Hate Speech Detection Task (HaSpeeDe) - Evalita 2018
– Task Guidelines –

Cristina Bosco[⋆], Felice Dell'Orletta[◇], Fabio Poletto[⋆], Manuela Sanguinetti[⋆], Maurizio Tesconi[•]
[⋆]Univeristy of Torino, Computer Science Department
[◇]CNR-ILC, Pisa, [•]CNR-IIT, Pisa
{bosco, msanguin}@di.unito.it, felice.dellorletta@ilc.cnr.it,
fabio.poletto@edu.unito.it, maurizio.tesconi@iit.cnr.it

June 4, 2018

Contents

1 Task Description

The task consists in automatically annotating messages from two popular micro-blogging platforms, Twitter and Facebook,with a boolean value indicating the presence (or not) of HS. It is proposed for the first time for Italian within the context of Evalita, following the success of similar tasks on sentiment analysis, such as those for polarity and subjectivity detection, organized in the two last editions of this campaign.

Considering the linguistic, as well as meta-linguistic, features that distinguish Twitter and Facebook messages, namely due to the differences in use between the two platforms, the task will be further organized into three sub-tasks, based on the dataset used for training and testing the participants' systems:

- **Task 1: HaSpeeDe-FB**, where only the Facebook dataset can be used to classify the Facebook test set
- **Task 2: HaSpeeDe-TW**, where only the Twitter dataset can be used to classify the Twitter test set
- **Task 3: Cross-HaSpeeDe**, which can be further subdivided into two sub-tasks:
  1. **Task 3.1: Cross-HaSpeeDe_FB**, where only the Facebook dataset can be used to classify the Twitter test set
  2. **Task 3.2: Cross-HaSpeeDe_TW**, where only the Twitter dataset can be used to classify the Facebook test set

  Cross-HaSpeeDe has been proposed as an out-of-domain task that specifically aims on one hand at highlighting the challenging aspects of using social media data for classification purposes, and on the other at enhancing the systems' ability to generalize their predictions with different datasets. We therefore encourage you to participate in both sub-tasks.
  However, we point out that they can also be performed separately, and you are free to choose to participate in just one of them.

NOTE: You are also allowed to use other additional resources for training, BUT you have to declare it during submission, specifying exactly which resources you used, besides the training set we provided for the task.

Below we provide more details on the data and the submission instructions.

2 Dataset

The dataset proposed for this task is the result of a joint effort of two research groups on harmonizing the annotation previously applied to two different datasets, in order to allow their exploitation in the task.

The first dataset is a collection of Facebook comments developed by the group from CNR-Pisa and created in 2016 [1], while the other one is a Twitter corpus developed in 2018 by the group from the Computer Science Department of Turin University [3], [2].

Data Split Both Facebook and Twitter datasets consist of a total amount of 4,000 comments/tweets, randomly split into development and test set, of 3,000 and 1,000 messages respectively.

Training Set Format The annotation format is the same for both datasets used for this task, and it consists of a simplified version of the schemes adopted in the corpora briefly introduced above.

The data has been encoded in a UTF-8 plain-text file with three tab-separated columns, each one representing the following information:

1. the ID of the Facebook comment or tweet,
2. the text,

3. the class: `1` if the text **contains** hate speech, and `0` otherwise (see Tables 1 and 2 for a few examples).

id	text	hs
8	Io voterò NO NO E NO	0
36	Matteo serve un colpo di stato. Qua tra poco dovremo andare in giro tutti armati come in America.	1

Table 1: Annotation examples from the Facebook dataset.

id	text	hs
811099342593462272	Corriere: Tangenti, Mafia Capitale dimenticata Mazzette su buche e campi rom https://t.co/RTPu3LNTzF	0
821056358283902977	@investigator113 altro che profughi? sono zavorre e tutti uomini	1

Table 2: Annotation examples from the Twitter dataset.

Test Set and Submission Format The test set will be formatted the same way as the training set, but it will consist of the first two columns only (described above in the Tables), while each participant is asked to submit the complete file, with the predicted class for each post in the third column.

3 Dataset Release to Participants

The dataset will be made available via this GitHub repository:

https://github.com/msang/haspeede2018

We remind you that the training set will be made available on **May 28th 2018**. It is associated with a license that defines the terms that the users will be bind by in data exploitation and citation.

Important: If some update is applied to the released data, a notification will be published in the task website. Please check it regularly to be sure that you are using the last released training set.

The test set will be instead released on **3rd September 2018** (see also the "Important dates" section in the task web page[1]).

4 Submission of Results

Once you have run your system/s over the test data, you will have to send it to us following these recommendations:

1. choose a team name and name the files containing your runs in the following way: `haspeede_teamName_systemID.tsv`
2. send the file/s to the following email addresses: `felice.dellorletta@ilc.cnr.it` and `msanguin@di.unito.it`, using the subject "`haspeede-taskName - teamName`", where the "`taskName`" suffix namely stands for the task type for which you are submitting the results:
   - `FB` should be used for Task 1 (**Haspeede-FB**)
   - `TW` should be used for Task 2 (**Haspeede-TW**)
   - `Cross_FB` should be used for Task 3.1 (**Cross-Haspeede_FB**)
   - `Cross_TW` should be used for Task 3.2 (**Cross-Haspeede_TW**)

[1] http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html

Important: If you used even other additional resources for training, you have to declare it during submission, specifying exactly which resources you used, besides the training set we provided for the task.

5 Evaluation

Each participating team will initially have access to the training data only. Later, the unlabeled test data will be released. After the assessment, the complete test data will be released as well.

Participants will be allowed to **submit up to 2 runs for each sub-task**.

For each sub-task a separate official ranking will be provided, and the evaluation will be performed according to the standard metrics known in literature, i.e Precision, Recall and F-score. Given the imbalanced distribution of hateful *vs* not hateful messages, and in order to get more useful insights on the system's performance on a given class, the scores will be computed for each class separately, and finally the F-score will be macro-averaged, so as to get the overall results.

For all sub-tasks, the baseline score will be computed as the performance of a classifier using a zeroR classification method, where each tweet or Facebook comment is always classified into the majority class.

6 Contacts

For any question or problem, please start a topic on our googlegroups mailing list:

haspeede-evalita2018@googlegroups.com

References

[1] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017.*, pages 86–95, 2017.

[2] Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. Hate speech annotation: Analysis of an italian twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR, december 2017.

[3] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of LREC 2018*, May 2018.