HATE-ITA: Hate Speech Detection in Italian Social Media Text

Debora Nozza

Federico Bianchi

Giuseppe Attanasio

Bocconi University, Milan

debora.nozza@unibocconi.it

@debora_nozza

erc

INTEGRATOR

European Research Council

Fondazione CARIPLO

Binary Hate Speech Classifier in Italian Social Media Text

HATE-ITA

HATE-ITA

MilaNLProc/**hate-ita** like 1

it | arxiv:2104.12250 | mit | text classification | abusive language | hate speech | offensive language

Model card | Files and versions | Community | Settings

Edit model card

Debora Nozza • Federico Bianchi • Giuseppe Attanasio

Abstract

Online hate speech is a dangerous phenomenon that can (and should) be promptly counteracted properly. While Natural Language Processing has been successfully used for the purpose, many of the research efforts are directed toward the English language. This choice severely limits the classification power in non-English languages. In this paper, we test several learning frameworks for identifying hate speech in Italian text. We release HATE-ITA, a set of multi-language models trained on a large set of English data and available Italian datasets. HATE-ITA performs better than mono-lingual models and seems to adapt well also on language-specific slurs. We believe our findings will encourage research in other mid-to-low resource communities and provide a valuable benchmarking tool for the Italian community.

Model

This model is the fine-tuned version of the XLM-T model.

Results

This model had an F1 of 0.83 on the test set.

HATE-ITA

Installing

```
pip install -U hate-ita
```

Important: If you want to use CUDA you need to install the correct version of the CUDA systems that matches your distribution, see PyTorch.

Features

```
from hate-ita.classifier import HateSpeechClassifier
hc = HateSpeechClassifier()

hc.predict(["ti odio", "come si fa a rompere la lavatrice porca puttana"])

>> ["hate", "non-hate"]
```

Models

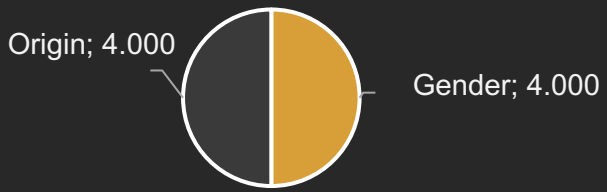
Model	Link	Macro F1 on Test Set
HATE-ITA	https://huggingface.co/MilaNLProc/hate-ita	0.83
HATE-ITA-L	TBD	TBD

https://github.com/MilaNLProc/hate-ita

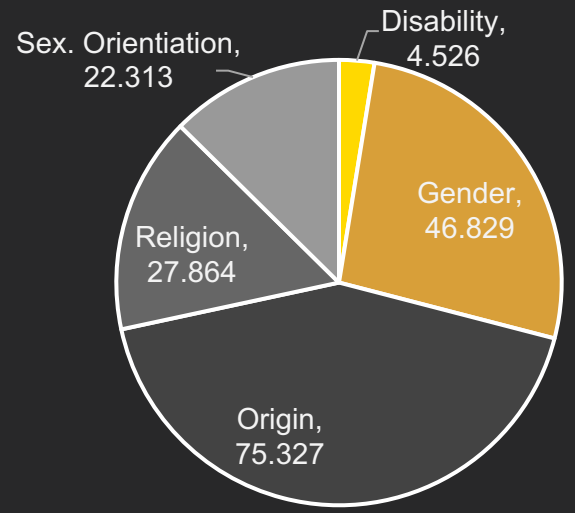
Data

Fersini et al., 2018
Sanguinetti et al., 2018

Italian



English



~22x

Ousidhoum et al., 2019
Kennedy et al., 2020a
Kennedy et al., 2020b
Mathew et al., 2021
Kiela et al., 2021
Mollas et al., 2022

NLP models

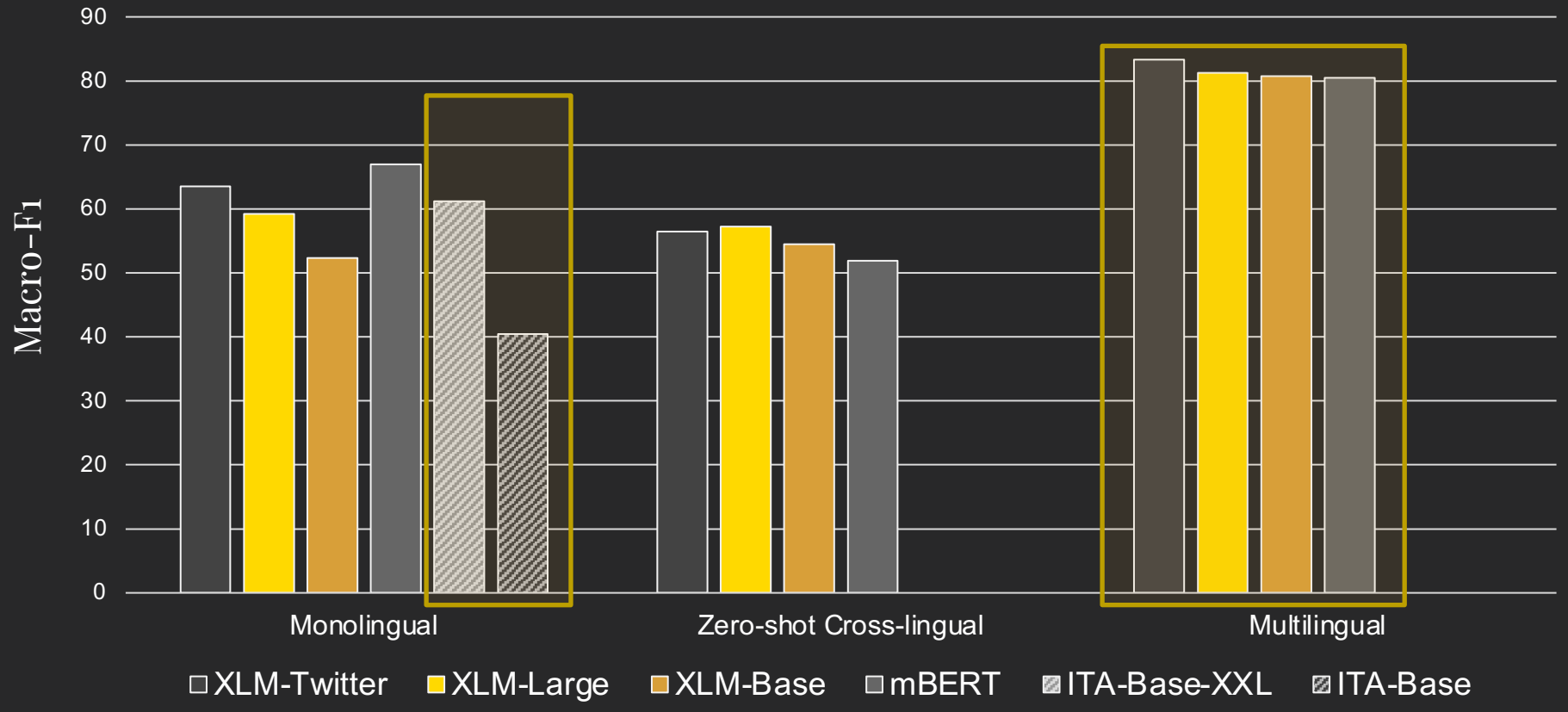
Italian

Bert-Base-Italian
Bert-Base-Italian-XXL

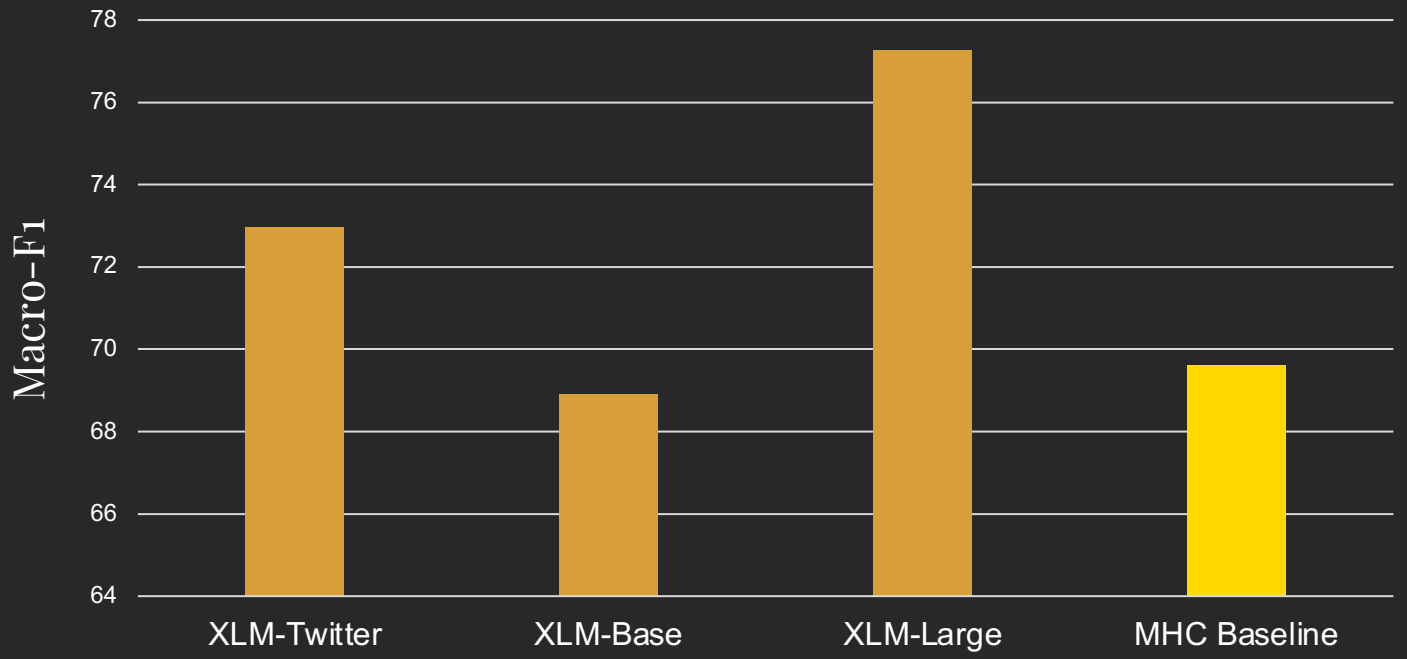
English

XLM-Roberta-Twitter-Base
XLM-Roberta-Base
XLM-Roberta-Large
Multilingual BERT

Results – Test set



Results – Italian HateCheck



Take-home points

HATE-ITA

- We release a suite of multi-language models for hate speech detection in Italian Social Media
- We demonstrate that the joint learning is beneficial and necessary
- We aim to encourage more research in other mid-to-low resource communities

Thank you!

debora.nozza@unibocconi.it

@debora_nozza