A REVIEW OF THE DATASETS

ASSESSING RELIABILITY OF ANNOTATIONS IN THE CONTEXT OFMODEL PREDICTIONS AND EXPLANATIONS

Abstract

In this report, we've gathered all the important information regarding the three datasets Task10 & Task 11 @SemEval2023 and EXIST.Also, we compared them together.

Shahedi, T. (Tina)

t.shahedi@uu.nl

Table of Content

Introduction

In this comparative analysis, we delve into a detailed examination of the annotation processes of three datasets: Task 10 and Task 11 from SemEval 2023, and the EXIST dataset. Our goal is to identify and compare the annotation processes across various dimensions, including the nature of the tasks, the annotation guidelines, the evaluation measures, the profiles of the annotators, and more.

Overall description of each dataset

1. 3rd Shared Task (CLEF 2023); or sEXism Identification in Social neTworks (**EXIST**)**,** the third shared task at CLEF 2023, is a series of scientific events and shared tasks on sexism identification in social networks. The task contains three hierarchical subtasks:

1. **TASK 1 - Sexism Identification**
2. **TASK 2 - Source Intention**
3. **TASK 3 - Sexism Categorization**

These tasks will be described in detail later.

---

2. Task 10 (SemEval 2023); or Explainable Detection of Online Sexism (**EDOS**) supports the development of English-language models for sexism detection that are more accurate as well as explainable. The task contains three hierarchical subtasks:

1. **TASK A - Binary Sexism Detection**
2. **TASK B - Category of Sexism.**
3. **TASK C - Fine-grained Vector of Sexism**

These tasks will be described in detail later.

---

3. Task 11 (SemEval 2023); or Learning With Disagreements (**Le-Wi-Di**) dataset combines four distinct datasets into a unified format. These data set are listed in below:

1. **MD-Agreement**
2. **ConvAbuse**
3. **HS-Brexit**
4. **ArMIS**

This harmonization is achieved using JavaScript Object Notation (JSON) format. Each entry in the datasets mentioned above features several key fields that are common across the four datasets. However, some of these fields contain information specific to each dataset. These specifics will be described in detail later.

1. EXIST shared tasks

Objective	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.
Language	1. English 2. Spanish
Origin/Source	Twitter (tweets)
No. entries	At least 5,000 tweets in two different languages
Sub tasks	**1. TASK 1 - Sexism Identification** **2. TASK 2 - Source intention classification** **3. TASK 3 - Sexism Categorization**
Sampling and data gathering process	consider sources of **bias in data**: **1. Seed,** a wide range of terms that are employed in both sexist and non-sexist contexts. 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 **2. 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 **3. User bias,** ensure an appropriate balance in the contribution of the different types of users
The annotators	1. The selection of annotators for the development of the EXIST 2023 dataset will take into account the heterogeneity necessary **to avoid bias**. 2. 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.
Annotation Process	1. Consider some sources of **label bias** base on **gender**, **ethnicity**, **country**, **education** and **age** of the annotators. 2. Adopt the "**learning with disagreements**" paradigm
Evaluation	Simmilar to SemEval 2021's approach, with **"hard" (single-label)** and **"soft" (label distribution comparison)** evaluations, using distinct metrics for participant comparison and consistency with past studies.
Disagreement	1. The key innovation is using the "learning with **disagreements**" approach for dataset development and potentially system evaluation. 2. preserve the multiple labels assigned by an heterogeneous and representative group of annotators

1.1. EXIST task 1 - Sexism Identification

Description	The first task is a binary classification task where systems must decide whether or not a given tweet is sexist
Example	**SEXIST**: Woman driving, be careful!.
	NOT SEXIST: Just saw a woman wearing a mask outside…..

1.2. EXIST task 2 - Source intention classification

<table>
<tr><td>Description</td><td colspan="2">This task aims to categorize the <u>sexist message</u> according to the intention of the author.</td></tr>
<tr><td rowspan="6">Classification tasks</td><td rowspan="2">1. Direct</td><td>to write a message that is sexist by itself or incites to be sexist</td></tr>
<tr><td>Example: A woman needs love, to fill the fridge</td></tr>
<tr><td rowspan="2">2. Reported</td><td>to report and share a sexist situation suffered by a woman or women in first or third person</td></tr>
<tr><td>Example: ….. He’d lost a race against a girl.</td></tr>
<tr><td rowspan="2">3. Judgemental</td><td>since the tweet describes sexist situations or behaviours with the aim of condemning them</td></tr>
<tr><td>Example: the woman was the one quitting her job for the family’s welfare</td></tr>
</table>

1.3. EXIST task 3 - Sexism Categorization

<table>
<tr><td>Description</td><td colspan="3">Each <u>sexist tweet</u> categorized in one or more categories.</td></tr>
<tr><td rowspan="14">The categories</td><td rowspan="4">1. Ideological and inequality</td><td colspan="2">1. The text discredits the feminist movement,</td></tr>
<tr><td colspan="2">2. Rejects inequality between men and women, or presents men as victims of gender-based oppression.</td></tr>
<tr><td rowspan="2">Example:</td><td>1. Feminism is a war on men….</td></tr>
<tr><td>2. Think the whole equality thing is getting out of hand…</td></tr>
<tr><td rowspan="2">2. Stereotyping and dominance</td><td colspan="2">Expresses false ideas about women that suggest they are more suitable to fulfill certain roles[1], or inappropriate for certain tasks[2], or claims that men are somehow superior to women.</td></tr>
<tr><td colspan="2">Example: I feel like everytime I flirt with a girl they start to imagine all the ways they can utilize me.</td></tr>
<tr><td rowspan="4">3. Objectification</td><td colspan="2">1. Women depicted as objects, separate from their dignity and personal characteristics.</td></tr>
<tr><td colspan="2">2. Assumptions or descriptions of certain physical qualities that women must have in order to fulfill traditional gender roles, e.g., women should maintain a standard and ideal of beauty or attacks on a woman’s physical appearance.</td></tr>
<tr><td rowspan="2">Example:</td><td>1. I just want women for sex….</td></tr>
<tr><td>2. No offense but I’ve never seen an attractive african american hooker</td></tr>
<tr><td rowspan="2">4. Sexual violence</td><td colspan="2">Sexual suggestions, requests or harassment of a sexual nature</td></tr>
<tr><td colspan="2">Example: I wanna touch your tits..</td></tr>
<tr><td rowspan="2">5. Misogyny and non-sexual violence</td><td colspan="2">Non-sexual expressions of hatred and violence towards women.</td></tr>
<tr><td colspan="2">Example: Some woman are so toxic…</td></tr>
</table>

[1] Mother, wife, family caregiver, faithful, tender, loving, submissive, etc.
[2] Driving, hardwork, etc

2. EDOS shared tasks

Objective	Development of language models for sexism detection that are more accurate as well as explainable, with fine-grained classifications for sexist content
Language	English
Origin/Source	Social media platforms: 1. Gab 2. Reddit
No. entries	A pool of 1M entries created for each platform, from which 10,000 entries are sampled for labeling (Totally 20,000 entries).
Sub tasks	1. **TASK A - Binary Sexism Detection** 2. **TASK B - Category of Sexism.** 3. **TASK C - Fine-grained Vector of Sexism**
Data gathering process	1. **Gab**:34M publicly available **Gab posts** from August 2016 to October 2018 are collected, and 1M entries are randomly sampled to create the pool. 2. **Reddit**:A list of 81 subreddits likely containing sexist content is compiled, from which comments from August 2016 to October 2018 are collected **using the Reddit API**. These subreddits are categorized into four groups: Incels, Men Going Their Own Way, Men's Rights Activists, and Pick Up Artists. Sampling is restricted to 24 subreddits with at least 100k comments, forming a dataset of 42M comments. From this, 250k comments from each category are randomly sampled to create the final pool. 3. A high-quality dataset annotated by women experts, utilizing diverse sampling techniques. This is paired with a larger unlabelled dataset to optimize the balance between labelling costs and the effectiveness of trained systems.
Data Preparation and Sampling	1. **Data Cleaning:** The text from Gab and Reddit pools underwent cleaning, involving the replacement of URLs and usernames with generic tokens, removal of empty entries, entries containing only URLs or emoji, non-English entries, and duplicates. 2. **Boosted Data Sampling:** To counter class imbalance in random sampling results, a mix of community-based sampling on Reddit and various other methods is employed, avoiding user-based information for privacy reasons. 3. **Ensemble of Data Sampling Methods:** To mitigate biases and ensure diverse coverage of 11 fine-grained sexism vectors, 6 sampling techniques[3] are used on the cleaned Gab and Reddit totaling 20,000 entries.
The annotators	1. **Expert annotation** was chosen over crowdwork. 2. **19 highly-trained annotators**, having passed the screening process[4], were recruited. 3. To reduce implicit biases in labeling, only annotators who **self-identify as women** were recruited.

[3] The six sampling techniques include:
1. 1,000 entries featuring at least one sexist keyword.
2. 1,000 entries with a sexist and a topical keyword.
3. 1,000 entries distributed across toxicity deciles from the perspective model.
4. 1,000 entries from various deciles using a bespoke sexism detection classifier.
5. 1,000 entries where perspective's toxicity model scores differ significantly from our custom classifier.
6. 5,000 entries using a mix of topical keywords and other perspective attribute scores.

[4] All annotators passed a challenging 200-entry screening task that covered all 11 sexism vectors

Annotation Process	1. Each entry was labeled by three annotators. 2. Expert adjudication was used for disagreements, specifically for entries with less than unanimous 3/3 agreement in Task A and less than 2/3 agreement in Tasks B and C. 3. The expert team, provided labels for these cases. 4. Data was assigned to annotators in bi-weekly batches over two months. 5. Continuous collaboration with annotators allowed for feedback integration into guidelines and ongoing welfare monitoring.
Evaluation	To account for imbalance between classes, evaluated all systems with macro-average F1 score
Disagreement	All annotators were trained in multiple pilot tasks, and disagreements resolved by experts.

2.1. EDOS task A - Binary Sexism Detection

Description	For each entry, decide whether its primary label is **Sexist or Not Sexist**.
Sexist Category	Defined as abuse or negative sentiment directed towards women based on their gender or a combination of gender and other identity attributes. 1. Entries referring to a woman or women, including transgender women, or supporters of feminism. 2. Entries expressing negative sentiment based on gender, such as derogatory, threatening, or prejudicial comments. 3. The entry is labeled rather than the speaker, with adherence to specific criteria. 4. Quotes are to be taken at face value and jokes are to be evaluated for sexism, regardless of tone.
Not Sexist Category	It is determined whether entries contain abuse of protected characteristics other than gender.Confusing cases include: 1. Offensive language not specifically targeting women. 2. Abuse directed at individuals without a gender basis. 3. Abuse of protected characteristics other than gender. 4. Criticism of feminism as a theory is not labeled sexist; however, abuse of feminists is[5].

2.2. EDOS task B - Category of Sexism.

Description	For entries labeled as **Sexist**, each is assigned only **one**[6] Secondary Label.	
The categories	**1. Threats, plans to harm, and incitement**	Promoting harm or violence against women, including physical, sexual, and privacy threats.
	2. Derogation	Derogates or dehumanizes women, involving negative stereotypes and objectification.
	3. Animosity	Subtle or implicit sexist and stereotypes, sometimes appearing as benevolent sexism
	4. Prejudicial Discussions.	Denies gender discrimination and justifies sexism, including male victimhood narratives.

[5] A distinction is made between criticism of feminism and abuse of feminists; entries combining both are labeled as Sexist

[6] in cases with multiple applicable Secondary Labels, the most appropriate label is chosen. If uncertain, labels are selected based on their ranked order (i.e., 1.1 >> 1.2 >> ...)

2.3. EDOS task C - Fine-grained Vector of Sexism

Description	11 **Fine-Grained Sexism Vectors** are disaggregated into distinct categories, ensuring each vector is mutually exclusive and collectively exhaustive, covering all sexist content.	
The distinct categories	1) **1.1. Threats of harm**	Intent or desire to harm women, including physical, sexual, emotional, or privacy harm.
	2) **1.2. Incitement and encouragement of harm**	Incitement to harm women, rationalizing or justifying the act.
	3) **2.1. Descriptive attacks**	Derogatory characterizations of women, covering abilities, appearance, behavior, intellect, character, or morals.
	4) **2.2. Aggressive and emotive attacks**	Strong negative sentiment against women, through descriptions, accusations, or gendered slurs.
	5) **2.3. Dehumanising attacks and overt sexual objectification**	Comparing women to non-human entities or reducing them to sexual objects.
	6) **3.1. Causal use of gendered slurs, profanitie and insults**	Use of gendered slurs and insults, not necessarily as intentional attacks.
	7) **3.2. Immutable gender differences and gender stereotypes**	Assertions of inherent differences between genders, often used in sexist jokes.
	8) **3.3. Backhanded genderedcompliments**	Backhanded compliments to women, implying inferiority or reducing value to attractiveness.
	9) **3.4. Condescending explanations or unwelcome advice**	Unsolicited or patronizing advice to women on familiar topics.
	10) **4.1. Supporting mistreatment of individua women**	Support for individual mistreatment of women, including denial or justification.
	11) **4.2. Supporting systemic discrimination against women as a group**	Support for systemic discrimination against women, through denial or justification.

3. Le-Wi-Di shared tasks

Objective	To promote this approach to developing nlp models[7] by providing a unified framework for training and evaluating with such datasets.
Language	1. English 2. Arabic
Origin/Source	1. Twitter (tweets) 2. Conversations with AI systems
No. entries	12816 (tweets) 4050 (Conversations)
Sub Tasks	1. Misogyny and sexism detection 2. Abusiveness detection 3. offensiveness detectio
Sub Datasets	**1. MD-Agreement** **2. ConvAbuse** **3. HS-Brexit** **4. ArMIS**
Sampling and data gathering process	For each dataset is different
The annotators	1. Experts, 2. Specific demographic groups, 3. Amazon Mechanical Turk (AMT)-crowd
Annotation Process	For each dataset is different
Evaluation	Soft metrics were prioritized to evaluation, although both hard and soft evaluation metrics were employed.
Disagreement	It focuses on **subjective tasks**, instead of covering different types of disagreements

3.1. Le-Wi-Di 1st dataset- MD-Agreement[8]

Description	The Multi-Domain Agreement dataset (MD-Agreement), is a dataset of English tweets from three domains.
Language	English
Source of entries	Tweets
No. entries	10,753
Entries'Domain	1. Black Lives Matter (BLM) 2. Election2020 3. Covid-19
Task	Offensiveness detection
Annotators	The anonymized reference to the annotators that annotated the specific item. (**>800 different annotators**, via Amazon Mechanical Turk (AMT)-crowd)
No. Annotations	5

[7] it focuses entirely on nlp, instead of both nlp and computer vision tasks

[8] Leonardelli, E., Menini, S., Aprosio, A. P., Guerini, M., & Tonelli, S. (2021). Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. arXiv preprint arXiv:2109.13563.

Data Selection and Annotation Process	Annotation contains the disaggregated annotations from the crowd-annotators that annotatated this item (binary [0,1]). A set of domain-specific hashtags and keywords (e.g., #covid19, #election2020, #blm) was identified following an empirical analysis of online discussions. Tweets in English containing these keywords were gathered during specific periods using Twitter's public APIs. From this collection, a subset of tweets was randomly chosen and pre-processed. The processing included splitting hashtags into words using the Ekphrasis tool and replacing mentions of users and URLs.	
% of Agreement	Almost 30% of the dataset has then been annotated with a 2 vs 3 annotators disagreement, while almost another 30% of the dataset has an agreement of 1 vs 4 judgments.	
Evaluation	**hard_label**	**1 = offensive**, **0 = not offensive**. Assigned accordingly to the majority of annotations received in "annotations"
	soft_label 0	[0-1] Probability of label "0". The proportion of annotators that assigned 0 to the item.
	soft_label 1	[0-1] Probability of label "1". The proportion of annotators that assigned 1 to the item.

3.2. Le-Wi-Di 2nd dataset- ConvAbuse [9]

Description	Is a dataset of English dialogues conducted between users and two conversational agents.	
Language	English	
Source of entries	Conversations with AI systems	
No. entries	4,050	
Task	Abusive language detection	
Annotators	**8 experts** in gender studies	
No. Annotations	Varies across entries (min 3)	
Data Selection and Annotation Process	The user utterances have been annotated by experts in gender studies using a **hierarchical labelling scheme**, following categories: **1. Abuse binary** (0,1) **2. Abuse severity** (1,0,-1,-2,-3; 2) **3. Directedness** (explicit, implicit) **4. Target** (group, individual–system, individual–3rd party) **5. Type** (general, sexist, sexual harassment, homophobic, racist, transphobic, ableist, intellectual) Annotation contains the disaggregated annotations from the crowd-annotators. Comma-separated, range [-3,-2,-1, 0, 1]. From -3 to -1 is considered abusive, while 0 to 1 is not abusive.	
% of Agreement	Around 20% of the examples were found to be abusive	
Evaluation	**hard_label**	**1 = abusive, 0 = not abusive**. Assigned accordingly to the majority of annotations received. Note that, sometimes no majority existed. In this case, label has been assigned randomly (few cases).

[9] Curry, A. C., Abercrombie, G., & Rieser, V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. arXiv preprint arXiv:2109.09483.

	soft_label 0	[0-1] Probability of label "0". The proportion of annotators that considered the item abusive (0 or 1 in the field annotators)
	soft_label 1	[0-1] Probability of label "0". The proportion of annotators that considered the item abusive (-3 or -2 or -1 in the field annotators)

3.3. Le-Wi-Di 3rd dataset- HS-Brexit[10]

Description	Is a dataset of tweets on Abusive Language on Brexit	
Language	English	
Source of entries	Tweets	
No. entries	**1120**	
Task	**Offensiveness (hate speech) detection** in particular: **1.** Xenophobia and islamophobia **2.** Aggressiveness **3.** Offensiveness **4.** Sereotype	
Annotators	**1.** Target group of three Muslim immigrants in the UK (**three** annotators) **2.** Control group of three other individuals (**three** annotators)	
No. Annotators	**Six annotators** belonging to two distinct groups	
Data Selection and Annotation Process	Annotation contains the disaggregated annotations from the crowd-annotators (comma separted, binary [0,1]) The dataset, characterized by binary annotations, is skewed towards the negative class across all four dimensions, with positive class instances ranging from 7% in aggressiveness to 18% in offensiveness. Tweets where the target and control groups entirely disagreed often contained strongly connotated hashtags like #illegals and #rapists. In all cases of total disagreement, the presence of hate was indicated by the target group and its absence by the control group, with no instances where the control group indicated hate and the target group did not.	
Evaluation	**hard_label**	[0,1]. **0 = no HS**, **1 = HS**. Assigned accordingly to the majority of annotations received in "annotations". In case of no majority existance, this label has been assigned randomly (few cases).
	soft_label 0	[0-1] Probability of "0". Probability of label "0". The proportion of annotators that assigned 0 to the item.
	soft_label 1	[0-1] Probability of "1". Probability of label "0". The proportion of annotators that assigned 1 to the item.

[10] Akhtar, S., Basile, V., & Patti, V. (2020, October). Modeling annotator perspective and polarized opinions to improve hate speech detection. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 8, pp. 151-154).

3.4. Le-Wi-Di 4th dataset -ArMIS[11]

<table>
<tr><td>Description</td><td colspan="2">Is a dataset of Arabic tweets annotated for misogyny and sexism detection</td></tr>
<tr><td>Language</td><td colspan="2">Arabic</td></tr>
<tr><td>Source of entries</td><td colspan="2">Tweets</td></tr>
<tr><td>No. entries</td><td colspan="2">943</td></tr>
<tr><td>Task</td><td colspan="2">Misogyny and sexism detection, in particular, on where judges stand on the axis from conservative to liberal.</td></tr>
<tr><td>Annotators</td><td colspan="2">three people :
1. one self-identifying as a conservative male,
2. one moderate female,
3. one liberal female.</td></tr>
<tr><td>No. Annotators</td><td colspan="2">3 people</td></tr>
<tr><td>Data Selection and Annotation Process</td><td colspan="2">Tweets labeled for sexism using ami guidelines by Anzovino et al. (2018).</td></tr>
<tr><td rowspan="3">Evaluation</td><td>hard_label</td><td>[0,1]. 0 = not misogynistic/sexist, 1 = misogynistic/sexist.</td></tr>
<tr><td>soft_label 0</td><td>[0-1] Probability of "0". Probability of label "0". The proportion of annotators that assigned 0 to the item.</td></tr>
<tr><td>soft_label 1</td><td>[0-1] Probability of "1". Probability of label "0". The proportion of annotators that assigned 0 to the item.</td></tr>
</table>

3.5. Le-Wi-Di datasets Summary

Dataset used in Almanea and Poesio (2022) for:

- Comparing sexism detection models based on disagreement and soft-loss training (Uma et al., 2020).
- Comparing with models using the 'radical perspectivist' approach (Akhtar et al., 2021).

Dataset used in Uma et al. (2022) to analyze:

- Differences in subjective bias.
- Bias due to ambiguity.
- Bias due to noise.

[11] Almanea, D., & Poesio, M. (2022, June). ArMIS-the Arabic misogyny and sexism corpus with annotator subjective disagreements. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 2282-2291).