EVALITA 2020

Automatic Misogyny Identification

Task Guidelines

Elisabetta Fersini[1], Debora Nozza[2], Paolo Rosso[3]

[1] University of Milano-Bicocca, Milan
[3] Bocconi University, Milan
[2] Universitat Politècnica de Valencia, Valencia

TABLE OF CONTENTS

1 Task description

The AMI shared task proposes the automatic identification of misogynous content in Italian language in Twitter. More specifically, it is organized according to two main subtasks:

*__Subtask A - Misogyny & Aggressive Behaviour Identification__: a system must recognize if a text is misogynous or not, and in case of misogyny, if it expresses an aggressive attitude;

*__Subtask B - Unbiased Misogyny Identification__: a system must discriminate misogynistic contents from the non-misogynistic ones, while guaranteeing the fairness of the model (in terms of unintended bias) on a synthetic dataset.

Concerning the **aggressive behaviour**, the main goal is to classify each misogynous tweet as belonging to one of the following two categories:

- *Aggressive*: the text includes an aggressive misogynous message;
- *Non-Aggressive*: it refers to a non-aggressive misogynous message.

2 Description of the dataset

The data that will be provided to the participants for the AMI shared task comprises a *raw dataset* and a *synthetic dataset* for measuring bias. Each dataset is distinguished in Training Set and Test Set.

2.1 Raw dataset

The **raw dataset** is a balanced dataset of tweets manually labelled according to two levels:

- **Misogyny**: Misogyny vs Not Misogyny
- **Aggressiveness**: Aggressive vs Not Aggressive

The training data for this dataset are provided as **TSV** files (**tab-separated** files) and report the following fields:

"**id**" "**text**" "**misogynous**" "**aggressiveness**"

where:

- **id** denotes a unique identifier of the tweet.
- **text** represents the tweet text.
- **misogynous** defines if a tweet is misogynous or not misogynous; it takes values:
  - **0** if the tweet is not misogynous;
  - **1** if the tweet is misogynous.
- **aggressiveness** denotes if a misogynous tweet is aggressive or not; it takes value as:
  - **0** denotes a non-aggressive tweet (not misogynous tweets are labelled as 0 by default);
  - **1** if the tweet is aggressive.

Examples of all possible allowed combinations are reported in Appendix 1.

2.2 Synthetic dataset

The **synthetic dataset**, for measuring the presence of unintended bias, contains template-generated text labelled according to:

- **Misogyny**: Misogyny (1) vs Not Misogyny (0)

The training data for this dataset are provided as **TSV** files (**tab-separated** files) and report the following fields:

"**id**" "**text**" "**misogynous**"

where:

- **id** denotes a unique identifier of the template-generated text.
- **text** represents the template-generated text.
- **misogynous** defines if the template-generated text is misogynous or non-misogynous; it takes values as 1 if the tweet is misogynous, 0 if the tweet is non-misogynous.

3 Submission format

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

- **Subtask A - Misogyny & Aggressive Behaviour Identification**: we will consider the annotations provided for the fields "**misogynous**" and "**aggressiveness**" for the raw test dataset.
- **Subtask B - Unbiased Misogyny Identification**: we will consider the annotations provided for the field "**misogynous"** for the raw test dataset and the field "**misogynous**" for the synthetic test dataset**.**

For each task, we distinguish between constrained and unconstrained runs:

- for a **constrained run**, teams must use the provided training data only (lexicons are admitted for constrained runs);
- for an **unconstrained run**, teams can use additional data for training, e.g., additional annotated tweets.

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**" "**misogynous**" "**aggressiveness**"

Following, 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	0	0
2	1	1
3	1	1
4	1	0

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

3.1 Submission for Subtask B

Participants to this Subtask must submit **TWO run files**, one related to the prediction on the raw test set and one related to the prediction on the synthetic test set.

3.1.1 Raw test set submission

The format for submitting the prediction on the raw data is the following one:

"**id**" "**misogynous**"

We report below 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	0
2	1
3	1
4	1

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

3.1.1 Synthetic test set submission

The format for the submission related to the synthetic test set is the following one:

"**id**" "**misogynous**"

Following, 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 synthetic text that you find in the Test Set and that you have to include for the evaluation phase.

1	1
2	0
3	0
4	1

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

4 How to submit your runs

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

teamName.subtaskname.dataType.runType.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 Substask B;
- **dataType** denotes the evaluated dataset type and could be "r" for *raw* or "s" for *synthetic*;
- **runType** denotes the type of the run and could be "c" for *constrained* or "u" for *unconstrained*;
- **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.r.c.run1	bestTeam.Br.u.run1
bestTeam.B.s.c.run1	bestTeam.A.r.u.run2

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

(!) Submissions for a run for Subtask B **must** comprise two files related to the same run (e.g. bestTeam.B.r.c.run1 and bestTeam.B.s.c.run1) where the same prediction model is used both on the raw and synthetic dataset.

Once you have created your ZIP files, submit them to submissions.ami@gmail.com using the subject "AMI@EVALITA2020 - teamName".

5 Evaluation

Subtask A. The ranking will be computed by averaging the F1 measures estimated for the Misogynous and Aggressiveness classes.

$$score_A = \frac{F_1(\text{Misogynous}) + F_1(\text{Aggressiveness})}{2}$$

Subtask B. The ranking will be computed by the weighted combination of AUC computed on the test raw dataset AUC_{raw} and three per-term AUC-based bias scores computed on the synthetic dataset ($AUC_{Subgroup}$, AUC_{BPSN}, AUC_{BNSP}). Let s be an identity-term (e.g. "girlfriend" and "wife") and be the number of identity-terms, the evaluation will be performed according to the following metric:

$$score_B = \frac{1}{2}\text{AUC}_{\text{raw}} + \frac{1}{2}\frac{\sum_s \text{AUC}_{\text{Subgroup}}(s) + \sum_s \text{AUC}_{\text{BPSN}}(s) + \sum_s \text{AUC}_{\text{BNSP}}(s)}{N}$$

Unintended bias can be uncovered by looking at differences in the score distributions between data mentioning a specific identity-term s (**subgroup** distribution) and the rest (**background** distribution). The three per-term AUC-based bias scores are related to specific subgroups as follows:

- $AUC_{Subgroup}(s)$: calculates AUC only on the data within the subgroup . This represents model understanding and separability within the subgroup itself. A low value in this metric means the model does a poor job of distinguishing between misogynous and non-misogynous comments that mention the identity.
- $AUC_{BPSN}(s)$: Background Positive Subgroup Negative (BPSN) calculates AUC on the misogynous examples from the background and the non-misogynous examples from the subgroup. A low value in this metric means that the model confuses non-misogynous examples that mention the identity-term with misogynous examples that do not, likely meaning that the model predicts higher misogynous scores than it should for non-misogynous examples mentioning the identity-term.
- $AUC_{BNSP}(s)$: Background Negative Subgroup Positive (BNSP) calculates AUC on the non-misogynous examples from the background and the misogynous examples from the subgroup. A low value here means that the model confuses

misogynous examples that mention the identity with non-misogynous examples that do not, likely meaning that the model predicts lower misogynous scores than it should for misogynous examples mentioning the identity.

6 Final remarks

Visit the website for updates and news (https://amievalita2020.github.io/).

If you have any question or problem, please open a thread on the Google Group mailing list (https://groups.google.com/d/forum/amievalita2020).

References

1. Poland, B. (2016). Haters: Harassment, Abuse, and Violence Online. University of Nebraska Press.

2. Hewitt, S., Tiropanis, T., and Bokhove, C. (2016). The problem of identifying misogynist language on Twitter (and other online social spaces). In Proceedings of the 8th ACM Conference on Web Science, pp. 333-335.

3. Anzovino M., Fersini E., and Rosso P. (2018). Automatic Identification and Classification of Misogynistic Language on Twitter. In Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB 2018). Lecture Notes in Computer Science vol 10859, pp. 57-64.

4. Dixon L., Li J., Sorensen J., Thain N., and Vasser-man L. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 67–73. ACM 2018.

5. Nozza D., Volpetti C., and Fersini E. (2019). Unintended Bias in Misogyny Detection. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '19), pp. 149–15.

6. Fersini E., Anzovino M., and Rosso P. (2018). Overview of the task on Automatic Misogyny Identification at IberEval. In Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). CEUR Workshop Proceedings, pp. 214-228.

7. Fersini E., Nozza D., and Rosso P. (2018). Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018). CEUR Workshop Proceedings.

8. Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F.M., Rosso P., and Sanguinetti M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation , pp. 54–63.

9. Borkan D., Dixon L., Sorensen J., Thain N., and Vasserman L. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In Companion of The 2019 World Wide Web Conference (WWW 2019). ACM.

Appendix: Examples of all possible combinations

A.1 Submission for Subtask A - Examples

Additionally to the field "id", we report in the following all the combinations of labels to be predicted, i.e. "**misogynous**", "**aggressiveness**"

```
0	0
1	1
1	0
```

A.2 Submission for Subtask B - Examples

Additionally to the field "id", we report in the following all the combinations of labels to be predicted, i.e. "**misogynous**", "**aggressiveness**"

For the raw test set, the possible label combinations are:

```
0	0
1	1
1	0
```

For the synthetic test set, the possible labels are:

```
0
1
1
```