Racism is a virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 crisis

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, Srijan Kumar

Paper Introduction

Context

- Anti-Asian hate speech escalated during pandemic
- Racially motivated

Definitions

Hate speech	*F*ck Chinese scums of the Earth disgusting pieces of sh*t learn how to not kill off your whole population of pigs, chickens, and humans. coronavirus #wuhanflu #ccp #africaswine #pigs #chickenflu nasty nasty China clean your f*****g country.*
Counterspeech	*The virus did inherently come from China but you can't just call it the Chinese virus because that's racist. or KungFlu because 1. It's not a f*****g flu it is a Coronavirus which is a type of virus. And 2. That's also racist.*
Neutral speech	*COVID-19: #WhiteHouse Asks Congress For $2.5 Bn To Fight #Coronavirus: Reports #worldpowers #cli- matesecurity #disobedientdss #senate #politics #news #unsc #breaking #breakingnews #wuhan #wuhanvirus https://t.co/XipNDc*

COVID-HATE Dataset: Tweets

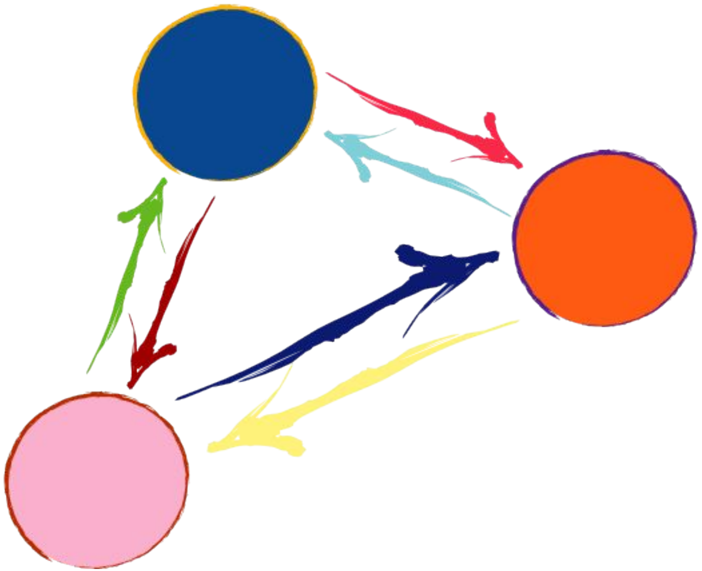
- Used hashtag keywords for three categories of speech to scrape 206M Tweets
- Two annotators annotated sample of ~3K Tweets for three categories

COVID-HATE Dataset: Social Network

- Create social network of 1.3M user nodes who made at least one COVID-19 Tweet and their neighbor nodes
- Categorize users based on their Tweets into categories
  - Hate speech user
  - Counterspeech user
  - Dual speech user
  - Neutral speech user

Paper Contribution

- **Novel contribution**
  - Previous literature: spread of hate speech
  - Interaction between counterspeech and hate speech, dynamics on social media

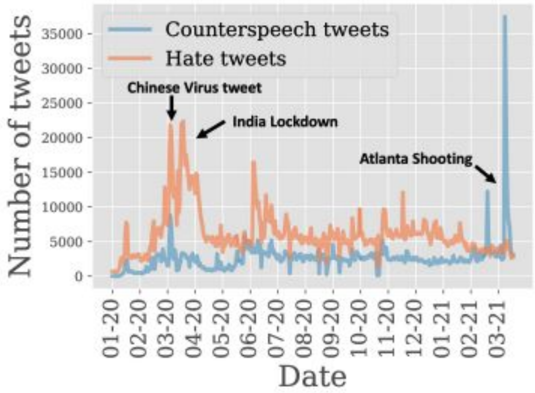


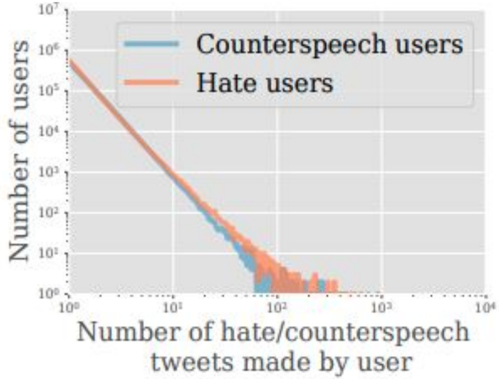
Hate/Counterspeech Classification Model

- Classification task: classify Tweet as hate speech, counterspeech, or neutral speech
- Features: Linguistic, hashtag keyword occurrences, BERT embeddings
- Best performing model: BERT model fine-tuned on labeled Tweets dataset
- Final model used to label the 206M Tweets

Descriptive Analysis

- Number of hate speech and counterspeech Tweets correlates with historical events
- Distribution of hate speech and counterspeech Tweets forms a long-tail
  - A few users generate the majority of the hate speech and counterspeech Tweets





Social Network Connectivity Structure

Intragroup and intergroup connectivity can be explained by

(1) the network graph's inherent structural properties

OR

(2) unique properties/behaviors of nodes in the observed network

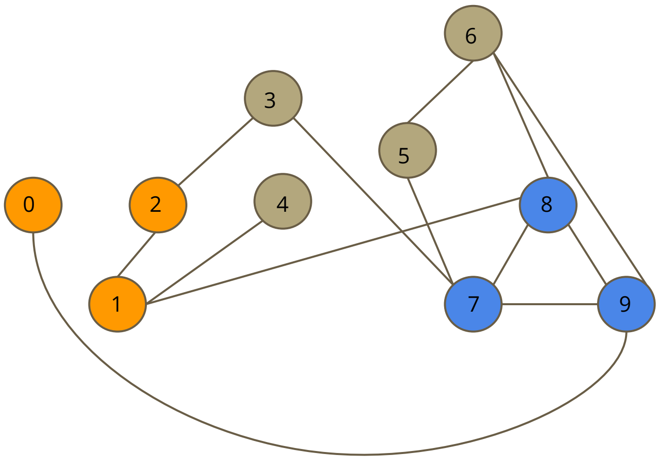
Method: Degree-preserving randomization [1]

To isolate effect due to variable 2, create *baseline networks* by sampling over networks with <u>same graph structure</u> as the **observed network** to estimate and control for effect of variable 1 on connectivity.

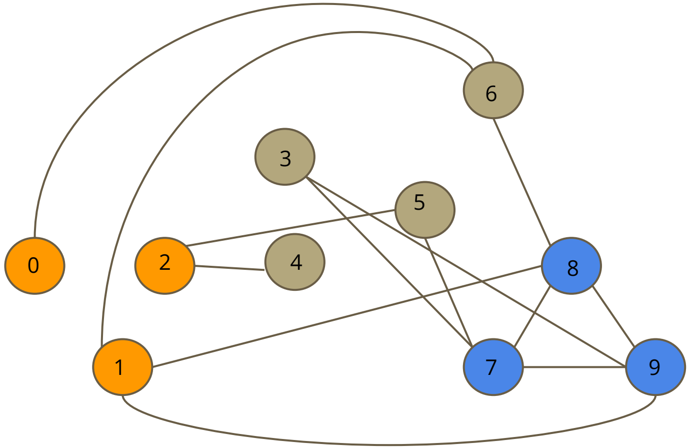
[1] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in Proceedings of the SIGCHI conference on human factors in computing systems, 2010, pp. 1361–1370

Social Network Connectivity Structure

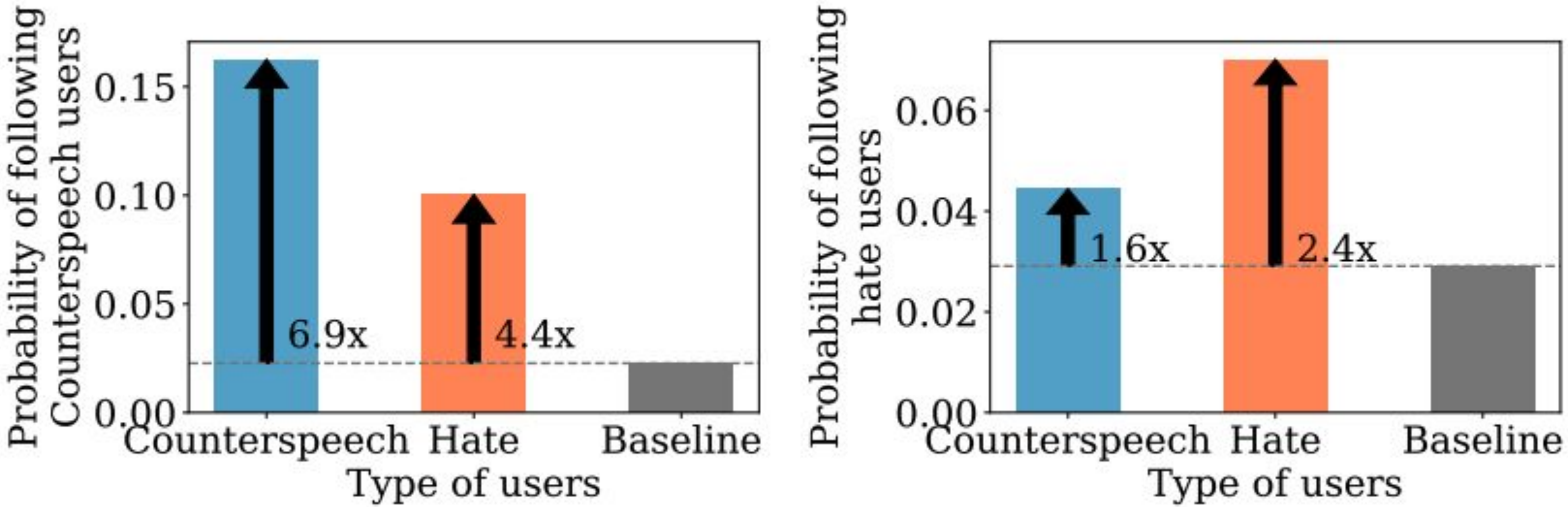
Observed Network



Baseline Network (1 run)



Users display homophily and are highly interconnected



Influence of Counterspeech on Spread of Hate

Method: Event Cascade

Model dynamics of hate/counterspeech infection as an event cascade

- Cascade: temporally-ordered sequence of events of nodes that transition from neutral to hate/counterspeech states
- Each cascade associated with risk function

$$Risk_{s \rightarrow s'}(n) = \frac{|Infected_{s'} \cap Exposed_s(n)|}{|Exposed_s(n)|}$$

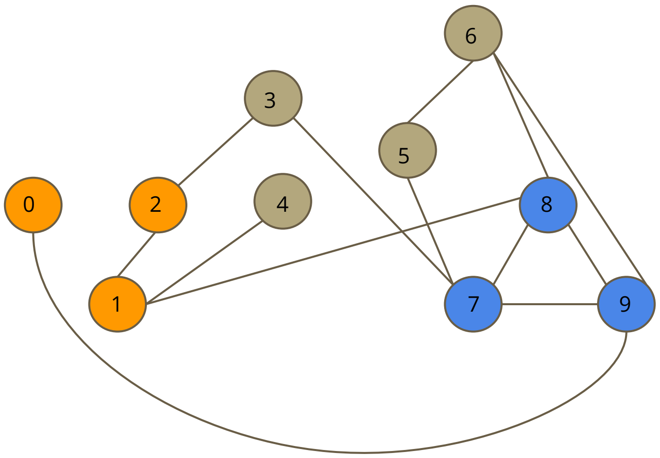
$s \in \{hate, counterspeech\}$

$s' \in \{hate, counterspeech\}$

Probability of user in transitioning from neutral to s' state after exposure to n neighbors in s state

Influence of Counterspeech on Spread of Hate

Observed Network



$\text{Risk}_{\text{hate->hate}}(0) = 1 / 4$

$\text{Risk}_{\text{hate->hate}}(1) = 2 / 6$

Influence of Counterspeech on Spread of Hate

Infection risk can be explained by

(1) Homophily

OR

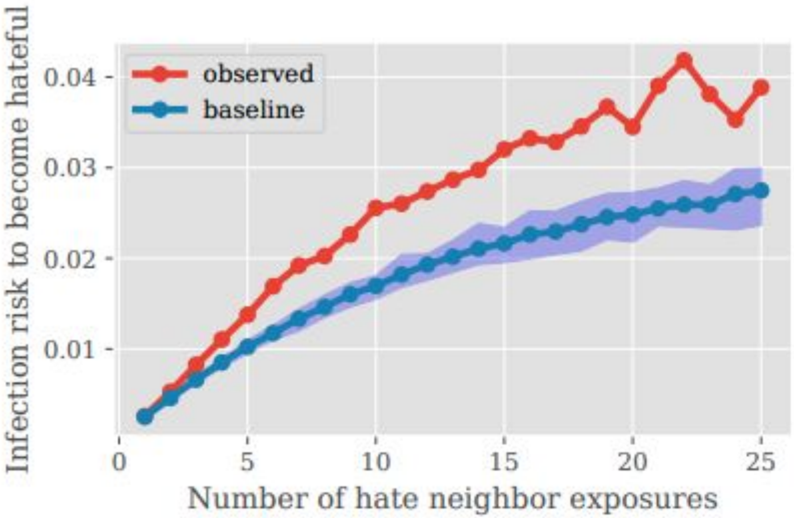
(2) Users' influence on one another in the observed network

Method: Homophily-preserving randomization [2]

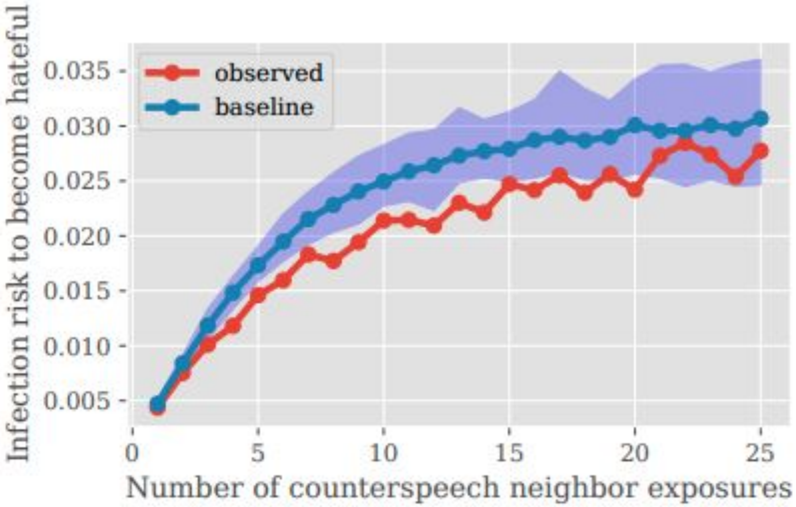
Similar to degree-preserving randomization method for connectivity

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in ACM SIGKDD, 2008.

Exposure to Counterspeech Deters Hate Speech



(a) Hate $\rightarrow$ hate



(b) Counterspeech $\rightarrow$ hate

Peer Review

Strengths

- Large-scale dataset with text and network data for a specific type of hate speech
- Annotation by members of the targeted outgroup, inter-rater agreement validation
- Statistically significant result on the effect of counterspeech on deterring hate speech

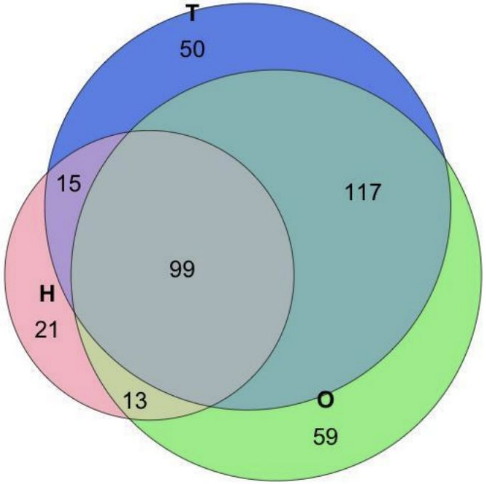
[3]

[3] L. Breitfeller, E. Ahn, D. Jurgens, Y. Tsvetkov. "Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1664–1674, Hong Kong, China, November 3–7, 2019.

Peer Review

Critiques

- Precision in the definition of hate speech



[4]

- Weighting edges of graph by strength of ties may affect outcomes
- For influence model, does following a user to who writes hate/counterspeech imply exposure to hate/counterspeech, given the long-tail distribution?

[4] A. Schöpke-Gonzalez, S. Wu, S. Kumar, P. J. Resnick, L. Hemphill. "How We Define Harm Impacts Data Annotations: Explaining How Annotators Distinguish Hateful, Offensive, and Toxic Comments," arXiv preprint arXiv:2309.15827

Peer Review

Questions

- What are some of the possible failure cases of the text classification model?
  - False positives: Why might counterspeech be misclassified as hate speech?
  - False negatives: Why might hate speech be misclassified as counterspeech?
- How can we reconcile heterogeneous definitions of harmful speech? What factors should affect the degree of intervention?

Follow-up Project

Counterspeech: Integrative Strategies for Combating Online Hate

- **Objectives**
  - Delve deeper into mechanisms of counterspeech
  - Test whether counterspeech could be a viable solution to curb hate
- **Project phases**
  - AI tool development for counterspeech identification + generation
    - Using the same COVID-HATE dataset + designing a new one using the same method
  - Community workshops / pilot programs → data collection on effectiveness of the tool
  - Policy memo based on data collected
- **Expected results**
  - Improved efficiency + impact of counterspeech
  - Longer term: reduced instances of hate speech, shaping public policy

Follow-up Project

Counterspeech: Integrative Strategies for Combating Online Hate

- **Discussion Questions**
  - What metrics would be most informative to measure the “efficiency and impact” of counterspeech?
  - How might the effectiveness of counterspeech vary depending on the platform or context?
  - How can the project balance the need for effective counter-speech with the risk of suppressing free speech?