Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit

Ted Grover, Gloria Mark
University of California, Irvine
Department of Informatics
grovere@uci.edu, gmark@uci.edu

Abstract

Over the past few years, new ideological movements like the Alt-Right have captured the attention and concern of both mainstream media, policy makers, and scholars alike. Today, the methods by which right-wing extremists are radicalized are increasingly taking place within social media platforms and online communities. However, no research has yet investigated methods for proactively detecting online communities that may be displaying overall warning signs of mass ongoing ideological and political radicalization. In our work, we use a variety of text analysis methods to investigate the behavioral patterns of a radical right-wing community on Reddit (r/altright) over a 6-month period until right before it was banned for violation of Reddit terms of service. We find that this community showed aggregated behavioral patterns that aligned with past literature on warning behaviors of individual extremists in online environments, and that these behavioral patterns were not seen in a comparison group of eight other online political communities, similar in size and user engagement. Our research helps build upon the established literature on the detection of extremism in online environments, and has implications for proactive monitoring of online communities.

Introduction

Over the past ten years, instances of violent extremism and terrorism across the world have more than doubled, and the number of fatalities from acts of terrorism have nearly tripled (Roser et al. 2018). Although the majority over this period have occurred in the middle east, a substantial proportion of these attacks have occurred within the United States. From 2016 to 2017, the number of annual domestic terrorist attacks within the U.S. nearly tripled, particularly from attackers motivated by right wing ideologies (Lowery, Kindy, and Tran 2018). Governments and scholars have both noted the threats posed by growing domestic right-wing extremism across North America.

Today, the routes by which right wing extremists are being radicalized is evolving, and are increasingly taking place within social media platforms and online communities. For instance, research has shown how online underground communities like the /pol/ forum on 4Chan have contributed to the growth and propagation of the alt-right and other extreme right-wing ideologies (Hine et al. 2017; Zannettou et al. 2017). In today's digital age, the ability to proactively detect signs of ideological radicalization and extremism within online communities is an open research area with important societal consequences.

Although past work has already shown the feasibility of detecting individual extremists in online environments (Brynielsson et al. 2013; Johansson, Kaati, and Sahlgren 2017), no research has yet investigated if and how theories and methods of detecting potentially dangerous individuals could be refactored and applied to the aggregated text of online communities. These can aid human judgement in detecting and distinguishing groups and communities suspected of becoming increasingly ideologically radical.

The contribution of this study is as follows. Using past established methods of detecting 'warning behaviors' of individual radicalization and violence as a theoretical basis, we investigate if these same warning behaviors appear in the aggregated language use of an online community, associated with extreme right-wing ideology. By employing various text analysis methods, we show that this online community presented many behavioral patterns that align quite seamlessly with established theories of warning behaviors of individual extremists. In doing so, we present initial evidence that the aggregated behavioral patterns of radical political communities may mirror the known psychopathology of individual extremists in online environments. We present the implications of our work for proactive online community monitoring.

Theory

In our work, we draw upon and employ existing theoretical frameworks to detect behavioral patterns that may be indicative of ideological radicalization and extremism in online environments. We focus on warning behaviors that can be readily analyzed and quantified with textual analysis of social media data (Cohen et al. 2014).

Warning signs of Ideological Violence

Past work has aimed to identify behavioral markers of radicalization, deemed 'warning behaviors', to assess the individual threats of of terrorism or violence (ie. 'lone wolf terrorism') (Cohen et al. 2014; Meloy et al. 2012). In the context of lone wolf terrorism, warning behaviors can be defined as distinct behaviors that "precede an act of targeted violence, is related to it, and may, in certain cases, predict it" (Cohen et al. 2014, p.248). Eight distinct warning behaviors have been identified (Meloy et al. 2012), three of which have been shown to be realistically detected through text analysis methods, *fixation*, *identification*, and *leakage*, described next.

Fixation is defined as behavior indicating an 'increasingly pathological preoccupation with a person or cause' (Cohen et al. 2014, p.249). Important characteristics of fixation behaviors relevant and applicable for text analysis (Cohen et al. 2014) are an: (1) an increasing perseveration on the person or cause, (2) increasingly negative accounts of the object of fixation, and (3) increasingly strident opinions and angry emotional undertones.

Identification is comprised of three sub-categories of distinct behaviors which include identification with radical action (ie. framing oneself as a 'hero' or 'warrior'), identification with a role model (e.g. a past lone-wolf terrorist), and identification with, and moral obligation towards their ideological in-group (ie. an us vs. them mentality) (Cohen et al. 2014).

Leakage involves the declaration of intent to do harm to a desired target, including the planning, research, or implementation of an attack (Cohen et al. 2014). This intent may or may not be specific, and can vary in directness. Leakage usually entails preoccupation with the target.

Related Work

Using this theory of warning behaviors for radicalization as a theoretical foundation to draw from, we report on literature of these theories in practice, as well as of the alt-right, reddit communities, and automated detection of hate speech. We identify a gap in research, namely how these theories can be applied to detect potential warning signs in ideologically radical and potentially dangerous online communities.

Detection of Possible Terrorists Online

A major application of the theory of warning behaviors is to proactively detect possible violent individuals and 'lone-wolf terrorists' (Cohen et al. 2014). These theories have been used to develop tools and protocols to detect and track possible violent extremists in online environments (e.g. social media and forums), e.g. an end-to-end tool called the 'Impactorium' that employed warning behavior detection (fixation, identification, and leakage), to aid in analysis of extremist forums by assigning 'threat' scores to individuals and monitoring their activity (Brynelsson et al. 2013). Scrivens, Davies, and Frank (2018) built a replicable system to identify radical authors in Islamist forums using some principles from the theories on warning behaviors. Although past work has already shown the feasibility of detecting individual extremists in environments already known to contain content related to violent extremism, no research has investigated if these theories could be applied at a higher-level analysis, to proactively detect and distinguish if online communities show signs of ongoing ideological radicalization.

Increasingly more common, people form and participate in online communities united around shared ideological or political aims. Research has shown that the online social networks of individuals on platforms like Twitter show strong political homiphily, across the political spectrum (Halberstam and Knight 2016). Although mostly benign places for people to discuss current events, some online communities have emerged as spaces that can advertently or inadvertently radicalize their members. Past research has demonstrated the Islamic State (ISIS) actively employ Facebook and Twitter to spread propaganda, and recruit and radicalize its young members who seek community and purpose (Blaker 2015). Online underground communities like the 'politically incorrect' forum on 4Chan (/pol/) has contributed to the growth and propagation of the alt-right and other extreme right-wing ideologies (Hine et al. 2017; Zannettou, 2017).

We are interested in discovering methods for proactively detecting signs of ideological radicalization within online communities, ideally before violent action can occur in the real world. Recent research has investigated using patterns of inter-community conflict on Reddit to identify novel behavioral patterns of problematic subreddits that have been previously banned (Datta and Adar 2018), and to proactively identify communities that may instigate conflicts (Kumar et al. 2018). In our work, we are interested in extending this line of research by examining if the psycholinguistic patterns *within* an online community could be an additional indicator of a community becoming increasingly radical. To do so, we present a case study of a prominent online community that, during its existence, was a hub for extreme right-wing ideology: the r/altright subreddit.

The Alt-Right

As a political and ideological movement, the alt-right is quite young, only emerging as a topic of mainstream discourse in 2015, coinciding with the U.S. presidential campaign of Donald Trump (Hawley 2017). Its boundaries and beliefs remains loosely defined (Hawley 2017); thus, scholarly literature on the movement itself remains somewhat limited.

The alt-right emerged primarily through a multi-faceted integration of various internet subcultures, and its organization is highly decentralized, existing through dedicated online websites and organizations (Forscher and Kteily 2017). Despite its loose structure, a major foundation of alt-right thought is its concern with the preservation of the white identity (Hawley 2017). The beliefs of the altright have been described as promoting anti-globalist, socially conservative, islamophobic, misogynistic, racist, and xenophobic ideas (Hawley 2017). Research investigating the psychological profiles of the alt-right showed its adherents to have high levels of dark triad traits, high aggression, extreme intergroup bias, and show dehumanization of racial minorities (Forscher and Kteily 2017).

Research on the alt-right behavior has primarily focused on how fringe alt-right news sites spread misinformation across social networks like Reddit, Twitter, and 4chan (Zannettou et al. 2017; Starbird 2017). Research has also investigated how Gab, an emerging alternative social network that promotes itself as a bastion for free speech, has attracted a significant alt-right and right-wing extremist presence on the platform (Lima et al. 2018; Zannettou et al. 2018). However, analysis of the textual content of alt-right social media users is limited. Morstatter et al. (2017), using hierarchical topic modeling of networks of alt-right twitter users, found that these users focused on hot-button social issues like immigration and race. To our knowledge, no work has investigated psycholinguistic characteristics of alt-right communities like sentiment, tone, use of morally loaded language, and patterns of word choice. We believe that our research is the first to investigate how in an isolated alt-right community, psycholinguistic characteristics may reflect theories of ideological extremism and radicalization.

Hate Speech Detection

In recent years, there has been a concerted effort to develop methods to automatically detect instances of hate speech online. Machine learning algorithms have detected hate speech in written text using crowdsourced or manually annotated data (e.g. Davidson et al. 2017; Salminen et al. 2018). Davidson et al. (2017) distinguished between offensive language (e.g. use of swear words), hate speech (e.g. language that has intent to harm members of a group), or neither (ie. not hateful or offensive). With an overall F1 score of 0.90, their model misclassified almost 40% of hate speech, suggesting the inherent difficulty in clearly defining hate speech, and the important role that context and the speaker plays in human perceptions of it. Salminen et al. (2018) developed a comprehensive granular taxonomy for hateful online comments, and built a machine learning model to categorize hateful comments into their different categories with an F1 score of 0.79. Other work from Chandrasekharan et al. (2017), used a collection of large-scale text data from both relatively neutral and previously banned problematic subreddits, which they call a 'Bag of Communities' (BoC) model, to predict the presence of abusive content on other communities. Their model had a 75% accuracy of detecting abusive content in an unseen target community, and 91% accuracy when the model was iteratively used more and more data from the target community to aid prediction (Chandrasekharan et al. 2017).

Hate speech detection is a difficult problem given the different rhetorical forms and styles that can be used, and the inherent role that context plays (Davidson et al. 2017; Salminen et al. 2018). The relatively modest predictive accuracy of most machine learning based approaches reflects this. Although detecting more instances of hate speech than usual or expected may be indicative of a possibly dangerous and radical online community, this method can be inaccurate, unreliable, or incomplete. For instance, some communities that engage in hate speech may be ultimately focused on trolling, but have no other underlying nefarious goals or aims. We propose that the presence of hate speech is just one factor of several that contributes towards determining if an online community shows trends associated with radicalization and extremism. We use both theory-driven methods and data-driven models to detect and approximately quantify the presence of hate speech, for a multi-faceted and comprehensive account of trends towards radicalization and extremism in an online community.

Research Questions

We chose to focus on the components of the described theories of warning behaviors that we judged to be the most applicable to aggregated group level behaviors. For instance, detecting 'Leakage' warning behaviors is something we judge to be most applicable to individual behaviors within online environments already known to harbor violent extremists. Yet judging and detecting purposeful intent is difficult due to the role that context, and slight differences in word choice can have on meaning. For instance, there are a very wide variety of words and phrases that can signal intent (e.g. 'I am going to…', 'we should do…'), and the presence of negation terms can make detecting intent prone to potentially detecting many false

positives (Brynielsson et al. 2013). In these authors work, they restricted their analysis to sites and forums already known to harbor violent extremists, which likely helped mitigate false positives to an acceptable level. As detecting individual instances of leakage in previously unexplored online environments like in our work can be quite difficult and prone to both precision problems, we contend that quantifying the aggregate presence of leakage behaviors would be an even harder and potentially infeasible task. As a test, we constructed more than 50 instances of statements of direct or indirect intent (e.g. 'we are going to….', 'we should…', 'I want to…') and filtered the r/altright subreddit comment data for comments that included these phrases. A qualitative analysis of these comments showed no distinct emergent patterns by which we may detect harmful intent to a reasonable degree of accuracy. For brevity, we exclude the results of this experiment from our analysis. We also concluded that some aspects of 'Identification' warning behavior, particularly identification with radical action, and identification with a role model as also somewhat unfeasible to accurately detect and quantify at scale, due to similar problems of the presence of negation terms and problems with ambiguity. Because of this, we focus our research questions on detecting behavioral markers that we believe to be operably and accurately detectable and quantifiable at a community wide level: fixation behaviors and group identification behaviors, as follows:

1. **RQ1: Fixation:** To what extent does the alt-right subreddit show fixation warning behaviors?
2. **RQ2: Group Identification:** To what extent does the alt-right subreddit show markers of in-group and out-group identification?

In the next section, we outline the data sources and methodology we use to address our research questions.

Data and Methods

Data Collection

All datasets used in our research are drawn from Reddit, a popular discussion forum website where users can submit, comment on, upvote and downvote on posted content (Singer et al. 2014). Reddit also allows users to create and manage communities within the platform called subreddits, which typically focus on particular topics. Often, they employ their own rules and norms for proper participation in the community (Singer et al. 2014).

In our work we use comments taken from the r/alt-right subreddit (AR) as our main data source for our analysis. During its existence, AR was the largest subreddit explicitly associated with the alt-right, and listed its description as 'Discussion of Alternative Right related news and theories' (redditmetrics.com). AR was first formed in March of 2010, but only started gaining popularity and subscribers during 2016, reaching 1,000 subscribers by July 15th, 2016, up to 16,007 subscribers the day before the subreddit was banned by Reddit for violating their content policy (Statt 2017).

To collect comment data from AR we use the pushshift API (pushshift.io). The last day of complete data from AR available from the API was January 11th, 2017. For our work, we chose to collect comment data from AR for a 6-month (181 day) period from the point AR reached 1,000 subscribers (July 15th, 2016), until January 11th, 2017. We chose 1,000 subscribers as the minimum starting benchmark, as our analysis showed that users produced enough aggregated text data per day to reliably perform automated text analysis (at least 1000 words), as recommended by past literature (Pennebaker et al. 2015).

In total, we collected 123,360 unique comments from AR before data cleaning. Comments were cleaned to remove comments marked as deleted (shown as '[deleted]' in

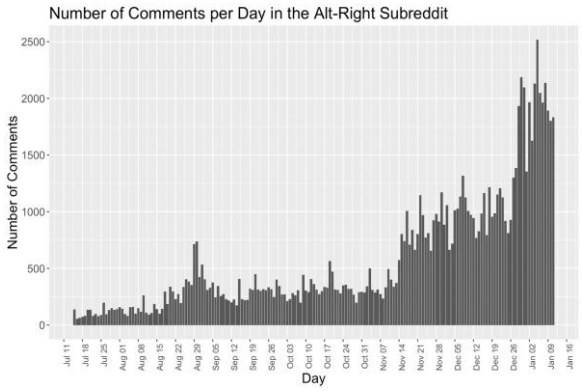

Figure 1: Number of unique comments per day in the Alt-Right subreddit (AR).

Subreddit	Number of Comments
r/altright (AR)	103,722
r/conservative (CON)	141,689
r/libertarian (LBT)	153,914
r/democrats (DEM)	21,522
r/republican (REP)	28,615
r/progressive (PRO)	20,720
r/socialist (SOC)	86,215
r/anarchism (ANA)	70,099
r/anarchocapitalist (ANC)	111,570

Table 1: Summary of number of comments per subreddit for the 6-month data collection period after data cleaning.

the API comment body), remove URLs, and non-trivial symbols (e.g. hashtags and @ symbols), and to change all words to lower-case capitalization. We also removed duplicate instances of text that occurred when another comment was quoted and replied to. Last, we filtered for and removed comments that used the automated reddit bot script (*'I'm a bot, bleep, bloop. Someone has linked to this thread from another place on reddit…'*).

Cleaning the comment data from AR resulted in 103,722 unique comments. Although the peak number of subscribers to AR over this time period was just 13,341, the subscribers had high overall levels of engagement with the subreddit, especially within its final weeks. For the last two weeks of extracting data, an average of 1963.3 new comments were posted per day (see Figure 1 for the distribution of comments per day over the sample period).

For comparison, we collected data on 8 other large and active subreddits existing over this period that focus specifically on distinct political ideologies. We divided into two groups for comparison: (1) a group (**CG**) composed of more mainstream politically ideological subreddits: *r/conservative (C)*, *r/libertarian (L)*, *r/democrats (D)*, *r/republican (R)*, *r/progressive (P)*, and *r/socialist (S)*, and (2) a group we label as radical (**RG**) composed of subreddits with more radical or fringe ideologies: *r/anarchism (A)*, and *r/anarchocapitalist (AC)*.

Our goal was to have as fair as possible a comparison as possible across representative ideological communities. The selection of subreddits in both groups were chosen to meet three criteria. First, we wanted to compare results in different politically ideological communities. Second, we only chose subreddits that had a substantial amount of user engagement (more than 20,000 comments over this 6-month period), to perform statistical analyses on relatively similar and large sample sizes. Last, we limited our focus to subreddits that were oriented towards general discussion of political ideologies, rather than communities oriented around specific politicians, political candidates, political or economic theorists, or specific political policies. For each subreddit in the control group, we collected comment data from the subreddit over the same 6-month period, and applied the same data cleaning procedure to the comment data as applied to AR. Table 1 provides a summary of all the subreddits we analyzed, and the total number of unique comments in each after data cleaning.

Methodological Tools

To detect the presence of hate speech in our data collections, we use a python library called *HateSonar*, which is an open-source version of one of the hate speech detection models mentioned in the related work (Davidson et al. 2017). We chose to employ this model as it was open-source, already pre-trained using social media data, and was designed to try and distinguish between actual hateful language and just offensive language (e.g. use of swear words for emphatic effect). HateSonar detects the presence of both hate speech and offensive language in any excerpt of text, and provides a probability score between 0 and 1 for the hate speech, offensive language, or neither category. Davidson et al. (2017) report an overall precision of 0.91 and recall of 0.90, and an F1 score of 0.90 for the entire model. For hate speech specifically, the model has a precision of 0.44, and recall of 0.61. However, they report that just 5% of offensive and 2% of innocuous (neither category) language is misclassified as hate speech.

To quantify the presence of different psycho-linguistic properties of the subreddit comments, we use LIWC (Linguistic Word County and Inquiry) (Pennebaker et al. 2015). LIWC calculates a rate, or percentage, of words in a sample of text belonging to several predefined categories. Past empirical studies on the validity of LIWC have found that it is reliably able to detect meaning from text in a wide variety of contexts, as well as detect emotional states, intentions, motivations, thinking styles, and individual differences (Pennebaker et al. 2015). We use the latest available LIWC dictionary (LIWC 2015), which provides a broader range of word categories relating to different social and psychological processes compared to earlier versions (Pennebaker et al. 2015).

Research Procedures

Table 2 presents an overview of the methods used to address our research questions, described next.

Fixation: Increasing perseveration on a person or cause. We first look at most frequent terms within the AR comment text data, filtering out common stop words (e.g. 'the', 'is', 'at' etc.). We also look at the term frequency-inverse document frequency (TF-IDF) across the AR subreddit and the 8 control subreddits filtering out stop words. TF-IDF is widely used in text-mining analysis to determine

Research Question	**Sub-Component (if applicable)**	**Methods**
RQ1: Fixation	Increasing perseveration on the person or cause	Term Frequency, TF-IDF → LIWC
	Increasingly negative account of the object of fixation	LIWC, HateSonar
	Increasingly strident opinion and angry emotional undertone	LIWC, HateSonar
RQ2: Group Identification	N/A	LIWC

Table 2: Methods used to address the research questions.

the keywords that are most important to a collection of text (Aizawa 2003). The TF-IDF algorithm weighs how often a term appears in a document (comment), but balances this weighting by the frequency of the word in the entire collection (Aizawa 2003). Using both the most frequent overall terms and highest weighted TF-IDF terms, we look for any emergent patterns and coherent themes that emerge in these top terms in AR, and compare the highest TF-IDF terms in the control subreddits to AR to determine if first, a fixation on a distinct collection of similar terms is shared by the other control subreddits, and second, if the most important terms in AR show a clear fixation on the more hateful and problematic themes associated with the alt-right (e.g. racism, misogyny, xenophobia), that is not shared by the comparison ideological subreddits.

Next, if through a qualitative interpretation of these top terms, if we find a fixation on a specific person or cause appears in the AR data, we manually inspect the top 2000 most frequent terms and the top 500 TF-IDF terms in the AR comment data, and select terms that are clearly related to the object of fixation, to create a dictionary of related fixation terms. We then use this dictionary with the LIWC engine, and calculate the percentage of words in the comments related to the object of fixation per day in AR over the 6-month period. We finally build a linear model to determine if there has been a significant linear increase in these LIWC scores over the 6-month period, which signifies an increasing preoccupation with this person or cause.

Fixation: Increasingly negative account of the object of fixation. For this analysis, we first filter the AR comment data into a smaller subset that includes only comments that mention one or more of these terms in our fixation dictionary from the previous step. Next, using this smaller subset of comments, we use the LIWC *Negative Emotion* category, which is composed of separate sub-categories for *Anger*, *Anxiety*, and *Sadness* to the data for each day over the 6-month period. We also use the LIWC *Emotional Tone* variable to each day of data, which produces an overall metric between 0 and 100 quantifying the overall emotional tone of a body of text. Higher scores indicate a positive and upbeat tone, while lower scores indicate a more negative and hostile tone. Last, we also apply the HateSonar to determine the proportion of comments per day that are categorized as hate speech and offensive language (model probability greater than 0.5). Last, we build linear models to determine if there has been a significant linear increase in negative emotion, hate speech, offensive language, and a linear decrease in emotional tone in comments discussing the object of fixation, over the 6-month period.

Fixation: Increasingly strident opinion and angry emotional undertone. Using the full set of AR comment data, we create a dictionary category which we label as *Core-Hostility* (CH) which is a summation of all the relevant LIWC *Anger*, *Swear*, and *Sexual* categories and apply it to the data. Again, we use HateSonar to determine the proportion of comments per day that are categorized as hate speech as well as offensive language. Using these categories as proxies for anger and strident opinion, we build linear models to determine if there are global trends in the rhetoric of AR that align with the warning behaviors over the 6-month period.

However, it is possible that any overall linear increases in anger and strident opinion may not be unique to just AR, and may be indicative of rising tension due to political tension and polarization generally. To test this, we apply the same dictionary categories and models to the comment data in each subreddit in the comparison groups. We then use Welch's independent sample t-tests to compare the dictionary category scores and F-tests for difference in best-fit slopes for linear trends over time between AR and the control subreddits, to see if there exist any statistical differences in anger and strident opinion between AR and the comparison groups.

Group Identification: To investigate group identification warning behaviors in AR, we use the method suggested by Cohen et al. (2014), and apply the LIWC first-person plural ('we') and third person plural categories ('they') to each day in the AR comment data. These markers act as effective proxies for measuring the degree to which in-group (proportion of 'we' used) vs. outgroup (proportion of 'they' used) dynamics are referred. We also create a new dictionary category which is the ratio of first person plural to first person plural singular ('I'). This measure can show how group identity (ie. 'we') has changed with respect to expressions of individual identity over time, potentially signifying a growing collective identity. The higher value of this ratio, the stronger is the group identity (Cohen et al. 2014; Matthiesen 2003).

We again apply these same dictionary categories to the comment data in each subreddit in the comparison group, and use Welch's independent sample t-tests to compare the dictionary category scores and F-tests for difference in best-fit slopes for linear trends over time between AR and the control subreddits, to see if there exist any statistical differences in group identification between AR and the comparisons over the 6-month period.

Results

Fixation

Increasing perseveration on a person or cause: Table 3 shows a table of top eight most frequent terms in AR (which comprised approximately 4% of all terms), while Table 4 presents a summary of the top TF-IDF terms in both AR and the control subreddits. Six of the top eight

Term	Percentage of all words
white	1.3%
jews	0.5%
alt	0.5%
race	0.4%
whites	0.4%
time	0.4%
black	0.3%
jewish	0.3%

Table 3: Top 8 most frequent terms in AR (filtering out stop words).

most frequent terms (excluding stop words) were, white (1.3%), jews (0.5%), race (0.4%), whites (0.4%), black (0.3%), jewish (0.3%). Looking at the top TF-IDF terms for AR we also see similar themes related to race appear with the top terms, including words like *goy*, *goyim*, *goys* which are all derogatory terms for non-jewish people. In addition, the third highest term is *jq*, short for 'the jewish question', which is a term for a conspiracy theory that jewish people have disproportionate control over the business world, media, and politics, etc (Hawley 2017). In addition, we see terms like *dindu* which is also a derogatory term related to African American culture (Hawley 2017). We also observe the term *ethnostate*, which is a term for a

SR	Top 10 TF-IDF Terms
AR	goy, trs, jq, goyim, pilled, goys, ethnostate, redpilled, degeneracy, dindu
CON	nevertrump, nevertrumpers, towin, shapiro, cruz's, sga, moslems, gowdy, nevertrumper, espn
LBT	ancap, gajo, nap, rothbard, petersen, gary's, mcafee, garyjohnson, amash, iava,
DEM	anotherquery, anotherstring, query, trumpler, voterview, perez, voterinfo, mcfaul, pollingplacesearch, kamala
REP	nevertrump, nevertrumpers, rino, mccrystal, aca, keypuncher, conventionofstates, crowdstrike, akbhar, nehlen
PRO	batchelder, stine, pacifica, chanos, trumpo, grayson, baraka, tulsi's, daveweigel, lucids
SOC	socialism, prin, iww, ableism, monthlyreview, marxists, dsa, engels, bourgeoisie, megathread
ANA	ancaps, fash, iww, ableism, tankie, kropotkin, tankies, stirner, libcom, riseup
ANC	ancap, zerotalk, seronet, ancaps, zeronetwork, anarchocapitalism, goldandblack, ancapistan, greasemonkey, tampermonkey

Table 4: Top 10 TF-IDF terms for all subreddits (SR). Assigned acronyms for each subreddit are used.

Category	Terms
WhiteEthno	white*, ethno*, aryan*
JewOrBlack	jews, jewish, jewry, judaism, jew, ashkenazi, zion*, black*
OtherRacial	racist*, racial*, race*, racism, ethnic* muslim*, islam*, arab*, mestizo*, asian*
RacialSlang	jq, goy*, wuz, kang*, shekel*, dindu*, (((, kebab*, moonman, negro*, nignog, nog*, blm*, nigg*, oy vey

Table 5: Categories and terms for Racial Dictionary in AR.

proposed state where residence is determined by race (Haw ley 2017). Synthesizing the results from the most frequent terms and top TF-IDF terms, we see clear emergent themes suggesting a fixation with race and racial degradation, which aligns with some of the more problematic themes associated with alt-right movements. Although the top TF-IDF terms in the other control subreddits do show some emergent themes, these themes appear oriented around certain specific politicians, concepts or platforms (e.g. *nevertrump, nevertrumpers* in CON and REP*; zerotalk, zeronet, zeronetwork* in the ANC; *tankie, tankies* in ANA) rather than entire groups characterized by immutable characteristics (by race, sexuality, etc.).

Given that a fixation on race and racial hierarchy appears to be present in AR, we next build a dictionary of terms associated with race and racial concepts by inspecting the top 2000 most frequent terms and top 500 TF-IDF terms in AR. Table 5 provides an overview of the words we defined for our dictionary. We found four distinct themes of racially oriented terms, which we categorized based on their overall frequency in the data. First and most popular were terms related to terms associated with white identity or the development of an ethnostate (e.g. *white, aryan, ethno**). Second, most popular were terms associated with Jewish or Black identity (e.g. *Ashkenazi, zion*, black*). Third most popular were terms associated with Arab, Asian, or Mexican identity (e.g. *asian, mestizo*) Last, there were several slang terms and symbols for racial concepts unique to the community (e.g. *(((, kebab**). All slang terms were cross referenced for correct interpretation, using the alt-right behavior tracking website 'Angry White Men' (angrywhitemen.org). We labelled these four themes respectively as *WhiteEthno*, *JewOrBlack*, *OtherRacial*, and *RacialSlang*.

A linear model on the proportion of all racial terms in the AR comment data showed a modest increasing linear trend for the entire 6-month period (R^2 = .09, β = .002, p <.0001).

Measure	R^2	β	p
Negative Emotion	.04	.002	<.01
Anger	.11	.002	<.0001
Sadness	.003	-.0001	.42
Anxiety	.01	-.0003	.06
Emotional Tone	.14	-.07	<.0001
Hate Speech	.04	<.0001	<.01
Offensive Language	.24	.001	<.0001

Table 6: Summary of linear model results for comments containing racial terms.

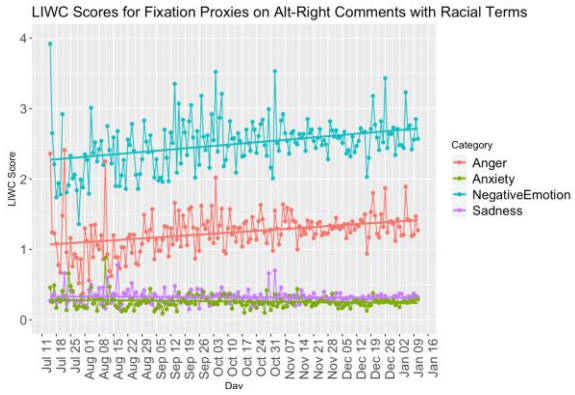

Figure 2: Trends in LIWC Negative Emotion, Anger, Sadness, and Anxiety categories over time for comments in AR containing racial terms.

Increasingly negative account of the object of fixation: Next, we build linear models for the entire 6-month period using this subset of AR comment data containing these racial terms. Table 6 presents a summary of the linear model results, and Figure 2 shows the trends in the proportion of LIWC *Negative Emotion*, *Anger*, *Sadness*, and *Anxiety* categories over time. For the LIWC *Negative Emotion*, *Anger*, *Sadness*, and *Anxiety* categories, only Anger (R^2 = .11, β = .002, p <.0001) and Negative Emotion (R^2 = .04, β = .002, p <.01) have a significant linear increase. For the LIWC *Emotional Tone* variable, there was a significant linear decrease over this period (R^2 = .14, β = -.07, p <.0001), indicating increasingly negative tone. In addition, there was a weak linear increase in the proportion of hate speech (R^2 = .04, β < .0001, p <.01), and a significant line ar increase in the proportion of offensive language (R^2 = .24, β = .001, p <.0001).

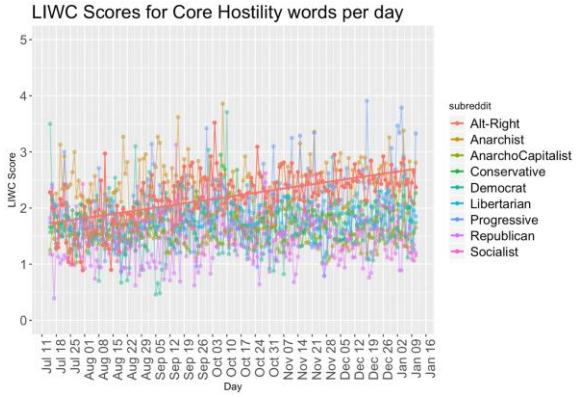

Figure 3: Trends in Core Hostility (CH) over time in AR and all comparison subreddits. AR is colored red.

Increasingly strident opinion and angry emotional undertone: Table 7 shows the results of an F-test comparing differences in slopes for the same metrics. Table 8 shows the results of Welch's independent t-tests comparing AR against all other subreddits in both CG and RG, for the proportion of CoreHostility (CH) words, as well as the proportion of comments classified as Hate Speech (Hate), and Offensive Language (OL). All p-values in these tables have been adjusted using the Bonferroni correction for multiple comparisons.

For CH, we observe that there was a strong and significant linear increase over the 6-month period in AR (R^2 =

SR	CH			Hate			OL		
	R^2	β	F, p	R^2	β	F, p	R^2	β	F, p
AR	.39	.005	N/A	.04	<.001	N/A	.26	.001	N/A
CON	**.01**	**<.001**	**52.59, <.0001**	-.005	<.001	6.44, .09	**.001**	**<.001**	**33.42, <.0001**
LBT	**-.005**	**<.001**	**3.46, <.05**	.08	<.001	2.77, .78	**.03**	**<.001**	**26.14, <.0001**
DEM	**-.005**	**<.001**	**39.66, <.0001**	-.002	<.001	5.36, .17	**.0001**	**<.001**	**16.94, <.001**
REP	**.02**	**.001**	**26.09, <.0001**	.02	<.001	3.44, .53	**.02**	**<.001**	**17.93, <.001**
PRO	.08	.003	4.66, .20	-.005	<.001	6.43, .09	**-.003**	**<.001**	**17.92, <.001**
SOC	**.02**	**<.001**	**93.62, <.0001**	-.006	<.001	4.77, .24	**-.004**	**<.001**	**46.40, <.0001**
ANA	**.01**	**.001**	**28.08, <.0001**	.004	<.001	2.65, .83	**.03**	**<.001**	**15.97, <.001**
ANC	**-.01**	**<.001**	**85.20, <.0001**	-.005	<.001	4.92, .22	**-.004**	**<.001**	**40.28, <.0001**

Table 7: F-tests for difference in best fit slopes for AR against subreddits (SR) for CoreHostility (CH), Hate Speech (Hate), and Offensive Language (OL). Bold implies that the slope in AR is statistically greater than that of the comparison subreddit (p<.05). All p-values are adjusted using the Bonferroni correction for multiple comparisons.

SR	CH (t, p, df)	Hate (t, p, df)	OL (t, p, df)
AR	N/A	N/A	N/A
CON	**9.09, <.0001, 302.33**	**25.87, <.0001, 232.67**	**7.68, <.0001, 312.07**
LBT	**3.73, <.01, 247.10**	**28.22, <.0001, 209.13**	**4.91, <.0001, 299.49**
DEM	**11.83, <.0001, 357.56**	**28.69, <.0001, 243.42**	**6.72, <.0001, 356.95**
REP	**18.226, <.0001, 359.55**	**28.42, <.0001, 244.48**	**18.02,<.0001, 353.98**
PRO	**4.179, <.001, 338.78**	**23.583, <.0001, 322.77**	1.33, .99, 352.88
SOC	**13.81, <.0001, 317.24**	**25.79, <.0001, 253.59**	**7.88, <.0001, 303.01**
ANA	-3.298, <.01, 359.86	**20.52,<.0001, 314.41**	-8.39, <.0001, 351.75
ANC	**19.195, <.0001, 249.13**	**18.79, <.0001, 274.52**	**2.77, <.05, 298.93**

Table 8: Welch's Independent t-test results comparing AR against all other subreddits (SR) for CoreHostility (CH), Hate Speech (Hate), and Offensive Language (OL). Bold implies that the mean AR value was significantly greater than the comparison subreddit (p<.05). All p-values are adjusted using the Bonferroni correction.

SR	We (t, p, df)	They (t, p, df)	We/I (t, p, df)
AR	N/A	N/A	N/A
CON	**20.57, <.0001, 279.19**	**9.69, <.0001, 349.92**	**12.70, <.0001, 269.82**
LBT	**21.78, <.0001, 262.06**	**14.37, <.0001, 308.85**	**14.409,<.0001, 273.47**
DEM	**7.2091, <.0001, 308.08**	**11.18, <.0001, 317.75**	-2.53, p=.10, 273.1
REP	**12.64, <.0001, 353.86**	**12.79, <.0001, 349.05**	**6.47,<.0001, 346.94**
PRO	**10.63, <.0001, 339.18**	**5.432, <.0001, 322.77**	.83, .99, 291.34
SOC	**16.38, <.0001, 325.17**	**12.88 <.0001, 294.36**	**13.358,<.0001, 294.96**
ANA	**19.21, <.0001, 316.9**	**9.11,<.0001, 359.07**	**19.87, <.0001, 283.7**
ANC	**27.574, <.0001, 251.72**	**8.01, <.0001, 320.05**	**20.63, <.0001, 237.77**

Table 9: Welch's Independent t-test results comparing AR against all subreddits (SR) for First Person Plural (We), Third Person Plural (They), and ratio of First Person Plural to First Person Singular (We/I). Positive test statistics imply that the mean value for this subreddit was less than AR. Column Entries in bold imply that the mean AR value was significantly greater than that of the comparison subreddit. All p-values are adjusted using the Bonferroni correction.

.39, β = .005, p<.0001). The results of the F-test for difference in slopes showed that this slope was significantly greater than all the comparison subreddits in both CG and RG, except for PRO (r/progressive) (R^2 = .08, β = .003, p<.02). However, the goodness-of-fit of this linear increase was much stronger in AR (R^2 of .39 vs. .08). In addition, we see that the mean proportion of CH words in AR was significantly greater than all the comparison subreddits, except for ANA (r/anarchism) (t=-3.298, p<.01, df=249.13). Figure 3 shows the trends in the proportion of CH over time in AR and all comparison subreddits.

For hate speech, we find that that over the 6-month period, the mean proportion of hate speech comments in AR is significantly greater than all the comparison subreddits. However, we only find a weak linear increase in hate speech over time in AR (R^2 = .04, β < .001, p=.03), and F-tests for the difference in slopes showed that this trend was not significantly different than the comparison subreddits,

However, we observe that there was a strong and significant linear increase over the 6-month period in the proportion of offensive language in AR (R^2 = .26, β = .001, p<.0001). The results of the F-test for difference in slopes showed that this slope was significantly greater than all the comparison subreddits in both CG and RG. Finally, we also see that the mean proportion of offensive language comments in AR was significantly greater than other comparison subreddits, except for PRO (t=1.33, p=.99, df=249.13) and ANA (-8.39, <.0001, 351.75).

Group Identification

Table 9 shows the results of Welch's independent t-tests comparing AR against all other subreddits in both CG and RG, for the proportion of first person plural words, third person plural words, and the ratio of the proportion of first person plural words to first person singular words. For both first person plural and third person plural words, we find that over the 6-month period, we see that the mean proportion of these words in AR is significantly greater than all the comparison subreddits. Last, we find that the mean ratio of the proportion of first person plural to first person singular words is significantly greater than all comparison subreddits except DEM (t=-2.53, p=.10, df=273.1) and PRO (t=-.83, p=.99, df=291.34). Figure 4 shows the trends in the proportion of both first person plural and third person plural words over time.

The linear trends in the proportion of first person plural words (R^2 = .02, β = -.001, p=.04), third person plural words (R^2 = .06, β = .001, p<.001), and ratio of first person plural to first person singular words (R^2 = .06, β = .001, p=.13) were either relatively weak or non-significant, which suggests that the degree of group identification within AR remained relatively stable over time. Moreover, F-tests for difference in slopes for these categories between AR and all comparison subreddits yielded no major significant differences.

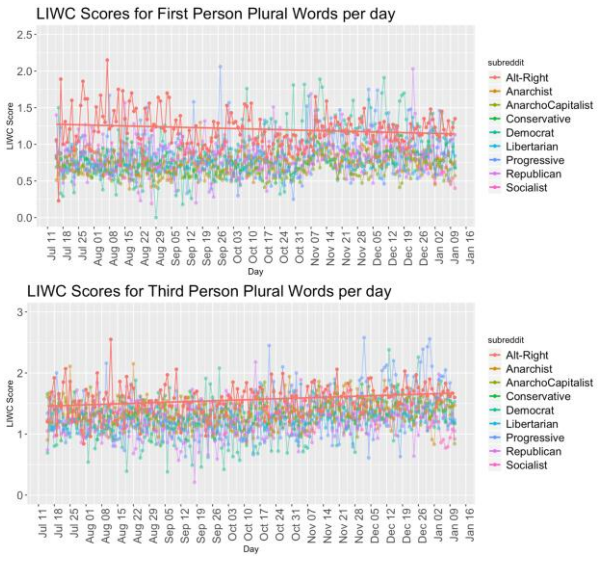

Figure 4: Trends in First Person Plural words (top graph) and Third Person Plural words (bottom graph) over time in AR and all comparison subreddits. AR is colored red.

Summary

Overall, our results suggest that over the course of the 6-month period examined, AR showed behavioral attributes well aligned with both fixation and group identification warning behaviors. AR showed a clear fixation on racial concepts and ideas (primarily directed towards jewish and-black people), and showed increasing negative emotion, particularly anger, over time in comments mentioning these objects of fixation. Furthermore, AR showed elevated levels of hate speech and hostile language not seen in all comparison subreddits (except ANA), and a strong linear increase in hostile language (R^2 = .39) that was not present in any comparison subreddits. In addition, AR showed high levels of in-group and out-group identification markers that were not present in any comparison subreddits. However, we saw no linear increases or decreases in the use of language associated with group identification over this period.

Discussion and Conclusions

Our work is the first to present initial evidence that some warning behaviors that have been previously associated with the psychopathology of individual ideological extremists in online environments, may also be present in the aggregated behavior of entire online communities focused on extreme or radical political ideologies. Using r/altright subreddit (AR) as a case study, we show that this community presented clear signs of known warning behaviors of violent extremism, particularly fixation and group identification, during a 6-month period prior to its ban for violating Reddit's terms of service. For fixation behaviors, we observed an increasing fixation in AR on race and racial concepts and an increasing anger in comments that mentioned these racial terms. Furthermore, we saw a strong linear increase in hostile language use (R^2 = .39). For group identification, we saw that AR used language associated with in-group (e.g. 'we') and out-group (e.g. 'they') at a much higher rate than all comparison subreddits. However, we saw no significant linear increases or decreases in group identification language over the 6-month period, which suggests that group identity has been an important and stable aspect of Alt-Right thought even from its earlier emergence into political discourse. This observation aligns with past research noting the important foundational role white identity plays in Alt-Right ideology (Hawley 2017). The combination of both these behavioral patterns (fixation and group identification) was not seen in other political subreddits we compared against.

In our work, we chose to employ some relatively straight forward theory driven dictionary methods to identify warning behaviors, either based on recommendations from related literature (Cohen et al. 2014), or through choosing the dictionary categories that we judged to be the best proxy for the descriptions of each warning behavior. As these methods just rely on simple word count metrics, the effectiveness of these dictionary proxies in detecting warning behaviors is not certain. Therefore, we chose to perform a brief evaluation of the effectiveness of these methods. First, we randomly sampled 800 comments from the entire set of AR comments. We then manually scored each comment on an ordinal scale of 1 to 5 based on the perceived degree the comment presented properties that aligned with the warning behaviors and their subcomponents, if applicable (5 being the strongest alignment). Only comments containing racial terms were used to score the second fixation behavior (i.e. negative account of racial fixations), while all sample were used to score the remaining behaviors. For each measure, we then divide the comments into ten subsets from the sample ordered by increasing assigned score. Finally, for each warning behavior, we calculated the corresponding LIWC scores for the aggregated comments of each subset, and performed a Pearson correlation between the LIWC scores and the mean of the human-assigned scores across all subsets. Correlation results are shown in Table 10. We observe that most of the dictionary proxies have a strong and significant association with their relevant warning behaviors ($r > |.70|$), except for third person plural words, which indicates that this measure may not be a reliable indicator of group identification. Despite these strong associations, dictionary methods like LIWC unfortunately have no ability to consider the context in which these words are used. Although LIWC has been employed previously in similar research contexts for detection of general temporal trends in large collections of text (De

	Neg	Ang	Tone	CH	We	They
Fixation: negative account	**.79***	**.75***	**-.72***	N/A	N/A	N/A
Fixation: strident/angry opinion	N/A	N/A	N/A	**.90***	N/A	N/A
Group Identification	N/A	N/A	N/A	N/A	**.98***	.48

Table 10: Pearson correlations between human scored warning behaviors and LIWC scores for negative emotion (Neg), Anger (Ang), Emotional Tone (Tone), Core Hostility (CH), first person plural (We), and third person plural (They).
* - p <.05

Choudhury et al. 2014), we foresee future opportunities to investigate if distributional approaches to language use differences (e.g. word embeddings), or other context-sensitive natural language processing techniques may be more effective at detecting aggregated warning behaviors. Future work should also aim to determine if the warning behaviors we did not attempt to detect in our work (e.g. leakage, identification with role model) could be readily detected through more context-sensitive techniques.

A next step is to also investigate how replicable our findings are with other communities associated with radical and extreme ideologies. Even though we did not see as fixation behaviors as clear as in AR, the r/anarchism (ANA) subreddit also showed elevated levels of hostile language over this period, which suggests that more radical ideologies may show aspects of these warning behaviors to varying degrees. In recent years, radical political groups loosely associated with anarchist ideologies, like Antifa, have been implicated in acts of public violence and vandalism (Pyrooz and Densley 2018). Future work should investigate if similar psycholinguistic patterns of language use appear in other online communities oriented around political and ideological extremism, as well as in other domains where extremism may harbor itself (e.g. religious extremism or conspiracy theorist communities). One example of another controversial and now banned subreddit that could be analyzed through a similar methodology is r/incels, a community oriented around the grievances of involuntarily celibate men. In the past four years, four mass killings resulting in 45 deaths have been committed by men who have self-identified with the 'incel' subculture (Kelshall, 2019).

As political and ideological stratification in society continues to grow (Halberstam and Knight 2016), and online communities focused on ideological commitments become more numerous, moderators of online platforms, like Reddit, which will inevitably harbor more of these radical communities, face difficult challenges in how to balance the right to free expression, with broader concerns of public safety and wellbeing. Research has shown that the more individuals see their extreme views as stigmatized offline, the more likely they are to build a sense of community in the online sphere (De Koster and Houtman 2008). However, it is clear from past research that these radical online communities can also over time create violent extremists. Therefore, careful monitoring of these communities going forward is crucial. Through our research, we aim to contribute a replicable and malleable methodological framework for more objective means in long-term proactive monitoring of warning behaviors of ideological extremism in online communities, ideally before violent action may take place in the real world.

Limitations

The API we used to query Reddit (pushshift.io) is known to have issues with missing data (Gaffney and Matias 2018). However, the worst problems with missing data are for submissions and comments prior to 2011, and the rate of missing submissions was found to be considerably worse overall than for comments, which we use in our work (Gaffney and Matias 2018). As we use only comment data in our research, the chance that missing data issues may skew our results are considerably less. The relative abundance of comment data compared to submission data, and greater concerns over missing submission data were the primary drivers for our decision to only use comment data in our analyses. Ideally, future work using a more complete Reddit API should analyze both comment and submission data together, and perhaps even compare the relative presence of warning signs in comments and submissions. Despite any potential issues with missing data, Gaffney and Matias (2018), suggest that work focusing on user history or network analysis between users faces the greatest risks of any kind when using this API, which are methods which we did not employ in our research.

References

Aizawa, A. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management 39*(1): 45-65.

Blaker, L. 2015. The Islamic State's use of online social media. *Military Cyber Affairs*, *1*(1): 4.

Brynielsson, J.; Horndahl, A.; Johansson, F.; Kaati, L.; Mårtenson, C.; and Svenson, P. 2013. Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics 2*(1): 11.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* 3175-3187. ACM.

Cohen, K.; Johansson, F.; Kaati, L.; and Mork, J.C. 2014. Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence, 26*(1): 246-256.

Datta, S., and Adar, E. 2018. Extracting Inter-community Conflicts in Reddit. *arXiv preprint arXiv:1808.04405*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

De Choudhury, M.; Monroy-Hernández, A.; and Mark, G. 2014, April. Narco emotions: affect and desensitization in social media during the mexican drug war. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* 3563-3572. ACM.

De Koster, W., and Houtman, D. 2008. 'Stormfront is like a second home to me' On virtual community formation by right-wing extremists. *Information, Communication & Society 11*(8): 1155-1176.

Forscher, P.S., and Kteily, N. 2017. A Psychological Profile of the Alt-right.

Gaffney, D., and Matias, J.N. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one 13*(7): e0200162.

Halberstam, Y., and Knight, B., 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics 143*: 73-88.

Hawley, G. 2017. *Making Sense of the Alt-right*. Columbia University Press.

Hine, G.E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017, May. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.

Johansson, F.; Kaati, L.; and Sahlgren, M. 2017. Detecting linguistic markers of violent extremism in online environments. In *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications* 2847-2863. IGI Global.

Kelshall, C., 2019. Violent Transnational Social Movements and their Impact on Contemporary Social Conflict. *The Journal of Intelligence, Conflict, and Warfare, 1*(3).

Kumar, S.; Hamilton, W.L.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* 933-943.

Lima, L.; Reis, J.C.; Melo, P.; Murai, F.; Araujo, L.; Vikatos, P.; and Benevenuto, F. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 515-522. IEEE.

Lowery, W.; Kindy, K.; and Tran, A. 2018. In the United States, right-wing violence is on the rise. Washington Post. https://www.washingtonpost.com/national/in-the-united-states-right-wing-violence-is-on-the-rise/

Matthiesen, K. 2003. On collective identity. *Protosociology* 1(18): 66-88.

Meloy, J.; Hoffmann, J.; Guldimann, A.; and James, D. 2012. The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral sciences & the law 30*(3): 256-279.

Morstatter, F.; Shao, Y.; Galstyan, A.; and Karunasekera, S. 2018. From alt-right to alt-rechts: Twitter analysis of the 2017 german federal election. In *Companion of the The Web Conference 2018 on The Web Conference 2018* 621-628.

Pennebaker, J.W.; Boyd, R.L.; Jordan, K.; and Blackburn, K. 2015. *The development and psychometric properties of LIWC2015*.

Pyrooz, D.C., and Densley, J.A. 2018. On Public Protest, Violence, and Street Gangs. *Society 55*(3): 229-236.

Roser ,M.; Mohamed, N.; and Ritchie , H. 2018. Terrorism. *Published online at OurWorldInData.org*. Retrieved from: https://ourworldindata.org/terrorism.

Salminen, J.; Almerekhi, H.; Milenković, M.; Jung, S.G.; An, J.; Kwak, H.; and Jansen, B.J. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*.

Scrivens, R.; Davies, G.; and Frank, R. 2018. Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. *Behavioral sciences of terrorism and political aggression 10*(1): 39-59.

Singer, P.; Flock, F.; Meinhart, C.; Zeitfogel, E.; and Strohmaier, ¨ M. 2014. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? In WWW (Companion).

Starbird, K. 2017. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *Eleventh International AAAI Conference on Web and Social Media* 230-239.

Statt, N. 2019. *Reddit bans two prominent alt-right subreddits*. The Verge. https://www.theverge.com/2017/2/1/14478948/reddit-alt-right-ban-altright-alternative-right-subreddits-doxing.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtelris, N.; Leontiadis, I.; Sirivianos, M.; ... and Blackburn, J. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference* 405-417. ACM.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G. and Blackburn, J., 2018, April. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion of the The Web Conference 2018 on The Web Conference 2018* 1007-1014. International World Wide Web Conferences Steering Committee.