Evaluating the Effectiveness of **Deplatforming** as a Moderation Strategy on Twitter

Shagun Jhaver
Christian Boylston
Diyi Yang
Amy Bruckman

RUTGERS
Georgia Tech School of Interactive Computing

Background

Methods

Findings

Implications

Background

Methods

Findings

Implications

Background

Methods

Findings

Implications

Background

Methods

Findings

Implications

Background

What is Deplatforming?

- Permanent ban of controversial public figures with large followings on social media sites.


Explore
Settings
Profile
@realDonaldTrump
Account suspended
Twitter suspends accounts which violate the Twitter Rules


Alex Jones


Milo Yiannopoulus


Owen Benjamin

Research Questions

RQ 1: How does deplatforming affect the number of conversations about banned influencers?

RQ 2: How does deplatforming affect the spread of offensive ideas held by banned influencers?

RQ 3: How does deplatforming affect the activity and toxicity levels of supporters of these banned influencers?

Methods

Data

Examined observational data from Twitter through a temporal analysis of:

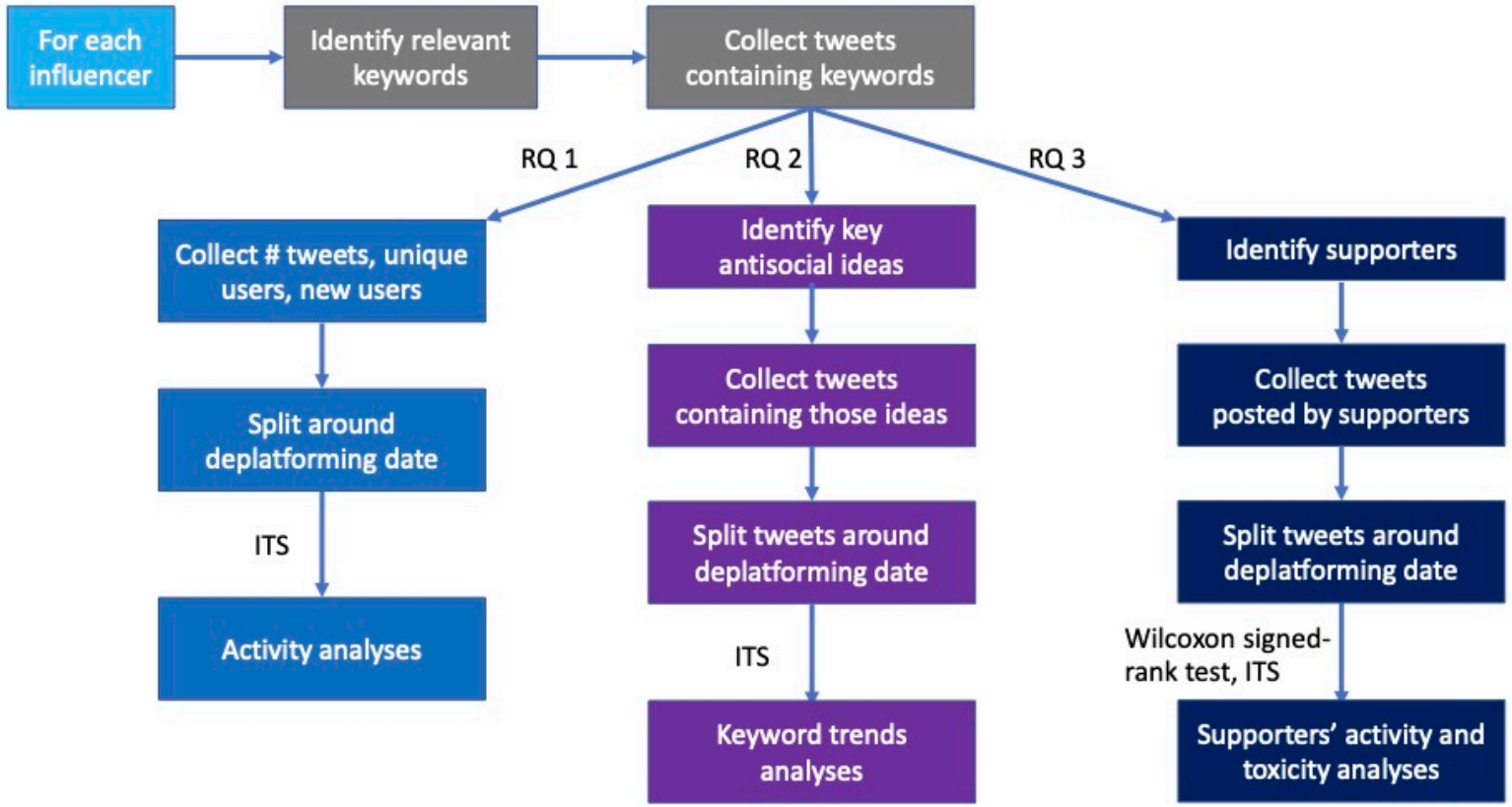
1. Tweets directly referencing deplatformed influencers,
2. Tweets referencing their offensive ideas, and
3. All tweets posted by their supporters.

Data

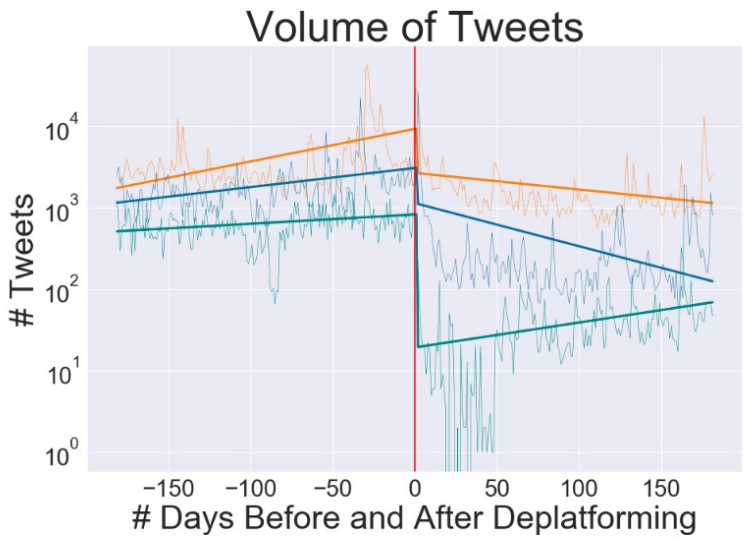
Influencer	# Followers	Deplatforming Date	# Tweets	# Supporters	# Supporters Tweets
Alex Jones	898,610	2018-09-06	1,157,713	2,935	17,050,653
Milo Yiannopoulos	338,000	2016-07-19	442,655	5,827	30,000,335
Owen Benjamin	122,634	2018-04-05	127,855	304	822,022

49 M tweets

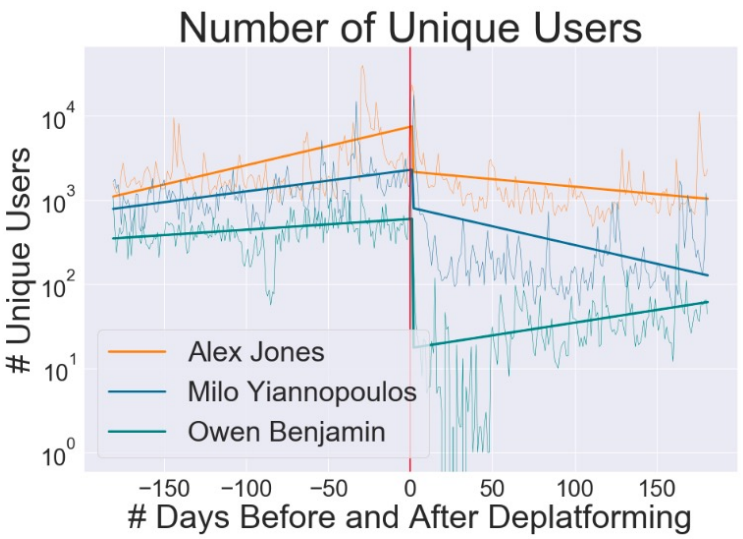
Framework


For each influencer
Identify relevant keywords
Collect tweets containing keywords
RQ 1
RQ 2
RQ 3
Collect # tweets, unique users, new users
Split around deplatforming date
ITS
Activity analyses
Identify key antisocial ideas
Collect tweets containing those ideas
Split tweets around deplatforming date
ITS
Keyword trends analyses
Identify supporters
Collect tweets posted by supporters
Split tweets around deplatforming date
Wilcoxon signed-rank test, ITS
Supporters' activity and toxicity analyses

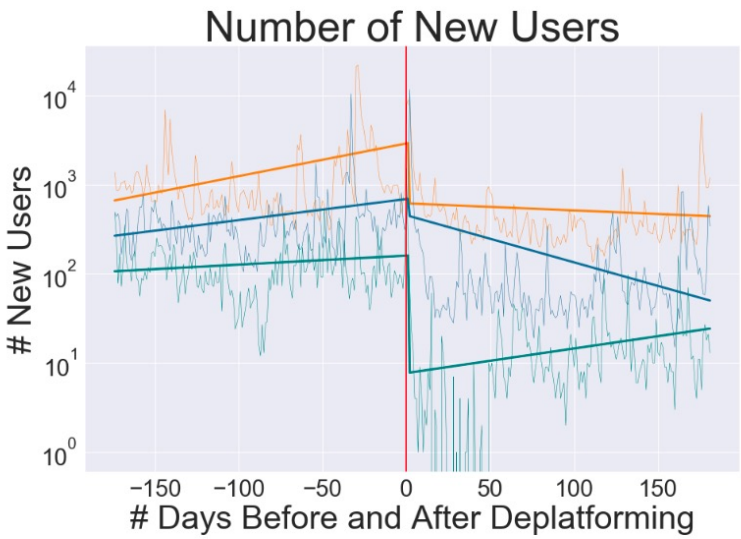
Findings



(a) Posting Activity Levels

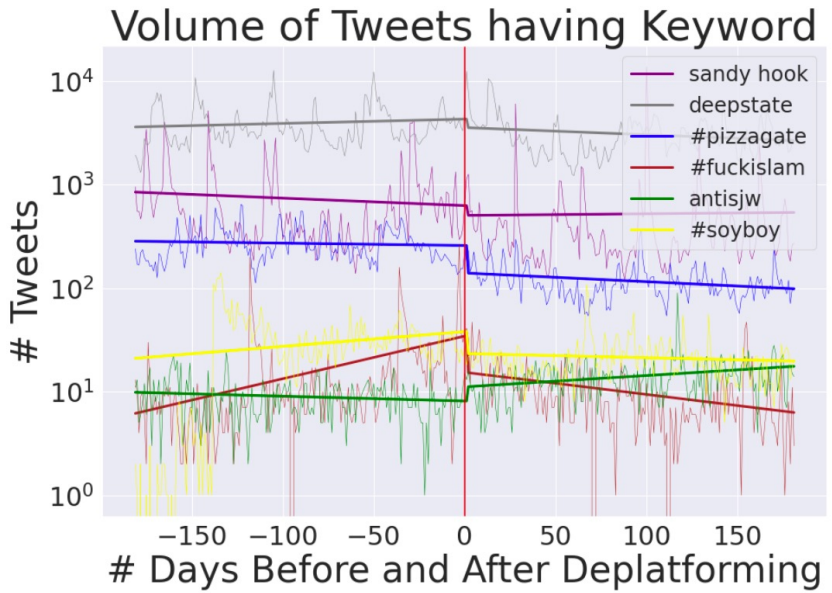


(b) Number of Unique Users

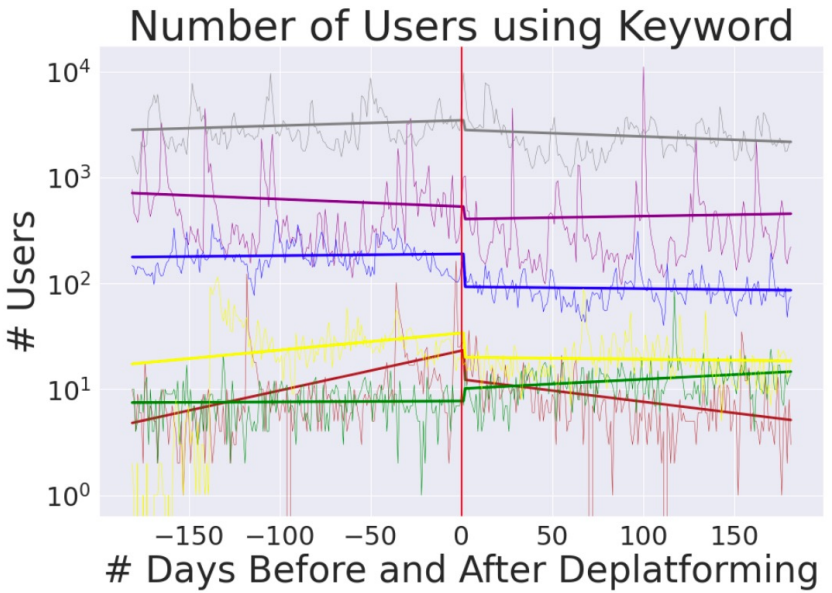


(c) Number of New Users

- Posts declined significantly, by 91.77% on average
- # Unique users diminished significantly, by 89.51%
- # New users declined significantly, by 89.65%

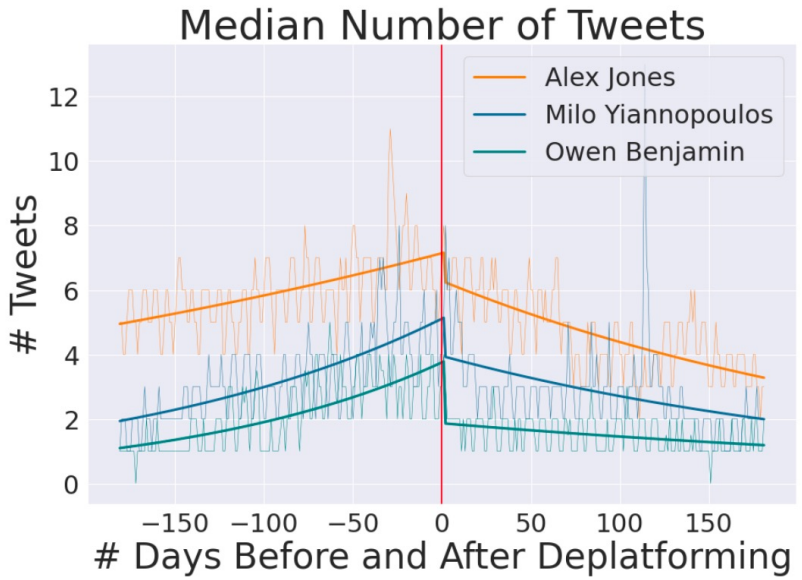


(a) Posting Activity Levels

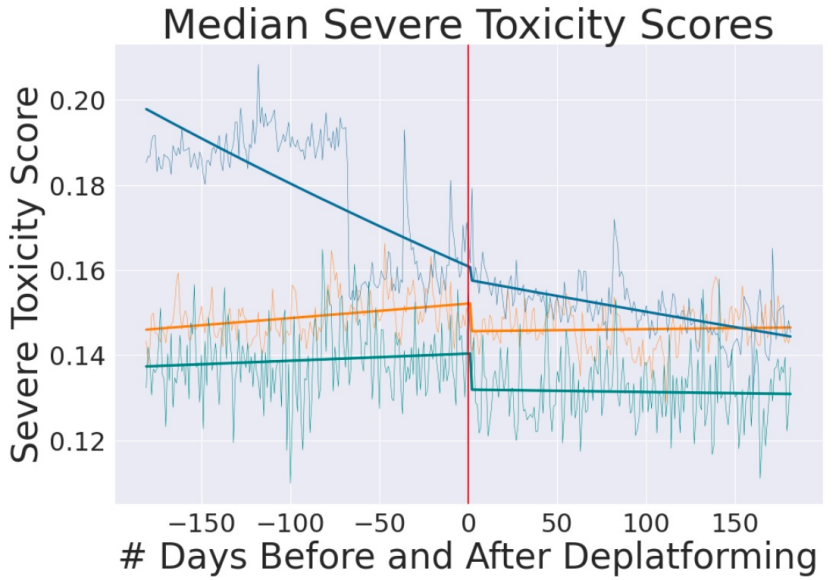


(b) Number of Unique Users

- Deplatforming helped significantly reduce the spread of many offensive ideas and conspiracy theories



(a) Median Posting Activity Levels



(b) Median Severe Toxicity Levels

- Deplatforming significantly reduced the overall posting activity levels of supporters for each influencer
- Median drop in supporters' tweets averaged 12.59%
- Median decline in supporters' toxicity averaged 5.89%

Implications

Deplatforming as a Moderation Tool

Effectively Reduces Offensive Influencers' Impact and Lessens Toxic Rhetoric

Platforms Must Defend Against Second-Order Harms of Deplatforming

- Deplatforming increased the prevalence of some offensive ideas
- A small group of supporters significantly increased both their activity and toxicity levels.
- Regulating in the aftermath is necessary
- Our approach can assist

Deplatforming versus Losing Advertising Revenue

- Deplatforming influencers reduced the posting activity levels of hundreds of their supporters.
- Financial benefits from advertising dollars tied to allowing toxic content
- Allowing toxic speech degrades vulnerable groups

Acknowledge

- Colleagues (Aaron Jiang, Amanda Baughan, Amy Zhang) and Reviewers
- Facebook Oversight Board Research Award

What is Deplatforming?

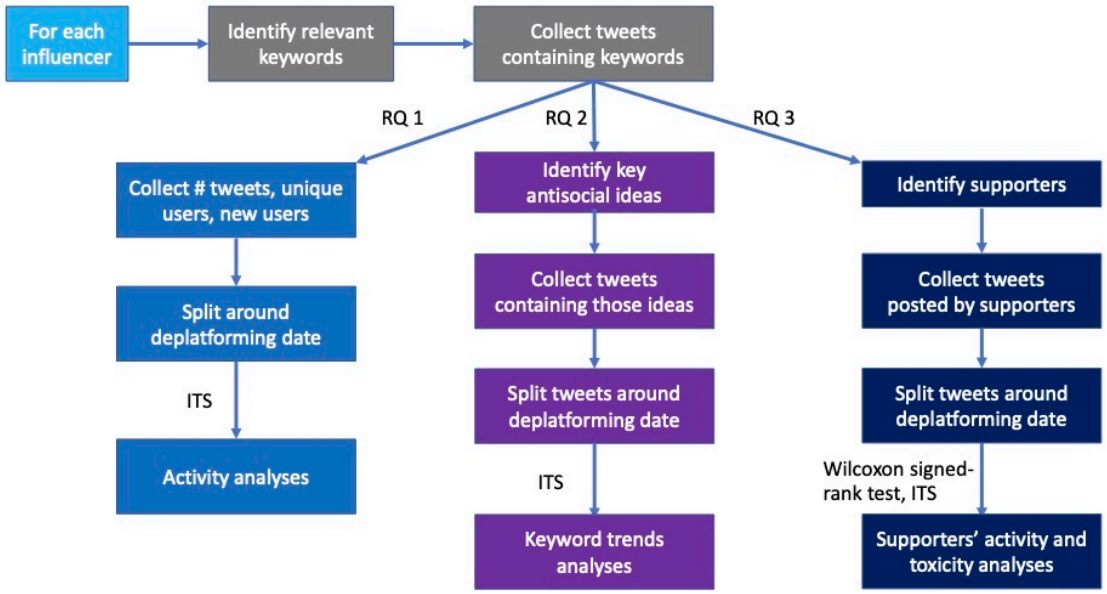
- Permanent ban of controversial public figures with large followings on social media sites.
- Once deplatformed, influencers are barred from making another account using their real names

Influencers we used as case studies

Research Questions

RQ 1: How does deplatforming affect the number of conversations about deplatformed influencers?
RQ 2: How does deplatforming affect the spread of offensive ideas held by deplatformed influencers?
RQ 3: How does deplatforming affect the overall activities of supporters of these deplatformed influencers?

Methodological Framework



Findings and Implications

Deplatforming Effectively Reduces Offensive Influencers' Impact and Lessens Toxic Rhetoric:

1. Posts referencing influencers reduce
2. Posts discussing ideas popularized by influencers reduce
3. Supporters become less active overall
4. Supporters become less toxic overall

Platforms Must Defend Against Second-Order Harms:

1. Deplatforming increases prevalence of some offensive ideas
2. Some supporters increased activity and toxicity levels

Email: shagun.jhaver@rutgers.edu