What [young] people think we should be allowed to say

CIVICA Youth Gender Polarization and Digital Fragmentation Conference

Simon Munzert · Hertie School, Berlin

Oct 18, 2025

Hertie School · King's College London · Max Planck Institute for Human Development · Universität Mannheim

Co-funded by: facebook research

Where I'm coming from

Content moderation is big and... bad?

Hertie School

- Platforms moderate content opaquely, often without oversight (Gillespie 2018)
- Content moderation with commercial services is **massive** (e.g., 500m requests per day to JigSaw's Perspective API in 2021; 200m weekly users of OpenAI's Moderation API)
- **Content Moderation APIs** by Google, Microsoft, Amazon, OpenAI, and others are supposedly trained on human labeled data but produce lots of questionable decisions (Hartmann et al. 2025; FPR/FNR up to 75%, ACC as low as 60% in balanced toxic/non-toxic samples)

Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations

David Hartmann
TU Berlin
Berlin, Germany
Weizenbaum Institute for the Networked Society
Berlin, Germany
d.hartmann@tu-berlin.de

Amin Oueslati
Hertie School Berlin
Berlin, Germany
amin.m.oueslati@gmail.com

Dimitri Staufer
TU Berlin
Berlin, Germany
staufer@tu-berlin.de

Lena Pohlmann
TU Berlin
Berlin, Germany
Weizenbaum Institute for the Networked Society
Berlin, Germany
l.pohlmann@tu-berlin.de

Simon Munzert
Hertie School Berlin
Berlin, Germany
munzert@hertie-school.org

Hendrik Heuer
Center for Advanced Internet Studies (CAIS) gGmbH
Bochum, Germany
University of Wuppertal
Wuppertal, Germany
hendrik.heuer@cais-research.de

Abstract

Commercial content moderation APIs are marketed as scalable solutions to combat online hate speech. However, the reliance on these APIs risks both silencing legitimate speech, called over-moderation, and failing to protect online platforms from harmful speech, known as under-moderation. To assess such risks, this paper introduces a framework for auditing black-box NLP systems. Using the framework, we systematically evaluate five widely used commercial content moderation APIs. Analyzing five million queries based on four datasets, we find that APIs frequently rely on group identity terms, such as "black", to predict hate speech. While OpenAI's and Amazon's services perform slightly better, all providers under-moderate implicit hate speech, which uses codified messages, especially against LGBTQIA+ individuals. Simultaneously, they over-moderate counter-speech, reclaimed slurs and content related to Black, LGBTQIA+, Jewish, and Muslim people. We recommend that API providers offer better guidance on API implementation and threshold setting and more transparency on their APIs' limitations.
Warning: *This paper contains offensive and hateful terms and concepts. We have chosen to reproduce these terms for reasons of transparency.*

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; Empirical studies in collaborative and social computing; • **General and reference** → *Measurement*; • **Social and professional topics** → **Hate speech**.

Keywords

Content Moderation APIs, Audit, AI Transparency and Accountability, Human-AI Interaction in Content Moderation, Algorithmic Bias in Hate Speech Detection

1 Introduction

Content moderation has become a widely used tool in combating online hate speech. While human moderators play an essential role in hate speech removal, human-based moderation is expensive, difficult to scale, and often exposes outsourced workers to distressing content that affects their mental health [34]. To address these challenges, companies such as Google, Microsoft, Amazon, Jigsaw, and OpenAI offer commercial, automated content moderation services. These API-based solutions are marketed as scalable, efficient solutions to tackle the growing challenges faced by social media platforms and other websites that deal with user-generated content [85]. For most major platforms, content moderation decisions are – to a large extent – partially or fully automated [21]. However, these decisions are potentially fallible. When harmful content is not moderated (under-moderation; reflected in a high False Negative Rate (FNR) of the content classifier), users are left unprotected from hate speech [23]. Conversely, when legitimate content is moderated (over-moderation; reflected in a high False Positive Rate (FPR)), this limits users' opportunities to express themselves and participate in public discourse. Both issues become aggravated when they systematically affect selected social groups, particularly those defined by protected characteristics such as gender, race, or religion.

While over- and under-moderation across different forms of hate speech has been extensively researched for off-the-shelf NLP models, research on over- and under-moderation of *commercial content moderation APIs* is limited. This is an important research gap since on off-the-shelf – meaning locally accessible – NLP models such as Sap et al. [90], Wiegand et al. [102] and Röttger et al. [87], have

Why hate speech moderation at scale is so challenging

Hertie School

It's a hard problem

- Volume and velocity of content
  - Billions of posts per day
  - Error amplification at scale
- Adversarial users
  - Coded language, meme shifts
  - Evasion of detection systems
- Context is everything
  - Meaning depends on intent
  - Hard for AI to judge nuance
- Legal and reputational constraints
  - Balance free speech vs harm
  - Laws differ across countries

And we might not even agree on what the problem is...

- The fundamental subjectivity of hate speech
  - Likely varies by culture, context of speech, and individual prefs
  - Why it's impossible to agree on what's allowed (Dan Luu) and the "No vehicles in the park" rule

Why hate speech moderation at scale is so challenging

Hertie School

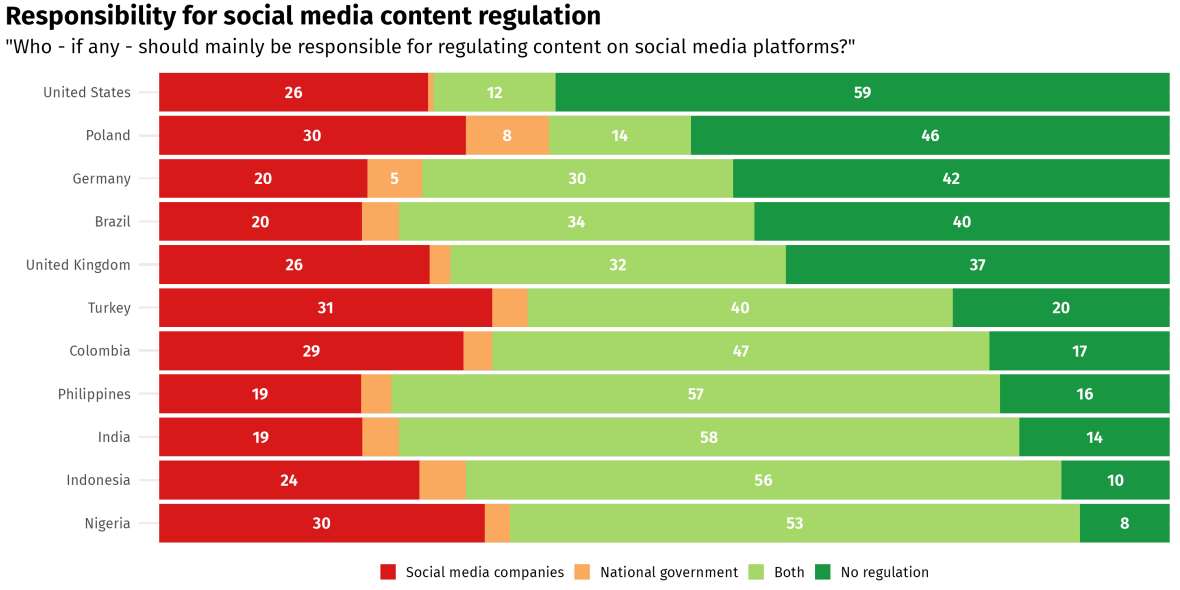
It's a hard problem

- Volume and velocity of content
  - Billions of posts per day
  - Error amplification at scale
- Adversarial users
  - Coded language, meme shifts
  - Evasion of detection systems
- Context is everything
  - Meaning depends on intent
  - Hard for AI to judge nuance
- Legal and reputational constraints
  - Balance free speech vs harm
  - Laws differ across countries

And we might not even agree on what the problem is...

- The fundamental subjectivity of hate speech
  - Likely varies by culture, context of speech, and individual prefs
  - Why it's impossible to agree on what's allowed (Dan Luu) and the "No vehicles in the park" rule

... and how it should be tackled



Evidence on variation in hate speech moderation preferences

Hertie School

- Support for moderation of severe hate speech (threats, violence) (Pradel et al. 2024, Munzert et al. 2025, Rasmussen 2022)
- Persistent gender gap in content moderation preferences (Pradel and Theocharis 2024; Munzert et al. 2025)
- Ideological differences in tolerance for hate speech (Munzert et al. 2025)
- Partisan norms and identity signaling drive sanctioning preferences (Ahn et al. 2024, Dias et al. 2024)

American Political Science Review (2024) 118, 4, 1895–1912
doi:10.1017/S000305542300134X

Toxic Speech and Limited Demand for Content Moderation on Social Media

FRANZISKA PRADEL *Technical University of Munich, Germany*
JAN ZILINSKY *Technical University of Munich, Germany*
SPYROS KOSMIDIS *University of Oxford, United Kingdom*
YANNIS THEOCHARIS *Technical University of Munich, Germany*

When is speech on social media toxic enough to warrant content moderation? Platforms impose limits on what can be posted online, but also rely on users' reports of potentially harmful content. Yet we know little about what users consider inadmissible to public discourse and what measures they wish to see implemented. Building on past work, we conceptualize three variants of toxic speech: incivility, intolerance, and violent threats. We present results from two studies with pre-registered randomized experiments (Study 1, N = 5,130; Study 2, N = 3,734) to examine how these variants causally affect users' content moderation preferences. We find that while both the severity of toxicity and the target of the attack matter, the demand for content moderation of toxic speech is limited. We discuss implications for the study of toxicity and content moderation as an emerging area of research in political science with critical implications for platforms, policymakers, and democracy more broadly.

PNAS NEXUS

PNAS Nexus, 2025, **4**, pgaf032
https://doi.org/10.1093/pnasnexus/pgaf032
Advance access publication 12 February 2025
Research Report

Citizen preferences for online hate speech regulation

Simon Munzert[a,*], Richard Traunmüller[b], Pablo Barberá[c], Andrew Guess[d] and JungHwan Yang[e]

[a]Data Science Lab, Hertie School, 10017 Berlin, BE, Germany
[b]School of Social Sciences, University of Mannheim, 68159 Mannheim, BW, Germany
[c]Department of Political Science and International Relations, University of Southern California, Los Angeles, CA 90089, USA
[d]Department of Politics and School of Public and International Affairs, Princeton University, Princeton, NJ 08544, USA
[e]Department of Communication, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
*To whom correspondence should be addressed: Email: munzert@hertie-school.org
Edited By Katherine Ognyanova

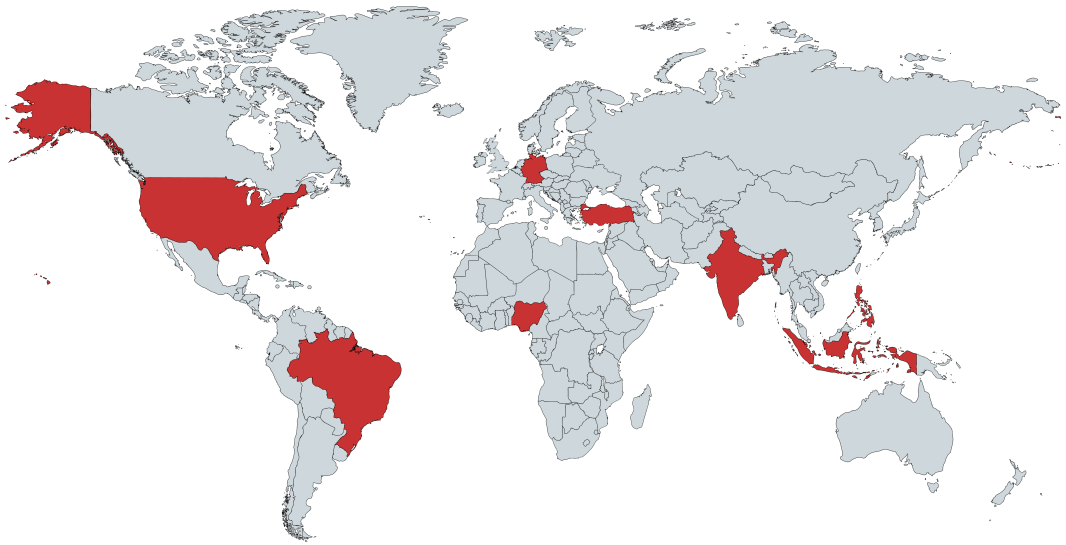
Abstract

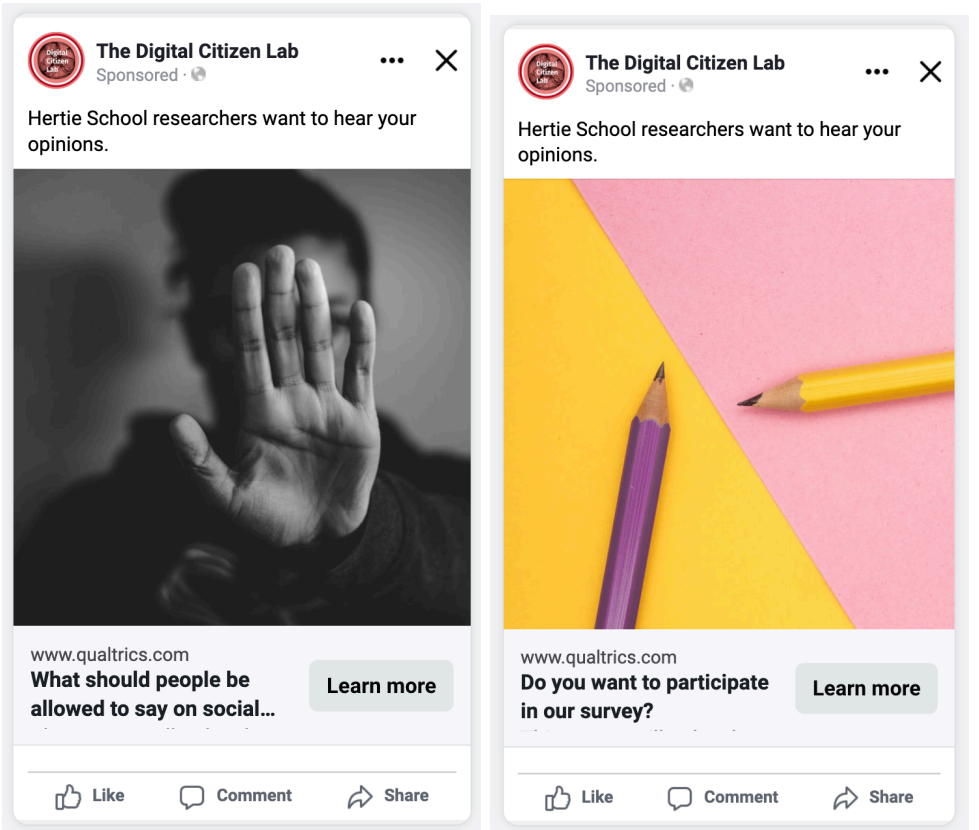
The shift of public discourse to online platforms has intensified the debate over content moderation by platforms and the regulation of online speech. Designing rules that are met with wide acceptance requires learning about public preferences. We present a visual vignette study using a sample ($n = 2,622$) of German and US citizens that were exposed to 20,976 synthetic social media vignettes mimicking actual cases of hateful speech. We find people's evaluations to be primarily shaped by message type and severity, and less by contextual factors. While focused measures like deleting hateful content are popular, more extreme sanctions like job loss find little support even in cases of extreme hate. Further evidence suggests in-group favoritism among political partisans. Experimental evidence shows that exposure to hateful speech reduces tolerance of unpopular opinions.

Data collection

Hertie School

- Online survey fielded in 11 countries
- Recruitment via FB/Instagram ads
- ~18.5k completed interviews
- Pre-registered at OSF

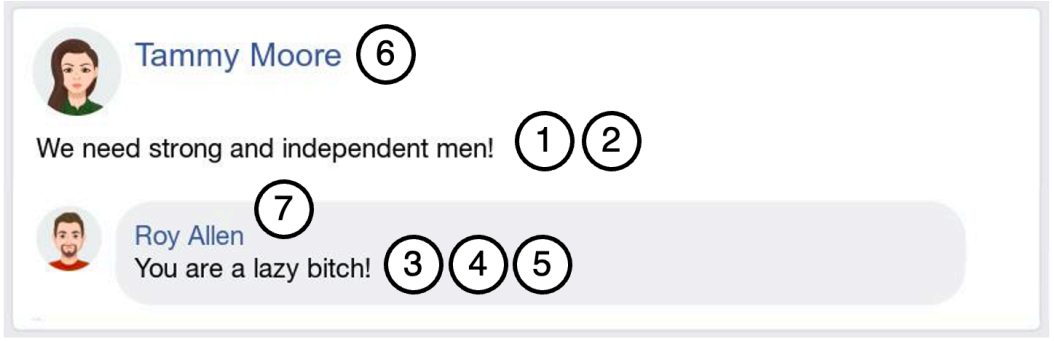




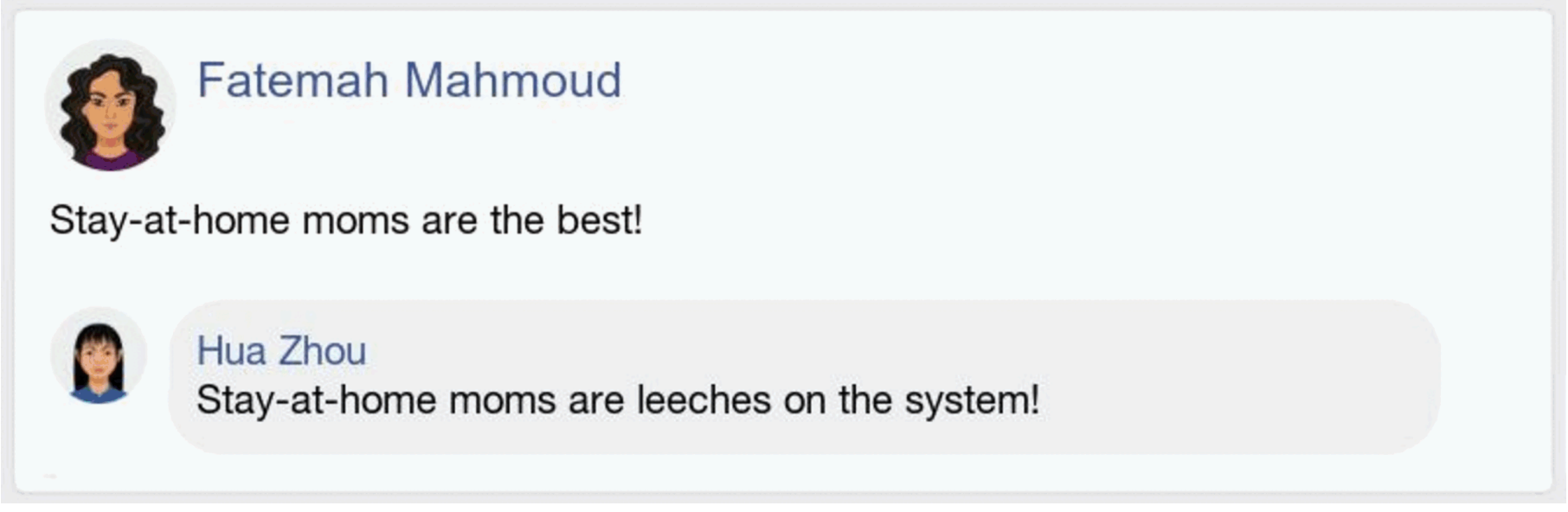
Vignette attributes and levels

Hertie School

Attribute	Levels
① Topic	14 topics (feminism, religiosity, partisanship, ...)
② Target stance	left, right
③ Message category	opinion, meme, mocking, insult, threat
④ Message severity	moderate, extreme
⑤ Message scope	personal, group
⑥ Target's identity	white/non-white, female/male
⑦ Sender's identity	white/non-white, female/male

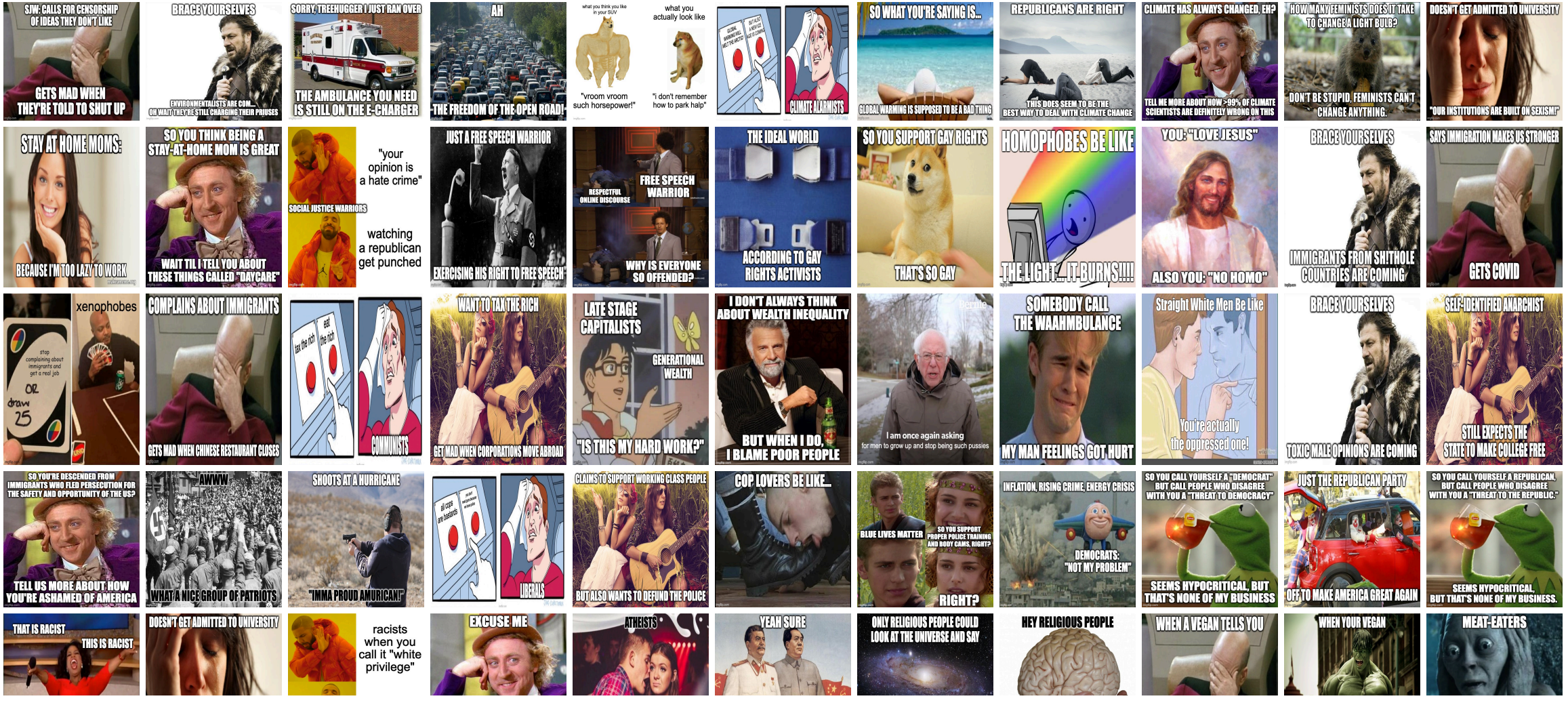


Some of the vignettes



All the (U.S.) memes

Hertie School



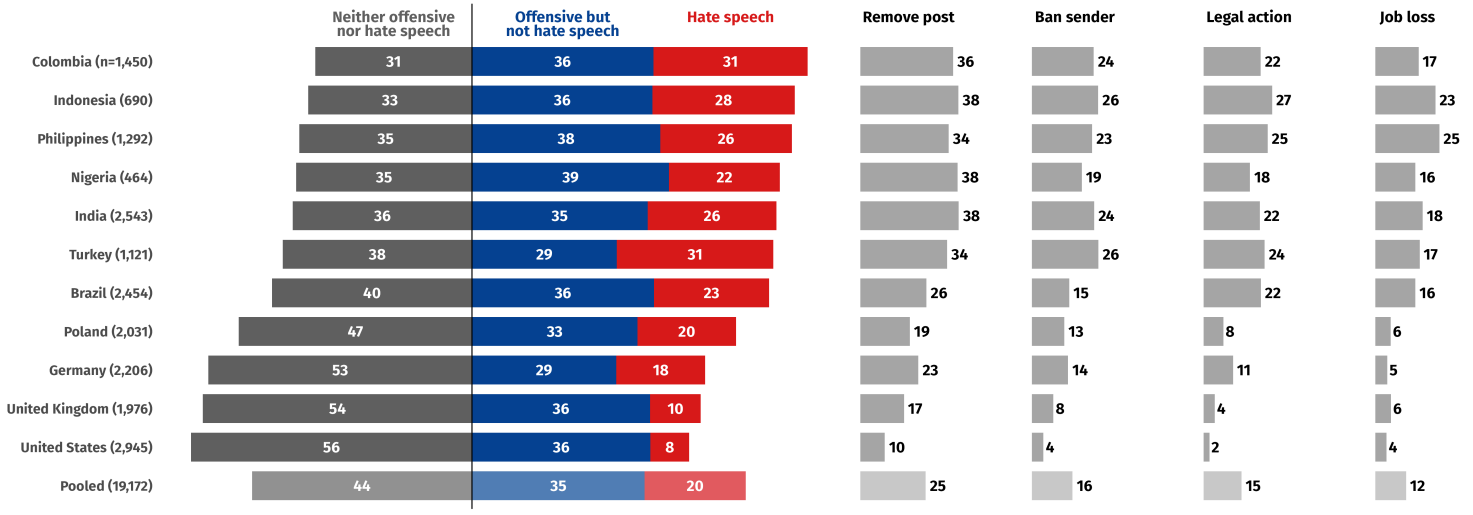
Messages most often flagged as hate speech

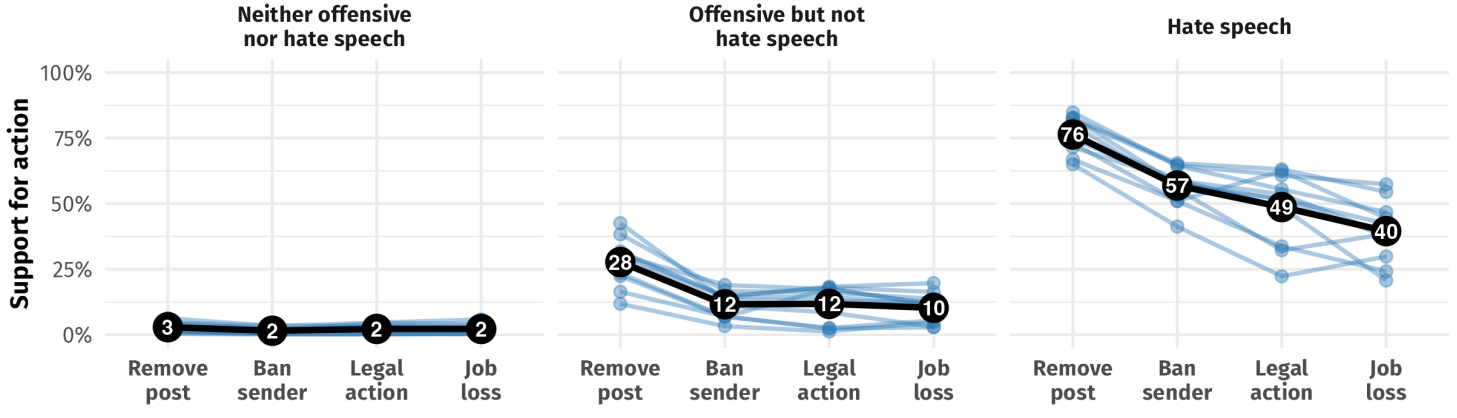
Hertie School

Message features			Perceptions			Preferred actions				
Message (US version)	Sender	Position	Neither offensive nor hate speech	Offensive but not hate speech	Hate speech	Remove post	Ban sender	Legal action	Job loss	Number of selected actions
Die you homo!	Threat	Left	8%	18%	74%	69%	49%	40%	32%	
Death to all homosexuals!	Threat	Left	8%	16%	74%	64%	52%	48%	35%	
Death to all cops!	Threat	Right	8%	19%	72%	66%	48%	47%	32%	
You should be killed like vermin!	Threat	Left	6%	22%	70%	66%	54%	46%	36%	
Protect your family, kick a homo's ass today	Threat	Left	9%	21%	70%	61%	47%	45%	38%	
Immigrants should be punched in the face.	Threat	Left	7%	22%	70%	57%	38%	38%	31%	
You deserve to be raped!	Threat	Left	3%	27%	70%	73%	65%	60%	48%	
Stay-at-home moms should get their houses burned down and die!	Threat	Right	6%	24%	69%	65%	52%	46%	36%	
Feminists deserve to be raped!	Threat	Left	4%	26%	68%	72%	57%	55%	44%	
I hope someone burns your house down and you die!	Threat	Right	7%	25%	67%	70%	57%	50%	40%	
Leftie Democrats should be shot on sight!	Threat	Left	11%	21%	67%	62%	47%	46%	31%	
I hope someone shoots you in the street!	Threat	Right	6%	29%	65%	67%	51%	49%	35%	
Death to all homophobes!	Threat	Right	10%	25%	65%	56%	41%	39%	25%	
Vegans should be rounded up and slaughtered like cattle!	Threat	Left	11%	26%	63%	63%	44%	42%	29%	
Execute all atheists!	Threat	Left	14%	24%	62%	56%	41%	35%	24%	

Perceived nature of speech and support for moderation

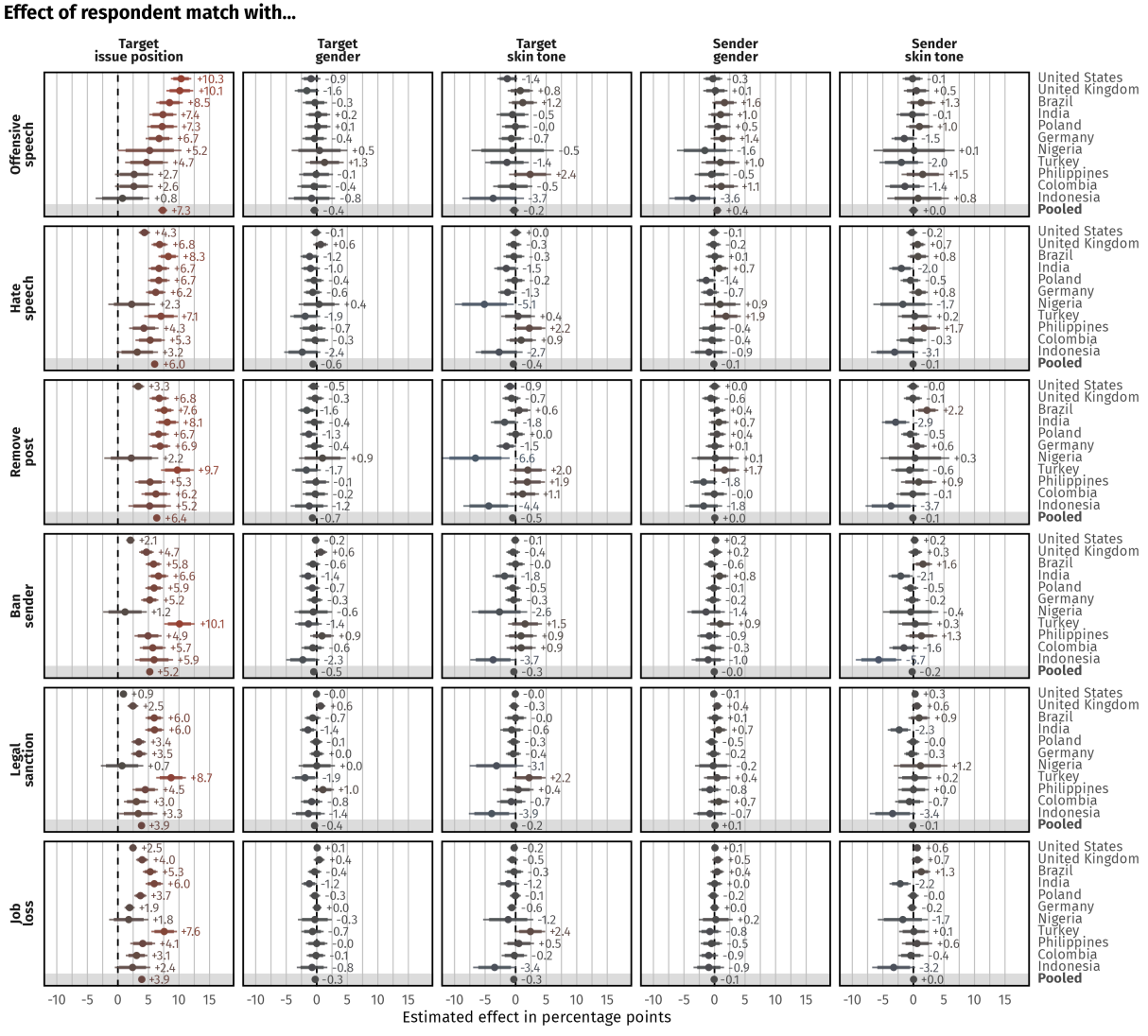
Hertie School





Myside bias in content moderation I

Hertie School



Myside bias in content moderation II

Hertie School

Respondent characteristics matter

Hertie School

	Perceptions		Actions			
Pooled	Offensive	Hate speech	Remove post	Ban sender	Legal sanction	Job loss
Gender (vs. Male)						
Female	+5.3pp	+3.2pp	+7.2pp	+4.6pp	+2.7pp	+2.0pp
Other	+1.0pp	+3.4pp	+4.4pp	+2.9pp	+2.1pp	+1.9pp
Age (vs. 18-29)						
70+	+2.2pp	+1.0pp	+1.3pp	+3.0pp	+4.2pp	+2.6pp
50-69	-2.2pp	-0.7pp	0.0pp	+0.9pp	+2.7pp	+0.6pp
30-49	-3.5pp	-1.5pp	-0.5pp	-0.1pp	+1.2pp	-0.3pp
Education (vs. Low)						
High	+0.2pp	-1.7pp	-3.6pp	-3.2pp	-2.5pp	-2.2pp
Intermediate	+0.1pp	-1.5pp	-3.1pp	-2.3pp	-1.6pp	-1.5pp
Ethnicity (vs. Non-White)						
White	-0.4pp	-0.9pp	-1.0pp	-0.3pp	-1.1pp	-0.8pp
Minority (vs. Non-Minority)						
Minority	-0.7pp	-0.2pp	-0.3pp	+0.2pp	+1.0pp	+1.4pp
Political interest (vs. Low)						
High	+0.4pp	-0.9pp	-1.4pp	-1.3pp	+0.6pp	-0.3pp
Political ideology (vs. Left)						
Center	-1.9pp	-1.8pp	-2.8pp	-2.1pp	-1.4pp	-1.8pp
Right	-4.3pp	-2.3pp	-4.2pp	-2.2pp	-1.6pp	-1.0pp
Being able to speak freely, general (vs. More free)						
Just as free	-0.7pp	-1.5pp	-1.9pp	-1.7pp	-2.1pp	-1.9pp
Less free	-2.2pp	-1.8pp	-3.8pp	-2.8pp	-2.9pp	-3.6pp
Being able to speak freely, personal (vs. More free)						
Less free	+1.0pp	+0.1pp	+0.6pp	-0.5pp	-0.1pp	-1.0pp
Just as free	+0.6pp	+0.5pp	+1.7pp	+1.0pp	+1.9pp	-0.4pp

	Perceptions		Actions			
Pooled	Offensive	Hate speech	Remove post	Ban sender	Legal sanction	Job loss
Silencing speech score (vs. Low)						
Medium	+6.6pp	+3.4pp	+6.7pp	+3.5pp	+2.0pp	+1.6pp
High	+6.5pp	+5.2pp	+13.7pp	+8.8pp	+5.0pp	+5.7pp
Empathy score (vs. Low)						
Medium	+1.8pp	-0.1pp	-0.8pp	-1.3pp	-1.4pp	-1.5pp
High	+1.0pp	+1.0pp	+0.5pp	+0.1pp	-0.8pp	-0.6pp
Toxic experience score (vs. Low)						
High	+3.6pp	+2.3pp	+2.6pp	+2.1pp	+1.3pp	+1.6pp
Medium	+2.9pp	+0.7pp	+0.6pp	+0.4pp	+0.2pp	+0.2pp
Hostile engagement score (vs. Low)						
Medium	-1.1pp	-1.6pp	-3.4pp	-2.6pp	-3.0pp	-2.1pp
High	-2.6pp	-2.1pp	-4.3pp	-2.6pp	-3.1pp	-1.8pp
Model summary						
$N_{Observations}$	93,237	116,064	115,067	115,082	115,018	114,808
$N_{Respondents}$	14,485	14,508	14,492	14,492	14,487	14,479
$N_{Vignette\ decks}$	4,439	4,444	4,441	4,441	4,440	4,439
$N_{Countries}$	11	11	11	11	11	11
σ	0.41	0.33	0.34	0.29	0.28	0.25

Respondent characteristics matter

Hertie School

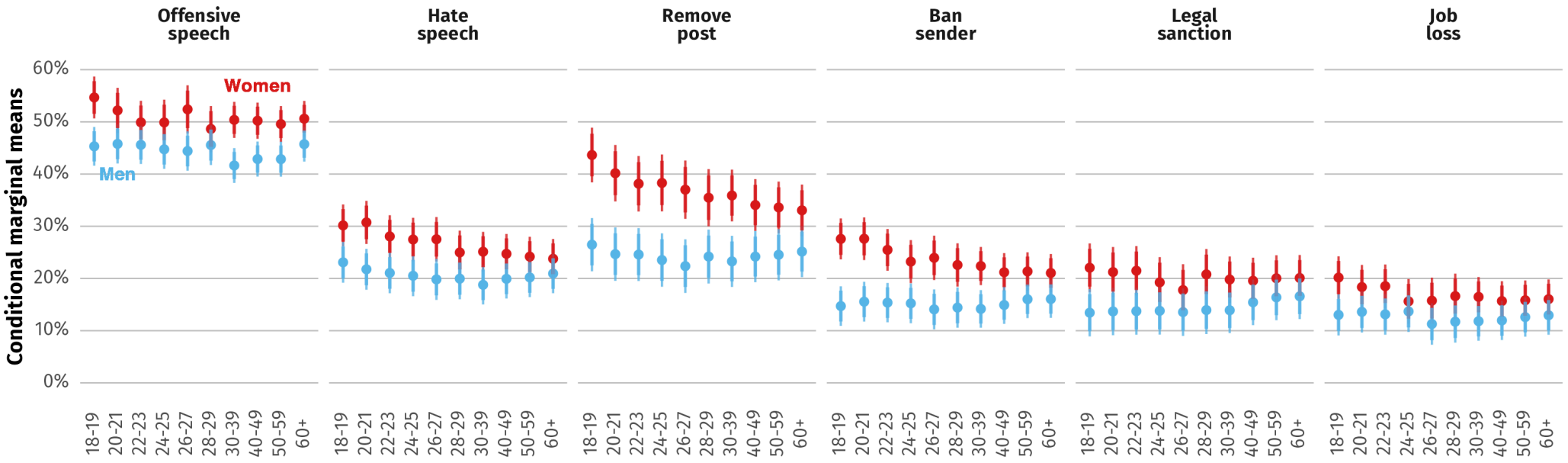
What's going on here?

	Perceptions		Actions			
Pooled	Offensive	Hate speech	Remove post	Ban sender	Legal sanction	Job loss
Gender (vs. Male)						
Female	+5.3pp	+3.2pp	+7.2pp	+4.6pp	+2.7pp	+2.0pp
Other	+1.0pp	+3.4pp	+4.4pp	+2.9pp	+2.1pp	+1.9pp
Age (vs. 18-29)						
70+	+2.2pp	+1.0pp	+1.3pp	+3.0pp	+4.2pp	+2.6pp
50-69	-2.2pp	-0.7pp	0.0pp	+0.9pp	+2.7pp	+0.6pp
30-49	-3.5pp	-1.5pp	-0.5pp	-0.1pp	+1.2pp	-0.3pp
Education (vs. Low)						
High	+0.2pp	-1.7pp	-3.6pp	-3.2pp	-2.5pp	-2.2pp
Intermediate	+0.1pp	-1.5pp	-3.1pp	-2.3pp	-1.6pp	-1.5pp
Ethnicity (vs. Non-White)						
White	-0.4pp	-0.9pp	-1.0pp	-0.3pp	-1.1pp	-0.8pp
Minority (vs. Non-Minority)						
Minority	-0.7pp	-0.2pp	-0.3pp	+0.2pp	+1.0pp	+1.4pp
Political interest (vs. Low)						
High	+0.4pp	-0.9pp	-1.4pp	-1.3pp	+0.6pp	-0.3pp
Political ideology (vs. Left)						
Center	-1.9pp	-1.8pp	-2.8pp	-2.1pp	-1.4pp	-1.8pp
Right	-4.3pp	-2.3pp	-4.2pp	-2.2pp	-1.6pp	-1.0pp
Being able to speak freely, general (vs. More free)						
Just as free	-0.7pp	-1.5pp	-1.9pp	-1.7pp	-2.1pp	-1.9pp
Less free	-2.2pp	-1.8pp	-3.8pp	-2.8pp	-2.9pp	-3.6pp
Being able to speak freely, personal (vs. More free)						
Less free	+1.0pp	+0.1pp	+0.6pp	-0.5pp	-0.1pp	-1.0pp
Just as free	+0.6pp	+0.5pp	+1.7pp	+1.0pp	+1.9pp	-0.4pp

	Perceptions		Actions			
Pooled	Offensive	Hate speech	Remove post	Ban sender	Legal sanction	Job loss
Silencing speech score (vs. Low)						
Medium	+6.6pp	+3.4pp	+6.7pp	+3.5pp	+2.0pp	+1.6pp
High	+6.5pp	+5.2pp	+13.7pp	+8.8pp	+5.0pp	+5.7pp
Empathy score (vs. Low)						
Medium	+1.8pp	-0.1pp	-0.8pp	-1.3pp	-1.4pp	-1.5pp
High	+1.0pp	+1.0pp	+0.5pp	+0.1pp	-0.8pp	-0.6pp
Toxic experience score (vs. Low)						
High	+3.6pp	+2.3pp	+2.6pp	+2.1pp	+1.3pp	+1.6pp
Medium	+2.9pp	+0.7pp	+0.6pp	+0.4pp	+0.2pp	+0.2pp
Hostile engagement score (vs. Low)						
Medium	-1.1pp	-1.6pp	-3.4pp	-2.6pp	-3.0pp	-2.1pp
High	-2.6pp	-2.1pp	-4.3pp	-2.6pp	-3.1pp	-1.8pp
Model summary						
$N_{Observations}$	93,237	116,064	115,067	115,082	115,018	114,808
$N_{Respondents}$	14,485	14,508	14,492	14,492	14,487	14,479
$N_{Vignette\ decks}$	4,439	4,444	4,441	4,441	4,440	4,439
$N_{Countries}$	11	11	11	11	11	11
σ	0.41	0.33	0.34	0.29	0.28	0.25

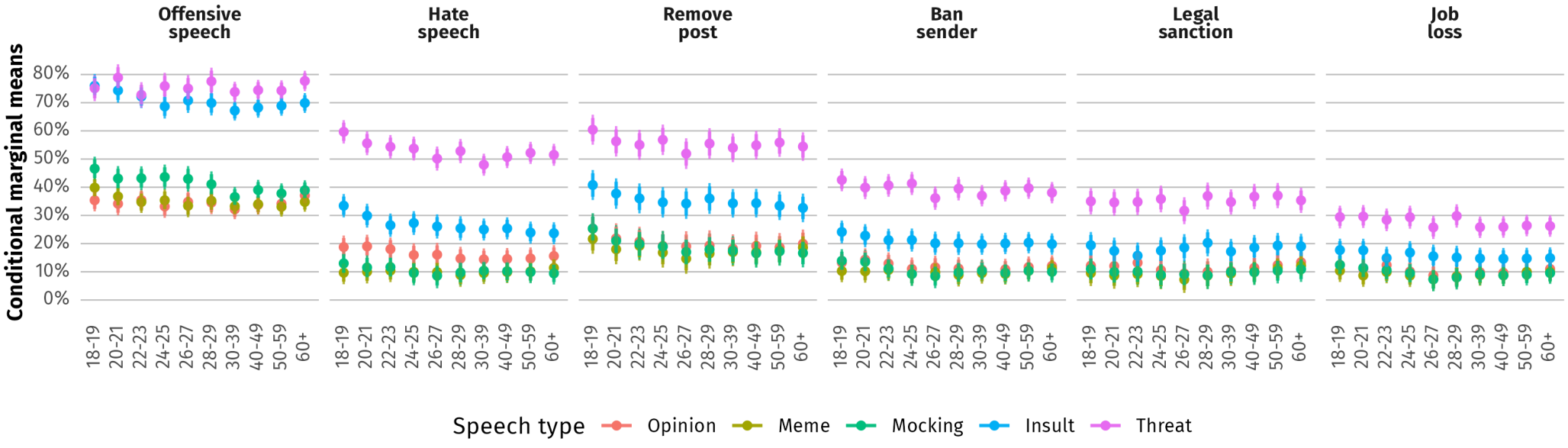
Zooming in: age and gender differences

Hertie School



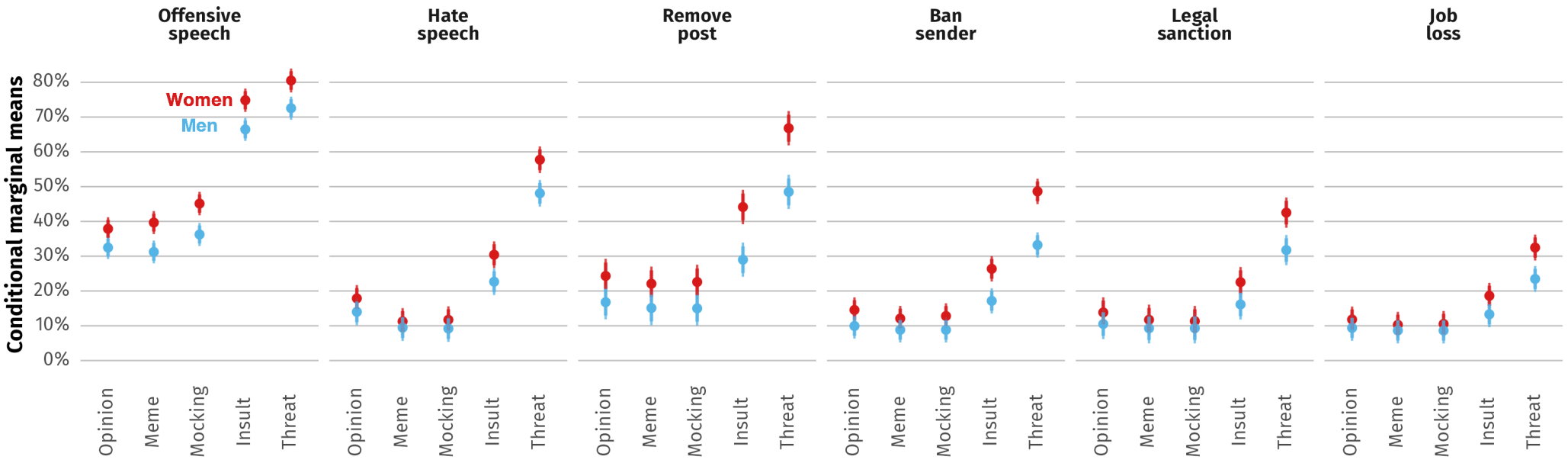
Zooming in: age differences across content types

Hertie School



Zooming in: gender differences across content types

Hertie School



Time for a new agenda

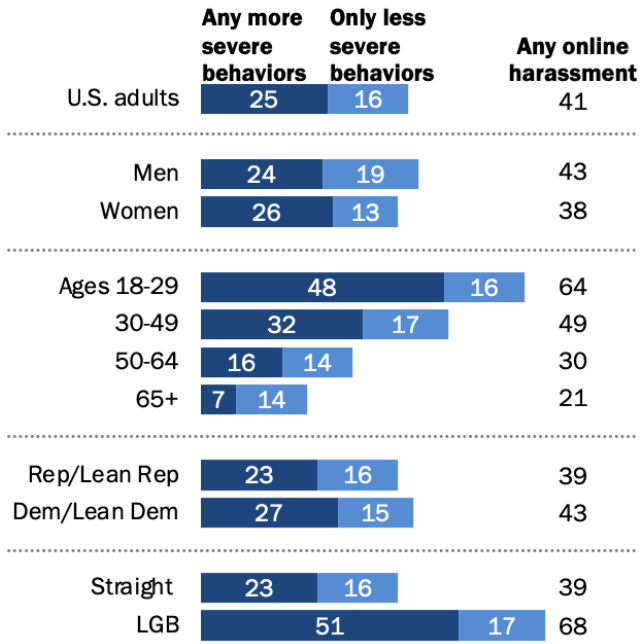

MUSIC BAND
How do you do, fellow kids?
peacock

Online harassment: a problem of the young...

Hertie School

Roughly two-thirds of adults under 30 have been harassed online

% of U.S. adults who say they have personally experienced ___ online

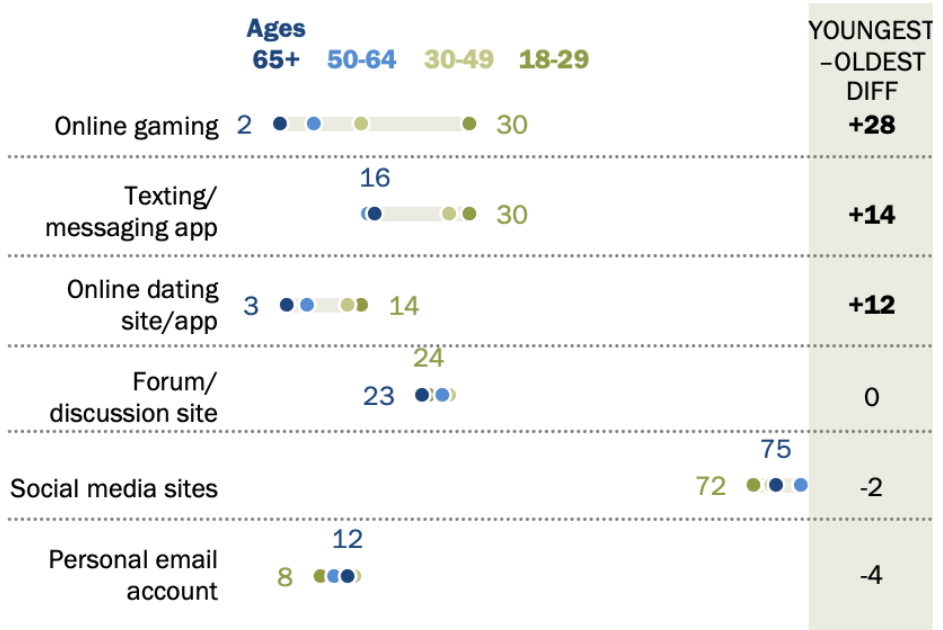


Note: More severe behaviors include being physically threatened, stalked, sexually harassed or harassed for a sustained period of time. Less severe behaviors include being called an offensive name or having someone trying to purposefully embarrass them. LGB indicates those who identify as lesbian, gay or bisexual. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

PEW RESEARCH CENTER

Younger online harassment targets more likely to say they were harassed online most recently while gaming, messaging or online dating

Among the 41% of U.S. adults who have personally experienced online harassment, % who say their most recent experience occurred in the following online environments



Note: Statistically significant differences in **bold**. Figures may not subtract to the DIFF value due to rounding. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

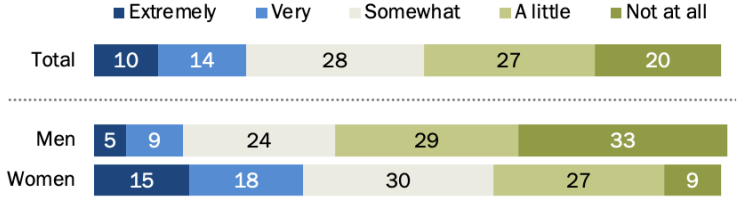
PEW RESEARCH CENTER

… and with variation across genders

Hertie School

Women targeted in online harassment are more than twice as likely as men to say most recent incident was very or extremely upsetting

Among the 41% of U.S. adults who have personally experienced online harassment, % who say their most recent experience was ___ upsetting

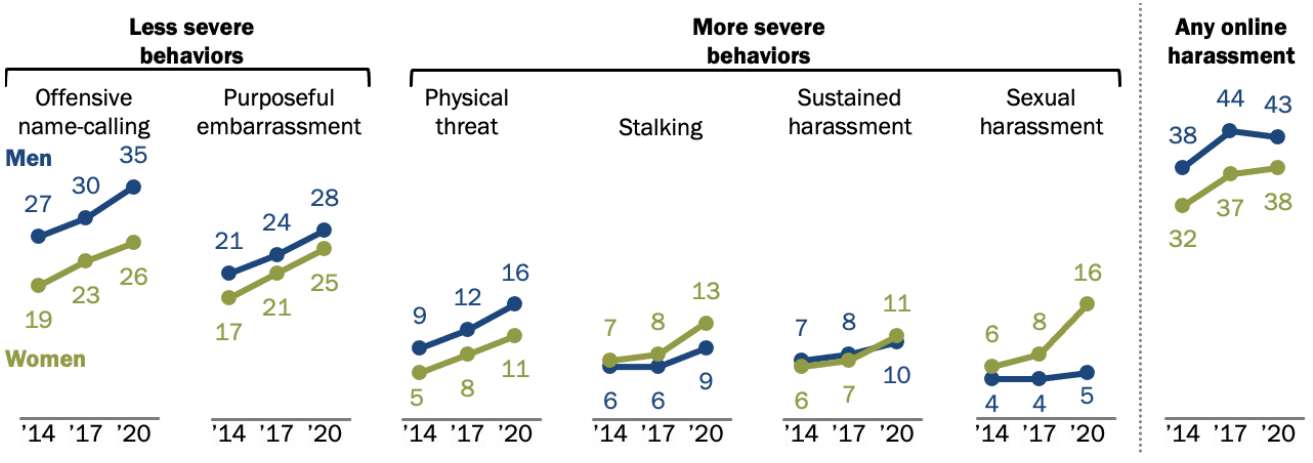


Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

PEW RESEARCH CENTER

Share of women who report being sexually harassed online has doubled since 2017

% of U.S. adults who say they have personally experienced the following behaviors online



Note: For 2020 estimates, respondents are grouped according to their gender. Prior to 2020, groupings were defined according to their biological sex. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

PEW RESEARCH CENTER

Youth language and culture is different and changing fast

Hertie School

We need to broaden our data basis

Hertie School

ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, I. Budak Arpinar

The convenience of social media has also enabled its misuse, potentially resulting in toxic behavior. Nearly 66% of internet users have observed online harassment, and 41% claim personal experience, with 18% facing severe forms of online harassment. This toxic communication has a significant impact on the well-being of young individuals, affecting mental health and, in some cases, resulting in suicide. These communications exhibit complex linguistic and contextual characteristics, making recognition of such narratives challenging. In this paper, we provide a multimodal dataset of toxic social media interactions between confirmed high school students, called ALONE (AdoLescents ON twittEr), along with descriptive explanation. Each instance of interaction includes tweets, images, emoji and related metadata. Our observations show that individual tweets do not provide sufficient evidence for toxic behavior, and meaningful use of context in interactions can enable highlighting or exonerating tweets with purported toxicity.

ChildGuard: A Specialized Dataset for Combatting Child-Targeted Hate Speech

Gautam Siddharth Kashyap, Mohammad Anas Azeez, Rafiq Ali, Zohaib Hasan Siddiqui, Jiechao Gao, Usman Naseem

Hate speech targeting children on social media is a serious and growing problem, yet current NLP systems struggle to detect it effectively. This gap exists mainly because existing datasets focus on adults, lack age specific labels, miss nuanced linguistic cues, and are often too small for robust modeling. To address this, we introduce ChildGuard, the first large scale English dataset dedicated to hate speech aimed at children. It contains 351,877 annotated examples from X (formerly Twitter), Reddit, and YouTube, labeled by three age groups: younger children (under 11), pre teens (11--12), and teens (13--17). The dataset is split into two subsets for fine grained analysis: a contextual subset (157K) focusing on discourse level features, and a lexical subset (194K) emphasizing word-level sentiment and vocabulary. Benchmarking state of the art hate speech models on ChildGuard reveals notable drops in performance, highlighting the challenges of detecting child directed hate speech.

Hateful Messages: A Conversational Data Set of Hate Speech produced by Adolescents on Discord

Jan Fillies, Silvio Peikert, Adrian Paschke

With the rise of social media, a rise of hateful content can be observed. Even though the understanding and definitions of hate speech varies, platforms, communities, and legislature all acknowledge the problem. Therefore, adolescents are a new and active group of social media users. The majority of adolescents experience or witness online hate speech. Research in the field of automated hate speech classification has been on the rise and focuses on aspects such as bias, generalizability, and performance. To increase generalizability and performance, it is important to understand biases within the data. This research addresses the bias of youth language within hate speech classification and contributes by providing a modern and anonymized hate speech youth language data set consisting of 88.395 annotated chat messages. The data set consists of publicly available online messages from the chat platform Discord. ~6,42% of the messages were classified by a self-developed annotation schema as hate speech. For 35.553 messages, the user profiles provided age annotations setting the average author age to under 20 years old.

Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text

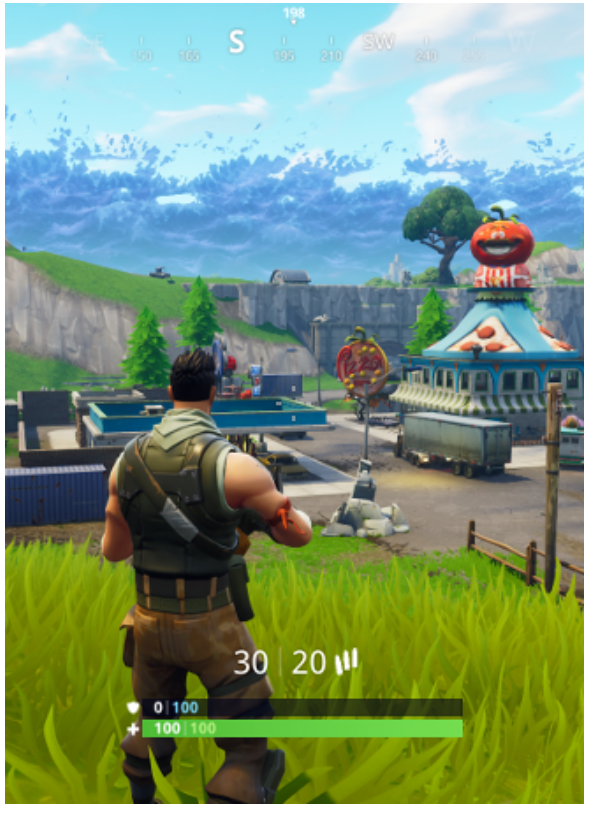
Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar
Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland Galway
{shardul.suryawanshi, bharathi.raja, mihael.arcan, paul.buitelaar}@insight-centre.org

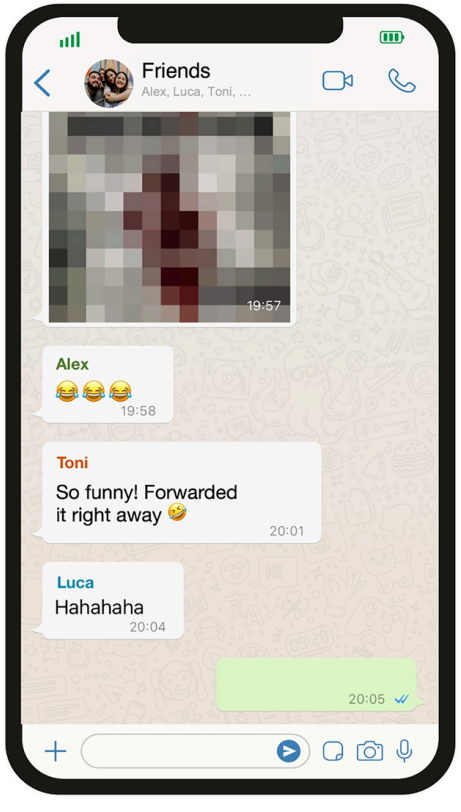
Abstract

A meme is a form of media that spreads an idea or emotion across the internet. As posting meme has become a new form of communication of the web, due to the multimodal nature of memes, postings of hateful memes or related events like trolling, cyberbullying are increasing day by day. Hate speech, offensive content and aggression content detection have been extensively explored in a single modality such as text or image. However, combining two modalities to detect offensive content is still a developing area. Memes make it even more challenging since they express humour and sarcasm in an implicit way, because of which the meme may not be offensive if we only consider the text or the image. Therefore, it is necessary to combine both modalities to identify whether a given meme is offensive or not. Since there was no publicly available dataset for multimodal offensive meme content detection, we leveraged the memes related to the 2016 U.S. presidential election and created the MultiOFF multimodal meme dataset for offensive content detection dataset. We subsequently developed a classifier for this task using the MultiOFF dataset. We use an early fusion technique to combine the image and text modality and compare it with a text- and an image-only baseline to investigate its effectiveness. Our results show improvements in terms of Precision, Recall, and F-Score. The code and dataset for this paper is published in *https://github.com/bharathichezhiyan/Multimodal-Meme-Classification-Identifying-Offensive-Content-in-Image-and-Text*

We need to go places where this happens

Hertie School





File: ass in the ass panda bears.jpg (125 KB, 1088x726)

/uhg/ - Ukraine Happening General #18292

Anonymous (ID: DKPMqBua) 10/16/25(Thu)01:23:29 No.518993518 [Reply]

Previous: >>518980474

Day 1,331 – Daily assessment: https://understandingwar.org/analysis/russia-ukraine/#research

▶Latest
>Russia to mobilise from reserve soldiers of 2 million
>Trump says russia has taken 1.5million casualties.
>Ukraine has liberated Alekseevka
>Russia's current account in Q2 falls to just $5.4B the lowest level of the war
>The Ukrainian defense forces pushed the enemy back further near Nowe Schachowe, Solotyj Kolodjas, Kutscherowo Jar, and Nowopawliwka, - DeepState
>No arab countries attended russia's summit, despite nearly all being invited
>Khoroshe, Novoselivka, and Sichneve have been liberated
>Mali Scherbaky & Sherbaky proper has been liberated
>Pokvorsk is no longer our focus, the main focus now is to capture Prymorske - gerasimov, supreme commander of the russian army
>Feodosia oil terminal is still burning for 5 days now

Comment too long. Click here to view the full text.

283 replies and 134 images omitted. Click here to view.

Anonymous (ID: ZgWm6VLk) 10/16/25(Thu)09:52:36 No.519019501 ▶
File: 1665668893968539.png (802 KB, 941x520)

So how GOOOOOOOOOOOOOOOOOD was he MOOOOOOOOOOOOOOOOOOOOOORNING so far comrades? How is Total Hohol Deelectrification coming along?

Anonymous (ID: CtKPGSWC) 10/16/25(Thu)09:53:22 No.519019534 ▶ >>519019693

it's telling how uncreative you fags are.
totally fueled by soviet nostalgia
top kek honestly
meanwhile we enjoy the collapse of rotten poccia

Towards a better understanding of youth speech preferences

Hertie School

Why this matters

- Language is identity for the youth; context (irony, memes!) is crucial
- Slang evolves faster than a millenial can scroll
- Platform use and communication styles are often gendered
- Current regulation rarely reflects young people's lived experiences

Where we need to go

- **Better data:** systematic mapping of how adolescents use and interpret toxic or provocative speech
- **Better ethics:** frameworks that take seriously the sensitivities of studying youth speech norms
- **Better dialogue:** between researchers, platforms, and young people themselves

Data on toxic youth speech + moderation preferences

Hertie School

What we have

ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, I. Budak Arpinar

The convenience of social media has also enabled its misuse, potentially resulting in toxic behavior. Nearly 66% of internet users have observed online harassment, and 41% claim personal experience, with 18% facing severe forms of online harassment. This toxic communication has a significant impact on the well-being of young individuals, affecting mental

Hateful Messages: A Conversational Data Set of Hate Speech produced by Adolescents on Discord

Jan Fillies, Silvio Peikert, Adrian Paschke

With the rise of social media, a rise of hateful content can be observed. Even though the understanding and definitions of hate speech varies, platforms, communities, and legislature all acknowledge the problem. Therefore, adolescents are a new and active group of social media users. The majority of adolescents experience or witness online hate speech. Research in

Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar
Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland Galway

Plus: Tools to generate synthetic content at scale, solid preference evaluation pipeline, comparative surveys on preferences (gender 👍, youth 👎)

What I'd like us to have

- 📣 Longitudinal behavioral **data on youth speech** and **platform use**
- 📣 **Platform moderation data**, ideally linked to user demographics
- 📣 **Everything all at once:** An ongoing, comparative, panel+cross-section youth survey infrastructure that integrates behavioral data and a module on speech tolerance