IberEval 2018

Automatic Misogyny Identification

Task Guidelines

Maria Anzovino[1], Elisabetta Fersini[1], Paolo Rosso[2]

[1] University of Milano-Bicocca
[2] Universitat Politècnica de Valencia

TABLE OF CONTENTS

1 Task description

The AMI task proposes the automatic identification of misogynous content both in Spanish and English languages in Twitter. The AMI shared task is organized according to two main sub-tasks:

*__Subtask A – Misogyny Identification__: discrimination of misogynist contents from the non-misogynist ones;

*__Subtask B - Misogynistic Behaviour and Target Classification__: recognition of the targets that can be either specific users or groups of women together with the identification of the type of misogyny against women.

Regarding the **misogynistic behaviour**, a tweet must be classified as belonging to one of the following categories:

- *Stereotype & Objectification:* a widely held but fixed and oversimplified image or idea of a woman; description of women's physical appeal and/or comparisons to narrow standards.
- *Dominance:* to assert the superiority of men over women to highlight gender inequality.
- *Derailing:* to justify woman abuse, rejecting male responsibility; an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men.
- *Sexual Harassment & Threats of Violence:* to describe actions as sexual advances, requests for sexual favours, harassment of a sexual nature; intent to physically assert power over women through threats of violence.
- *Discredit:* slurring over women with no other larger intention.

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

- *Active (individual)*: the text includes offensive messages purposely sent to a specific target;
- *Passive (generic)*: it refers to messages posted to many potential receivers (e.g groups of women).

2 Description of the dataset

For the AMI task, we provide one corpus for Spanish and one corpus for English. Each corpus is distinguished in Training Set and Test Set. Regarding the training data, the Spanish corpus is composed of 3307 tweets, while the English one is composed of 3251 tweets.

The training data provided are **tab-separated**, reporting the following fields:

“**id**” “**tweet**” “**misogynous**” “**misogyny_category**” “**target**”

where:

- **id** denotes a unique identifier of the tweet
- **tweet** represents the tweet text
- **misogynous** defines if the tweet is misogynous or not misogynous; it takes values as 1 if the tweet is misogynous, 0 if the tweet is misogynous.
- **misogyny_category** denotes the type of misogynistic behaviour; it takes value as:
    - **stereotype**: denotes the category “*Stereotype & Objectification”*;
    - **dominance**: denotes the category “*Dominance”*;
    - **derailing**: denotes the category “*Derailing”*;
    - **sexual_harassment**: denotes the category “*Sexual Harassment & Threats of Violence”*;
    - **discredit**: denotes the category “*Discredit”;*
    - **0** if the tweet is not misogynous.
- **target** denotes the subject of the misogynistic tweet; it takes value as:
    - **active**: denotes a specific target (individual);
    - **passive**: denotes potential receivers (generic);
    - **0** if the tweet is not misogynous.

Examples of all possible combinations are reported in Appendix 1.

3 Submission format

Results for both tasks should be submitted in a plain text file with **tab-separated** fields. The format of the run files submitted by participants must be as follows:

“id" “misogynous” “misogyny_category” “target”

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	0
2	1	stereotype	active
3	1	stereotype	passive
4	1	discredit	passive

Specifically, submitted runs must contain one tweet per line including the original values provided in Test Set for what concerns the **id** field, plus the annotations for the fields that are relevant with respect to the tasks. In particular:

- **Subtask A - Misogyny Identification**: we will consider only annotations provided for the field “**misogynous”** (2nd column of the submission format)
- **Subtask B - Misogynistic Behaviour and Target Classification**: we will consider only annotations provided for the fields “**misogyny_category”** and “**target**” (3rd and 4th columns of the submission format)

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

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 five runs in total (!).** Runs can be constrained (the provided training data and lexicons are admitted) and unconstrained (additional data for training, e.g. additional annotated tweets, are allowed).

4 How to submit your runs

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

teamName.runType.runID

where:

- **teamName** represents the name of your team
- **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", "run4", "run5"

Examples of some possible submissions are reported in the following:

bestTeam.c.run1	bestTeam.u.run1
bestTeam.c.run2	bestTeam.u.run2

Send all relevant files to submissions.ami@gmail.com using the subject "AMI@IberEval2018 - teamName".

5 Evaluation

Subtask A. Systems will be evaluated on the field "**misogynous"** using the standard accuracy measure, and ranked accordingly. Accuracy will be computed as follows:

$$\text{Accuracy} = \frac{\text{number of correctly predicted instances}}{\text{total number of instances}}$$

Subtask B. Each field to be predicted will be evaluated independently on the other using a macro-average F1-score. In particular, the macro-average F1-score for the "**misogyny_category"** field will be computed as average of F1-scores obtained for each category (stereotype, dominance, derailing, sexual_harassment, discredit), estimating F_1(misogyny_category). Analogously, the macro-average F1-score for the "**target"** field will be computed as average of F1-scores obtained for each category (active, passive), F_1(target).

The final ranking of the systems participating to subtask B will be based on the macro-average F1-score (F_1), computed as follows:

$$F_1 = \frac{F_1(\text{misogyny\_category}) + F_1(\text{target})}{2}$$

6 Final remarks

If you have any questions or problems, please open a thread on the Google Groups mailing list (https://groups.google.com/forum/#!forum/amiibereval2018).

References

1. M. Anzovino, E. Fersini, P. Rosso. Automatic Identification and Classification of Misogynistic Language on Twitter. In Proceedings of the 23rd International Conference on Natural Language & Information Systems, 2018.
2. Hewitt, S., Tiropanis, T. and Bokhove, C., 2016, May. 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). ACM.
3. Poland, B., 2016. Haters: Harassment, Abuse, and Violence Online. U of Nebraska

Appendix: Examples of all possible combinations

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

0	0	0
1	stereotype	active
1	stereotype	passive
1	dominance	active
1	dominance	passive
1	derailing	active
1	derailing	passive
1	sexual_harassment	active
1	sexual_harassment	passive
1	discredit	active
1	discredit	passive