You too Brutus! Trapping Hateful Users in Social Media: Challenges, Solutions & Insights

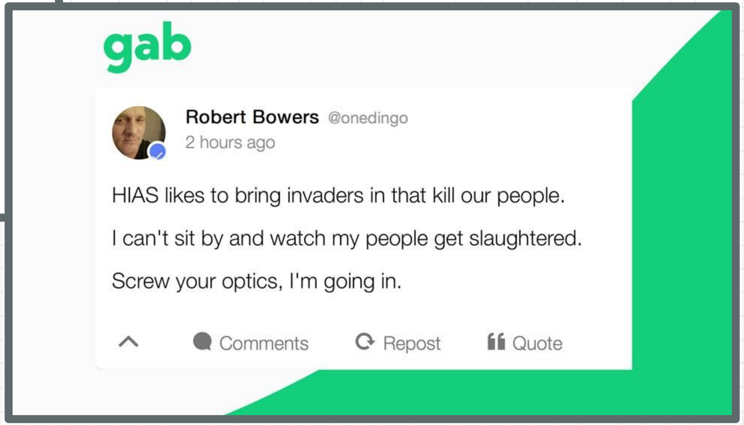
- **Mithun Das**, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, Binny Mathew

Warning!

⚠️ This presentation contains material that many will find offensive or hateful; however this cannot be avoided owing to the nature of the work.

Hate speech across platforms

Twitter



Gab

Effect of Hate speech?

- The public expression of hate speech promotes the devaluation of minority members[1]
- Frequent and repetitive exposure to hate speech could increase an individual’s outgroup prejudice[2]

[1] Jeff Greenberg and Tom Pyszczynski. 1985. The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease. Journal of Experimental Social Psychology 21, 1 (1985), 61–72.
[2] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. Aggressive behavior 44, 2 (2018), 136–146.

Real World Consequence

Bulandshahr Violence

Pittsburg Shooting

Christchurch Shooting

Rohingya Genocide

Sri Lanka riot

What Can be the Solution?

- Detecting Hateful posts
    - Individual posts can be automatically detected

Most of the work so far tried to detect hateful posts on social media.

Challenges of Post Level Detection

- If the context of a post is ambiguous, it is difficult to decide whether a post is hateful or not.
- Adversarial attack can fool the hate speech detection system.

What Can be other Solution?

- Detecting hateful posts
  - Individual posts can be automatically detected
- Detecting hateful users
  - The users who engage in spreading hateful content

Advantages to Detect Hateful Users

- Majority of the hateful posts are generated by a few hateful users.
- Hateful users are densely connected.

Table of Contents

01

Data collection

How we got the data from Gab and Twitter?

Datasets

- GAB
    - Gab is a social media platform which promotes itself as a "Champion of free speech”
- Twitter
    - More mainstream social media platform with relatively stricter moderation policies

Gab Data Collection

- We use the existing crawled Gab dataset by Mathew et al[3].
- The dataset has been crawled using Gab's API and standard snowball technique.
- The dataset contain **381K users** and their followership network.

[3] Mathew, Binny & Dutt, Ritam & Goyal, Pawan & Mukherjee, Animesh. (2019). Spread of Hate Speech in Online Social Media. 173-182. 10.1145/3292522.3326034.

Gab Data Sampling

- To ensure sufcient representation of hateful and non-hateful users a **lexicon** of **45 high-precision hate terms** are used (like 'kike', 'ni*ger') to identify hateful posts[3].

Gab Data Sampling

- To ensure sufcient representation of hateful and non-hateful users a **lexicon** of **45 high-precision hate terms** are used (like 'kike', 'ni*ger') to identify hateful posts[3].
- Initial seed-set of **2,769 hateful users** was created considering the users who have made **at least 10 such posts**

Gab Data Sampling

- To ensure sufcient representation of hateful and non-hateful users a **lexicon** of **45 high-precision hate terms** are used (like ‘kike’, ‘ni*ger’) to identify hateful posts[3].
- Initial seed-set of **2,769 hateful users** was created considering the users who have made **at least 10 such posts**
- Then a **repost network** was created where nodes represent users and edge-weights denote the **reposting frequency**.

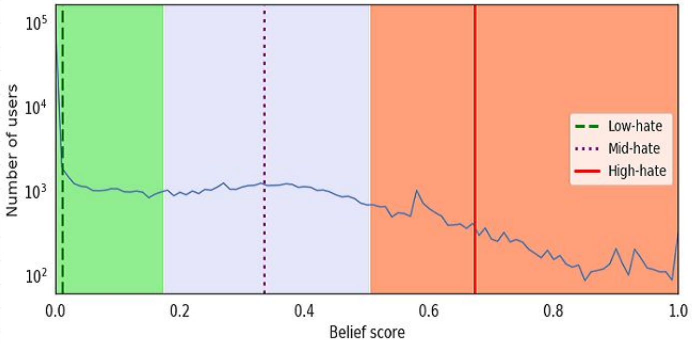
Gab Data Sampling

- To ensure sufcient representation of hateful and non-hateful users a **lexicon** of **45 high-precision hate terms** are used (like 'kike', 'ni*ger') to identify hateful posts[3].
- Initial seed-set of **2,769 hateful users** was created considering the users who have made **at least 10 such posts**
- Then a **repost network** was created where nodes represent users and edge-weights denote the **reposting frequency**.
- Using **DeGroot's model** a **belief score** has been **assigned** to each user.

User Selection for Annotation

- **Users** are then **clustered** on the basis of this score using k-means algorithm into three tiers – "**high**", "**medium**" and "**low**" .

- We randomly sample 300 users from each of these three tiers with the additional constraint that the user must have posted at least 10 times.



Twitter Data Collection & Sampling

- We use the existing crawled and annotated Twitter dataset by Ribeiro et al[4].
- The dataset has been crawled using Twitter API.
- The data sampling process is similar to the method we used for Gab.
- Unlike Gab, instead of using followeship network, retweet network has been used.

[4] Ribeiro, Manoel & Calais, Pedro & dos Santos, Yuri & Almeida, Virgilio & Meira Jr, Wagner. (2018). Characterizing and Detecting Hateful Users on Twitter.

02

Annotation

How we annotated the data?

Annotation Guidelines for Gab

- A user is defined has hateful, if

  The user endorses content that is humiliating, attacking or insulting, some groups or individuals based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease[5].

[5] ElSherief, Mai & Kulkarni, Vivek & Nguyen, Dana & Wang, William & Belding, Elizabeth. (2018). Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media.

Annotating the Gab data

- Using the annotation guidelines, two experts annotated the 900 users selected based on the data sampling discussed earlier.
- Dubious cases which arose as a result of conflict were dropped.
- This yields a final count of **423 hateful** and **375 non-hateful** users and constitutes our set of a total of **798 labelled** instance.

Followership Network Creation (Gab)

- Constructed 1.5-degree network of these labeled users that consists of their immediate followers, followings and connections among themselves.

Followership Network Creation (Gab)

- Constructed 1.5-degree network of these labeled users that consists of their immediate followers, followings and connections among themselves.
- The nodes in the network represent the user accounts and the edges represent follower-following relationship.

Followership Network Creation (Gab)

- Constructed 1.5-degree network of these labeled users that consists of their immediate followers, followings and connections among themselves.
- The nodes in the network represent the user accounts and the edges represent following relationship.
- The graph is further pruned by removing users with less than 10 posts.

Followership Network Creation (Gab)

- Constructed 1.5-degree network of these labeled users that consists of their immediate followers, followings and connections among themselves.
- The nodes in the network represent the user accounts and the edges represent following relationship.
- The graph is further pruned by removing users with less than 10 posts.
- Filtered graph has **47K** users and **13.8M edges** and this constitutes the final network.

Final Dataset

	Gab	Twitter
No. of hateful users	423	544
No. of non hateful users	375	4427
Total users in the network	47K	100K
Edges in the network	13.8M	2.28M

03

Detection

Automatic detection of hateful users

Detection Methods

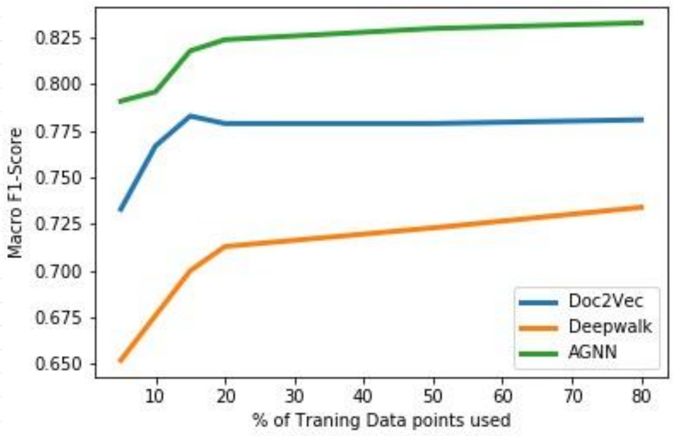
- Text based models
  - (fastText+LR), (Glove+LR), LSTM, (Doc2vec+LR), BERT, TSVM
- Network based models
  - Deepwalk, Node2vec.
- Graph neural network based models
  - GraphSAGE, GCN, AGNN, ARMA, ChebNet, GAT

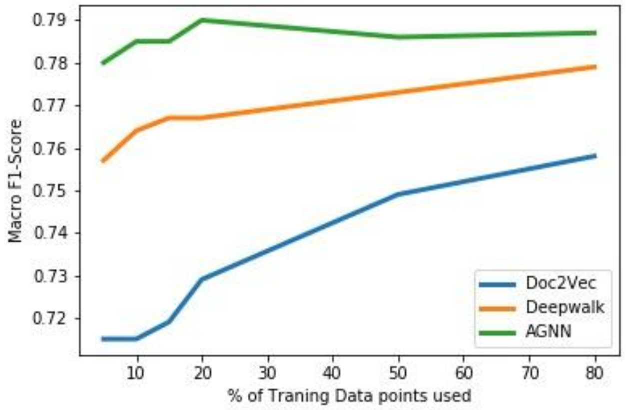
Results

Method	Inputs	Gab 5%	Gab 10%	Gab 15%	Gab 20%	Gab 50%	Gab 80%	Twitter 5%	Twitter 10%	Twitter 15%	Twitter 20%	Twitter 50%	Twitter 80%
fastText	Y, X_L	0.492	0.537	0.571	0.603	0.690	0.709	0.624	0.634	0.648	0.651	0.670	0.676
Glove	Y, X_L	0.695	0.720	0.745	0.750	0.778	0.784	0.650	0.666	0.674	0.681	0.691	0.695
LSTM	Y, X_L	0.579	0.600	0.605	0.608	0.622	0.645	0.514	0.487	0.567	0.564	0.592	0.608
Doc2vec	Y, X_L	0.733	0.767	0.783	0.779	0.779	0.781	0.715	0.715	0.719	0.729	0.749	0.758
BERT	Y, X_L	0.631	0.660	0.682	0.701	0.740	0.764	0.603	0.665	0.690	0.709	0.729	0.740
TSVM	Y, X	0.686	0.704	0.712	0.712	0.739	0.753	0.480	0.520	0.533	0.533	0.585	0.611
DeepWalk	Y, G	0.652	0.676	0.700	0.713	0.723	0.734	0.757	0.764	0.767	0.767	0.773	0.779
Node2vec	Y, G	0.647	0.672	0.695	0.704	0.725	0.744	0.692	0.720	0.732	0.734	0.749	0.748
GraphSAGE	Y, X, G	0.778	**0.808**	0.806	0.811	0.827	0.828	0.762	0.773	0.774	0.780	0.782	0.777
GCN	Y, X, G	0.721	0.735	0.730	0.738	0.751	0.758	0.756	0.759	0.767	0.773	0.776	0.770
AGNN	Y, X, G	**0.791**	0.796	**0.818**	**0.824**	**0.830**	**0.833**	**0.780**	**0.785**	**0.785**	**0.790**	0.786	**0.787**
ARMA	Y, X, G	0.765	0.778	0.783	0.797	0.809	0.805	0.757	0.760	0.761	0.762	0.770	0.769
ChebNet	Y, X, G	0.778	0.802	0.796	0.798	0.805	0.812	0.746	0.750	0.754	0.762	0.761	0.766
GAT	Y, X, G	0.683	0.718	0.725	0.726	0.745	0.758	0.757	0.774	0.781	0.777	**0.787**	0.782

GNNs which combine both textual and network features exhibit an improved performance over the individual text based classifiers and the network embeddings

Results





Gab

Twitter

AGNN which combine both textual and network features exhibit an improved performance over the individual text based classifiers and the network embeddings

Cross Platform Evaluation

Methods	Train	Test	F1	F1(H)	P(H)	R(H)
AGNN	**Twitter**	**Gab**	0.75	**0.81**	0.71	**0.94**
Doc2Vec			**0.77**	0.79	**0.76**	0.77
AGNN	**Gab**	**Twitter**	**0.74**	**0.54**	**0.58**	**0.50**
Doc2Vec			0.58	0.31	0.22	0.45

Observations and Insights

- AGNN is able to make correct predictions as the user (to be classified) has several hateful neighbors in its vicinity
- GNN based classification is less beneficial while detecting isolated hateful nodes

04

Post-Facto Analysis

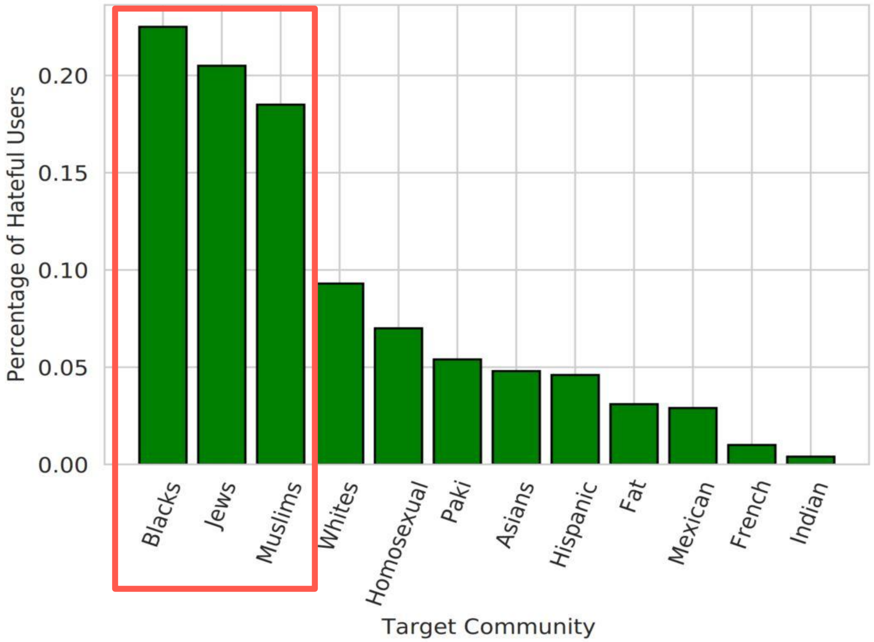
Target of hateful users

Post-Facto Analysis on Gab

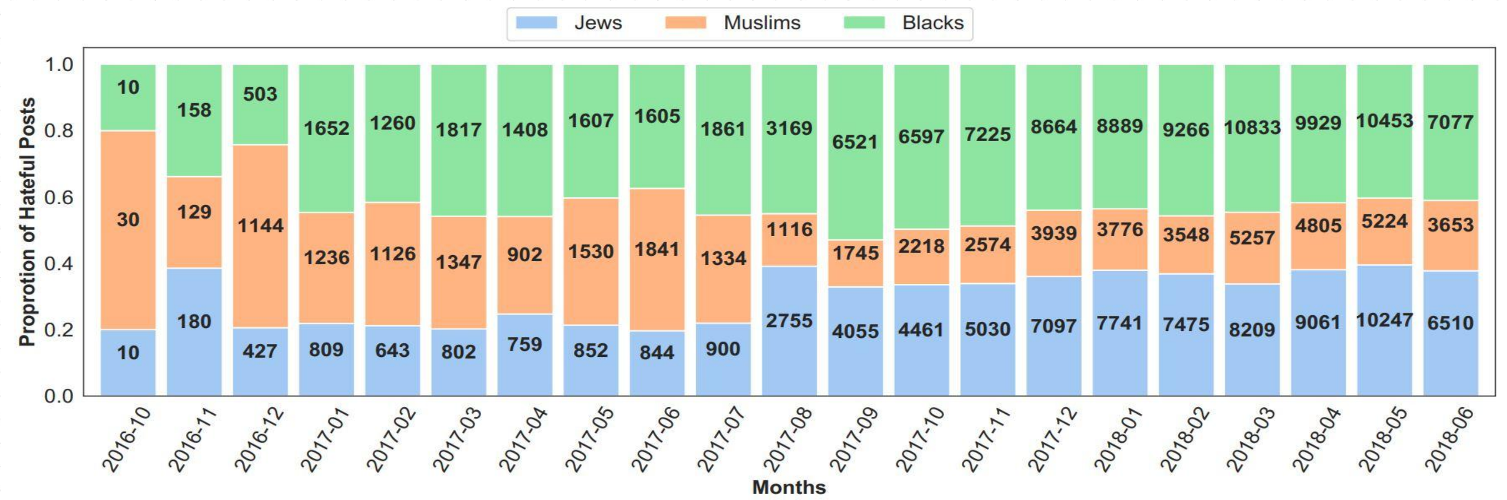
- Reasons for choosing the Gab dataset for this analysis are
  - availability of the full longitudinal data
  - loose moderation policies of the platform that enables the use of high precision keywords for obtaining reasonable results, which is not true for Twitter
- We divide our entire dataset into 21 snapshots ranging from October 2016 to June 2018.
- We take the best-performing AGNN model trained on the entire Gab data and use it to label the users present in each snapshot as hateful or not
- Using high-precision lexicon we find the target of the hateful users.

Overall Target Distributions

'Blacks', 'Jews' and 'Muslims' are the most prominent targets on Gab.

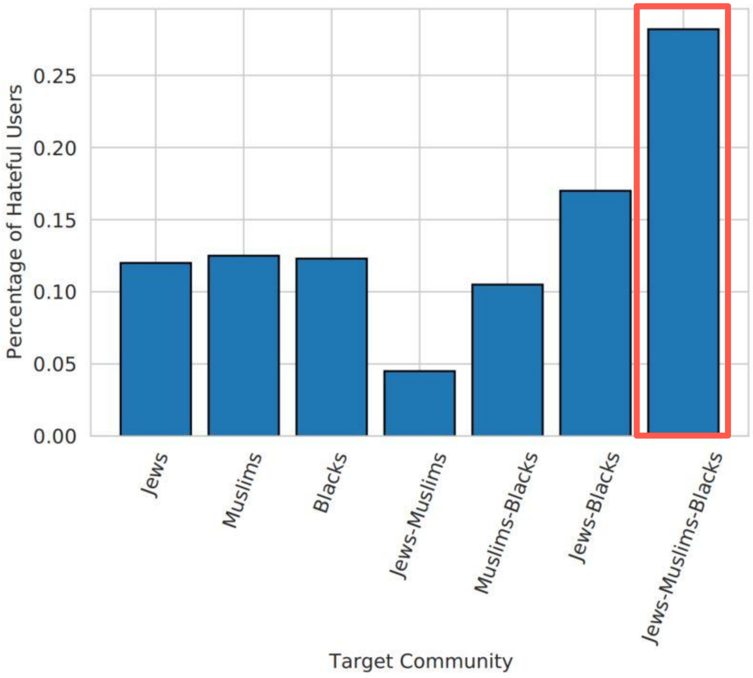


Rise and Rise of Hatred



- Rise in the gross number of posts over time.
- Since August' 17, 'Jews' and 'Blacks' become slightly more prominent targets

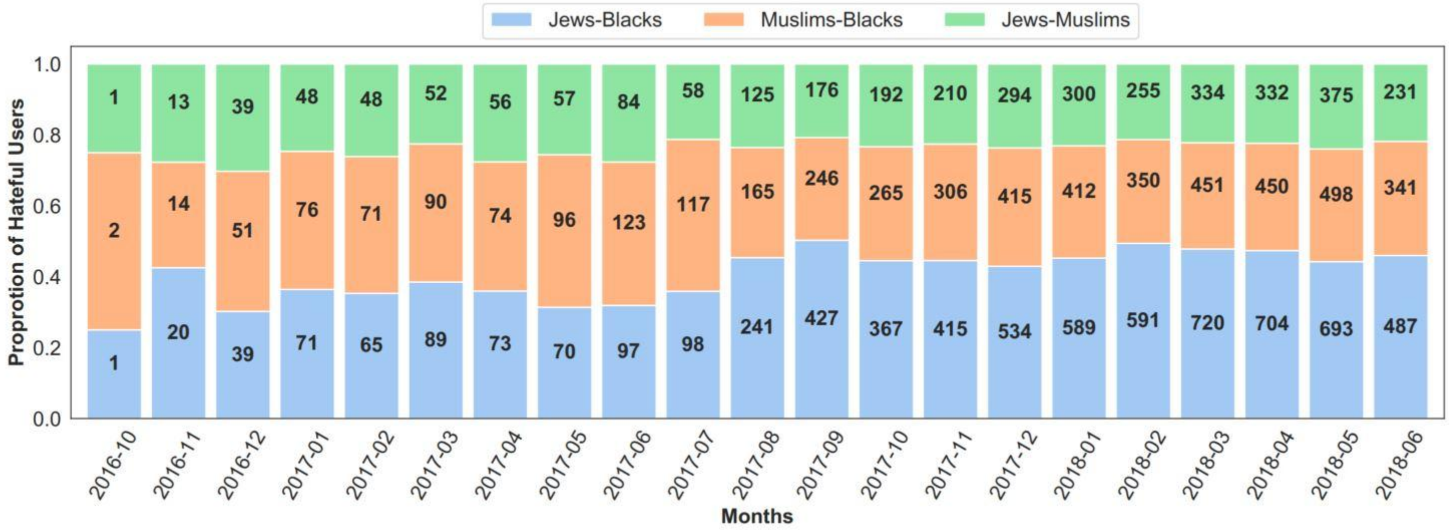
Multiple Targets



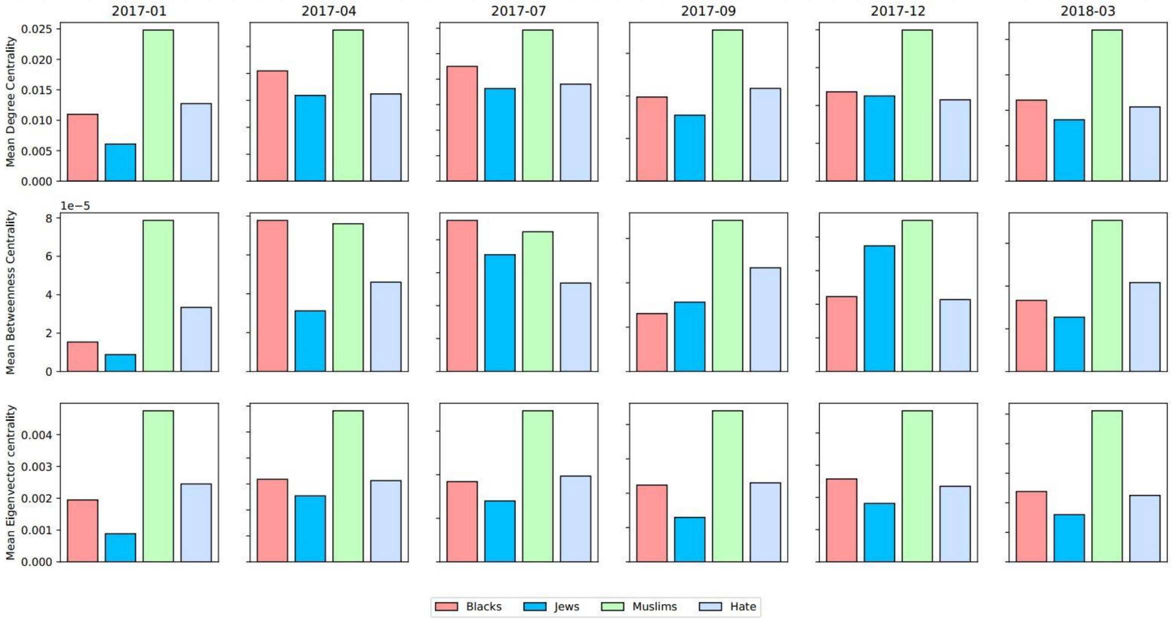
- Multi-target users are **more** in Gab. (Categories are mutually exclusive)

Multiple Targets Over Time

'Jews-Blacks' are the **most targeted** communities, followed by 'Muslims-Blacks' and 'Jews-Muslims'.



Centrality values of hateful users



Most central positions in the overall follower-followee network are occupied by the 'Muslim' targetting hateful users. Lot of them → inciting fear against Muslims

Trending Hashtags

Months	Trending Hashtags
Dec 2016	#BanIslam, #FakeNews, #StopWhiteGenocide, #WhiteGenocide, #MerryChrist-mas, #Israel, #Islam, #FreeSpeech
Jul 2017	#CNNBlackmail, #TheGoyimKnow, #JewBusiness, #ShoahBusiness, #Stop-WhiteGenocide, #Jesus, #DefendEurope, #CNN, #AmericaFirst
Jun 2018	#Islam, #Gab, #Muslim, #SpeakFreely, #PresidentTrump, #FreeTommyRobinson, #Potus, #Terrorism

Takeaways

- A **dataset** of 423 hateful users and 375 non-hateful users from the social media platform **Gab**.
- Textual and network features together can improve the performance of hateful users detection.
- Cross-platform results show AGNN model is generalizable.
- ‘**Blacks**’, ‘**Jews**’ and ‘**Muslims**’ are the most prominent target in Gab. Most of the hateful users target multiple target communities.

Dataset and Code:
https://github.com/hate-alert/Hateful-users-detection

Thank You!

Send your questions at mithundas@iitkgp.ac.in

Find more about us here !
https://hate-alert.github.io/