



Case Studies of Hate Speech on Social Media: Analysis and Automatic Detection

Sylvia Jaki

jakisy@uni-hildesheim.de

[@sylviajaki](https://twitter.com/sylviajaki)

Die Mutter aller Probleme! Das Problem aller Opfer!

Cyberbullying

Profanity

Dangerous Speech

HATE SPEECH

Flaming

“speech that increases the risk for violence targeting certain people because of their membership in a group, such as an ethnic, religious, or racial”
(Brown 2016, 7)

“any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic”
(Nockleby 2000)

“messages showing attributes such as hostility, aggression, intimidation, insults, offensiveness, unfriendly tone, uninhibited language, and sarcasm”
(Turnage 2007, 44)

Gendertrolling

Shitposting

Part I: Introduction of the case studies



Misogynist hate
speech in the forum
Incels.me

German right-wing
hate speech on
Twitter

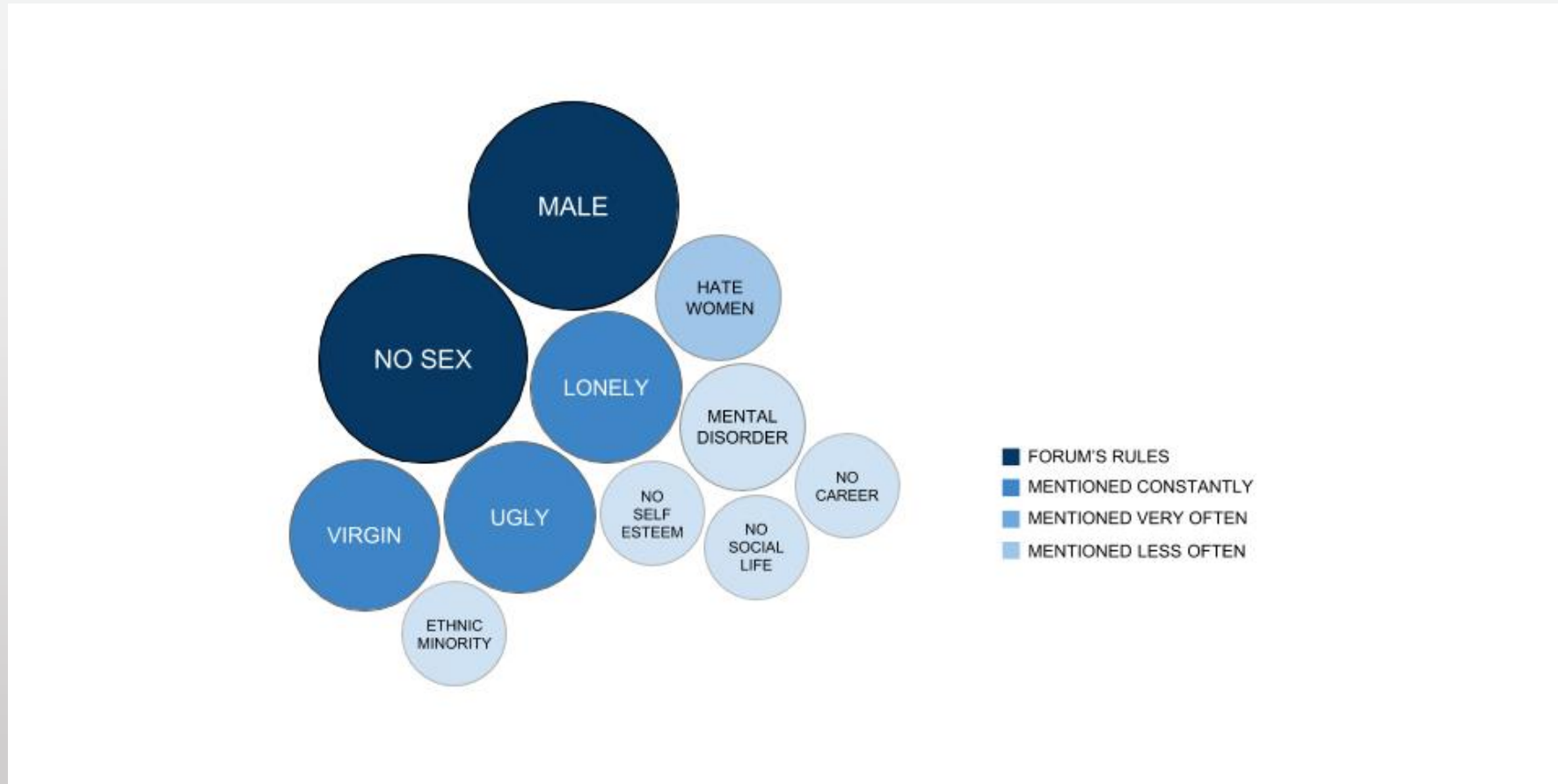
Facebook
comments before
the German federal
elections 2017

1 Misogynist hate speech on incels.me

Involuntary celibates (incels)

- On April 23 2018, 25-year old Alek Minassian killed 10 and injured 16 in Toronto by driving a van into pedestrians
- Shortly before the attack, he posted a message on Facebook stating: “The Incel Rebellion has begun!”
- Incels = involuntary celibates: (often adolescent) men, mainly from US
- Many incels experience loneliness, desperation, suicidal thoughts
- They attribute their lack of success with women to the “degeneracy” of women and good-looking men
- Result: online forums with highly inflammatory comments, extremism, radicalisation

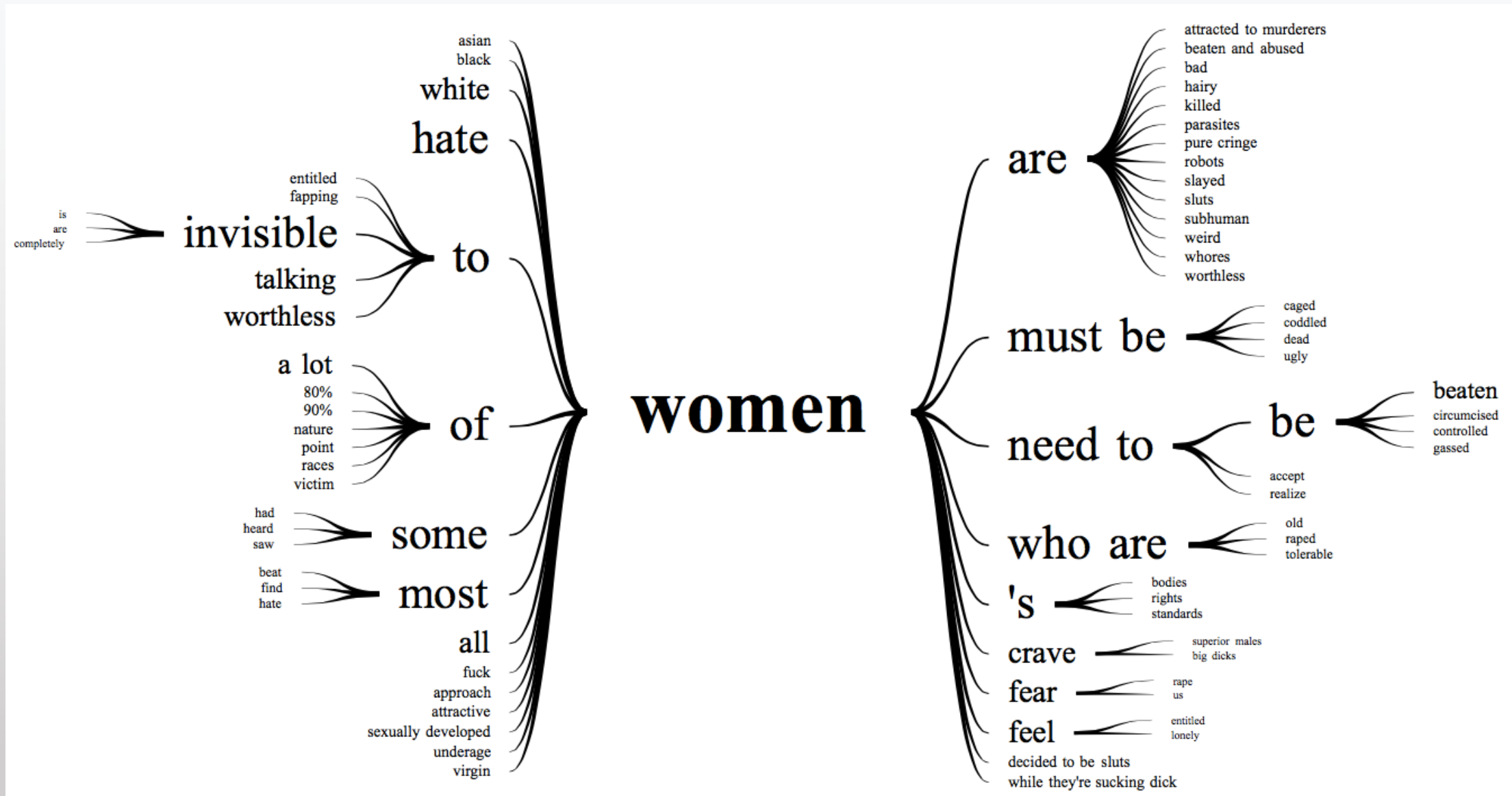
Core and peripheral aspects of incelism



Jaki et al. (2018)

Data set and observations

- 50,000 messages from the forum incels.me
- The domain .me suspended the forum in October 2018 for spreading violence and hate speech
- About 30% of the threads we analysed were **misogynist**; about 15% were **homophobic**, and 3% were **racist**
- Hatred of women as well as feminism as one of the most important discussion topics
- Large number of **disparaging designations for women** (*cum dumpster, cum rag, it, roastie, slut, whore*, etc.).



- Examples

Female HAVE TO become property again. They should not have the right to even SPEAK without male permission

Better a few dead than all of them living their carefree lives. Enjoy the little things

I loathe fat women. Bunch of useless fucking hogs

Roastie slut got what she deserved

- A lot of features of incel misogyny extends to online misogyny in general, such as detailed depictions of sexual violence, extreme insults, or the shaming of women for bodily flaws
- The forum also contains allusions to potential terroristic attacks

I have no problem if any of us starts killing as many people as possible. The more young women, Chads, Chadlites, normies, cucks, and Boomers who get slaughtered, the better


2 Right-wing hate speech on Twitter

Data Set

- 55,000 hateful tweets from 112 users (collection period: August 2017 – April 2018)
- Manual selection of Twitter profiles and extraction of their tweets
- Right-wing background is often directly visible in the profiles
 - ✓ coded numbers or abbreviations
 - ✓ references to German mythology
 - ✓ pictures of Nazi frontmen
 - ✓ profile descriptions with anti-Islam messages, references to White Power, Reconquista Germanica, etc.

Twitter, Inc. [US] | https://twitter.com/


Home About Search Twitter Have an account? Log in



Tweets 271 Following 39

Joined January 2016

93 Photos and videos



New to Twitter?

Twitter, Inc. [US] | https://twitter.com/

Home About Search Twitter Have an account? Log in



Tweets 1,141 Following 201

Die Schonzeit ist vorbei! Nationalen Sozialismus durchsetzen! Mit allen Mitteln & auf allen Ebenen! #NSjetzt

Joined November 2014

114 Photos and videos



Twitter, Inc. [US] | https://twitter.com/

Home About Search Twitter Have an account? Log in



Tweets 254 Following 104 Followers 14 Likes 73

Follow

Thule

Joined February 2017

Photos and videos

New to Twitter?
Sign up now to get your own personalized timeline!

Tweets Tweets & replies Media

Retweeted

Dec 15

Alle 18 Hundertschaften (5.700 Beamte) der Bereitschaftspolizei #NRW sind am 31. Dezember 2017 im Dienst. Das gibt ganz gut wieder, in welche Zustände sich das Land befindet. Mehr Worte braucht es im Grunde nicht.



Tweets Tweets & replies Media

Pinned Tweet

Mar 30

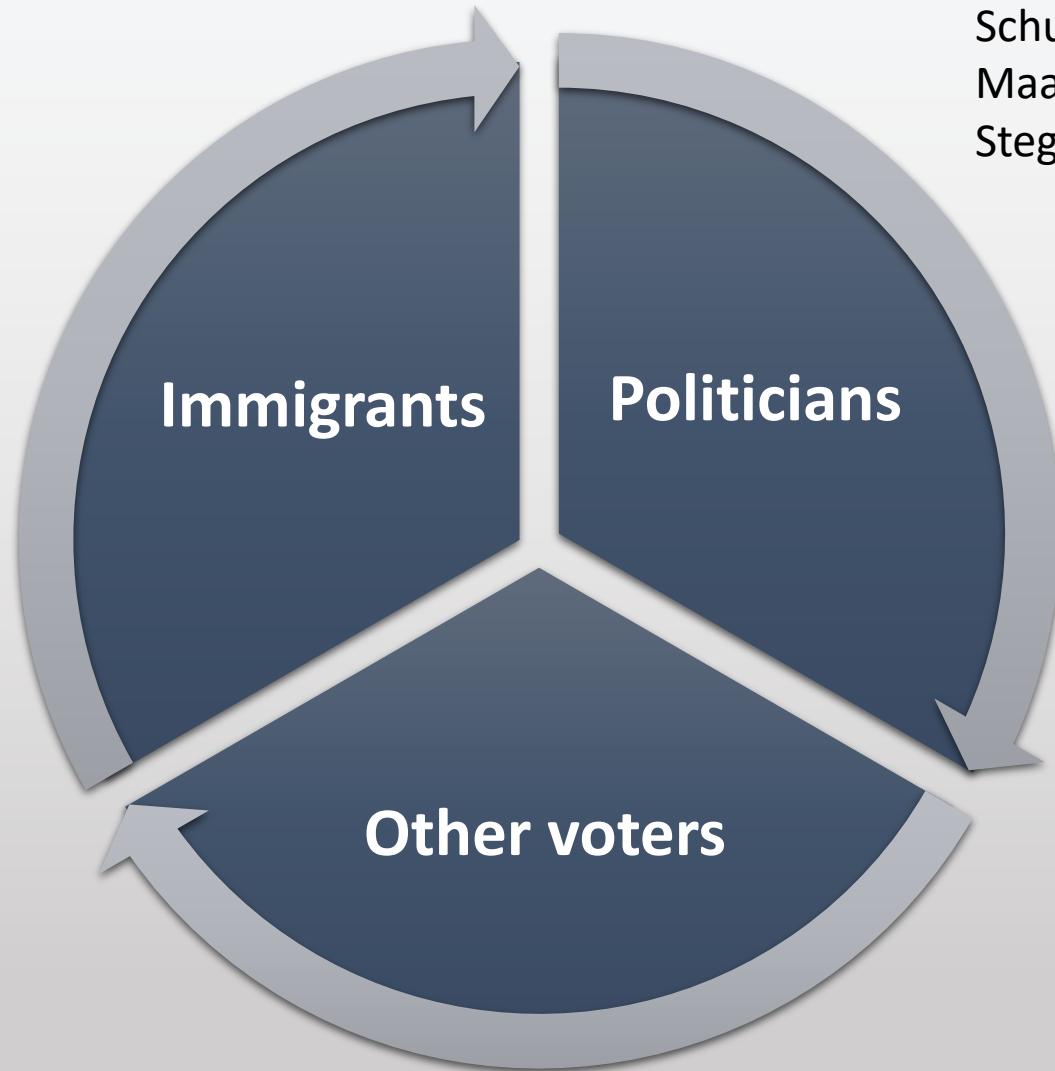
Wir kämpfen in der Nacht für bessere Tage! #Widerstand #NSjetzt #AntiKap #ResistCapitalism #KapitalismusZerschlagen



Targets

*Nafris, Invasoren, Asyltouristen,
Merkel-Gäste, Illegale,
Wohlstandsflüchtlinge, Bunte,
Zudringlinge, Abfall, Müll, Abschaum,
Pack, Parasiten, Gesindel;
Islamesianer, Salafistenschwester,
Kampfmuslimas, Burka-Frauen,
Vollbärte, Mutombo, Bongo, Kloneger
...*

The ten nationalities that were
mentioned most often in hate tweets:
German, African, Turkish, Israeli,
Syrian, Russian, Afghan, Saudi-Arabian,
Romanian, Iraqi, Moroccan



Individuals such as Angela
Merkel (*Volksverräterin*,
Bauerntrampel), Martin
Schulz (*Arschkriecher*), Heiko
Maas (*Vollpfosten*), Ralf
Stegner (*Einzeller*) ...

Left-wingers in general
(*Verbrecher, Sozi Clowns*,
linkes Faschistenpack,
Grünfaschisten ...)

Linksfaschisten, Traumtänzer, Gutmenschen ...

Language

- **Dehumanising metaphors** (*Abfall, Abschaum, Müll, Parasiten*)
- **Profanity** (*diese scheiß Zecke, Besser tot als mit Kacke am Arsch leben*)
- **Stereotyped compounds** (*Museldiebe, Asyltouristen, Kanakenstadt*)
- Appearance of **stigma words**, especially as hashtags (*Rapefugees, Krimigranten, Buntland, Religioten*)
- **Resemantisation** of words (*bunt, Fachkräfte*)
- **Capitalization** for intensification (*DEUTSCH, ISLAM, LINKS*)

Gibt es nen Unterschied?

#Flüchtlinge = #Krimigranten

#Merkels #Fachkräfte haben mal wieder einen #Asyl-#Gangbang durchgeführt und zu acht ein 13-jähriges #deutsches #Mädchen #vergewaltigt. Von #Reue keine Spur. Sie meinten: "Sie hat es doch gewollt!" - Unfassbar, was in #Buntland mittlerweile los ist.

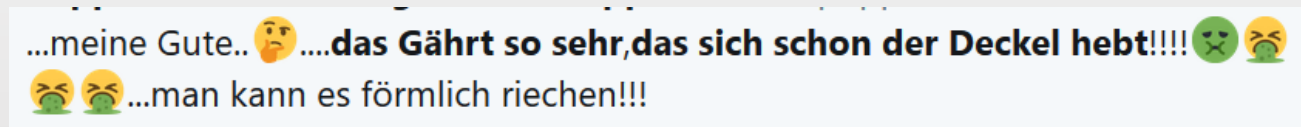
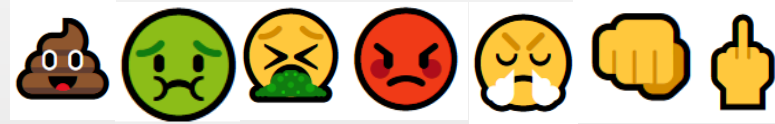


Asyl-Gang-Bang in NRW: 8 Fachkräfte vergewaltigen 13-Jährige und ...
OMAL GETEILT 0 0 0 0 0 0 0 0 Nachdem 8 junge Ausländer bereits im April dieses Jahres ein 13-jähriges Mädchen brutal vergewaltigt haben, beginnt ...

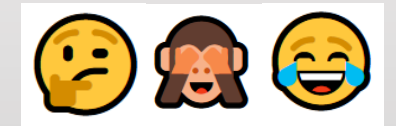
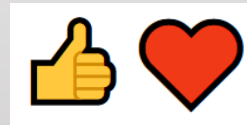
Visual elements, e.g. Emojis

- Generally a wide selection of emojis appearing in hate tweets

- Particularly frequent:



- Function: expression of emotions, intensification, judgement
- Also emojis which have more of a positive polarity in other contexts;
in part used ironically



3 Facebook comments during the federal
elections 2017

Or: Who polluted the debate?

W&SN #WasNBTW17
www.uni-hildesheim.de/wahlkampfanalyse
BTW17

A cooperation with Prof. Wolf Schünemann
(Universität Hildesheim)



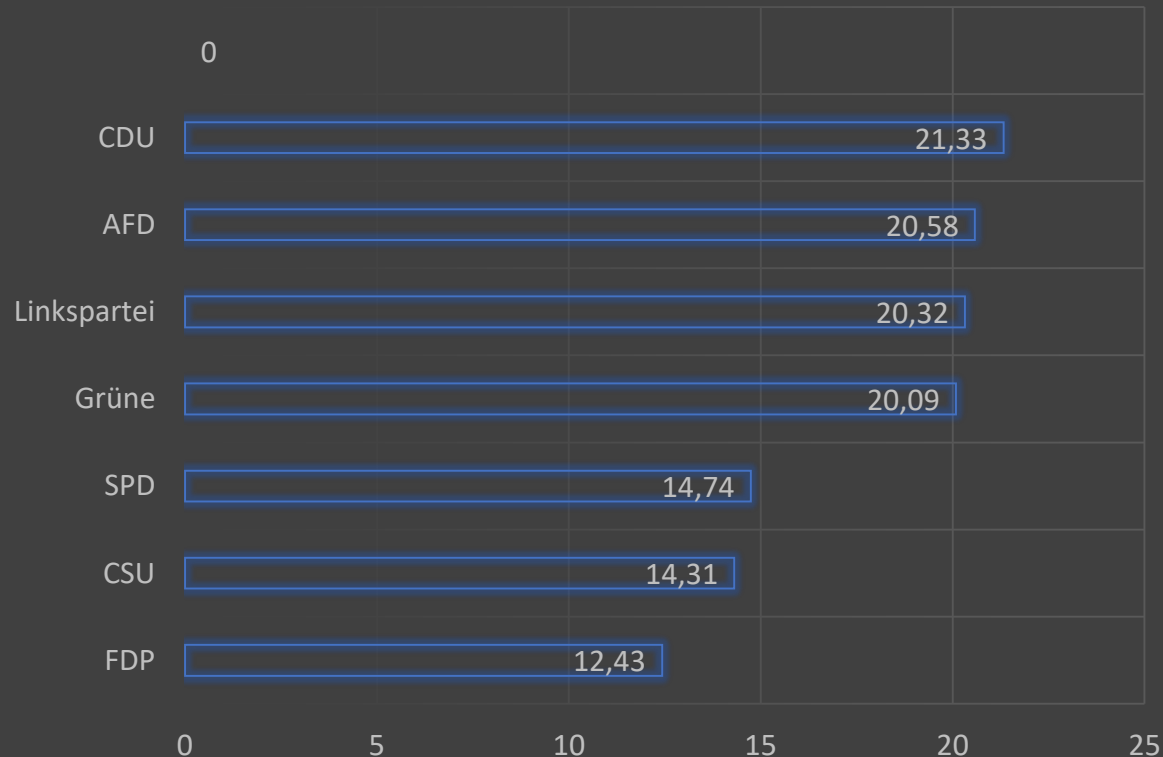
The data

- Data collected for WasN Project 1 (“Online-Diskurs-Daten aus sozialen Netzwerken”)
- 2.9 mio. comments users made on the public Facebook pages of major German parties and their leading candidates
- Period: between 29/1/2017 and 24/9/2017
- Analysis of words than can be categorised as offensive or profane as a (rough) indicator of hate speech

Party political comparison of profane and offensive words

POWs for parties

POWs 2-4 in %



POWs for politicians

POWs 2-4 in %



The targets of the hate speech

- A selective comparison of the following profiles:
 - ✓ Katrin Göring-Eckardt
 - ✓ Alice Weidel
 - ✓ B90/Grüne
 - ✓ AfD
- Observation: Hate Speech not always targeted at the person/party associated with the profile
- Hate Speech can also be used to support the person/party (AfD!)



Overview

AFD

- **Support from fans**
- Hate between users
- **Little politically sound argumentation**
- A lot of right-wing hate speech, but also anti-right-wing hate speech
- Promoting anxiety

B90/Grüne

- **Little support from fans**
- Hate between users
- A lot of “concerned citizens”, but also anti-right-wing hate speech

- **Support from fans**
- **Little hate speech targeted at AW**
- Hate between users
- A lot of anti-left-wing hate speech

Alice Weidel

- **Little support from fans**
- Hate between users
- **Hate speech often targeted at the party as a whole (or at Claudia Roth)**

Katrin Göring-Eckardt

Part II: Methods of hate speech detection

	Incels.me	Twitter	Facebook
Data size (messages)	50,000	55,000	2.9 mio.
Language	English	German	German
Type	Misogynist	Right-wing	Mixed
Quantitative and qualitative analysis	✓	✓	✓
Automatic detection (ML)	✓	✓	✗
Lexicon	✗	✗	✓

Facebook data: Profanity and Offensive Words (POWs)

- **Manually annotated dictionary** which allows for the quantitative analysis of hate speech in a dataset
- Decision to work with a dictionary a result of Germeval Shared Task 2018 on offensive tweets
- List of **2852 lexemes** that mainly consist of words taken from German Twitter Embeddings (Ruppenhofer 2018)
- These words are either often used tendentiously in political contexts or are vulgar/offensive

Types of words

Word classes (mainly):

- Nouns (*Lüge, Wesen, Arsch, Firlefanz*), incl. compounds (*Fremdenfeind, Lügenpresse*)
- Also: adjectives (*blöd, links-grün*) and participles (*verblendet*)
- Infinitives (*hetzen, spucken*) and imperatives (*lutsch, laber*), also verbs in 1st or 3rd person singular
- Interjections (*mimimi, boah*)

Listed separately (tokens):

- Declensions (*Dreckschwein, Dreckschweine*)
- Conjugations (*labern, laber, labert*)
- Spelling variations (*schreien/schrein, scheiß/scheiss/scheis/chice*)









Annotation of intensity

Degrees

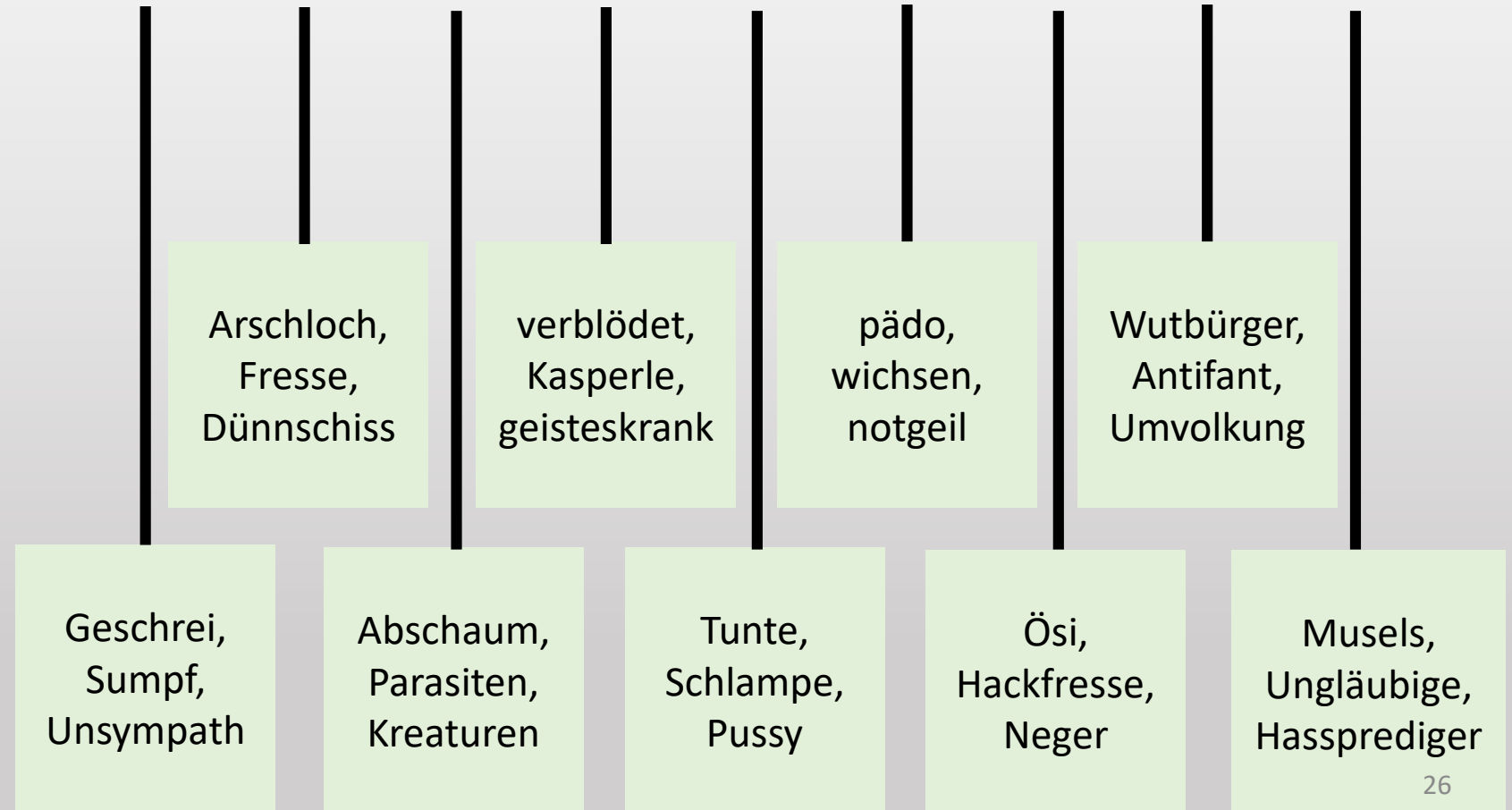
- 0 - tendentious
- 1 - tendentious, sensational
- 2 - demeaning
- 3 - offensive (vulgar, racist)
- 4 - offensive (extremely so)

Examples

- 0 - *nichtmal, religiös, AfDler, Staub, Übergriffe*
- 1 - *heulen, unkontrolliert, Extremisten, beschimpft*
- 2 - *Schnauze, stupide, Systemparteien, antideutsch*
- 3 - *verblödet, Dreck, Honk, Lügenpresse, Invasoren*
- 4 - *Hure, Untermenschen, Ungeziefer, Drecksau*

H	I	J	K	L	M	N	O	P
								
HATE	SHIT	SCUM	FOOL	SLUT	FUCK	GOOK	HEIL	HELL

Annotation of type



Difficulties

or: “You shall know a word by the company it keeps” (Firth 1957:11)

Degree

- Not entirely independent of the concrete situation in which a word is used



Intensity

*honk,
verrecken,
hurensöhne*

Polarity

*bunt,
willkommenskultur,
kulturbereicherer*

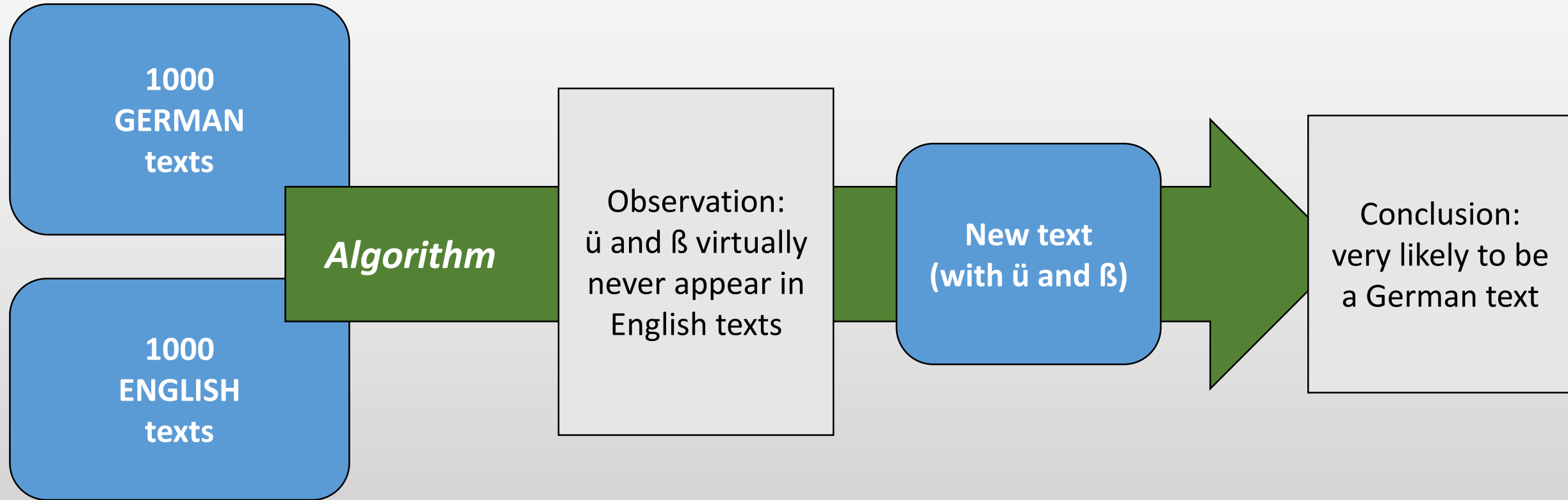
Type

- Lexical ambiguity e.g. *geil, sack, fickt, würgen, schwuler, dödel, muschi, pralle*
- Grammatical ambiguity
e.g. *quatsch, blase, leeren, ritze*

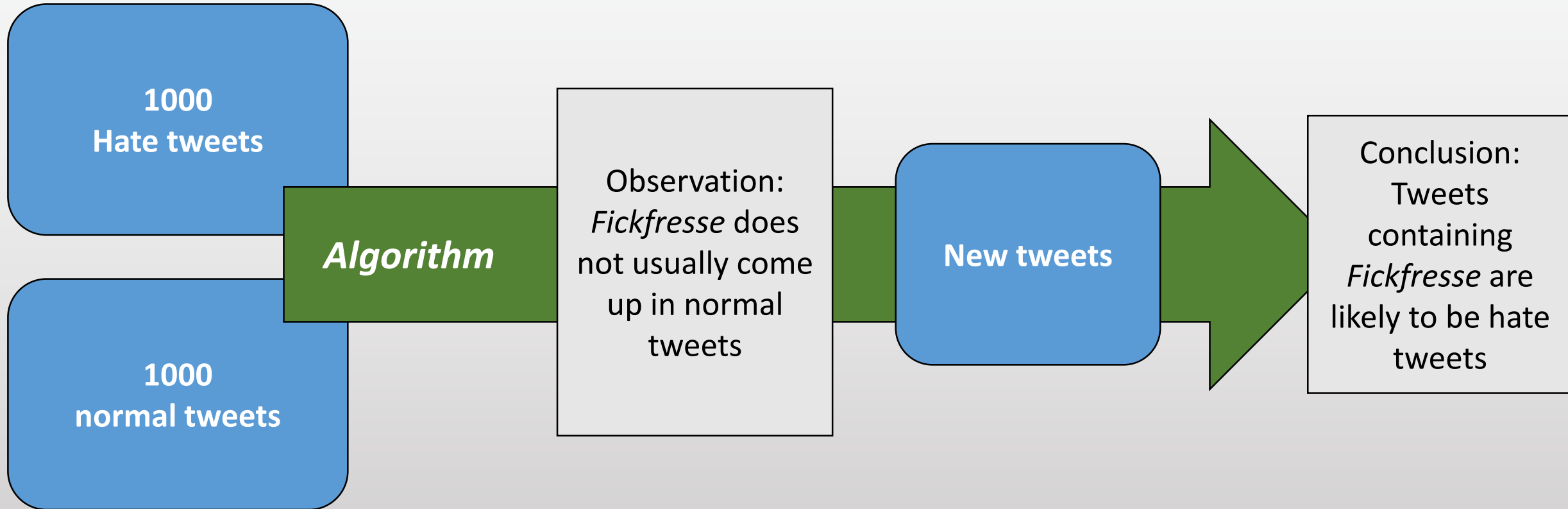
Pragmatic solution:

Possibility for contextualisation by direct link to social media

Machine Learning: A simple example



Automatic detection of hate tweets



The classifiers

- Right-wing German hate speech on Twitter:
 - ✓ 55K hateful tweets + 50K “safe” tweets (20K tweets by more than 35 elected German politicians; 30K German tweets by Twitter users, talking about family, work, holidays, etc.)
 - ✓ We trained a Perceptron machine learning algorithm
 - ✓ Feature selection: character trigrams as features; additional features such as character bigrams, character unigrams, word unigrams and word bigrams
 - ✓ Overall accuracy of **84.08%** (more specifically, precision = 84.21 %, recall = 83.97 %)
- Incel language on Incels.me:
 - ✓ 50K Incels.me messages + 50K neutral texts (40K paragraphs from English Wikipedia articles and 10K random English tweets)
 - ✓ We used a CNN (multi-layer neural network, deep learning)
 - ✓ Overall accuracy of **92.5%** (more specifically, precision = 92%, recall = 93%)

Problems

- Detection systems still **not reliable enough** ➡ risk of over- and underblocking due to false positives and negatives
- To enable a **wide applicability** of these systems, it is important to know on what basis the classifier takes decisions
- ANNs may outperform other approaches, but they are often not very transparent
- An algorithm trained on a specific type of hate speech may not perform well on a different data set (Jaki & De Smedt 2018 vs. De Smedt & Jaki 2018b)
- Automatic detection tools usually disregard the **multimodality** of hate speech in the social media

Automatic hate speech detection to counter hate speech

Legal framework in Germany

- Freedom of speech is guaranteed by Article 5 of German Grundgesetz
- However, it is restricted as soon as other people's personal rights are infringed
- Infringements of personality rights as per StGB are, for example:
 - *Öffentliche Aufforderung zu Straftaten* (incitement to criminal behaviour), §111 StGB
 - *Volksverhetzung* (incitement of the masses), §130 StGB
 - *Bedrohung* (intimidation), §241 StGB

German NetzDG (Netzwerkdurchsetzungsgesetz)

- Law that fosters the removal of illegal hate speech and fake news
- Social media platforms have 24 hours to react to messages that users reported to them
- Problem: Some cases are clearly illegal, such as *Findet diese Drecksau und entmannt sie an Ort und Stelle* (§111 StGB), but where does libel, incitement of the masses, intimidation, etc. begin?
- Our German data only show few cases which are clearly illegal

Food for thought and future directions

- When is the removal of hate speech censorship?
- How much freedom of expression do we want and need?
- Especially overblocking will lead to the impression of censorship
- The removal of hate speech may lead to even more polarisation
- The attitudes behind the comments do not disappear with a removal of the message
- Need for the combination with other approaches
- Need for explainable AI

Thank you very much
for your attention.



References

Brown, R. (2016): *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech*. <https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf>.

De Smedt, T./Jaki, S. (2018a): The Polly corpus: Online political debate in Germany. *Proceedings of CMC and Social Media Corpora*, 33-36.

De Smedt, T./Jaki, S. (2018b): Challenges of Automatically Detecting Offensive Language Online: Participation Paper for the GermEval Shared Task 2018 (HaUA). *Proceedings of GermEval 2018 Workshop*, 27-32.

Jaki, S./De Smedt, T. (2018, in prep.): Right-Wing Hate Speech on Twitter: Analysis and Automatic Detection.

Jaki, S./De Smedt, T./Gwózdź, M./Panchal, R./Rossa, A./De Pauw, G. (2018, in prep.): Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*.

Nockleby, J. T. (2000): Hate speech. *Encyclopedia of the American Constitution*. Vol. 3, 2nd ed., edited by L. W. Levy & K. L. Karst, 1277-79. Detroit: Macmillan Reference US.

Ruppenhofer, J. (2018): German Twitter Embeddings, https://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/GermanTwitterEmbeddings/GermanTwitterEmbeddings_data.shtml?fbclid=IwAR2SzbT54f_IXGvQVbLQlwV5X6yrkXIg7ZZ40mnZTjs7_BGJ_kavWnnxZ_w

Turnage, A. K. (2007): Email Flaming Behaviors and Organizational Conflict. *Computer-Mediated Communication* 13(1):43-59.