Privacy & Surveillance in Content Moderation

03 February 2023

Zeerak Talat
Floating
zeerak_talat@sfu.ca | @zeeraktalat

“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.)

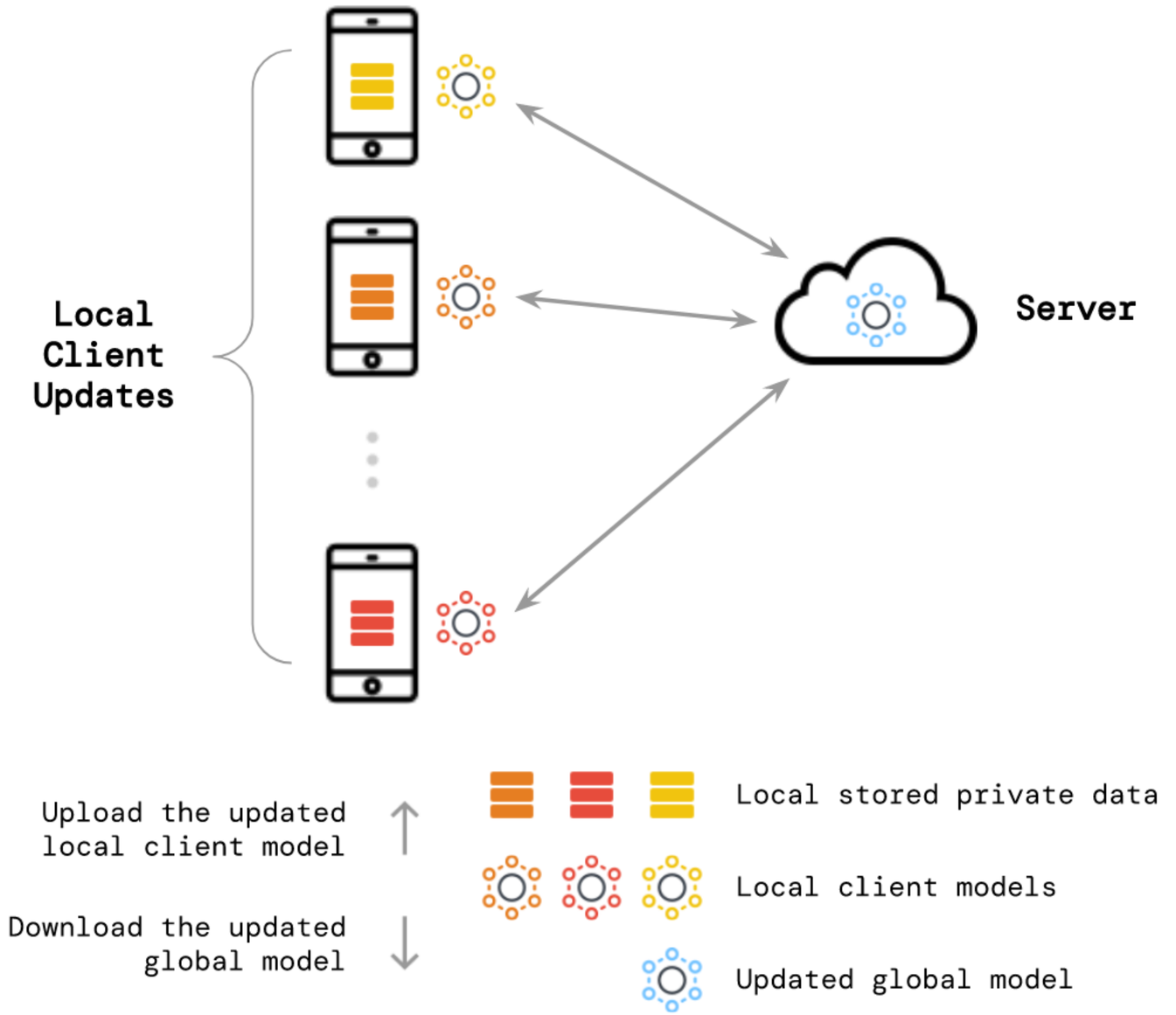
Reception Theory

- Two parties engaged in communicative acts
  - Speaker
  - Audience
- Decoding
  - Dominant
  - Oppositional
  - Negotiated

Reception Theory & CM

Prison of Content Moderation

Source: Inside an Abandoned Panopticon Prison in Cuba. Atlas Obscura. 2017.



Federated Learning Method. Source: Federated Hate Speech Detection. Gala et al. EACL. 2023

	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.

References

1. *Studio ATAO | Understanding Respectability Politics*. (n.d.). Studio ATAO. Retrieved June 13, 2022, from https://www.studioatao.org/respectability-politics
2. Bentham, J. (2010). *The Panopticon writings* (M. Božovič, Ed.). Verso.
3. Hall, S. (1973). Encoding and Decoding in the television discourse. *University of Birmingham*.
4. Jay Gala, Deep Gandhi, Jash Mehta, & Zeerak Talat. (2023). A Federated Approach for Hate Speech Detection. *Proceedings of EACL*.
5. Tod Seelie. (2017). Inside an Abandoned Panopticon Prison in Cuba. *Atlas Obscrua*. https://www.atlasobscura.com/articles/panopticon-prison-cuba