EVALITA 2023

Homotransphobia Detection in Italian (HODI)

Task Guidelines

Debora Nozza[1], Greta Damo[1], Alessandra Teresa Cignarella[2],

Tommaso Caselli[3], Viviana Patti[2]

[1] Bocconi University, Milan
[2] University of Turin, Turin
[3] University of Groningen, Groningen

TABLE OF CONTENTS

1 Task description

The HODI shared task will focus on identifying homotransphobia in Italian tweets. HODI is organized according to two main subtasks:

- **Subtask A - Homotransphobia detection**: the objective is to detect if a text is homotransphobic or not.
- **Subtask B - Explainability**: the objective is to extract the rationales of the classification models trained for Subtask A. **Participation to subtask B is conditional upon participation to subtask A.**

2 Description of the dataset

The HODI dataset includes messages directed at minority groups who are frequently targets of homotransphobia. We collected tweets in a specific time period from May 1, 2022 to August 31, 2022, using a set of 21 keywords (that will be made available at the end of the evaluation window). The dataset is divided into two sets: training and testing.

2.1 Training data - Subtask A

The training dataset for Subtask A is a dataset of tweets manually labeled according to:

- **Homotransphobia**: Homotransphobic vs. Not Homotransphobic

The datasets are provided as **TSV** files (**tab-separated** files) and report the following fields:

"**id**" "**text**" "**homotransphobic**"

where:

- **id** denotes a unique identifier of the tweet.
- **text** represents the tweet text.
- **homotransphobic** defines if a tweet is homotransphobic or not homotransphobic; it takes values:
  - **0** if the tweet is not homotransphobic;
  - **1** if the tweet is homotransphobic.

2.2 Training data - Subtask B

The training dataset for Subtask B contains tweets manually labeled according to:

- **Rationales**: the span of text in which words considered homotransphobic are highlighted.

The datasets are provided as **TSV** files (**tab-separated** files) and report the following fields:

"**id**" "**text**" "**rationales**"

where:

- **id** denotes a unique identifier of the template-generated text.
- **text** represents the template-generated text.
- **rationales** denote the textual span that is considered homotransphobic and is represented as a list of characters' positions; if the text is not homotransphobic the list is empty.

As follows, we provided an illustrative example of a subset of training data instances:

id	text	rationales
1	i hate all gays	[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14]
2	LGBT+ ppl are generally quite nice	[]
3	@user_abcdefghjk cuz he's a faggot, that's why he should die	[11,12,13,14,15,16,30,31,32,33,34,35, 36,37,38,39,40,41,42,42]
4	science says homosexuals are inferior beings [URL]	[13,14,15,16,17,18,19,20,21,22,23,24, 25,26,27,28,29,30,31,32,33,34,35,36, 37,38,39,40,41,42,42]

NOTE to better understand the RATIONALES FORMAT:

The example with id=1 contains the characters of a homotransphobic span from character 0 to 14. The example with id=2 does not contain any homotransphobic span. The example with id=3 contains the characters of a homotransphobic span from character 11 to 16 and again from character 30 to 42. Lastly, the example with id=4 contains a homotransphobic span of text from character 13 to 42. **IMPORTANT: character counting starts from zero!**

3 Submission format

Results for both tasks should be submitted as a **tab-separated** file. Submitted runs must contain one result per line, including the **id** field provided in the test sets. In particular:

- **Subtask A - Homotransphobia detection**: we will consider the annotations provided for the fields "**homotransphobic**" for the test dataset.
- **Subtask B - Explainability**: we will consider the annotations provided for the field "**rationales"** for the test dataset**.**

Additional corpora The use of external corpora is allowed. Especially for Subtask B, we encourage users to exploit existing corpora in explainable hate speech detection to counteract the limited amount of data. In particular, participants can reuse data in English from HateXplain[1] (Mathew et al., 2021), the SemEval-2021 Task 5: Toxic Spans Detection[2] (Pavlopoulos et al., 2021) and the Task C HaSpeeDe2[3] at EVALITA 2020 (Sanguinetti et al., 2020) (nominal utterances and syntactic realizations of hateful messages).

IMPORTANT: **Each team can submit up to 3 runs for each subtask, i.e., at most 3 runs for Subtask A and 3 runs for Subtask B.**

3.1 Submission for Subtask A

Participants will submit a run file with the following format:

"**id**" "**homotransphobic**"

Below, we report a toy example of a submitted run. You can see in blue the values you will have to provide, and in black the id of the tweet that you find in the Test Set and that you have to include for the evaluation phase.

id	rationales
1	1
2	0
3	1
4	1

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

IMPORTANT: The submission file should not include the header

3.1 Submission for Subtask B

Participants will submit a run file with the following format:

"**id**" "**rationales**"

Below, we report a toy example of a submitted run. You can see in blue the values you will have to provide, and in black the id of the tweet that you find in the Test Set and that you have to include for the evaluation phase.

[1] Data available here https://github.com/hate-alert/HateXplain
[2] Data available here https://github.com/ipavlopoulos/toxic_spans
[3] Data available here http://www.di.unito.it/~tutreeb/haspeede-evalita20/data.html

id	rationales
1	[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14]
2	[]
3	[11,12,13,14,15,16,30,31,32,33,34,35,36,37,38,39,40,41,42,42]
4	[13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35, 36,37,38,39,40,41,42,42]

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

IMPORTANT: The submission file should not include the header.

4 How to submit your runs

Once you have run your system on the test set, you must send us your output naming your prediction files as follows:

teamName.subtaskname.runID

where:

- **teamName** represents the name of your team;
- **subtaskName** represent the name of the subtask and could be "A" for Subtask A and "B" for Subtask B;
- **runID** represents a progressive identifier of your runs and could be "run1", "run2", "run3".

Examples of some possible submissions are reported in the following:

bestTeam.A.run1	bestTeam.A.run2
bestTeam.B.run1	bestTeam.B.run2

(!) All relevant runs must be compressed as a single ZIP file named **teamName.zip** (e.g., *bestTeam.zip*)

Once you have created your ZIP file, submit them to hodi.evalita@gmail.com using the subject "HODI - EVALITA2023 - teamName".

5 Evaluation

Subtask A. Systems will be evaluated using a standard evaluation metric: macro-averaged F1-score.

Subtask B. Systems will be evaluated using *agreement* metrics well-known in explainable NLP. Following Pavlopoulos et al. (SemEval 2021), we will evaluate systems in terms of F1 computed on character offsets. For each system, we computed the F1 score per post, between the predicted and the ground truth character offsets. Then, we returned the macro-averaged (over test posts) score.

6 Final remarks

Visit the website for updates and news (https://hodi-evalita.github.io/).
If you have any question or problem, please open a thread on the Google Group mailing list (https://groups.google.com/g/hodievalita2023).

References

1. Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14867–14875, 2021

2. John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. SemEval-2021 task 5: Toxic spans detection. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 59–69, Online, August 2021. Association for Computational Linguistics.

3. Manuela, Sanguinetti, et al. "Haspeede 2 @ EVALITA 2020: Overview of the EVALITA 2020 hate speech detection task." Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020). CEUR, 2020.