The Effect of Extremist Violence on Hateful Speech Online

Alexandra Olteanu (IBM Research)
Carlos Castillo (UPF)
Jeremy Boy (UN Global Pulse)
Kush Varshney (IBM Research)

Presented by
Miguel Luengo-Oroz
(UN Global Pulse)
Hate speech is pervasive and can have serious consequences.

Ban Muslims, and you won’t have Islamic terrorism

Islam is the problem and everyone knows this

Muslim savages brainwash their kids into hating and killing non believers since really young

we should deport Muslim [expletive] and their families

#IllRideWithYou indicates one should not be scared to be a Muslim. One should be scared to be a racist

Paraphrased for anonymity!
How **extremist violence events** impact the prevalence of various **types of speech** online?

Quebec mosque shooting
January 2017
Islamophobic

Brussels bombings
March 2016
Islamist terrorism
This Study

19 months observation period

2 different social platforms: Twitter (107 M) and Reddit (45 M)

4 dimensions of hate speech: stance, intensity, target group, and frame

13 extremist attacks involving Arabs and Muslims as perpetrators or victims
Step 1: Data acquisition & lexicon construction

Methodology Overview

Step 1: Longitudinal data collection & lexicon construction (through iterative query expansion)
**Methodology Overview**

**Step 1:** Longitudinal data collection & lexicon construction (through iterative query expansion)

**Steps 2 & 3:** Data categorization & event selection (13 events)
Step 1: Longitudinal data collection & lexicon construction (through iterative query expansion)

Steps 2 & 3: Data categorization & event selection (13 events)

Step 4: Impact analysis, by employing causal inference techniques
Operationalizing Hate Speech

Speech that could be perceived as offensive, derogatory, or in any way harmful, and that is motivated, in whole or in a part, by someone’s bias against an aspect of a group of people, or related to commentary about such speech by others, or related to speech that aims to counter any type of speech that this definition covers.
<table>
<thead>
<tr>
<th>Target</th>
<th>Stance (re: individuals, groups, ideas)</th>
<th>Commentary</th>
<th>Framing (re: a potential problem)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabs - Muslims/Islam</td>
<td><strong>Favorable</strong> stance, supports</td>
<td><strong>Neutral</strong>, factual, or unclear</td>
<td>Diagnoses possible <strong>causes</strong></td>
</tr>
<tr>
<td>Other <strong>religious</strong> groups</td>
<td><strong>Unfavorable</strong> stance, against</td>
<td></td>
<td>Suggests possible <strong>solutions</strong></td>
</tr>
<tr>
<td>Other <strong>ethnic/national</strong></td>
<td></td>
<td></td>
<td>Both</td>
</tr>
<tr>
<td>immigrants/refugees, ...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others: gender, age, ...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severity</td>
<td><strong>Severity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offends or discriminates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intimidates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Promotes <strong>violence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. **Predict the counterfactual**: what would have happened had no event taken place?

2. **Estimate the effect**: what is the difference among observed behavior (the “factual”) and the predicted one (the “counterfactual”)

3. **Aggregate results**: what is the distribution of effects across platforms and types of speech

**Query**: evil muslim

![Graphs showing observation vs prediction and difference for Orlando nightclub shooting and Olathe Kansas shooting.](image-url)
Estimating Relative Effects

The counterfactual approach reveals substantial changes in the frequency of different markers of hate speech.

Each event is different, but there are some regularities.

Orlando nightclub shooting, June 2016
Islamist terrorist (& homophobic)
Regularities: Targeted Groups and Hate Speech Types

Overall, after an attack we observe stronger effects for:

- (increase) violent speech targeting Muslims, Arabs, and Islam
- (increase) counter speech related to religion e.g., promoting religious tolerance
- (increase) commentary about negative actions particularly towards immigrants
Regularities: Hate Speech and Counter-Speech

Following **Islamist terrorist attacks**, our estimates indicate:

- increases in hate speech targeting Muslims, Arabs, or Islam
  - **Twitter**: +3.0, 95%CI [1.7, 4.4], **Reddit**: +2.9, 95%CI [2.4, 3.3]

- increases in high-severity hate speech targeting Muslims, Arabs, or Islam
  - **Twitter**: +10.1, 95%CI [1.4, 18.9], **Reddit**: +6.2, 95%CI [3.9, 8.4]

- an overall increase in counter-speech terms *(more salient when related to religion)*
  - **Twitter**: +1.8, 95%CI [0.7, 3.0], **Reddit**: +2.9, 95%CI [2.4, 3.4]
Regularities: Hate Speech and Counter-Speech

Do Islamophobic attacks elicit a similar reaction? Our estimates suggest NO.

- **hate speech**: we do not see a consistent reaction neither across events, nor across platforms
- **counter-speech**: we do not see an overall increase across events
<table>
<thead>
<tr>
<th>Takeaways</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Violence leads to hate speech and counter-hate speech</td>
<td>● &quot;Seed&quot; terms may lead to bias</td>
</tr>
<tr>
<td>● Islamist terrorism attacks are followed by increases in hate speech against Muslims and Arabs</td>
<td>● Query-level annotations may introduce biases and noise</td>
</tr>
<tr>
<td>● Methodology based on counter-factual is useful for dealing with this type of series</td>
<td>● Analyzed messages were in English and related to attacks in the &quot;West&quot;</td>
</tr>
<tr>
<td></td>
<td>● Our analysis is retrospective and platforms actively delete harmful content</td>
</tr>
</tbody>
</table>

Studies like ours benefit from being **replicated** using different data sources, as well as different data collection & hate operationalization strategies!
Questions & Contact

Alexandra Olteanu (IBM Research)
alexandra@aolteanu.com / @o_saja

Carlos Castillo (UPF)
chato@acm.org / @ChaToX

Jeremy Boy (UN Global Pulse)
jeremy@unglobalpulse.org / @myjyby

Kush Varshney (IBM Research)
krvarshn@us.ibm.com / @krvarshney

Data will be released at:
https://github.com/sajao/EventsImpactOnHateSpeech

- Query terms
- Example time series
- Detailed crowdsourcing instructions

More results & details in the paper!