Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection

Debora Nozza
Bocconi University, Milan

debora.nozza@unibocconi.it
@debora_nozza
This presentation contains examples of offensive language; they do not represent the views of the author.
Challenges

1) Hate Speech covers a wide range of target types

2) Lack of consistency across available corpora

3) Research is mainly conducted on English
1) Hate Speech covers a **wide range of target types**

2) **lack of consistency** across available corpora

3) Research is mainly conducted exclusively **only on English**

*Zero-shot, cross-lingual solutions based on multilingual contextual embeddings!*
But is it?

Zero-shot, cross-lingual learning hate speech detection does not transfer to different target types and is strongly limited by the high presence of language- and target-specific taboo interjections in non-hateful contexts.

how the hell can you break the washing machine
Zero-shot, cross-lingual learning hate speech detection does not transfer to different target types and is strongly limited by the high presence of language- and target-specific taboo interjections in non-hateful contexts.
Experimental Settings

- **Zero-shot Cross-lingual VS Monolingual**

- **Dataset**: benchmark corpora of hate speech against *immigrants* and *women* in Italian, English, and Spanish (Bosco et al., 2018; Fersini et al., 2018; Basile et al., 2019)

- **Multilingual contextual embeddings**: multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-R (Conneau et al., 2020)
Results: HS against immigrants and women

Training language

- train-IT
- train-EN
- train-ES

Test languages:
- test-IT
- test-ES
- test-EN
Results

Training language
- train-IT
- train-EN
- train-ES

HS against immigrants
HS against women
Results

Performance does not transfer to different hate speech target types.
ma poi come si fa a rompere la lavatrice porca puttana

Correct prediction by monolingual model

Misclassified prediction trained on English and Spanish data

ma poi come si fa a rompere la lavatrice porca puttana
Take-home points

◈ Zero-shot, cross-lingual transfer learning is not a feasible solution for hate speech detection, which is language- and target-specific

◈ Few-shot, cross-lingual learning provides improvements (+16%)

◈ There is the need of more conscious research
Thank you!

debonozza@unibocconi.it
@debora_nozza