Ethics in Natural Language Processing – SS 2022

Lecture 5
InCivility and Hate Speech I

TECHNISCHE UNIVERSITÄT DARMSTADT

Dr. Thomas Arnold
Aniket Pramanik

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Slides and material from Yulia Tsvetkov

Carnegie Mellon University
Language Technologies Institute

Syllabus (tentative)

Nr.	Lecture
01	Introduction, Foundations I
02	Foundations II
03	Bias I
04	Bias II
05	Incivility and Hate Speech I
06	NO LECTURE – Christi Himmelfahrt
07	Incivility and Hate Speech II
08	Low-Resource NLP, NLP for Social Good
09	NO LECTURE - Fronleichnam
10	Privacy and Security I
11	Privacy and Security II
12	Language of Manipulation I
13	Language of Manipulation II

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

What is Hate Speech?

Effects of Hate Speech

Hate Speech Identification

Recap from last week

TECHNISCHE UNIVERSITÄT DARMSTADT

- Debiasing a dataset is NOT done by just deleting the relevant feature
  - Information often correlates with (a combination of) other features
- Equal precision OR recall does not ensure unbiased ML models
  - What is more important? What is the cost of False Positives / False Negatives?
- Bias can be amplified by machine learning
- Word Embeddings can reflect / measure social bias

Learning Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

After hearing this lecture, you should be able to…

- **give examples for characteristics of Hate Speech**
- **differentiate Hate Speech from Bias**
- **discuss potential challenges in collecting a dataset for Hate Speech**
- **describe difficulties in automatic hate speech detection**
- **give examples how abusive words can be obfuscated**

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

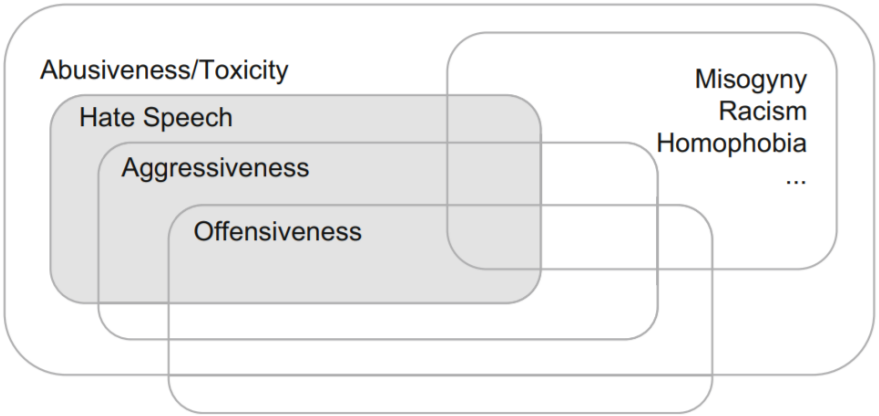
What is Hate Speech?

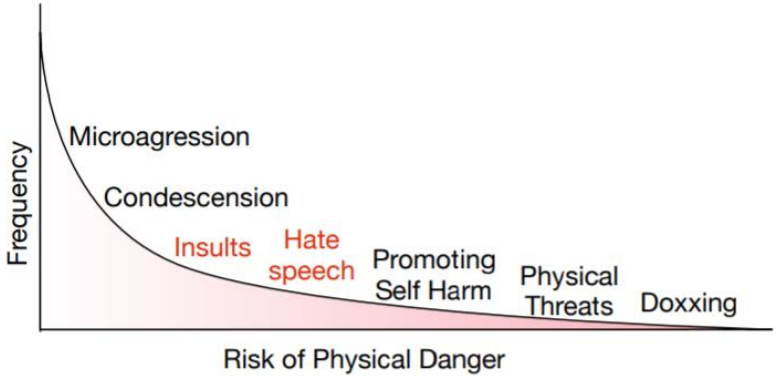
Effects of Hate Speech

Hate Speech Identification

Warning About Hate Speech Examples

Hate speech has many shades





F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti (2021) Resources and benchmark corpora for hate speech detection: a systematic review. *In Lang Resources & Evaluation*

Jurgens D., Chandrasekharan E., and Hemphill L. (2019) A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *Proc. ACL*

Hate speech has many shades

TECHNISCHE UNIVERSITÄT DARMSTADT

- Umbrella term: Abuse
- Hate speech
- Offensive language
- Sexist and racist language
- Aggression
- Profanity

- Cyberbullying
- Harassment
- Toxic language
- Trolling
- Anti-social behavior
- ...

What is Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

“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, J. Encyclopedia of the American Constitution 2000)

What is Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

"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, J. Encyclopedia of the American Constitution 2000)

TARGET

What is Hate Speech?

“language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group”

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

What is Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

*“language that is used **to expresses hatred** towards a targeted group or is **intended to be derogatory, to humiliate, or to insult** the members of the group”*

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

INTENT

What is Hate Speech?

“language that threatens or incites violence”

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

What is Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

*“language that **threatens** or **incites violence**”*

(Davidson et al., Automated Hate Speech Detection and the Problem of Offensive Language, *ICWSM* 2017)

EFFECT

What is Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

*"any offense motivated, in whole or in a part, by the offender's **bias** against an aspect of a group of people"*

(Silva et al., Analyzing the Targets of Hate in Online Social Media, *ICWSM* 2016)

THE CAUSE

How is hate speech different from the topics of bias we discussed last week?

How is Hate Speech Different from Bias?

TECHNISCHE UNIVERSITÄT DARMSTADT

Veiled vs Overt?

(Hidden vs. Open)

How is Hate Speech Different from Bias?

TECHNISCHE UNIVERSITÄT DARMSTADT

POLITICS 12/13/2017 04:00 pm ET | **Updated** Dec 13, 2017

This Is The Daily Stormer's Playbook

A leaked style guide reveals they're Nazis about grammar (and about Jews).

By Ashley Feinberg

https://www.huffingtonpost.com/entry/daily-stormer-nazi-style-guide_us_5a2ece19e4b0ce3b344492f2

How is Hate Speech Different from Bias?

TECHNISCHE UNIVERSITÄT DARMSTADT

POLITICS 12/13/2017 04:00 pm ET | Updated Dec 13, 2017

This Is The Daily Stormer's Playbook

A leaked style guide reveals they're Nazis about grammar (and about Jews).

By Ashley Feinberg

While racial slurs are allowed/recommended, not every reference to non-white should not be a slur and their use should be based on the tone of the article. Generally, when using racial slurs, it should come across as half-joking - like a racist joke that everyone laughs at because it's true. This follows the generally light tone of the site.

It should not come across as genuine raging vitriol. That is a turnoff to the overwhelming majority of people.

DAILY STORMER STYLE GUIDE

https://www.huffingtonpost.com/entry/daily-stormer-nazi-style-guide_us_5a2ece19e4b0ce3b344492f2

How is Hate Speech Different from Bias?

~~Veiled vs Overt?~~

Intentional

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

What is Hate Speech?

Effects of Hate Speech

Hate Speech Identification

Who Are Targets of Hate Speech?

Who Are Targets of Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

`I <intensity> <userintent> <hatetarget>`

*"I f*cking hate white people"*

Twitter	% posts	Whisper	% posts
I hate	70.5	I hate	66.4
I can't stand	7.7	I don't like	9.1
I don't like	7.2	I can't stand	7.4
I really hate	4.9	I really hate	3.1
I fucking hate	1.8	I fucking hate	3.0
I'm sick of	0.8	I'm sick of	1.4
I cannot stand	0.7	I'm so sick of	1.0
I fuckin hate	0.6	I just hate	0.9
I just hate	0.6	I really don't like	0.8
I'm so sick of	0.6	I secretly hate	0.7

Table 1: Top ten hate expressions in Twitter and Whisper.

(Silva et al., Analyzing the Targets of Hate in Online Social Media, ICWSM 2016)

Who Are Targets of Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

Twitter		*Whisper*	
Hate target	**% posts**	**Hate target**	**% posts**
Nigga	31.11	Black people	10.10
White people	9.76	Fake people	9.77
Fake people	5.07	Fat people	8.46
Black people	4.91	Stupid people	7.84
Stupid people	2.62	Gay people	7.06
Rude people	2.60	White people	5.62
Negative people	2.53	Racist people	3.35
Ignorant people	2.13	Ignorant people	3.10
Nigger	1.84	Rude people	2.45
Ungrateful people	1.80	Old people	2.18

Table 2: Top ten targets of hate in Twitter and Whisper

Who Are Targets of Hate Speech?

TECHNISCHE UNIVERSITÄT DARMSTADT

Categories	Example of hate targets
Race	nigga, black people, white people
Behavior	insecure people, sensitive people
Physical	obese people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people

https://www.hatebase.org/

Why Hate Speech Online is More Visible than Offline?

Online Disinhibition Effect (Suler'04)

TECHNISCHE UNIVERSITÄT DARMSTADT

Benign disinhibition and *Toxic* disinhibition

- ➔ Dissociative anonymity ("You don't know me")
- ➔ Invisibility ("You can't see me")
- ➔ Asynchronicity ("See you later")
- ➔ Solipsistic Introjection ("It's all in my head")
- ➔ Dissociative Imagination ("It's just a game")
- ➔ Minimization of Status and Authority ("Your rules don't apply here")

Why Don’t They Solve The Problem?

TECHNISCHE UNIVERSITÄT DARMSTADT

Hate Speech vs Spam

Hate Speech Platforms

Hate Speech Platforms

Who is Responsible?

TECHNISCHE UNIVERSITÄT DARMSTADT

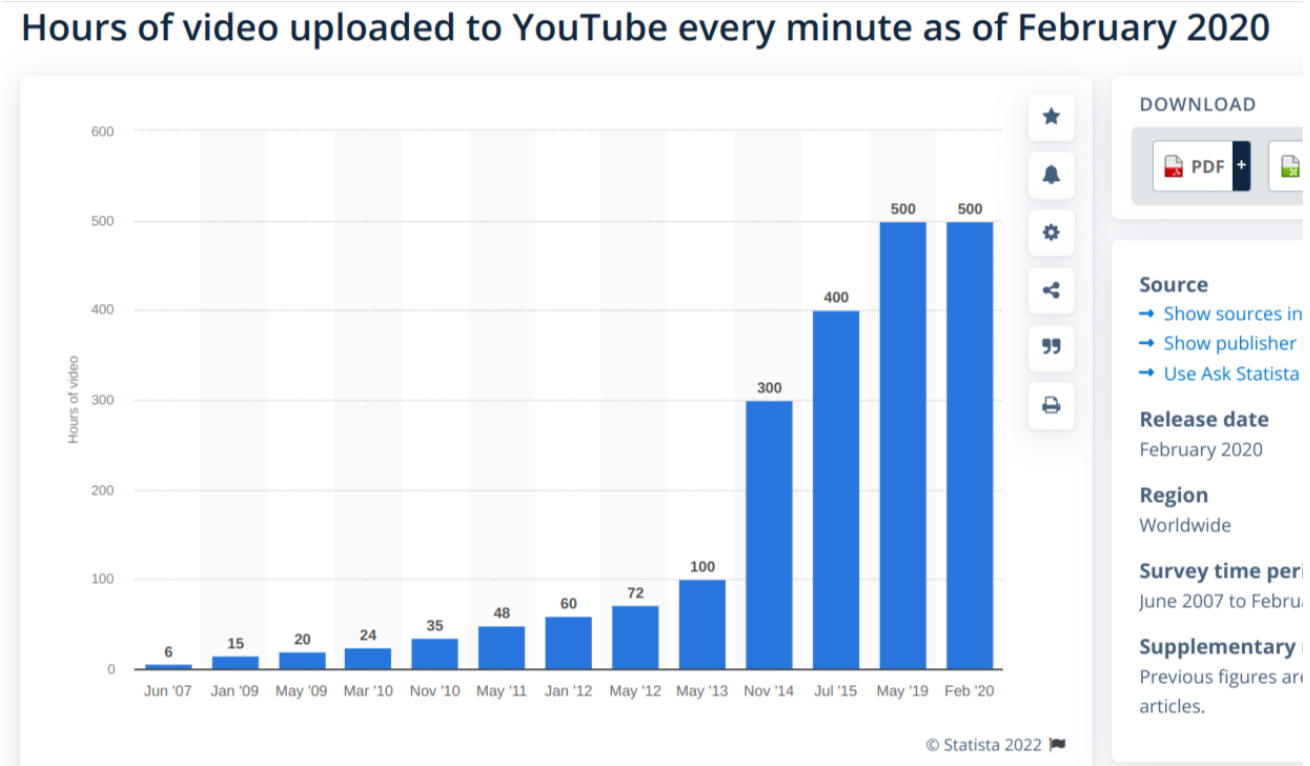
- Imagine that you are an engineer or a product manager at Google.
- You need to first come up with these policies and then enforce them.
- And you ask: why me?
- Online platforms created this problem but should they be responsible to solve it?
- Then, who should be solving the problem of hate speech?

Regulation

- https://www.facebook.com/legal/terms
- https://www.google.com/policies/terms/
- https://twitter.com/en/tos
- ...

Companies talk to activists, free speech lawyers, social groups…
to create TOS (Terms Of Service) for their platform

Impossible to monitor manually



Why Do Companies Regulate the Problems?

In 2017 alone, the company spent $9.4 billion on marketing

https://goo.gl/3AnREe

"These advertisers include five of the top 20 U.S. advertisers, who collectively make up 7.5 percent of total United States ad spend: AT&T, General Motors, Verizon, Walmart, and Johnson & Johnson."

https://goo.gl/XXgzG4

What Actions Should Be Taken?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ok, if we can identify it's hate speech, what action can we do?
- Close their account? Delete their messages?

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

What is Hate Speech?

Effects of Hate Speech

Hate Speech Identification

Hate Speech vs. Free Speech

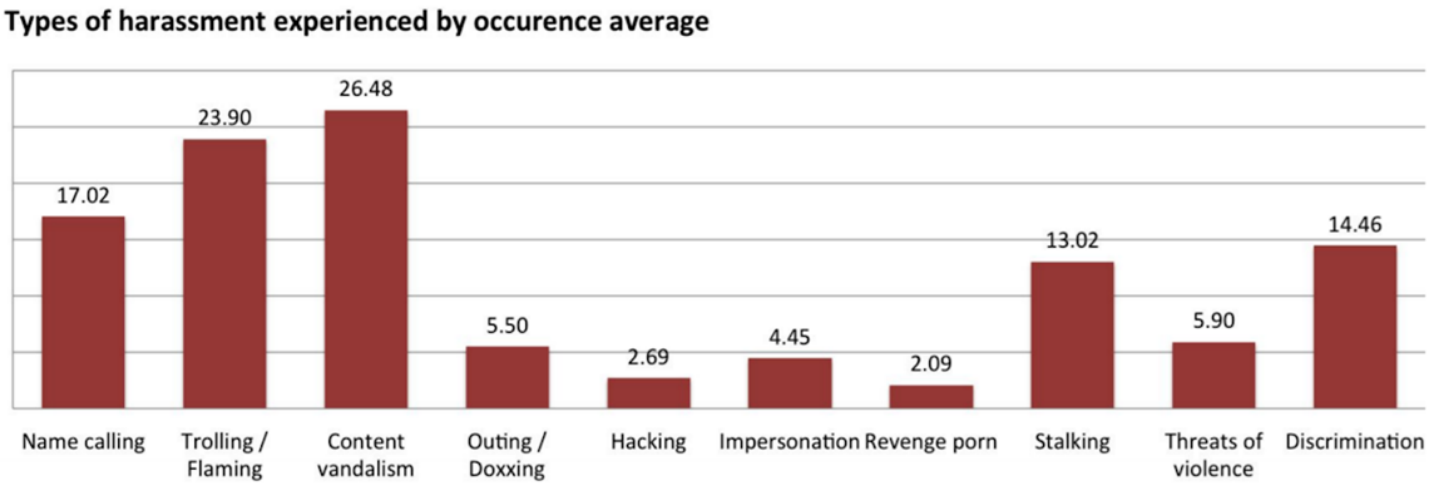
"I know it when I see it."

-- Supreme Court Justice Potter Stewart to describe his threshold test for obscenity in *Jacobellis v. Ohio* (1964)

It’s Hard

it can be done to anyone by anyone at any platform
it’s a hard problem, that many companies are trying to solve right now
there is no solution yet

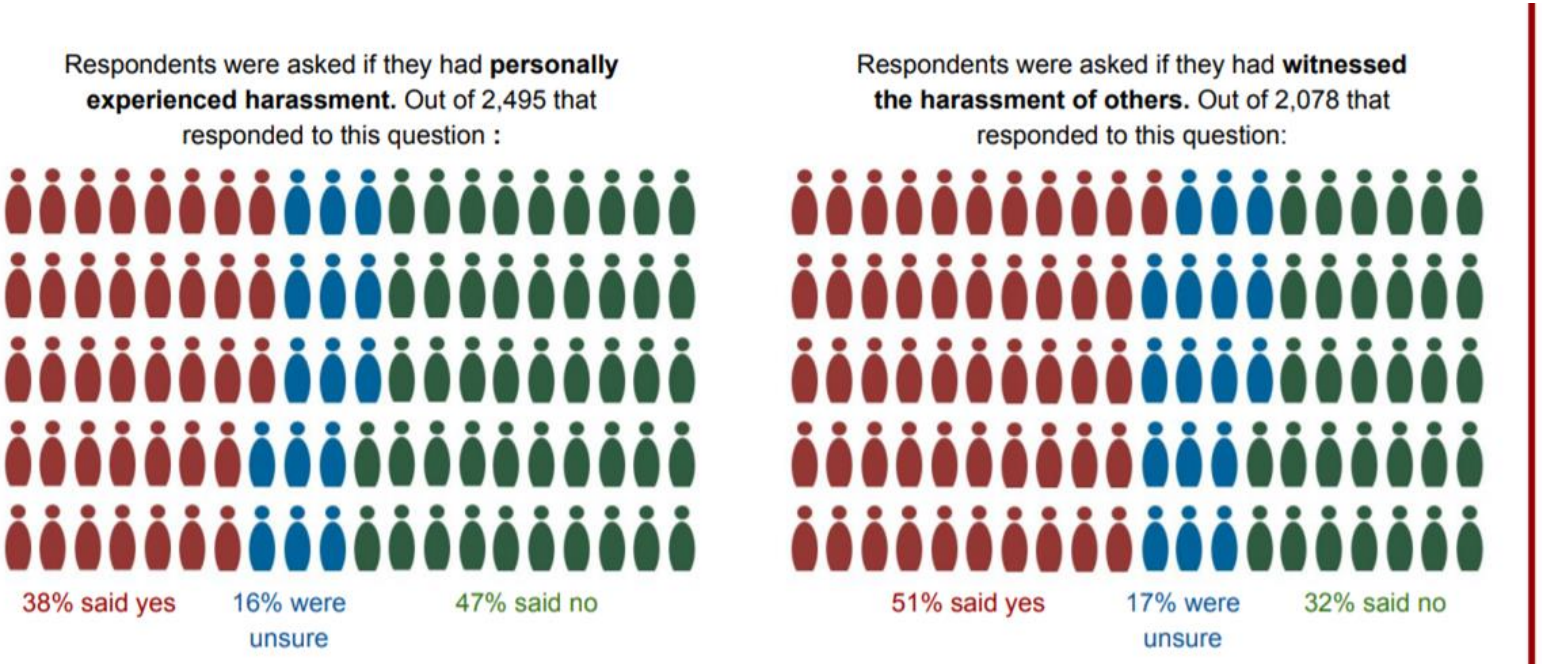
Types of Hate Speech



Support and Safety Team. Harassment Survey. Wikimedia Foundation, 2015

Of the **3,845** Wikimedia users who participated, 38% of the respondents could confidently recognise that they had been harassed, while 15% were unsure and 47% were confident that they had not been harassed. Similarly, 51% witnessed others being harassed, while 17% were unsure and 32% did not witness harassment.

Pew Research Center survey 2017



Element AI survey 2018

- manual analysis of 228,000 tweets sent to 778 women politicians and journalists in the UK and USA in 2017
- women politicians and journalists are assaulted every 30 seconds on Twitter
- 7.1% of tweets sent to the women in the study were problematic or abusive.
- black women were disproportionately targeted, being 84% more likely than white women to be mentioned in abusive or problematic tweets. One in ten tweets mentioning black women was abusive or problematic, compared to one in fifteen for white women;
- black and minority ethnic women were 34% more likely to be mentioned in abusive or problematic tweets than white women;

https://www.amnesty.org.uk/press-releases/women-abused-twitter-every-30-seconds-new-study

Computational approaches

Computational approaches

TECHNISCHE UNIVERSITÄT DARMSTADT

A simple classifier?

- AMT to label existing comments as abusive/non-abusive
- Lexicon+BOW features
- Regular expressions: “you are” , “I hate”
- Sentiment
- Brown clusters, fasttext embeddings, BERT embeddings
- Or fine-tune large pretrained model

Will it work?

Why Hate Speech Identification is Hard

TECHNISCHE UNIVERSITÄT DARMSTADT

Intentional obfuscation of abuse words, short forms etc

- Single character substitution: *nagger* (W&H'12)
- Homophone *joo* (W&H'12) *JOOZ* (NTTMC'16)
- Expanded spelling *j@e@w* (W&H'12)
- *Ni99er* (NTTMC'16)
- Tokenization *Woopiuglyniggeratgoldberg* (NTTMC'16)

Why Hate Speech Identification is Hard

TECHNISCHE UNIVERSITÄT DARMSTADT

Coded language that appears to mean one thing to the general population but has an additional meaning in in-group

- http://www.diversityinc.com/news/alt-right-trolls-devise-racist-codes-social-media/

Codewords

Table 1: Some common codewords

Code word	Actual word
Google	Black
Yahoo	Mexican
Skype	Jew
Bing	Chinese
Skittle	Muslim
Butterfly	Gay

Table 2: Top 10 most correlated terms

Term	Pearson correlation coefficient
#MAGA	0.149
#ALTRIGHT	0.140
gas	0.136
((( )))	0.136
white	0.136
war	0.118
hate	0.100
#MAWA	0.098
destroy	0.083
goy	0.083

(Magu et al., Detecting the Hate Code on Social Media, ICWSM 2017)

Why Hate Speech Identification is Hard

TECHNISCHE UNIVERSITÄT DARMSTADT

In- and out-group lexicons : rap lyrics contains curse words.

Words like *black, jew, women* are used in various contexts more frequently than in hateful speech.

- Keyword spotting will yield false positives

Related issue: bias in hate speech detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Train/test two different classifiers
  - TWT-HATEBASE (Davidson et al, 2017)
  - TWT-BOOTSTRAP (Founta et al., 2018)
- Rates of **false flagging of toxicity**
  - Broken down by dialect group on heldout set

Predictions by both classifiers
biased against AAE tweets

Within dataset proportions

			% false identification	
DWMW17 Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

			% false identification	
FDCL18 Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Maarten Sap et al.(2019) The Risk of Racial Bias in Hate Speech Detection. *ACL*

Why is hate speech detection a hard problem

- In- and out-group lexicons: *nigg*s a* ↔

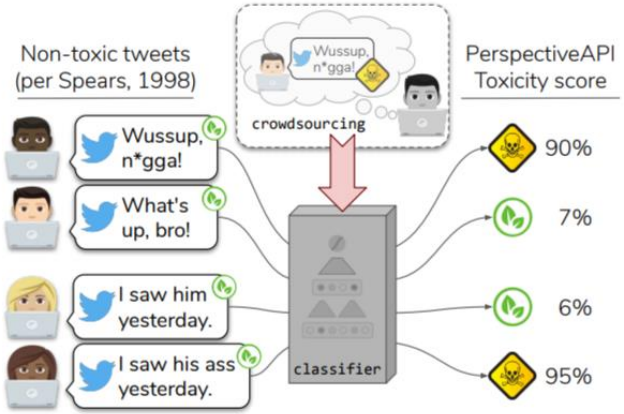


Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

Maarten Sap et al.(2019) The Risk of Racial Bias in Hate Speech Detection. *ACL*

Why is hate speech detection a hard problem

- Hateful comments can be fluent and grammatical and do not evoke any blacklisted word

I am surprised they reported on this, who cares about another dead woman? (NTTMC'16)

- Doesn't have to be in the scope of one sentence

Why Hate Speech Identification is Hard

TECHNISCHE UNIVERSITÄT DARMSTADT

Sarcasm: hateful messages said in sarcastic way in non-hateful context (NTTMC'16)

Data collection is Also Hard

TECHNISCHE UNIVERSITÄT DARMSTADT

- News outlets and online communities remove this content
- Hard to obtain due to privacy issues
- Possibility to flag content? But part of trolling is to go to non-abusive content and flag it as abusive.
- This is why it is difficult even for companies to identify automatically abusive content even using feedback from users

Datasets

Dataset	# Tweets	Labels	Annotators/Tweet
Chatzakou et al. (2017)	9,484	aggressive, bullying, spam, normal	5
Waseem and Hovy (2016)	16, 914	racist, sexist, normal	1
Davidson et al. (2017)	24, 802	hateful, offensive (but not hateful), neither	3 or more
Golbeck et al. (2017)	35,000	the worst, threats, hate speech, direct harassment, potentially offensive, non-harassment	2 to 3
Founta et al. (2018)	80, 000	offensive, abusive, hateful speech, aggressive, cyberbullying, spam, normal	5 to 20
ElSherief et al. (2018)	28,608	directed, generalized + target = archaic, class, disability, ethnicity, gender, nationality, religion, sexual orientation	3
Ours	13,000	Labels for five different aspects	5

Table 1: Comparative table of some of the available hate speech and abusive language corpora in terms of labels and sizes.

Ousidhoum N. et al. (2019) Multilingual and Multi-Aspect Hate Speech Analysis. EMNLP

Datasets

- B. Vidgen & L. Derczynski (2020) Directions in Abusive Language Training Data: Garbage In, Garbage Out *PLOS One*
  - https://hatespeechdata.com/
- Poletto et al. (2021) Resources and benchmark corpora for hate speech detection: a systematic review In *Language Resources and Evaluation*

Categorization

William Warner & Julia Hirschberg. (2012) **Detecting Hate Speech on the World Wide Web.** *Workshop on Language in Social Media*

- Stereotype-based
  - Racist
  - Anti-semitic

Categorization

Nobata et al. (2016) **Abusive Language Detection in Online User Content.** *WWW*

Table 4: Annotation Instructions

Given the passage below, first rate the passage on whether the passage is clean or abusive. And if abusive, • whether it contains hate speech • whether it contains derogatory languge • whether it contains profanity	
Hate Speech:	language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity. Some examples: • The JEWZ are Market Manipulators Hate them freaking Jews !! To bad Hitler did not finish them off !!! • I hate the mall because there are too many unemployed apefricans and mudsharks. • 300 missiles are cool! Love to see um launched into TelAviv! Kill all the gays there! • EGYPTIANS ARE DESCENDED FROM APES AND PIGS. NUKE EGYPT NOW Ah me tu se ku sh
Derogatory:	language which attacks an individual or a group, but which is not hate speech. Some examples: • All you perverts (other than me) who posted today, needs to leave the O Board. Dfasdfdasfadfs • yikes...another republiCUNT weighs in....
Profanity:	language which contains sexual remarks or profanity. Some examples: • T.Boone Pickens needs to take a minimum wage job in FL for a week. I guarantee he shuts the f up after that. • Somebody told me that Little Debbie likes to take it up the A.$.$. • So if the pre market is any indication Kind of like the bloody red tampons that you to suck on all day??

Categorization

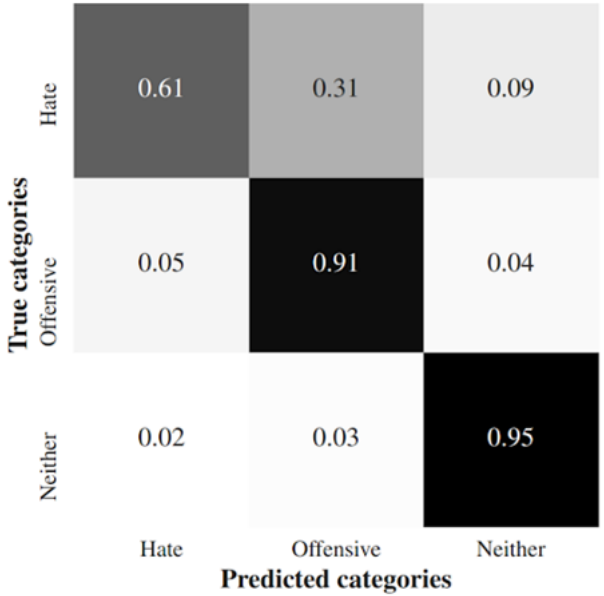
Waseem et al. (2017) **Understanding Abuse: A Typology of Abusive Language Detection Subtasks.** *Workshop on Abusive Language Online*

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

Table 1: **Typology of abusive language.**

Categorization

Davidson et al. (2012) **Automated Hate Speech Detection and the Problem of Offensive Language.** *ICWSM*



Categorization

Cheng et al. (2015) **Antisocial Behavior in Online Discussion Communities.** *AAA*

- Future-banned users
- Never-banned users

Identification Approaches

William Warner & Julia Hirschberg. (2012) **Detecting Hate Speech on the World Wide Web.** *Workshop on Language in Social Media*

- Hateful language directed towards a minority or a disadvantaged group
- Stereotype-based
  - Racist
  - Anti-semitic
- Standard SVM classifiers with standard and pattern features

Classification features

- Character n-grams to account for spelling variations
- From prior research on text normalization:
  - length of comment in tokens
  - average length of word
  - number of punctuations
  - number of periods, question marks, quotes, and repeated punctuation
  - number of one letter tokens
  - number of capitalized letters
  - number of URLS
  - number of tokens with non-alpha characters in the middle
  - number of politeness words
  - number of unknown words
  - number of insult and hate blacklist word

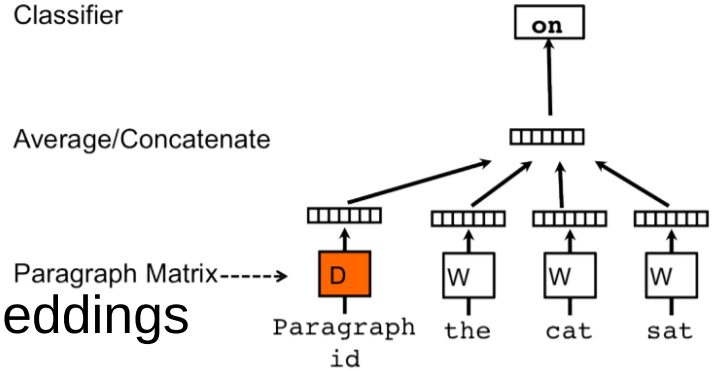
Classification features

- Character n-grams to account for spelling variations
- From prior research on text normalization
- Syntactic features
  - parent of node
  - grandparent of node
  - POS of parent
  - POS of grandparent
  - tuple consisting of the word, parent and grandparent
  - ...
- Distributional semantic features
  - word2vec
  - paragraph2vec

Identification Approaches

Djuric et al. (2015) **Hate Speech Detection with Comment Embeddings.** *WWW*

- “abusive speech targeting specific group characteristics, such as ethnicity, religion, or gender”
- Yahoo Finance comments
  - 56K hate speech, 895K clean comments
- paragraph2vec Le&Mikolov’14
- train a binary classifier with paragraph embeddings



Next Lecture

TECHNISCHE UNIVERSITÄT DARMSTADT

InCivility and HateSpeech II