NLP for the People or the People for NLP?

25-Nov-2025

Zeerak Talat
z@zeerak.org | @zeeraktalat
www: zeerak.org

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
Edinburgh Futures Institute | **Centre for Technomoral Futures**

Thingstätte (Yesterday)

Is NLP for the People or are the People for the consumption of NLP technologies?

Content Moderation

Hateful Symbols or Hateful People?
Predictive Features for Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Dirk Hovy
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Understanding Abuse:
A Typology of Abusive Language Detection Subtasks

Zeerak Waseem
Department of Computer Science
University of Sheffield
United Kingdom
z.w.butt@sheffield.ac.uk

Thomas Davidson
Department of Sociology
Cornell University
Ithica, NY
trd54@cornell.edu

Dana Warmsley
Department for Applied Mathematics
Cornell University
Ithica, NY
dw457@cornell.edu

Ingmar Weber
Qatar Computing Research Institute
HBKU
Doha, Qatar
iweber@hbku.edu.qa

Automated Hate Speech Detection and the Problem of Offensive Language

Thomas Davidson,[1] **Dana Warmsley,**[2] **Michael Macy,**[1,3] **Ingmar Weber**[4]
[1]Department of Sociology, Cornell University, Ithaca, NY, USA
[2]Department of Applied Mathematics, Cornell University, Ithaca, NY, USA
[3]Department of Information Science, Cornell University, Ithaca, NY, USA
[4]Qatar Computing Research Institute, HBKU, Doha, Qatar
{trd54, dw457, mwmacy}@cornell.edu, iweber@hbku.edu.qa

Are You a Racist or Am I Seeing Things?
Annotator Influence on Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

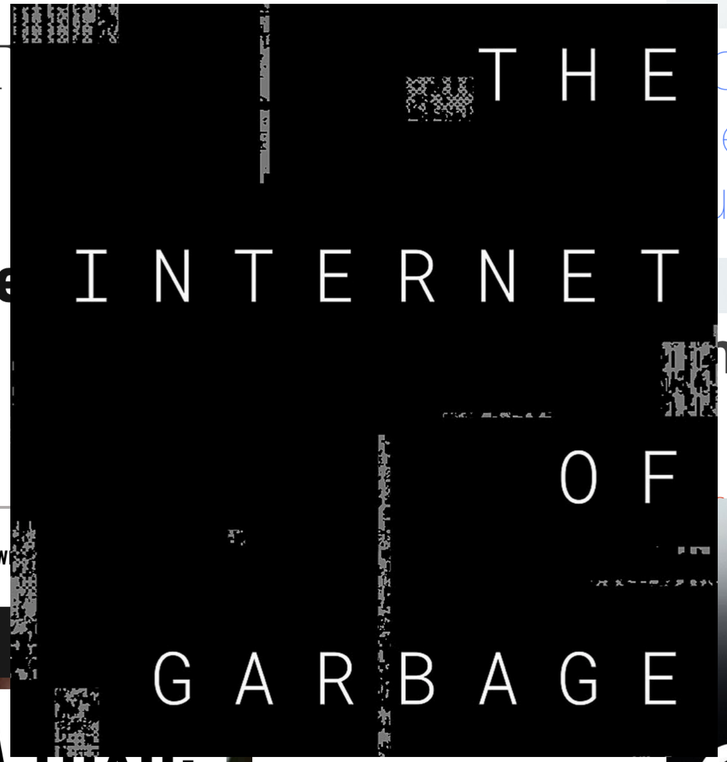
Clean up the Internet

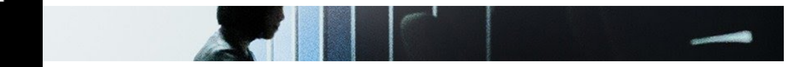
Clean up the internet is an independent, UK-based organisation concerned about the degradation in online discours its implications for democracy. We campaign for evidence-based action to increase civility and respect online, and to online bullying, trolling, intimidation, and misinformation.

Defund Hate Speech: T
Clean Is Upon Us

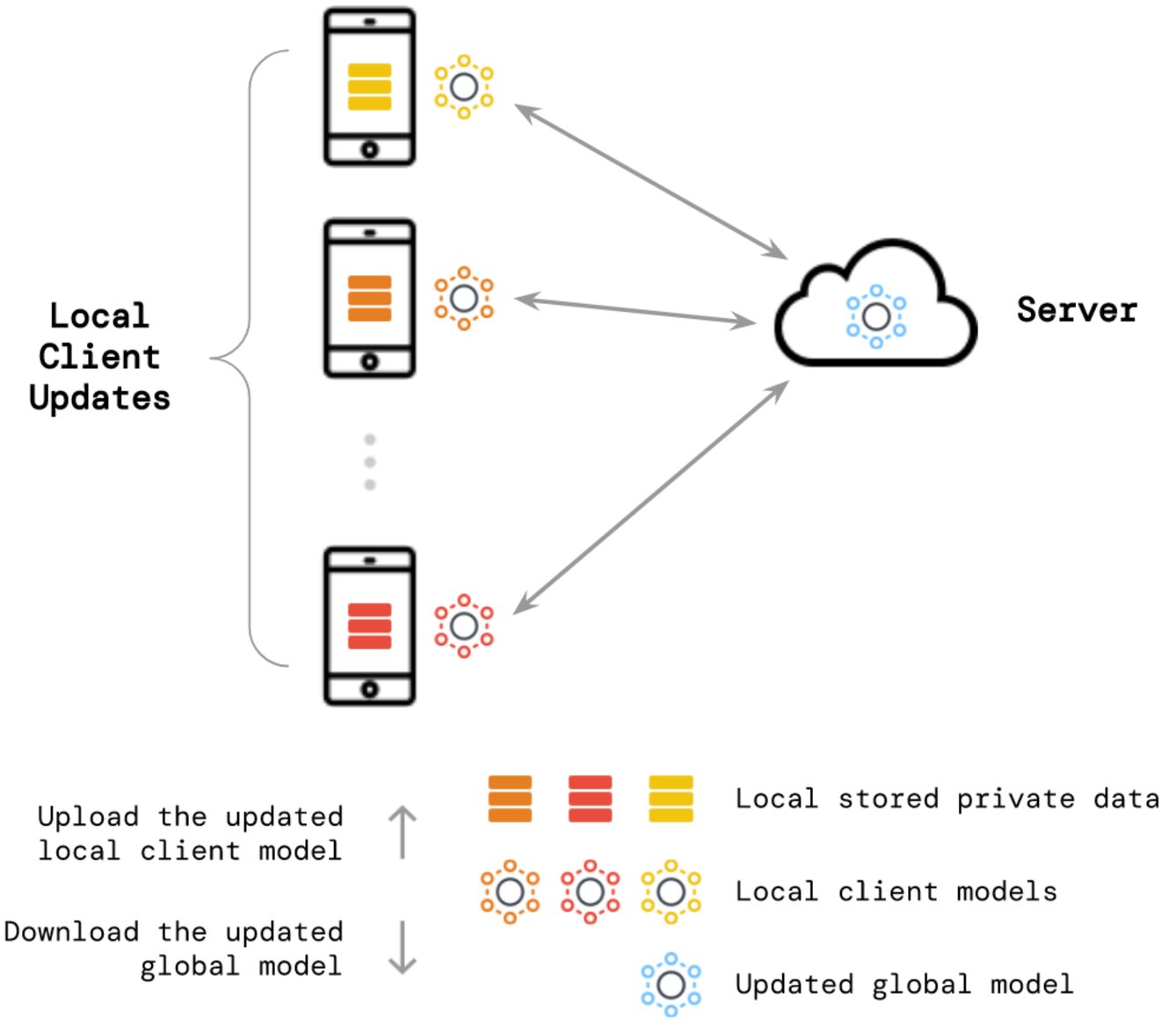
Content on Digital Platforms is on
e. How Can Brands Rebuild
umer Trust?

Abuse, racism and hate spe
the comments online



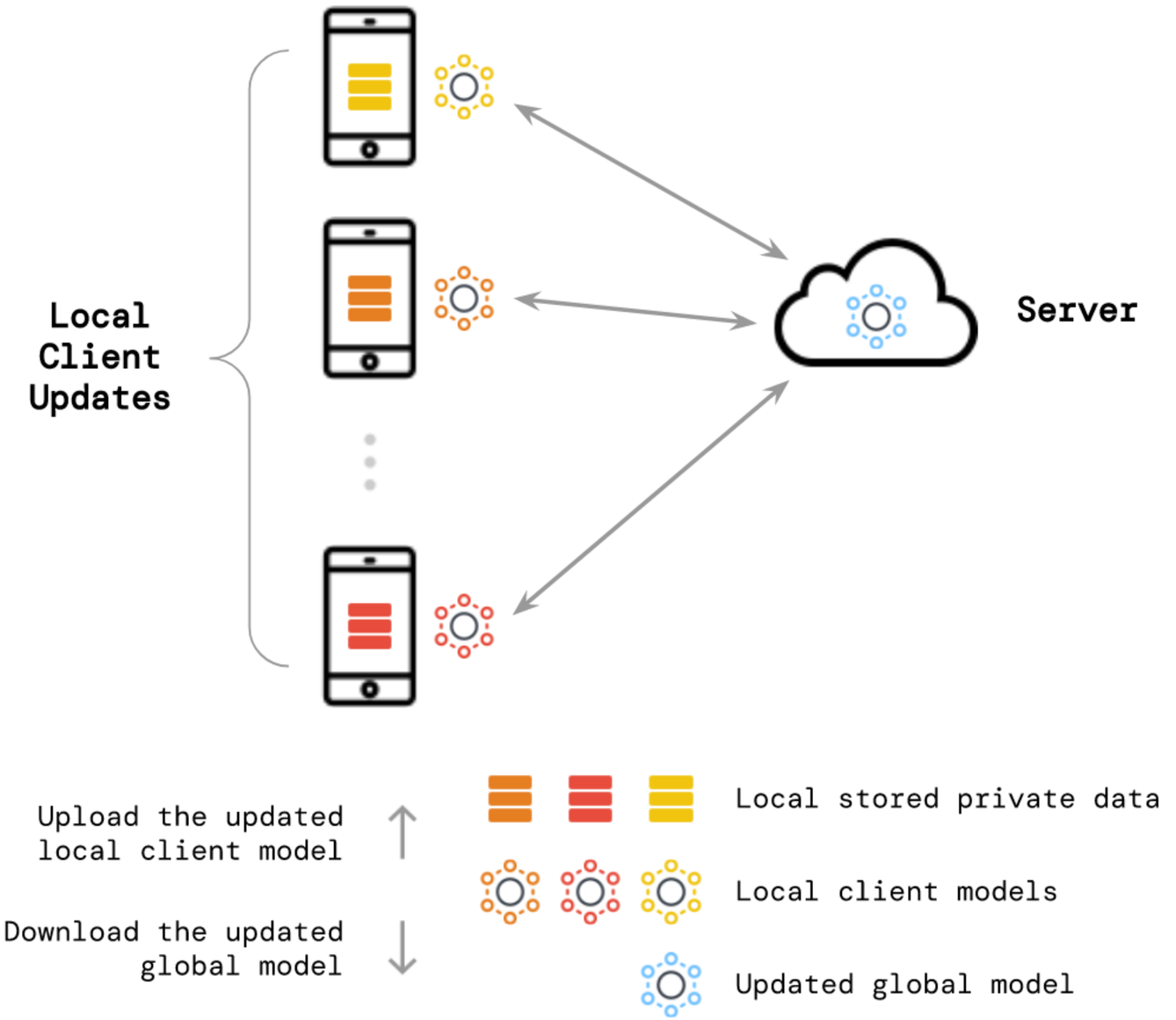


nilever warns social media
clean up "toxic" content



Federated Learning Method

	Centralised			Federated	
	Precision	Recall	F1	F1	
LogReg	69.11	57.45	62.20	69.09	30% of data
Bi-LSTM	71.43	66.64	67.90	69.15	30% of data
FNet	71.35	64.73	66.58	71.15	10% of data
DistilBERT	73.99	69.01	69.39	72.34	50% of data
RoBERTa	**75.45**	**70.58**	**71.03**	**72.61**	10% of data



Federated Learning Method

- The Unreasonable Actor
  - Enforcement problem
- Individual vs. Systemic issues

NLP and the Social Sphere

A Survey on Gender Bias in Natural Language Processing

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models

A Survey on Gender Bias in Natural Language Processing

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

Measuring Bias in Contextualized Word Representations

Mitigating Gender Bias in Natural Language Processing: Literature Review

Multi-Dimensional Gender Bias Classification

ammer as Woman is to Homemaker?

Identifying and Reducing Gender Bias in Word-Level Language Models

Assessing Gender Bias in Machine Translation – A Case Study with Google Translate

er Bias in Abusive Language Detection

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender in Word Embeddings But do not Remove Them

Mitigating Gender Bias Amplification in Distribution by Posterior Regularization

Examining Gender Bias in Languages with Grammatical Gender

English: Templates	English: Biased Sentences	French: Templates	French: Biased Sentences
GENDER-PL talk a lot.	Women talk a lot.	Les GENDER:FEM-PL sont bavardes.	Les femmes sont bavardes.
GENDER-PL talk a lot.	Men talk a lot.	Les GENDER:MASC-PL sont bavards.	Les hommes sont bavards.

Mitchell et al., (2025) SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models. NAACL.

- ~200 reports of impacts of pretrained models
  - Drops in reporting after Q4 2023
    - Climate impacts
    - Biases in models
- Models degrade (Curry et al., 2024) for
  - Lower class dialects (of English)
  - Variations of African American English (AAE)

Impacts on AAE reported in 2016 (Blodgett et al. & Jørgensen et al.) …

“[L]anguage models’ attitudes about AAE are even more negative than the most negative experimentally recorded human attitudes about African Americans, i.e., the ones from the 1930s.”

Hoffman et al., Dialect prejudice predicts AI decisions about people’s character, employability, and criminality. Nature. 2025

FEMINIST DATA MANIFEST-NO

The Manifest-No is a declaration of refusal and commitment. It refuses harmful data regimes and commits to new data futures.

Photo of Banksy piece. Adam Bowie. CC-BY-NC-SA
https://www.flickr.com/photos/adambowie/
2421922870/in/photostream/

The Future?

Liverpool lineup - Wigan Athletic v Liverpool, 9 March 2010 by Dan Farrimond
https://www.flickr.com/photos/illarterate/4438489374

Man in Repsol Orange White and Blue Motorcycle Racing Gear Riding Sports Bike By Keong Racun. https://www.pexels.com/photo/man-in-repsol-orange-white-and-blue-motorcycle-racing-gear-riding-sports-bike-167569/


WALL·E