FACULTY OF ARTS
AND PHILOSOPHY

LANGUAGE AND TRANSLATION TECHNOLOGY TEAM – LT³

Nina Bauwelinck & Els Lefever

HATE SPEECH DETECTION ON TWITTER

Defining hate speech

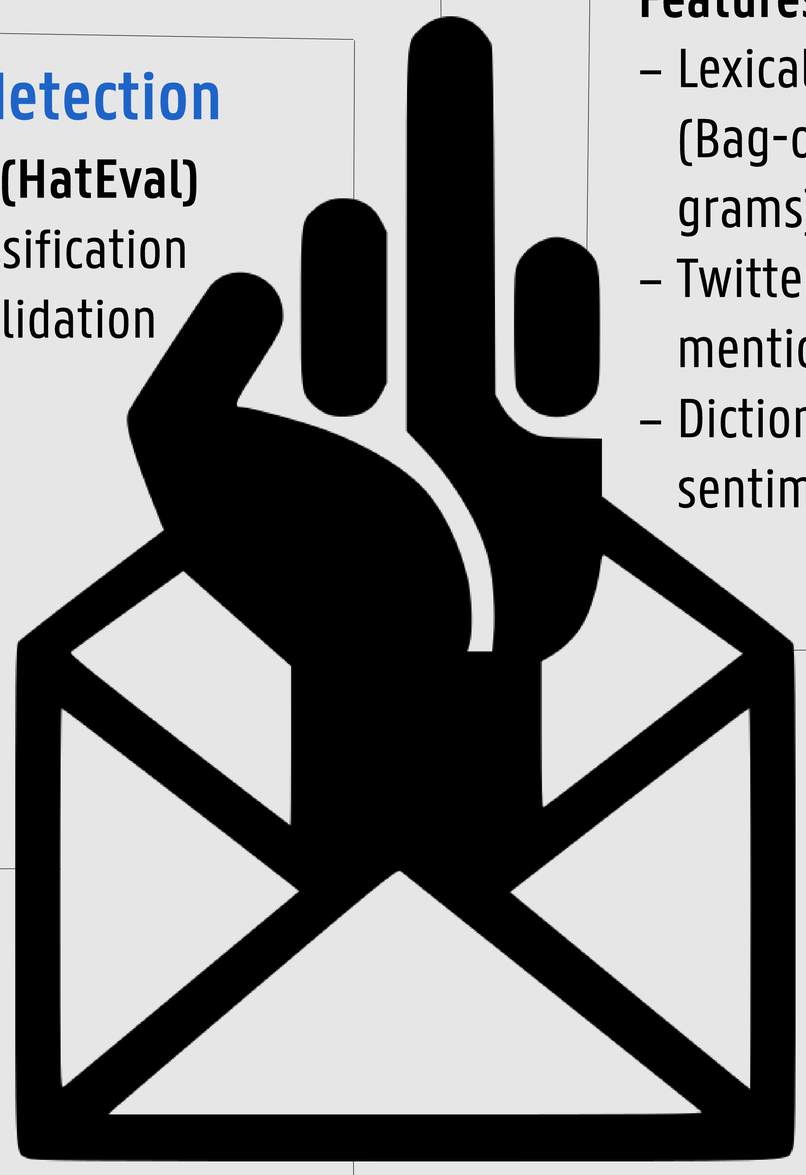
- Insulting, intimidating, harassing (different intensities).
- Individuals or groups,
- Race, gender, sexual orientation, religion, nationality, etc.

Twitter Policy

- Hateful Conduct Policy
- Abusive Behaviour Policy

Automatic detection

SemEval 2019 (HatEval)
Supervised classification
5-fold Cross-Validation



Features

- Lexical surface level features (Bag-of-words / character and token n-grams)
- Twitter-specific features (hashtags, URLs, mentions)
- Dictionary features: syntactic and sentiment information

Results

Strong baseline (77,71% F-score)
Twitter-specific features best (78,59% F-score)
Sentiment features underperforming

Future research

- Features: extra-linguistic and context
- Expanding profanity lexicon
- Offenders' strategies for avoiding detection
- Related subtasks: directed vs. generalized hate speech; othering language.

Recurring errors

Metaphors

Open your mouth & take the **meat** like a hoe you bitch ass

Consensual use

happiest of birthdays to the main hoe I hope you have a wonderful day angel USERNAME

Self-directed

@USERNAME 1o million, one cent less, i am a liei8ng son of a bitch and my mom is a whore

Quotes

Vines: Bitch disgusting

Memes: Fat whore!!! Ugly bitch!!! Me: URL

References

V. Basile et al., "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics, 2019, pp. 54–63.
N. Bauwelinck, G. Jacobs, V. Hoste, and E. Lefever, "Lt3 at semeval-2019 task 5 : multilingual detection of hate speech against immigrants and women in twitter (hateval)," in Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics, 2019,
N. Bauwelinck & E. Lefever, "Measuring the Impact of Sentiment for Hatespeech Detection on Twitter," Proceedings of HUSO 2019, The Fifth International Conference on Human and Social Analytics, IARIA, 2019, Rome, Italy.

Contact

nina.bauwelinck@ugent.be
www.lt3.ugent.be

Universiteit Gent
@ugent
Ghent University

GHENT
UNIVERSITY