Rise of Hate Content in Social Media

Animesh Mukherjee

The rise of social media

The social media sites have become the default go to for any sort of information

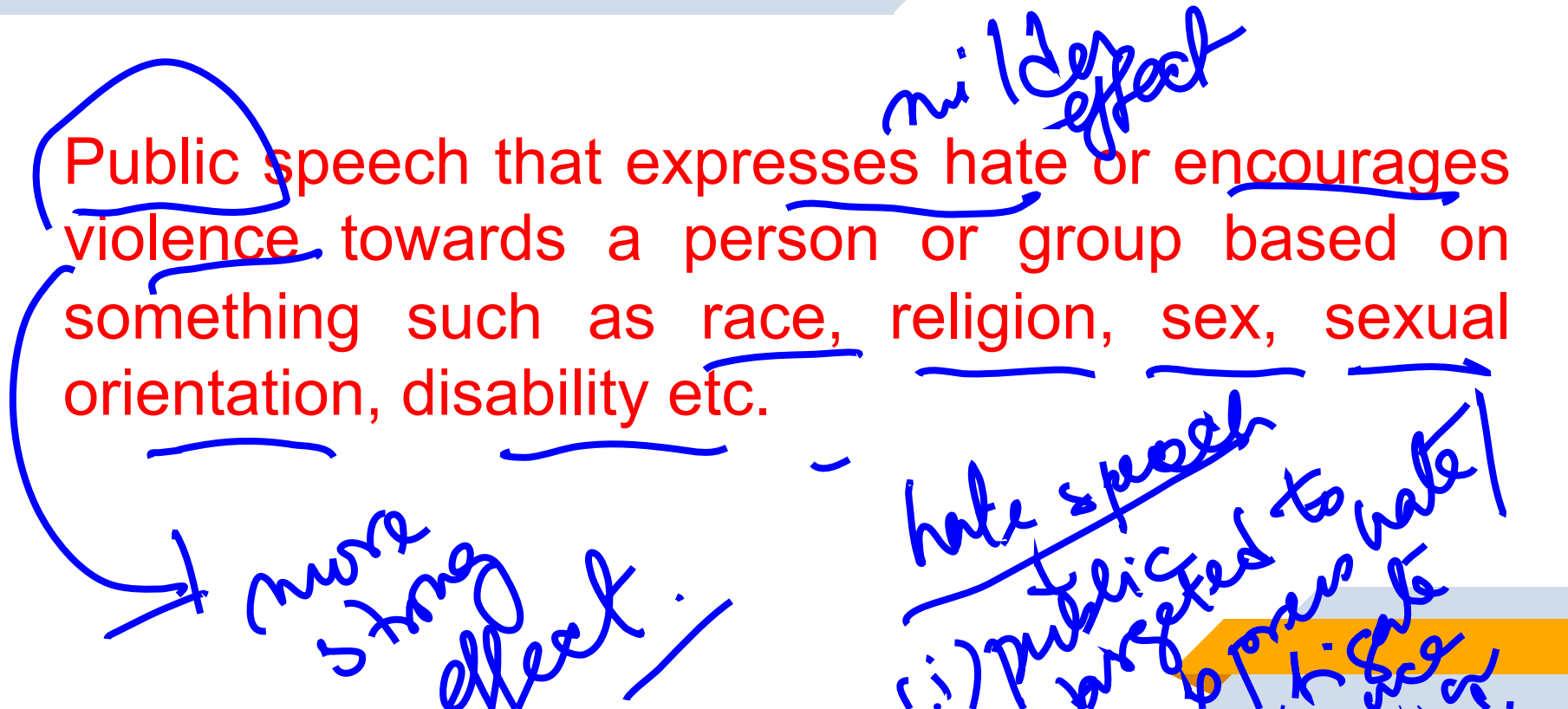
People rate social media as the source of news

Negative consequences

- Increased polarization
- Abuse
- Hate speech

Definition

Public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, sexual orientation, disability etc.



Negative consequences

Bulandshahr Violence

Pittsburg Shooting

Christchurch Shooting

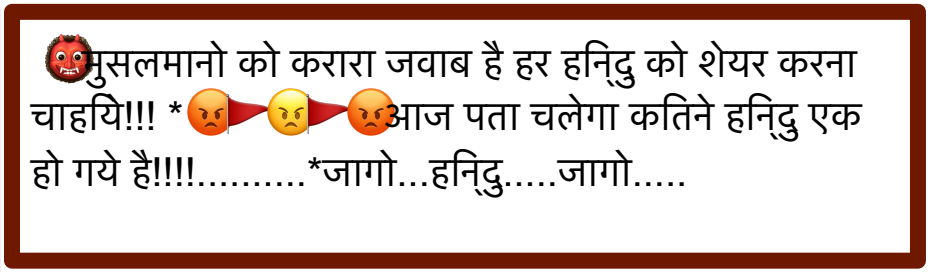
Rohingya Genocide

Sri Lanka riot

Hate speech across platforms

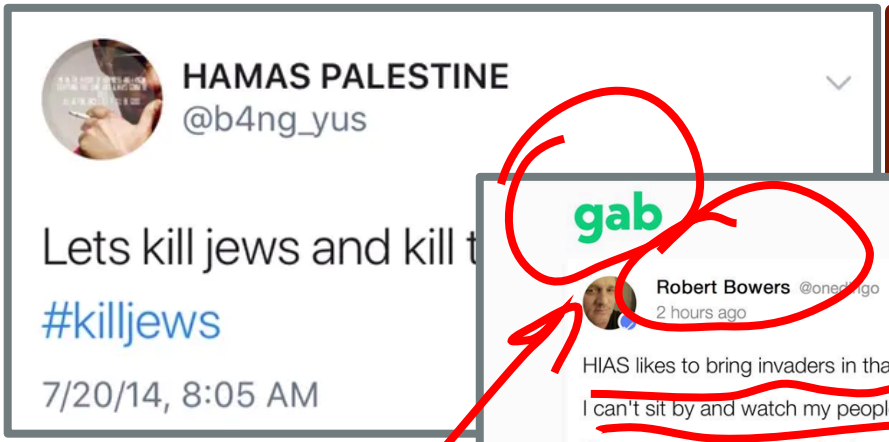
Hate speech across platforms

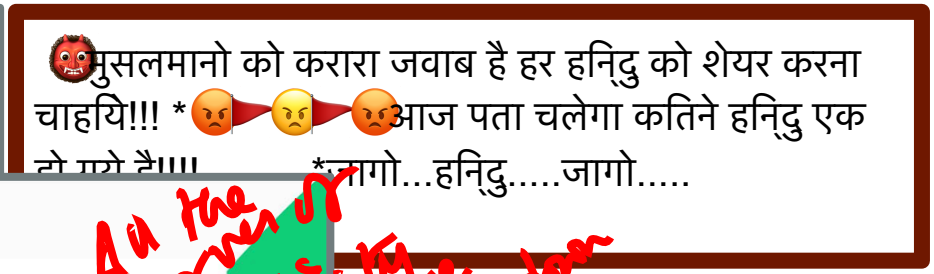
Twitter

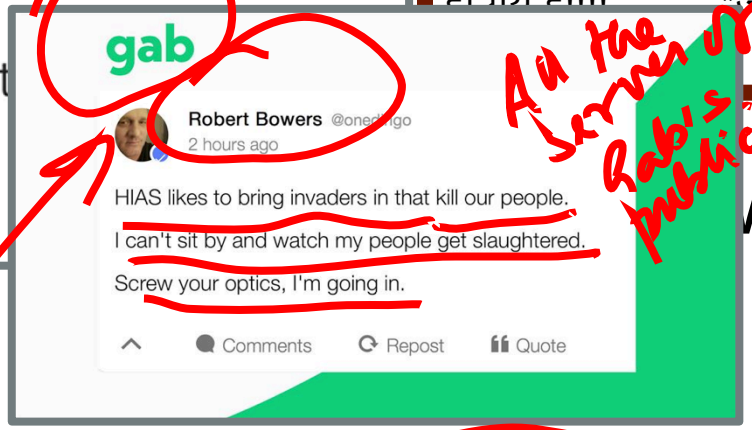


Whatsapp

Hate speech across platforms







Twitter

Whatsapp

Gab . com

Amplification of hate?

- The public expression of hate speech promotes the devaluation of minority members
- Frequent and repetitive exposure to hate speech could increase an individual's outgroup prejudice
- Echo chamber of hate
- State sponsored hate

Spread of hate speech in online social media

ACM WebSci 2019, Best Paper Honorable Mention

Unmoderated spread of hate speech

Gab.com

What is Gab?

- Promotes itself as **“Champion of free speech”**.
- Criticised as an echo-chamber for **“alt-right users”.**
- Similar to Twitter in terms of posting content but has loose (almost no) moderation.
- Gab promotes “free-speech”, allowing users to post hateful content without fear of repercussion.

Data collection

Curated a massive dataset of 21M posts and 343K users using the Gab API.

Have information about:

- Basic details of each user like username
- All the posts of the user
- Followers and followings of the user.

Identify hateful users

- Create a **seed-set** of hateful users (each doing 10+ hate posts).
- Create a **repost network** of the users and corresponding **belief network**.
- Initialize the seed-set of hateful users with **score 1** and rest **with 0**.
- Run a **diffusion model** on the belief network[1,2] to update the beliefs of the users iteratively.
- Users with **belief scores [.75, 1]** → hateful **(KH)** = 2,290.
- Users with **belief scores [0, .25]** → non-hateful **(NH)** = 58,803.

[1] Golub, B. and Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. American Economic Journal: Microeconomics, 2(1): 112–49

[2] Ribeiro, M., Calais, P., Santos, Y., Almeida, V., and Jr., W. M. (2018a). Characterizing and detecting hateful users on Twitter

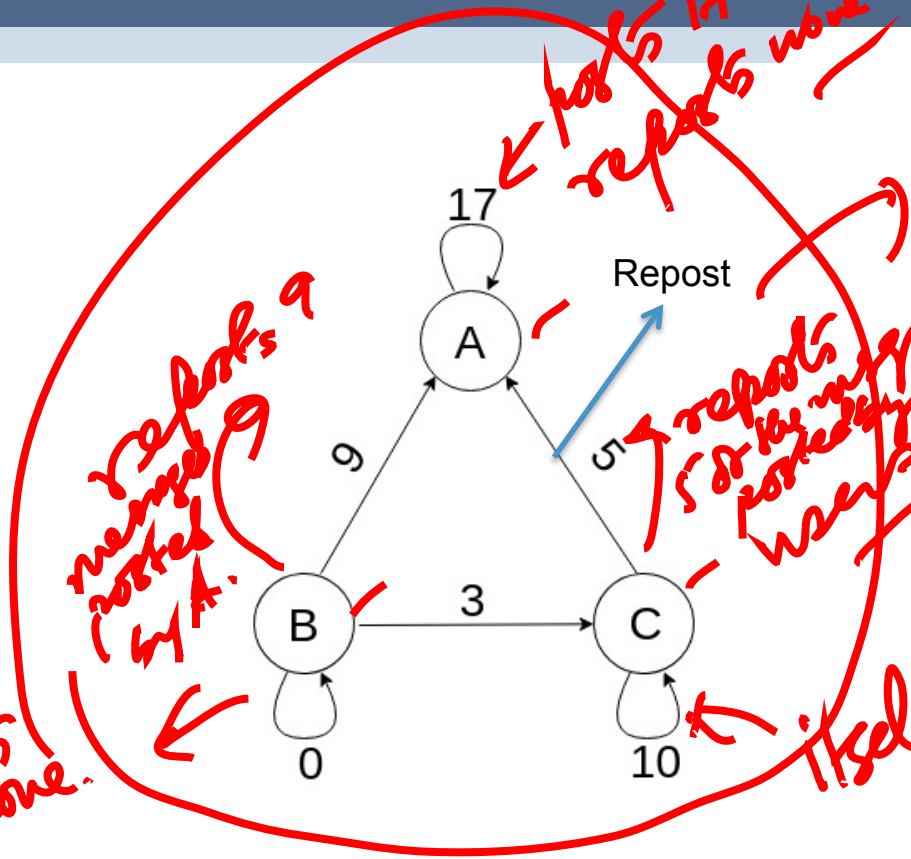
Seed set of hateful users

Lexicon-based filtering

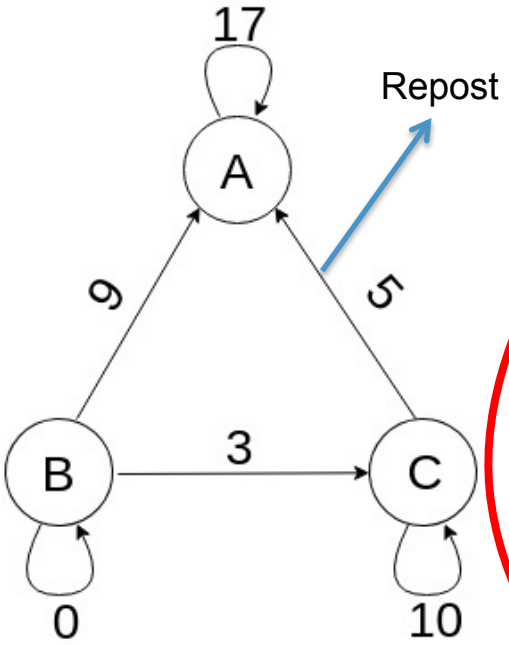
- Employed a lexicon of 45 high-precision keywords indicative of hate
  - Kike: racial slur against Jews
  - Beached whale: racial slur against fat people
  - Paki: racial slur against Muslims

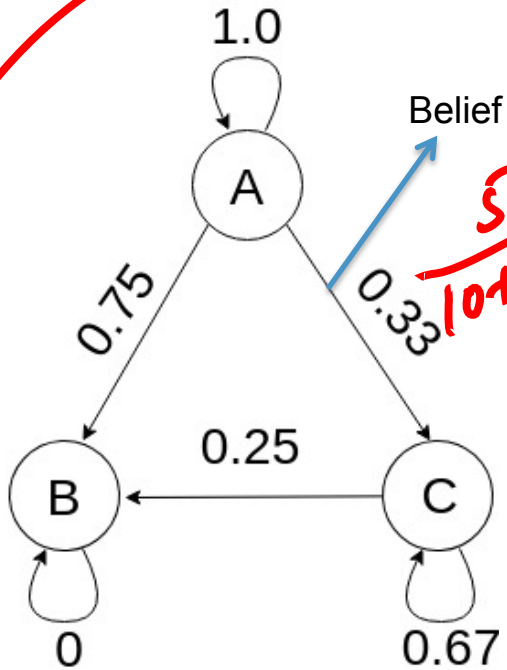
- Find users who have 10+ posts with one or more of these keywords (on Gab use of such words are almost surely indicative of hate content)

Repost network



Belief network

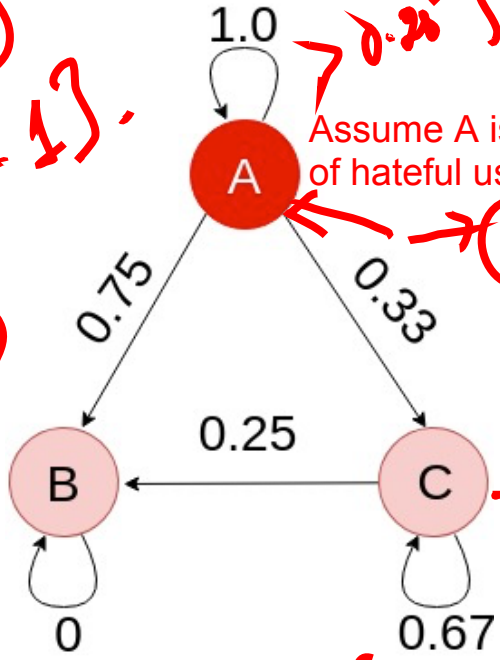




DeGroot's model

$$b_C^{i+1} = 0.33 \times b_A^i + 0.67 \times b_C^i$$

- Model progresses in an iterative fashion till convergence.
- At convergence: KH: nodes with scores: [0.75,1], NH: nodes with scores: [0.0, 0.25]



Assume A is a node from the seed set of hateful users –> score = 1

How good is the diffusion model?

	Annotator 1	Annotator 2	Kappa
Hate (KH)	86.9%	93.2%	0.70
Not Hate (NH)	92.2%	99.4%	0.87

How good is the diffusion model (more evidence)?

- Topics represent specific niche communities in GAB
- HT: Topics in which KH users pre-dominantly post
- NT: Topics in which NH users pre-dominantly post

HT	Jews Are The Synagogue Of Satan, The Black Race SUCKS, Street Shitter, Israel Holocaust Remembrance Day
NT	Xenoblade Chronicles 2(Spoilers), 2018 memes to amuse you, What's Going On?, Landscape, Classic Cars and Trucks,

How good is the diffusion model (more evidence)?

- Domains shared
- The most popular KH domains indulges in spreading conspiracy theories.[3]

User	Domains
KH	*dailystormer*, imageshack, *radioaryan*, endculturalmarxism, *christophercantwell, infostormer*, rationalwiki, skepdic
NH	xxxbios, bring-back-america, yourlawyer, Energy-Ingenuity, petreporters, internetmarketingexperience, strippersforyou

Many of these are extreme rights

[3] Vosoughi, S., Roy, D., and Aral, S. (2018a). The spread of true and false news online. Science, 359(6380):1146–1151.

Cascades: Tracing the influence path

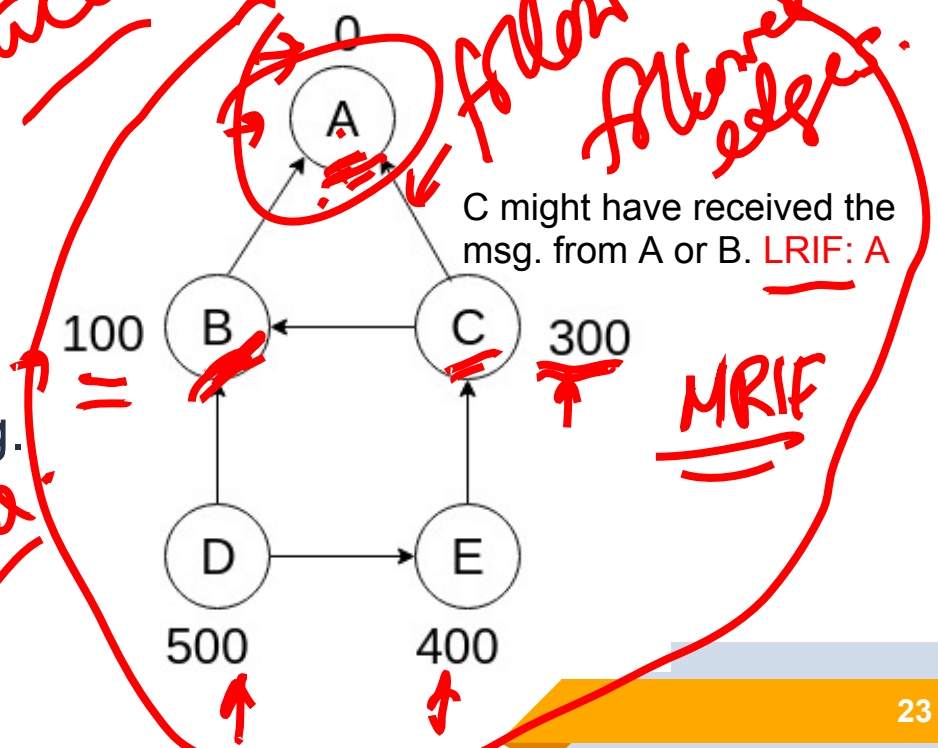
- Cascade - Path traced by a post as it is re-posted by users.
- Impossible to trace the exact influence path.
- Leverage the social connections between users.
- Employ the LRIF[5] model to create a DAG to trace the path

[5] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 65–74. ACM.

Tracing the influence path

Links indicate follower relationship between users.

Numbers indicate time of reposting.



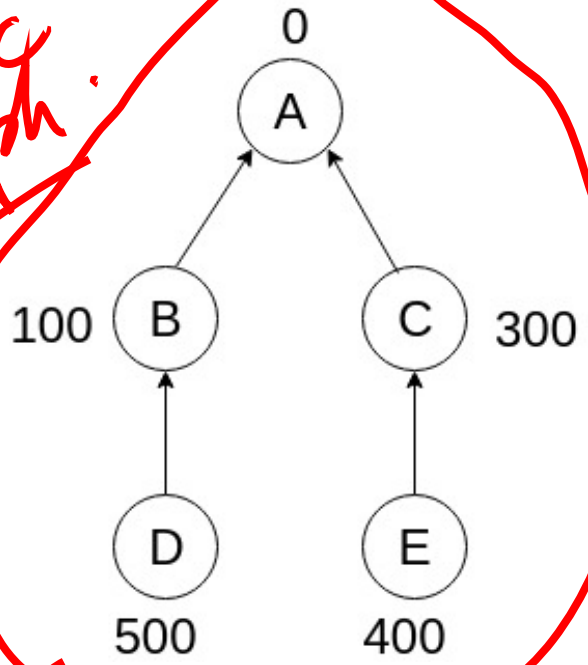
Imposing LRIF model

DAG under the LRIF assumpation.

Cascade is this DAG.

Cascades generated by posts of KH users

Cascades generated by posts of NH users



Cascade properties

- **Size** - number of unique users
- **Depth** - length of the largest path in the cascade

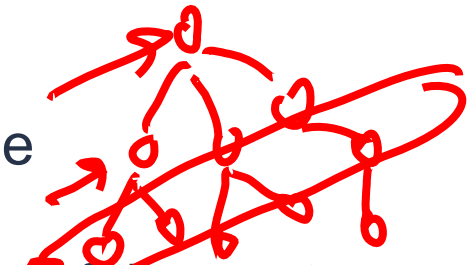
$$D = \max(d_i), 0 \leq i \leq n \quad d_i \text{ is the depth of node } i.$$

- **Average Depth** - average depth of each path from root node

$$AD = \frac{1}{n-1} \sum_{i=1}^{n} d_i$$

- **Breadth** - maximum width of the cascade
- **Structural virality** - indicates viral nature of the post

$$SV = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$



Experiments

Cascade properties observed for the KH and NH users w.r.t

- Original posts of the users.
- Posts containing attachments like media or images.
  - Compare the virality of visual vs textual information.
- Posts belonging to topics.
  - Observe the effect of community on virality.

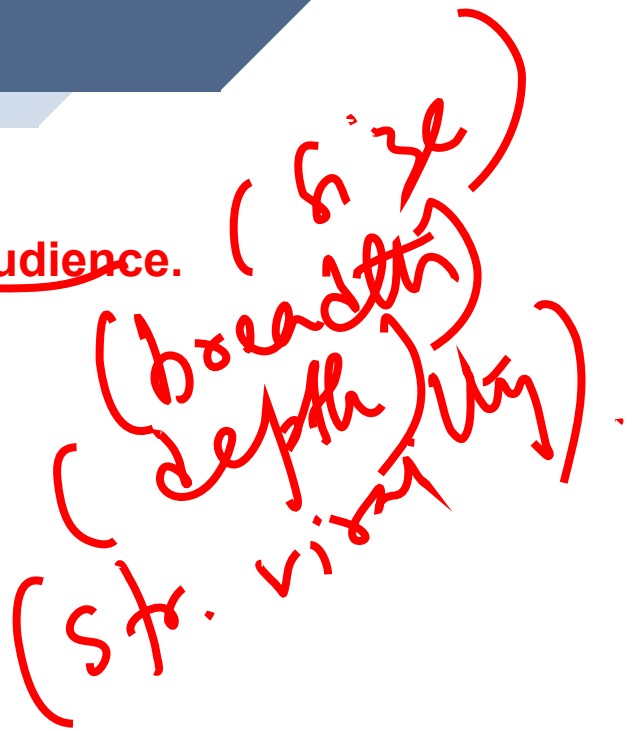
Analysis of posts

	Posts		Attachments		Topics	
Feature	**KH**	**NH**	KH	NH	KT	NT
Size	**1.28**	**1.21**	1.34	1.23	1.68	1.51
Depth	**0.13**	**0.09**	0.16	0.11	0.30	0.24
Breadth	**1.13**	**1.10**	1.15	1.11	1.30	1.24
AD	**0.11**	**0.08**	0.14	0.10	0.26	0.22
SV	**0.13**	**0.09**	0.16	0.11	0.31	0.25

- Differences in diffusion properties are **statistically significant (KS test, $p<0.01$).**

Analysis of posts

- Posts of hateful users have a **larger audience.**
- Posts of hateful users **spread wider.**
- Posts of hateful users **spread deeper.**
- Posts of hateful users are **more viral.**



Analysis of attachments

	Posts		**Attachments**		Topics	
Feature	KH	NH	**KH**	**NH**	KT	NT
Size	1.28	1.21	**1.34**	**1.23**	1.68	1.51
Depth	0.13	0.09	**0.16**	**0.11**	0.30	0.24
Breadth	1.13	1.10	**1.15**	**1.11**	1.30	1.24
AD	0.11	0.08	**0.14**	**0.10**	0.26	0.22
SV	0.13	0.09	**0.16**	**0.11**	0.31	0.25

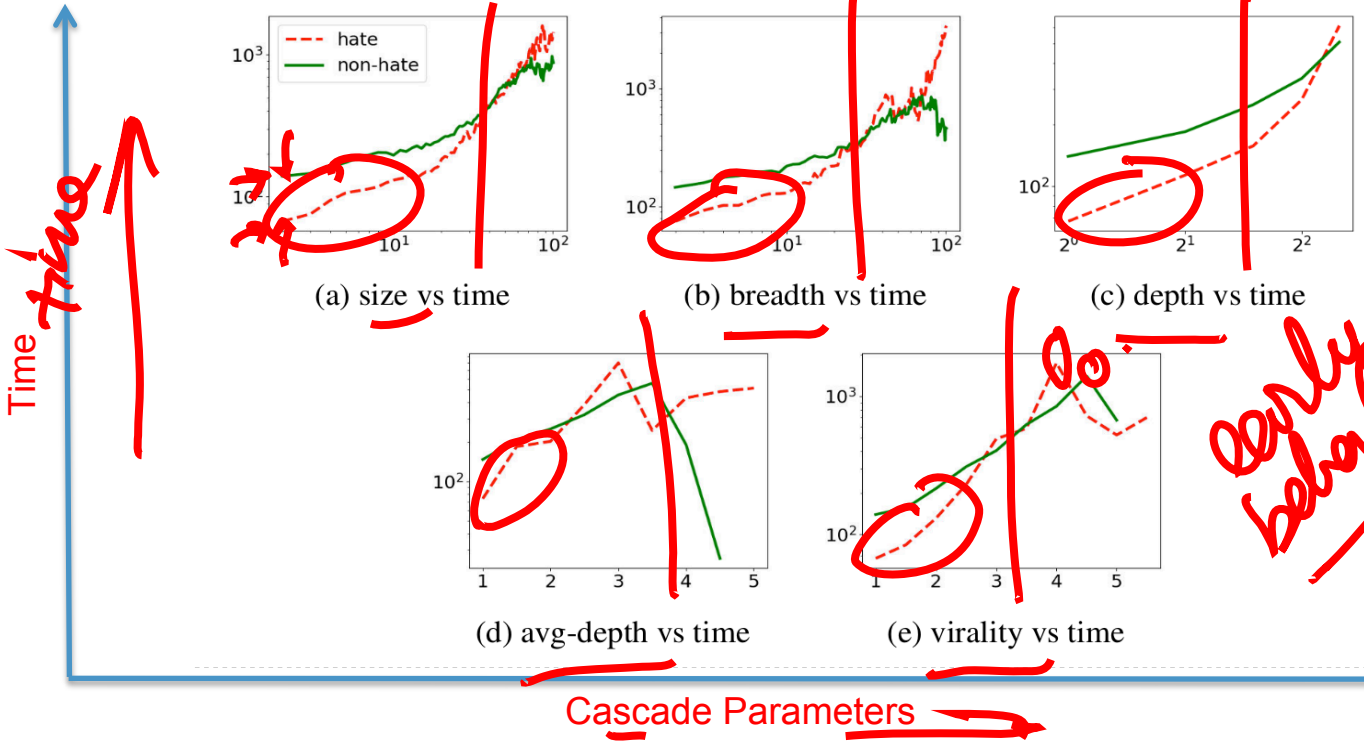
- Differences in diffusion are **more pronounced.**

Analysis of topics

	Posts		Attachments		**Topics**	
Feature	KH	NH	KH	NH	**KT**	**NT**
Size	1.28	1.21	1.34	1.23	**1.68**	**1.51**
Depth	0.13	0.09	0.16	0.11	**0.30**	**0.24**
Breadth	1.13	1.10	1.15	1.11	**1.30**	**1.24**
AD	0.11	0.08	0.14	0.10	**0.26**	**0.22**
SV	0.13	0.09	0.16	0.11	**0.31**	**0.25**

- Differences in diffusion are **more pronounced.**

Temporal evolution of cascade parameters



Design of online platforms

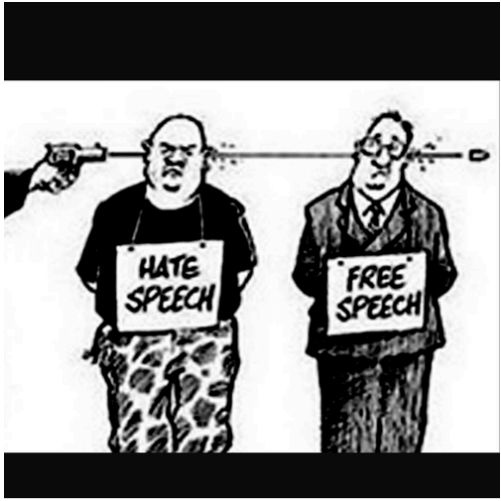
- Need methods to stop/slow the spread of hateful messages

Popular strategies to handle hate speech?

- Block/suspend the hateful message/account
- Several governments have established hate speech laws to prevent its spread
- Several social media sites have also tried to take strict action against hate speech

Popular strategies to handle hate speech?

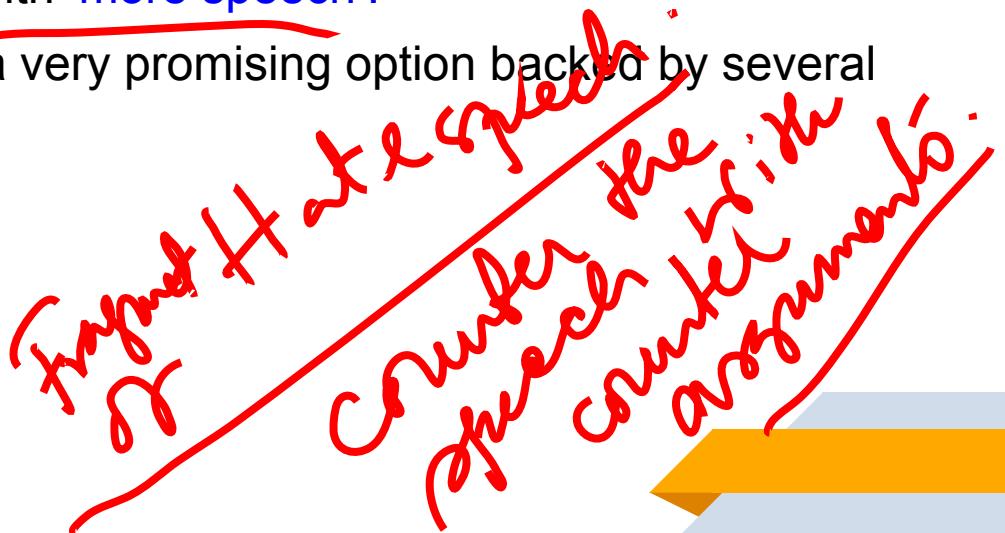
- Block/suspend the hateful message/account
- Several governments have established hate speech laws to prevent its spread
- Several social media sites have also tried to take strict action against hate speech
- However it could lead to free speech violation and doesn't stop the hate speaker



I may not always agree with what you say but I'll always support your right to say it

Possible alternative?

- Counter the hateful messages with ‘more speech’.
- Counterspeech is emerging as a very promising option backed by several organizations and NGOs.

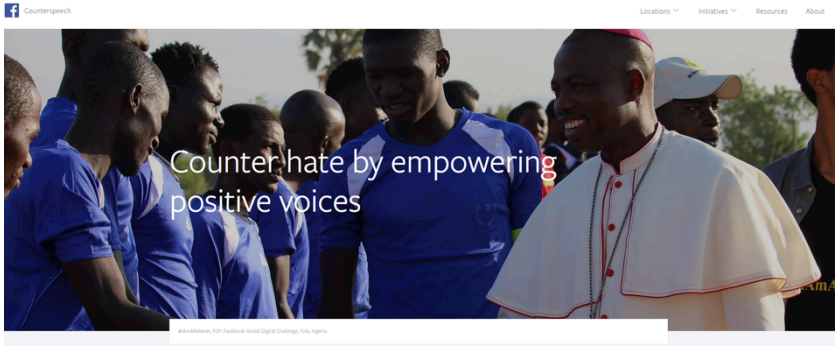


Thou shalt not hate: Countering online hate speech

ICWSM 2019

What is counterspeech?

- A common, crowd-sourced response to extremism or hateful content.
- Social platforms like Facebook have started counterspeech programs to tackle hate speech.
- Facebook has even publicly stated that it believes counterspeech is not only potentially more effective, but also more likely to succeed in the long run.



Our definition of counterspeech

- For creating the dataset, we focus on the **comments of hateful videos** posted on **YouTube**

- We define counterspeech as *a direct response/comment (not reply to a comment) that counters the hateful or harmful speech.*

- Taking the YouTube videos that contain hateful content toward three target communities: Jews, African-American (Blacks), and LGBT, we collect user comments to create a dataset which contains counterspeech.

Dataset collection

Scraped comments from 31 videos hateful videos on YouTube

Hamas cleric: "Destruction of Israel and Jews will bring blessing to world"

The American Preacher Spreading Hate

BBC Three

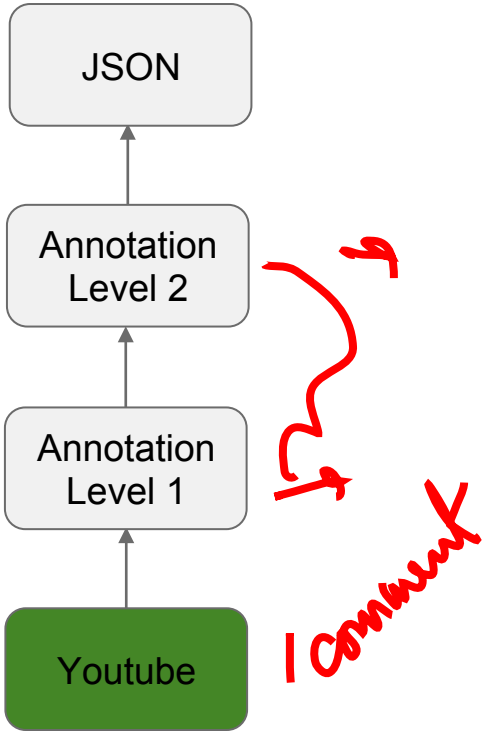
Subscribe 1.5M

62,232 views

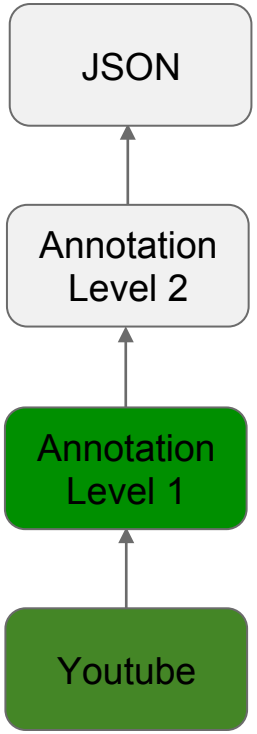
Types of counterspeech

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Warning of offline or online consequences
4. Affiliation
5. Denouncing hateful or dangerous speech
6. Humor and sarcasm
7. Positive tone
8. Hostile language

Data collection



Data collection



Each Comment

Counterspeech

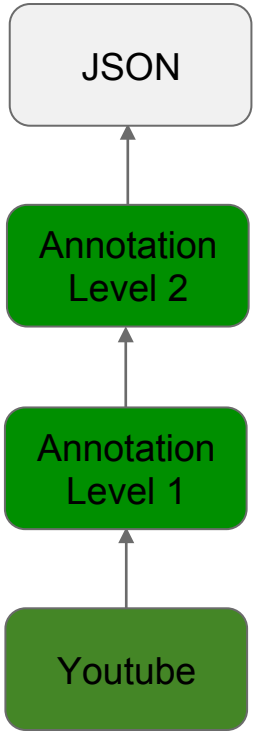
6898

Non-Counterspeech

7026

Kappa: 0.80

Data collection



Counterspeech comments

Presenting facts to correct misstatements or mis-perceptions

Pointing out hypocrisy or contradictions

Warning of offline or online consequences

Affiliation

Denouncing hateful or dangerous speech

Humor and sarcasm

Positive tone

Hostile language

Kappa: 0.87

6 898 counter speech

→ a particular counterspeech can have multiple labels

Dataset details

- **Hostile language** is the major category among all the classes and is present in around **39.74%** of the counterspeech.

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	526	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	618	1100
Positive tone	359	237	268	864
Hostile	712	946	1083	2741
Total	2582	2811	3726	9119

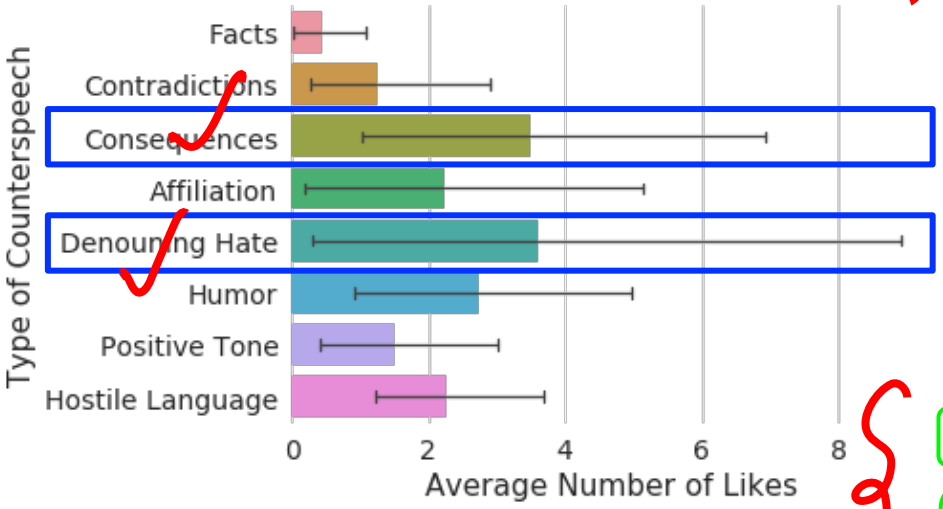
Dataset details

- Different communities attract different types of counterspeech.

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	526	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	618	1100
Positive tone	359	237	268	864
Hostile	312	946	1083	2741
Total	2582	2811	3726	9119

For different communities different types of counterspeech are more effective

African-American community



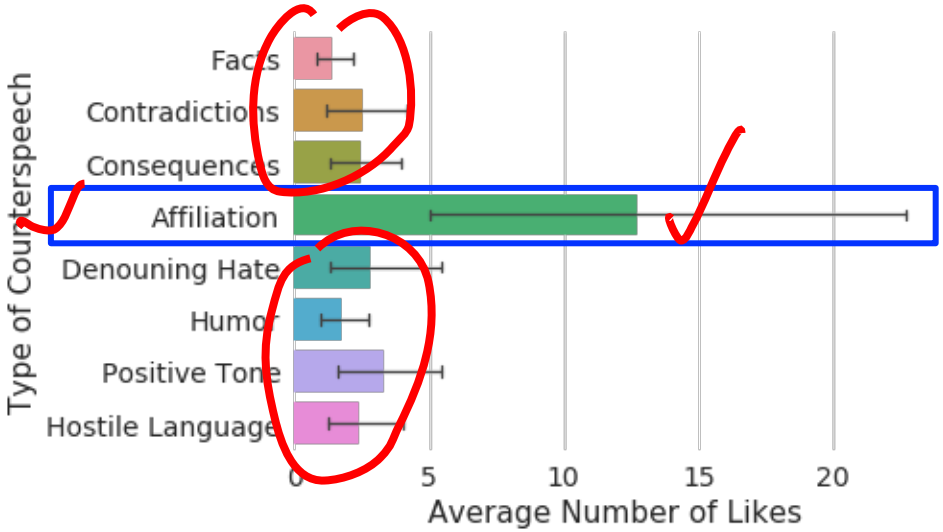
In case of the African-American community, the counterspeakers call out for racism and talk about consequences of their actions

Example:

"i hope these cops got fired! this is bullshit"

"Sad to see the mom teaching her children to be racist and hateful. The way the guy handled it was great."

Jewish community

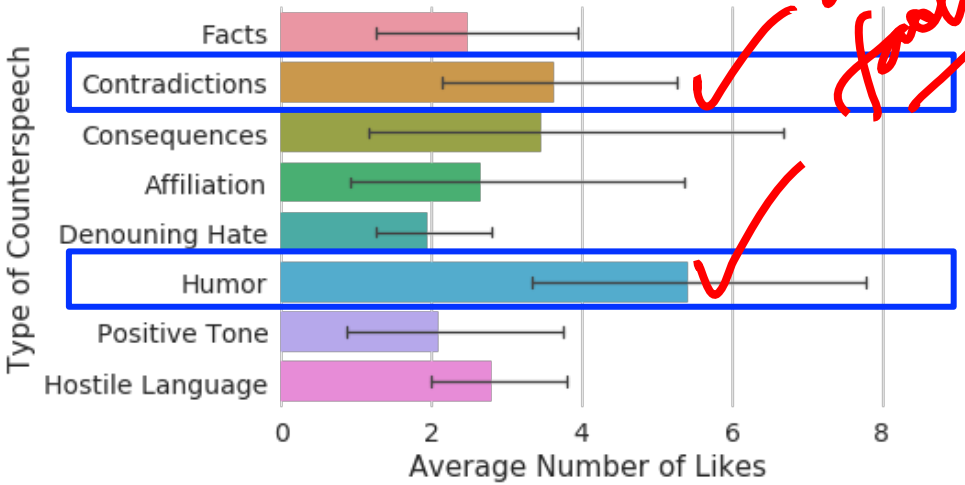


In case of the Jews community, we observe that the people affiliate with both the target and the source community (‘Muslims’, ‘Christians’) to counter the hate message.

Example:

“I'm Jewish And I'm really glad there some people that stand up for us And I have no problems with Muslims. We're all brothers and sisters”

LGBT community



In case of the LGBT community, the counterspeakers make use of sarcasm and provide several points which contradict the statements expressed by the hate speaker.

Example:

"Marriage was defined in a magical garden with a talking snake by an invisible man in the sky.. Makes perfect sense!"

"It's ironic to me that a man who is supposed to teach love and acceptance would hand down this type of advice"

Classification task

Binary classification: Counter Vs. Non-counter

Method	Precision	Recall	F1-Score	Accuracy
XGB+SV+TF-IDF+BOWV	0.716(+/-0.038)	0.715(+/-0.039)	0.715(+/-0.04)	0.716(+/-0.038)
MLP+SV+TF-IDF	0.714(+/-0.031)	0.713(+/-0.033)	0.713(+/-0.033)	0.714(+/-0.032)
CB+SV+TF-IDF+BOWV	0.708(+/-0.04)	0.706(+/-0.043)	0.705(+/-0.043)	0.707(+/-0.042)
RF+SV+TF-IDF+BOWV	0.697(+/-0.043)	0.693(+/-0.045)	0.692(+/-0.046)	0.695(+/-0.044)
SVC+SV+TF-IDF+BOWV	0.693(+/-0.029)	0.691(+/-0.03)	0.691(+/-0.03)	0.692(+/-0.029)

Multi-label classification: Types of counterspeech

Method	Accuracy	Precision	Recall	F1-Score	Hamming Loss
$General_B$	0.322	0.397	0.322	0.356	0.191
XGB+SV+TF-IDF-BOWV	0.472(+/-0.012)	0.509(+/-0.012)	0.733(+/-0.011)	0.601(+/-0.011)	0.212(+/-0.015)
MLP+SV+TF-IDF	0.44(+/-0.014)	0.504(+/-0.013)	0.527(+/-0.021)	0.515(+/-0.015)	0.295(+/-0.018)
LR+SV+TF-IDF	0.469(+/-0.014)	0.5(+/-0.014)	0.734(+/-0.02)	0.595(+/-0.015)	0.212(+/-0.015)
GNB+SV	0.339(+/-0.014)	0.357(+/-0.016)	0.71(+/-0.02)	0.475(+/-0.018)	0.072(+/-0.012)
DT+SV+TF-IDF	0.301(+/-0.015)	0.356(+/-0.02)	0.361(+/-0.02)	0.358(+/-0.019)	0.193(+/-0.012)

Can be used for automatic generation of counterspeech

Why I do this?

- Have been a victim myself
- Algorithm – I (2019); Numerical Feedback: 4.21/5
- But,

Why I do this?

- Have been a victim myself.
- Algorithm – I (2019); Numerical Feedback: 4.21/5
- But, while reading the subjective feedback
- “His legs are weak 😂, can beat his up easily actually lol.”