Hate Speech Detection Task (HaSpeeDe) Second Edition @ Evalita 2020 – Task Guidelines –

Cristina Bosco[⋆], Tommaso Caselli[⋄], Gloria Comandini[⊕], Elisa Di Nuovo[⋆], Simona Frenda[⋆,°], Viviana Patti[⋆], Irene Russo[⊙], Manuela Sanguinetti[⋆], Marco Stranisci[•]
[⋆]University of Torino
[°]Universitat Politècnica de València
[⋄]University of Groningen
[⊕]University of Trento
[⊙]CNR-ILC, Pisa
[•]Acmos, Torino
{bosco, dinuovo, frenda, msanguin, patti}@di.unito.it, t.caselli@rug.nl, gloria.comandini@unitn.it, irene.russo@ilc.cnr.it, marco.stranisci@acmos.net

August 3, 2020

Contents

1 Task Description

The first edition of HaSpeeDe [1] consisted in automatically annotating messages in Social Media with a boolean value indicating the presence of Hate Speech (HS). The high participation and the promising results encouraged this second edition at EVALITA 2020.
HaSpeeDe 2020 focuses on three main phenomena of HS online that are reflected along three tasks.

- **Task A**
  Hate Speech Detection (Main Task)
  Task A consists in a binary classification task aimed at determining the presence or the absence of hateful content in the text towards a given target (among Immigrants, Muslims or Roma people).
  Given a message, decide whether the message contains Hate Speech or not.

- **Task B**
  Stereotype Detection (Pilot Task 1)
  Task B consists in a binary classification task aimed at determining the presence or the absence of a stereotype toward a given target (among Immigrants, Muslims or Roma people). As defined in Merriam Webster Dictionary, Stereotype is "a standardized mental picture that is held in common by members of a group and that represents an oversimplified opinion, prejudiced attitude, or uncritical judgment". Considering these characteristics, Stereotype can be used to express hateful or offensive messages in a more subtle manner, hindering the correct recognition of Hate Speech [4]. In this perspective, Task B aims to boost the investigation of its occurrences especially in hateful context.
  Given a message, decide whether the message expresses Stereotype or not.

- **Task C**
  Identification of Nominal Utterances (Pilot Task 2)
  Task C consists in a sequence labeling task aimed at recognizing Nominal Utterances in hateful tweets.
  Nominal utterances (NUs) are intended as independent syntactic constructions built around a non-verbal head and, as showed in [2], they play an important role in the expression of hateful messages conveying also populist slogans or sharp adherence to a particular political or social point of view. Task C, indeed, aims to encourage the analysis and the identification of NUs in Hate Speech detection.
  Given a hateful message, label Nominal Utterances.

Finally, another challenging novelty of HaSpeeDe 2020 is the fact that the performances of the participating systems will be evaluated on a set of mixed text genres. The spread of Hate Speech takes place on different platforms online, such as Social Media and Newspaper sites, therefore, the test set consists of tweets and newspapers headlines. Although both genres usually involve different registers (informal and formal styles), the annotation scheme about hate speech recognition showed to be suitable for both text varieties. Moreover, the shortness and the sharp style of both types of text could help the classification task.

NOTE:

- Participation is allowed to all three tasks (Task A, Task B and Task C), two tasks (Task A and Task B or Task A and Task C) or only Task A;

- Participants are allowed to use other additional resources for training, BUT they have to declare it during submission, specifying exactly which resources were used, besides the training set provided for the task.

2 Development and Test Data

The dataset of HaSpeeDe 2020 includes texts targeting minority groups such as Immigrants, Muslims and Roma communities, whose relative social problems constantly feed the public and political debate triggering HS. In this perspective, we collected tweets and newspapers headlines using specific keywords related to the mentioned Italian minorities.

Data Split The entire dataset is split in a training set composed of tweets, and a test set comprising mixed genres (tweets and newspapers headlines).

Important Dates

- The Training Set is made available on **May 29th 2020**. It is associated with a license that defines the terms that the users will be bind by in data exploitation and citation. If some **update** is applied to the released data, a notification will be published in the task website[1] and on the Google Group[2]
- The Registrations will be closed on **September 4th 2020**[3].
- The Test Set will be released on **September 18th, 2020**.
- The System Results will be due to organizers by **September 25th, 2020**.
- The Evaluation Results will be due to participants on **October 2nd, 2020**
- The Technical Reports will be due to organizers on **October 16th, 2020** and the camera-ready version on **November 6th, 2020**.
- The pre-recorded video presentations should be sent to the EVALITA chairs by **November 27th, 2020**.
- The Final Workshop will be on **December 16th-17th 2020**[4].

2.1 Training Set Distribution and Format

The training set is composed of the HaSpeeDe2018 tweets with the addition of brand new data. Overall, it includes 6839 tweets distributed as in tables 1 and 2.

HS	NOT HS	TOTAL
2766	4073	6839

Table 1: Distribution of Hate Speech in the Training set.

STEREOTYPE	NOT STEREOTYPE	TOTAL
3042	3797	6839

Table 2: Distribution of Stereotype in the Training set.

Task A and B The single training set provided for both tasks A and B is a tab-separated values (TSV) file including the following fields:

1. a number that substitutes the original tweet ID
2. the text
3. the hate speech class: `1` if the text **contains** hate speech, and `0` otherwise
4. the stereotype class: `1` if the text **contains** stereotypes, and `0` otherwise.

The data format in the training set is as follows:

```
id text hs stereotype
```

id	text	hs	stereotype
8783	Via tutti i campi Rom e disinfettare per bene il lerciume che si lasciano dietro. Mai più campi Rom in Italia NO NO E NO	1	1
9254	Vanno affondate. Hanno rotto i c.....i Aquarius vuol dettare ancora legge: carica migranti e rifiuta gli ordini libici	1	0
9414	Istat conferma: migranti vengono in Italia a farsi mantenere	0	1
7242	Il casino libico torna comodo per certi magheggi col petrolio? Continuate a parlarci dell'emergenza migranti...	0	0

Table 3: Examples from the Training set.

and the mention of usernames in the tweets and the URLs are replaced with the `@user` and `URL` placeholders.
The test set will contain values only for the first two fields (`id` and `text`).
Table 3 shows some examples of the possible combinations in accord with our annotation scheme.

Task C The training set provided for task C has been annotated using WebAnno and exported in the WebAnno TSV v. 3.2 format and eventually modified in order to simplify the annotation fields and to have the NUs' annotation in a IOB (Inside-Outside-Beginning) format. The annotation includes the following tab-separated fields:

1. a number that substitutes the original tweet ID (`6386` in the example) followed by an hyphen that divide it from the the token number (`1` in the example)
2. token (word) (`Buongiorno` in the example)
3. annotation of the NUs (`B-NU-CGA` in the example)

The data format in the training set is as follows:

```
TweetID-TokenNumber Token IOB_annotation

6386-1 Buongiorno B-NU-CGA
```

The test set will contain all the values of the training set, except for the last column (`annotation of NUs`).
The annotation of NUs will be labeled with **B-NU-CGA** (*Begin*-NU-CGA), **I-NU-CGA** (*Inside*-NU-CGA) or **O** (*Outside*). A NU starts with a token labeled as B-NU-CGA and all its other tokens are labeled as I-NU-CGA; every token that is not part of a NU is annotated as O.

Here are some examples taken from the Training set.

#Text=Vergogna!
8482-1 Vergogna B-NU-CGA
8482-2 ! I-NU-CGA
#Text=È UNA PROVOCAZIONE...ORA BASTA..
NESSUNO SBARCHI IN #ITALIA

[1] http://di.unito.it/haspeede20
[2] Join the Google Group https://groups.google.com/d/forum/haspeede-evalita2020, so please check it regularly to stay up to date with the news regarding the task.
[3] Here the url for the registration: http://www.evalita.it/2020/taskregistration
[4] http://www.evalita.it/2020

9602-23	È	O
9602-24	UNA	O
9602-25	PROVOCAZIONE	O
9602-26	.	O
9602-27	.	O
9602-28	.	O
9602-29	ORA	B-NU-CGA
9602-30	BASTA	I-NU-CGA
9602-31	.	I-NU-CGA
9602-32	.	I-NU-CGA
9602-33	NESSUNO	O
9602-34	SBARCHI	O
9602-35	IN	O
9602-36	#	O
9602-37	ITALIA	O

The annotation of NUs is Coarse-Grain and it can join two or more verbless constructions, if they are not separated by a verbal utterance or by a full stop.

#Text=Padroni in casa nostra padroni a casa nostra via la merda negra islamica e zingara dall'Italia Matteo premier

4747-1	Padroni	B-NU-CGA
4747-2	in	I-NU-CGA
4747-3	casa	I-NU-CGA
4747-4	nostra	I-NU-CGA
4747-5	padroni	I-NU-CGA
4747-6	a	I-NU-CGA
4747-7	casa	I-NU-CGA
4747-8	nostra	I-NU-CGA
4747-9	via	I-NU-CGA
4747-10	la	I-NU-CGA
4747-11	merda	I-NU-CGA
4747-12	negra	I-NU-CGA
4747-13	islamica	I-NU-CGA
4747-14	e	I-NU-CGA
4747-15	zingara	I-NU-CGA
4747-16	dall'Italia	I-NU-CGA
4747-17	Matteo	I-NU-CGA
4747-18	premier	I-NU-CGA

If separated by a full stop, two or more NUs can be adjacent and the second NU can be recognized from the label B-NU-CGA.

#Text=Bastaaaaaa. Fuori gli immigrati e dentro gli italiani.

9152-1	Bastaaaaaa	B-NU-CGA
9152-2	.	I-NU-CGA
9152-3	Fuori	B-NU-CGA
9152-4	gli	I-NU-CGA
9152-5	immigrati	I-NU-CGA
9152-6	e	I-NU-CGA
9152-7	dentro	I-NU-CGA
9152-8	gli	I-NU-CGA
9152-9	Italiani	I-NU-CGA
9152-10	.	I-NU-CGA

2.2 Training set Release

The datasets for tasks A, B and C will be made available in a password-protected zip archive in the following GitHub repository:

https://github.com/msang/haspeede/tree/master/2020

In order to unzip the file, participants are asked to fill in this form:

https://forms.gle/BJQy6ciiXXtPCCJdA

An email notification will be sent with the password.

3 Submission of Results

3.1 Submission Format

For each task, participants will be allowed to **submit up to 2 runs**.

Task A and B Participants are asked to submit the results of the runs in a plain text file with tab-separated values (TSV) using the format as below:

`id text hs stereotype`

The last column (`stereotype`) can be left blank for those who participate only in Task A.

Task C Participants are asked to submit the results of the runs following the same file format of the training set, since the NUs in the test set will be labeled relying on the provided tokenization.

3.2 How to submit your runs

Once participating systems have been run over the test data, the file with the run predictions should be sent to the organizers following the recommendations below:

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: `t.caselli@rug.nl` and `msanguin@di.unito.it`, using the subject "`haspeede-taskName-teamName-runNumber`", where the "`taskName`" suffix namely stands for the task type for which you are submitting the results:
   - `A` should be used for Task A (**Hate Speech Detection**)
   - `B` should be used for Task B (**Stereotype Detection**)
   - `C` should be used for Task C (**Identification of Nominal Utterances**)

   while the "`runNumber`" just indicates the number of the run the team is willing to submit (hence its value should be "`1`" or "`2`").

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

4 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. All the data will be made available in the GitHub repository mentioned in Section 2.2.
For each 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.
For Task A and Task B, 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. Precision, Recall, and F-score will be computed as follows:

$$precision_{class} = \frac{\#correct\_class}{\#assigned\_class}$$

$$recall_{class} = \frac{\#correct\_class}{\#total\_class}$$

$$F_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}}$$

For Task C, a NU is considered correct only if it is an exact match, i.e., if all tokens that compose it are correctly identified. Precision, Recall, and F-score are computed as follows:

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F_{class} = 2\frac{precision recall}{precision + recall}$$

Where *tp* is the tokens that are part of an extent in both key and response, *fp* is the number of tokens that are part of an extent in the response but not in the key, and *fn* is the number of tokens that are part of an extent in the key but not in the response.
Different baseline systems will be built according to the task type:

- Task A and B (Hate Speech and Stereotype Detection): a Linear SVM with TF-IDF of word and char-grams will be used.
- Task C (Identification of Nominal Utterances): the baseline replicates the one presented for the COSMIANU corpus [3] which identifies as correct in the test the NUs that appear in the training (memory-based approach).

5 Contacts

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

https://groups.google.com/d/forum/haspeede-evalita2020

References

[1] Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR, 2018.

[2] Gloria Comandini and Viviana Patti. An Impossible Dialogue! Nominal Utterances and Populist Rhetoric in an Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 163–171, 2019.

[3] Gloria Comandini, Manuela Speranza, and Bernardo Magnini. Effective Communication without Verbs? Sure! Identification of Nominal Utterances in Italian Social Media Texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[4] Chiara Francesconi, Cristina Bosco, Fabio Poletto, and Manuela Sanguinetti. Error analysis in a hate speech detection task: the case of haspeede-tw at evalita 2018. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), Bari, Italy, November 13-15, 2019*, 2019.