Social Approval and Network Homophily as Motivators of Online Toxicity

Julie Jiang (USC), Luca Luceri (USC), Joe Walther (UCSB), Emilio Ferrara (USC)

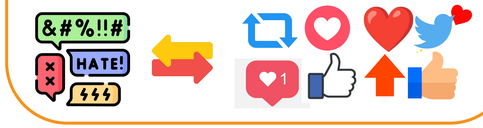
USC Viterbi School of Engineering

USC Information Sciences Institute

Hate Speech
targeted, offensive discourse based on race, religion, gender, etc.

Theory of online hate
(Walther, 2022)

Hateful behavior may be fueled, reinforced, and exacerbated by social approvals from their network



We use a collection of anti-immigration Twitter users in 2020-2021

Gather all of their historical tweets

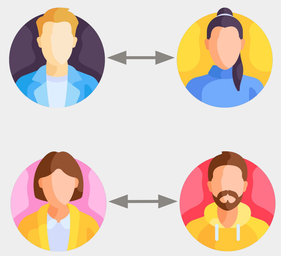
Data

Detect hate scores of tweets and users (Perspective API)

Compile social network interaction data

Icons created by Freepik - Flaticon

RQ1: Is a user's hatefulness related to how hateful their social network is?



Yes, there is network homophily (McPherson, 2001):
Users tend to preferentially associate themselves with **similarly hateful** users

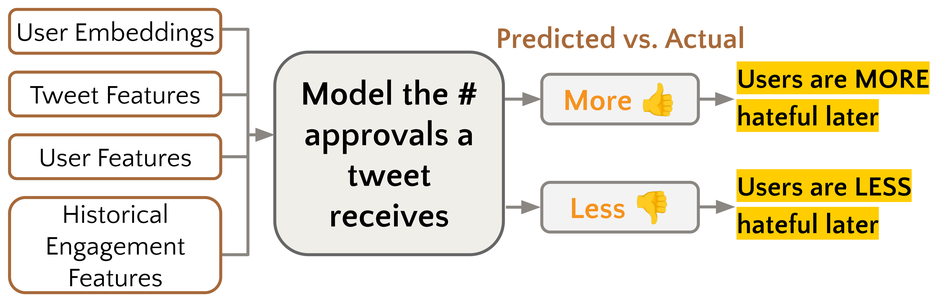
- Significant network assortativity
- High predictability of user hate score from the social network using ML methods (Social-LLM by Jiang et al., 2024)

RQ2: Does receiving more social approval increase a user's subsequent hateful?

We show that toxic behavior could be a socially motivated behavior.

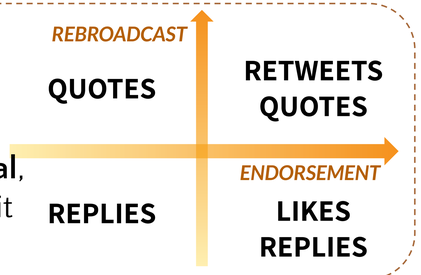
How:

1. Build ML model to predict how many 👍 a tweet receives, based on user, tweet, and relevant historical features
2. Compare predicted # 👍 with actual # 👍, and locate instances where
   a. the tweet got **much more** 👍 than predicted (maybe social approvals)
   b. the tweet got **much fewer** 👍 than predicted (maybe social disapprovals)
3. Evaluate Δ change in average toxicity before and after the social (dis)approval

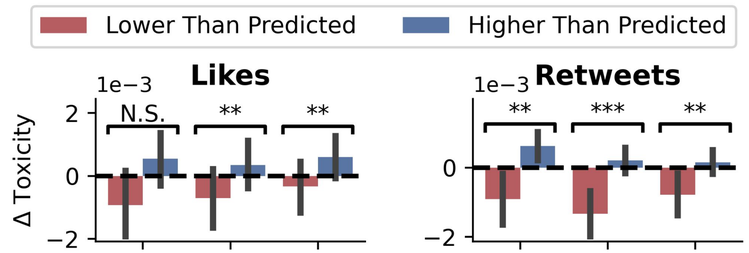


Social approval or disapproval?

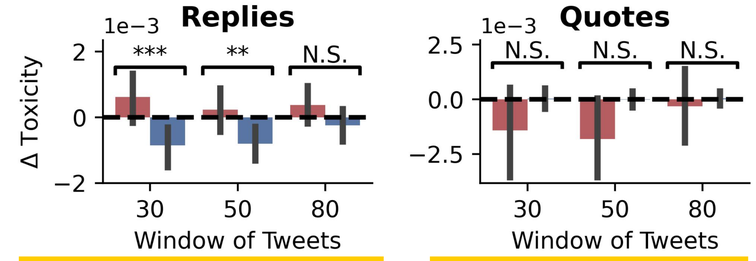
Social engagement can represent social **approval**, but can also be a conduit for social **disapproval**



Effect of Social Engagement on Toxicity



Likes and *retweets*–likely approvals–increase future toxicity



Replies–likely disapprovals–decrease future toxicity

Quotes–could be either–has no significant impact

Implications

- Hate posting is not a solo endeavor but collective behavior. It is "*networked*"
- Content moderation strategies: shadowban, disable visibility of some social approvals, etc.

April 2024