Social Network Analysis and Mining (2024) 14:32
https://doi.org/10.1007/s13278-023-01187-5

ORIGINAL ARTICLE

Intellectual dark web, alt-lite and alt-right: Are they really that different? a multi-perspective analysis of the textual content produced by contrarians

Breno Matos[1] · **Rennan C. Lima**[1] · **Jussara M. Almeida**[1] · **Marcos A. Gonçalves**[1] · **Rodrygo L. T. Santos**[1]

Received: 17 July 2023 / Accepted: 17 December 2023 / Published online: 25 January 2024

Abstract
Contrarian groups, notably Intellectual Dark Web, Alt-lite, and Alt-right, are present across the Web, ranging from fringe websites to mainstream social media. Such massive presence raises major concerns as contrarians often engage in the spread of conspiracy theories and hate speech toward particular groups of people. Historically, there is a general sense that these groups exhibit different degrees of extremism, with Alt-right standing out as the most extremist one. In particular, prior work often takes participation in Alt-right communities as a proxy for radicalization. Yet, *to which extent are these groups really different?* While most previous analyses have focused on a *content consumption* (i.e., viewer) standpoint, no prior work analyzed these groups (i.e., contrarians) from a **content production** perspective. *Are there significant differences in the content produced by them?* Toward tackling this question, we here analyze the textual data associated with videos shared by the three aforementioned groups. Specifically, we analyze 14 years of content produced by contrarians on YouTube with data from 355,000 videos. Firstly, we assess the degree of toxicity of the content created by each contrarian group, comparing them to one another and, for control purposes, against traditional media content. The results show that all contrarian groups have a more skewed toxicity distribution than traditional media. Yet, all three groups exhibit very similar textual toxicity properties. Further analyses based on psycholinguistic properties and semantic (text) classification reinforce the observation that indeed there is great similarity among the content created by all three contrarian groups. These results suggest that, despite the different definitions, the three contrarian groups are indeed much more similar, in terms of the content produced and shared by them, than the general wisdom (and literature) seems to suggest. Moreover, we also identify a significant temporal increase in content toxicity in all three groups, corroborating prior observations regarding the escalation in the harmfulness of online speech over the years.

Keywords Natural language processing · Radicalization · Social media

1 Introduction

Online social networks (OSNs), such as YouTube and Twitter, have made publishing and sharing content online easier. These platforms rely on millions of users posting content and interacting with each other to keep the network active. Notably, the literature has reported a continuous decrease in the use of traditional media as source of information, while the consumption of news and opinion content from social media increases (Newman et al. 2021; Ingram 2018), which potentializes the reach and importance of ideas spread by online personalities, such as YouTubers and digital influencers. In this context, much has been debated about the OSNs' role in spreading toxic content (Obadimu et al. 2019) and

✉ Breno Matos
brenomatos@dcc.ufmg.br

Rennan C. Lima
rennancordeiro@dcc.ufmg.br

Jussara M. Almeida
jussara@dcc.ufmg.br

Marcos A. Gonçalves
mgoncalv@dcc.ufmg.br

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

[1] Universidade Federal de Minas Gerais, P.O. Box 1212, Belo Horizonte, Brazil

increasing user radicalization in recent years (Ribeiro et al. 2020).

Radicalization in OSNs is a challenging subject to study, as there is no consensus regarding what “hateful" or “extreme" means (Sellars 2016). We here consider the definition proposed by McCauley and Moskalenko (2008) which relates radicalization to *"change in beliefs, feelings, and behaviors in directions that increasingly justify intergroup violence and demand sacrifice in defense of the ingroup"*. We choose this definition as it is widely used in the literature (King and Taylor 2011; Bartlett and Miller 2012; Dalgaard-Nielsen 2010; Hafez and Mullins 2015; O'Malley 2022; Wolfowicz et al. 2020; Borum 2011; Ribeiro et al. 2020).

Radicalization in OSNs has been the topic of several studies. For example, Lima et al. (2020) compared moderated textual data from Twitter with data from Gab, an un-moderated social network heavily used by fringe/extreme communities, finding that Gab content presents higher toxicity and negativity. However, as reported in prior work, even highly moderated platforms also struggle with user radicalization. For instance, by performing a large-scale audit of YouTube channels, Ribeiro et al. (2020) found a progressive migration of users from traditional to more extreme content. The authors focused on analyzing channels of the so-called *contrarian groups*, namely the Intellectual Dark Web (IDW), Alt-lite, and Alt-right, identifying radical online communities based on defining features, such as ties to white supremacy (e.g., Alt-right) (Ribeiro et al. 2020). Contrarians set themselves apart from traditional media and, as shown by prior work, share content that can create gateways to radicalization (Lewis 2018; Ribeiro et al. 2020).

Considering the current literature on radicalization in OSNs, we observe a general assumption, followed by a large number of scientific studies (Ribeiro et al. 2020; Winter 2019; McClernan 2019; Kelsey 2020; Finlayson 2021; Atkinson 2018; Lewis 2018; Hosseinmardi et al. 2021; Moffitt 2023; Das 2023) as well as articles published by the mass media (Marantz 2017; Roose 2019a, b; Weiss and Winter 2018; ADL 2019, 2017), that the aforementioned contrarian groups are different and have varying degrees of extremism by definition, with IDW being the least extreme, Alt-right being the most extreme, and Alt-lite lying between the two. Yet, to our knowledge, whether such differences indeed exist remain to be evaluated.

Moreover, focusing particularly on YouTube as a major platform for information dissemination online (Grant 2022; Milmo 2022; Coleman 2022; Li et al. 2020; Tang et al. 2021), prior studies, including (Ribeiro et al. 2020), focused on user viewing patterns, i.e., patterns of **content consumption**. No properties of the (textual) content generated and spread by these groups were indeed analyzed. As such, defining properties of the **content produced** by channels owned by contrarians remain to be characterized.

To which extent is such content really different? Does the content produced by these groups reflect distinct degrees of extremism, as often assumed in existing literature? These are the questions we aim to tackle in this work. In particular, we investigate online radicalization from a **content production** perspective. In particular, we analyze textual properties of content produced by contrarian groups on YouTube. Following (Ribeiro et al. 2020), we consider channels by the three aforementioned contrarian groups, namely IDW, Alt-lite, and Alt-right, and include traditional media (mainly news channels) content as a control group for comparison purposes. Our goal is to assess the similarities and differences among the three contrarian groups, as well as between them and the control Media group, which will help us understand if the boundaries separating the contrarians are meaningful: that is, if contrarians are distinct enough in respect to the content produced by them to be considered different groups, or if instead the separation between them is artificial. Notably, we are interested in quantifying the differences in the level of *toxicity* of their content, and assessing whether such differences (if any indeed exist) agree with the varying degrees of extremism expected from the analyzed communities (i.e., traditional media should report the lowest toxicity scores and Alt-right the highest ones). Notably, our study is driven by three main research questions (RQs):

RQ1: What are the differences in the content produced by the four groups with respect to various toxicity-related features? To answer this question, we use Google's Perspective API [1] to characterize the channels' metadata with respect to a number of toxicity-related features (e.g., insult, profanity). We also perform a temporal analysis of toxicity-related features of content produced by all four groups over a 14-year period. Our results reveal that: (i) there are no large differences across the three contrarian groups with respect to most toxicity features analyzed (with only marginal differences in a few cases), implying that differences in the level of extremism, as suggested in (or assumed by) the literature (Ribeiro et al. 2020; Hosseinmardi et al. 2021; Kelsey 2020; Finlayson 2021; Atkinson 2018; Lewis 2018; Winter 2019; Moffitt 2023; Das 2023), do not clearly emerge in the degree of toxicity of the content produced by these groups; (ii) all three contrarian groups produce much more toxic content than the control Media group, which sets itself apart with significantly lower toxicity; and (iii) all three contrarian groups exhibit a significant increase in the level of content toxicity (average values for multiple features) over the years.

RQ2: From a psycholinguistic perspective, what are the main differences among the groups? We further investigate the characteristics of the toxic content

[1] https://www.perspectiveapi.com/.

produced by contrarians by analyzing psycholinguistic attributes. To that end, we employ LIWC (Tausczik and Pennebaker 2010), a psycholinguistic dictionary that classifies words into 73 categories expressing multiple traits, such as writing style, cognitive and affective concepts, and grammatical classes, offering an alternative look into the content produced by each community. Our analysis reveals additional contrasts between Media and all contrarian groups, consistently with our RQ1 results. For example, all contrarians present more negative emotions, swear words, and sexual related terms than Media. Moreover, regarding particular LIWC attributes, we do observe some differences among the contrarians, with, for example, Alt-lite presenting more sexual-related terms than both Alt-right and IDW. Yet, such differences are subtle and much more marginal than the observed differences between contrarians and Media.

RQ3: From a semantic (contextual representation) and lexical perspectives, what are the main differences among the groups? We employ a state-of-the-art contextual language model based on the Transformer architecture (Devlin et al. 2018) to produce content embeddings and compute their similarities. We also analyze lexical features via an information gain classifier to showcase the most discriminative terms. Once again, we find that Media content exhibits very different textual patterns, setting itself apart from content produced by contrarians, which in turn present much more similar embeddings. Further results of a clustering analysis of the contextually embedded representations reveal that the contents produced by contrarians exhibit weak separability (de Andrade et al. 2023), suggesting a reasonable degree of semantic similarity in their discourses.

In sum, our results show that, at least in terms of the properties analyzed, which are widely studied properties when it comes to textual content (notably hate speech) analysis, the three contrarian groups offer quite similar content to their audiences. We acknowledge that the concept of radicalization has different facets and may manifest itself in different ways. Yet, our results seem to suggest that the separation of these groups into increasing levels of extremism, as often assumed in the literature, does not manifest itself in the content properties analyzed and might indeed be artificial. As such, our work sheds a new light into radicalization in OSNs (notably YouTube), where the boundaries between groups of users often associated to such practice are clearly much more blurry than previously assumed.

In the remainder of this article, Sect. 2 summarizes prior studies on contrarian groups and related work. Section 3 describes the dataset used, while Sects. 4, 5, and 6 present the methodology and results for our three research questions, respectively. Section 7 discusses limitations of our work and, finally, Sect. 8 concludes the paper, discussing possible extensions for future work.

2 Background

We use (Ribeiro et al. 2020) as a basis for our work. Ribeiro et al. (2020) provide substantial quantitative evidence on the alleged radicalization pipeline on YouTube, analyzing over 330,000 videos and 72 million comments. According to the study, the researchers analyzed 349 channels on YouTube, which they broadly classified into four types: Media, the Alt-lite, the Intellectual Dark Web (IDW), and the Alt-right. The radicalization hypothesis in their paper suggests that the IDW and the Alt-lite channels serve as gateways to fringe far-right ideology represented by Alt-right channels, with these three communities varying in the extremity of their content. The study found that users consistently migrate from milder to more extreme content on YouTube and a large percentage of users who consume Alt-right content now, consumed Alt-lite and IDW content in the past. Specifically, the study analyzed user migration patterns and found that the commenting user bases among the IDW, the Alt-lite, and the Alt-right are increasingly similar, indicating a growing percentage of users consuming extreme (Alt-right) content on YouTube while also consuming content from other milder communities (Alt-lite/IDW).

2.1 Contrarian groups

Following (Ribeiro et al. 2020), we consider three YouTube's notable communities: Intellectual Dark Web (IDW), the Alt-lite, and the Alt-right. These communities grew during the "anti-politically correct (PC)" culture in the 2010s (Nagle 2017) and have been described as contrarians (Ribeiro et al. 2020) due to views that often oppose mainstream ideas.

Alt-right, which is short for *alternative right*, is a segment of the white-supremacist movement (ADL 2017) that openly supports antisemitic and racist views (ADL 2019). It is a far-right group with a younger demographic and a significant online presence on multiple websites, especially forums and imageboards (e.g., 4chan and 8chan) (ADL 2017). The term Alt-lite, in turn, was coined to differentiate individuals who do not openly condone racist opinions from other right-wing groups. As argued by Alt-right writer Greg Johnson, "the Alt-lite is defined by civic nationalism as opposed to racial nationalism, which is a defining characteristic of the Alt-right" (ADL 2019). Intellectual Dark Web (IDW), in turn, is a term used to describe iconoclast thinkers, political commentators, podcast hosts, and academics that discuss controversial topics (Weiss and Winter 2018). The term was popularized in a New York Times article (Weiss and Winter 2018): "iconoclastic

thinkers, academic renegades and media personalities who are having a rolling conversation about all sorts of subjects, [. . .] touching on controversial issues such as abortion, biological differences between men and women, identity politics, religion, immigration, etc." Examples of personalities considered part of the IDW are Jordan Peterson, Ben Shapiro, and Joe Rogan (Weiss and Winter 2018).

In summary, openly racist views distinguish the Alt-lite and the Alt-right more clearly, while both are right-wing communities. IDW can pertain to any part of the political spectrum, while Alt-lite and Alt-right are both entirely right-wing. Although controversial, IDW personalities do not necessarily support extreme views, which are fundamental in defining the Alt-right, which openly promotes white supremacy. In between IDW and the Alt-right is the Alt-lite, which often flirts with concepts related to white supremacy (such as globalist conspiracies) and has a blurry line (ADL 2019) setting itself apart from the Alt-right, with many Alt-liters being accused of attenuating their views and softening their positions to appeal to a broader public (Ribeiro et al. 2020). Thus, by definition, the three contrarian groups target different user populations, and such differences may reflect how they behave online (e.g., properties of content shared by them). Indeed, according to the literature, IDW is considered the least extreme of the three groups, Alt-right the most extreme one, while Alt-lite lies in the middle of the two (Ribeiro et al. 2020; Winter 2019; McClernan 2019; Kelsey 2020; Finlayson 2021; Atkinson 2018; Marantz 2017; Lewis 2018; Roose 2019a; Weiss and Winter 2018; ADL 2019; Hosseinmardi et al. 2021), although no prior study analyzed whether such claimed differences in extremism are indeed reflected in the content produced by them. As stated in Sect. 1, we aim to shed light into the boundaries between the contrarians to assess whether the definitions mentioned reflect meaningful separations

2.2 Related work

A plethora of studies have investigated the presence of toxic content (toxicity, for short) online (Lima et al. 2020; Obadimu et al. 2019; Guimarães et al. 2020; de Andrade and Gonçalves 2021; Matos et al. 2022; Chipidza 2021), often employing Google's Perspective API, which assigns toxicity scores (capturing various toxicity-related features) to an input text. Another body of work, more closely related to ours, investigated radicalization and extremism online, driven by the recent increase in right-wing communities and their role as breeding grounds for radical and Alt-right content (Ribeiro et al. 2020; Zannettou et al. 2018). For instance, Mamié et al. (2021) performed a large-scale audit of subreddits and YouTube channels that are part of the Manosphere, a collection of websites with anti-feminist and misogynistic content. Their findings unraveled the dynamics connecting the Manosphere and the Alt-right groups, uncovering a considerable user overlap. Similarly, Bryant (2020) investigated the filter bubble effect on YouTube and its role as a recruiting tool for the Alt-right. The presence of extremism has also been studied on other platforms such as Gab (Arnold et al. 2021) and Twitter (Thorburn et al. 2018; Morstatter et al. 2018).

Others have linked toxicity sharing to contrarian and radical groups. For example, some authors analyzed the presence of hate speech and toxicity on Gab, observing that the platform draws Alt-right users and conspiracy theorists (Zannettou et al. 2018; Lima et al. 2020). Ribeiro et al. (2021b), in turn, studied Reddit communities that were banned and migrated to their own self-contained platforms, finding an increase in toxicity and radicalization indicators in the content shared by them after migration. The same authors also studied the evolution of Manosphere-related communities, revealing increased activity level and toxicity over the years (Ribeiro et al. May 2021). Like Mamié et al. (2021), they also observed a large overlap in user bases, supporting the hypothesis that these communities are a pathway to radicalization (Ribeiro et al. 2020). Ottoni et al. (2018) proposed an analysis of right-wing content on YouTube, which can overlap with contrarian content (e.g., far-right content), although not the same. We analyze a much larger amount of videos (355,000 versus 7000) while focusing on characterizing contrarian communities. More specifically, our present effort relies on data gathered by Ribeiro et al. (2020), which analyzed comments, video, and channel recommendations from media channels, IDW, Alt-lite, and Alt-right). Based on user comments on videos from each group, the authors observed a great overlap in user consumption of content by both Alt-lite and Alt-right groups.

While Ribeiro et al. (2020) and related studies (Mamié et al. 2021; Ribeiro et al. May 2021; Arnold et al. 2021) focused primarily on analyzing user interactions (e.g., through comments), which is intrinsically associated with patterns of *content consumption*, we here explore an orthogonal approach, focusing on the **content production** perspective. Specifically, we analyze the toxicity, psycholinguistic, semantic, and lexical properties of the content that YouTube channels from different communities expose to their users. We aim to answer questions such as: "Is the content produced by different contrarian groups essentially different, notably concerning toxicity?" Or, "is the separation enforced by community definition reflected in the content produced by them?" To the best of our knowledge, these questions have not been suitably and thoroughly answered in the literature regarding contrarians. By addressing them, we provide a complementary view of the content spread by right-wing and contrarian groups on YouTube—a view that has not been fully exploited by prior work.

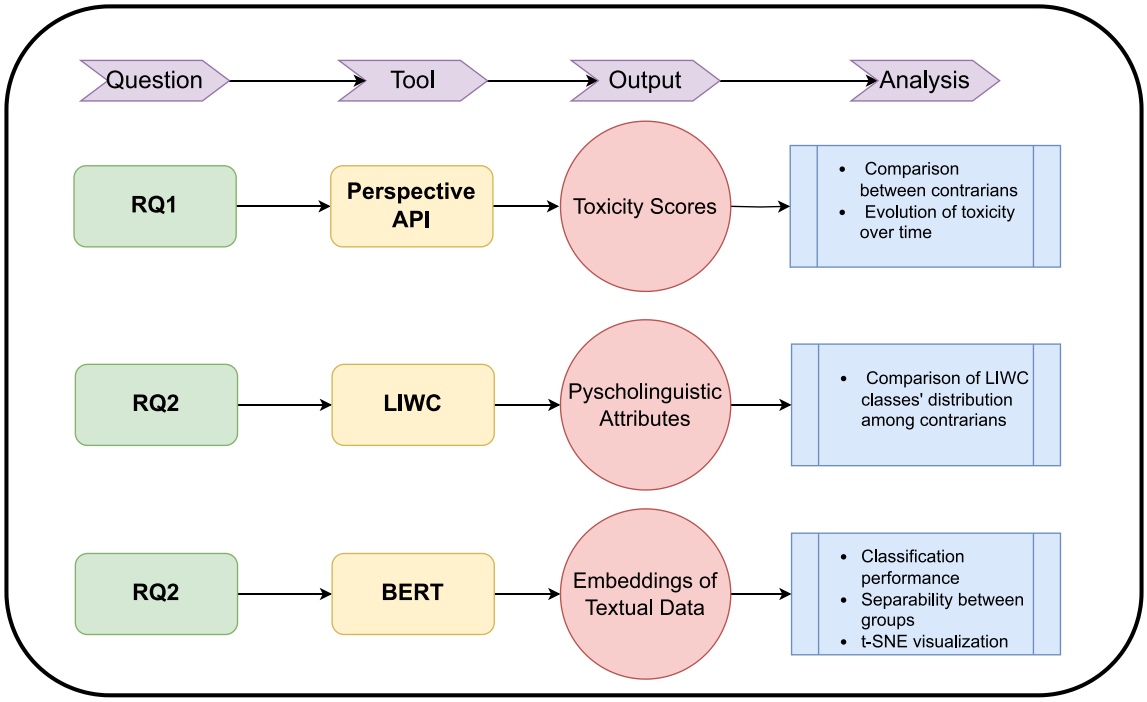

Fig. 1 Overview of the experimental methodology that supports our investigations

In sum, despite a large body of work on far-right and fringe communities' online activity, few of them analyzed content properties. In particular, no prior study focused on the properties of content produced and shared by contrarian YouTube channels. Some YouTubers speak to broader audiences and indeed can impact a community's dynamics and encourage real-world coordinated actions (Niu et al. 2021). Thus, looking into the contents they are sharing on the platform is key to a better assessment of how such contrarian groups may affect society as a whole. This investigation, which is lacking in the literature, is the aim of our present effort.

3 Methodology

Recall that our goal is to address three key research questions, posed in Sect. 1. To tackle each such question, we employ different techniques, as illustrated in Fig. 1, over two types of input that represent the videos in our dataset:(I) captions and (II) the concatenation of titles and descriptions.[2]

More specifically, toward tackling RQ1, we rely mostly on the Google's Perspective API to characterize the content produced by each contrarian group with respect to a number of toxicity-related features: we study how the scores for each feature compare between contrarians and analyze how such features evolved over time. Next, to analyze the differences among the groups with respect to psycholinguistic features (RQ2), we employ the Linguistic Inquiry and Word Count (LIWC), which categorizes words into multiple linguistic (e.g., first-person singular pronouns, conjunctions), psychological (e.g., anger, achievement), and topical (e.g., leisure, money) groups. Finally, toward delving into the differences in semantic and lexical aspects (RQ3), we use a state-of-the-art contextual language model to retrieve dense embedding representations for our inputs. More details on how each technique is employed and the results obtained are presented in the following sections. Next, we focus on the dataset used in our study.

We use a dataset of textual content from YouTube channels gathered by Ribeiro et al. (2020). This dataset includes metadata (video descriptions, titles, and captions) of over 355,000 videos shared by 350 channels published during a 14-year period, from 2006 to 2019. The videos are grouped into four categories, namely Media, IDW, Alt-lite, and Alt-right. According to Ribeiro et al. (2020), the list of *Media* channels, used for comparison purposes, was obtained from *mediabiasfactcheck.com*.[3] The lists of IDW, Alt-lite, and Alt-right channels, in turn, were gathered by starting with a list of seed channels and then collecting recommendations

[2] https://github.com/brenomatos/contrarians.

[3] https://mediabiasfactcheck.com/.

Table 1 Dataset overview

Category	# Channels	# Videos	% Videos with descriptions	# Videos with captions (%)	% Captions under 20 KB	Example channels
Media	68	220801	98 %	121374 (55%)	95 %	L.A. Times, New Yorker, Washington Post
IDW	125	59509	95 %	43472 (73%)	72 %	JordanPetersonVideos, JRE Clips, StevenCrowder
Alt-lite	73	57361	97 %	51369 (90%)	86 %	Rebel Media, Revenge of the Cis, Western Man
Alt-right	84	17417	83 %	15162 (87%)	65 %	AltRight.com, Stand Up Europe, American Pride

provided by YouTube for these channels. The list of seed channels is based on Anti Defamation League's report on the Alt-lite and the Alt-right (ADL 2019), Data & Society's report on YouTube Radicalization (Lewis 2018) and the IDW unofficial website (Manifest yyyy). Ribeiro et al. (2020) then manually assigned each collected channel to the IDW, Alt-lite, or Alt-right groups. Table 1 shows an overview of the dataset. Media is the category with most videos, but with the fewest channels. IDW and Alt-lite have a similar amount of videos, with IDW having the largest number of channels. Lastly, Alt-right is the category with the fewest videos, but roughly the same number of channels as Alt-lite's.

Our analyses rely on the following metadata associated with each video: title, description, and caption. Though the first is present in all videos, descriptions and captions (i.e., audio transcripts) are available for only a fraction of the videos, reported in the 4th and 5th columns of Table 1. Most videos have descriptions, but the fraction of videos with captions available varies from 65 to 95% across the categories. Moreover, the majority of the available captions are under 20 KB, as shown in the 6th column of the table. Examples of channels from each category are presented in the rightmost column of the table: Media contains popular news outlets, such as the Washington Post. IDW contains famous channels such as those from Jordan Peterson and Joe Rogan (JRE). Alt-lite and Alt-right contain more niche channels.

4 RQ1: Content toxicity

We start our analysis by tackling RQ1: *What are the differences in the content produced by the four groups with respect to various toxicity-related features?* To that end, we follow prior work (Lima et al. 2020; Ribeiro et al. May 2021, 2021b; Guimarães et al. 2020; Obadimu et al. 2019) and employ Google's Perspective API to evaluate text *toxicity-related features*. Given an input text, the API returns multiple scores, one for each Perspective attribute, indicating how likely a reader would perceive the input text as containing the given attribute. Perspective allows evaluation regarding multiple attributes. We selected six attributes, namely, toxicity, severe toxicity, insult, profanity, threat, and inflammatory, which, as per the definition presented in Appendix A.1 closely relate to our scope.[4] As input to Perspective, we consider two representations of each video's textual metadata: (i) the concatenation of the title and description and (ii) the caption. Also, Perspective does impose a 20 KB upper limit on the input. Thus, for (ii), we constrain the input to the initial 20KB of text for longer captions. As shown in Table 1, most videos in each category (especially for contrarian groups) have captions and, out of those, less than 35% of the videos have captions with size exceeding that limit. For both (i) and (ii), we keep our pre-processing to a minimum, removing only emojis and URLs, as Perspective is sensitive to various features such as the use of upper/lowercase and punctuation (Hosseini et al. 2017).

We analyze the scores of each Perspective attribute associated with the metadata of videos created by each community from two standpoints. First, we perform an aggregated analysis, including all videos in our dataset. We then analyze how the Perspective attribute scores of each community evolved over time using yearly average values for each feature and community. In both cases, we assess the differences in attribute scores across the communities from both statistical and practical perspectives. For the former, we first perform a Kruskal-Wallis test, a nonparametric method commonly used to compare multiple samples, to assess whether any of the four communities dominates the others with respect to the scores of each Perspective attribute. Next, we employ pairwise Mann–Whitney U tests to identify significant differences between the scores computed for specific pairs of communities. Moreover, regarding the yearly analysis, we employ the Mann–Whitney U test to assess differences between the means of the first and last years for all communities and attributes. The idea is that the test, paired with graphs for the yearly evolution of each attribute, will shed light into the changes over time. All tests are performed

[4] Further details on Perspective's attributes are available at https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages.

Table 2 Average scores of each perspective attribute

Attribute	Title and description				Caption			
	Media	IDW	Alt-lite	Alt-right	Media	IDW	Alt-lite	Alt-right
Toxicity	0.12733	0.20892	**0.22074**	**0.22028**	0.18613	0.26937	**0.28536**	0.26445
Severe toxicity	0.06269	0.12196	0.13039	**0.13709**	0.02102	**0.05355**	0.05132	0.04946
Insult	0.11523	0.20670	**0.21312**	0.19893	0.15753	0.24149	**0.26388**	0.22486
Profanity	0.06981	0.13219	**0.14780**	0.14566	0.1165	0.17936	**0.19354**	**0.19128**
Threat	0.14706	0.17379	0.19079	**0.21073**	0.05952	0.08283	0.09222	**0.10276**
Inflammatory	0.15744	0.23464	**0.29518**	0.29315	0.28203	0.4916	**0.55381**	0.53191

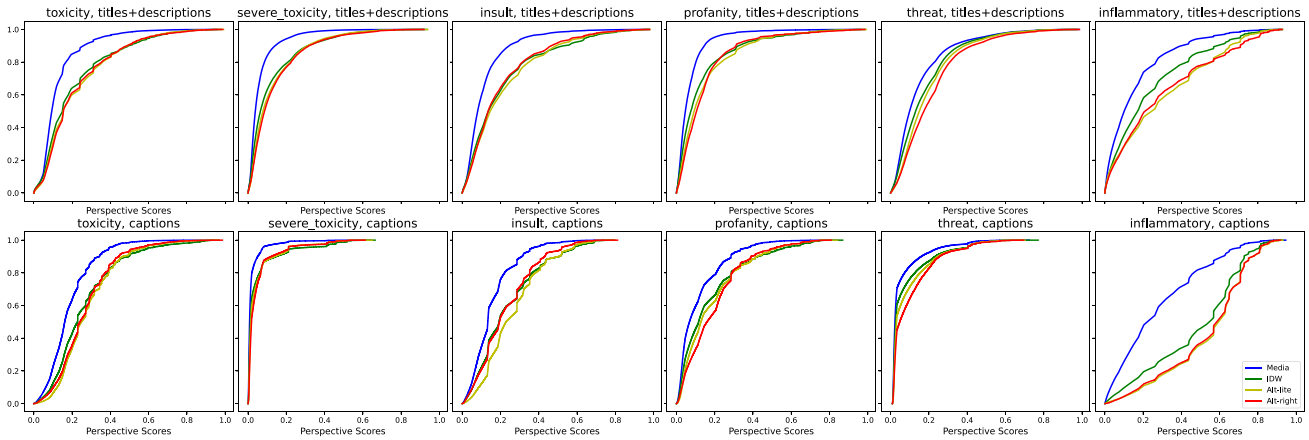

Fig. 2 Distribution of Persective attribute scores computed over the complete dataset. Results for titles and descriptions are shown in the top row, whereas results for captions are shown in the bottom row

with a level of significance α of 0.05 (i.e., statistical significance with p value < 0.05).

We start by presenting, in Table 2, the average scores for each Perspective attribute computed for each textual metadata, namely titles + descriptions and captions, and for each community, for the whole dataset. For all combinations of attribute and textual representation, the highest average values, along with other average values statistically tied with the highest, are shown in bold. We start by noting that in all cases the attribute values passed the Kruskal–Wallis test, implying that statistically significant differences (with p-value < 0.05) exist across the four communities. Indeed, as shown in the table, videos in the Media category stand out as much less toxic, according to all six attributes and for both metadata representations.

Focusing on the three contrarian groups, we observe statistically significant pairwise differences for some attributes and a few statistical ties. Yet, from a practical perspective, such differences are quite marginal in most cases, with the exception of the Inflammatory and, to a lesser extent, Threat attributes. Moreover, the category with the highest average score varies depending on the particular Perspective attribute and, in some cases, the textual representation used. For example, Alt-right exhibits the highest average severe toxicity for titles + descriptions (followed closely by Alt-lite). Yet, when captions are considered, the highest average is for IDW (with Alt-right having the smallest score). Also, for both textual representations, Alt-lite has the highest Insult score (on average) whereas Alt-right has the smallest one. For the inflammatory attribute, for which larger differences across the three groups are observed, Alt-lite stands out as the most toxic group, followed by Alt-right and IDW, for both metadata.

In addition to analyzing average attribute scores, we also look into the distributions of score values for all six Perspective attributes for the two textual representations and four communities. The cumulative distribution functions (CDFs) of the scores are shown in Fig. 2. In agreement with the results in Table 2, these graphs suggest that: (1) the Media category clearly exhibits much less toxic content than the contrarian groups according to all six Perspective attributes,[5] and (2) differences across the three contrarian groups, despite any statistical significance, are mostly marginal, with

[5] Interestingly, differences between Media and the contrarian groups are less noticeable for the Threat attribute, which might be due to nature of the news content often broadcasted by the Media channels.

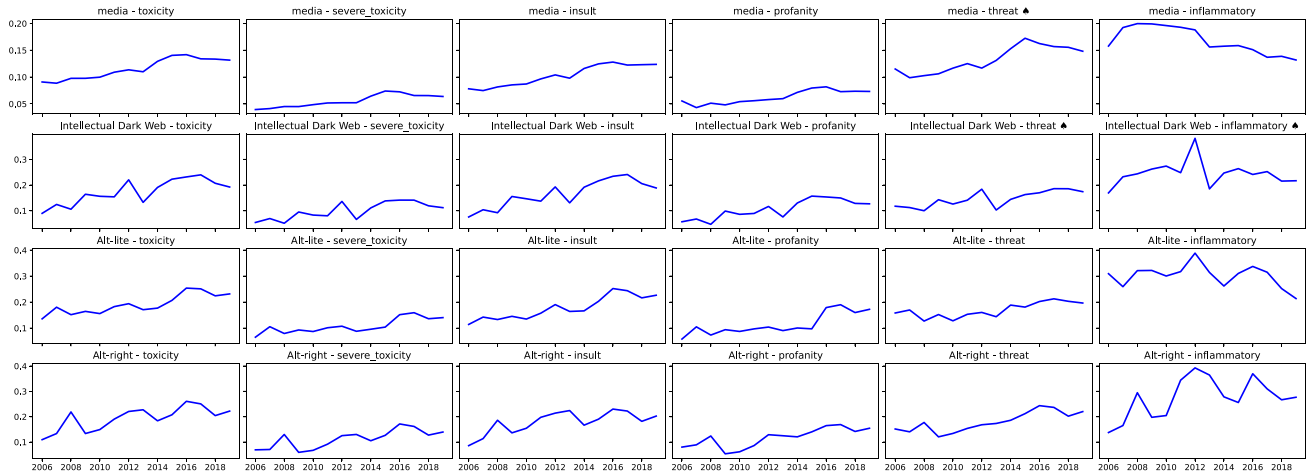

Fig. 3 Yearly average for features for titles and descriptions for all communities. *Y*-axis scale varies for each row of graphs. The ♠ symbol in graphs' titles emphasize cases with no statistically significant difference between the average values in 2006 and 2019

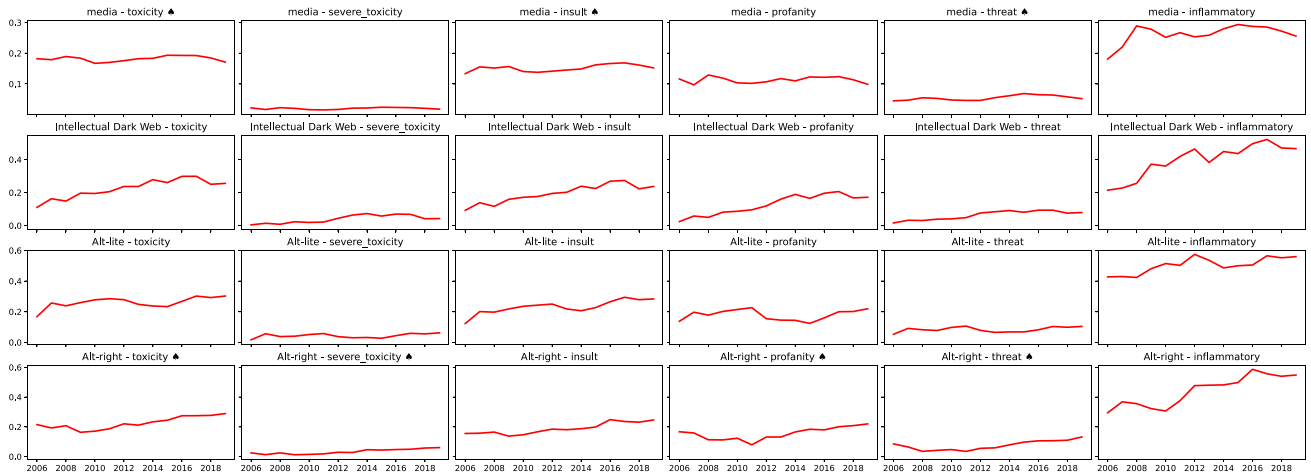

Fig. 4 Yearly average for features for captions for all communities. *Y*-axis scale varies for each row of graphs. The ♠ symbol in graphs' titles emphasize cases with no statistically significant difference between the average values in 2006 and 2019

the exception of the Inflammatory attribute, for which IDW is more clearly distinguished as less toxic than the other two communities.

Overall, these results suggest that there is no clear dominance (in toxicity) of any contrarian group over the other two. In particular, they suggest that the relative differences in the degree of extremism often reported in prior work (Ribeiro et al. 2020; Winter 2019; McClernan 2019; Kelsey 2020; Finlayson 2021; Atkinson 2018; Marantz 2017; Lewis 2018; Roose 2019a; Weiss and Winter 2018; Hosseinmardi et al. 2021) do not manifest themselves in the degree of toxicity of the content (according to Perspective) produced by these groups. Even more, depending on the specific toxicity-related attribute, the relative order among the three groups may be quite different from the one reported in the literature (Alt-right as the most extreme, followed by Alt-lite and IDW). The results discussed so far refer to all videos in our dataset, thus providing an aggregated toxicity analysis. We now analyze the toxicity of content produced yearly, to assess its evolution over time. We compare the average values of toxicity attribute scores for videos published each year for all four communities, six attributes and two metadata representations.

Figures 3 and 4 present yearly average values across all features for the four communities. Each row corresponds to a unique community. Figure 3 shows results for titles and descriptions, and Fig. 4 shows results for captions. We conducted the Mann–Whitney U test (with a p-value threshold of < 0.05) between each graph's initial and final years to evaluate statistical significance.

Except for threat for IDW and Media and Inflammatory for IDW, all other instances of titles and descriptions show statistical differences between average values of first and last years for features considering titles and descriptions. We also highlight two cases where values decreased over the years: Inflammatory for Media and IDW. However, overall, we see a trend of value increase over the years for most features. Regarding captions, seven of the 24 graphs show no statistical difference between the first and last years: 4 are from the Alt-right community, namely toxicity, severe toxicity, profanity, and threat. This result points to consistent feature scores, contrasting with the other communities over the years. For instance, IDW and Alt-lite showed increased toxicity over the years. All contrarians also show an increase in insult. Finally, all four communities showed an increase in inflammatory scores.

Previous works have pointed to an increase in radical content online from far-right communities and terrorist groups (Ribeiro et al. 2020; Neumann 2013; Ul Rehman et al. 2021; Nouh et al. 2019; Hosseinmardi et al. 2021). Our work adds to the literature by proposing a temporal analysis of the textual properties of the content spread by radical/extremist communities. The results of our temporal analyses point to an increase in toxicity on YouTube across all communities analyzed, considering titles and descriptions, and an escalation in the harmfulness of online speech over the years. Considering the amounts of years and videos analyzed, this result shows strong evidence of a shift in the type of language used online, specifically on YouTube. The change also impacts media channels, which, at first, should not be subject to statistically relevant differences in toxicity-related features due to the scope of their work. Recall that the list of media channels proposed by Ribeiro et al. (2020) is extensive and comprehensive: it contains channels from many parts of the political spectrum.

In summary, the results of our first research question, in addition to the temporal analyses, point to minor differences between contrarians, as presented in Fig. 2 and Table 2, which further question the validity, as referenced in previous works, of considering that these groups are different in terms of content produced.

5 RQ2: Psycholinguistic attributes

We now turn to our RQ2: *From a psycholinguistic perspective, what are the main differences among the groups?* To address this question, we follow previous work (Resende et al. 2019; Caetano et al. 2022; Malagoli et al. 2021; Matos et al. 2022) and employ the Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker 2010) lexicon to analyze psycholinguistic attributes of an input text. LIWC classifies words into 73 attributes that express multiple traits, such as writing style, cognitive and affective concepts, and grammatical classes.

Specifically, we employ LIWC to the metadata of each video created by each of the four communities, taking each textual representation, namely title and description concatenation or caption, separately. For each input (metadata) content, we compute the frequency of occurrence of each LIWC attribute. We then perform pairwise comparisons across the four communities with respect to the distribution of frequency of occurrence of each LIWC attribute, aiming at identifying significant differences in specific attributes. To that end, we employ the Kolmogorov–Smirnov (KS) test (with $p < 0.05$), which is a nonparametric test of the equality of continuous distributions.

Finally, for the sake of presentation, we focus on those cases (LIWC attribute and pair of communities) for which the compared distributions are considered statistically different and compute the overall frequency across all metadata associated with videos of each involved community. We then report the relative difference of such aggregated frequencies between the pair of communities.

For several cases, although the attribute distributions are statistically different (according to the KS test), the relative difference is too small to be significant from a practical perspective. Thus, we chose to focus on those cases for which such differences are above 10%. Figure 5 shows these results for comparisons based on titles + descriptions (graphs in the top row) and captions (graphs in bottom row). Note that the title of each graph identifies the pair of communities to which the results refer. For example, a graph entitled *A x B* reports all attributes for which the relative difference of frequencies between communities *A* and *B* (i.e., $\frac{A-B}{A}$) exceeds 10%.

We start by noting that the relative differences are much larger for comparisons between any contrarian group and Media than for comparisons involving two contrarian communities (note the different scales in the *x*-axes of the graphs). That is, once again, we see a very clear distinction between Media and the contrarians. Notably, contrarians tend to use more swear words, sexual terms as well as terms related to religion, anger, and other negative emotions. Interestingly, the largest discrepancies occur between Alt-lite and Media. These observations hold for both titles + descriptions and captions.

Among the contrarians, we do observe some distinctions, though these are much more subtle. Some of them refer to the use of specific linguistic properties, such as 3rd person singular pronouns (*she/he)* and articles, which are much more frequently used in the titles + descriptions of the videos created by the IDW community than in those for Alt-lite and Alt-right. Others refer to specific topics or jargon. For example, Alt-lite uses swear words, sexual, death, and anger-related terms much more frequently

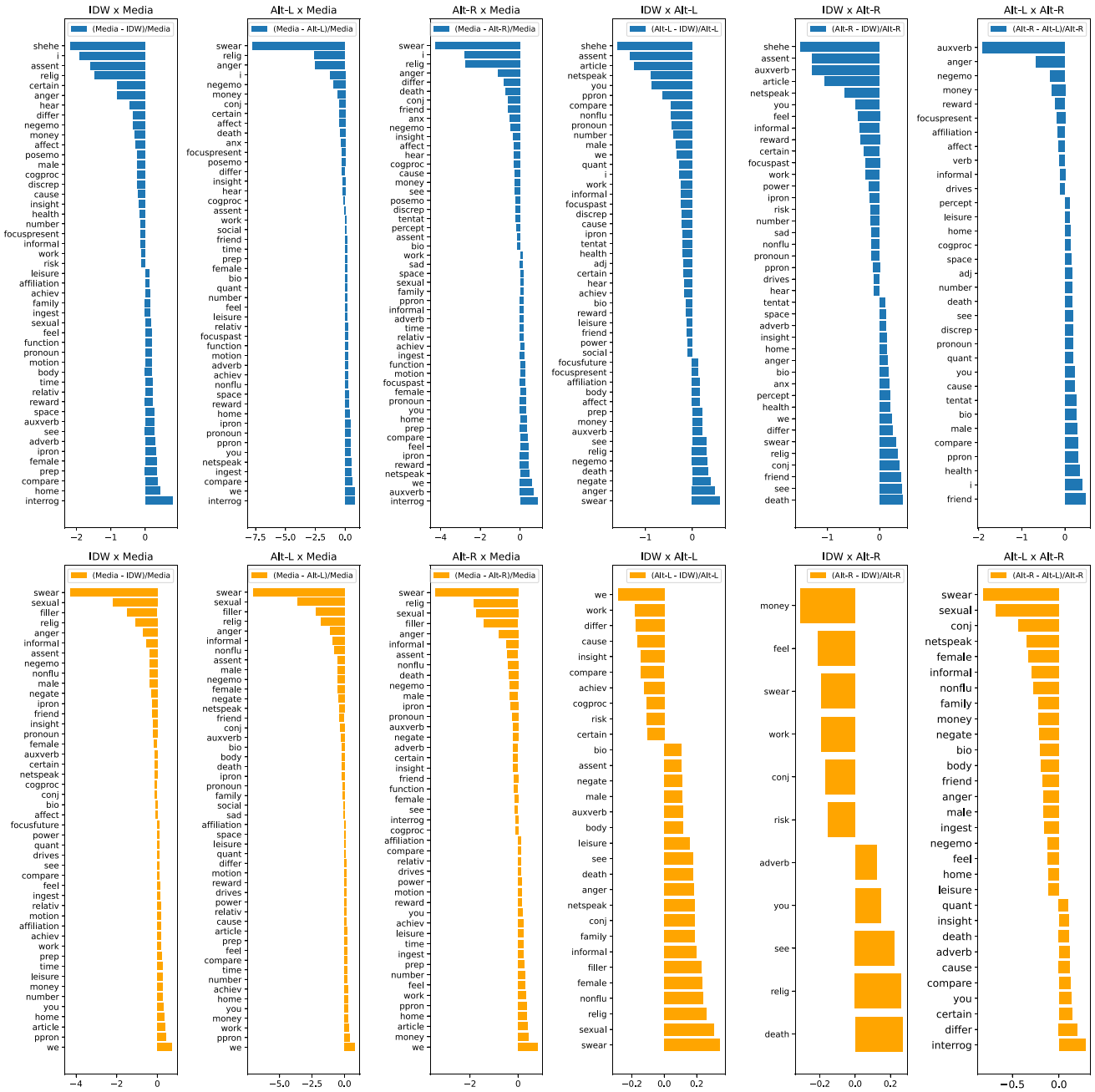

Fig. 5 LIWC attributes for which the pairwise comparison across communities suggest significant differences, with a relative difference in the overall occurrence frequency exceeding 10%. Top row refers to titles + descriptions, while bottom row refers to captions

than IDW. Similarly, Alt-right also uses terms related to death, anger and religion more often than IDW. IDW, in turn, uses Internet slang (*netspeak*) more often. Comparing Alt-lite and Alt-right, the former uses terms related to negative emotions and anger more often, whereas the latter makes more use of death-related words. Nevertheless, as mentioned, differences among contrarians are quite nuanced. Indeed, as shown in the figure, they exhibit much more attribute similarities than differences. The Media category, in turn, once again sets itself apart with very distinct usage patterns. In consonance with the outcomes of our first research question, our results for RQ2 show that differences between contrarians, as described in the previous works, are not clearly present and may be completely artificial.

6 RQ3: Semantic (contextual) and lexical analysis using text classifiers

We finally turn to our RQ3: *From semantic (contextual representation) and lexical perspectives, what are the main differences among the groups?* To tackle this question, we aim to assess differences among the textual features of the four considered groups based on contextual representations associated with their videos. We fine-tune a state-of-the-art contextual language model (BERT (Devlin et al. 2018)) to produce embeddings for the video's textual content. More details regarding the fine-tuning process can be found in the Appendix A.3. These embedded representations are projections of the videos in a vector space, in which textually similar videos are close in said space.

We use the fine-tuned representation combined with the TSNE algorithm (van der Maaten and Hinton 2008) to generate a two-dimension plot of a 1000 video sample. This algorithm projects high-dimensional embeddings into a two-dimensions space maintaining their relative positions on the original space. To further analyze class separability, we evaluate the Silhouette Score over the textual embeddings, considering their categories (Media, IDW, Alt-lite, or Alt-right) as clusters. This metric of cluster quality captures both intra-cluster cohesion and inter-cluster separability. Values range from −1 (worst) to +1 (best), with values close to 0 suggesting overlapping clusters (Rousseeuw 1987). We also measure the Separability Index (Zighed et al. 2002), which computes the percentage of entries in the data with the nearest neighbor of the same class. Finally, we compute Information Gain[6] for each pair of categories using unigrams to identify the top k ($k = 10$) terms that better discriminate videos across categories on the concatenation of video titles and descriptions.

Table 3 shows Precision, Recall, F1, and Separability Index per class for the average of the 5 test folds used in our experimental procedure. Precision measures the percentage of correct entries, Recall measures the percentage of all entries correctly predicted, and F1 is the harmonic mean of both measures. As we can see, the Media group is very distinguishable from the others, with high Precision and moderate Recall values, resulting in a good F1 as well as a high separability index. This means that Media presents different contextual features, such as terms, expressions, and themes for the videos when compared to the other groups. Regarding the contrarian groups, results are usually very low for the three classification metrics, approaching the random classification for Alt-Right and Alt-Lite and being a bit better for IDW. They also present a lower separability index

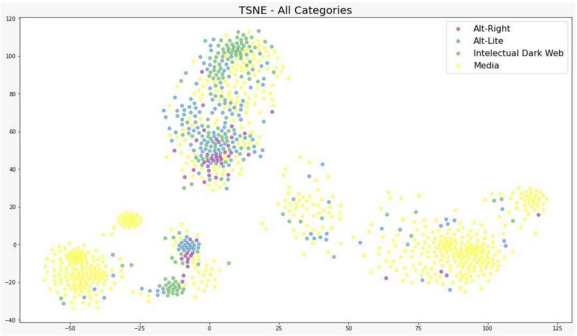

Fig. 6 TSNE Plots for a sample of 1000 videos using fine-tuned embeddings

Table 3 Precision, recall, F1-score, and separability index (SP) for the classes

	Precision	Recall	F1-score	SP
Alt-Right	0.21	0.63	0.31	0.72
Alt-Lite	0.36	0.31	0.33	0.84
IDW	0.39	0.53	0.45	0.83
Media	0.92	0.72	0.81	0.94
Weighted Avg	0.70	0.62	0.65	0.89

Table 4 Confusion matrix for the classifier

		Predicted class			
		Alt-Right	Alt-Lite	IDW	Media
True class	Alt-Right	3.08	0.82	0.69	0.29
	Alt-Lite	4.56	4.98	4.69	1.90
	IDW	3.15	2.92	8.82	1.85
	Media	3.94	4.93	8.44	44.8

Values in percentage

when compared to Media, specially Alt-Right. This means that the classifier could not effectively distinguish among these classes, which are hard to separate in the embedding space due to high similarity in content.

To further analyze these results, Table 4 shows the confusion matrix regarding the average of the 5 test folds. The values shown in the table correspond to the percentage of all entries in each case. Thus, the sum of all cells represents 100% of the data. The cases represent the combination of predicted and correct classes for each entry. Therefore, correct cases are in the diagonal, and all misclassified entries are in the remaining cells. For instance, considering the Alt-Lite group, 28,27% of true Alt-Lite instances were wrongly classified as Alt-Right. Thus, in consonance with the medium average classifier's F1 of 0.65, the matrix shows low generalization across all classes except for media. This

[6] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html.

Table 5 Silhouette coefficient for all pairs of communities

Representation	Media/IDW	Media/Alt-lite	Media/Alt-right	IDW/Alt-lite	IDW/Alt-right	Alt-lite/Alt-right
BERT embeddings	0.102	0.112	0.165	0.021	0.062	0.030
Perspective API	0.179	0.22	0.285	0.009	0.028	0.003

Table 6 Top 10 terms with the largest score

Categories pair	Most discriminative terms
Alt-Right vs. Alt-Lite	Ice, red, therebel, platforms, intended, 0x956e7af6706c3b5e2cf7e15c16c7018c4f42af79litecoin, 17q1bff2up8orekn8dqgpepx83rfbaz5qlethereum, reports, media, literary
Alt-Right vs. IDW	Ice, red, radio, follow, rogan, joe, copyright, bloggingheads, facebook, alt
Alt-Right vs. Media	Subscribe, facebook, paypal, follow, news, patreon, support, gab, instagram, books
Alt-Lite vs. IDW	Therebel, platforms, follow, rogan, wordpress, 0x956e7af6706c3b5e2cf7e15c16c7018c4f42af79litecoin, 17q1bff2up8orekn8dqgpepx83rfbaz5qlethereum, donation, blogger, joe
Alt-Lite vs. Media	Patreon, support, gab, follow, therebel, bitchute,\\subscribe, paypal, minds, wordpress
IDW vs. Media	Feed, shapiro, bloggingheads, rogan, instagram,\\copyright, read, support, paypal, recorded

demonstrates that identifying the groups is a hard task, given the videos' text lexical and contextual features. To further illustrate this point, Fig. 6 shows a representation of the embedding space using TSNE, and the classes overlap. Each plot dot represents a video (content); the color corresponds to the respective classes. As we can see, Media (in yellow) presents more cohesive clusters and is more spread across the space due to their higher volume and diverse channels. Contrarian videos do not really form cohesive clusters and are usually placed mixed with videos of other contrarian classes as well as of Media. This visualization is consistent with all other previous results.

Finally, Table 5 shows our silhouette analysis using the contextual embeddings for each group as a cluster. The results show weak separability between contrarian groups. For instance, the Silhouette coefficients for the **Alt-lite vs. IDW** and **Alt-lite vs. Alt-Right** pairs of clusters using BERT embeddings are close to zero, indicating that the clusters are almost indistinguishable. The highest separability scores for pairs of groups are obtained between the Media and the other groups, indicating that this group is more distinguishable. In sum, contrarian groups share significant semantic (contextual) similarities in our analysis, which is also consistent with the previous analyses.

We analyzed the pairwise information gain of the terms present in all classes to gain insights into the features that the classifier emphasizes to predict the groups. Table 6 shows the results, with the top 10 terms with the highest information gain scores, considering unigrams, for each pair of classes. This analysis can provide insights into the inner workings of the classifier. We highlight some insightful terms for all group pairs. **Alt-right/Alt-lite**: Terms include far-right media conglomerates: "red ice" and "therebel media"; Red Ice has been described as a white supremacist company (Willingham 2018). We also observe cryptocoins (Litecoin and Ethereum) wallet hashes. **Alt-right/IDW**: highlighted terms include Joe Rogan (famous IDW podcast host) and bloggingheads (podcast-like video blog). **Alt-right/Media**: Discriminative terms include funding platforms (Patreon, PayPal) and Gab, a fringe social network. **Alt-lite/IDW**: mentions to The Rebel Media, Joe Rogan, and cryptocoin wallet hashes, which consonates with previous reports of these groups using cryptocurrencies to raise funds (League 2017). **Alt-lite/Media**: Terms mention funding platforms (Patreon, PayPal) and Gab. Overall, most of the discriminative terms are proper nouns, indicating differences predominantly in entities belonging to the specific categories. In other words, most differences are lexical. This is consistent with the low accuracy of the classifier and the low separability results found in the silhouette score analysis of the clustering performed over the BERT's contextual embeddings, which are more semantically oriented.

In summary, these results consistently show that, unlike what is considered in previous works, all contrarian groups share a significant number of similarities. These similarities are not limited the toxicity and psycholinguistic attributes of their content, as RQ1 and RQ2 have shown, but also extend to contextual and lexical features. This finding challenges the common belief that these groups are not only different but also exhibit varying degrees of extremism. We believe that these observed similarities suggest that the differences among these groups might not be as significant as previously believed. More research is needed to fully understand the nature and implications of these shared characteristics.

Table 7 Perspective's definitions of analyzed attributes

Attribute name	Description
Toxicity	"A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
Severe Toxicity	"A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words."
Insult	"Insulting, inflammatory, or negative comment toward a person or a group of people."
Profanity	"Swear words, curse words, or other obscene or profane language."
Threat	"Describes an intention to inflict pain, injury, or violence against an individual or group."
Inflammatory	"Intending to provoke or inflame."

7 Limitations

Although our analyses rely on a large number of videos from multiple channels, we focus on three contrarian groups, which, as communities, are only partially representative of YouTube's ecosystem. Furthermore, we only analyze YouTube content, a moderated platform. The presence of moderation in an OSN may impact the analysis of extremists by preventing them from sharing more fringe and controversial opinions (due to fear of being banned), or lightening the kind of speech they propagate. Specific to our methods, for instance, this may impact the toxicity-related scores inferred from collected data. Additionally, we acknowledge that radicalization is a multi-faceted concept with no universal definition, prompting us to use a specific, but widely referenced one (McCauley and Moskalenko 2008) in our work.

Another limitation of our work is that our analyses ignore comments and other metadata associated with the video content, focusing only on video titles, descriptions, and captions. Further investigations using comments can unravel additional viewpoints, including what kind of textual patterns and features users usually use to interact with contrarian content on YouTube.

Finally, part of our analyses relies heavily on the scores produced by Google Perspective API for multiple features. However, previous work has shown that the API can be deceived and assign much lower toxicity scores for a sentence if it is subtly modified (Hosseini et al. 2017), and even present racial biases (Sap et al. 2019). Moreover, the scores are hard to interpret. Nevertheless, Perspective remains the mostly used tool to assess toxicity-related features in textual content (Guimarães et al. 2020; Mittos et al. 2020; Obadimu et al. 2019; Matos et al. 2022; Hoseini et al. 2023; Alharthi 2021; Ribeiro et al. 2021b; Rye et al. 2020; Zannettou et al. 2020; Ali et al. 2021), in addition to being freely accessible. The latter facilitates reproducibility, which is highly desirable for a scientific work. Lastly, although our classifier delivers meaningful insights about the indistinguishability of content produced across contrarian groups, it provides little insight into why that is the case, which can be improved by adding explainability frameworks, such as BertViz (Vig 2019).

8 Conclusions and future work

Contrarian communities are an integral part of online fringe groups. Our work focuses on contrarian groups (namely Intellectual Dark Web, Alt-lite, and Alt-right), analyzing their patterns of **content production**, setting itself apart from previous works, which focused on **content consumption**. In this context, we performed the first large-scale analysis of textual metadata produced by these contrarian groups. We analyzed titles, descriptions, and captions from over 355,000 videos from 350 YouTube channels. Our results indicate similar distributions of toxicity and psycholinguistics features among contrarians' textual content, despite claims in the literature that they are different groups, contrasting with the results obtained comparing contrarians with Media. We also performed a textual content classification, which revealed that contrarian groups share significant similarities from a lexical and semantic perspective, which also set them apart from traditional Media. Overall, our results indicate that from a perspective of the content produced by contrarians, these groups are similar, as the speeches propagated by them are closely entangled, even when analyzed using different techniques. Thus, we suggest that this discovery should prompt the community to reconsider this hypothesis of differences across the three contrarian groups, in terms of content produced, and cease endorsing it without adequate validation and analysis. Even more, we offer evidence that it may indeed be a misconception and deserves further investigation. Finally, we observed an increase in scores for toxicity-related features over time, suggesting an increase in online abusive language over the years, even for traditional Media.

For future work, we want to explore additional techniques, such as sentiment analysis (Viegas et al. 2020) and topic modeling (Viegas et al. 2020b), to gain insights from additional standpoints. Also, in future work, we will propose

an analysis of comments to assess toxicity, sentiment, and others. Finally, we will expand our research with model interpretability frameworks, such as BertViz (Vig 2019).

A Appendix

This appendix aims to provide additional information on some the techniques used in this paper.

A.1 Perspective attributes

In this section of the Appendix, we will go into more detail on Perspective API. Perspective API scores text on the impact said text may have on the reader and is widely used in the literature, as mentioned in Sect. 3 (Methodology). Perspective can provide scores for many attributes, and in this work, we used Perspective to infer scores for "toxicity," "severe toxicity," "insult," "profanity," "threat," and "inflammatory." Table 7 displays definitions of each attribute. Although Perspective allows users to use experimental (i.e., not thoroughly tested yet), we pertained to attributes used in production. Further details on additional attributes are available in Perspective's documentation.[7]

Finally, although Perspective's implementation is not open-source, their team has released information on how the current system was trained and deployed, including the pretraining of the model (Lees et al. 2022).

A.2 LIWC

The foundation of Linguistic Inquiry and Word Count (LIWC) stems from extensive scientific research spanning decades, showcasing the capacity of language to offer profound insights into individuals' psychological states, encompassing emotions, cognitive styles, and social concerns. While some connections are straightforward, like the use of positive words indicating happiness, such as "happy," "excited," and "elated," many relationships between verbal expression and psychology are less apparent. For instance, higher social standing and confidence are linked to elevated use of "you" words and reduced use of "me" words. LIWC relies on decades of empirical research and provides specialized means to comprehend, elucidate, and quantify psychological, social, and behavioral phenomena.

LIWC is a text analysis program that analyzes individual or multiple language files quickly and efficiently. It is designed to be transparent and flexible, allowing users to explore word use in various forms. LIWC is used in research to analyze the ways people use words when communicating, which can provide rich information about their beliefs, fears, thinking patterns, social relationships, and personalities. Further details on how LIWC was built are available in its documentation (Pennebaker et al. 2015). The extensive research employed in developing LIWC motivated us to use it in our methodology. In our work, we employed LIWC to analyze each word of an input text automatically, attributing it to a psycholinguistic class. Then, it calculates the overall frequency of each one of its categories in the input text. We relied on the frequency report returned by LIWC for the analyses of our second research question, implementing minor pre-processing, namely the removal of URLs and covert all text inputs to lowercase.

A.3 Embeddings

For the fine tuning, the first step is the prepare data for training. Given that, the training data for the fine-tuning is very skewed (see Table 1), with the Media category containing the most entries. To avoid learning biases, we employed an under-sampling strategy to build a balanced subsample of the training set prior to the classification analysis. The sampling strategy randomly selects 17k entries from each category based on the size of the smallest category (Alt-right), resulting in 68k entries for fine-tuning. Rather than focusing on the final accuracy of the classifier, which could benefit from more data, our main interest is in evaluating the model's capability of discriminating among the categories under similar conditions. The model consists of a classification layer over the BERT Tiny pre-trained model,[8] which is chosen over Vanilla BERT due to resource limitations. The model consists of a classification layer over the BERT Tiny pre-trained model, which has slightly superior classification effectiveness compared to BERT Tiny (Jiao et al. 2020), but has a much higher training cost.

We employ a five-fold cross-validation procedure to assess the classification model's discriminative capability. Data is split into five partitions, with four used for training and one for testing. The procedure is repeated five times with different training/test partitions, and the reported results are averages over the 5 test partitions. The model is trained for five epochs with 512 as the max input size of tokens, the standard maximum BERT-like model implementations, and a batch size of 16 entries as the maximum allowed due to resource restrictions. Other parameters are the default of the HuggingFace's trainer,[9] representing standard values. We use the [CLS] token output to capture contextual embedding representations for all entries of the balanced dataset sample. BERT represents a sentence as a sequence of hidden states,

[7] https://perspectiveapi.com/how-it-works/.

[8] https://huggingface.co/prajjwal1/bert-tiny.

[9] https://huggingface.co/docs/transformers/main_classes/trainer.

which must be reduced to a single vector for downstream tasks. Therefore, BERT prepends a [CLS] token (short for "classification") at the beginning of each sentence and uses a more straightforward method of taking the hidden state corresponding to the first token.

Acknowledgements We thank Ribeiro et al. (2020) for kindly sharing the dataset with us. This work was partially supported by the authors' individual grants from CNPq, CAPES, and FAPEMIG.

Author contribution All authors contributed to the study conception and design. Material preparation and data collection were performed by [BM] . Analyses for RQ1 and RQ2 were performed by [BM], while analyses for RQ3 were performed by [RCL]. The first draft of the manuscript was written by [BM] and [RCL] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

ADL (2017) Glossary terms alt-right. https://www.adl.org/resources/glossary-terms/alt-right

ADL (2019) From alt right to alt lite: naming the hate. https://www.adl.org/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate

Alharthi R (2021) Recognizing hate-prone characteristics of online hate speech targets. In: Companion publication of the 13th ACM web science conference 2021, WebSci '21 companion, pp 153–155, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450385251. https://doi.org/10.1145/3462741.3466676

Ali S, Saeed MH, Aldreabi E, Blackburn J, De Cristofaro E, Zannettou S, and Stringhini G (2021) Understanding the effect of deplatforming on social networks. In: Proceedings of the 13th ACM web science conference 2021, WebSci '21, pp 187–195, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450383301. https://doi.org/10.1145/3447535.3462637

Arnold NA, Steer B, Hafnaoui I, H A Parada G, Mondragón RJ, Cuadrado F, and Clegg RG (2021) Moving with the times: Investigating the alt-right network gab with temporal interaction graphs. In: Proceedings of the ACM on human–computer interaction, 5(CSCW21). https://doi.org/10.1145/3479591

Atkinson DC (2018) Charlottesville and the alt-right: a turning point? Politics Groups Identities 6(2):309–315. https://doi.org/10.1080/21565503.2018.1454330

Bartlett J, Miller C (2012) The edge of violence: towards telling the difference between violent and non-violent radicalization. Terror Political Violence 24(1):1–21

Borum R (2011) Radicalization into violent extremism I: a review of social science theories. J Strateg Secur 4(4):7–36

Bryant LV (2020) The youtube algorithm and the alt-right filter bubble. Open Inf Sci 4(1):85–90. https://doi.org/10.1515/opis-2020-0007

Caetano J, Guimarães S, Araújo MMR, Silva M, Reis JCS, Silva APC, Benevenuto F, Almeida JM (2022) Characterizing early electoral advertisements on twitter: a Brazilian case study. In: Hopfgartner F, Jaidka K, Mayr P, Jose J, Breitsohl J (eds) Social informatics. Springer International Publishing, Cham, pp 257–272

Chipidza W (2021) The effect of toxicity on Covid-19 news network formation in political subcommunities on reddit: an affiliation network approach. Int J Inf Manag 61:102397. https://doi.org/10.1016/j.ijinfomgt.2021.102397

Coleman A (2022) Fact-checkers label youtube a "major conduit of online disinformation". https://www.bbc.com/news/technology-59967190

Dalgaard-Nielsen A (2010) Violent radicalization in Europe: What we know and what we do not know? Stud Confl Terror 33(9):797–814

Das S (2023) Laughing bodies and the tickle machine: understanding the youtube pipeline through alt-right humour. J Cult Res, pp 1–15

de Andrade CM, Belém FM, Cunha W, França C, Viegas F, Rocha L, Gonçalves MA (2023) On the class separability of contextual embeddings representations—or "the classifier does not matter when the (text) representation is so good!''. Inf Process Manag 60(4):103336. https://doi.org/10.1016/j.ipm.2023.103336

de Andrade CM, Belém FM, Cunha W, França C, Viegas F, Rocha L, Gonçalves MA (2023) On the class separability of contextual embeddings representations—or "the classifier does not matter when the (text) representation is so good!''. Inf Process Manag 60(4):103336. https://doi.org/10.1016/j.ipm.2023.103336

Devlin J, Chang M-W, Lee K, and Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

Finlayson A (2021) Neoliberalism, the alt-right and the intellectual dark web. Theory Cult Soc 38(6):167–190

Grant N (2022) Youtube may have misinformation blind spots, researchers say. https://www.nytimes.com/2022/11/05/technology/youtube-misinformation.html

Guimarães SS, Reis JCS, Ribeiro FN, and Benevenuto F (2020) Characterizing toxicity on facebook comments in Brazil. In WebMedia, WebMedia '20, pp 253–260, New York, NY, USA. ACM. ISBN 9781450381963. https://doi.org/10.1145/3428658.3430974

Hafez M, Mullins C (2015) The radicalization puzzle: a theoretical synthesis of empirical approaches to homegrown extremism. Stud Confl Terror 38(11):958–975

Hoseini M, Melo P, Benevenuto F, Feldmann A, and Zannettou S (2023) On the globalization of the qanon conspiracy theory through telegram. In: Proceedings of the 15th ACM web science conference 2023, WebSci '23, pp 75–85, New York, NY, USA. Association for Computing Machinery. ISBN 9798400700897. https://doi.org/10.1145/3578503.3583603

Hosseini H, Kannan S, Zhang B, and Poovendran R (2017) Deceiving Google's perspective api built for detecting toxic comments. arXiv:1702.08138

Hosseinmardi H, Ghasemian A, Clauset A, Mobius M, Rothschild DM, and Watts DJ (2021) Examining the consumption of radical content on youtube. In: Proceedings of the national academy of sciences, 118(32): e2101967118. https://doi.org/10.1073/pnas.2101967118. https://www.pnas.org/doi/abs/10.1073/pnas.2101967118

Ingram M (2018) Most Americans say they have lost trust in the media. https://www.cjr.org/the_media_today/trustin-media-down.php

Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, and Liu Q (2020) TinyBERT: Distilling BERT for natural language understanding. In EMNLP, pp 4163–4174. ACL. https://doi.org/10.18653/v1/2020.findings-emnlp.372. https://aclanthology.org/2020.findings-emnlp.372

Kelsey D (2020) Archetypal populism: The "Intellectual Dark Web" and the "Peterson Paradox", pp 171–198. Springer International Publishing, Cham. ISBN: 978-3-030-55038-7. https://doi.org/10.1007/978-3-030-55038-7_7

King M, Taylor DM (2011) The radicalization of homegrown Jihadists: a review of theoretical models and social psychological evidence. Terror Political Violence 23(4):602–622

League A-D (2017) Funding hate: How white supremacists raise their money. New York

Lees A, Tran VQ, Tay Y, Sorensen J, Gupta J, Metzler D, and Vasserman L (2022) A new generation of perspective API: efficient

multilingual character-level transformers. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, KDD '22, pp 3197–3207, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450393850. https://doi.org/10.1145/3534678.3539147

Lewis R (2018) Alternative influence: broadcasting the reactionary right on Youtube

Li HO-Y, Bailey A, Huynh D, Chan J (2020) Youtube as a source of information on Covid-19: a pandemic of misinformation? BMJ Glob Health 5(5):e002604

Lima L, Reis JC, Melo P, Murai F, and Benevenuto F (2020) Characterizing (UN) moderated textual data in social systems. In: ASONAM, pp 430–434. IEEE

Malagoli LG, Stancioli J, Ferreira CHG, Vasconcelos M, Couto da Silva AP, and Almeida JM (2021) A look into covid-19 vaccination debate on twitter. In: 13th ACM web science conference 2021, WebSci '21, pp 225–233, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450383301. https://doi.org/10.1145/3447535.3462498

Mamié R, Horta Ribeiro M, and West R (2021) Are anti-feminist communities gateways to the far right? Evidence from reddit and youtube. In: Proceedings of the ACM WebSci '21

Manifest I The intellectual dark web. Undisclosed. https://web.archive.org/web/20190407170300/, http://intellectualdark.website/

Marantz A (2017) The alt-right branding war has torn the movement in two. https://www.newyorker.com/news/news-desk/the-alt-right branding-war-has-torn-the-movement-in-two

Matos B, Lima RC, Almeida JM, Gonçalves MA, Santos RLT (2022) On the presence of abusive language in MIS/disinformation. In: Hopfgartner F, Jaidka K, Mayr P, Jose J, Breitsohl J (eds) Social informatics. Springer International Publishing, Cham, pp 292–304 (**ISBN 978-3-031-19097-1**)

McCauley C, Moskalenko S (2008) Mechanisms of political radicalization: pathways toward terrorism. Terrorism and political violence 20(3):415–433

McClernan N (2019) Steven pinker's right-wing, alt-right & hereditarian connections

Milmo D (2022) Youtube is major conduit of fake news, factcheckers say. https://www.theguardian.com/technology/2022/jan/12/youtube-is-major-conduit-of-fake-news-factcheckers-say

Mittos A, Zannettou S, Blackburn J, De Cristofaro E (2020) "and we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. In: Proceedings of the international AAAI ICWSM 14:452–463

Moffitt B (2023) What was the 'alt' in alt-right, alt-lite, and alt-left? on 'alt' as a political modifier. Political Stud 00323217221150871. https://doi.org/10.1177/00323217221150871

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 Proceedings of the WWW '18, WWW '18, pp 621–628, Republic and Canton of Geneva, CHE. WWW '18. ISBN: 9781450356404. https://doi.org/10.1145/3184558.3188733

Nagle A (2017) Kill All Normies: Online Culture Wars From 4Chan And Tumblr To Trump And The Alt-Right. Zero Books, Alresford, GBR. 1785355430

Neumann PR (2013) Options and strategies for countering online radicalization in the United States. Stud Confl Terror 36(6):431–459

Newman N, Fletcher R, Schulz A, Andi S, Robertson CT, and Nielsen RK (2021) Reuters institute digital news report 2021. Reuters Institute for the study of Journalism

Niu S, Mai C, McKim KG, and McCrickard S (2021) #teamtrees: Investigating how youtubers participate in a social media campaign. In: Proceedings of the ACM on human–computer interaction, 5(CSCW21). https://doi.org/10.1145/3479593

Nouh M, Nurse JR, and Goldsmith M (2019) Understanding the radical mind: identifying signals to detect extremist content on twitter. In: 2019 IEEE international conference on intelligence and security informatics (ISI), pp 98–103. https://doi.org/10.1109/ISI.2019.8823548

Obadimu A, Mead E, Hussain MN, and Agarwal N (2019) Identifying toxicity within Youtube video comment text data. In: SBP-BRiMS '19, pp 214–223. Springer

O'Malley RL, Holt K, and Holt TJ (2022) An exploration of the involuntary celibate (incel) subculture online. https://doi.org/10.1177/0886260520959625. PMID: 32969306

Ottoni R, Cunha E, Magno G, Bernardina P, Meira Jr W, and Almeida V (2018) Analyzing right-wing Youtube channels: Hate, violence and discrimination. In: Proceedings of ACM WebSci, WebSci '18, pp 323–332, New York, NY, USA. ACM. ISBN: 9781450355636. https://doi.org/10.1145/3201064.3201081

Pennebaker JW, Boyd RL, Jordan K, and Blackburn K (2015) The development and psychometric properties of liwc2015. Technical report

Resende G, Melo P, Reis JCS, Vasconcelos M, Almeida JM, and Benevenuto F (2019) Analyzing textual (MIS)information shared in Whatsapp groups. In: Proceedings of the 10th ACM conference on web science, WebSci '19, pp 225–234, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450362023. https://doi.org/10.1145/3292522.3326029

Ribeiro MH, Ottoni R, West R, Almeida VA, and Meira Jr W (2020) Auditing radicalization pathways on Youtube. In: Proceedings of ACM FAT*, pp 131–141

Ribeiro MH, Blackburn J, Bradlyn B, De Cristofaro E, Stringhini G, Long S, Greenberg S, Zannettou S (2021) The evolution of the manosphere across the web. Proc ICWSM 15(1):196–207

Ribeiro MH, Jhaver S, Zannettou S, Blackburn J, Stringhini G, De Cristofaro E, and West R (2021b) Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. In: Proceedings of the ACM on human–computer interaction, 5(CSCW2). https://doi.org/10.1145/3476057

Roose K (2019a) The making of a Youtube radical. The New York Times, 8

Roose K (2019b) The making of a Youtube radical. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rye E, Blackburn J, and Beverly R (2020) Reading in-between the lines: an analysis of dissenter. In: Proceedings of the ACM internet measurement conference, IMC '20, pp 133–146, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450381383. https://doi.org/10.1145/3419394.3423615

Sap M, Card D, Gabriel S, Choi Y, Smith AN (2019) The risk of racial bias in hate speech detection. In: Proc ACL

Sellars A (2016) Defining hate speech. Berkman Klein Center Research Publication 2016–20:16–48

Tang L, Fujimoto K, Amith MT, Cunningham R, Costantini RA, York F, Xiong G, Boom JA, Tao C (2021) "down the rabbit hole'' of vaccine misinformation on youtube: network exposure study. J Med Internet Res 23(1):e23262. https://doi.org/10.2196/23262

Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54

Thorburn J, Torregrosa J, and Panizo Á (2018) Measuring extremism: validating an alt-right twitter accounts dataset. In: IDEAL 2018, pp 9–14. ISBN: 978-3-030-03496-2

Ul Rehman Z, Abbas S, Khan MA, Mustafa G, Fayyaz H, Hanif M, and Saeed MA (2021) Understanding the language of ISIS: an empirical approach to detect radical content on twitter using machine learning. Comput Mater Continua, 66(2)

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605

Viegas F, Alvim MS, Canuto SD, Rosa T, Gonçalves MA, Rocha L (2020) Exploiting semantic relationships for unsupervised expansion of sentiment lexicons. Inf Syst 94:101606. https://doi.org/10.1016/j.is.2020.101606

Viegas F, Cunha W, Gomes C, Júnior APDS, Rocha L, and Gonçalves MA (2020b) Cluhtm—semantic hierarchical topic modeling based on cluwords. In: D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, Online, July 5–10, 2020, pp 8138–8150. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.724

Vig J (2019) A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations, pp 37–42, Florence, Italy. Association for Computational Linguistics. 10.18653/v1/P19-3007. https://www.aclweb.org/anthology/P19-3007

Weiss B, and Winter D (2018) Meet the renegades of the intellectual dark web. The New York Times, 8

Willingham A (2018) Middle school teacher secretly ran white supremacist podcast, says it was satire. CNN News

Winter A (2019) Online Hate: from the Far-Right to the 'Alt-Right' and from the Margins to the Mainstream, pp 39–63. Springer International Publishing, Cham. ISBN: 978-3-030-12633-9. https://doi.org/10.1007/978-3-030-12633-9_2

Wolfowicz M, Litmanovitz Y, Weisburd D, Hasisi B (2020) A field-wide systematic review and meta-analysis of putative risk and protective factors for radicalization outcomes. J Quant Criminol 36:407–447

Zannettou S, Bradlyn B, De Cristofaro E, Kwak H, Sirivianos M, Stringini G, and Blackburn J (2018) What is gab: a bastion of free speech or an alt-right echo chamber. In: Proceedings of WWW '18, WWW '18, pp 1007–1014, Republic and Canton of Geneva, CHE. WWW '18. ISBN: 9781450356404. https://doi.org/10.1145/3184558.3191531

Zannettou S, Elsherief M, Belding E, Nilizadeh S, and Stringhini G (2020) Measuring and characterizing hate speech on news &websites. In: Proceedings of the 12th ACM conference on web science, WebSci '20, pp 125–134, New York, NY, USA. Association for Computing Machinery. ISBN: 9781450379892. https://doi.org/10.1145/3394231.3397902

Zighed DA, Lallich S, and Muhlenbach F (2002) Separability index in supervised learning. In: PKDD, volume 2, pp 475–487. Springer