Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter

Shagun Jhaver
Christian Boylston
Diyi Yang
Amy Bruckman
Background

Methods

Findings

Implications
<table>
<thead>
<tr>
<th>Background</th>
<th>Methods</th>
<th>Findings</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>Methods</td>
<td>Findings</td>
<td>Implications</td>
</tr>
</tbody>
</table>
Background

Methods

Findings

Implications
Background
What is Deplatforming?

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

Twitter suspends accounts which violate the Twitter Rules
<table>
<thead>
<tr>
<th>Background</th>
<th>Methods</th>
<th>Findings</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex Jones</td>
<td>Milo Yiannopoulos</td>
<td>Owen Benjamin</td>
<td></td>
</tr>
</tbody>
</table>
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

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

49 M tweets
Framework

For each influencer
- Identify relevant keywords
- Collect tweets containing keywords

RQ 1
- Collect # tweets, unique users, new users
- Split around deplatforming date
- Activity analyses

RQ 2
- Identify key antisocial ideas
- Collect tweets containing those ideas
- Split tweets around deplatforming date
- ITS
- Keyword trends analyses

RQ 3
- Identify supporters
- Collect tweets posted by supporters
- Split tweets around deplatforming date
- Wilcoxon signed-rank test, ITS
- Supporters’ activity and toxicity analyses
Findings
• Posts declined significantly, by 91.77% on average
• # Unique users diminished significantly, by 89.51%
• # New users declined significantly, by 89.65%
• Deplatforming helped significantly reduce the spread of many offensive ideas and conspiracy theories
• 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

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

<table>
<thead>
<tr>
<th>For each influencer</th>
<th>Identify relevant keywords</th>
<th>Collect tweets containing keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect # tweets, unique users, new users</td>
<td>Identify key antisocial ideas</td>
<td>Collect tweets containing those ideas</td>
</tr>
<tr>
<td>Split around deplatforming date</td>
<td>Split tweets around deplatforming date</td>
<td>Identify supporters</td>
</tr>
<tr>
<td>Activity analyses</td>
<td>Keyword trends analyses</td>
<td>Collect tweets posted by supporters</td>
</tr>
<tr>
<td>ITS</td>
<td>ITS</td>
<td>Split tweets around deplatforming date</td>
</tr>
<tr>
<td>ITS</td>
<td>ITS</td>
<td>Wilcoxon signed-rank test, ITS</td>
</tr>
</tbody>
</table>

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