Monitoring Hate Speech in the US Media

JANUARY 2019

Media and Peacebuilding Project
THE GEORGE WASHINGTON UNIVERSITY

Table of Contents

Abstract

Hate speech targeting minority groups is often associated with acts of political violence. One of the main disseminators of hate speech is the mass media, which continues to draw large audiences in Western states, but also act as a key source of content on social media. This report is based on a research project that aims to create awareness and accountability regarding hate speech by identifying the sources, targets and intensity of hate speech in leading US media political talk/news shows, focusing on the top 10 conservative and top 10 liberal shows by audience size across radio, cable news and YouTube. The study uses a keyword-based automated extraction method to identify potential cases of hate speech, which are then validated by human coders on a novel 6-point hate speech intensity scale.

Research Team

Babak Bahador, Author
Babak Bahador is an Associate Research Professor and Director of the Media and Peacebuilding Project at The George Washington University. He has published and taught in the area of media, peace and conflict for over a decade. He holds a PhD in International Relations from the London School of Economics and Political Science.

Daniel Kerchner, Author
Daniel Kerchner is a Senior Software Developer and Librarian for The George Washington University Libraries. Recent projects include GW ScholarSpace, Social Feed Manager, and others. Dan holds degrees from Cornell University and the University of Virginia, and is a certified Project Management Professional and a certified Software Carpentries/Data Carpentries instructor.

Leah Bacon, Research Assistant
Leah Bacon is pursuing a masters in strategic communication and media at The George Washington University's School of Media and Public Affairs. Leah completed her undergraduate studies in communication and sociology from Boston College.

Amanda Menas, Research Assistant
Amanda Menas is a graduate of The George Washington University with a double major in Political Communication and Human Services/Social Justice. As a research assistant she has transcribed national survey data, developed best practices for social media, and established templates for future grants.

Introduction

Hate speech is defined in this study as negative speech that targets individuals or groups (defined as individuals who share a commonality such as gender, age bracket, ethnicity, nationality, profession, socio-economic status, sexual orientation and religion). Hate speech can take on many forms, including speech, text, images, videos, gestures and other forms of communication. Hate speech is based on the human emotion of hate, which is an enduring dislike involving a loss of empathy and possible desire for harm by the in-group (us) against the targeted out-group (them) (Waltman and Mattheis 2017). There is no internationally accepted definition of hate speech and the term is problematic for many reasons (Bartlett et al., 2014, Benesh 2015, Saleem et al., 2017). In our use of the term, we do not assume that all members of the in-group audience respond in a uniform manner to the speech and assume limited, nuanced effects that vary by individual. Hate speech, therefore, should only be understood as speech that has the potential to increase hate in the recipient audience. Furthermore, even when hate speech increases hate, it does not necessarily mean that the out-group is in greater danger of physical harm as other moral, cultural, political and legal inhibitions can deter a shift from the emotion of hate to the act of violence. Factors that can increase the risk of violence include the speaker's influence, audience susceptibility, the medium and the social/historical context (Benesch 2013, Brown 2016).

Jurisdictions vary in their interpretation of what constitutes hate speech and remedies to address it. Article 20 of the International Covenant on Civil and Political Rights (United Nations) states that "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." In our study, we include a broader range of group categories such as profession/industry (e.g. media/tech) and countries (as an extension of nationality)[1] as such groups are increasingly targeted by hate speech. The notion of hate speech inevitably runs into the principle of free speech, which is a human right that is vital to the healthy functioning of a democracy. In the United States, free speech is protected by the First Amendment of the US Constitution, limiting potential prosecution against speech to cases in which "imminent lawless action" is advocated (Tucker, 2015).

Group hate speech, if successfully sold to the "in-group," can justify collective responses and punishment against all members of the targeted "out-group", who increasingly become viewed as

ENEMIES.

This study only examines hate speech targeted at groups. Such hate speech is significantly more dangerous than hate speech targeting individuals. This is because when a group is associated with extremely negative actions or characteristics, all members of the group become guilty by association, even though only a few, or even none, may be responsible. Group hate speech, if successfully sold to the in-group, can justify collective responses and punishment against all members of the targeted out-group, who increasingly become viewed as enemies. Such speech, in fact, has an established history of use in pre-violence scenarios to justify and socially prepare in-groups for upcoming violence in their name (Bahador 2015, Carruthers 2011, Dower 1986, Keen 1991). According to Sam Keen, "We think others to death and then invent the battle-axe or the ballistic missiles with which to actually kill them" (1991, p.10).

Hate Speech Intensity Scale

While much of the literature on hate speech views the concept as a single category, it is important to acknowledge that variations of intensity are distinguishable. There are clearly differences, for example, between denouncing an action associated with a group versus calling for genocide against a group. To this end, this report introduces a 6-point color-coded hate speech intensity scale. This scale was devised from an extensive content analysis of US news media and a review of the academic literature on hate speech and related topics. The scale is presented in the table below and shows a color, number, title, description and examples for each point on the scale. Importantly, within the description category, a distinction is made between "rhetorical language," which are negative words/phrases associated with a group occurring in the past, present and future and "responses," which describes what the in-group should do (or has done or is doing) against the out-group.

Color	Title	Description	Examples
⬛	6. Death	Rhetoric implies literal killing by group. Responses include the literal death/elimination of a group.	Kill, annihilate, destroy
🟥	5. Violence	Rhetoric implies infliction of physical harm or metaphoric/aspirational physical harm or death. Responses include literal violence or metaphoric/aspirational physical harm or death.	Hurt, rape, starve, torturing, mugging
🟧	4. Demonizing and Dehumanizing	Rhetoric includes sub-human and superhuman characteristics. There are no responses for #4.	Alien, demon, monkey, Nazi, cancer, monster, germ
🟨	3. Negative character	Rhetoric includes non-violent characterizations and insults. There are no responses for #3.	Stupid, aggressor, fake, crazy
🟨	2. Negative actions	Rhetoric includes negative non-violent actions associated with the group. Responses include non-violent actions including metaphors.	Threaten, stop, outrageous behavior, poor treatment, alienate, hope for their defeat
🟩	1. Disagreement	Rhetoric includes disagreeing at the idea/mental level. Responses include challenging groups claims, ideas, beliefs, or trying to change them.	False, incorrect, wrong; challenge, persuade, change minds

At the most basic level (#1) are statements of disagreement, such as indications that the group is wrong, what they claim is false or what they believe is incorrect. In general, this is the mildest form of negative speech directed at groups and deals more with their ideas versus their actions and characteristics. Responses are also at the claim/idea level and deal with rejecting them or trying to persuade and change their position.

The next level (#2) deals with negative non-violent actions associated with the group. Rhetorical words/phrases that are ambiguous on violence (e.g. defeat/stop) are included in #2 and only moved to #5/6 if their context clearly refers to violence. Responses are non-violent responses or independent actions by the in-group towards the out-group. Responses/independent actions that are ambiguous on violence are included in #2 and only moved to # 5/6 if their context clearly refers to violence. This category also includes non-violent negative metaphors.

The next level (#3) refers to non-violent negative characterization or insults. This is worse than #2 (negative non-violent actions), as it suggests that the negativity is an inherent part of the group and less likely to change, whereas actions, being episodic, could be an anomaly not intrinsic to the group's nature. There are no responses in this category as it is not action oriented.

The next level (#4) refers to negative characterizations that are either dehumanizing or demonizing. Dehumanization refers to despised sub-human entities that are considered inferior such as pigs, rats, monkeys and even germs. Demonization involves portraying an enemy as superhuman, such as a monster, robot or even diseases like cancer that are a mortal threat to the in-group's survival. When presented this way, the destruction of the adversary is not only acceptable, but even desirable and beneficial for the in-group and its survival (Bar-Tel 1990; Merskin 2004). Demonization/dehumanization is a particularly extreme type of negative characterization (#3) and a well-established tool for justifying political violence, and thus merits its own category beyond more standard negative characterizations. Like #3, there are no responses in this category as it is not action oriented.

The next 2 levels (#5 and #6) represent the worst hate speech, as they refer to violence, either allocated to the group's actions or about what should be done to the group (responses). Category #5 refers to literal violence that is non-lethal, either based on past/current/future actions, or metaphorical/aspirational violence that is either nonlethal or lethal. While literal references to violence can certainly increase hate and support for violence, research shows that even metaphorical violence can increase support for violence (Kalmoe 2014). Category #6 refers to literal violence that is lethal, either based on past/current/future actions, or responses that call for literal death/destruction.

Methods

Sources

This project aims to track the sources, targets and intensity of hate speech across the US political talk/news shows, especially regarding shows that have daily audiences of at least 1 million listeners/followers. To this end, we identified the top 10 conservative and liberal media shows based around personalities across cable news, radio and YouTube (Katz, 2018; Talkers; YouTube). While all the 20 shows identified had over 1 million daily viewers[2], the top personality/show belonged to Sean Hannity, with a combined daily audience of just over 16 million. The following tables display the media figures, their shows, their platforms and their audience size. The first table displays the conservative shows and the second the liberal shows:

Leading Conservative Political Talk/News Shows

Figure	Show	Platform	Audience Size (000s)
Sean Hannity	Hannity (TV) and The Sean Hannity Show (Radio)	Fox (TV) and Premiere	TV: 2,663 Radio: 13,500
Rush Limbaugh	The Rush Limbaugh Show	Premiere	14,000
Michael Savage	The Savage Nation	Westwood One	11,000
Glenn Beck	The Glenn Beck Program	Premiere	10,500
Laura Ingraham	The Ingraham Angle (TV), The Laura Ingraham Show (Radio)	Fox	TV: 2,155 Radio: 8,000
Mark Levin	The Mark Levin Show	Westwood One	10,000
Alex Jones	Infowars	YouTube	5,900
Joe Pags	The Joe Pags Show	The Joe Pags Show	4,000
Tucker Carlson	Tucker Carlson Tonight	Fox	2,223
Bret Baier	Special Report with Bret Baier	Fox	1,782

Leading Liberal Political Talk/News Shows

Figure	Show	Platform	Audience size (000s)
Thom Hartmann	The Thom Hartmann Program	WYM Media	6,250
Stephanie Miller	The Stephanie Miller Show	WYM Media	5,750
Cenk Uygur & Ana Kasparian	The Young Turks	YouTube	4,200
Rachel Maddow	The Rachel Maddow Show	MSNBC	2,403
Lawrence O'Donnell	Last Word with O'Donnell	MSNBC	1,924
Chris Hayes	All In With Chris Hayes	MSNBC	1,579
Chris Matthews	Hardball with Chris Matthews	MSNBC	1,420
Brian Williams	11th Hour with Brian Williams	MSNBC	1,457
Ari Melber	The Beat with Ari Melber	MSNBC	1,358
Chuck Todd	Meet the Press Daily	MSNBC	1,223

For each show, transcripts were gathered for the first seven weekdays of June 2018 (June 1st to June 11th) and turned into PDFs. Only the first hour of each show was included in the study. For cable news, transcripts were gathered from Lexis Nexis. For radio, shows were gathered either from YouTube or recorded from podcasts/show websites etc. and transcribed via YouTube closed captioning. YouTube show transcripts were gathered from YouTube closed captioning. Shows were reviewed to ensure that transcripts were accurate.

Units of Analysis

To identify potential examples of hate speech (based on the 6-point scale), two dictionaries were created. The first called "Subjects" identified groups there were potential targets of hate speech. The second called "Keywords" identified negative words and phrases that could be applied to the subjects. Subjects and keywords dictionaries were created through an extensive review of the transcripts gathered and subsequent online research. Initially, approximately 200 subjects and 4,000 keywords were identified. However, during the coding process, more groups and keywords were added, bringing the total at the end of the study period to 289 groups and 5,451 keywords. Some terms used for groups were similar, allowing for them to be combined into larger umbrella groups. For example, terms such immigrant, refugee, migrant, alien and dreamer could potentially fit into one larger umbrella group. Overall, 69 umbrella groups were identified. A Unit(s) of analysis (UOA) for the purpose of this study is when a negative keyword applies (is relevant) to a subject.

Semi-automated Extraction and Identification of UOA

To facilitate and automate key aspects of the work of discovering instances of hate speech in the text of show transcripts, the project team partnered with software developer librarians[3] from the George Washington University Libraries. The tool developed in this collaboration searches through transcripts and identifies instances of keywords, subjects and co-located pairs of keywords and subjects. The human coder then analyzes these instances to determine and notate if they are semantically relevant within the context of the transcript. The result is a CSV file where instances of potential hate speech UOA are data observations recorded in rows, and characteristics of the instances are in columns associated with variables. This "tidy data" format (Wickham, 2014) is ideal for facilitating further exploration and quantitative analysis of the data.

The software tool is a Python program that expects the following inputs:

1. PDF file(s) containing transcripts.
2. A CSV-formatted "keywords" file containing a list of keywords; accompanying each keyword is an intensity score and a group identifier, both assigned by analysts. Multi-word keywords are represented in normalized format (e.g. witch_hunt rather than witch hunt; see #4 below).
3. A CSV-formatted "subjects" file containing a list of subjects; accompanying each subject is a group identifier. Multi-word subjects are represented in normalized format (e.g. low_income rather than low income; see #4 below).
4. A CSV-formatted "normalize_terms" file containing a list of multi-word terms; for each term is an equivalent term with underscore ("_") characters substituted for spaces. An example would be gun owners/gun_owners.

The tool is run at the command line. There are also several parameters which can optionally be modified from their default values, including:

- window: The maximum number of words apart (for this analysis, a window value of 5 words was used).
- context: The number of words before and after the window to extract in order to provide context for the coder/analyst.

The program's algorithm can be summarized as the following series of steps:

1. Use the "pdfminer" Python library, specifically the pdfminer.six fork[4], to extract text from the transcript PDF files.
2. Tokenize the text (i.e. break it into discrete words).
3. Normalize multi-word terms, by replacing any instances of multi-word terms in the text with their single-token equivalents.

4. Use a sliding "window" to search for instances of keywords, subjects, and co-located keywords and subjects.

To elaborate on the search algorithm in step 4, for each phrase evaluated, the code executes the following algorithm:

If the left-most word matches a word in the subject words list:

- Check for a keyword (i.e. a word matching a word on the keywords list) in the phrase, to the right.
    - If there are no keywords, write out a row with just the subject word.
    - If there is a keyword, write out a row with both the subject and keyword word.

If the right-most word matches a word in the subject words list:

- Check for a keyword in the phrase, to the left.
    - If a keyword is found, write out a row with both the subject and keyword word.

As an example, if our window size is 5 words and we have the following text:

we had a bunch of globalist vampires drinking our blood

then the algorithm will review phrases in the following sequence:

```
we had
we had a
we had a bunch
we had a bunch of
   had a bunch of globalist
       a bunch of globalist vampires
         bunch of globalist vampires drinking
               of globalist vampires drinking our
                  globalist vampires drinking our blood
                            vampires drinking our blood
                                     drinking our blood
                                              our blood
```

In the above example, "globalist" is a word on the subject list, and "vampires" is a word on the keywords list, so the program will write out a row documenting an instance of these words being found together.

Data associated with each observation is written to the output CSV file to facilitate analysis, as follows:

extract date	The date/time the program was run
file	The file name of the PDF where the instance was found
show_date	The date of the show (extracted from the file naming schema)
show_id	The show ID (extracted from the file naming schema)
subject	The subject word found
subject_code	The subject code (used to group similar subject words together)
keyword	The keyword found (if present)
keyword_code	The keyword code (the intensity score assigned to the keyword, from the keywords file)
keyword_id	The keyword ID (used to group similar keywords together)
relevant?	A blank column to be used by the human coder to mark whether the match is semantically relevant in context
extract	A larger section of text around the window is provided, to provide the human coder with context

The CSV file becomes the basis for the coding process described in the next section.

A higher rate of discovery of terms may be achievable by further cleaning the text after it is extracted from the PDF files. The current method of creating PDFs of the show transcripts often results in extraneous markings (such as timestamps) between words as well as other features which can result in two or more words concatenated in a way such that they are not broken apart by the tokenizer. By identifying some common patterns among content that is not germane to the transcript text, researchers may be able to clean (i.e. remove) some text which interferes with tokenization. Additionally, exploring improved approaches to creating the PDFs may also lead to extracted text which is more accurately tokenized.

The code and instructions for the Python program are publicly available on Github at https://github.com/gwu-libraries/vopd. Version 1.0 of the code[5] was used for this analysis.

Coding Process

The CSV file contained all extracted examples from the transcripts in which a subject and keyword were co-located within 5 words. When this occurred, twenty words from both sides of the co-located words were gathered and presented as a potential UOA. Coders then either accepted, rejected or added in relation to these co-located words. Accepting meant that the keyword applied to the subject. In such cases, the coder also entered a number on the 6-point hate speech scale and decided if it was an example of rhetoric (designated by A) or response (designated by B). If the coder rejected the co-located words, this would not count as a UOA. Finally, the coder could add up to 3 additional UOA from the text if new cases of hate speech (keywords applying to subjects) were found in the same text. This could involve adding new keywords that applied to identified subjects, new subjects that applied to identified keyword or new subjects and keywords that were not identified in the automated extraction. When new keywords and subjects were found, they were added to the dictionaries.

For the purposes of this study, coders added new words to the dictionary daily over the 7-day period of this study. The updated dictionary was then used for finding potential UOA the next day for each day of the 7-day duration of the study. Then, a second round of searches was conducted using only the updated dictionaries for each day, revealing more new UOA each day. The totals from both rounds of coding were included in the final data findings.

Intercoder Reliability

To test for intercoder reliability, two coders separately coded 1,321 units generated by the system from the June 4th and 5th data. This followed an extensive effort to fine tune the code book by coding the data from June 1st collectively and clarifying coding instructions to remove ambiguities. In total, each coder had 11 choices to make for each text generated by the system, as outlined below:

Option #	Definition
1	Rejected potential UOA presented and could not find new UOA in text
2	Accept UOA presented and/or add new UOA; coded 1A - 1 intensity, rhetoric (A)
3	Accept UOA presented and/or add new UOA; coded 1B - 1 intensity, response (B)
4	Accept UOA presented and/or add new UOA; coded 2A - 2 intensity, rhetoric (A)
5	Accept UOA presented and/or add new UOA; coded 2B - 2 intensity, response (B)
6	Accept UOA presented and/or add new UOA; coded 3A - 3 intensity, rhetoric (A)
7	Accept UOA presented and/or add new UOA; coded 4A - 4 intensity, rhetoric (A)
8	Accept UOA presented and/or add new UOA; coded 5A - 5 intensity, rhetoric (A)
9	Accept UOA presented and/or add new UOA; coded 5B - 5 intensity, response (B)
10	Accept UOA presented and/or add new UOA; coded 6A - 6 intensity, rhetoric (A)
11	Accept UOA presented and/or add new UOA; coded 6B - 6 intensity, response (B)

Using the ReCal online intercoder reliability web service (Freelon 2010), we found percentage agreement at 90.1% and Scott's Pi, Cohen's Kappa and Krippendorff's Alpha coefficients all at 0.781, which at above 0.7, indicates reliable data findings (Lombard et al, 2002).

Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha(nominal)	N Agreements	N Disagreements	N Cases
90.1%	0.781	0.781	0.781	1,190	131	1,321

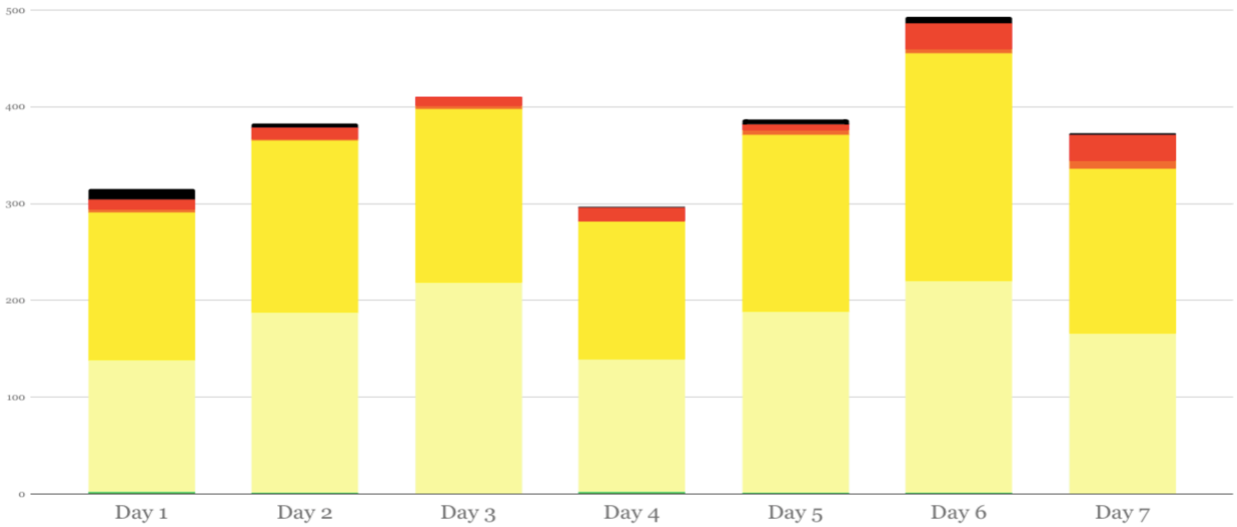
Findings

Over the 7-day period of this study, the system identified 11,390 potential units of analysis. Human coders, however, found that only 1,047 (9.2%) of these were relevant UOA (called "system relevant"), while 90.8% were not. However, upon reviewing the text pulled up by the system, coders found an additional 1,611 relevant UOA (called "Added Relevant"), for a total of 2,658 UOA. The table below shows the total UOA for this study over the study timeline.

Date	System Generated	System Relevant	Added Relevant	Total UOA
June 1	1,647	172	143	315
June 4	1,616	119	264	383
June 5	1,798	152	258	410
June 6	1,351	141	156	297
June 7	1,671	175	212	387
June 8	1,713	169	324	493
June 11	1,594	119	254	373
Total	11,390	1,047	1,611	2,658

Daily Hate Speech Monitoring

One of the long-term goals of this project is to identify the volume and intensity of hate speech in the US media on a daily basis, with the goal of producing automatically-generated reports each day. This could act like a thermometer on the level of hate in the country and as a possible early-warning signal if subsequent research can correlate hate speech intensity with hate crimes and even political violence. The following chart and table highlight our findings over the 7-day study period.



Date	1	2	3	4	5	6	Total/Day
Day 1	2	136	153	3	10	11	315
	0.6%	43.2%	48.6%	1.0%	3.2%	3.5%	
Day 2	1	186	178	2	12	4	383
	0.3%	48.6%	46.5%	0.5%	3.1%	1.0%	
Day 3	0	218	180	3	9	0	410
	0.0%	53.2%	43.9%	0.7%	2.2%	0.0%	
Day 4	2	137	143	0	14	1	297
	0.7%	46.1%	48.1%	0.0%	4.7%	0.3%	
Day 5	1	187	183	5	6	5	387
	0.3%	48.3%	47.3%	1.3%	1.6%	1.3%	
Day 6	1	219	235	5	26	7	493
	0.2%	44.4%	47.7%	1.0%	5.3%	1.4%	
Day 7	0	166	170	8	27	2	373
	0.0%	44.5%	45.6%	2.1%	7.2%	0.5%	

Targets of Hate Speech

Another goal of this project is to identify the targets of hate speech. To this end, the following table columns show the top 10 targets of hate speech over the 7-day period on a daily basis and for the total 7-day period (Total). The table also shows the average score (on the 6-point scale) for each day and for the 7-day period and for each of the top 10 categories. The two columns on the far right of the table show the number (#) of UOA allocated to each of the top 10 groups and the percentage (%) each represented from total UOA. For example, the media was the top target over the 7-day period with 744 UOA, which represent 28.0% of the total UOA in the study. Definitions for each target group are presented in the Appendix.

	Day 1 (2.7)	**Day 2 (2.6)**	**Day 3 (2.5)**	**Day 4 (2.6)**	**Day 5 (2.6)**	**Day 6 (2.7)**	**Day 7 (2.7)**	**Total (2.6)**	**#**	**%**
1	Russia	Media	Media	Media	Media	Media	Media	Media (2.4)	744	28.0%
2	Media	Russia	Russia	USG	USG	Russia	NKorea	Russia (2.2)	428	16.1%
3	NKorea	USG	USG	Russia	Russia	USG	China	USG (2.6)	298	11.2%
4	USG	Asian	Athlete	Migrant	Migrant	China	Russia	NKorea (2.6)	123	4.6%
5	Elite	Migrant	Tech	America	China	Migrant	Elite	China (2.5)	121	4.6%
6	China	Elite	America	Law	Foreign	Elite	America	Migrant (3.0)	100	3.8%
7	Latinx	Law	Elite	CA	America	Edu.	Asian	Elite (3.0)	91	3.4%
8	Foreign	Foreign	Muslim	Athlete	Women	Corp.	Women	America (2.5)	78	2.9%
9	America	America	Women	Elite	Edu.	Foreign	Migrant	Foreign (3.0)	55	2.1%
10	Comic	Muslim	Migrant	Women	NKorea	Latinx	Canada	Women (2.3)	53	2.0%
								Other (2.6)	567	21.3%

Targets of Hate Speech by Conservative Shows

	Day 1 (2.6)	Day 2 (2.6)	Day 3 (2.5)	Day 4 (2.6)	Day 5 (2.6)	Day 6 (2.7)	Day 7(2.7)
1	Media	Media	Media	Media	Media	Media	Media
2	Russia	USG	Russia	USG	USG	Russia	China
3	USG	Russia	USG	Migrant	Russia	USG	NKorea
4	Elite	Elite	Tech	America	Migrant	China	Russia
5	Latinx	Migrant	Athlete	Russia	China	Migrant	Elite
6	America	Muslim	America	CA	Women	Elite	America
7	China	Asian	Elite	Women	Edu.	Edu.	Women
8	NKorea	CA	Women	Law	Elite	Foreign	Migrant
9	Tech	America	Muslim	Athlete	Foreign	Canada	Asian
10	Law	Tech	LGBT+	Elite	Latinx	Latinx	Tech

Total Seven-Day Ranking	UOA	% of Total UOA	Intensity Scale Average
Media	630	33.8%	2.4
Russia	208	11.2%	2.2
USG	190	10.2%	2.2
China	94	5.0%	2.6
Migrants	92	4.9%	2.7
Elite	86	4.6%	2.5
America	55	3.0%	3.2
Women	49	2.6%	3.0
NKorea	38	2.0%	2.8
Tech	36	1.9%	2.4
Other	385	20.7%	2.6
Total	**1,863**	**100.0%**	**2.6**

Targets of Hate Speech by Liberal Shows

	Day 1 (2.8)	Day 2 (2.4)	Day 3 (2.5)	Day 4 (2.6)	Day 5 (2.4)	Day 6 (2.7)	Day 7 (2.8)
1	NKorea	Russia	Russia	Media	USG	Russia	NKorea
2	Russia	Media	USG	Russia	Media	USG	Media
3	Media	USG	Athlete	USG	Russia	Media	China
4	Foreign	Law	Foreign	Law	America	Corp.	Russia
5	USG	Foreign	Law	America	Foreign	China	Asian
6	China	Asian	White	Foreign	NKorea	NKorea	Canada
7	Comics	America	Media	Black	Asian	Ukraine	USG
8	Men	China	Tech	Asian	Migrant	Foreign	Foreign
9	Women	Migrant	NKorea	Christian	EU	America	Ukraine
10	Corp.	White	America	German	Christian	Asian	America

Total 7-Day Ranking	UOA	% of Total UOA	Intensity Scale Average
Russia	220	27.7%	2.3
Media	114	14.3%	2.7
USG	109	13.7%	2.5
NKorea	85	10.7%	2.6
Foreign	39	4.9%	2.6
China	27	3.4%	2.5
Law	24	3.0%	3.4
America	23	2.9%	2.5
Asian	17	2.1%	3.1
Corp.	16	2.0%	2.3
Other	121	15.2%	2.5
Total	**795**	**100.0%**	**2.6**

Media Subject Category Analysis

As mentioned, the highest volume of hate speech was targeted at the media group over the period of this study. For our analysis, the term “media” itself was the most frequently mentioned at 54.7% of the total 744 UOA. This was followed by “news” and “CNN,” as we included individual media outlets in this category.

Target	UOA	Percentage
Media	407	54.7%
News	118	15.8%
CNN	41	5.5%
Hollywood	20	2.6%
Journalists/ Journalism	20	2.6%
Press	20	2.6%
MSNBC	18	2.4%
Fox	17	2.2%
New York Times	12	1.6%
Media Matters	11	1.4%
Other	60	8.0%
Total	**744**	**100.0%**

The following bullet points show examples for the media group on the 6-point hate speech intensity scale, with keywords in blue and subjects in red:

- June 6, 2018, Hannity, “Obama’s justice department and FBI--well mishandled the criminal case into Hillary Clinton. In other words, we were right and your **media** was **wrong**. **- Code 1**
- June 1, 2018, The Beat, "He will tell a big tale and a big story, and so now the **media** is **conflating** that with absolute lying.” - **Code 2**
- June 7, 2018, Ingraham Angle, “we are always fair and balanced, we are not the **destroy-Trump** **media**. let not your heart be troubled.” - **Code 3**
- June 1, 2018, Savage Nation, “the bible is the most offensive book in the world to the **vermin** in the **media**” **- Code 4**
- June 5, 2018, Stephanie Miller show, “Sean Hannity demands every honest patriot take to the streets, right-wing **media** **calls for war** and insurrection” - **Code 5**
- There were no examples of #6 for the media category. However, this is an example from another group: June 4, 2018, Infowars, “Bernie supported **black lives matter**, the **cop killers**, when they killed cops that's the type of monster a cold-blooded person that Sanders is”**- Code 6**

Future Research

With the abundance of political talk shows in the US, there is no shortage of potential data to analyze. However, the semi-automated human coding effort required to analyze just 7 days worth of 20 shows was formidable. Our goal is to implement machine learning techniques and a way to automate classification of UOA. Existing natural language processing (NLP) classifiers trained on English-language text may have the semantic and syntactic knowledge to be able to classify a UOA, answering the coder's question as to whether, in the context of the sentence, speaker is intending to associate the keyword with the subject. Ideally, a classifier would be able to classify each potential UOA generated by the system using one of the 11 codes, but initial work may be best started on a simple, binary scheme to determine whether a UOA is or is not an instance of rhetoric/response at all.

> This could act like a ***thermometer*** on the level of hate in the country, and as a possible **EARLY WARNING SIGNAL...**

As the goal of monitoring these media sources is more one of observing trends over time, and of comparing sources relative to each other, as long as the same classifier is applied to all of the sources, a less-than perfect rate of missed instances will still yield meaningful results. Our current methodology only captures a sample of more obvious cases of hate speech and likely misses more subtle forms that require a deeper reading. This limitation will improve over time as we continue to build the two dictionaries. However, we also plan to look for additional means of capture to complement the existing approach to gather larger samples over time.

Finally, we aim to broaden our research focus in two ways. First, we plan to expand our subjects to include rival political groups (Liberal, Conservative, Republican, Democrat etc.), which are currently excluded from our study. In observing the media content, it is clear that much of the hate in political talk/news is targeted at rival political groups. This expansion could meaningfully contribute to the understanding of rhetorical political polarization in the United States and its intensification.

Second, we aim to expand our sources to eventually include any media program with at least 1 million followers/listeners etc. and all influential US politicians, such as members of Congress. We also plan to capture a wider range of texts including political speeches and social media messages from these same key media and political sources.

In our hyper-mediated world, those who hold large audiences have a responsibility to avoid building hate amongst their followers towards various minority and rival political groups. The goal of our project is to identify and display instances and trends in group-targeted hate – especially the most severe kind involving demonization, dehumanization and violence advocacy – to foster awareness, accountability and de-escalation.

Notes

[1] The inclusion of countries as subjects posed a dilemma for our research team because the intent of the media source when mentioning a country is often to explain the negative actions of the state, not to negatively describe the nation and its people. However, our research focused on the likely impact on the source's audience (in-group) and not the intent of the source. To this end, research shows that public sentiment towards other countries follows their media framing (Brewer, Joseph and Willnat 2003). As such, even when sources do not intend to build hate for a country by mentioning them in a negative way, they are likely inadvertently contributing to building negative public sentiment. It is important to note that we only included UOA in our data when the country or its people (Russia, Russians) was mentioned alone. When the state or its leadership was mentioned (e.g. Russian government, Putin), it was not included as a UOA.

[2] While the data on daily viewers/listeners is more robust for radio and cable news, it is difficult to know exact daily viewers for YouTube shows. We did, however, include the top show on the left (The Young Turks) and right (Alex Jones' Infowars, which was available at the time of this study, but removed from YouTube on August 6, 2018).

[3] The authors wish to attribute credit to Justin Littman, who produced the initial working version of the code, including much of the core logic upon which the code still relies.

[4] https://github.com/pdfminer/pdfminer.six

[5] DOI: 10.5281/zenodo.1482912

References

Bahador, B. (2015) The Media and Deconstruction of the Enemy Image. In V. Hawkins and L. Hoffmann (eds), *Communication and Peace: Mapping an emerging field*. New York: Routledge, 120-132.

Bar-Tel, D. (1990) Causes and Consequences of Delegitimization: Models of Conflict and Ethnocentrism. *Journal of Social Issues* 46 (1): 65-81.

Bartlett, J., Reffin, J., Rumball, N., and Williamson, S. (2014). *Anti-social media*. Demos: 1-51. At: https://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf?1391774638 (accessed November 1, 2018).

Benesch, S. (2013). Dangerous Speech: A Proposal to Prevent Group Violence. *Dangerous Speech Project.* At: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf. (accessed December 18, 2018).

Benesh, S. (2014) Defining and diminishing hate speech. State of the World's Minorities and Indigenous Peoples 2014. Minority Rights International: 18-25. At: https://minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities-2014-chapter02.pdf (accessed November 1, 2018)

Brewer, P.R., G. Joseph and L. Willnat (2003) Priming or Framing: Media Influence on Attitudes Toward Foreign Countries. *International Communication Gazette* 65(6): 493-508.

Brown, R. (2016) *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech.* At: https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf (accessed December 18, 2018).

Carruthers, S. (2011) *The Media and War*. Basingstoke, UK: Palgrave MacMillan.

Dower, J.W. (1986) *War Without Mercy: Race and Power in the Pacific War*. New York: Pantheon Books.

Freelon, D. (2010) ReCal: Intercoder reliability calculation as a web service. International Journal of Internet Science, 5(1): 20-33.

Katz, A.J. (2018) Q3 2018 Ratings: Fox News Marks 67 Straight Quarters as No.1 Cable News Network; Hannity Becomes No.1 Basic Cable Series. *Adweek: TVNewser.* October 2. At: https://www.adweek.com/tvnewser/q3-2018-ratings-fox-news-marks-67-straight-quarters-as-most-watched-cable-news-network-hannity-becomes-no-1-basic-cable-series/378271 (accessed November 5, 2018).

Keen, S. (1991) *Faces of the Enemy: Reflections on the Hostile Imagination*. San Francisco: Harper & Row.

Lombard M., Snyder-Duch J., and Bracken C.C. (2002) Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research* 28 (4): 587-604.

Merskin, D. (2004) The Construction of Arabs as Enemies: Post-September 11 Discourse of George W. Bush. *Mass Communication and Society* 7(2): 157-175.

Nathan, P.K. (2014) Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence, *Political Communication*, 31 (4): 545-563.

Tucker, E. (2015) How federal law draws a line between freedom of speech and hate crimes. PBS News Hour. December 31. At: https://www.pbs.org/newshour/nation/how-federal-law-draws-a-line-between-free-speech-and-hate-crimes (accessed November 2, 2018)

Saleem, H.M., Dillon, K.P., Benesch, S., and Ruths, D. (2017) A Web of Hate: Tackling Hateful Speech in Online Social Spaces. At: https://dangerousspeech.org/a-web-of-hate-tackling-hateful-speech-in-online-social-spaces/
(accessed November 1, 2018).

United Nations. (1966) International Covenant on Civil and Political Rights. At: https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx (accessed October 30, 2018)

Talkers (2018) Top Talk Audiences. At: http://www.talkers.com/top-talk-audiences/ (accessed November 5, 2018).

Waltman, M. S., & Mattheis, A. A.(2017). Understanding hate speech. In *Oxford encyclopedia of communication*. Oxford, England: Oxford University Press.

Wickham, H. (2014) Tidy Data. Journal of Statistical Software, 59(10).

YouTube (n.d.) News and Politics Channels. At: https://www.youtube.com/channels/news_politics (accessed November 5, 2018).

Appendix

Legend	
Word/Symbol	**Definitions**
Asian	Asian, Japanese, South Korean, Philippines
Athletes	Eagles, Players, Coaches, Football
Corp.	Banks, Corporations, JPMorgan, Lenders
Black	African Americans, Black Lives Matter, Blacks, Black Women
CA	California
Canada	Canadians, Canada
China	China, Chinese
Christian	Bible Thumper, Christian, Christians
Comic	Comedians
Edu.	Universities, Students, College
Elite	Elites, Globalists, Establishment, Bureaucrat
EU	Europe, European Union, European Leaders
Foreign	Foreign entities
German	German, Germans, Germany
Latinx	Latinos, Hispanics, Mexico, Guatemalans
Law	Cops, Judges, Law Enforcement, Lawyers, Police
LGBT+	Gay, Homosexual, Transgender
Media	Media, News, CNN, MSNBC, Fox News
Men	Men, Patriarchy
Migrant	Immigrant, Refugee, Immigrants, Alien
Muslim	Muslims, Hijab
NKorea	North Korea, North Korean, North Koreans
Russia	Russians, Moscow, Kremlin
Tech	Tech Companies, Apple, Twitter, Google

Ukraine	Ukraine, Ukranian
USG	US Government, Federal Agencies
White	Whites, White Men
Women	Feminists, Females, Woman, Girl