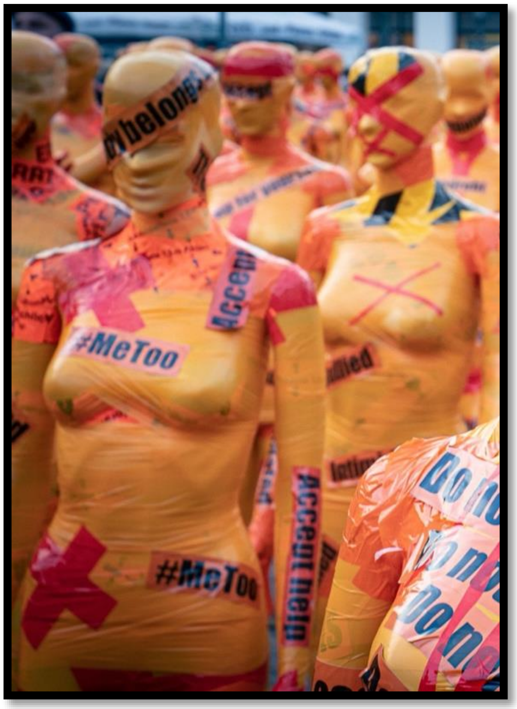
EXIST 2024

sEXism Identification in Social neTworks

Lab Guidelines

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



Source: unsplash

http://nlp.uned.es/exist2024/

[1] Universidad Nacional de Educación a Distancia
[2] Universitat Politècnica de València
[3] RMIT University

Task Description

Participants will be asked to classify either “tweets” (tasks 1 to 3) or “memes” (tasks 4 to 6), in English and Spanish, according to the following six tasks:

TASK 1: Sexism Identification in Tweets

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 in Tweets

The second subtask is a multi-class classification. For the tweets that have been predicted 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 sexist 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 in Tweets

The third subtask is a multi-label classification. For the tweets that have been predicted 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 <u>multi-label task</u>, 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!".*

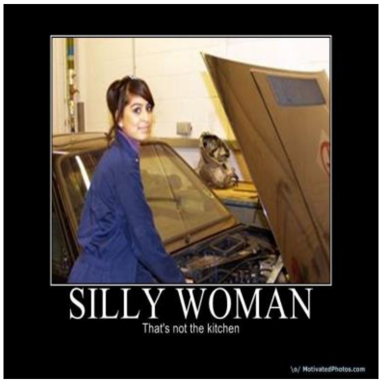
- **STEREOTYPING-DOMINANCE**: the text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hard work, etc.), or claims that men are somehow 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".*

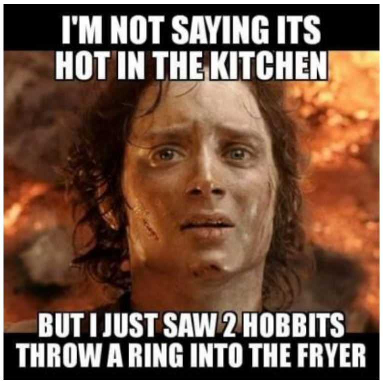
- **SEXUAL-VIOLENCE**: the text includes or describes sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault).
  - *"Wanna touch your tits..you can't imagine what I can do on your body."*
  - *"Fuck that cunt, I would with my fist".*

- **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".*

TASK 4: Sexism Identification in Memes
This is a binary classification task consisting of deciding whether or not a given meme is sexist. The following figures are some examples of sexist and not sexist memes, respectively.

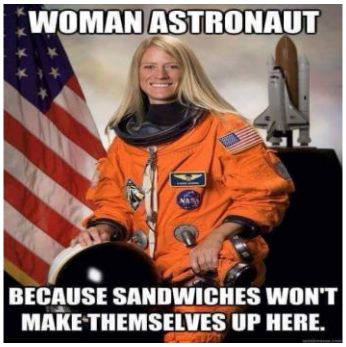


(a) Sexist

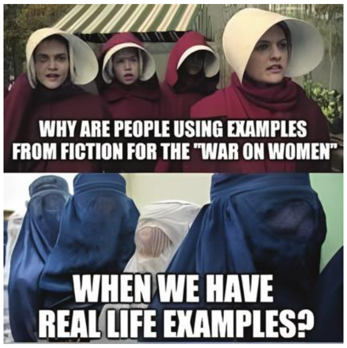

I'M NOT SAYING ITS HOT IN THE KITCHEN
BUT I JUST SAW 2 HOBBITS THROW A RING INTO THE FRYER

(b) Not sexist

TASK 5: Source Intention in Memes
As in task 2, this task aims to categorize the meme 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. Due to the characteristics of the memes, the **REPORTED** label is virtually null, so in this task systems should only classify memes with **DIRECT** or **JUDGEMENTAL** labels. The following figures are some examples of them, respectively.

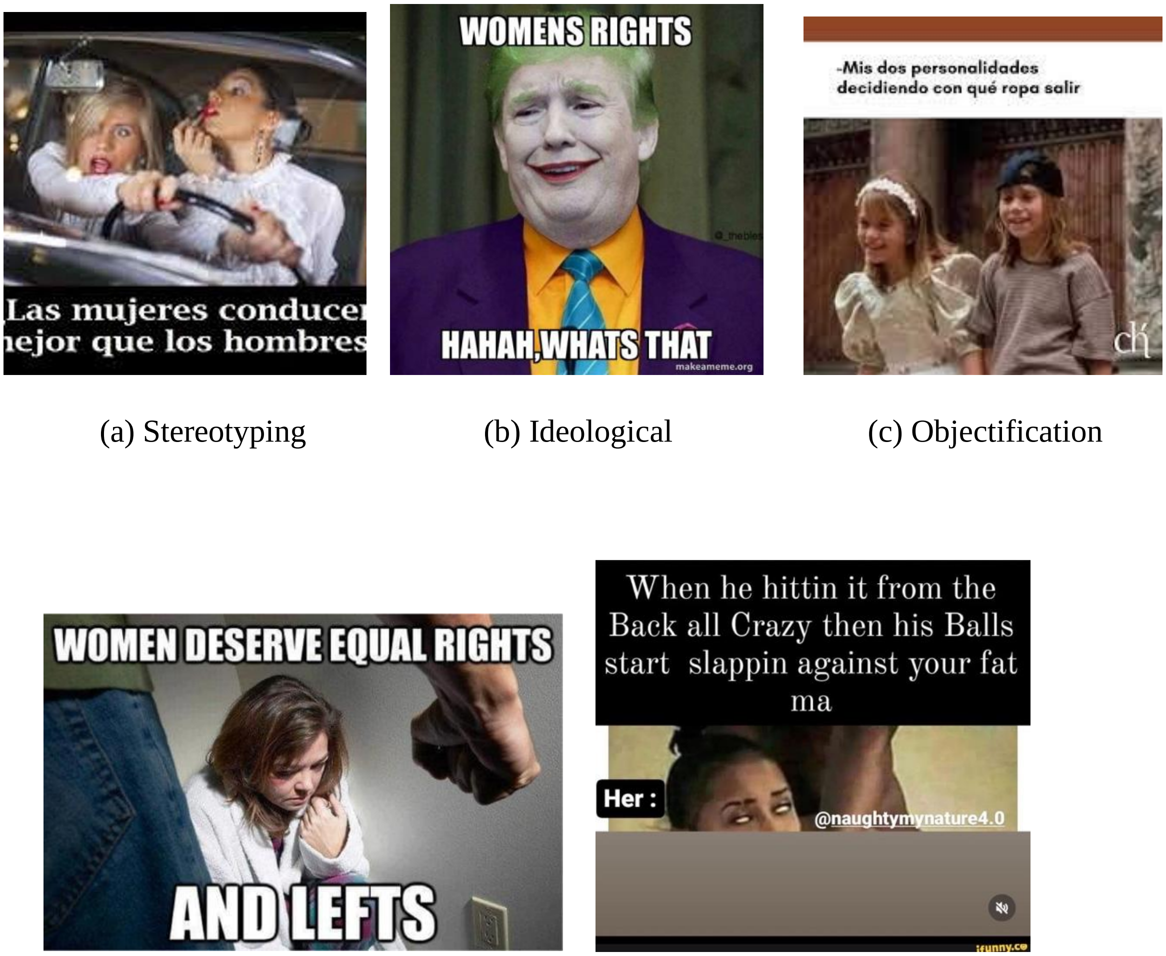


(a) Direct


WHY ARE PEOPLE USING EXAMPLES FROM FICTION FOR THE "WAR ON WOMEN"
WHEN WE HAVE REAL LIFE EXAMPLES?

(b) Judgemental

TASK 6: Sexism Categorization in Memes
This subtask is a multi-label classification. 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**. The following figures are some examples of categorized memes.


Las mujeres conduce
ejor que los hombres
WOMENS RIGHTS
HAHAH,WHATS THAT
makeameme.org
-Mis dos personalidades decidiendo con qué ropa salir
WOMEN DESERVE EQUAL RIGHTS
AND LEFTS
When he hittin it from the Back all Crazy then his Balls start slappin against your fat ma
Her :
@naughtymynature4.0
ifunny.co

(a) Stereotyping (b) Ideological (c) Objectification

(d) Misogyny (e) Sexual violence

More details and examples can be found at the EXIST 2024 website (http://nlp.uned.es/exist2024/).

Datasets Description

The EXIST 2024 Tweets Dataset will be employed in Tasks 1-3, while the EXIST 2024 Memes Dataset will be used in Tasks 4-6.

EXIST 2024 Tweets Dataset

The EXIST 2024 Tweets 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. **"ethnicity_annotators":** the self-reported ethnicity of the different annotators. Possible values are: “Black or African America”, “Hispano or Latino” , “White or Caucasian”, “Multiracial”, “Asian”, “Asian Indian” and “Middle Eastern”.
9. **"study_level_annotators":** the self-reported level of study achieved by the different annotators. Possible values are: “Less than high school diploma”, “High school degree or equivalent”, “Bachelor’s degree”, “Master’s degree” and “Doctorate”.
10. **"country_annotators":** the self-reported country where the different annotators live in.
11. **"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”.
12. **"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”.
13. **"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”.
14. **“split”:** subset within the dataset the tweet belongs to (“TRAIN”, “DEV”, “TEST” + “EN”/”ES”).

EXIST 2024 Memes Dataset

The EXIST 2024 Memes Dataset contains more than 5,000 labeled memes, both in English and Spanish. In particular, the training set contains 4,044 memes and the test set contains 1,053 memes. Distribution between both languages has been balanced.

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

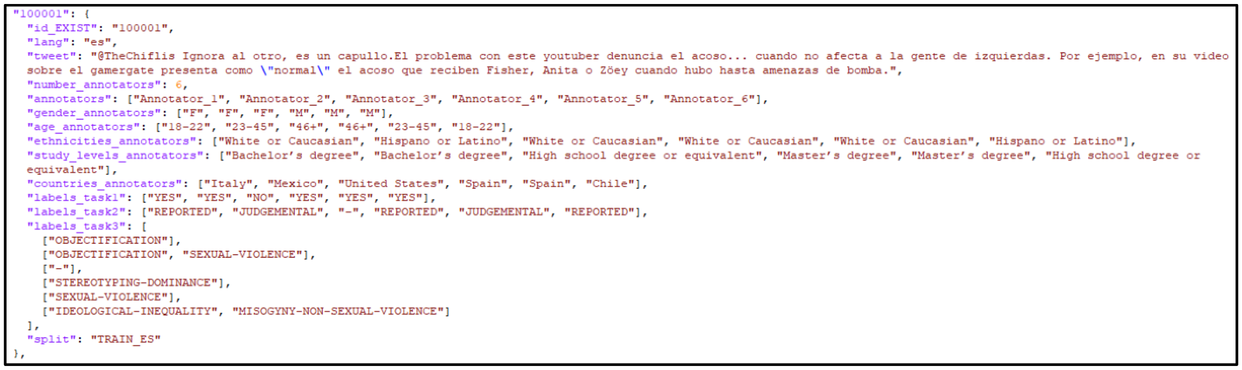
1. "**id_EXIST**": a unique identifier for the meme.
2. "**lang**": the languages of the meme (“en” or “es”).

3. "**text**": the text automatically extracted from the meme.
4. "**meme**": the name of the file that contains the meme.
5. "**path_memes**": the path to the file that contains the meme.
6. **"number_annotators":** the number of persons that have annotated the meme.
7. **"annotators":** a unique identifier for each of the annotators.
8. **"gender_annotators":** the gender of the different annotators. Possible values are: "F" and "M", for female and male respectively.
9. **"age_annotators":** the age group of the different annotators. Possible values are: 18-22, 23-45, and 46+.
10. **"ethnicity_annotators":** the self-reported ethnicity of the different annotators. Possible values are: "Black or African America", "Hispano or Latino" , "White or Caucasian", "Multiracial", "Asian", "Asian Indian" and "Middle Eastern".
11. **"study_level_annotators":** the self-reported level of study achieved by the different annotators. Possible values are: "Less than high school diploma", "High school degree or equivalent", "Bachelor's degree", "Master's degree" and "Doctorate".
12. **"country_annotators":** the self-reported country where the different annotators live in.
13. **"labels_task4":** a set of labels (one for each of the annotators) that indicate if the meme contains sexist expressions or refers to sexist behaviours or not. Possible values are: "YES" and "NO".
14. **"labels_task5":** a set of labels (one for each of the annotators) recording the intention of the person who created the meme. Possible labels are: "DIRECT", "JUDGEMENTAL", "", and "UNKNOWN".
15. **"labels_task6":** 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 meme. Possible labels are: "IDEOLOGICAL-INEQUALITY","STEREOTYPING-DOMINANCE","OBJECTIFICATION", "SEXUAL-VIOLENCE"**,** "MISOGYNY-NON-SEXUAL-VIOLENCE", "-", and "UNKNOWN".
16. **"split":** subset within the dataset the meme belongs to ("TRAIN-MEME", "TRAIN-MEME" + "EN"/"ES").

IMPORTANT: Since labels for Tasks 2, 3, 5 and 6 are only assigned if the tweet/meme has been labeled as sexist (label "YES" for Task 1 and 4), the label "-" is assigned to not sexist instances in these tasks. The label "UNKNOWN" is assigned to tweets/memes for which the annotators did not provide a label. Note that "UNKNOWN" is not a target class and therefore should not be predicted by the systems.

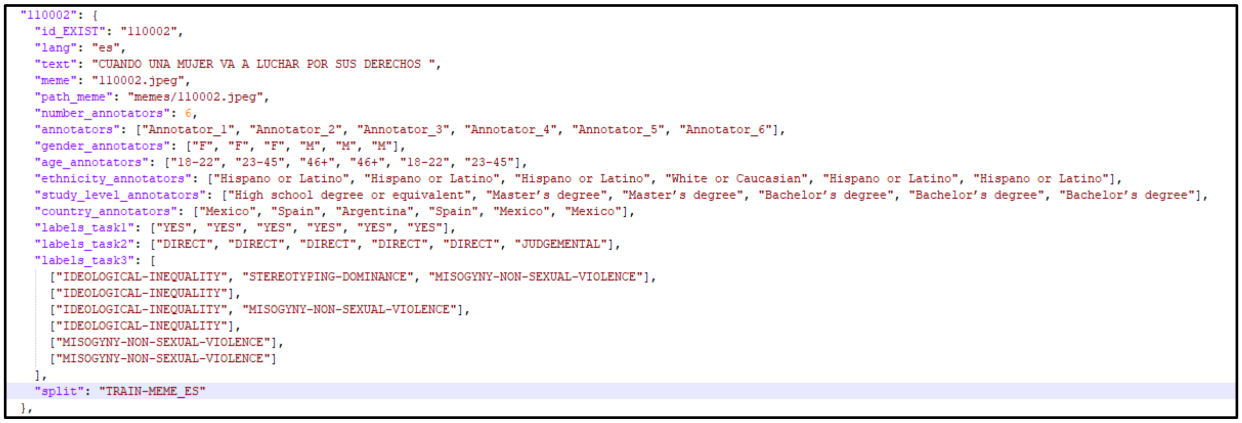
For the test set, labels for the different tasks are not provided.

Examples of the annotations of a tweet and a meme from the training sets are given in Figures 1 and 2.



```
"100001": {
  "id_EXIST": "100001",
  "lang": "es",
  "tweet": "@TheChiflis Ignora al otro, es un capullo.El problema con este youtuber denuncia el acoso... cuando no afecta a la gente de izquierdas. Por ejemplo, en su video
  sobre el gamergate presenta como \"normal\" el acoso que reciben Fisher, Anita o Zöey cuando hubo hasta amenazas de bomba.",
  "number_annotators": 6,
  "annotators": ["Annotator_1", "Annotator_2", "Annotator_3", "Annotator_4", "Annotator_5", "Annotator_6"],
  "gender_annotators": ["F", "F", "F", "M", "M", "M"],
  "age_annotators": ["18-22", "23-45", "46+", "46+", "23-45", "18-22"],
  "ethnicities_annotators": ["White or Caucasian", "Hispano or Latino", "White or Caucasian", "White or Caucasian", "White or Caucasian", "Hispano or Latino"],
  "study_levels_annotators": ["Bachelor’s degree", "Bachelor’s degree", "High school degree or equivalent", "Master’s degree", "Master’s degree", "High school degree or
  equivalent"],
  "countries_annotators": ["Italy", "Mexico", "United States", "Spain", "Spain", "Chile"],
  "labels_task1": ["YES", "YES", "NO", "YES", "YES", "YES"],
  "labels_task2": ["REPORTED", "JUDGEMENTAL", "-", "REPORTED", "JUDGEMENTAL", "REPORTED"],
  "labels_task3": [
    ["OBJECTIFICATION"],
    ["OBJECTIFICATION", "SEXUAL-VIOLENCE"],
    ["-"],
    ["STEREOTYPING-DOMINANCE"],
    ["SEXUAL-VIOLENCE"],
    ["IDEOLOGICAL-INEQUALITY", "MISOGYNY-NON-SEXUAL-VIOLENCE"]
  ],
  "split": "TRAIN_ES"
},
```

Figure 1 Example of the annotations for a tweet



```
"110002": {
  "id_EXIST": "110002",
  "lang": "es",
  "text": "CUANDO UNA MUJER VA A LUCHAR POR SUS DERECHOS ",
  "meme": "110002.jpeg",
  "path_meme": "memes/110002.jpeg",
  "number_annotators": 6,
  "annotators": ["Annotator_1", "Annotator_2", "Annotator_3", "Annotator_4", "Annotator_5", "Annotator_6"],
  "gender_annotators": ["F", "F", "F", "M", "M", "M"],
  "age_annotators": ["18-22", "23-45", "46+", "46+", "18-22", "23-45"],
  "ethnicity_annotators": ["Hispano or Latino", "Hispano or Latino", "Hispano or Latino", "White or Caucasian", "Hispano or Latino", "Hispano or Latino"],
  "study_level_annotators": ["High school degree or equivalent", "Master’s degree", "Master’s degree", "Bachelor’s degree", "Bachelor’s degree", "Bachelor’s degree"],
  "country_annotators": ["Mexico", "Spain", "Argentina", "Spain", "Mexico", "Mexico"],
  "labels_task1": ["YES", "YES", "YES", "YES", "YES", "YES"],
  "labels_task2": ["DIRECT", "DIRECT", "DIRECT", "DIRECT", "DIRECT", "JUDGEMENTAL"],
  "labels_task3": [
    ["IDEOLOGICAL-INEQUALITY", "STEREOTYPING-DOMINANCE", "MISOGYNY-NON-SEXUAL-VIOLENCE"],
    ["IDEOLOGICAL-INEQUALITY"],
    ["IDEOLOGICAL-INEQUALITY", "MISOGYNY-NON-SEXUAL-VIOLENCE"],
    ["IDEOLOGICAL-INEQUALITY"],
    ["MISOGYNY-NON-SEXUAL-VIOLENCE"],
    ["MISOGYNY-NON-SEXUAL-VIOLENCE"]
  ],
  "split": "TRAIN-MEME_ES"
},
```

Figure 2 Example of the annotations for a meme

Submission Format

This year, the assessment will be conducted utilizing the PyEvALL library, a specialized Python library for evaluating information systems. Access to the PyEvALL library can be obtained through code implementation or via a web interface known as EvALL. The evaluation web portal will be released from May with the test sets available for evaluation, both EXIST 2023 and EXIST 2024, so new proposals can be evaluated.

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.
  - **Soft labels:** A probabilistic distribution over the different possible classes.

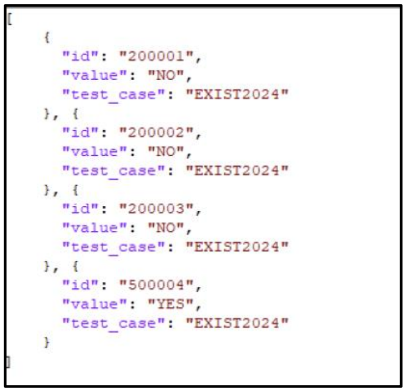
Each team is allowed to send up to **3 runs per task and type of output (i.e., hard vs soft)**. That is, each team is allowed to send up to **36 runs in total (3x2x6)**.

Submitting Results for Tasks 1 and 4: Sexism Identification in Tweets/Memes

Participants submitting results for tasks 1 and 4 must format the runs in JSON format. Each tweet/meme must be represented as a JSON object with the following attributes:

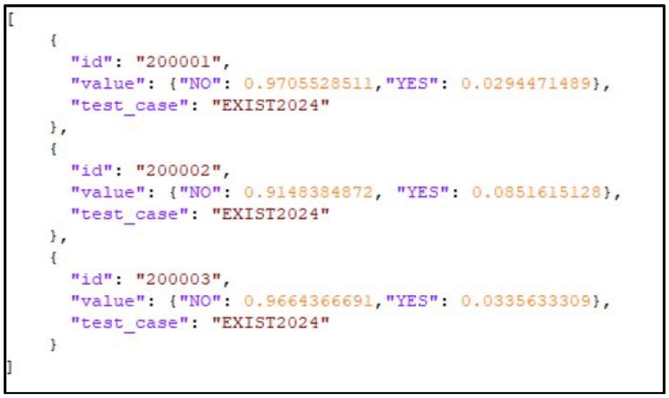
1. "**id**": The unique identifier for the tweet/meme.
2. **"value":** Possible values are:
    - “YES” or “NO”, if a hard output is submitted, or
    - A probability value for each of the two possible labels (“YES” and “NO”), if a soft output is submitted. Note that the sum of the probabilities must be 1.0.
3. **“test_case”:** “EXIST2024”

The following images show two examples for two runs (one submitting hard results and the other submitting soft results).



```
[
    {
      "id": "200001",
      "value": "NO",
      "test_case": "EXIST2024"
    }, {
      "id": "200002",
      "value": "NO",
      "test_case": "EXIST2024"
    }, {
      "id": "200003",
      "value": "NO",
      "test_case": "EXIST2024"
    }, {
      "id": "500004",
      "value": "YES",
      "test_case": "EXIST2024"
    }
]
```

(a) Run with hard outputs.



```
[
    {
      "id": "200001",
      "value": {"NO": 0.9705528511,"YES": 0.0294471489},
      "test_case": "EXIST2024"
    },
    {
      "id": "200002",
      "value": {"NO": 0.9148384872, "YES": 0.0851615128},
      "test_case": "EXIST2024"
    },
    {
      "id": "200003",
      "value": {"NO": 0.9664366691,"YES": 0.0335633309},
      "test_case": "EXIST2024"
    }
]
```

(b) Run with soft outputs

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

Submitting Results for Tasks 2 and 5: Source Intention in Tweets/Memes

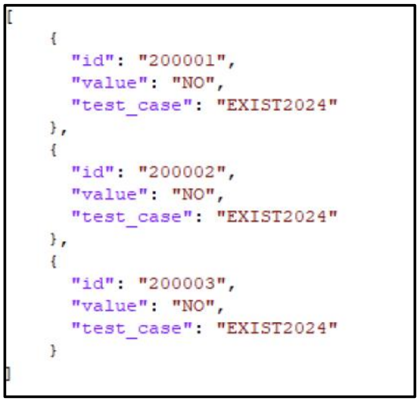
Participants submitting results for tasks 2 and 5 must format the runs in JSON format. Each tweet/meme must be represented as a JSON object with the following attributes:

1. "**id**": The unique identifier for the tweet/meme.
2. **"value":** Possible values are:

- “NO”, “DIRECT”, “REPORTED” or “JUDGEMENTAL”, for task 2; only “NO”, “DIRECT” or “JUDGEMENTAL”, for task 4.
- A probability value for each of the 4/3 possible labels, if a soft output is submitted. Note that the sum of the probabilities must be 1.0.

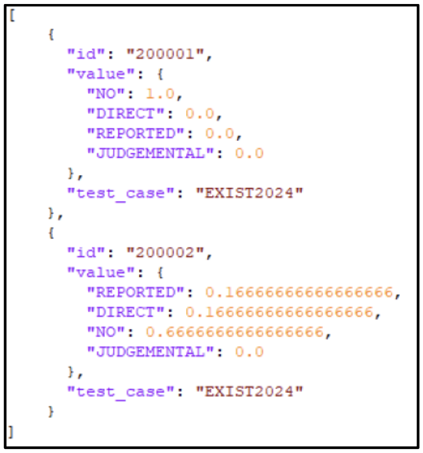
3. **“test_case”:** “EXIST2024”

The following images show two examples for two runs (one submitting hard results and the other submitting soft results).



```
[
    {
      "id": "200001",
      "value": "NO",
      "test_case": "EXIST2024"
    },
    {
      "id": "200002",
      "value": "NO",
      "test_case": "EXIST2024"
    },
    {
      "id": "200003",
      "value": "NO",
      "test_case": "EXIST2024"
    }
]
```

(a) Run with hard outputs.



```
[
    {
      "id": "200001",
      "value": {
        "NO": 1.0,
        "DIRECT": 0.0,
        "REPORTED": 0.0,
        "JUDGEMENTAL": 0.0
      },
      "test_case": "EXIST2024"
    },
    {
      "id": "200002",
      "value": {
        "REPORTED": 0.16666666666666666,
        "DIRECT": 0.16666666666666666,
        "NO": 0.6666666666666666,
        "JUDGEMENTAL": 0.0
      },
      "test_case": "EXIST2024"
    }
]
```

(b) Run with soft outputs

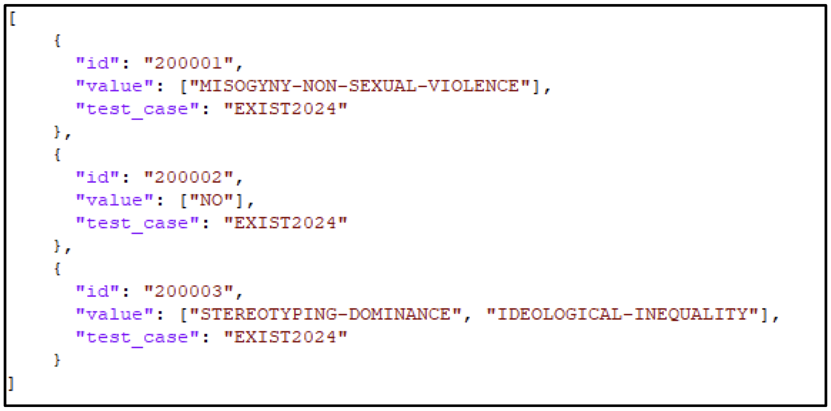
Again, the participants can provide only hard or soft labels, or both labels.

Submitting Results for Tasks 3 and 6: Sexism Categorization in Tweets/Memes

Participants submitting results for tasks 3 and 6 must format the runs in JSON format. Each tweet/meme must be represented as a JSON object with the following attributes:

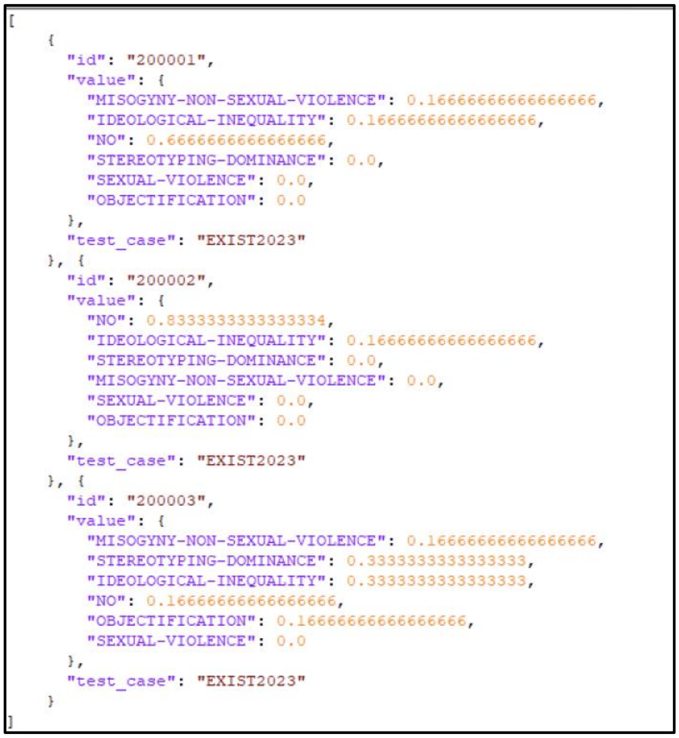
1. "**id**": the unique identifier for the tweet/meme.
2. **"hard_label":** Possible values are:
   - “NO”, ““IDEOLOGICAL-INEQUALITY”, “STEREOTYPING-DOMINANCE”, “OBJECTIFICATION”, “SEXUAL-VIOLENCE” and “MISOGYNY-NON-SEXUAL-VIOLENCE”.
   - A probability value for each of the six possible labels, if a soft output is submitted. Note that, since this is a multi-label classification task, the sum of the probabilities does not have to be 1.0.
3. **“test_case”:** “EXIST2024”

The following images show two examples for two runs (one submitting hard results and the other submitting soft results).



```
[
    {
      "id": "200001",
      "value": ["MISOGYNY-NON-SEXUAL-VIOLENCE"],
      "test_case": "EXIST2024"
    },
    {
      "id": "200002",
      "value": ["NO"],
      "test_case": "EXIST2024"
    },
    {
      "id": "200003",
      "value": ["STEREOTYPING-DOMINANCE", "IDEOLOGICAL-INEQUALITY"],
      "test_case": "EXIST2024"
    }
]
```

(a) Run with hard outputs.



```
[
    {
      "id": "200001",
      "value": {
        "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.16666666666666666,
        "IDEOLOGICAL-INEQUALITY": 0.16666666666666666,
        "NO": 0.6666666666666666,
        "STEREOTYPING-DOMINANCE": 0.0,
        "SEXUAL-VIOLENCE": 0.0,
        "OBJECTIFICATION": 0.0
      },
      "test_case": "EXIST2023"
    }, {
      "id": "200002",
      "value": {
        "NO": 0.8333333333333334,
        "IDEOLOGICAL-INEQUALITY": 0.16666666666666666,
        "STEREOTYPING-DOMINANCE": 0.0,
        "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.0,
        "SEXUAL-VIOLENCE": 0.0,
        "OBJECTIFICATION": 0.0
      },
      "test_case": "EXIST2023"
    }, {
      "id": "200003",
      "value": {
        "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.16666666666666666,
        "STEREOTYPING-DOMINANCE": 0.3333333333333333,
        "IDEOLOGICAL-INEQUALITY": 0.3333333333333333,
        "NO": 0.16666666666666666,
        "OBJECTIFICATION": 0.16666666666666666,
        "SEXUAL-VIOLENCE": 0.0
      },
      "test_case": "EXIST2023"
    }
]
```

(b) Run with soft outputs.

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

```
exist2024_<team_name>
```

The directory will contain one file per run named

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

where run_id is a number between 1 and 3, evaluation context can be *hard* or *soft*, and task may be *task1, task2, task3, task4, task5 or task6*.

For instance:

- exist2024_UNED/task1_hard_UNED_1
- exist2024_UNED/task2_soft_UNED_3

The (compressed) directory with your runs must be submitted to the competition by filling up the following form:
https://docs.google.com/forms/d/e/1FAIpQLSc7riSBK4En-DqBTb0YCZSy4AlD3h9haXydRdHoaDM-iCv6KA/viewform

Notice that only one submission per team is allowed.

Evaluation

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

- **Tasks 1 and 4 (sexism identification)**: binary classification, mono label.
- **Tasks 2 and 5 (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/two mutually exclusive subcategories: *direct*, *reported* and *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.
- **Tasks 3 and 6 (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/meme 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, 2, 4 and 5, which are mono label problems, the sum of the probabilities of each class must be one. But in tasks 3 and 6, which are multi label, each annotator may select more than one category for a single item. Therefore, the sum of the 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 tasks 3 and 6, 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, two types of evaluation will be reported:

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

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

Where 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:

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

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):

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

where lca(a,b) is the lowest common ancestor of categories a and b.

EVALUATION VARIANTS FOR EACH TASK

For each of the tasks, the evaluation will be performed in the two 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 tasks 1 and 4, the class annotated by more than 3 annotators is selected; for tasks 2 and 5, the class annotated by more than 2 annotators is selected; and for tasks 3 and 6 (multi-label), the class 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 tasks 1 and 4, we will use F1 for the positive class. In the remaining tasks, 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 consider the relationships between classes: 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.

- **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.

Description of the PyEvALL Evaluation Library

This year we will use the PyEvALL evaluation library, which includes all the metrics used in the EXIST 2024 competition. PyEvALL is available at https://github.com/UNEDLENAR/PyEvALL. PyEvALL (The Python library to Evaluate ALL) is an evaluation tool for information systems that allows assessing a wide range of metrics covering various evaluation contexts, including classification, ranking, or LeWiDi (Learning with disagreement). The documentation included in the PyEvALL Readme explains how to install it (via pip or source code), how to use it, and the input and output formats (the different PyEvALL reports).

How to use the official PyEvALL python evaluation package

PyEvALL allows to evaluate several metrics from different evaluation contexts using only 8 lines of code. As an example, the following code will execute an evaluation over a prediction file for the first EXIST task, and will show the results in an embedded report:

```
predictions = "test/hard/EXIST2024_test_task1_baseline_2.json"
gold = "test/hard/EXIST2024_test_task1_gold_hard.json"
test = PyEvALLEvaluation()
params= dict()
params[PyEvALLUtils.PARAM_REPORT]= PyEvALLUtils.PARAM_OPTION_REPORT_EMBEDDED
metrics=["ICM", "ICMNorm" ,"FMeasure"]
report= test.evaluate(predictions, gold, metrics, **params)
report.print_report()
```

Similarly, PyEvALL allows to evaluate several prediction files and generate a meta report. As an example, the following code will execute an evaluation over different prediction files for the second EXIST task, and will show the results in a table in tsv format:

```
predictions = "test/hard/"
lst_pred=[]
onlyfiles = [f for f in listdir(predictions) if isfile(join(predictions, f))]
for file in onlyfiles:
     lst_pred.append(predictions+ file)
```

```
gold = "test/hard/EXIST2024_test_task2_gold_hard.json"
TASK_2_HIERARCHY = {"YES":["DIRECT","REPORTED","JUDGEMENTAL"], "NO":[]}
params= dict()
prueba = PyEvALLEvaluation()
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_2_HIERARCHY
params[PyEvALLUtils.PARAM_REPORT]= PyEvALLUtils.PARAM_OPTION_REPORT_DATAFRAME
metrics=["ICM", "ICMNorm" ,"FMeasure"]
report= prueba.evaluate_lst(lst_pred, gold, metrics, **params)
report.print_report_tsv()
```

Notice that, as task 2 is a hierarchical classification problem, the parameter PyEvALLUtils.PARAM_HIERARCHY has been set with the appropriate hierarchy.

The evaluation of all tasks in the EXIST 2024 challenge can be done using the code previously described by changing the configuration, metrics and hierarchy, according to the different context evaluations:

Hard-hard task 1 and 4: in this configuration, there is no hierarchy and the measures used are ICM, ICM Normalized and F1.

```
metrics=["ICM", "ICMNorm" ,"FMeasure"]
```

Hard-hard task 2: in this configuration, the hierarchy parameter should be set as well as the measures ICM, ICM Normalized and F1.

```
metrics=["ICM", "ICMNorm" ,"FMeasure"]
TASK_2_HIERARCHY = {"YES":["DIRECT","REPORTED","JUDGEMENTAL"], "NO":[]}
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_2_HIERARCHY
```

Hard-hard task 5: in this configuration, the hierarchy parameter should be set as well as the measures ICM, ICM Normalized and F1.

```
metrics=["ICM", "ICMNorm" ,"FMeasure"]
TASK_5_HIERARCHY = {"YES":["DIRECT, "JUDGEMENTAL"], "NO":[]}
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_5_HIERARCHY
```

Hard-hard task 3 and 6: in this configuration, the hierarchy parameter should be set as well as the measures ICM, ICM Normalized and F1.

```
metrics=["ICM", "ICMNorm" ,"FMeasure"]
TASK_3_HIERARCHY = {"YES":["IDEOLOGICAL-INEQUALITY","STEREOTYPING-
DOMINANCE","OBJECTIFICATION", "SEXUAL-VIOLENCE", "MISOGYNY-NON-SEXUAL-
VIOLENCE"], "NO":[]}
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_3_HIERARCHY
```

Soft-soft task 1 and 4: in this configuration, there is no hierarchy and the measures used are ICM-soft, ICM-soft Normalized and F1.

```
metrics=["ICMSoft", "ICMSoftNorm", "CrossEntropy"]
```

Soft-soft task 2: in this configuration the hierarchy parameter should be set as well as the measures ICM-soft, ICM-soft Normalized and F1.

```
metrics=["ICMSoft", "ICMSoftNorm", "CrossEntropy"]
```

```
TASK_2_HIERARCHY = {"YES":["DIRECT","REPORTED","JUDGEMENTAL"], "NO":[]}
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_2_HIERARCHY
```

Soft-soft task 5: in this configuration the hierarchy parameter should be set as well as the measures ICM-soft, ICM-soft Normalized and F1.

```
metrics=["ICMSoft", "ICMSoftNorm", "CrossEntropy"]
TASK_5_HIERARCHY = {"YES":["DIRECT, "JUDGEMENTAL"], "NO":[]}
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_5_HIERARCHY
```

Soft-soft task 3 and 6: in this configuration the hierarchy parameter should be set as well as the measures ICM-soft, ICM-soft Normalized and F1.

```
metrics=["ICMSoft", "ICMSoftNorm"]
TASK_3_HIERARCHY = {"YES":["IDEOLOGICAL-INEQUALITY","STEREOTYPING-
DOMINANCE","OBJECTIFICATION", "SEXUAL-VIOLENCE", "MISOGYNY-NON-SEXUAL-
VIOLENCE"], "NO":[]}
params[PyEvALLUtils.PARAM_HIERARCHY]= TASK_3_HIERARCHY
```

Content of the evaluation folder

The content of the folder evaluation is:

- **“golds”**: this folder includes the official gold standards for all tasks and evaluation contexts. In particular, the “hard_gold” allows participants to evaluate their system outputs in a hard-hard evaluation context, while the “soft_gold” must be used in the soft-soft evaluations.
- **“baselines":** this folder includes the official baselines for each task. Notice that, with only one file, all evaluation contexts can be addressed. The “majority_class” baseline is basically a non-informative system where all instances are labeled with the majority class, while the “minority_class” is a non-informative system where all instances are classified as the minority class. Notice that these baselines are provided only as an example of system output, and not as a state-of-the-art approximation.

Questions

If you have any questions or problems, please open a thread on the Google Groups
https://groups.google.com/g/exist2024atclef2024

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.