# HAte SPEEch DEtection
# HaSpeeDe


**Cristina Bosco**, **Manuela Sanguinetti**, Department of Computer Science, University of Turin
**Felice Dell'Orletta**, ILC–CNR, Pisa
**Fabio Poletto**, Acmos, Turin
**Maurizio Tesconi**, IIT–CNR, Pisa

# Task Description and Motivation

Shared task on **hate speech** detection on Italian social media

= automatically annotate messages with a **binary value**

(**0** or **1**) indicating the presence (or not) of hate speech


**HATE SPEECH: any expression ``*that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth*“ (Erjavec and Kovačič, 2012)**

# Task Description and motivation

Online platforms discourage hateful content, but its removal mainly relies on trusted flaggers and common users reports

- Hate speech identification requires a multidisciplinary approach (psychology, law, social sciences, etc.) but **Computational Linguistics** plays a key role
- **Task main purpose**: encourage and promote the participation of several research groups to allow an advancement in the state of the art for Italian

# Task organization and datasets

4 Sub-tasks that depend on the dataset used:

- **HaSpeeDe-FB** = Facebook Train + Facebook Test
- **HaSpeeDe-TW** = Twitter Train + Twitter Test

**Cross-HaSpeeDe**:

- **Cross-HaSpeeDe_FB**
  = Facebook train + Twitter Test
- **Cross-HaSpeeDe_TW**
- = Twitter train + Facebook Test

# The Facebook dataset

Retrieved from the corpus described in Del Vigna et al. (2016)

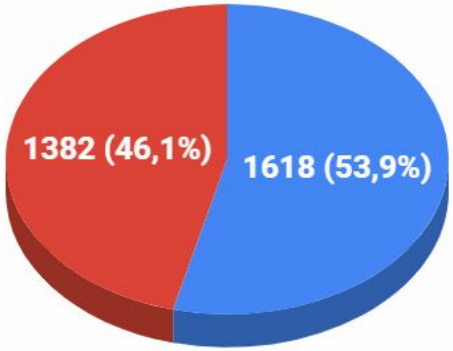
**Content**: comments from web pages and groups with potentially hateful content

**Targets**: Religion, Physical and/or mental handicap, Socio–economical status, Politics, Race, Sex and Gender issues, and Other

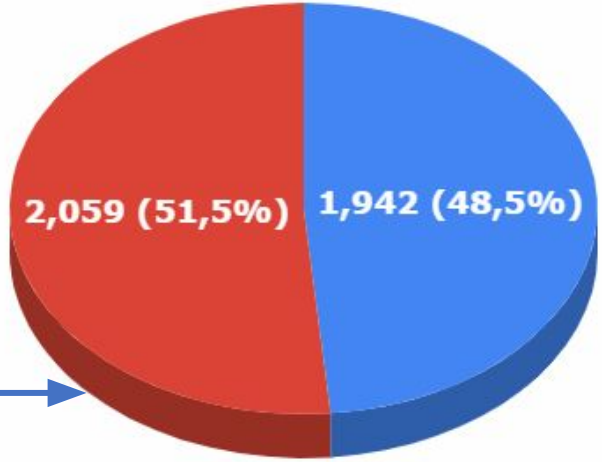
**Annotation**: Experts

# The Facebook dataset

Label Distribution

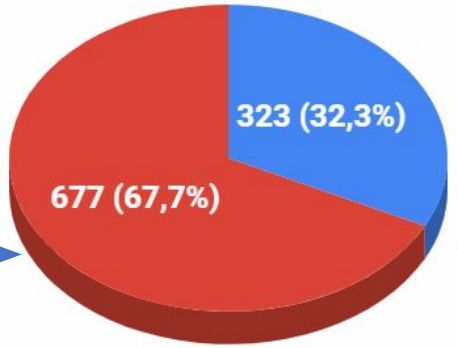



train (3,000)




all (4,000)

test (1,000)




not hate speech

hate speech

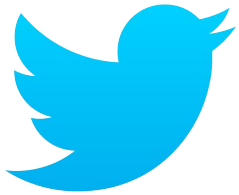
# The Facebook dataset

Format of the annotated items:

| 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 |

# The Twitter dataset

Retrieved from the corpus described in Poletto et al. (2017) – Sanguinetti et al. (2018)



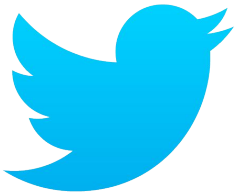
**Content**: tweets filtered using *neutral* keywords related to three HS targets

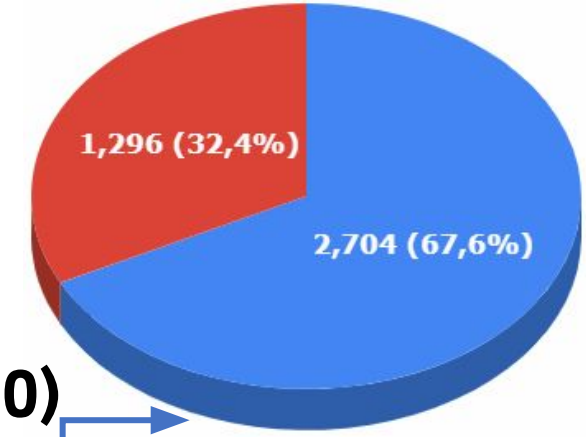
**Targets**: Immigrants, Muslisms, Roma

**Annotation**: Experts + CrowdFlower (now FigureEight)

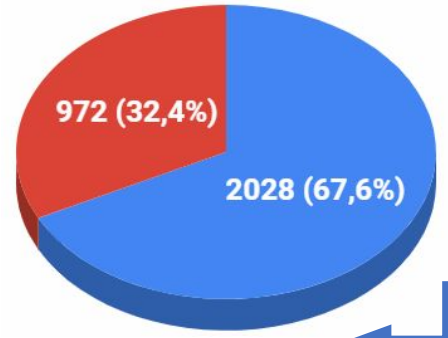
# The Twitter dataset

## Label Distribution



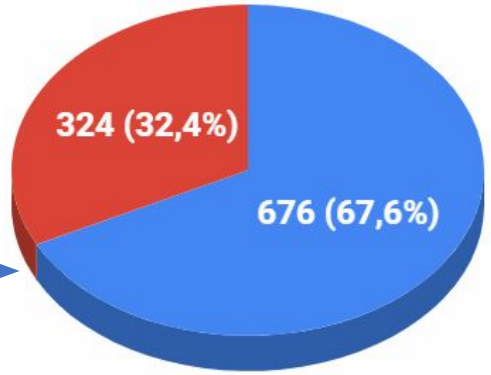



all (4,000)




train (3,000)

test (1,000)




not hate speech

hate speech

# The Twitter dataset

Format of the annotated items

| Id | Text | HS |
|---|---|---|
| 1783 | *Corriere: Mafia Capitale, 4 patteggiamenti Gli appalti truccati dei campi rom* | 0 |
| 3290 | *altro che profughi? sono zavorre e tutti uomini* | 1 |

# Evaluation and metrics

- Precision, Recall, F1–score for each class (0|1) + Macro–F1
- Baseline: classification based on the most frequent class
- Up to 2 submissions allowed for each task
- Separate official rankings provided for each task

# Participation: 9 teams

| Team | Affiliation |
|---|---|
| **GCRP** | Universitat Politècnica de València + CERPAMID, **SPAIN–CUBA** |
| **INRIA–FBK** | Université Côte d'Azur, CNRS, Inria + FBK, Trento **FRANCE–ITALY** |
| **ItaliaNLP** | ILC–CNR, Pisa, **ITALY** |
| **Perugia** | University for Foreigners of Perugia + University of Perugia + University of Florence, **ITALY** |
| **RuG** | University of Groningen+ University of Salerno, **THE NETHERLANDS–ITALY** |
| **sbMMP** | Zurich University of Applied Sciences, **SWITZERLAND** |
| **StopPropagHate** | INESC TEC + University of Porto + Eurecat, Centre Tecnique de Catalunya, **PORTUGAL–SPAIN** |
| **HanSEL** | University of Bari "Aldo Moro", **ITALY** |
| **VulpeculaTeam** | University of Perugia, **ITALY** |

# Results: HaSpeeDe-FB

| Team | Macro F1-score |
|---|---|
| baseline | 0.2441 |
| ItaliaNLP_2 | 0.8288 |
| ItaliaNLP_1 | 0.8106 |
| InriaFBK_1 | 0.8002 |
| InriaFBK_2 | 0.7863 |
| Perugia_2 | 0.7841 |
| RuG_1 | 0.7751 |
| HanSEL | 0.7738 |
| VulpeculaTeam* | 0.7554 |
| RuG_2 | 0.7428 |
| GRCP_2 | 0.7147 |
| GRCP_1 | 0.7144 |
| StopPropagHate_2* | 0.6532 |
| StopPropagHate_1* | 0.6419 |
| Perugia_1 | 0.2424 |

**FB train + FB test**

complete results at goo.gl/8gX7xc

Highest score:
Cimino and De Mattei
*"Multitask Learning in Deep Neural Networks for Hate Speech Detection in Facebook and Twitter"*

# Results: HaSpeeDe-TW

| Team | Macro F1-score |
|---|---|
| baseline | 0.4033 |
| ItaliaNLP_2 | 0.7993 |
| ItaliaNLP_1 | 0.7982 |
| RuG_1 | 0.7934 |
| InriaFBK_2 | 0.7837 |
| sbMMMP | 0.7809 |
| InriaFBK_1 | 0.78 |
| VulpeculaTeam* | 0.7783 |
| Perugia_2 | 0.7744 |
| RuG_2 | 0.753 |
| StopPropagHate_2* | 0.7426 |
| StopPropagHate_1* | 0.7203 |
| GRCP_1 | 0.6638 |
| GRCP_2 | 0.6567 |
| HanSEL | 0.6491 |
| Perugia_1 | 0.4033 |

**TW train + TW test**

complete results at
goo.gl/8gX7xc

Highest score:
Cimino and De Mattei
*"Multitask Learning in Deep Neural Networks for Hate Speech Detection in Facebook and Twitter"*

# Results: Cross-HaSpeeDe_FB

| Team | Macro F1-score |
|---|---|
| baseline | 0.4033 |
| InriaFBK_2 | 0.6541 |
| InriaFBK_1 | 0.6531 |
| VulpeculaTeam | 0.6542 |
| Perugia_2 | 0.6279 |
| ItaliaNLP_1 | 0.6068 |
| ItaliaNLP_2 | 0.5848 |
| GRCP_2 | 0.5436 |
| RuG_1 | 0.5409 |
| RuG_2 | 0.4845 |
| GRCP_1 | 0.4544 |
| HanSEL | 0.4502 |
| StopPropagHate | 0.443 |
| Perugia_1 | 0.4033 |

**FB train + TW test**

complete results at goo.gl/8gX7xc

Highest score:
Corazza, Menini, Arslan, Sprugnoli, Cabrio, Tonelli, and Villata
*"Comparing Different Supervised Approaches to Hate Speech Detection"*

# Results: cross-HaSpeeDe_TW

| Team | Macro F1-score |
|---|---|
| baseline | 0.2441 |
| ItaliaNLP_2 | 0.6985 |
| InriaFBK_2 | 0.6802 |
| ItaliaNLP_1 | 0.6693 |
| InriaFBK_1 | 0.6547 |
| VulpeculaTeam* | 0.6189 |
| RuG_1 | 0.6021 |
| RuG_2 | 0.5545 |
| HanSEL | 0.4838 |
| Perugia_2 | 0.4594 |
| GRCP_1 | 0.4451 |
| StopPropagHate* | 0.4378 |
| GRCP_2 | 0.318 |
| Perugia_1 | 0.2441 |

**TW train + FB test**

complete results at goo.gl/8gX7xc

Highest score:
Cimino and De Mattei
*"Multitask Learning in Deep Neural Networks for Hate Speech Detection in Facebook and Twitter"*

# Some remarks

- The performance of in-domain tasks is better than the one in the cross-domain tasks (expected)
- Tasks with FB test set (HaSpeeDe-FB/ Cross-HaSpeeDe_TW) better than those with TW test set (HaSpeeDe-TW/ Cross-HaSpeeDe_FB) (less expected)

# Some remark about results

Tentative explanations:

- Facebook comments are (in general) more grammatically correct and longer than tweets
- Higher distribution of HS comments in the FB dataset than in TW
- Fewer HS targets in the TW dataset than in FB
- TW corpus created filtering tweets with *neutral* keywords related to the selected targets (HS conveyed in subtler ways)

# Some remark about approaches

- Almost all participants used deep learning methods and/or pre-trained word embeddings BUT **polarity/subjectivity lexicons** still play a key role (see best-performing team in 3 out of 4 tasks)

- Overall, promising results have been achieved, despite the small datasets provided for training

# Future work

- The group from UniTO is now organizing in cooperation with Paolo Rosso (University Politecnica of Valencia) and Elisabetta Fersini (University of Milano Bicocca) the

**SemEval 2019 Task 5 – Shared Task on Multilingual Detection of Hate**

https://competitions.codalab.org/competitions/19935

# Thank you!

HATE SPEECH AND SOCIAL MEDIA by Fondazione CRT

and

I HATE PREJUDICE by Compagnia di San Paolo

allowed the development of the Twitter dataset