UNIVERSITÄT DUISBURG ESSEN
Open-Minded

What does this imply? – Examining the Impact of Implicitness on the perception of Hate Speech

Darina Benikova, **Michael Wojatzki & Torsten Zesch**

Doctoral Researcher
Language Technology Lab
University of Duisburg-Essen

LANGUAGE TECHNOLOGY LAB

Warning and disclaimer

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

TRIGGER WARNING: Slides contain expressions of racial and religious discrimination and violence.

Disclaimer: The used text pieces do not in any way reflect the views or opinions of the authors

Real world

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

German politician says migrants should be shot at border if they're trying to enter country illegally

- Frauke Petry, 40, said German border police should shoot illegal migrants
- Leader of right-wing party added that use of armed force was 'last resort'
- Alternative for Germany (AfD) party now has 11 per cent support in polls

Social Media Reaction

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Shoot and kill in case of illegal border crossing #afd

Shoot and kill immigrants in case of illegal border crossing #afd

Hate Speech - Example

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Shoot and kill in case of illegal border crossing #afd

Implicit

Shoot and kill immigrants in case of illegal border crossing #afd

Explicit

Hate Speech – Implicit Stance

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Implicit stance
Target of sentiment is **not** explicitly mentioned

Here: negative stance against immigrants

Shoot and kill in case of illegal border crossing #afd

Hate Speech – Explicit Stance

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Explicit stance
Target of sentiment is explicitly mentioned

Here: negative stance against immigrants

Shoot and kill immigrants in case of illegal border crossing #afd

Real world

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Germany train crash: Deaths as two trains collide

Police say at least ten people killed and scores injured as trains collide near the town of Bad Aibling in Bavaria.

Hate Speech - Example

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Are Mustafa and Ali responsible for the train accident?

Implicit

The Muslims are responsible for the train accident!

Explicit

Why does it matter?

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Bundestag passed a law against hate speech in June 17

- Forbids hate and slander in social media
- Obliges operators to delete hate speech within 24 h

Delete Hate Speech or Pay Up, Germany Tells Social Media Companies

By MELISSA EDDY and MARK SCOTT JUNE 30, 2017

RELATED COVERAGE

Twitter Fails E.U. Standard on Removing Hate Speech Online MAY 31, 2017

Facebook and Twitter Could Face Fines in Germany Over Hate Speech Posts MARCH 14, 2017

Facebook Runs Up Against German Hate Speech Laws NOV. 28, 2016

Motivation for our Study

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Implementation of this law has consequences for CL/NLP:

Problems with definition of hate speech

- [Ross et al. 2016] show that definition has little influence on perception

Influence of implicit/explicit dimension largely unknown

→ Focus of this study

Does implicitness affect the perception of hate speech?

Source Corpus

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Source corpus [Ross et al., 2016]

- 541 German tweets
- Binary annotation if hate speech or not
- Gradual annotation of abusiveness (target: Muslims and immigrants)

Are Mustafa and Ali responsible for the train accident?

Creating the Dataset

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

1. Identify implicit tweets
2. Paraphrasing rules to make hate speech explicit
3. Compare perceived intensity

Are Mustafa and Ali responsible for the train accident?

The Muslims are responsible for the train accident!

Filtering and Annotating the Corpus

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

- Filtering process for hateful implicit Tweets
  → 36 Tweets

- Implicit tweets paraphrased to explicit according to rules
  → 72 instances

- This set was annotated by 100 participants

Re-annotation similar to [Ross et al. 2017]

- Binary annotation of hate speech
- Gradual annotation of abusiveness

Filtering process

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Only tweets marked as hate speech by >=1 annotator in [Ross et al. 2016])

These were further filtered for implicit:

- We found that **rapefugee** has strongest correlation with hate speech
- Strong association for cognates (rapist, rape, rapes, …)

http://iconizer.net/files/Plastic_XP/orig/filter_data.png

Paraphrasing Rules

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

If an **implicit** hateful comment is

- Softened through hedges or epistemic modality in 1st per. Sg.
- Implicit generalizations
- Rhetorical questions
- Contains implicit stance

Then we make it more **explicit** by

- Deleting hedges or making them more explicit
- Make it explicit through the use of quantifiers
- Reformulate them to statements
- Make stance explicit

Conversion Process

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Implicit

Are Mustafa and Ali responsible for the train accident?

Rules

Explicit

- Ex.1 […] -*Are **Muslims** responsible for the train accident?*
- Ex.2 […] -*Mustafa and Ali are responsible for the train accident!*
- All rules - ***The Muslims** are responsible for the train accident!*

Human study

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Anonymous online survey

Two survey sets (explicit/implicit)

Participants (=annotators)

- given definition of hate speech
- Binary rating of hate speech
- 6-point scale Likert rating of abusiveness

Results of Human Hate Speech Annotation

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Binary Hate Speech Rating:

- 31.6% rated as hate speech in explicit condition
- 40.1% rated as hate speech in the implicit condition

High standard deviation for both conditions (aligns with findings of [Ross et al. 2016])

χ^2-test: no significant difference in distribution of two conditions

Individual instances: several significant polarity changes

Intensity of Perceived Hate Speech

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

- T-test: no statistically significant differences between the explicit and the implicit condition (on average)
- Detailed analysis on instance level:

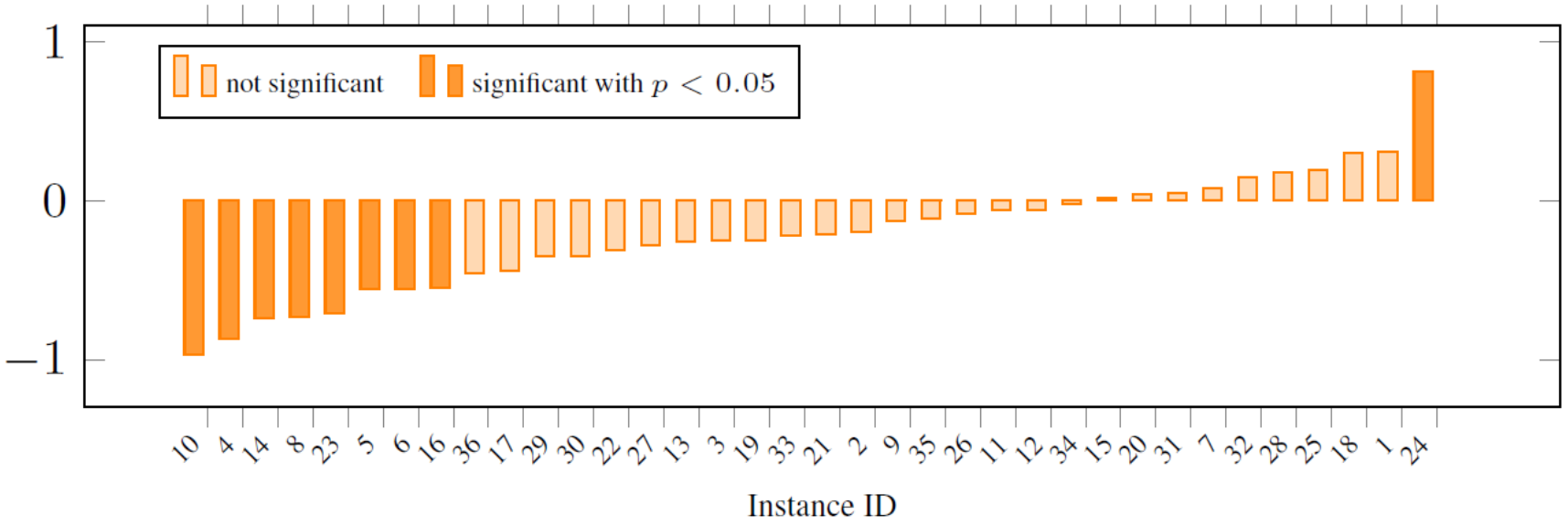


Figure 1: Change in hate speech intensity between implicit and explicit versions.

Results of human hate speech detection

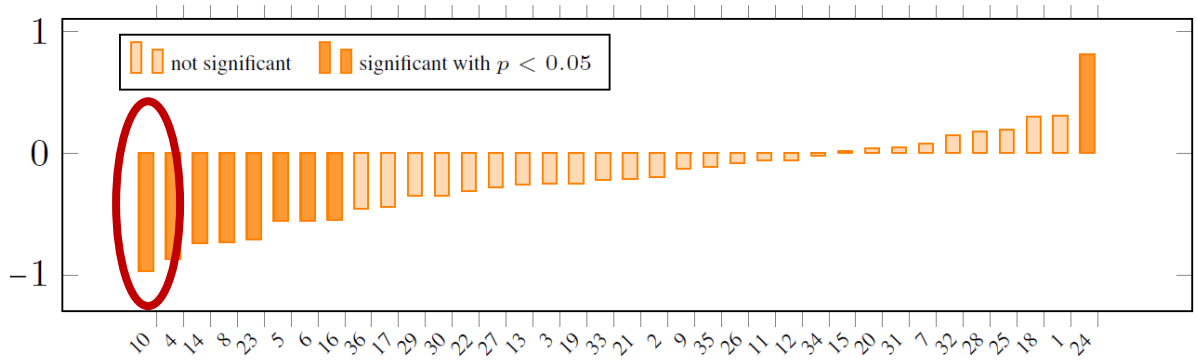
UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Wir brauchen einen Aufstand der Anständigen in Deutschland. #Abmerkeln #Abschieben #Grenzkontrollen #Islamisierung stoppen

Implicit

Wir brauchen einen Aufstand der Anständigen in Deutschland. #Abmerkeln **Muslime** #Abschieben #Grenzkontrollen #Islamisierung stoppen

Explicit



Results of human hate speech detection

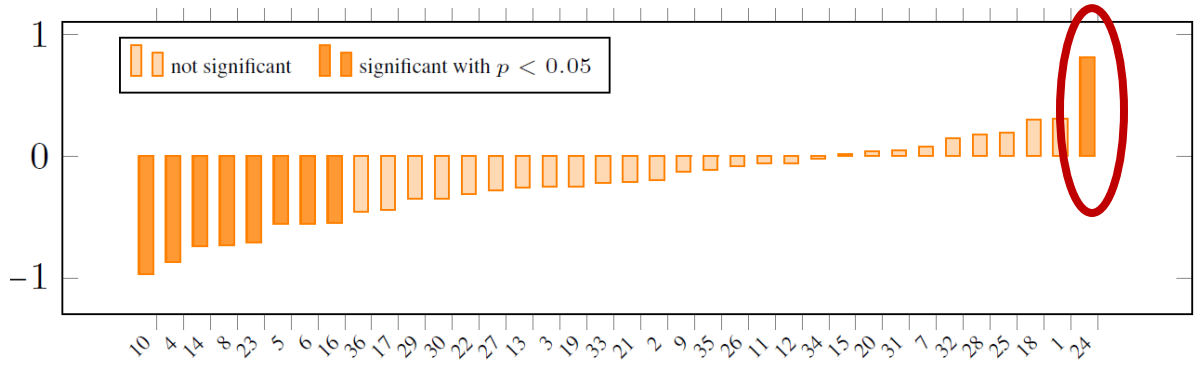
UNIVERSITÄT DUISBURG ESSEN
Open-Minded

#Sonderrecht für #radikale, #kriminelle #Islamisten installieren? -nach d #Scharia aburteilen? -#Diebstahl: Hand ab? -schnell Ruhe im Karton

Implicit

kein #Sonderrecht für #radikale, #kriminelle #Islamisten installieren**!** -nach d #Scharia aburteilen**!** -#Diebstahl: Hand ab**!** -schnell Ruhe im Karton

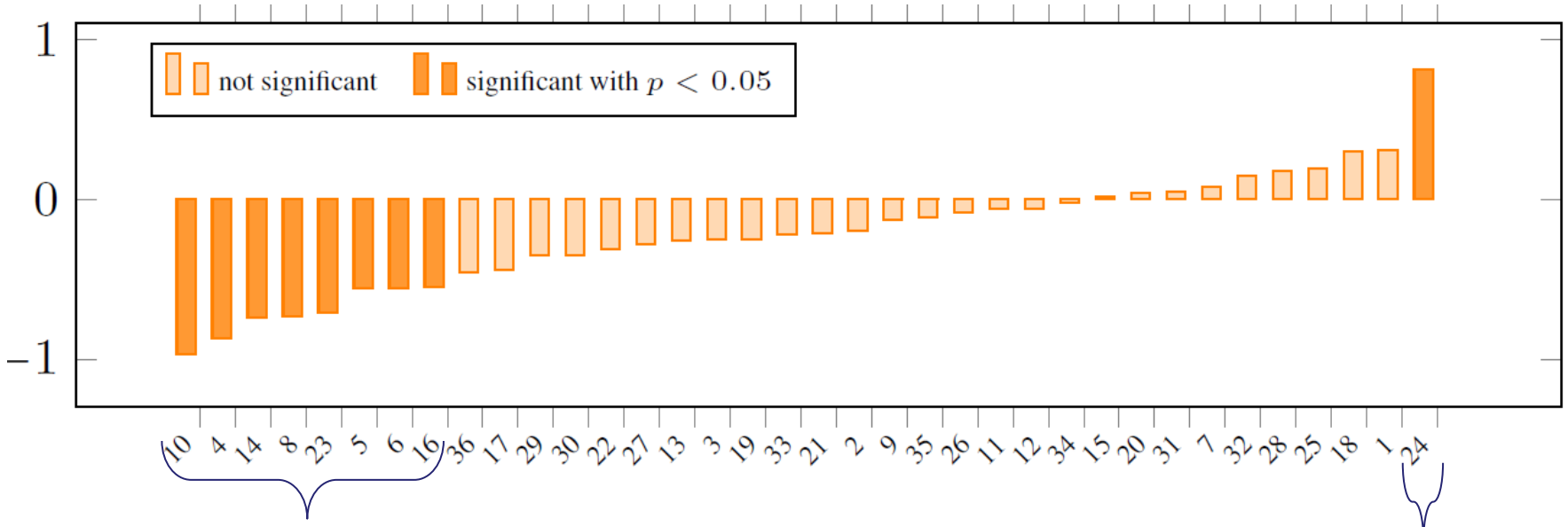
Explicit



Results of human hate speech detection

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

- Intensity of Hate Speech:
- Detailed analysis
  - Change in HS intensity between implicit and explicit versions:



The implicit version is more global and less directed

only instance containing direct threat of violence in explicit version

Results of human hate speech detection

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

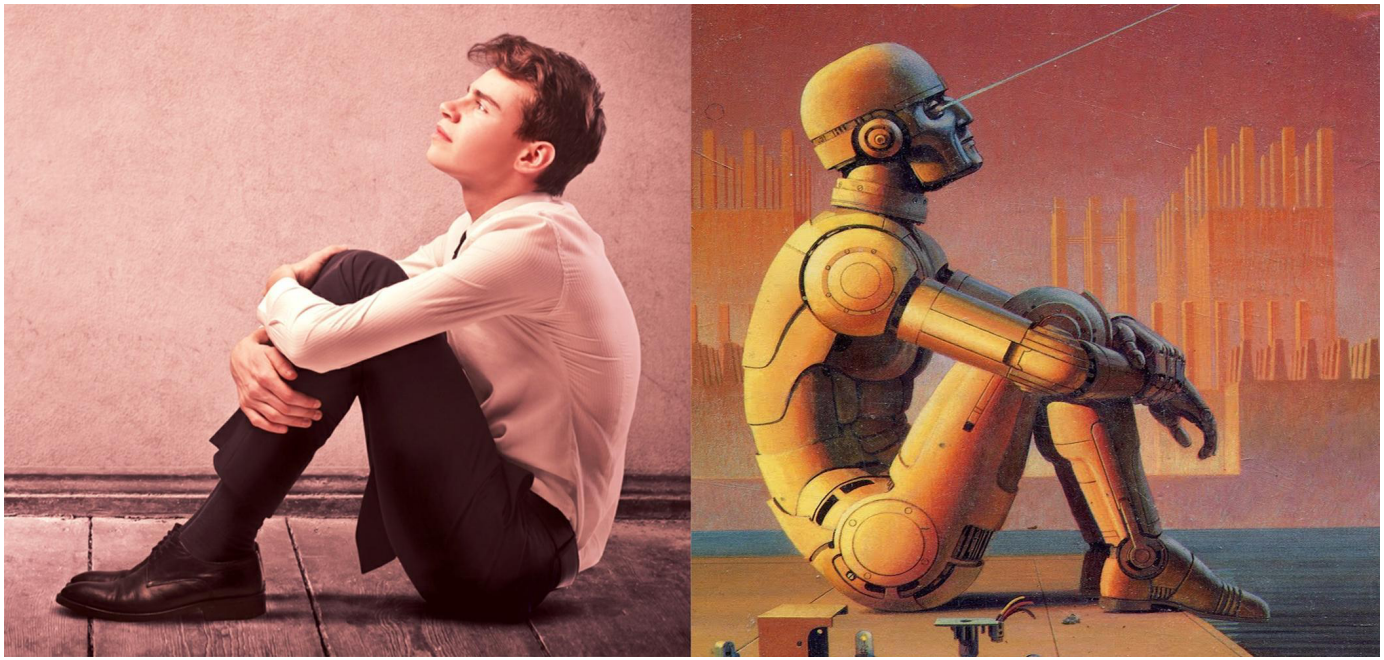
- Intensity & Binary Rating:
- 3 of the 8 significantly less offensive explicit stances are also significantly less often considered being hate speech in the binary decision
- instance 24, which is perceived significantly more offensive is more frequently considered as hate speech

- relationship between the offensiveness and the hate speech rating
- both are equally affected by implicitness

Hate speech detection: humans and machines

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Are automatic techniques as affected by implicitness in hate speech as humans?



https://www.americaninno.com/boston/machine-learning-marketing/

Supervised Machine Learning

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Examine the influence of implicitness on automatic hate speech detection:

Adaptation of SoA-system (Waseem and Hovy, 2016)

- SVM; features: type-token ratio, emoticon ratio, character, token, and POS uni-, bi-, and trigams features

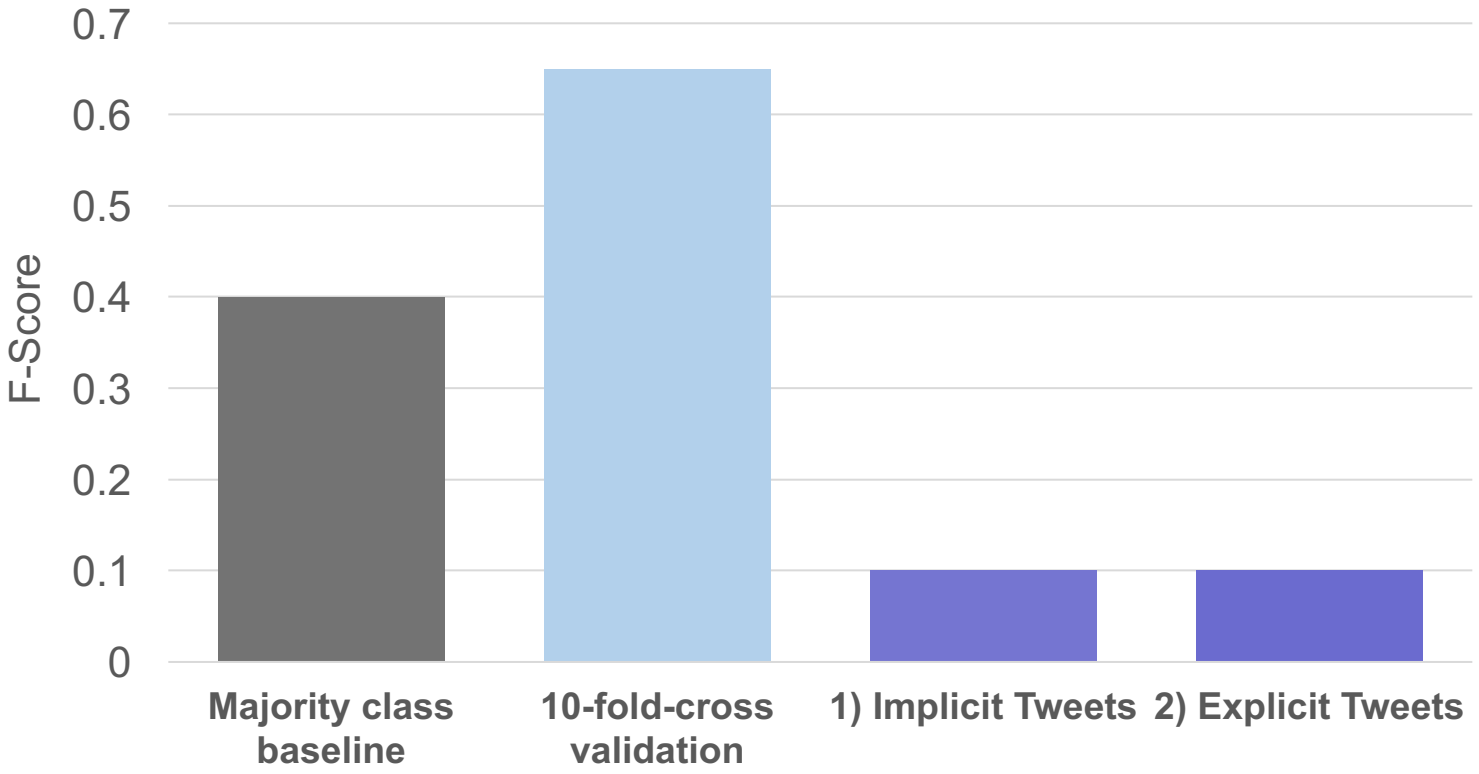
Evaluation of system output:

- Using Ross et al. corpus
- Comparison with majority class baseline
- 10-fold cross validation
- Train/test split with implicit instances

Results of Automatic Hate Speech Detection

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

- Implicit tweets are especially hard to detect
- Classifiers are blind for implicit-explicit shift



Conclusions

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Implicitness affects rating of hate speech

The phenomenon is invisible to automatic classifiers.

→ Severe problem for automatic hate speech detection, as it opens door for more intense hate speech hiding behind the phenomenon of implicitness

Delete Hate Speech or Pay Up, Germany Tells Social Media Companies

Conclusions

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Implicitness affects rating of hate speech

The phenomenon is invisible to automatic classifiers.

→ Severe problem for automatic hate speech detection, as it opens door for more intense hate speech hiding behind the phenomenon of implicitness

Direction of relationship needs further investigation:

- When implicit version is perceived as more hateful, the Tweets were rather insulting than threatening.

Future Work

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

- Larger Dataset (Both data and participants)
- More diverse participant group
- Annotating source of hate speech e.g. threat, insult
- Paraphrasing of explicit hate speech to implicit hate speech

Sources

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of HateSpeech Annotations: The Case of the European Refugee Crisis. In Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, pages 6–9, Bochum, Germany.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of NAACL-HLT , pages 88–93.

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Thank you for your attention!

Any questions?

Data and Guidelines:
https://github.com/MeDarina/HateSpeechImplicit