

EXIST 2023

sEXism Identification in Social neTworks

Task Guidelines

Jorge Carrillo-de-Albornoz^{1,3}, Laura Plaza^{1,3}, Enrique Amigó¹, Roser Morante¹, Julio Gonzalo¹, Paolo Rosso², Damiano Spina³



Source: unsplash

<http://nlp.uned.es/exist2023/>

¹ Universidad Nacional de Educación a Distancia

² Universidad Politécnica de Valencia

³ RMIT University

Task Description

Participants will be asked to classify “tweets” (in English and Spanish) according to the following three tasks:

TASK 1: Sexism Identification

The first subtask is a binary classification. The systems must decide whether a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour), and classify it according to two categories: **YES** and **NO**.

Examples of sexist tweets (“YES”) are:

- *“It’s less of #adaywithoutwomen and more of a day without feminists, which, to be quite honest, sounds lovely.”*
- *“I’m sorry but women cannot drive, call me sexist or whatever but it is true.”*
- *“You look like a whore in those pants” - My brother of 13 when he saw me in a leather pant.*

Examples of not sexist tweets (“NO”) are:

- *“Where all the white women at?”*
- *“The shocking video of a woman at the wheel who miraculously escapes an assassination attempt”*

TASK 2: Source Intention

For the tweets that have been classified as sexist, the second task aims to classify each tweet according to the intention of the person who wrote it. One of the three following categories must be assigned to each tweet:

- **DIRECT:** The intention is to write a message that is sexist itself, as in:
 - *“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:** The intention of the author is to report or describe a sexist situation or event suffered by a woman or women in first or third person, as in:
 - *“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”.*
- **JUDGEMENTAL** The intention of the author is to be judgemental since the tweet describes sexist situations or behaviors with the aim to condemning them.
 - *“As usual, the woman was the one quitting her job for the family’s welfare”.*

TASK 3: Sexism Categorization

For the tweets that have been classified as sexist, the third task aims to categorize them according to the type of sexism. We propose a five-class classification task: This is a multi-label task, so that more than one of the following labels may be assigned to each tweet:

- **IDEOLOGICAL-INEQUALITY:** the text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
 - *"#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."*
 - *"I think the whole equality thing is getting out of hand. We are different, thats how were made!"*
- **STEREOTYPING-DOMINANCE:** the text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hard work, etc.), or claims that men are somehow superior to women.
 - *"Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to."*
 - *"I feel like everytime I flirt with a girl they start to imagine all the ways they can utilize me"*
- **OBJECTIFICATION:** the text presents women as objects apart from their dignity and personal aspects or assumes or describes certain physical qualities that women must have to fulfill traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women's bodies at the disposal of men, etc.).
 - *"No offense but I've never seen an attractive african american hooker. Not a single one"*
 - *"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:** the text includes or describes sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault).
 - *"Wanna touch your tits..you can't imagine what I can do on your body."*
 - *"Fuck that cunt, I would with my fist"*
- **MISOGYNY-NON-SEXUAL-VIOLENCE:** the text expresses hatred and violence towards women, different to that with sexual connotations.
 - *"Domestic abuse is never okay.... Unless your wife is a bitch"*
 - *"Some women are so toxic they don't even know they are draining everyone around them in poison. If you lack self awareness you won't even notice how toxic you really are"*

More details and examples can be found at the EXIST 2023 website (<http://nlp.uned.es/exist2023/>).

Dataset Description

The EXIST 2023 dataset contains more than 10,000 labeled tweets, both in English and Spanish. In particular, the training set contains 6,920 tweets, the development set contains 1,038 tweets and the test set contains 2,076 tweets. Distribution between both languages has been balanced.

The data sets are provided in **JSON format**. Each tweet is represented as a JSON object with the following attributes:

1. **"id_EXIST"**: a unique identifier for the tweet.
2. **"lang"**: the languages of the text ("en" or "es").
3. **"tweet"**: the text of the tweet.
4. **"number_annotators"**: the number of persons that have annotated the tweet.
5. **"annotators"**: a unique identifier for each of the annotators.
6. **"gender_annotators"**: the gender of the different annotators. Possible values are: "F" and "M", for female and male respectively.
7. **"age_annotators"**: the age group of the different annotators. Possible values are: 18-22, 23-45, and 46+.
8. **"labels_task1"**: a set of labels (one for each of the annotators) that indicate if the tweet contains sexist expressions or refers to sexist behaviours or not. Possible values are: "YES" and "NO".
9. **"labels_task2"**: a set of labels (one for each of the annotators) recording the intention of the person who wrote the tweet. Possible labels are: "DIRECT", "REPORTED", "JUDGEMENTAL", "-", and "UNKNOWN".
10. **"labels_task3"**: a set of arrays of labels (one array for each of the annotators) indicating the type or types of sexism that are found in the tweet. Possible labels are: "IDEOLOGICAL-INEQUALITY", "STEREOTYPING-DOMINANCE", "OBJECTIFICATION", "SEXUAL-VIOLENCE", "MISOGYNY-NON-SEXUAL-VIOLENCE", "-", and "UNKNOWN".
11. **"split"**: subset within the dataset the tweet belongs to ("TRAIN", "DEV", "TEST" + "EN"/"ES").

IMPORTANT: Since labels for Tasks 2 and 3 are only assigned if the tweet has been labeled as sexist (label "YES" for Task 1), the label "-" is assigned to not sexist tweets in Tasks 2 and 3. The label "UNKNOWN" is assigned to tweets for which the annotators did not provide a label.

For the test set, only attributes **"id_EXIST"**, **"lang"**, **"tweet"** and **"split"** are provided. An example of the annotations of a tweet from the training set is given in the Figure 1.

```

"102442": {
  "id_EXIST": "102442",
  "lang": "es",
  "tweet": "@melixtitans Y donde esta la misoginia ahi? Justo a la gente meramente por
el reallity show y porque me caen mal adentro y ya jaja pero no ando de cotorra y
dando prioridades para hacer visible una causa solo cuando atacan a mis favs.",
  "number_annotators": 6,
  "annotators": ["Annotator_472", "Annotator_473", "Annotator_474", "Annotator_475",
  "Annotator_476", "Annotator_477"],
  "gender_annotators": ["F", "F", "F", "M", "M", "M"],
  "age_annotators": ["18-22", "23-45", "46+", "46+", "23-45", "18-22"],
  "labels_task1": ["YES", "NO", "YES", "NO", "YES", "YES"],
  "labels_task2": ["DIRECT", "-", "DIRECT", "-", "JUDGEMENTAL", "REPORTED"],
  "labels_task3": [
    ["STEREOTYPING-DOMINANCE", "MISOGYNY-NON-SEXUAL-VIOLENCE"],
    ["-"],
    ["IDEOLOGICAL-INEQUALITY"],
    ["-"],
    ["STEREOTYPING-DOMINANCE"],
    ["IDEOLOGICAL-INEQUALITY"]
  ],
  "split": "TRAIN_ES"
},

```

Figure 1 Example of the annotations for a tweet

Submission Format

The participants have the possibility to choose:

- The task or tasks they want to participate in.
- For each of the tasks, whether their system will provide:
 - **Hard labels:** A unique “hard” (single or multiple) label for each instance, as traditionally done in ML.
 - **Soft labels:** A probabilistic distribution over the different possible classes.

Each team is allowed to send up to **3 runs per task**. That is, each team is allowed to send up to **9 runs in total**.

Submitting Results for Task 1: Sexism Identification

Participants submitting results for Task 1 should format the runs in JSON format. Each tweet must be represented as a JSON object with the following attributes:

1. **"id_EXIST"**: The unique identifier for the tweet.
2. **"hard_label"**: Possible values are “YES” or “NO”.

3. **"soft_label"**: For each of the two possible labels ("YES" and "NO"), a probability is indicated. Note that the sum of the probabilities must be 1.0.

The following image shows an example of a run composed of two tweets.

```
{
  "100001": {
    "hard_label": "YES",
    "soft_label": {
      "YES": 0.7,
      "NO": 0.3
    }
  },
  "100324": {
    "hard_label": "NO",
    "soft_label": {
      "YES": 0.2,
      "NO": 0.8
    }
  }
}
```

Note that, if desired, the participants can provide only hard or soft labels, or both labels.

Submitting Results for Task 2: Source Intention

Participants submitting results for Task 2 should format the runs in JSON format. Each tweet must be represented as a JSON object with the following attributes:

1. **"id_EXIST"**: The unique identifier for the tweet.
2. **"hard_label"**: Possible values are "NO", "DIRECT", "REPORTED" or "JUDGEMENTAL".
3. **"soft_label"**: For each of the four possible labels ("NO", "DIRECT", "REPORTED" and "JUDGEMENTAL"), a probability is indicated. Note that the sum of the probabilities must be 1.0.

The following image shows an example of a run composed of two tweets.

```
{
  "100001": {
    "hard_label": "DIRECT",
    "soft_label": {
      "DIRECT": 0.5,
      "REPORTED": 0.3,
      "JUDGEMENTAL": 0.2,
      "NO": 0.0
    }
  },
  "100324": {
    "hard_label": "REPORTED",
    "soft_label": {
      "DIRECT": 0.2,
      "REPORTED": 0.6,
      "JUDGEMENTAL": 0.1,
      "NO": 0.1
    }
  }
}
```

Again, the participants can provide only hard or soft labels, or both labels.

Submitting Results for Task 3: Sexism Categorization

Participants submitting results for Task 3 should format the runs in JSON format. Each tweet must be represented as a JSON object with the following attributes:

1. **"id_EXIST"**: the unique identifier for the tweet.
2. **"hard_label"**: Possible values are "NO", "IDEOLOGICAL-INEQUALITY", "STEREOTYPING-DOMINANCE", "OBJECTIFICATION", "SEXUAL-VIOLENCE" and "MISOGYNY-NON-SEXUAL-VIOLENCE".
3. **"soft_label"**: For each of the six possible labels ("NO", "IDEOLOGICAL-INEQUALITY", "STEREOTYPING-DOMINANCE", "OBJECTIFICATION", "SEXUAL-VIOLENCE" and "MISOGYNY-NON-SEXUAL-VIOLENCE"), a probability is indicated. Note that, since this is a multi-label classification task, the sum of the probabilities does not have to be 1.0.

The following image shows an example of a run composed of two tweets.

```
{
  "100001": {
    "hard_label": ["STEREOTYPING-DOMINANCE", "OBJECTIFICATION"],
    "soft_label": {
      "NO": 0.1,
      "IDEOLOGICAL-INEQUALITY": 0.0,
      "STEREOTYPING-DOMINANCE": 0.2,
      "OBJECTIFICATION": 0.95,
      "SEXUAL-VIOLENCE": 0.87,
      "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.1
    }
  },
  "100324": {
    "hard_label": ["IDEOLOGICAL-INEQUALITY"],
    "soft_label": {
      "NO": 0.3,
      "IDEOLOGICAL-INEQUALITY": 0.76,
      "STEREOTYPING-DOMINANCE": 0.89,
      "OBJECTIFICATION": 0.2,
      "SEXUAL-VIOLENCE": 0.15,
      "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.0
    }
  }
}
```

Again, the participants can provide only hard or soft labels, or both labels.

How to Submit your Runs

Each team must pack all the runs in a directory named

`exist2023_<team_name>`

The directory will contain one file per run named

`<task>_<team_name>_<run_id>`

where `run_id` is a number between 1 and 3, and task may be *task1*, *task2* or *task3*.

For instance:

- `exist2023_UNED/task1_UNED_1`
- `exist2023_UNED/task2_UNED_3`

The (compressed) directory with your runs must be sent to icalbornoz@lsi.uned.es, lplaza@lsi.uned.es and enrique@lsi.uned.es together with a separate excel file containing metadata about your runs (i.e., brief description about techniques used, resources, etc.), and using the subject “EXIST2023@CLEF2023 - teamName”.

Evaluation

From the point of view of evaluation metrics, our three tasks can be described as:

- **Task 1 (sexism identification):** binary classification, mono label.
- **Task 2 (source intention):** multiclass hierarchical classification, mono label. The hierarchy of classes has a first level with sexist/not sexist, and a second level for the sexist category with three mutually-exclusive subcategories: direct/reported/judgemental. A suitable evaluation metric must reflect the fact that a confusion between not sexist and a sexist category is more severe than a confusion between two sexist subcategories.
- **Task 3 (sexism categorization):** multiclass hierarchical classification, multi label. Again the first level is a binary distinction between sexist/not sexist, and there is a second level for the sexist category that includes *ideological & inequality, stereotyping and dominance, objectification, sexual violence, misogyny* and *non-sexual violence*. These classes are not mutually exclusive: a tweet may belong to several subcategories at the same time.

The learning with disagreements paradigm can be considered in both sides of the evaluation process:

(i) The ground truth. In a “hard” setting, variability in the human annotations is reduced to a gold standard set of categories, **hard labels**, that are assigned to each item (e.g. using majority vote). In a “soft” setting, the gold standard is the full set of human annotations with their variability. Therefore, the evaluation metric incorporates the proportion of human annotators that have selected each category, **soft labels**. Note that in tasks 1 and 2, which are mono label problems, the sum of probabilities of each class must be one. But in task 3, which is multi label, each annotator may select more than one category for a single item. Therefore, the sum of probabilities of each class may be larger than one.

(ii) The system output. In a “hard”, traditional setting, the system predicts one or more categories for each item. In a “soft” setting, the system predicts a probability for each category, for each item. The evaluation score is maximized when the probabilities predicted match the actual probabilities in a soft ground truth. Again, note that in task 3, which is a multi label problem, the probabilities predicted by the system for each of the categories do not necessarily add up to one.

For each of the tasks, three types of evaluation will be reported:

1. *Hard-hard*: hard system output and hard ground truth.
2. *Hard-soft*: hard system output and soft ground truth.
3. *Soft-soft*: soft system output and soft ground truth.

OFFICIAL METRIC: ICM & ICM-soft

For all tasks and all types of evaluation (hard-hard, hard-soft and soft-soft) we will use the same official metric: ICM (Information Contrast Measure) (Amigó and Delgado, 2022). ICM is a similarity function that generalizes Pointwise Mutual Information (PMI), and can be used to

evaluate system outputs in classification problems by computing their similarity to the ground truth categories. The general definition of ICM is:

$$\text{ICM}(A, B) = \alpha_1 \text{IC}(A) + \alpha_2 \text{IC}(B) - \beta \text{IC}(A \cup B)$$

Where $\text{IC}(A)$ is the Information Content of the item represented by the set of features A , etc. ICM maps into PMI when all parameters take a value of 1.

In Amigó and Delgado, the general ICM definition is applied to cases where categories have a hierarchical structure and items may belong to more than one category. The resulting evaluation metric is proved to be analytically superior to the alternatives in the state of the art. The definition of ICM in this context is:

$$\text{ICM}(s(d), g(d)) = 2I(s(d)) + 2I(g(d)) - 3I(s(d) \cup g(d))$$

Where $I()$ stands for Information Content, $s(d)$ is the set of categories assigned to document d by system s , and $g(d)$ the set of categories assigned to document d in the gold standard.

As there is not, to the best of our knowledge, any current metric that fits hierarchical multi-label classification problems in a learning with disagreement scenario, we have defined an extension of ICM (*ICM-soft*) that accepts both soft system outputs and soft ground truth assignments. ICM-soft works as follows: first, we define the Information Content of a single assignment of a category c with an agreement v to a given item:

$$I(\{\langle c, v \rangle\}) = -\log_2(P(\{d \in D : g_c(d) \geq v\}))$$

Note that the information content of assigning a category c with an agreement v grows inversely with the probability of finding an item that receives category c with agreement equal or larger than v .

The system output and the gold standards are sets of assignments. Therefore, in order to estimate their information content, we apply a recursive function similar to the one described in (Amigó and Delgado, 2022):

$$\begin{aligned} I\left(\bigcup_{i=1}^n \{\langle c_i, v_i \rangle\}\right) &= I(\langle c_1, v_1 \rangle) + I\left(\bigcup_{i=2}^n \{\langle c_i, v_i \rangle\}\right) \\ &\quad - I\left(\bigcup_{i=2}^n \{\langle \text{lca}(c_1, c_i), \min(v_1, v_i) \rangle\}\right) \end{aligned}$$

where $\text{lca}(a, b)$ is the lowest common ancestor of categories a and b .

EVALUATION VARIANTS FOR EACH TASK

For each of the tasks, evaluation will be performed in the three modes described above, as follows:

- **Hard-hard evaluation.** For systems that provide a hard, conventional output, we will provide a hard-hard evaluation. To derive the hard labels in the ground truth from the different annotators' labels, we use a probabilistic threshold computed for each task. As a result, for task 1, the class annotated by more than 3 annotators is selected; for task 2, the class annotated by more than 2 annotators is selected; and for task 3 (multi-label), the annotated by more than 1 annotator are selected. Items for which there is no majority class (i.e. no class receives more probability than the threshold) will be removed from this evaluation scheme. **The official metric will be the original ICM** (as defined in (Amigó and Delgado, 2022)). We will also report and compare systems with F1 (the harmonic average of precision and recall). In task 1, we will use F1 for the positive class. In tasks 2 and 3, we will use the average of F1 for all classes. Note, however, that F1 is not ideal in our experimental setting: although it can handle multi-label situations, it does not take into account the relationships between classes. In particular, a mistake between not sexist and any of the sexist subclasses, and a mistake between two of the positive subclasses, are penalized equally, although the former is a more severe error.
- **Hard-soft evaluation.** For systems that provide a hard output we will also provide a hard-soft evaluation, comparing the categories assigned by the system with the probabilities assigned to each category in the ground truth. **We will use ICM-soft as the official evaluation metric in this variant.** The probabilities of the classes for each instance are calculated according to the distribution of labels and the number of annotators for that instance. It is important to notice that some instances are labeled as "UNKNOWN". In those cases, the number of annotators is decreased according to the number of "UNKNOWN" found for that instance. As the soft evaluation context is less restrictive all instances of the set are included. At this point only ICM-soft will be included in the evaluation script, although we may report additional metrics in the final report.
- **Soft-soft evaluation.** For systems that provide probabilities for each category, we will provide a soft-soft evaluation that compares the probabilities assigned by the system with the probabilities assigned by the set of human annotators. As in the previous case, **we will use ICM-soft as the official evaluation metric in this variant.** We may also report additional metrics in the final report.

Content of the evaluation package

The evaluation package is a python script, "exist2023evaluation.py", included in the folder "evaluation". The content of the folder evaluation is:

- **"exist2023evaluation.py"**: the official python script to evaluate all the EXIST 2023 tasks in all the evaluation contexts.
- **"golds"**: this folder includes the official gold standards for all tasks and evaluation contexts calculated as described above. In particular, the "hard_gold" allows participants

to evaluate their system's outputs in the hard-hard evaluation context, while the "soft_gold" must be used in the "hard-soft" and "soft-soft" evaluations.

- **"baselines"**: this folder includes the official baselines for each task, that can be employed in all the evaluation contexts. The "baseline_1" is basically a non-informative system where all instances are labeled with the majority class, while the "baseline_2" is also a non-informative system where all instances are classified as the majority positive class. Notice that these baselines are provided only as an example of system output, and not as a state-of-the-art approximation.

How to use official python evaluation package

In order to use the evaluation python module, participants must execute the following command in a prompt:

```
python exist2023evaluation.py
"-p baselines/EXIST2023_training_task1_baseline_1.json"
"-g golds/EXIST2023_training_task1_gold_soft.json"
"-e golds/EXIST2023_training_task1_gold_hard.json"
"-t task1"
```

Where the parameters are:

- -p: this parameter is mandatory, and indicates the **path to the system output** with the predictions that we want to evaluate. Notice that no information about the evaluation, hard or soft, must be provided. The script automatically will deal with this.
- -g: this parameter is mandatory, and indicates the **path to the gold standard** used in the evaluation. Notice that this gold standard can be the hard or soft gold standard if we only provide one gold standard, that is we do not use the optional parameter -e. If the user provides two gold standards, the -g parameter must indicate the path to the soft gold standard.
- -e: this parameter is optional and indicates the **path to the hard gold standard** used in the evaluation.
- -t: this parameter is mandatory and indicates the task addressed. Options are: "task1", "task2", "task3".

Questions

If you have any questions or problems, please open a thread on the Google Groups

<https://groups.google.com/g/exist2023atclef2023>

References

Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. (2021). **Overview of EXIST 2021: sEXism Identification in Social neTworks**. *Procesamiento del Lenguaje Natural* 67, 195-207.

Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, D., and Rosso, P. (2022). **Overview of EXIST 2022: sEXism Identification in Social neTworks**. *Procesamiento del Lenguaje Natural* 69, 229-240.

Laura Plaza, Jorge Carrillo de Albornoz, Roser Morante, Julio Gonzalo, Enrique Amigó, Damiano Spina, Paolo Rosso. (2023). **Overview of EXIST 2023: sEXism Identification in Social neTworks**. *Proceedings of ECIR'23*.

Enrique Amigó and Agustín Delgado. 2022. **Evaluating Extreme Hierarchical Multi-label Classification**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.