All that Glitters ain’t Gold

Examining Machine Learning as Socio-technical Infrastructure

27 Apr 2023

Zeerak Talat
zeerak_talat@sfu.ca | @zeeraktalat
www: zeerak.org

- B.Sc. Computer Science @ University of Copenhagen (UCPH)
- M.Sc. IT & Cognition @ UCPH
  - *Advisor: Dirk Hovy (Bocconi)*
- Ph.D. Computer Science @ University of Sheffield (USFD)
  - *Advisor: Kalina Bontcheva (USFD)*

"Criticisms of the field [of AI], no matter how sophisticated and scholarly they might be, are certain to be met with the assertion that the author simply fails to understand a basic point."

Philip Agre (1997)

[This is a replacement of an image from Disney's Mulan reading "Let's get down to business" because no one is as excited about litigating for copyright as Disney]

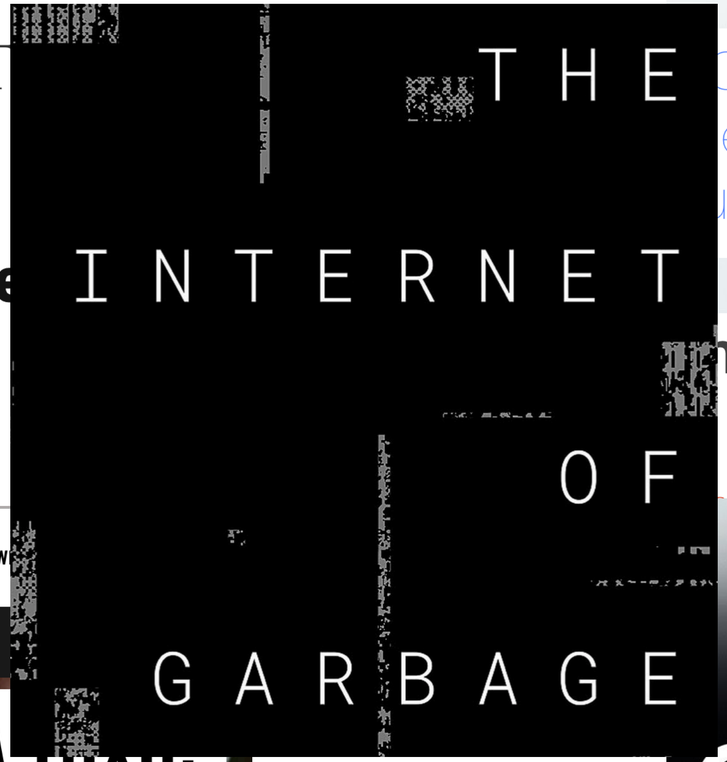
What Even is Content Moderation?

Clean up the Internet

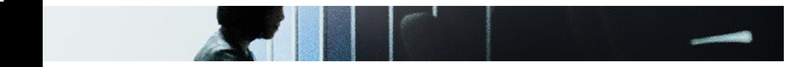
Clean up the internet is an independent, UK-based organisation concerned about the degradation in online discours its implications for democracy. We campaign for evidence-based action to increase civility and respect online, and to online bullying, trolling, intimidation, and misinformation.

Two ways social networks could control toxic content

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

Questions for Content Moderation

1. Who is most affected by Sanitisation Efforts?

   A: Marginalised communities

2. What does it mean to be sanitised?

   A: To be made invisible, as if one does not exist

3. Who can determine what is "dirty"

   A: Those who impose structure on others

4. What does it mean to be dirty?

[I]deas about separating, purifying, demarcating and punishing transgressions have as their main function to impose system on an inherently untidy experience.

Mary Douglas (1978)

“Respectability politics upholds the idea that the supposed worthiness of a marginalized group should be evaluated—that is, by comparing the traits and actions of the marginalized group to the values of respectability set solely by the dominant group.”

Studio ATAO (n.d.)

‘No one ever accused the God of monotheism of objectivity, only of indifference’

Donna Haraway (1988)

Home → Policies → Facebook Community Standards

Hate speech

Policy details User experiences Data

Policy details

CHANGE LOG

Today
Current version

23 Nov 2022

28 Jul 2022

30 Jun 2022

Policy rationale

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

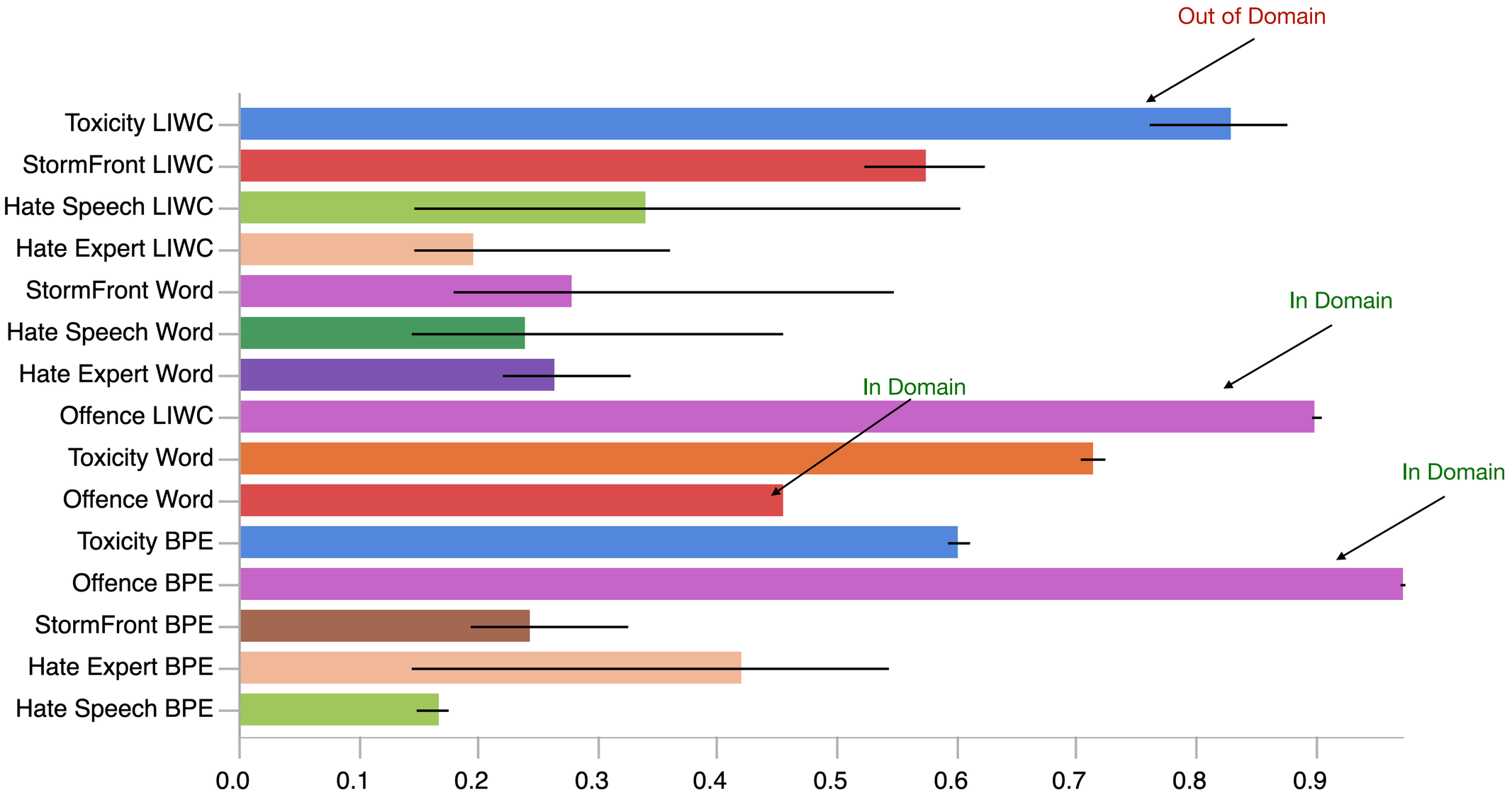
We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We

Image Source: Facebook Policy guidelines (22/04/2023)

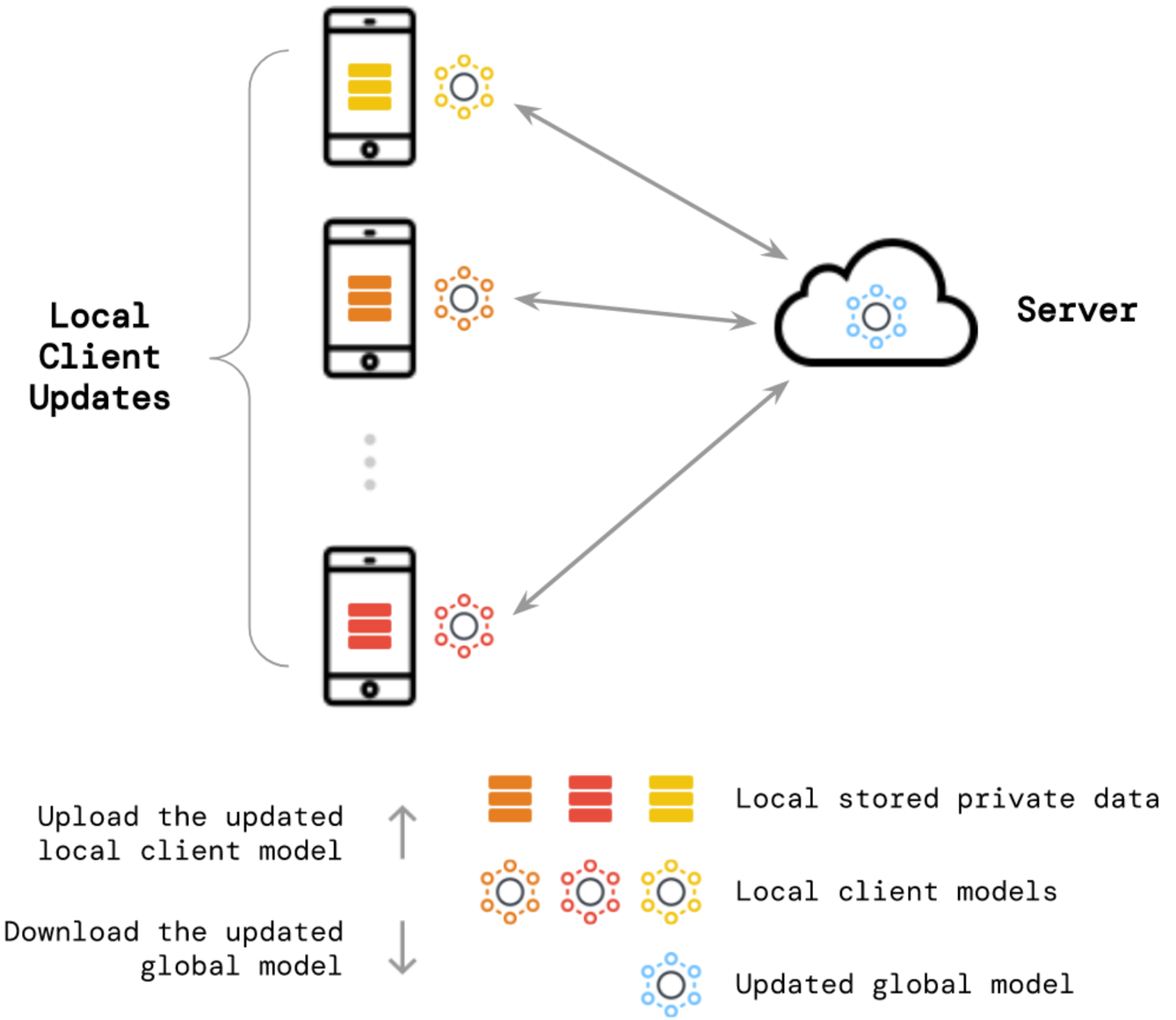
"Respectability politics upholds the idea that the supposed worthiness of a marginalized group should be evaluated—that is, by comparing the traits and actions of the marginalized group to the values of respectability set solely by the dominant group.

Studio ATAO (n.d.)

Action from Hopelessness



In-domain & cross-domain performance of MLP model trained on 'Offence" dataset in terms of F1-score.



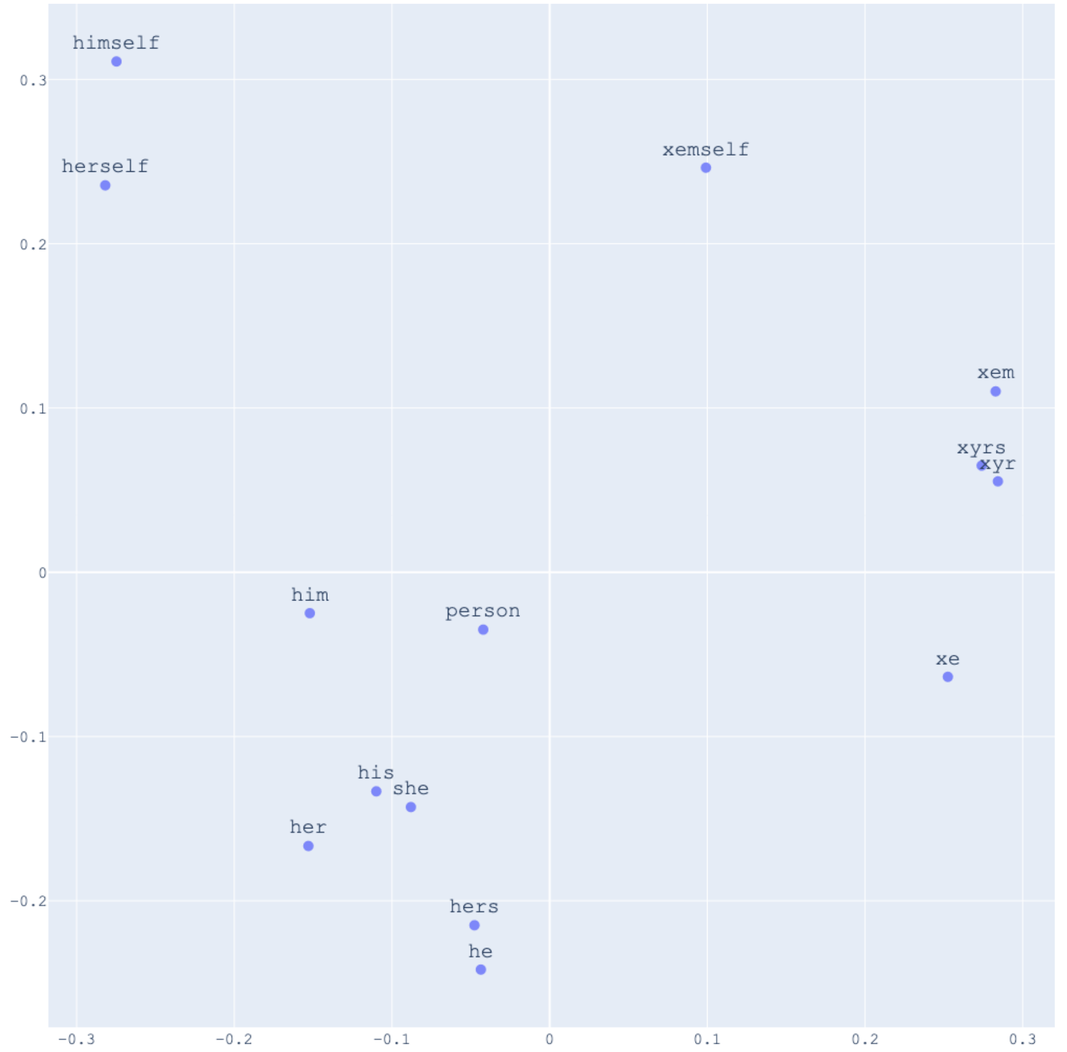
Federated Learning Method

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

Results on combined multi-class dataset using Federated Optimization.


STEVEN SPIELBERG Presents
BACK TO THE FUTURE
A ROBERT ZEMECKIS Film
He was never in time for his classes. . .
He wasn't in time for his dinner. . .
Then one day. . . he wasn't in his time at all.
"BACK TO THE FUTURE" Starring MICHAEL J. FOX
CHRISTOPHER LLOYD · LEA THOMPSON · CRISPIN GLOVER
Written by ROBERT ZEMECKIS & BOB GALE Music by ALAN SILVESTRI Produced by BOB GALE and NEIL CANTON
Executive Producers STEVEN SPIELBERG KATHLEEN KENNEDY and FRANK MARSHALL
AMBLIN ENTERTAINMENT
Directed by ROBERT ZEMECKIS
Soundtrack Available on MCA Records and Cassettes
DOLBY STEREO IN SELECTED THEATRES
Read the BERKLEY Book
A UNIVERSAL Picture
© 1985 Universal City Studios, Inc.
Coming soon to a theatre near you.

All that Glitters ain't Gold



Original

Xe Fiction: Full Fine-tuning

"Imperialism leaves behind germs of rot which we must clinically detect and remove from our land but from our minds as well."

Franz Fanon (1963)

References

1. Agre, P. (1997). Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In G. C. Bowker, S. L. Star, L. Gasser, & W. Turner (Eds.), *Social science, technical systems, and cooperative work: Beyond the great divide*. Lawrence Erlbaum Associates.
2. Birhane, A., Talat, Z. (Forthcoming) It's Incomprehensible: On Machine Learning and Decoloniality. In S. Lindgren (Eds.), Handbook of Critical Studies of Artificial Intelligence. Edward Elgar Publishers.
3. Costanza-Chock, S. "Design Justice, A.I., and Escape from the Matrix of Domination." *Journal of Design and Science*, 2018.
4. Fanon, F. *The Wretched of the Earth*. New York: Grove Press, 2002.
5. Douglas, M.. Purity and Danger: An Analysis of the Concepts of Pollution and Taboo. Repr. London: Routledge, 1978.
6. Foucault, M. (1969). *Archaeology of knowledge*. Routledge.
7. Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599.
8. Jay Gala, Deep Gandhi, Jash Mehta, & Zeerak Talat. (2023). A Federated Approach for Hate Speech Detection. *Proceedings of EACL*.
9. Kirtz, J. L., Talat, Z., Tomlinson, C., & Chun, W. H. K. (In Review). Definitely Maybe: On the specificity of ambiguity in content moderation.
10. Studio ATAO. (n.d.). *Understanding Respectability Politics*. Studio ATAO. Retrieved June 13, 2022, from https://www.studioatao.org/respectability-politics
11. Talat, Z. (2021). "It ain't all good:" Machinic abuse detection and marginalisation in machine learning [University of Sheffield].
12. Talat, Z., daayan, d., Bingel, J., & Augenstein, I. (In Review). Disembodied Machine Learning: On the Illusion of Objectivity in NLP.
13. Talat, Z., & Lauscher, A. (2022). *Back to the Future: On Potential Histories in NLP* (arXiv:2210.06245). arXiv.
14. Thylstrup, N., & Talat, Z. (Forthcoming). Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. First Monday.