)
10. Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., Poesio, M.: SemEval-2021 Task 12: Learning with Disagreement. Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 338–347 (2021)
11. Hovy, D., Prabhumoye, S.:Five Sources of Bias in Natural Language Processing. Language and Linguistics Compass, **15**(8), p. e12432 (2021)
12. Sheng, V. S., Provost, F., Ipeirotis, P. G.:Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08), pp. 614–622 (2008)
13. Rodrigues, F., Pereira. F. C.:. Deep Learning from Crowds. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 1611–1618 (2018)
14. Peterson, J. C., Battleday, R. M., Griffiths, T. L., Russakovsky, O.: Human uncertainty makes classification more robust. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, pp. 9616–9625 (2019)
15. Roselli, D., Matthews, J., Talagala, N.: Managing Bias in AI. In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19), pp. 539–544 (2019)