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)

IBM Research

upf. Universitat Pompeu Fabra Barcelona

GLOBAL PULSE

hate speech

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

Paraphrased for anonymity!

counter-hate speech

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

Hate speech is pervasive and can have serious consequences

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

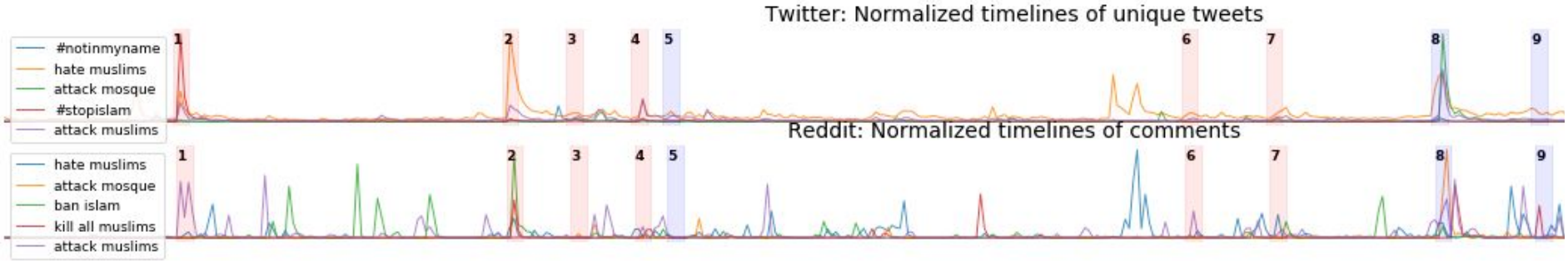
This Study

19 months observation period

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

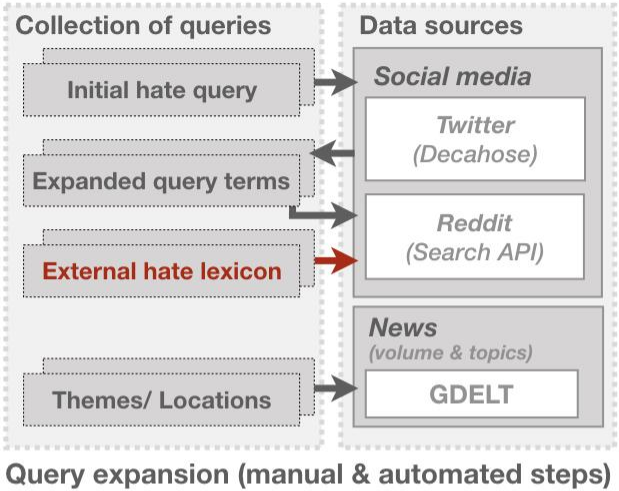
4 dimensions of hate speech: **stance**, **intensity**, **target group**, and **frame**

13 extremist attacks involving Arabs and Muslims as **perpetrators** or **victims**



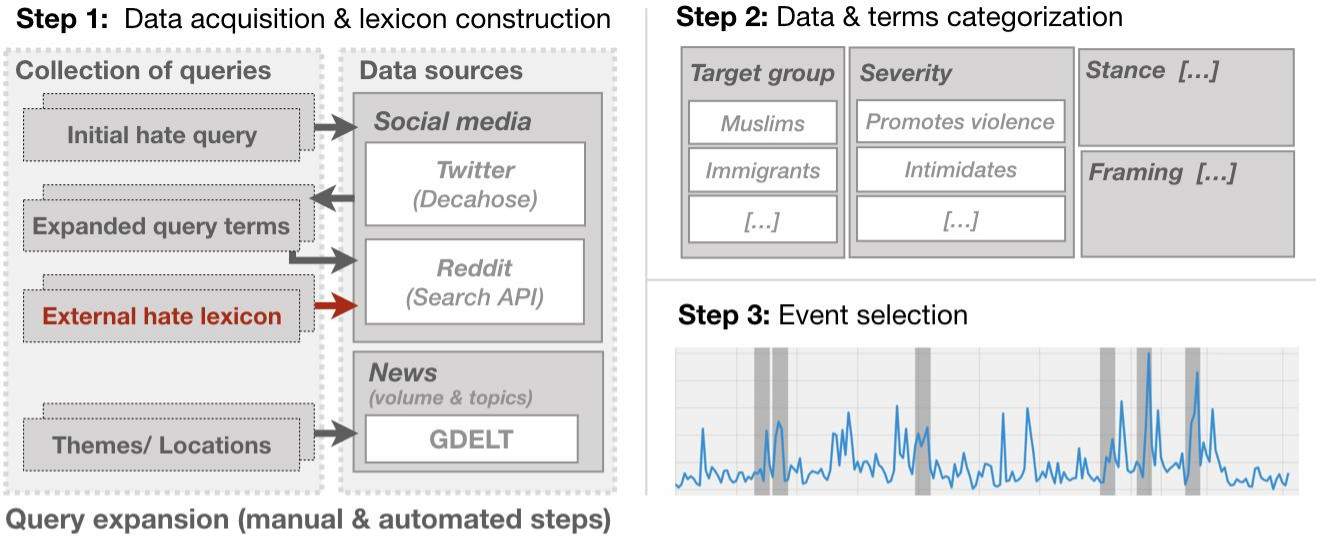
Methodology Overview

Step 1: Data acquisition & lexicon construction



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

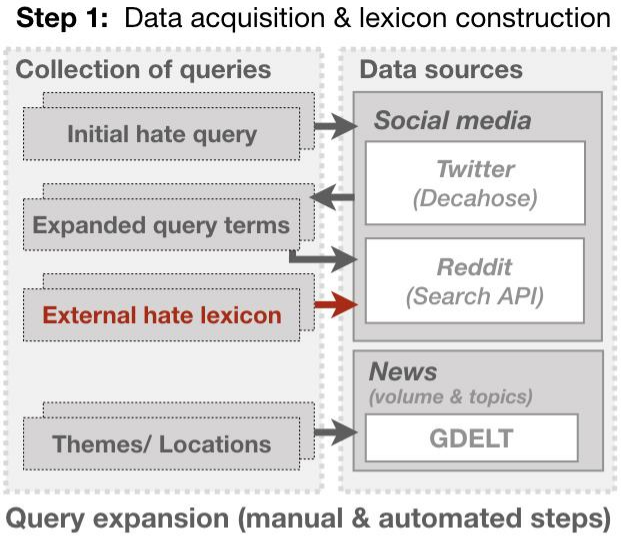
Methodology Overview

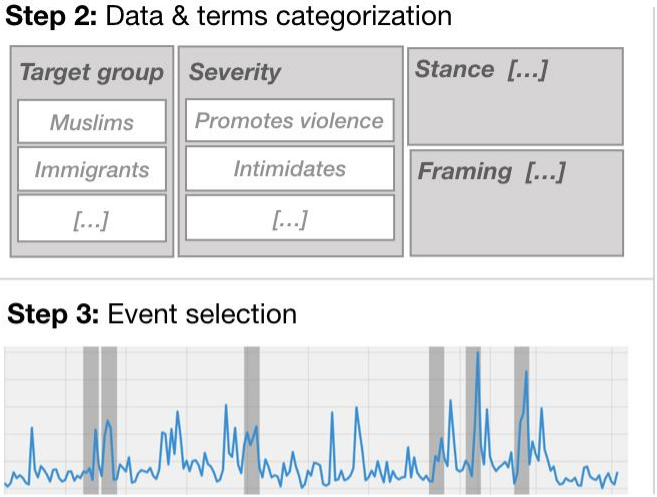


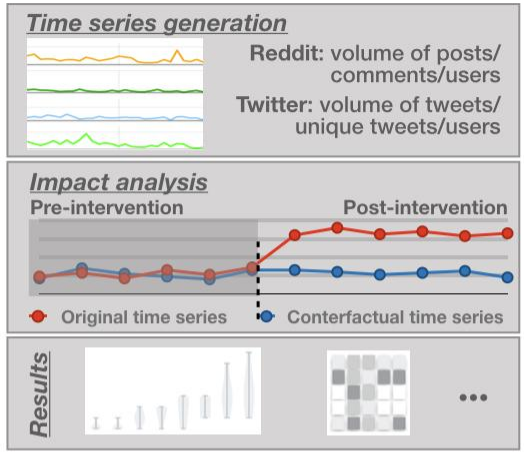
Step 1: Longitudinal data collection & lexicon construction (through iterative query expansion)

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

Methodology Overview







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

#agendaofevil, #attackamosque, #banislam, #bansharia, #cantcoexistwithislam, #deathcult, #deleteislam, #deportallmuslims, **#extremistsarenotmuslim**, #f***allah, #f***islam, **#illridewithyou**, #islamicinvasion, #islamistheproblem, #killallmuslims, #marchagainstsharia, #norapeugees, **#notinmyname**, #refugeesnotwelcome, #religionofhate, **#takeonhate**, #stopimportingislam, **#weareallmuslim,** #stopmoslemsinvasion, #islamiscrimmal, #islamisevil, **#terrorismhasnoreligion**

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.

Multidimensional Taxonomy of Hate Speech

Target

Arabs - **Muslims/Islam**

Other **religious** groups

Other **ethnic/national** groups

Immigrants/refugees, …

Others: gender, age, ...

Stance (re: individuals, groups, ideas)

Favorable stance, supports

Unfavorable stance, against

Commentary

Neutral, factual, or unclear

Severity

Offends or **discriminates**

Intimidates

Promotes **violence**

Framing (re: a potential problem)

Diagnoses possible **causes**

Suggests possible **solutions**

Both

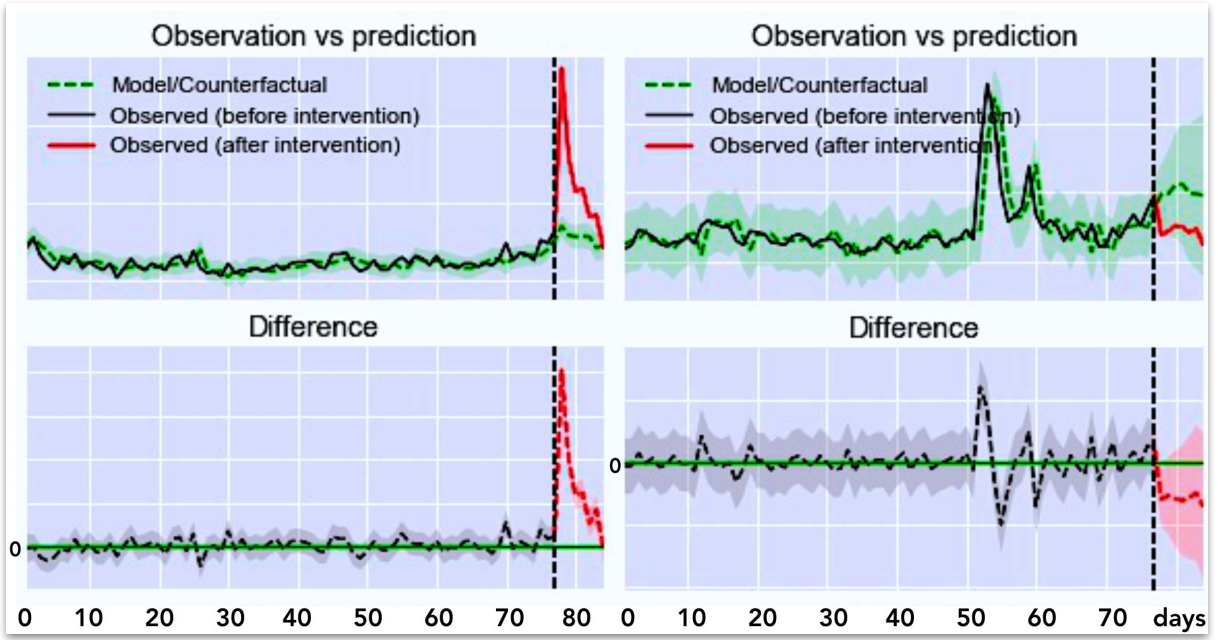
Quantifying the Impact of Events

1. **Predict the counterfactual**: what would had happen 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



Orlando nightclub shooting

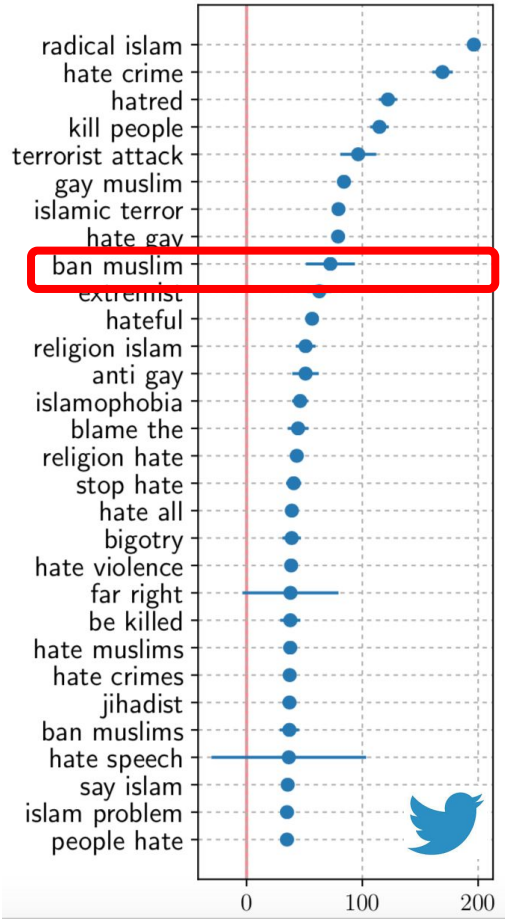
Olathe Kansas shooting

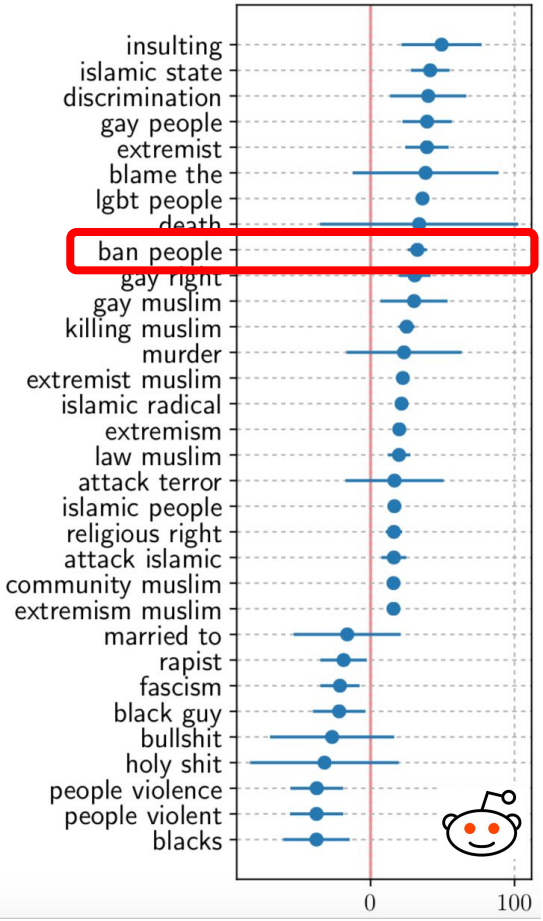
Estimating Relative Effects

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)

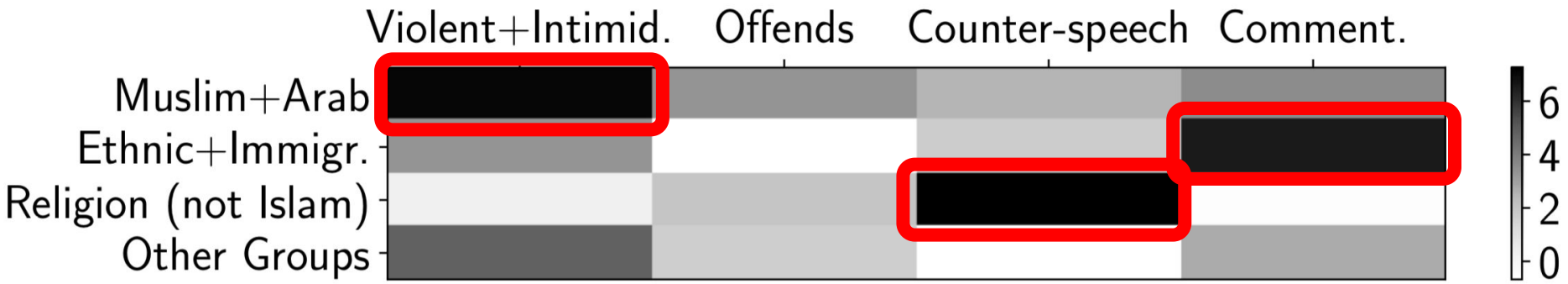




Regularities: Targeted Groups and Hate Speech Types

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

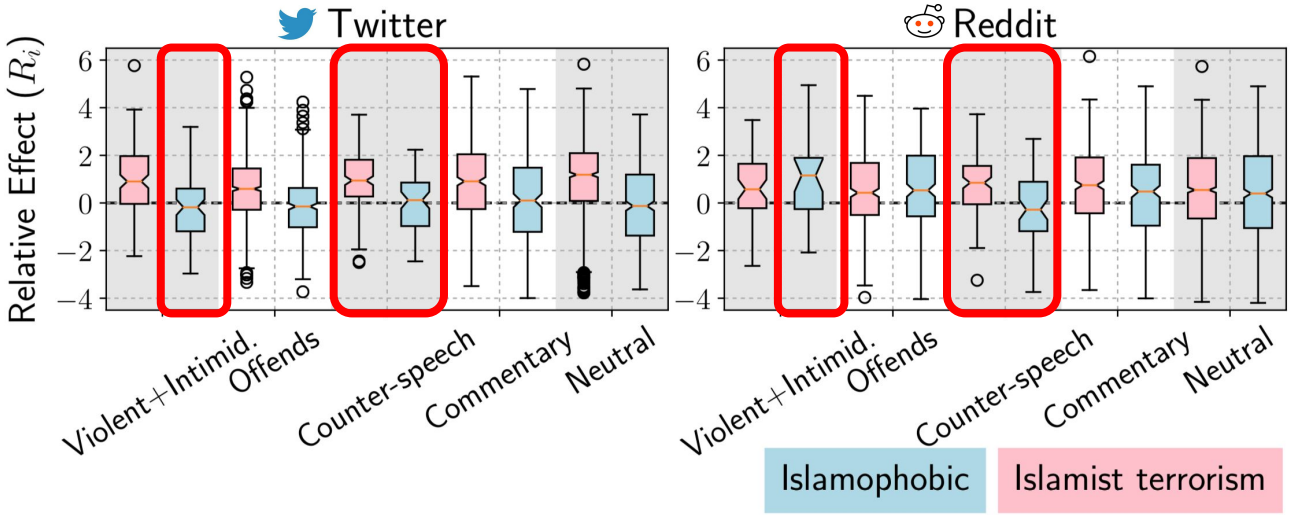
Following **Islamist terrorist attacks**, our estimates indicate:

- 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



Takeaways

- Violence leads to hate speech and counter-hate speech
- Islamist terrorism attacks are followed by increases in hate speech against Muslims and Arabs
- Methodology based on counter-factual is useful for dealing with this type of series

Limitations

- "Seed" terms may lead to bias
- Query-level annotations may introduce biases and noise
- Analyzed messages were in English and related to attacks in the "West"
- Our analysis is retrospective and platforms actively delete harmful content

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!