The Effect of Extremist Violence on Hateful Speech Online

Alexandra Olteanu
IBM Research
alexandra.olteanu@ibm.com

Carlos Castillo
UPF, Barcelona
chato@acm.org

Jeremy Boy
UN Global Pulse
jeremy@unglobalpulse.org

Kush R. Varshney
IBM Research
krvarshn@us.ibm.com

Abstract

User-generated content online is shaped by many factors, including endogenous elements such as platform affordances and norms, as well as exogenous elements, in particular significant events. These impact what users say, how they say it, and when they say it. In this paper, we focus on quantifying the impact of violent events on various types of hate speech, from offensive and derogatory to intimidation and explicit calls for violence. We anchor this study in a series of attacks involving Arabs and Muslims as perpetrators or victims, occurring in Western countries, that have been covered extensively by news media. These attacks have fueled intense policy debates around immigration in various fora, including online media, which have been marred by racist prejudice and hateful speech. The focus of our research is to model the effect of the attacks on the volume and type of hateful speech on two social media platforms, Twitter and Reddit. Among other findings, we observe that extremist violence tends to lead to an increase in online hate speech, particularly on messages directly advocating violence. Our research has implications for the way in which hate speech online is monitored and suggests ways in which it could be fought.

1 Introduction

Hate speech is pervasive and can have serious consequences. According to a Special Rapporteur to the UN Humans Rights Council, failure to monitor and react to hate speech in a timely manner can reinforce the subordination of targeted minorities, making them "vulnerable to attacks, but also influencing majority populations and potentially making them more indifferent to the various manifestations of such hatred" (Izsák 2015). At the individual level, people targeted by hate speech describe "living in fear" of the possibility that online threats may materialize in the "real world" (Awan and Zempi 2015). At the level of society, hate speech in social media has contributed to fuel tensions among communities, in some cases leading to violent clashes (Izsák 2015).

Following one of the most severe humanitarian crises in recent history, Europe has seen a high immigration influx, including Syrian, Afghan, and Iraqi refugees.[1] In the same period, several deadly terror attacks have occured in Western nations (Wang 2017; Global Terrorism Database 2017), leading to an increasingly alarming anti-Muslim rhetoric (TellMAMA 2017) by right-wing populist movements (Greven 2016) and right-leaning media outlets (Worley 2016), often conflating refugees and Muslims with Islamic fanatics (Diène 2006). This rhetoric has also gained adoption online (UNGP and UNHCR 2017), prompting governmental agencies[2] and NGOs to call on social media platforms to step up their efforts to address the problem of hate speech (Roberts 2017; TellMAMA 2017). The concern is that the increase in hateful narratives online led to an upsurge in hate crimes targeting Muslim communities (Roberts 2017). Insights into how online expressions of hate thrive and spread can help stakeholders' efforts to de-escalate existing tensions (Burnap and Williams 2014).

In this paper, we explore *how hate speech targeting specific groups on social media is affected by external events*. Anchoring our analysis in a series of Islamophobic and Islamist terrorism attacks in Western countries, we study their impact on the prevalence and type of hate and counter-hate speech targeting Muslims and Islam on two different social media platforms: Twitter and Reddit.

Our contribution. We conduct a quantitative exploration of the causal impact of specific types of external, non-platform specific events on social media phenomena. For this, we create a lexicon of hate speech terms, as well as a collection of 150M+ hate speech messages, propose a multidimensional taxonomy of online hate speech, and show that a causal inference approach contributes to understanding how online hate speech fluctuates. Among our findings, we observe that extremist violence attacks tend to lead to more messages directly advocating violence, demonstrating that concerns about a positive feedback loop between violence "offline" and hate speech online are, unfortunately, well-founded.

Paper Outline and Methodology Overview. Outlined in Figure 1, our approach consists of several steps:

Step 1: We create a longitudinal collection of hate speech messages in social media from Twitter and Reddit that covers a period of 19 months. This collection is based on a series of keywords that are obtained through an iterative expansion

[1] World Economic Forum (Dec. 2016) "Europe's refugee and migrant crisis" https://www.weforum.org/agenda/2016/12/europes-refugee-and-migrant-crisis-in-2016-in-numbers

[2] BBC News (Sep. 2017) "Social media warned to crack down on hate speech" http://bbc.com/news/technology-41442958

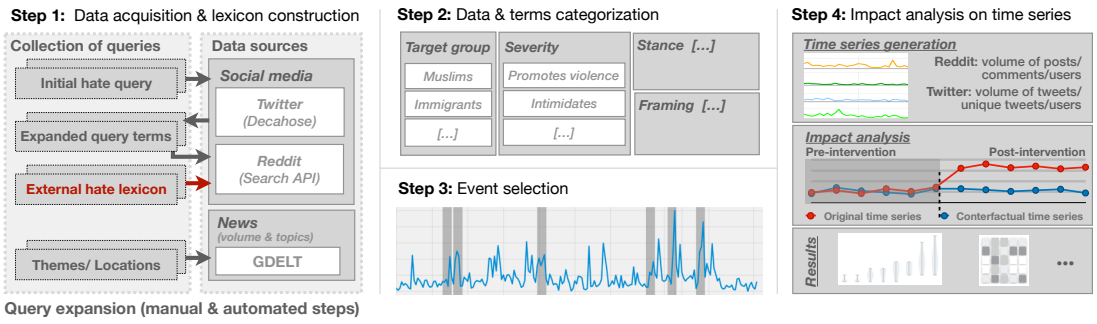

Figure 1: Steps in our analysis framework: (1) data acquisition and lexicon creation, §3; (2) data categorization, §4; (3) events selection, §6; (4) impact analysis on time series of hate related terms, §5 and results, §6.

of known hate speech terms (§3).

Step 2: We categorize the data along four dimensions: 1) the group each message refers to, 2) the attitude of speakers, 3) the severity of hateful expressions, particularly whether they advocate violence, and 4) the framing of content (§4).

Step 3: We select 13 extremist attacks involving Arabs and Muslims as perpetrators or victims, like the Berlin Christmas market attack on Dec. 2016, perpetrated by a follower of jihadist group ISIL, or the Quebec City mosque shooting on Jan. 2017, by a far-right white nationalist (Table 2).

Step 4: As evaluating the effect of such attacks on various slices of social media is a causal question, we frame it as measuring the impact of an intervention (event) on a time series (temporal evolution of speech). Following techniques for causal inference on time series (Brodersen et al. 2015), we estimate an event's impact on various types of hate and counter-hate speech by comparing the behavior of corresponding time series after an event, with counterfactual predictions of this behavior had no event taken place (§5).

The last sections present (§6) and discuss (§7) our results.

2 Background & Prior Work

We are interested in the relation between online hate speech and events. To ground our study, we first review work defining hate and counter-hate speech. Given our focus on anti-Muslim rhetoric in the context of extremist violence, we outline previous works on hate speech after terror attacks, and studies of hateful narratives targeting Muslims. We also cover observational studies on social media, particularly those focusing on harmful speech online.

Hate Speech Online and Offline

Defining hate speech. While hate speech is codified by law in many countries, the definitions vary across jurisdictions (Sellars 2016). As we study hate speech directed at followers of a religion (Islam) that is often conflated with a particular ethnicity and culture, we use the definition by the Council of Europe covering "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin." (Committee of Ministers, Council of Europe 1997). We extend this definition to further consider as hate speech "animosity or disparagement of an individual or a group on account of a group characteristic such as race, color, national origin, sex, disability, religion, or sexual orientation" (Nockleby 2000), allowing us to juxtapose hateful speech directed at Muslims with other groups (e.g., LGBTQ, immigrants).

Counter-hate speech. Censoring hate speech may clash with legal protections on free speech rights. Partially due to this tension, the position of international agencies like UNESCO is that "the free flow of information should always be the norm. Counter-speech is generally preferable to suppression of speech" (Gagliardone et al. 2015). Thus, it is important not only to study hate speech, but also to contrast it with counter-speech efforts—a rare juxtaposition in social media research (Benesch et al. 2016; Magdy et al. 2016). Magdy et al. (2016) estimate that a majority of Islam and Muslim related tweets posted in reaction to the 2015 terrorist attacks in Paris stood in their defense; an observation also made by UNGP and UNHCR (2017) following the 2016 terrorist attack in Berlin, and supported by our own results (§6).

Hate speech and violent events. The prevalence and severity of hate hate speech and crimes tends to increase after "trigger" events, which can be local, national, or international, often galvanizing "tensions and sentiments against the suspected perpetrators and groups associated with them" (Awan and Zempi 2015). For instance, Benesch et al. (2016) found extensive hate and counter-hate speech after events that triggered widespread emotional response like the Baltimore protests, the U.S. Supreme Court decision on same-sex marriage, and the Paris attacks during 2015; while Faris et al. (2016) found spikes in online harmful speech to be linked to political events. While these studies are related to ours, they focus on content posted during specific events, or on correlating changes in patterns (e.g., spikes) with events' occurrence. We focus on broader patterns, aiming to quantify changes across types of events and types of content by applying causal inference techniques.

Islamophobia. The conflation of Muslims and Islam with terrorism—particularly developed after September 11, 2001—is a key factor behind the increase in Islamophobic attitudes (Diène 2006). A significant increase in anti-Muslim hate crimes was observed after terrorist attacks by individu-

als that identify as "Muslim or acting in the name of Islam," with those having a "visible Muslim identity" being the most vulnerable to hostility, including online and offline intimidation, abuse and threats of violence (Awan and Zempi 2015).

Observational Studies Using Social Media

Hate speech on online social platforms. While social media platforms provide tools to meet new people, maintain relationships, promote ideas, and promote oneself; they have also opened up new avenues for harassment based on physical appearance, race, ethnicity, and gender (Duggan 2017). This has led to efforts to detect, understand, and quantify such harmful speech online, with goals such as modeling socially deviant behavior (Cheng et al. 2017), building better content filtering and moderation tools (Matias et al. 2015), and informing policy makers (Faris et al. 2016).

The main categorization criteria for online hate speech has been based on the group being targeted (e.g., "black people," "fat people"), the basis for hate (e.g., race, religion) (Silva et al. 2016; Mohammad et al. 2016), and the speech severity (Davidson et al. 2017). For instance, Silva et al. (2016) found "soft" targets like "fat people" to be among the top target groups; yet, these groups are often not included in the documentation of offline hate crimes. Davidson et al. (2017) further discuss challenges in distinguishing between hate speech and other types of offensive speech.

Observational methods applied to social data. Recent studies show that quasi-causal methods can be applied to social media data to e.g., distill the outcomes of a given situation (Olteanu, Varol, and Kıcıman 2016), measure the impact of an intervention (Chandrasekharan et al. 2018), or estimate the effect of online social support (Cunha, Weber, and Pappa 2017). The application of these methods to social data, including propensity score matching (De Choudhury et al. 2016), difference-in-differences (Chandrasekharan et al. 2018), and instrumental variables (Zhang, Li, and Hong 2016), was found to reduce confounding biases.

Chandrasekharan et al. (2018)'s work is closest to ours, as it employs techniques from the causal inference literature to quantify the impact of an intervention on hateful behavior on Reddit. Yet, the intervention they study is platform-specific– a ban on an existing community on Reddit–whereas we look at the impact of external (non-platform specific) events on both Reddit and Twitter. Our focus is on the overall prevalence of hate speech, rather than on the behavior of given groups of users, and we measure the effect of given interventions (events) on various types of hate speech (§4).

Operationalization of hate speech on social media. Due to lack of consensus on what constitutes hate speech and the challenges in operationalizing existing definitions at the scale of current online platforms, prior work has used a mix of manual and automated term selection strategies to identify terms that are likely to occur in hateful texts (Chandrasekharan et al. 2018; Davidson et al. 2017). While focusing on speech targeting Muslims and Islam, we similarly combine existing lexicons with terms obtained through a combination of manual and automated steps (§3).

3 Data Collection

Our goal is to characterize and measure online hate speech targeting Muslims and Islam in reaction to major Islamist terror attacks and Islamophobic attacks perpetrated in Western countries. Here, we describe our data collection process, which attempts to be inclusive (high-recall) and hence uses a broad definition of hate and counter-hate speech. We iteratively expand an initial query of keywords related to relevant items by identifying new keywords in the retrieved messages. Our base datasets contain messages from Twitter and Reddit, and a collection of news articles; these are not associated to any particular event, but cover messages potentially related to hate and counter-hate speech over a period of 19 months: from January 1, 2016 to August 1, 2017.

Data Sources

Twitter (https://twitter.com/) is one of the largest microblogging platforms used by hundreds of millions every month. To collect Twitter messages ("tweets") we use an archive representing 10% of the entire public stream, known as the "Decahose."

Reddit (https://reddit.com/) is a large social news aggregation platform used by millions every month. Users submit "posts" and "comments" that gain or lose visibility according to up- and down-votes. We collect posts through Reddit's Search API[3] (comments are not searchable via this API), retaining all comments to posts matching our queries.

News. Finally, we collect news articles from GDELT (Global Data on Events, Location, and Tone, http://gdeltproject.org/), the largest online catalog of global news events. We use these data as exogenous variables when modeling social media time series before and after a given event.

Query Construction

We collected data using keyword queries, a sampling method applicable to both Twitter's and Reddit's APIs. As our goal was to create a high-recall collection, our sampling procedure consists in formulating an initial query (bootstrapping), followed by an expansion of that query. This method is known to improve the coverage of social media data (Olteanu et al. 2014; Davidson et al. 2017).

Query Bootstrapping. We bootstrapped our query selection with an initial list of terms (keywords and hashtags) used in social media campaigns related to anti-Muslim hate and counter-hate speech. This list was assembled retrospectively (as was the rest of our data) using (i) news articles and blog posts discussing social media usage during hate and counter-hate campaigns,[4] (ii) resources from NGOs or governmental agencies tracking or analyzing hate speech on social media (Awan and Zempi 2015; UNGP and UNHCR 2017), and (iii) research articles (Magdy et al. 2016). Selected terms were individually validated by manually searching for them on both Twitter and Reddit. Additional co-occurring

[3] Using the PRAW library: https://praw.readthedocs.io/.

[4] E.g., http://huffingtonpost.com/entry/anti-muslim-facebook-twitter-posts-myths-wajahat-ali_us_57f55bb1e4b002a7312084eb or http://muslimgirl.com/29612/10-hashtags-literally-changed-way-muslims-clap-back/

Source	Terms	Messages	Users	URLs
Twitter	825	107M	26M	–
Reddit	1,257	45M	3.3M	–
News	–	–	–	3.3 M

Table 1: Summary of the final data collection.

terms found in this process were added to the list. This step resulted in a list of 91 terms, including "'#f***quran," "#nosharia," "ban islam," and "kill all muslims."

Query Expansion. We then employed a query expansion heuristic to identify further terms that may appear in messages expressing hate or counter-hate towards different groups, including, but not limited to, Arabs and Muslims. The heuristic considers terms frequently appearing in social media messages matched by the terms in our initial list. To obtain a high-recall collection, we considered any new term that may constitute hate or counter-hate speech, using an inclusive, broad definition inspired by Silva et al. (2016) and Chatzakou et al. (2017), and expanded to also cover commentary and counter-hate speech elements. We recorded all terms related to *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.*

This expansion was independently done in two iterations for both Twitter and Reddit. First, one of the authors did an annotation pass to identify new query terms. Second, as we favored recall, at least one other author did an additional annotation pass over the terms rejected by the first annotator.

External lexicon. To further expand our list of query terms, we added terms from a lexicon built using HateBase,[5] a website that compiles phrases submitted and tagged by internet users as constituting hate speech. Given that only an estimated 5% of messages containing HateBase terms were actually identified as hateful; instead of directly using these terms, we used 163 unique terms extracted from Twitter messages containing HateBase terms and manually annotated as hateful or offensive by Davidson et al. (2017).[6]

Data Acquisition

Table 1 presents a summary of the data we acquired.

Acquiring Twitter data. We first queried the bootstrap terms, and retrieved 958K messages posted by 413K users. We then expanded the query by manually annotating 2088 terms that appeared more frequently than an arbitrary threshold (ranging from 75 for tri-grams to 300 for uni-grams, which are typically less precise than tri-grams and noisier at lower frequencies), after removing stopwords using the Python NLTK package. We found an extra 612 terms. We queried these terms, growing our collection by 55M tweets posted by 12.5M users. The resulting dataset contains on average 4.5M tweets per month. Since we used the Twitter Decahose (a 10% sample of all Twitter content), we estimate this collection is in fact representative of a larger set of roughly 45M tweets per month. Finally, we retrieved tweets matching the 163 external hate terms (based on HateBase), resulting in an additional 51.6M tweets by 13.7M users. Altogether, we collected over 1TB of raw Twitter data.

Acquiring Reddit data. We again began by querying the bootstrap terms, and retrieved 3K posts with 140K comments written by 49K users. We then expanded the query by selecting high-frequency terms (thresholds ranging from 50 to 300 as these data were sparser than Twitter) across all posts and comments, and manually annotating them. Given that the Reddit Search API normalizes terms before running a query, we did not keep different inflections of the same terms. We annotated 4272 terms, and found 1002 related to hate and counter-hate speech. We queried these terms, and retrieved an extra 300K posts with 41M comments written by 3.1M users. Finally, we queried the external hate terms. Altogether, we collected 337K posts with 45M comments written by roughly 3.3M users.

Acquiring news data. We used GDELT's Global Knowledge Graph (GKG), as it provides the list of news articles covering each event in their database. This allowed us to compute the overall volume of news per day, amounting to over 130M URLs over our 19 months period of interest.

4 Characterizing Hate Speech

Here, we present example themes from messages posted in the aftermath of extremist events (listed in §6), and characterize them along four dimensions (stance, target, severity, and framing), which we then use to analyze the data.

Exploration of Post-Event Messages

To understand how the content and themes of messages vary with respect to who is mentioned, what is said, and how the content is framed, we review messages posted after one terrorist and two Islamophobic attacks: *Manchester Arena bombing*, an Islamist terrorist attack in Manchester that targeted concert goers, killing 23 people and wounding 512 others; *Portland train attack*, carried out by a man shouting racial and anti-Muslim slurs who fatally stabbed two people and injured a third; and *Quebec City mosque shooting* that targeted worshipers, leaving 6 dead and 19 injured. We focus on these particular events for their overall difference in nature. Table 2 includes example messages.

Who is mentioned? Naturally, many messages mentioned (directly or indirectly) Arabs, Muslims, or Islam, given how we collected our data and the focus of our study. Yet, we also found messages mentioning the victims of the attacks, the mainstream media, political and religious groups (e.g., "the left", "Christians"), immigrants in general, and high-profile individuals (e.g., politicians, journalists).

What is said, and why? The content of the messages ranged from blaming Arabs and Muslims for the attack, to providing context and defending Islam. Some messages made crude generalizations or included denigrating insults, while others appeared to either intimidate or incite violence.

[5] HateBase: https://www.hatebase.org/

[6] Given that our queries represent a conjunction of the words in each term and that for Reddit articles are ignored, we pre-process most terms to remove the article *a* and remove duplicates.

Theme	Paraphrased message/comment(s)	Theme	Paraphrased message/comment(s)
Blames Muslims	"*Ban Muslims, and you won't have Islamic terrorism*" (T) "*Islam is the problem and everyone knows this*" (R)	Defends Muslims	"*killing innocent people is not Islam, there were Muslims at that concert as well*" (T) "*#IllRideWithYou indicates one should not be scared to be a Muslim. One should be scared to be a racist*" (T)
Denigrates or intimidates	"*Muslim savages brainwash their kids into hating and killing non believers, as apes and pigs, since really young*" (R)	Incites violence	"*#StopIslam wipe its followers from the face of the earth*" (T)
Diagnoses causes	"*the left say, look they were not refugees; the fact is that this would never happen if we would have banned them*" (R)	Suggests a remedy	"*we should deport Muslim scumbags and their families*" (R)

Table 2: Example messages from Reddit (R) and Twitter (T) for some of the analyzed events, provided for illustration purposes. Messages have been (sometimes heavily) paraphrased for anonymity.

How is the content framed? According to Entman (1993), content may be framed according to whether it defines a problem, diagnoses its causes, makes a moral judgment, or suggests a remedy. Several messages echoed similar points of view about what the "problem" might be, what the causes are, and what the solutions should be. After the Manchester Arena Bombing, a repeated theme could be paraphrased as "*if Muslims were not allowed in the country, there would be no terrorist incidents*." A proposed solution was e.g., "*stop Islamic immigration before it is too late*." Some messages went further, linking the event to immigration more broadly. We also found posts framing the event (and what happened after) as a lesson on what Islam is, and what it stands for, e.g., "*Islam only wants to kill and rape, [Q]uran is a manual of evil*." Yet, other messages tried to push back on this framing by suggesting a counter-narrative.

Four Dimensions of Online Hate Speech

Based on prior work and our exploration of post-event messages, we derive four main dimensions of hate and counter-hate speech: **stance**, **target**, **severity**, and **framing**. While these are useful, we recognize these dimensions are unlikely to capture all aspects of online expressions of hate.

Stance. Magdy et al. (2016) make a distinction between online speech that *attacks* and blames, speech that *defends*, and speech that is *neutral* towards Islam and Muslims following a terrorist attack. Benesch et al. (2016) introduce a taxonomy for spontaneous expressions of counter-hate speech on social media platforms. We adapt these categorizations to define the following stances of speech for our study:

- *Takes a favorable stance in support of individuals, groups, or ideas:* defend, show solidarity, propose counter narratives, denounce, or comment on acts of hatred, or emphasize the positive traits of individuals, groups, or ideas (e.g., #ThisIsNotIslam, #NotInMyName);
- *Takes an unfavorable stance against individuals, groups, or ideas:* attack, blame, denigrate, demean, discriminate, employ negative stereotypes, seek to silence, or generally emphasizes the negative traits of an individual or group (e.g., "kill all Muslims," #RefugeesNotWelcome);
- *Commentary on negative actions or speech against individuals, groups, or ideas:* comment on or characterize acts of violence, hatred, harassment, or discrimination (e.g., "hate speech," "racial slur"); and
- *Neutral, factual, or unclear if it is in support or against a person or group:* none of the above; report news facts or comments, describe an event, or not related to a minority or vulnerable group (e.g., "seven injured," "white van").

Target. Hate speech can target any minority or vulnerable group by singling out its identifying characteristics. In the case of Muslims or Islam, these characteristics include religion, country of origin, immigration status, ethnicity, or a conflation of several or all characteristics. We identify the following targets of hate and counter-hate speech:

- *Muslims and Islam*;
- *Religious groups:* unspecified, any religion except Islam;
- *Arabs, Middle-Easterners, or North Africans:* descent without reference to religion;
- *Ethnic groups or groups of foreign descent:* unspecified, any foreign descent, except Arab;
- *Immigrants/refugees/foreigners in general:* without indicating a specific religion or descent; and
- *Other groups of non-immigrants:* based on e.g., gender, sexual orientation, appearance, disability, or age.

Severity. International organizations are concerned with how hate speech can lead to violent acts (Izsák 2015). Expressions of hate take many forms (Ghanea 2013; Matias et al. 2015); they can be ambiguous, and the perception of what is hateful varies between individuals (Olteanu et al. 2017). Capturing such subtleties is essential to understanding how severe the repercussions of online hate speech can be; for instance, the Jewish *Anti-Defamation League* defines a "Pyramid of Hate," showing how prejudice enables discrimination, which enables violence, which enables genocide.[7] We use the following levels of severity of hate speech:

- *Promotes violence:* threaten with violence, incite violent acts, and intend to make the target fear for their safety (e.g., "attack mosque," "kill muslims");
- *Intimidates:* harass or intimidate the target, or invite others to do so, while actively seeking to cause distress (e.g., "deport illegals," "Muslims not welcomed");
- *Offends or Discriminates:* defame, insult, or ridicule the target, showing bias, prejudice, or intolerance, while actively seeking to embarrass or harm the target's reputation, (e.g., "Muslim [expletive]," "sand [n-word]");

Framing. Kuypers (2010) defines framing as the "process whereby communicators, consciously or unconsciously, act to construct a point of view that encourages the facts of a

[7] See: https://sfi.usc.edu/lessons/pyramid-hate

given situation to be interpreted by others in a particular manner." Benford and Snow (2000) note that framing is critical to understand social movements and collective action; it can also operate in different ways (Entman 1993). For our analysis, from test annotations we noticed that two frames were quite distinguishable in the text and complementary:[8]

– *Diagnoses the cause or causes for a problem* (or elements seen as possible causes): identifies what creates a problem, suggests a diagnose or disagrees with a diagnose (e.g., "terrorists exist because they come from a place that, socially, is centuries behind");
– *Suggests a solution or solutions for a problem* (or actions seen as possible solutions): proposes or defends actions seen as solving or removing the problem (e.g., "we should target the mosques and [M]uslims, this is what you need to do when at war with these [expletive]");
– *Both diagnoses causes and suggests solutions*: if both of the above categories apply to the message.

Terms or sentences may perform multiple of these framing functions, but they may also perform none of them (Entman 1993). Thus, for annotation purposes we add a catch-all category for those cases where none of these functions apply.

5 Methodological Framework

To quantify how extremist violence events affect the prevalence of various types of speech, we treat these events as interventions on observed time series. Following existing techniques for causal inference on time series (Eichler 2012; Brodersen et al. 2015), we measure this effect by comparing the behavior of an observed time series (which we refer to as *treated*) after an event with a counterfactual time series of its behavior had the event not taken place. This synthetic unobserved counterfactual time series (which we refer to as *control*) is modeled from several observed time series that may be correlated to the treated time series (yet not affected by the event), as we describe below. The causal effect is then estimated based on the differences between the treated and the control time series. Broadly, since we model the counterfactual of the treated time series, this is a generalization of the application of the differences-in-differences techniques to time series (Brodersen et al. 2015).

Observed Time Series. We consider time series covering our 19-month observation period with a granularity of one day. For each of the 825 terms we have for Twitter, we experiment with three time series: one for the number of tweets, one for the number of tweets excluding re-tweets (i.e., unique after removal of "RT @user" prefixes), and one for the number of unique users. Similarly, for the 1,257 terms we have for Reddit, we experiment with three time series: one for the number of posts, one for the total number of comments in these posts, and one for the total number of unique users in the post and comments.

Synthetic Control Time Series. A synthetic control time series is a *counterfactual* that reflects behavior *had the extremist violence event not taken place*. For each treated time series, we build a control series for 1 week following the event based on several data sources:[9] (1) the observed series in the 11 weeks leading to the event; (2) the observed series exactly 1 year before the event, for 12 weeks (corresponding to the 11 weeks before and 1 week after the event, but a year earlier); (3) the observed series 23 weeks prior to the event, similarly for 12 week;[10] and (4) external information from Twitter, Reddit, and news sources.

The external information includes time series whose behavior is unlikely to be affected by the events (§6). First, we use the overall volume of news on GDELT (i.e., number of daily news article URLs per day) as it does not seem to be affected by any of our events during the observation window. Second, we use the overall number of tweets containing the word "news" which we also observe is not affected by any of our events (also a proxy for the overall volume of tweets). Third, we use the overall number of Reddit posts containing the general term "people," which we also observe is not affected by the events in our list (this is not the case for the series of posts in Reddit containing, e.g., the term "breaking news" which was affected by several of our events).

The methodology for synthesizing the control follows Brodersen et al. (2015), using a state space model to predict the counterfactual from the various sources we described above. However, our models are fit using maximum likelihood estimation (Fulton 2015) rather than Bayesian methods like Markov chain Monte Carlo preferred by Brodersen et al. (2015). Our implementation uses the state space model in the *UnobservedComponents* Python package to model and predict the series, following existing Python implementations of Brodersen et al. (2015).[11]

Impact Estimation. To estimate the effect of an event using the treatment and control time series, we compute the relative lift or drop as $\mathrm{rel}_{\mathrm{effect}} = 100 \times \frac{\sum t_k - c_k}{\sum c_k}$, where t_k is the value of the treated time series at time k, and c_k that of the control time series. The summations are over the days we observe after the event, seven in our case. We focus on relative effect as it better allows for comparison across events. For each event, we rank terms based on the relative effect.

Some of our time series have intervals of low volume (particularly for Reddit) that may lead to negative-valued synthetic controls and skewed estimates of the effect. To address this, we add a large constant C to all time series before synthesizing the control and estimating the effect. This transformation preserves the shape and amplitude of the impact.[12]

[8] The "makes a moral judgment" frame is present in some form in many messages, but often supporting another frame; the "defines a problem" frame is rarely seen without the "diagnoses its causes."

[9] The 1 week observation period allows us to observe the bulk of the effects, e.g., after a terror attack in UK, (Burnap et al. 2014) found that related messages were propagated for less than 2 days.

[10] Recall that our entire dataset spans 19 months. When information one year before is not available, we use one year later; when information 23 weeks before is not available, we use 5 weeks later.

[11] Python modules used: http://statsmodels.org/dev/generated/statsmodels.tsa.statespace.structural.UnobservedComponents.html; similar Python implementations https://github.com/tcassou/causal_impact, and https://github.com/jamalsenouci/causalimpact.

[12] We experimented with logarithmic transformations, but believe that a linear transformation leads to more interpretable confidence intervals and estimates. We arbitrarily set $C = 1000$.

Num.	Date	Name	Type	URL	Language	Country
1	March 22 2016	Brussels bombings	Islamist terrorism	http://r.enwp.org/2016_Brussels_bombings	French/Dutch	Belgium
2	June 12 2016	Orlando nightclub shooting	Islamist terrorism	http://r.enwp.org/2016_Orlando_nightclub_shooting	English	United States
3	June 28 2016	Istanbul Airport attack	Islamist terrorism	http://r.enwp.org/2016_Ataturk_Airport_attack	Turkish	Turkey
4	July 14 2016	Nice attack	Islamist terrorism	http://r.enwp.org/2016_Nice_attack	French	France
5	July 22 2016	Munich shooting	Islamophobic	http://r.enwp.org/2016_Munich_shooting	German	Germany
6	November 28 2016	Ohio State University attack	Islamist terrorism	http://r.enwp.org/2016_Ohio_State_University_attack	English	United States
7	December 19 2016	Berlin attack	Islamist terrorism	http://r.enwp.org/2016_Berlin_attack	German	Germany
8	March 22 2017	Westminster attack	Islamist terrorism	http://r.enwp.org/2017_Westminster_attack	English	United Kingdom
9	January 29 2017	Quebec City mosque shooting	Islamophobic	http://r.enwp.org/Quebec_City_mosque_shooting	English	Canada
10	February 23 2017	Olathe Kansas shooting	Islamophobic	http://r.enwp.org/2017_Olathe,_Kansas_shooting	English	United States
11	May 22 2017	Manchester Arena bombing	Islamist terrorism	http://r.enwp.org/2017_Manchester_Arena_bombing	English	United Kingdom
12	May 26 2017	Portland train attack	Islamophobic	http://r.enwp.org/2017_Portland_train_attack	English	United States
13	June 19 2017	Finsbury Park attack	Islamophobic	http://r.enwp.org/2017_Finsbury_Park_attack	English	United Kingdom

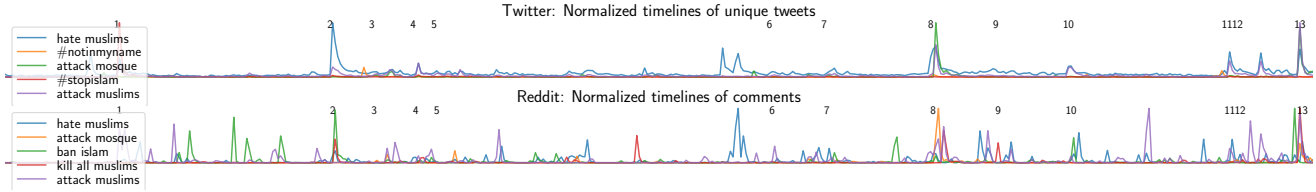

Figure 2: List of events we consider in this study (top), and examples of normalized time series corresponding to top 5 bootstrap terms by volume on both Twitter and Reddit (bottom). Markers in the time series correspond to event numbers in the table.

	Stance				**Severity**			**Target**					**Framing**		
	Favor.	Unfavor.	Comm.	Neutral	Violent	Intimid.	Offend	Muslims	Relig.	Ethnic	Immigr.	Non-Immigr.	Cause	Sol.	Both
Reddit	6.7%	31.6%	16.4%	43.3%	13.8%	9.8%	75.8%	47.4%	6.3%	9.4%	2.0%	6.7%	64.5%	2.5%	9.7%
Twitter	5.1%	48.3%	16.7%	27.0%	22.1%	9.8%	67.5%	40.6%	4.5%	11.2%	3.1%	8.2%	56.5%	8.3%	24.1%

Table 3: Distribution of annotations along the entire 19-month observation period, done at the term level (except framing, done at the message level). The percentages may not add to 100% as we omit the cases when none of the categories apply.

6 Experimental Results

In this section, we present experimental results that estimate how different types of events affect various forms of online speech. First, we select 13 extremist violence attacks (Islamist terrorist and Islamophobic), that occurred during our full 19-month observation period. Next, we annotate our data at query term level for stance, target, and severity, and at message level for framing, according to the hate speech taxonomy introduced in §4. Finally, we present results on various categories of hate speech across events and platforms.

Experimental Setup

Events Selection. We select a set of extremist violence attacks in Western countries involving Arabs and Muslims as perpetrators or victims, and covered by international news media. Our sources are two Wikipedia pages listing Islamist terrorist attacks and Islamophobic incidents.[13] When two events occur within the same week, we selected the one with the largest number of victims, also the most prominent in the news. The list of events is available in Figure 2, where we also display the time series of top-5 bootstrap terms (§3) on Twitter and Reddit, which shows that these events cover most of the peaks in these terms for Twitter and a majority of them for Reddit.

Crowdsourced Annotations. Our entire list of terms contains 1890 unique terms, which we annotate by employing crowdsource workers through the Crowdflower platform. We select workers from countries having a majority of native English speakers or that were affected by the events (e.g., Germany). Except for "framing," for cost and scalability purposes, we annotate each term with the most likely category the text containing them may fall under. For framing we annotate entire messages, as annotating at the term-level annotations does not produce reliable labels.

For each hate speech dimension and category, we provide detailed definitions and extensive examples; and, for each term we annotate, we show crowd workers clickable links to corresponding search results matching our queries, as returned by both social media platforms, Twitter and Reddit, as well as by two major search engines, Bing and Google. Following standard crowdsourcing practices, we gather at least 3 annotations per term (up to 5 when consensus was not reached), using a set of unambiguous test questions provided by the authors to catch inattentive workers, and resolving disagreements by majority voting. For framing, for each event we annotate samples of 5-6 messages matching the top 100 terms by relative effect, and posted around the time

[13] https://en.wikipedia.org/wiki/Islamophobic_incidents and https://en.wikipedia.org/wiki/List_of_Islamist_terrorist_attacks

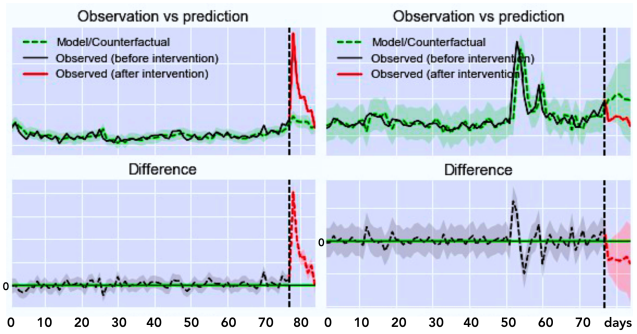

Figure 3: Example of impact estimation with counterfactual predictions, for the term "*evil muslims*." Black/red are the observed series before/after the event, green the counterfactual. Top: time series of tweets containing the term after an Islamist terrorism attack (left: Orlando nightclub shooting) and an Islamophobic attack (right: Olathe Kansas shooting). Bottom: differences between observed and counterfactual.

of the event.[14] To obtain the dominating frame of a term, we first determine the label of the messages it matches, and then assign by majority voting to each term the most prevalent frame, or if the "causes" or "solutions" frames are similarly prevalent, we assign the "causes and solutions" frame.

Table 3 shows the overall distribution of annotations; the annotations for frame provide only an approximation based on top terms as impacted by the events in our list. We observed that terms marked as unfavorable represent ≈30%-50% of our query terms, and only ≈20%-30% of those are identified as particularly severe (i.e., promoting violence or intimidating); corresponding to 15% on Twitter and 7% on Reddit. Given the recall-oriented nature of our collection, this supports the observation of Faris et al. (2016), who, using a similar taxonomy, also observed that the incidence of the most severe cases of hate speech is also typically small.

Pre- and Post-filtering. Our estimation method requires a minimum number of messages to produce a meaningful result; hence we filter out terms matching only a small number of messages, which we operationalize through arbitrary thresholds requiring a maximum of at least 30 users or messages per day during the event observation window.

Figure 3 shows an example of impact estimation on the "*evil muslims*" term, displaying the observed series, the control series, and their difference in two separate events. In the figure, the widening confidence interval of the forecast matches the intuition that predictions become less certain as we look further into the (counterfactual) future. In general, after applying this process, we consider there to be effect (increase or decrease) if the 90% confidence interval of the difference between treatment and control does not include zero, which means we consider there is no effect where the 90% confidence interval is too large or centered around zero.

[14]Due to restrictions to the use of our main collection, to annotate samples of tweets we used the 1% sample available at Internet Archive (https://archive.org/details/twitterstream), parsing through over 500 M tweets to locate those matching the query terms.

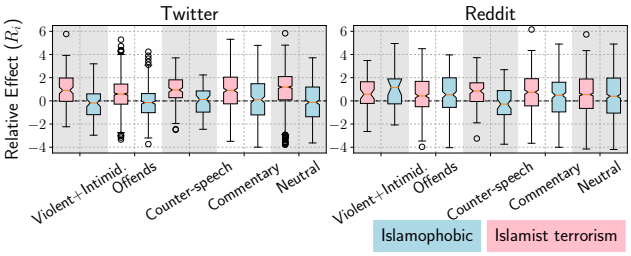

Figure 4: Distribution of impact on severity and stance across platforms (best seen in color). The absolute values of the relative effect on the y-axis are log transformed such that $R_i = sign(Rel_{effect}) \times ln(abs(Rel_{effect}))$.

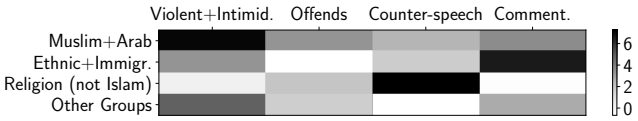

Figure 5: Aggregated variations in the mean relative effect across targets and types of speech on both platforms.

Results and Discussion

In this section, we want to quantify the increase or decrease of various types of speech according to the type of event and platform.

Do distinct types of events impact hate and counter-hate speech in different ways? Prior work found an increase in *hate speech* targeting Muslims following Islamist terrorist attacks (Faris et al. 2016), and our results agree (Twitter (T): +3.0, 95%CI [1.7, 4.4], Reddit (R): +2.9, 95%CI [2.4, 3.3]).[15] Looking at the intersection of high severity categories ("promotes violence" and "intimidates") with the target categories for Muslims and Arabs, we estimate an increase in the relative effects across events in both platforms (T: +10.1, 95%CI [1.4, 18.9], R: +6.2, 95%CI [3.9, 8.4]); also higher than the less severe category (offends or discriminates), with one exception, the 2016 Istanbul Airport attack.

The question is whether Islamophobic attacks elicit a similar, consistent reaction across platforms and events. The answer seems to be no: for instance, we only observe this pattern after one Islamophobic attack (the 2016 Finsbury Park attack), while after the 2017 Olathe Kansas shooting we estimate a decrease in high severity terms in both platforms. This observation is also supported at an aggregate level by Figure 4 (per-event figures omitted for brevity).

Similarly, our estimates indicate an overall increase in *counter-hate speech* terms following Islamist terrorist attacks (T: +1.8, 95%CI [0.7, 3.0], R: +2.9, 95%CI [2.4, 3.4]), but not after Islamophobic attacks. This effect of Islamist terror attacks on counter-speech is consistent with Magdy et al. (2016) who noticed a notable number of messages defending Muslims and Islam following the 2015 Islamist terror attack in Paris.

[15]Throughout the evaluation, these values indicate point estimates for the mean relative effect for each referenced category.

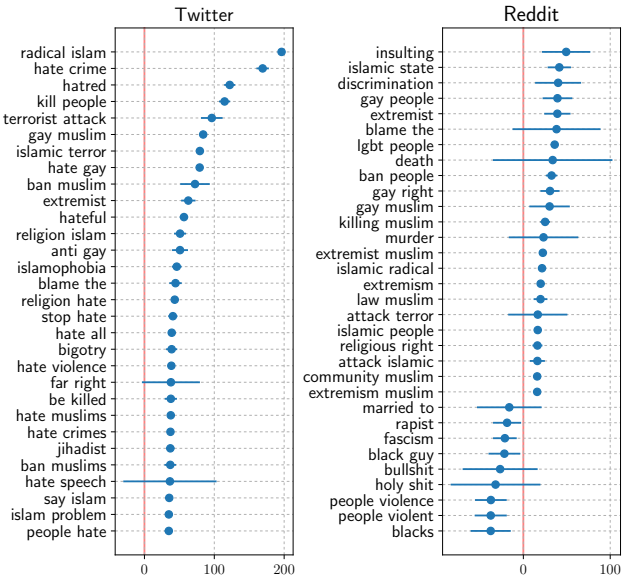

Figure 6: Increase and decrease of top terms after the 2016 Orlando nightclub shooting, where a gay venue was attacked by a self-identified "Islamic soldier."

	Islamist terrorism			Islamophobic		
	Cause	Solution	Both	Cause	Solution	Both
Reddit	64.4%	2.6%	7.5%	64.8%	2.5%	13.3%
Twitter	54.4%	9.2%	25.2%	59.6%	6.7%	22.4%

Table 4: Framing distribution for top terms (percentages do not add to 100% as we omit the "Does not apply" category).

Are these events more likely to lead to an increase in a specific type of speech? Figure 5 suggests that, on average, there is a higher increase in speech that both promotes violence or intimidates and focuses on Muslims and Arabs following extremist violence events; while there is an increase in counter-hate speech related to religion but not specifically mentioning Islam, e.g., focusing on religious tolerance or on positive aspects of religion in general (T: +2.8, 95%CI [2.1, 3.6], R: +7.8, 95%CI [1.9, 14.5]).

At the event-level, Figure 6 showcases an example of a complex interplay between hate and counter-hate speech terms in reference to different groups after the 2016 Orlando nightclub shooting. This was not only a deadly Islamist terrorist incident, but also the deadliest homophobic attack in the U.S., which means it was very prominently covered in media. It triggered a substantial increase in terms referring to both Islam and the gay/LGBT community. In general, our observations agree with an increase in mentions of Muslims, Islam, or Arabs, after Islamist terror attacks; but not after Islamophobic attacks (figures omitted for brevity).

Are there differences in how hate speech is shaped by the events across platforms? For Twitter, Figure 4 suggests an increase for the high-severity categories ("promotes violence" and "intimidates") after Islamic terrorist attacks, but not after Islamophobic attacks. In contrast, for Reddit this distinction is absent, as we see an overall increase after both Islamist terrorist and Islamophobic attacks.

Another aspect in which we see differences between Twitter and Reddit is in terms of the framing of messages, particularly with respect to messages including a "solution or something seen as a solution." In general, the terms that tend to increase the most in this frame call for banning or deporting immigrants/Muslims/Arabs, or waging war against Islam and/or Arabs. As shown in the "solution" and "both" columns in Table 4, this fraction is more prevalent for Twitter among the top 100 most impacted terms (about 34% in Islamist terrorist attacks, about 29% in Islamophobic attacks) than for Reddit (about 10% and 16% respectively).

7 Conclusions

Measuring the effect of external events on hate speech on social media is a challenging task, which needs to be approached with an appropriate methodology (causal inference in our case), and requires a combination of automated processes and manual annotations that balances the needs of large-scale analysis with a finite human annotation budget.

We used data from two social media sites, and from two classes of events (Islamist terrorism and Islamophobic attacks), performing a comparison of observed time series for various classes of online hate speech during such events, with counterfactual series that approximate their behavior had those events not taken place. This allows us to make more precise observations about the evolution of hate (and counter-hate) speech of various classes. Our methodology and observations provide a blueprint for better monitoring hate speech online, with particular attention to the relation between calls for violence online and deadly extremist attacks. Additionally, as we estimate increases in counter-hate speech during these attacks, social media platforms could intervene by boosting its visibility.

Future Work and Limitations. We hope that the evidence of variations in hate speech following certain events will lead to further research to understand why it happens, who it happens to, and what other qualities of an event may explain these variations; as well as research that delves into the source of the differences we observed across platforms.

Further, while our data collection is designed to maximize recall, aiming to provide a good coverage across several categorization dimensions, our bootstrap list of terms can still lead to bias in what gets included in our collections or not. The reliance on query-level annotations may as well introduce noise and biases due to ambiguous uses of some of the terms. We focused on English and only 13 events in the "West," yet future work includes explorations into how our observations may translate to other regions, languages, and type of events. Our frame analysis is also only a first stab at how hateful content is framed after extremist attacks; more in-depth analyses are needed.

Finally, our analysis is retrospective, and harmful content is actively deleted by many social media platforms when reported (Matias et al. 2015), which can result in incomplete data collections. As a result, we are more confident in results indicating an increase in certain types of speech, than on those indicating a decrease.

Reproducibility. The list of our query terms, several example time series, and the detailed instructions used in the crowdsourcing tasks, are available for research purposes at https://github.com/sajao/EventsImpactOnHateSpeech.

Acknowledgments. We thank Miguel Luengo-Oroz for early discussions and anonymous reviewers for their comments. This work was conducted under the auspices of the IBM Science for Social Good initiative. C. Castillo is partially funded by La Caixa project LCF/PR/PR16/11110009.

References

Awan, I., and Zempi, I. 2015. We fear for our lives: Offline and online experiences of anti-Muslim hostility. *Available at: https://tellmamauk.org/fear-lives-offline-online-experiences-anti-muslim-hostility/ [accessed: 18 December, 2017]*.

Benesch, S.; Ruths, D.; Dillon, K. P.; Saleem, H. M.; and Wright, L. 2016. Counterspeech on Twitter: A field study. *A report for Public Safety Canada under the Kanishka Project*.

Benford, R. D., and Snow, D. A. 2000. Framing processes and social movements: An overview and assessment. *Annual review of sociology* 26(1):611–639.

Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; Scott, S. L.; et al. 2015. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics* 9(1):247–274.

Burnap, P., and Williams, M. L. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. *Internet, Policy & Politics*.

Burnap, P.; Williams, M. L.; Sloan, L.; Rana, O.; Housley, W.; Edwards, A.; Knight, V.; Procter, R.; and Voss, A. 2014. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining* 4(1):206.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2018. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In *Proc. of CSCW*.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. *arXiv preprint arXiv:1702.06877*.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. *Proc. of CSCW*.

Committee of Ministers, Council of Europe. 1997. Recommendation no. r (97) 20 of the committee of ministers to member states on"hate speech". https://rm.coe.int/1680505d5b.

Cunha, T.; Weber, I.; and Pappa, G. 2017. A warm welcome matters!: The link between social feedback and weight loss in /r/loseit. In *Proc. of WWW Companion*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of CHI*.

Dièn̈e, D. 2006. Situation of Muslims and Arab peoples in various parts of the world. UN Economic and Social Council, E/CN.4/2006/17.

Duggan, M. 2017. Online harassment 2017. Pew Research Center.

Eichler, M. 2012. Causal inference in time series analysis. *Causality: Statistical perspectives and applications* 327–354.

Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication* 43(4):51–58.

Faris, R.; Ashar, A.; Gasser, U.; and Joo, D. 2016. Understanding harmful speech online. *SSRN:2882824*.

Fulton, C. 2015. Estimating time series models by state space methods in python: Statsmodels.

Gagliardone, I.; Gal, D.; Alves, T.; and Martinez, G. 2015. *Countering online hate speech*. UNESCO Publishing.

Ghanea, N. 2013. Intersectionality and the spectrum of racist hate speech: Proposals to the un committee on the elimination of racial discrimination. *Human Rights Quarterly* 35(4):935–954.

Global Terrorism Database. 2017. Western europe. https://www.start.umd.edu/gtd/search/Results.aspx?chart=overtime&region=8. Accessed: 2017-10-02.

Greven, T. 2016. The rise of right-wing populism in europe and the united states: A comparative perspective. *Freidrich Ebert Stiftung*.

Izsák, R. 2015. Hate speech and incitement to hatred against minorities in the media. UN Humans Rights Council, A/HRC/28/64.

Kuypers, J. A. 2010. Framing analysis from a rhetorical perspective. *Doing news framing analysis: Empirical and theoretical perspectives* 286–311.

Magdy, W.; Darwish, K.; Abokhodair, N.; Rahimi, A.; and Baldwin, T. 2016. #isisisnotislam or #deportallmuslims?: Predicting unspoken views. In *Proc. of WebSci*.

Matias, J. N.; Johnson, A.; Boesel, W. E.; Keegan, B.; Friedman, J.; and DeTar, C. 2015. Reporting, reviewing, and responding to harassment on Twitter. *SSRN:2602018*.

Mohammad, S. H.; Dillon, K. P.; Benesch, S.; and Ruths, D. 2016. A web of hate: Tackling hateful speech in online social spaces. In *Proc. of W. on Text Analytics for Cybersecurity and Online Safety*.

Nockleby, J. T. 2000. Hate speech. *Encyclopedia of the American constitution* 3:1277–1279.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.

Olteanu, A.; Talamadupula, K.; and Varshney, K. R. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proc.of WebSci*.

Olteanu, A.; Varol, O.; and Kıcıman, E. 2016. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. of CSCW*.

Roberts, R. 2017. Hate crime targeting uk mosques more than doubled in past year, figures show. The Independent.

Sellars, A. 2016. Defining hate speech. *SSRN:2882244*.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the targets of hate in online social media. *Proc. of ICWSM*.

TellMAMA. 2017. Anti-muslim hatred, terrorism, media sources, far right networks & spike points. TellMAMA.

UNGP, and UNHCR. 2017. Social media and forced displacement: Big data analytics and machine learning. Technical report, UN.

Wang, J. 2017. Attacks in Western Europe. Reuters Graphics.

Worley, W. 2016. Sun forced to admit '1 in 5 British Muslims' story was 'significantly misleading'. The Independent.

Zhang, Y.; Li, B.; and Hong, J. 2016. Understanding user economic behavior in the city using large-scale geotagged and crowdsourced data. In *Proc. of WWW*.