EXIST 2022

sEXism Identification in Social neTworks

Task Guidelines

Francisco Rodriguez-Sanchez[1], Jorge Carrillo-de-Albornoz[1], Laura Plaza[1], Julio Gonzalo[1], Paolo Rosso[2], Damiano Spina [3], Adrián Mendieta-Aragón[1], Guillermo Marco[1], María Plaza and Maryna Makeienko[1]

http://nlp.uned.es/exist2022/

[1] Universidad Nacional de Educación a Distancia
[2] Universidad Politécnica de Valencia
[3] Royal Melbourne Institute of Technology

Task Description

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

TASK 1: Sexism Identification

The first subtask is a binary classification. The systems have to decide whether or not a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour), and classifies it according to two categories: “sexist” and “non-sexist”.

TASK 2: Sexism Categorization

Once a message has been classified as sexist, the second task aims to categorize the message according to the type of sexism. In particular, we propose a five-classification task: “ideological-inequality”, “stereotyping-dominance”, “objectification”, “sexual-violence” and “misogyny-non-sexual-violence”.

- **IDEOLOGICAL AND INEQUALITY**: the text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
- **STEREOTYPING AND 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.
- **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 in order to fulfill traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women’s bodies at the disposal of men, etc.).
- **SEXUAL VIOLENCE**: the text includes or describes sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault).
- **MISOGYNY AND NON-SEXUAL VIOLENCE**: the text expresses hatred and violence towards women.

More details and examples can be found at the EXIST 2022 website (http://nlp.uned.es/exist2022/).

Dataset Description

In this new edition of EXIST 2022 challenge, we will use the EXIST 2021 dataset as training data. The entire EXIST 2021 dataset contains 11,345 labeled texts, tweets and gabs, both in English and Spanish. In particular, the EXIST 2021 training set contains 6977 tweets while the test set contains 3386 tweets and 982 gabs. Distribution between both languages has been balanced.

More details about the dataset crawling, generation and labeling are available in the task overview of EXIST 2021 (bias consideration, annotation process, quality experiments, inter-annotator agreement, etc.: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389).

For the test set, we will collect and label around 1000 tweets from Twitter following the procedure used in the EXIST 2021 dataset. More details are available in the dataset section at the EXIST 2022 website (http://nlp.uned.es/exist2022/).

The training and test sets are provided in a **tsv format**. In particular, the training dataset contains the following columns:

- "**Test_case**": tag needed in the EvALL framework for evaluating classification tasks. In EXIST 2021, this tag is set to "EXIST2021".
- "**Id**": denotes a unique identifier of the tweet or gab.
- "**Source**": defines the social network where the text was crawled, "twitter" or "gab".
- "**Language**": denotes the languages of the text ("en" or "es").
- "**Text**": represents the text of the tweet or gab.
- "**Task1**": indicates if the tweet or gab contains sexist expressions ("**sexist**") or not ("**non-sexist**").
- "**Task2**": categorize the message according to the type of sexism. Possible categories are: "**ideological-inequality**", "**stereotyping-dominance**", "**objectification**", "**sexual-violence**" and "**misogyny-non-sexual-violence**".

test_case	id	source	language	text	task1	task2
EXIST2021	000001	twitter	en	@RebelliousMinx_ Oh , Wonderful . Kay is gonna be...	sexist	sexual-violence
EXIST2021	000002	twitter	en	I understand all these, but what am saying is, if a woman...	non-sexist	non-sexist
EXIST2021	000003	twitter	en	v grateful to have a prof who is not just allowing but actively...	non-sexist	non-sexist

And the test dataset contains the following columns:

- "**Test_case**": tag needed in the EvALL framework for evaluating classification tasks. In EXIST 2022, this tag is set to "EXIST2022".
- "**Id**": denotes a unique identifier of the tweet.
- "**Source**": defines the social network where the text was crawled: "twitter".
- "**Language**": denotes the languages of the text ("en" or "es").
- "**Text**": represents the text of the tweet.

Submission format

Results for both tasks should be submitted in a plain text following the EvALL format for classification tasks (www.evall.uned.es) (Amigó et al., 2017). In particular, the classification task uses as input a 3 column tsv format **without headers**, where the first column represents the **TEST CASE**, the second column represents the **ID** of the item and the third column represents the **CATEGORY** assigned to the item, i.e. the category assigned for task1 and task2.

Participants submitting results for the first task, Task1, should format the runs as in the next figure, where correct values for the *category* column are "**sexist**" and "**non-sexist**".

```
EXIST2022   006978  non-sexist
EXIST2022   006979  sexist
EXIST2022   006980  sexist
```

Specifically, submitted runs must contain one tweet per line, with the information shown in the Figure above. As you can see in the image, the tag for the *test_case* in the EXIST 2022 shared tasks is **EXIST2022**, while the *id* represents a single tweet and should match exactly with the ids included in the test set. Finally, the *category* column represents the categories for the tasks.

Similarly, participants submitting results for the second task, Task2, should format the results as in the following Figure, where correct values for the *category* column are:

- **non-sexist**: if tweet does not contain sexist content.
- **ideological-inequality**: if the tweet is classified as sexist and categorized as idiolical-inequality.
- **stereotyping-dominance**: if the tweet is classified as sexist and categorized as stereotyping-dominant.
- **objectification**: if the tweet is classified as sexist and categorized as objectification.
- **sexual-violence**: if the tweet is classified as sexist and categorized as sexual-violence.
- **misogyny-non-sexual-violence:** if the tweet is classified as sexist and categorized as misogyny-non-sexual-violence.

```
EXIST2022   006978  non-sexist
EXIST2022   006979  exual-violence
EXIST2022   006980  ideological-inequality
EXIST2022   006981  stereotyping-dominance
EXIST2022   006982  non-sexist
```

IMPORTANT: Each line should NOT include the tweet's text in your submission.

Each team is allowed to send up to **3 runs per task**: Task1 and Task2. That is, each team is allowed to send up to **6 runs in total**.

How to submit your runs

Each team must pick up all runs and pack them in a directory named

exist2022_<team_name>

The directory will contain one file per run named

<tasks>_<team_name>_<run_id>

where run_id is a number between 1 and 3 and tasks may be *task1* or *task2*.

For instance:

- exist2022_UNED/task1_UNED_1.
- exist2022_UNED/task2_UNED_3.

The (compressed) directory with your runs must be sent to jcalbornoz@lsi.uned.es, lplaza@lsi.uned.es and frodriguez.sanchez@invi.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 "EXIST2022@IberLEF2022 - teamName".

Evaluation

In order to evaluate the performance of the different approaches proposed by the participants we will use the Evaluation Framework EvALL (www.evall.uned.es). Within this framework, we will evaluate the system outputs as classification tasks (binary and multiclass respectively) with the following measures: Accuracy, Precision, Recall and F-measure (using macro average with all classes for the three last).

In the first task, results of participants will be ranking using **accuracy**. Besides, other measures will be computed, such as F-measure, Precision and Recall. For the second task, we will use **F-measure** to rank the system outputs, analyzing the results according to the different categories and distributions. Similarly, we will compute other measures such as Precision and Recall, as well as modified version of CEM measure (Closeness Evaluation Measure) (Amigo et al., 2020).

The use of EvALL provides us with a robust, reliable and well tested evaluation framework, that allows the homogenous evaluation and comparison of the participant systems (Amigo et al., 2017; Amigo et al., 2018, Amigo et al., 2020). EvALL allows participants to evaluate their approaches under the same conditions and using the same data. Moreover, EvALL provides different formats for output systems, as well as different wrappers.

Questions

If you have any questions or problems, please open a thread on the Google Groups https://groups.google.com/g/exist2022atiberlef2022.

References

Amigó, E., Carrillo-de-Albornoz, J., Almagro-Cádiz, M., Gonzalo, J., Rodríguez-Vidal, J., and Verdejo, F. (2017). **EvALL: Open Access Evaluation for Information Access Systems**. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*.

Amigó, E., Spina, D., and Carrillo-de-Albornoz, J.. **An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric**. *In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18). ACM, New York, NY, USA, 625-634.*

Amigo, E., Gonzalo, J., Mizzaro, S., and Carrillo-de-Albornoz, J.. **An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results.** *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020).*