Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.1 (2023) 1–30
HTTPS://DOI.ORG/10.5117/CCR2023.1.7.ZINN

Political discussions in online oppositional communities in the non-democratic context

Aidar Zinnatullin
University of Bologna

Abstract
Taking into account YouTube's specific role in the Russian media system and the increasing level of political polarization in the country, this study examines the role of incivility in discussions and whether discussions in an anti-government community represent a place for disagreement between pro-opposition and pro-government users. I argue that an online environment helps these sides meet each other rather than creating echo chambers of like-minded users. Moreover, in the quite restrictive Russian context for political deliberation, the incivility of messages plays a role in further involving commenters in discussions. Using the corpus of comments posted in the discussion section of opposition leader Alexei Navalny's YouTube channel, I exploited class affinity modeling to identify pro-government and pro-opposition stances. Incivility was studied based on Google's Perspective API toxicity classifier. I found that users avoid extreme forms of incivility when interacting with other commenters, but uncivil comments are more likely to start discussion threads. Furthermore, the level of incivility in comments gets higher over time after a video release. Pro-government sentiments, on the one hand, are associated with a subsequent response from Navalny's supporters to the out-group criticism and, on the other hand, contribute to the further formation of hubs with a pro-government narrative. This research contributes to the extant literature on affective polarization on social media, shedding light on political discussions within an oppositional community in a non-democracy.

Keywords: cross-cutting disagreement, affective polarization, autocracy, Russia, YouTube

Introduction

According to the theory of *affective polarization*, emotions become the basis for discerning *"us vs. them"* and increasing intolerance toward the other side (Iyengar et al., 2012; Iyengar and Westwood, 2014). On the one hand, this phenomenon finds its manifestation in digital media when the level

of incivility follows offline events of contentious politics (Sun et al., 2021; Theocharis et al., 2020). On the other hand, the political talk itself on social media can increase polarization between its participants (Marchal, 2021; Yarchi et al., 2020).

However, political tensions in oppositional communities in the non-democratic context, where the state controls traditional offline media as the main source of political information but still allows relative freedom on the Internet, attract less attention (Bodrunova et al., 2021). To contribute to this nascent literature, I studied discussions in the community of the most vocal Russian opposition politician, Alexei Navalny, on YouTube. Although Navalny "escalated existing tensions rather than creating them in the first place" (Dollbaum et al., 2021, p. 171), his activity played a significant role in the launch of not only massive propaganda but also repressive campaigns against dissent by the ruling elite. As a result, both sides of the conflict see each other as an existential threat, resulting in affective polarization (Nugent, 2020) with a strong *"us vs. them"* division between the ruling elite and the opposition.

My focus on Navalny's YouTube channel is also because of the twofold role of this platform in the Russian media system. On the one hand, YouTube facilitates the promotion of the opposition's agenda and enlarges the political capital of independent activists (Litvinenko, 2021). Besides the range of monetization schemes for content creators, the foreign origin of the platform implies that the Russian government cannot access the personal data of users, which is important for the relatively free expression of thoughts by users. Through emotions of *affective attunement* (Papacharissi, 2014) caused by the extremely high level of corruption in the ruling elite and Russia's social inequality, which were regularly revealed and justified in video investigations, Navalny skillfully went beyond the already formed community, united, and mobilized different groups that were dissatisfied with the ruling elite. Active users' engagement with content (for instance, liking and commenting) facilitates the promotion of video investigations on YouTube through the recommendation system and the Trending service of the platform. Without the affordances of YouTube, which has become the main entertainment platform in Russia, Navalny could hardly have expanded his audience and created a political movement able to struggle with administrative political machines.

This study focuses on the period from 2013 to 2021 when Russia's political regime can be described as an "informational autocracy": when confronted with alternative visions of the country's political situation, the government

had been expending significantly more effort convincing the public that they have the necessary competence to implement effective policy rather than relying exclusively on repressive instruments (Guriev & Treisman, 2022). I address the following research questions: (1) What are the potential and limits of incivility, as a characteristic of affective polarization (Suhay et al., 2017), to engage in political discussions? (2) How do users interact with the pro-government narrative presented in the community of the most vocal opposition politician?

I use a variety of methods to answer these research questions. I start by providing information about (a) the comments' quantity change over time and (b) the level of inequality in the distribution of comments by users. Next, logistic regression models and local polynomial fits are used to study the relationship between conversations and incivility. Then, I show how the level of toxicity changes over time. Finally, to identify pro-government and pro-opposition comments and detect cross-cutting disagreement, I trained a supervised machine learning model—the class affinity model (Perry and Benoit, 2017)—based on a dictionary with derogatory words applied to Navalny and his supporters, Putin and the government. I detected such words on the basis of an iterated computer-assisted keyword selection approach suggested by King, Lam & Roberts (2017).

The main empirical findings are the following. First, top-level comments that open discussions tend to be more uncivil than those without discussion threads. But toxicity has its limits. Users are not willing to dispute with those who spread extreme forms of incivility with a null potential to deliberate. Third, the level of incivility of comments gradually goes up with time passing after a video release during the first 14 hours and then stabilizes for top-level comments that have discussion threads and thread comments themselves. Second, pro-government comments (1) attract Navalny's supporters, who respond to the out-group criticism, and (2) contribute to the emergence of pockets of a pro-government narrative.

This research contributes to the extant literature on affective polarization on social media, shedding light on patterns and peculiarities of online political discussions within an oppositional community in a non-democracy. The findings advance our understanding of the behavior of out-group commenters who attack the domain of their political opponents, eventually forming pro-government hotspots. Furthermore, I was able to identify a limited potential for incivility in order to initiate discussions and report on the dynamics of those discussions' incivility.

The remainder of this paper proceeds as follows. In the "Theoretical

Framework" section, I start with a link between the contextual peculiarities of Russia's political regime and the theory of affective polarization. Following that, I argue why the oppositional content on the Russian segment of YouTube attracts not only those who oppose the government. In the "Data and Methods" section, I present the research design, which is followed by the section where the results of the empirical analysis are described. In the "Discussion" section, I point out the limitations of the study, its implications for changing regime characteristics in Russia and for other non-democratic polities, and perspectives for future research. In the "Conclusion" section, I recap the main takeaways from the study.

Theoretical Framework

Affective Polarization and Non-Democratic Context

Affective polarization, understood as an individual's identification with a political party and the further division of the world into a group of like-minded individuals and those who represent the out-group entity, is based not on policy preferences but on emotions cultivated during political campaigns (Iyengar et al., 2012; Iyengar and Westwood, 2014). One of the components of this phenomenon is the *sorting* of social reality, including a non-political context, into those who share a common identity and "the other side". This process is exacerbated when different conflicts overlap, thereby creating a large, all-encompassing cleavage. Social media only facilitates sorting (Törnberg, 2022). It does not mean that people live in closed, homogeneous communities. Instead, social media facilitates their interaction with the "other side". But this interaction strengthens their orientation toward an initial identity in social and political spaces rather than creating a solid ground for comprehension of the "other side."

Sorting as a source of affective polarization has also been reported as a characteristic of non-Western contexts (Harteveld, 2021; Huang and Kuo, 2022). This process is present in Russian social media because the fragmentation of media platforms is limited (Pashakhin, 2021), which means that users with opposing political beliefs can encounter interpretations of events that do not suit their initial perceptions of political reality. As a result, they tend to orient toward already-formed identification rather than come closer to their opponents in evaluating political processes.

A distinctive feature of polarization in Russia is that it is formed on the "power-opposition" dimension (Urman, 2019; D. K. Stukal et al., 2022). A critical attitude toward the existing political regime in Russia distinguishes

Alexei Navalny's agenda. He began his career in 2007 using "green mailing" tactics. It is a legal practice for an individual to buy a small but sufficient number of shares to send inquiries to top management about a company's transactions and spending. Through such activities, Navalny quickly began to uncover corruption schemes in large state-owned corporations (e.g., Gazprom, Transneft, and Aeroflot). In the early 2010s, the main platform for publishing his investigations was a blog on LiveJournal. Mastery of YouTube by Navalny's Anti-Corruption Foundation began with the 2013 Moscow mayoral election campaign. Initially, these videos presented diary entries about how the campaign was going and short clips in which famous actors and public figures expressed support for him.

At the end of 2015, much more active work began on YouTube, where anti-corruption investigations began to be posted. The first resonant video was an investigation of the business of the sons of the Russian Prosecutor General, Yuri Chaika. Subsequently, more anti-corruption investigations about high-ranking Russian officials were conducted. During the 2018 presidential campaign, YouTube became Navalny's main communication channel for the audience. Although his content dealt with serious topics, it was presented to the viewer in an accessible manner, often with landmark drone footage. Videos maximally corresponded to the aesthetics of blogging on YouTube (Glazunova, 2022) by exploiting and producing memes relevant to the Russian online audience, which, coupled with Navalny's humour and self-irony, made it possible to classify such content as infotainment.

By 2020, Navalny had become a leader of the oppositional movement in the country, which promoted strategies of political action for a wide variety of events (from anti-corruption rallies to the effective "smart voting" (Turchenko and Golosov, 2020) campaign in regional elections), but with a focus on anti-corruption investigations (Kazun and Semykina, 2019). Navalny's leadership was based on creating a political infrastructure that could organize collective action. No other politician strained the regime much after the 2011-2012 protests.

Russian authorities had ignored Navalny's claims, while traditional media affiliated with the state had covered his activity rarely but strictly negatively (Kazun, 2019). Generally, protest in the official Russian discourse is framed as disorder and war. This explains the great attention paid by federal TV channels controlled by the Kremlin to events in pre-war Ukraine (Alyukov, 2021). The official media had been framing the protest as a clash of citizens with each other, during which aggressive radicals would surely become the winners, ready to kindle the fire of war against their citizens.

One consequence of Navalny's activity was an increasing level of polarization when the ruling elite launched a repressive and propaganda campaign trying to suppress the dissent (for instance, the law on foreign agents, tightening the rules for holding mass events, recognizing Navalny's Anti-Corruption Foundation as an extremist organization, pressure on social media platforms). This situation leads to affective polarization (Nugent, 2020) when the politicization of activists and supporters occurs with a strong *"us vs. them"* division based on emotions derived from an existential threat that both sides of the conflict see in each other (Dollbaum et al., 2021).

I focus on discussions in a community formed around the activities of Alexei Navalny, who presented himself as Putin's main competitor. The extant literature considers incivility and hate speech as features of affective polarization (Harel et al., 2020; D. K. Stukal et al., 2022). The perception of polarization manifests in incivility which is associated with lower expectations about online public deliberation (Hwang et al., 2014). But uncivil interactions with peers can have positive implications for community formation (Kosmidis and Theocharis, 2020) and the strengthening of solidarity among those who share common political beliefs when it comes to non-democratic political regimes. As Bodrunova et al. (2021) argue, uncivil comments remove barriers to opinion expression by users who, in a less liberating environment, might remain silent, keeping in mind that authorities do not welcome political expressions. Therefore, I check **Hypothesis 1A** in the context of discussions occurring in Navalny's YouTube community: *Uncivil comments are more likely to start discussion threads than civil ones.*

At the same time, hate speech and incivility mean disrespect to others (Kim et al., 2021), and it is reasonable to expect heterogeneous reactions. For some viewers of Navalny's videos, incivility can be liberating for expressing their opinions as shown in Bodrunova et al. (2021). But uncivil comments may also discourage other participants or potential commenters from engaging in a conversation. As such, in the long run, it has the potential to create a spiral of toxicity (Kim et al., 2021) through the synchronization of emotions between interlocutors and the effect of social interaction (Kwon and Gruzd, 2017). Thus, I check **Hypothesis 1B**: *The longer the time after a video release, the higher the degree of the incivility of comments.*

Navalny's YouTube channel as a Place for (Dis)similarity

YouTube in Russia is highly politicized, and communities form around political bloggers (Litvinenko, 2021). Alexei Navalny, being one of the most prominent examples of this tendency, could gain benefits from direct-casting

(Bastos et al., 2013), becoming the most vocal opposition politician in Russia (Titov, 2017). It is worth characterizing the content produced by his team.

YouTube as a platform creates its mediality of *affective attunement* that "permits people to feel their way into politics" (Papacharissi, 2014, p.118). Affect and emotion have the potential to reach out beyond the already-established community and form *affective publics* in this way (Papacharissi, 2014). Navalny's anti-corruption investigations were most often shocking Internet users, with the amount of money allegedly stolen from the state budget. Moreover, he tried to unite and mobilize different communities that were dissatisfied with the government because of its policies (truck drivers, employees of state organizations, medical workers, and many others whom he appealed to). Navalny skillfully reached different audiences, who sometimes stood for different ideals. Without consumers of these materials and YouTube's affordances (such as the Trending tab), it would be impossible to go beyond the relatively narrow group of political geeks (Glazunova, 2020). Hence, I consider the comments section as affordance (Evans et al., 2016) for Navalny's team to promote the video on the YouTube platform, for his supporters, to express their support, and for his opponents, to show that many people disagree with Navalny's position.

Why can discussions occur between those who believe in different political ideas in Navalny's YouTube community? The online environment frees people to express their points of view (Wojcieszak and Mutz, 2009), even when opinions are ideologically distant (Shugars and Beauchamp, 2019) and conflicting (Stromer-Galley, 2006). Politically controversial topics are accompanied by a moderate level of heterogeneity, i.e. it is more likely to observe dissimilar sentiments in comments on sensitive issues (Röchert et al., 2020). These observations go along with the idea of "corrective action" when individuals who perceive that media disproportionately affect public opinion are more likely to engage in political communication to make themselves more visible (Rojas, 2010). However, social media should not be considered solely as a space where deliberation between different groups occurs. Rather, it can be understood as a place for identity formation and strengthening (Törnberg and Uitermark, 2021). Social media intensifies existing contradictions owing to the sorting process in which individuals orient themselves to identities rather than opinions. Eventually, this process exacerbates the perception of profound cleavages in society and differences between in-group and out-group members (Törnberg, 2022).

The other line of literature reports that users may respond to social context cues e.g. likes or sentiment of comments (Li et al., 2015; Voggeser et

al., 2018; Cho and Kwon, 2015). This situation can also be considered through the lens of the well-known "spiral of silence" theory (Noelle-Neumann, 1974), in which an individual, seeing that dominant social attitudes propagated by the media or the social environment contradict his or her own opinion, tries to avoid expressing a point of view for fear of being isolated. However, the video format has the potential for much higher involvement than any other type of content. This type of storytelling may involve those who stand in apolitical positions and become spaces where informal political talk occurs (Coleman and Freelon, 2016). In some sense, it may resemble a leisure group (Wojcieszak and Mutz, 2009, p.50) and involve different people, not just opposition-minded users. Although apolitical users do not use social media as an entrance to political news (Möller et al., 2019), the YouTube channel of Navalny resembles not an average news broadcast but political infotainment. This simultaneously engages in the viewing process and, to some extent, creates a more relaxed atmosphere for opinion expression by different viewer categories.

When YouTube (via the Trending tab) contributes to the promotion of video content on a par with the media resonance that accompanies Navalny's investigations (Kazun, 2019), there are more chances that users can express thoughts that may doubt the arguments of a particular investigation, simply because such videos can go beyond the community of Navalny's supporters. YouTube provides a decent level of pseudonymity, which facilitates opinion expression contradictory to the dominant perspective (Halpern and Gibbs, 2013; Wu and Atkin, 2018), without affecting the quality of political discussion (Berg, 2016).

Toepfl (2020) defined Navalny's community as a leadership-critical public within a non-democratic context formed around a headliner who is not afraid of the country's ruler and actively criticizes him. I expect that cross-cutting disagreement can be observed when there are attacks on Navalny and his community or supporters. *Cross-cutting disagreement* is defined as a clash between those who criticize Putin or the authorities from one side and those who criticize Navalny or the opposition in response (and vice versa); this interaction must occur in the same pair of top-level and threaded comments (not exclusively within the thread).

Hypothesis 2: *Top-level comments attacking Navalny or the opposition are more likely to generate threaded comments from the opposite side than other types of top-level comments (e.g., pro-opposition and neutral stances).*

Data and Methods

Data. I collected comments for the videos uploaded to the YouTube platform on Navalny's channel from 2013 until July 2021 through the use of YouTube Data Tools (Rieder, 2015). The pooled dataset contains 8,980,313 comments from 407 videos. But I excluded several videos in different steps of the data collection because of issues related to the incorrect functioning of YouTube's API[1]. It decreased the overall number of comments to 7,985,548 without having a severe effect on the representation of discussions[2].

Discussion structure. On YouTube, users can comment on the content by posting top-level messages without responding to anyone. Such messages may eventually generate a discussion thread when other users comment underneath (n=579,556). I call such comment **top-level comments with threads**. Consequently, a **thread comment** appears, that is, one posted under a top-level comment as a response to it (n=2,018,973). Then, I also discerned **top-level comments without a thread** (n=5,387,019). This distinction allows me to discern different patterns of opinion expression and observe the interaction between users.

Method for testing Hypothesis 1A. As a measure of **incivility**, I use toxicity scores provided by Perspective API. Perspective is a convolutional neural network toxicity classifier trained on millions of comments in several languages, including Russian. Table A1 in the online appendices contain examples of the comments, their translation in English, and toxicity scores. I prefer not to focus on the debate about the distinction between incivility, intolerance, hate speech, toxicity, and other related concepts (Rossini, 2020) since my interest lies in the aggregate scope of conversational characteristics. Therefore, I follow Perspective API's definition of toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion"[3]. I use toxicity and incivility as interchangeable concepts.

I rely on Perspective API toxicity classifier to detect incivility for several reasons. First, it has been successfully used to detect toxicity in short comments written in Russian (Bogoradnikova et al., 2021) as a baseline for comparison and performed better than other methods compared by the authors[4]. Second, developers of the classifier revised the model, making it

[1]The service YouTube Data Tools did not allow me to retrieve comments from the video about Putin's palace (https://www.youtube.com/watch?v=ipAnwilMncI). Sending queries to the YouTube API directly returned comments that were not discerned into thread and top-level comments. As a result, comments about Putin's palace were excluded from the final dataset.

[2]I pseudonymized the IDs of the commenters using the encryptr and digest packages on R.

[3]https://developers.perspectiveapi.com/s/about-the-api

[4]The correlation coefficient between comment length and toxicity scores is 0.228, which is

more robust against adversary attacks (Lees et al., 2022). Third, social scientists actively organize human validations (Kim et al., 2021) of toxicity scores, reporting generally the better performance of Perspective API in comparison to human labelers in terms of accuracy and efficiency (Rajadesingan A., 2020; Vargo and Hopp, 2020).

To test Hypothesis 1A, I check the association between the toxicity of a top-level comment and whether it creates a discussion thread. I use logistic regression with a dependent variable operationalized as to whether a top-level comment has a thread (i.e., opens discussion) or not. The structure of the dataset derived from the YouTube API does not allow me to look at sub-threads when users respond to comments left in a thread. I marked all such comments as *thread* comments to a single top-level comment. As the main independent variable, I use a dichotomized version of the toxicity score where all comments with a score equal to or greater than 0.5 are defined as toxic, and messages with a score below 0.5 are non-toxic. I also exploit an alternative approach using local polynomial fits to predict the number of replies a top-level comment receives, taking Perspective's raw toxicity scores into account. OLS regression of the length of discussion threads on top-level comment toxicity is also presented (Table A4).

Method for testing Hypothesis 1B. To check Hypothesis 1B, I look at the association between a comment's toxicity and when it was posted (in 2- and 12-hour interval bins after the publication of the video). Here, the focus is on an average value of toxicity for three types of comments within a given time interval, with error bars representing 1.96 standard deviations of the mean. Moreover, I conducted a regression analysis of toxicity on the timing of comment posting (Tables A5 and A6 in the online appendices).

I also run OLS regressions of the time of posting (with 2- and 12-hour intervals) on the type of commenters according to (1) their frequency of interaction with Navalny's content (one-off vs. prolific commenters), (2) the average level of toxicity of their messages (after aggregating comments by their authors, if the average level of toxicity is 0.5 or more, a commenter is defined as "toxic", and otherwise "non-toxic") and (3) type of comment[5]. In addition, I added the interaction between these variables (Tables A7 and A8 in the online appendices).

negligible.

[5]I thank the anonymous reviewer for the suggestion to take into consideration different types of commenters and their contributions in different timeframes after a video release. I admit it is just a preliminary attempt to address this relevant question. It requires a much better, more elaborated research design. Therefore, I hope to continue this line of inquiry further in my research activities.

Discerning anti-/pro-government stances. For testing Hypothesis 2, I detected pro-government and pro-opposition sentiments in comments using class affinity modeling (Perry and Benoit, 2017). This method is appropriate in situations where most of the text messages are unlabeled but a small number of comments with extreme values on a hypothesized ideological spectrum are presented.

I started with the compilation of a dictionary containing derogatory references to the government and opposition. Such a focus on insults targeting the other side is related to the concept of affective polarization, driven by emotions and manifested in hate speech. The algorithm to compile such a dictionary was based on an iterated computer-assisted keyword selection approach suggested by King, Lam & Roberts (2017). First, I began with several derogatory words targeting Putin and Navalny that were detected after a close reading of YouTube comments and pages on Lurkmore, serving as an encyclopedia of political discourse on the Russian Internet. Second, I widen the scope of keywords in a snowball sampling manner, checking the sentiment of comments towards both the government and the opposition, either randomly reading some of them (when their number is huge) or the whole subset of comments with a particular word. In a snowball sampling, I rely on such metrics as the frequency of words and term frequency-inverse document frequency weighting. The final version of the dictionary contains approximately 530 string patterns that were used to search the corpus of comments for pro-government (240 string patterns) and opposition (290 string patterns) sentiment. Table A9 in the online appendices contains some of the words derived from this procedure.

Second, as a collection of comments that occupy extreme positions on the "pro-government-opposition" sentiment scale, I chose comments that (1) contain derogatory or mocking references to the government or opposition and (2) have scores greater than 0.5, according to Perspective API toxicity classifier. This value was chosen as a threshold from the civil to uncivil categories because of its probabilistic logic. When pro-government and opposition words from the dictionary were present in the same comment, they were excluded from the training set. The combination of derogatory or mocking words targeting political actors and uncivil sentiment in a comment serves as a training set for a class affinity model. I have checked the words included in my dictionary in Perspective API classifier; in most cases, these words are not recognized as toxic. In addition, the correlation between toxicity scores and affinity scores is weakly positive (0.12).

Class affinity scores range from 0 to 1. I define pro-government and

pro-opposition comments as having affinity scores equal to or greater than 0.8 (pro-opposition) and equal to or less than 0.2 (pro-government). Such thresholds are chosen because I am interested in extreme forms of attack on the other side. The training set contained 194,674 comments (2 percent of the comments corpus). Recall that the class affinity model allows one to work with a small number of labeled documents (Perry and Benoit, 2017). I applied the class affinity model to the full corpus of around 8 million comments. The distribution of comments according to their sentiment is presented in Table A10 in the online appendices.

Such an algorithm does not fit the canonical approaches in the supervised machine learning literature when coders annotate a sample of text that is then piped into classification models. My approach of combining Perspective API toxicity scores, in some sense, resembles an adaptation of supervision with the found data (Grimmer et al., 2022), which requires considerable effort to validate the results. I have already mentioned how the results of computer-assisted keywords used to identify pro-government and opposition cues were validated by a close reading of these comments. Regarding the relevance of Perspective API toxicity classifier for the Russian language, I have already speculated in the corresponding section on the analysis of the toxicity of the discussions.

The results of class affinity modeling were also validated against two independent coders who manually classified 1000 comments into pro-government, pro-opposition, and neutral categories, taking into account such features as attacks on the opposite side, the incivility of the message, and emotional expressions of commenters (the use of emojis, for example). Inter-coder reliability is 0.89. The model achieved a high prediction accuracy of 0.97 with the first coder and 0.95 with the second coder. But the precision metric for pro-government positions remains relatively low (0.71 for coder 1 and 0.73 for coder 2). I explain these results with the fact that the share of pro-government and pro-opposition comments is not very high (Table A10 in the online appendices). So, I take these results cautiously because of the specificity of the comments I need to discern. Other statistics are presented in Tables A11 and A12.

Method for testing Hypothesis 2. For testing Hypothesis 2, which checks an association between attacks on Navalny or the opposition and a subsequent response from their supporters, I create a scale of the discussion type and use multinomial logistic regression. The dependent variables have five values: (1) *no conversation*, which means that a top-level comment does not contain a reply, (2) *discussion*, a situation when a top-level comment

has a thread of reply comments but without attacks on the government or opposition, (3) *attack on the government*, a top-level comment has a reply comment with an attack on the government, (4) *attack on the opposition*, a top-level comment has a reply comment with an attack on the opposition, and (5) *attacks from both sides in a threaded comment*, i.e. a top-level comment provokes a reaction from both political camps. Table A13 in the online appendices contains the distribution of the discussion types.

The independent variable of interest is the comment type. It can have three values: pro-government, neutral, and pro-opposition. I also control the length of a comment, the number of likes it has, the time when a comment was posted (split into 2-hour intervals after the publication of a video), and whether a comment is toxic. Comment length and the number of likes it received were taken from a logged version ($log(1+x)$) rather than raw values to address the skewness in the distribution of both variables.

Additional tests were also performed with ordinal logistic regression split into two separate models, where the dependent variable has three levels: no discussion, discussion, and attack on the government or opposition (depending on the model) (Tables A16 and A17). This modification was not included in the main text of the article because it violated the assumption of proportional odds. I also check the results of multinomial logistic regression models, avoiding the assumption about the ordered nature of the dependent variable and dealing with two separate models (Tables A14 and A15).

Results

I begin with a general description of the comment dataset. Figure 1 shows how the number of comments changed monthly during the study period. It can be seen how commenting occurs unevenly over time. Overall, one percent of the most discussed videos (n=4) generated 23 percent of all the comments (2,078,519). Ten percent of the videos (n=41) contributed 50 percent of the comments (4,531,856). This distribution corresponds to a Gini coefficient of 0.67, which indicates a high concentration of comments from a limited number of videos. When this coefficient is close to one, it means that the comments have been distributed among a few videos (all comments are posted under a single video when the Gini coefficient is one). If the Gini coefficient is zero, then all comments are distributed equally among all the videos released by Navalny's team.

Comments are not only sporadic outbursts of emotion manifested in writing a single message, after which a user leaves the page (Figure 2). Instead, video begets conversations between the users. The Gini coefficient for

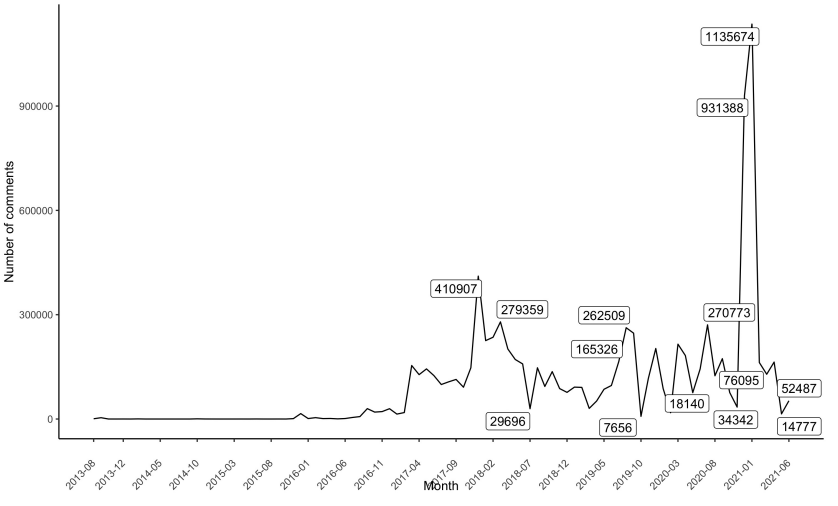

Figure 1: Number of comments on Navalny's YouTube channel by month of publishing

inequality in the distribution of comments by YouTube users confirms this point. The interpretation of these values is similar to what was previously said about the Gini coefficient for videos and comments. But here, we deal with the distribution of comments among users, not videos. According to Figure 2, the Gini index for threaded comments (red line) is higher than that for top-level comments (blue line). It means that the videos on Navalny's YouTube channel generated micro-discussions with an active exchange of opinions under the top-level comments. Some users comment on the statements of others, and their participation in these conversations varies. At the same time, top-level comments represent expressions of what commenters saw in the video clips rather than an exchange of opinions with peers, and the level of inequality in the distribution of this type of comment is lower.

Testing Hypothesis 1A*: Uncivil comments are more likely to start discussion threads than civil ones.* Table 1 contains the results of logistic regression with a dependent variable operationalized as whether a top-level comment has a thread (i.e., opens discussion) or not. All comments with a toxicity score equal to and over 0.5 are defined as toxic, and messages with a score below 0.5 are non-toxic.

Uncivil top-level comments have a higher chance of starting a conversation thread. After converting a coefficient from Column 1 in Table 1 to the probability of having a thread, we get 0.52. After adding into a model the interaction of toxicity with time measured through 2-hour intervals after a

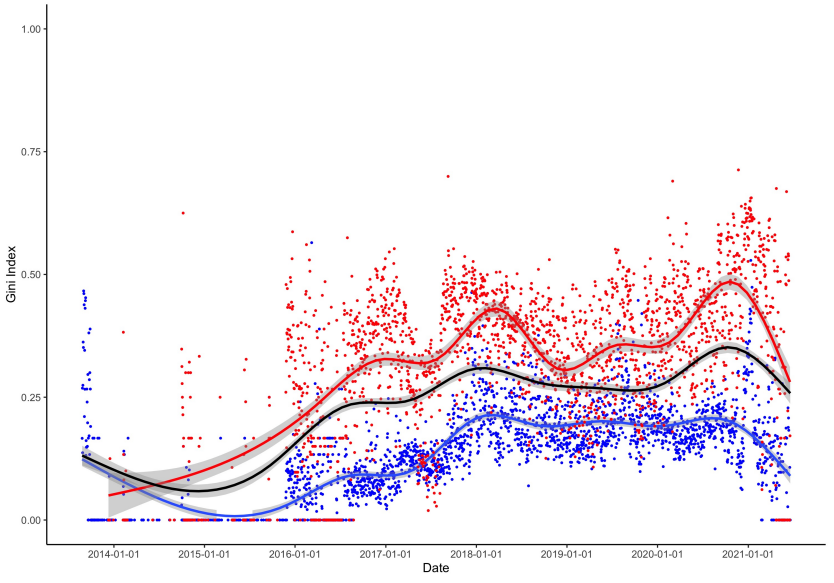

Figure 2: Inequality in the distribution of comments by YouTube users, a day-level snapshot: the red line and dots are for thread comments, the blue line and dots are for top-level comments, and the black line is for the whole corpus of comments

Table 1: Logistic regression results of a top-level comment having a discussion thread on toxicity, count of likes, comment length, 2-hour intervals

	Dependent variable:					
	Type of top-level comment: with or without thread					
	Model 1			Model 2		
	β	SE	p-value	β	SE	p-value
Toxicity (binary)	0.091	0.021	0.00001	0.178	0.025	0.000
Count of Likes (log)	0.889	0.015	0.000	0.889	0.015	0.000
Comment length (log)	0.650	0.013	0.000	0.650	0.013	0.000
2nd 2 hours	0.053	0.030	0.074	0.054	0.032	0.090
3rd 2 hours	0.142	0.029	0.00000	0.137	0.031	0.00002
4th 2 hours	0.222	0.045	0.00000	0.221	0.048	0.00001
5th 2 hours	0.331	0.043	0.000	0.325	0.045	0.000
6th 2 hours	0.453	0.068	0.000	0.452	0.072	0.000
7th 2 hours	0.532	0.070	0.000	0.546	0.074	0.000
More than 14 hours	0.605	0.081	0.000	0.632	0.080	0.000
Intercept	−5.135	0.044	0.000	−5.147	0.044	0.000
Interaction of toxicity with time	No			Yes		
Observations	5,966,575			5,966,575		
Log Likelihood	−1,308,195			−1,308,005		
Akaike Inf. Crit.	2,616,413			2,616,046		
Bayesian Inf. Crit.	2,616,562			2,616,291		

Note: Video clustered standard errors are presented

video release (Column 2), the coefficient for the variable of interest slightly increases (0.54 converted to probability). In Figure 1 in the online appendices[6], I present the marginal effects of the level of toxicity on the likelihood that a top-level comment opens a discussion. The likelihood that a top-level comment will open a discussion goes up from nearly 0.08 to approximately 0.2 in a toxicity range from 0 to 1.

There are more empirical tests of the hypothesis where I use centered toxicity scores, their squared term, video fixed effects, and add interactions of toxicity with other variables (Online Appendix B). Generally, the results do not contradict the ones reported in Table 1. However, controlling for the squared term of toxicity shows that toxicity has its limits: users do not engage with messages containing extreme incivility (Table A3 in the online appendices).

Further, I present the results of models that use the number of replies that a top-level comment receives as an outcome. Figure 3 shows how toxicity scores are associated with the number of replies that top-level comments get. In this analysis, I apply local polynomial fits. On the y-axis, fitting values of reply counts are presented, while on the x-axis are percentiles of toxicity scores. Figure 3 shows bimodality. Nearly the 60th percentile corresponds to a 0.1 toxicity score. The 80th percentile, which corresponds to a 0.31 toxicity score, gets the highest number of comments. It means that replies increase until toxicity reaches the 80th percentile of the distribution of comments and then declines. In other words, messages that get the most replies from other commenters do not demonstrate extreme forms of incivility. Moreover, according to Table A4, the average increase in the number of replies that a top-level comment attracts varies from 0.5 to 1.2 percent with a change from the category of civil to uncivil messages.

In addition, I run local polynomial regression of "like" counts on toxicity distribution (Figure 2 in the online appendices). The peak of the "like" counts is in the 50th percentile of the toxicity distribution, which corresponds to a toxicity score of 0.07. In other words, the comments that are predicted to be classified as toxic in only 7% of cases are the most popular in terms of endorsements from other users.

In general, Hypothesis 1A finds empirical evidence, but the association between toxicity and discussions has a more nuanced nature. According to the findings, uncivil top-level comments are more likely to have discussion threads. However, users are not willing to dispute with those who spread extreme forms of incivility towards others, probably seeing the limited po-

[6]Here, I use the original, not binary, version of a toxicity score variable.

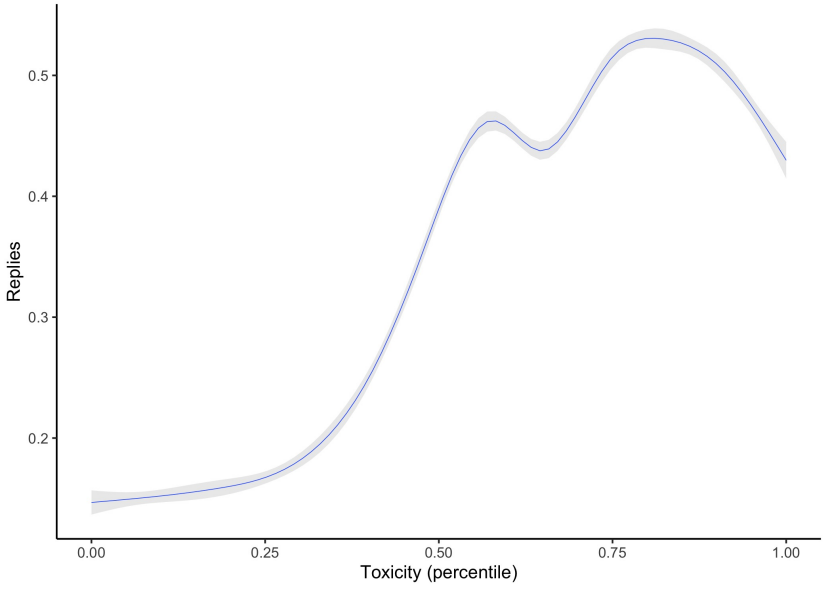

Figure 3: Comment replies and toxicity scores, n = 5,966,575 top-level comments

tential to deliberate.

Testing Hypothesis 1B*: The longer the time after a video release, the higher the degree of the incivility of comments.* Figures 4 and 5 show how the level of toxicity in three types of comments changes over time after the release of a video (by 2- and 12-hour intervals, respectively).

During the first hours, the level of toxicity of each type of comment is lower (Figure 4). Then there is a gradual increase in toxicity. After a few hours, the situation for top-level comments with threads and thread comments stabilizes, as Figure 5 with 12-hour intervals shows. However, the toxicity of top-level comments without threads continues to rise. Top-level comments with threads have the highest toxicity score in all periods, while comments in threads demonstrate lower toxicity. Top-level messages which do not open discussion are less toxic.

Regression results of toxicity on the timing of comment posting are displayed in Tables A5 and A6 in the online appendices. Users write more toxic top-level comments (but not reply comments) over time. Moreover, additional regression models of posting time on the types of commenters and their comments indicate that toxic and one-off commenters are more likely to write comments later, when a video loses its virality (Tables A7 and

A8).

Thus, the toxicity of comments increases with time following the release of a video, and, at a particular level, incivility reaches its peak. But the contribution to the growth of the toxicity of subsequent comments is made to a greater extent by (1) those commenters who, on average, write more toxic posts, (2) rarely do this (more inclined to be one-off), and (3) mostly when the existing discussions in the threads decelerate.

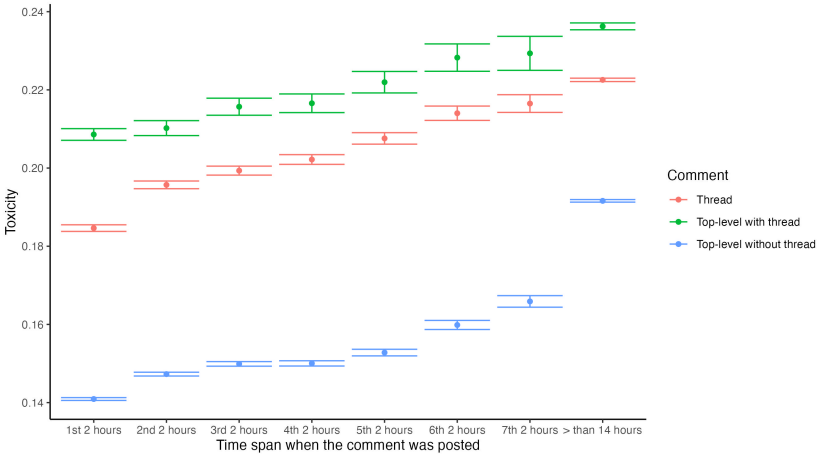

Figure 4: Toxicity score of comments by time of comment posting (2-hour bins)

Testing Hypothesis 2*: Top-level comments attacking Navalny or the opposition are more likely to generate comments from the opposite side than other types of top-level comments (e.g., pro-opposition and neutral stances).* I present the results of multinomial regression of the discussion type in the thread on the sentiment of a top-level comment (pro-government, neutral, or opposition), whether it is toxic, the comment length, the number of likes it receives, and the time when a top-level comment was posted. I use the lack of any discussion underneath a top-level comment as the baseline level of the outcome. Tables 2 and 3 contain columns corresponding to the other four levels of the dependent variable. All of them are compared to the "no conversation" baseline type of discussion. The standard errors are clustered on a video level. The main variable of interest—a top-level comment's sentiment—has three levels, where a neutral tone serves as a baseline and is hidden in the intercept.

We observe the association predicted in the hypothesis. The pro-government stance of the top-level comment is linked to the subsequent attack on the

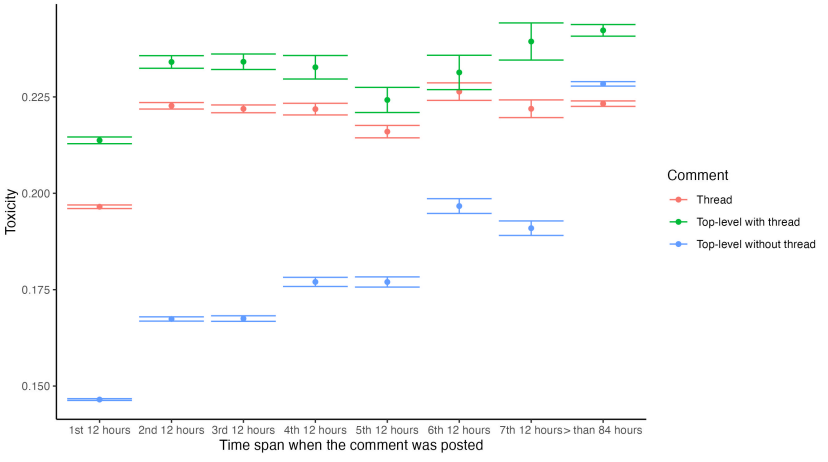

Figure 5: Toxicity score of comments by time of comment posting (12-hour bins)

government in the thread to this comment (Table 2, column 2). After converting log odds to the probability, we obtain 0.57. When a top-level comment is pro-government, this number indicates that the opposition is more likely to attack than compared to a neutral top-level comment as a baseline.

At the same time, the probability of attacks from both sides in a comment thread *vs.* without any discussion underneath is 0.76 if moving from a neutral comment category to a pro-government sentiment (Table 2, Column 1). Interestingly, attacks on the opposition are clustered in the sense that pro-government sentiment in a top-level comment is associated with the same sentiment in the thread underneath (see Table 3, Column 1). The probability of having a pro-government sentiment, in this case, is 0.77.

In the case of pro-opposition sentiment, pro-opposition stances articulated in a top-level comment are associated with the attack from the other side as well. But this association is weaker in terms of substantive significance. The probability of the criticism targeting the opposition in the thread *vs.* no conversation as a baseline will increase by 0.29 when a top-level comment moves from the neutral to the pro-opposition category (Table 3, Column 1). My interpretation of this observation is as follows: the comments section is replete with anti-government discourse, and the pro-government narrative that is out of the ordinary attracts the attention of both those who share these beliefs and those who support Navalny and the opposition. Online Appendix E shows that the predominant topics in the top-level comments are those that express sympathy for Navalny or contain a sentiment

attacking the ruling class. The association of pro-government discourse with discussions, in contrast to the pro-opposition narrative, confirms my expectation regarding the polarization in the community of Alexei Navalny on YouTube.

Online Appendix D contains the results of two multinomial regression models when a dependent variable has only three values: "attacks from both sides" were discarded, while attacks on the government and the opposition were checked in separate models (Tables A14 and A15). In addition to the Hausman-McFadden test for a multinomial logit model, this analysis also confirms that the independence of irrelevant alternatives assumption is true. Though the assumption of parallel lines does not hold, the results of ordered logistic regressions are also presented. The difference between Tables A16 and A17 lies in the event considered to be of the highest order. In Table A16, this is an attack on the government in the thread under the top-level comment. In Table A17, the dependent variable has the highest value when there is an attack on the opposition in the thread under the top-level comment. These results are similar to those presented in Table 2[7].

Table 2: Results of multinomial logistic regression of discussion type

	Reference 'No discussion underneath of a top-level comment'					
	Column 1 Attacks from both sides			Column 2 Attacks on government		
	β	SE	p-value	β	SE	p-value
Pro-Government (baseline: neutral)	1.152	0.056	0.000	0.296	0.056	0.00000
Pro-Opposition (baseline: neutral)	−0.685	0.053	0.000	−0.056	0.029	0.297
Toxicity (binary)	0.404	0.048	0.000	0.204	0.028	0.00002
Comment length (log)	1.273	0.021	0.000	0.788	0.019	0.000
Count of Likes (log)	1.634	0.022	0.000	1.124	0.017	0.000
Second 2 hours	−0.081	0.054	0.133	0.074	0.036	0.037
Third 2 hours	−0.016	0.060	0.793	0.137	0.042	0.001
Fourth 2 hours	0.085	0.078	0.278	0.136	0.050	0.006
Fifth 2 hours	0.250	0.077	0.001	0.226	0.045	0.004
Sixth 2 hours	0.282	0.101	0.005	0.411	0.072	0.00005
Seventh 2 hours	0.490	0.090	0.00000	0.418	0.064	0.00001
After more than 14 hours	0.569	0.091	0.000	0.458	0.078	0.00000
Intercept (Neutral)	−12.636	0.097	0.000	−8.112	0.070	0.000
Akaike Inf. Crit.		3,527,597			3,527,597	
Bayesian Inf. Crit.		3,528,304			3,528,304	
Observations		5,965,458			5,965,458	

Note: Video clustered standard errors are presented

[7]Only some of the model outputs are presented in Online Appendix C , whereas others can be requested from the author. In general, the results in all additional tests maintain the expected direction of association between the variables of interest and statistical significance.

Table 3: Results of multinomial logistic regression of discussion type

	Reference 'No discussion underneath of a top-level comment'					
	Column 1 Attacks on opposition			Column 2 Discussion		
	β	SE	p-value	β	SE	p-value
Pro-Government (baseline: neutral)	1.191	0.039	0.000	0.432	0.038	0.000
Pro-Opposition (baseline: neutral)	−0.893	0.042	0.000	−0.352	0.027	0.000
Toxicity (binary)	0.209	0.030	0.00002	0.061	0.018	0.202
Comment length (log)	0.939	0.016	0.000	0.603	0.013	0.000
Count of Likes (log)	1.045	0.019	0.000	0.837	0.013	0.000
Second 2 hours	0.001	0.042	0.979	0.065	0.027	0.017
Third 2 hours	0.102	0.042	0.015	0.153	0.028	0.000
Fourth 2 hours	0.236	0.061	0.000	0.233	0.046	0.000
Fifth 2 hours	0.426	0.069	0.00000	0.336	0.042	0.00002
Sixth 2 hours	0.516	0.093	0.00000	0.453	0.068	0.00001
Seventh 2 hours	0.563	0.081	0.000	0.536	0.075	0.000
After more than 14 hours	0.699	0.085	0.000	0.599	0.082	0.000
Intercept (Neutral)	−8.589	0.061	0.000	−5.142	0.043	0.000
Akaike Inf. Crit.		3,527,597			3,527,597	
Bayesian Inf. Crit.		3,528,304			3,528,304	
Observations		5,965,458			5,965,458	

Note: Video clustered standard errors are presented

Discussion

For obvious reasons, such as the sensitivity with which citizens present their true political preferences to interviewers, the phenomenon of affective polarization in non-democracies has received little attention using traditional methodological tools, most notably surveys. But we should not underestimate attempts to address this issue from the perspective of observational digital data. This study is one such effort, and it answers the question of to what extent polarized political discussions are in the online community of the most vocal Russian opposition politician, Alexei Navalny.

From my analysis of discussions in the comments section of Alexei Navalny on YouTube, three main contributions emerge to the literature on political communication within a non-democratic context. First, the role of incivility in affecting discussions has a two-fold nature. On the one hand, comments that attract reactions from other users in the form of text replies are more uncivil. This relationship between replies and toxicity is not linear. Incivility has limits in driving further discussion because users are not prone to react to extremely toxic messages. If previous studies provide evidence that initial posts increase the likelihood of getting uncivil replies in such a way, creating a spiral of toxicity (Kim et al., 2021; Rega and Marchetti, 2021; Unkel and Kümpel, 2022), top-level comments with threads in the dataset analyzed are more uncivil than comments that follow. This finding

highlights the role of incivility in the specific context of Russia, where the authorities restrict people's expressions about politics in general and regulate the way they have to communicate on the Web (for instance, swear words are prohibited by law) (Bodrunova et al., 2021). To be discussed, a comment must have some potential to signal that the environment is free to express opinions in a frank manner, without slipping into the direct abuse of the participants.

Second, the level of toxicity of comments goes up over time. However, commenters contributing to the increase in toxicity over time differ from those who engage in conversations when a video gains more attention from the audience (within several hours after its publication). The former are sporadic and use more toxic language, writing top-level comments (i.e., they do not engage in the presented discussion threads).

Third, pro-government messages posted as top-level comments in the community of Russia's opposition leader serve as a crossroads between two opposing camps. Navalny's supporters tend to respond to attacks from their opponents. At the same time, the pro-government sentiment expressed in the form of thread comments focuses on pockets of disagreement initiated by other pro-government commenters. I interpret this result as indicating that communities on social media do not fully create ideological silos where their members cannot encounter the opposite point of view. Rather, they may see ideas that are contrary to their views. In such a situation, this only contributes to the rallying of the group and the strengthening of a common identity. Therefore, it becomes important to react to attacks from the "other side." The fact that pro-government commenters are less active in commenting on pro-opposition posts can be explained by their intention to show their presence and disagreement with the dominant narrative. Replies to top-level comments get little attention due to YouTube's design of the comment section, while top-level comments are much better at attracting attention from other users.

Although my work adds to the general theory of political communication in informational and electoral authoritarianism (as defined by pre-war Russia), its implications for the new realities of a political system with tight control over dissent are limited. Undoubtedly, the Russian-Ukrainian conflict and the earlier imprisonment of Alexei Navalny changed the political regime in Russia towards more control over citizens and suppressing collective action and dissent in general. But, as the experience of other autocracies shows (Nugent, 2020), mass repression only contributes to polarization. With the confrontation of "*us vs. them*", suppression of dissent

generates more emotional anger. In addition, not all foreign-origin social media platforms, including YouTube, were banned at the time this text was written. Independent journalists and political activists actively migrate to YouTube to continue their work, even if they are forced to flee the country due to repressions. And the reason YouTube is still available in Russia is that pro-government content also finds an appreciative home on this platform. This is one perspective for future research: to compare pro-government and opposition YouTube channels both in terms of the content they disseminate, their popularity (for instance, their appearance in the Trending tab, which is understudied in the domain of computational communication research), and the comments they receive, taking into consideration different stages in the regime's evolution.

Personalistic autocracies show similarities in their communication strategies to tackle challenges to keeping the status quo. Pro-government discourse becomes simpler, with greater potential to polarize society by blaming someone for a country's economic woes (Rozenas and Stukal, 2019; AYTAC, 2021) or other issues (Alrababa'h and Blaydes, 2020; Laebens and Öztürk, 2020). Thus, political polarization does not have to be a corollary of social cleavages. Instead, we can consider it a by-product of political entrepreneurs' actions to pursue their goals (McCoy and Somer, 2018). Even under tighter regime constraints, the Internet and social media, in particular, provide a platform for citizens to disseminate information and raise awareness about the incompatibility of their interests and values with those who benefit from the current state of affairs. This thesis appears to be supported by Belarusian and Russian examples featuring Sviatlana Tsikhanouskaya (Mateo, 2022) and Alexei Navalny.

The snapshot nature of the data does not fully reflect the dynamic aspects of political discussion on Navalny's YouTube channel. Focusing on the period when the message was left by a user does not allow me to restore the whole context in which the user responded to someone. YouTube algorithms constantly change the configuration of comments when the user observes comments under the option "Top comments." Therefore, it is difficult to go beyond a simple description of discussions and make associations between different aspects of the political conversation.

The applied research design does not allow me to tell anything about who exactly comments on the posts of pro-government commenters in the threads while continuing to express pro-government discourse there. This may include other users responding to relevant signals as well as the authors of the original top-level comments. In general, the profile of commenters

needs to be covered in future studies. It is also necessary to look at the substance of political conversation in the comment section. Undoubtedly, the sentiment of messages and the timing of their posting are not the only reasons users engage with each other when discussing politics. Here, topics around which conversations evolve can reveal other aspects of online political communication between peers within oppositional communities in non-democracies. I also deliberately avoided framing pro-government comments as pro-government astroturfing activity in this study (Sanovich et al., 2018; D. Stukal et al., 2019; D. Stukal et al., 2017). I do not rule out this interpretation entirely, but attempts to identify inauthentic behavior require a different research toolkit and design.

Conclusion

In this study, I examined the discussions that take place in the YouTube community of Russia's most vocal opposition politician, Alexei Navalny. Based on a corpus of nearly eight million comments spanning the years 2013 to 2021, I demonstrated that uncivil comments are more likely to generate discussion threads than civil ones. But this relationship is not straightforward in terms of the correlation "more incivility, more discussion." It rather tells us that to be discussed, a comment must have some potential for deliberation in terms of signaling that the environment is conducive to the expression of opinions in a more frank manner. In addition, the toxicity of comments gets higher over time after a video is posted. This was observed during the first 14 hours after the release of the video. Then, the level of toxicity for top-level comments with threads and messages left in threads stabilizes, remaining approximately at the same level. My analysis also concluded that discussions in Navalny's YouTube comment section are not a manifestation of a bastion of like-minded users who have no opportunity to meet cross-cutting disagreements. Instead, critics of Navalny are visible, and their presence attracts both oppositional sentiments as a response and endorsements from like-minded commenters. Future research should consider a more detailed and in-depth examination of conversations, with a focus on topical aspects of the discussion and contexts controlling, for example, the sentiments of other messages, the profile of commenters, the effect of algorithms, and so on. Regardless, these results point to the need for the analysis of the digital trace data of Russia's political communication because self-reporting methods demonstrated their weaknesses with the start of a new level of the Russian-Ukrainian conflict.

Acknowledgments

I thank two anonymous reviewers whose constructive comments helped to significantly improve this paper. I am also grateful for useful comments on a previous version of the paper from participants of the ICA Regional Conference 2022 "Computational Communication Research in Central and Eastern Europe." Many thanks to the Chair of Digital Governance (holder Yannis Theocharis) at the Technical University of Munich where I was able to fruitfully discuss this paper. I would like to thank Augusto Valeriani for supervising me in this research project.

Data Availability Statement

Online appendices, the code and data underlying this article are available on Open Science Framework: https://osf.io/46rtp/

References

Alrababa'h, A., & Blaydes, L. (2020). Authoritarian media and diversionary threats: Lessons from 30 years of syrian state discourse. *Political Science Research and Methods*, *9*(4), 693–708. https://doi.org/10.1017/psrm.2020.28

Alyukov, M. (2021). News reception and authoritarian control in a hybrid media system: Russian TV viewers and the russia-ukraine conflict. *Politics*, 026339572110414. https://doi.org/10.1177/02633957211041440

AYTAC, S. E. (2021). Effectiveness of incumbent's strategic communication during economic crisis under electoral authoritarianism: Evidence from turkey. *American Political Science Review*, *115*(4), 1517–1523. https://doi.org/10.1017/s0003055421000587

Bastos, M. T., Raimundo, R. L. G., & Travitzki, R. (2013). Gatekeeping twitter: Message diffusion in political hashtags. *Media, Culture & Society*, *35*(2), 260–270. https://doi.org/10.1177/0163443712467594

Berg, J. (2016). The impact of anonymity and issue controversiality on the quality of online discussion. *Journal of Information Technology & Politics*, *13*(1), 37–51. https://doi.org/10.1080/19331681.2015.1131654

Bodrunova, S. S., Litvinenko, A., Blekanov, I., & Nepiyushchikh, D. (2021). Constructive aggression? multiple roles of aggressive content in political discourse on russian YouTube. *Media and Communication*, *9*(1), 181–194. https://doi.org/10.17645/mac.v9i1.3469

Bogoradnikova, D., Makhnytkina, O., Matveev, A., Zakharova, A., & Akulov, A. (2021). Multilingual sentiment analysis and toxicity detection for text messages in russian. *2021 29th Conference of Open Innovations Association (FRUCT)*. https://doi.org/10.23919/fruct52173.2021.9435584

Cho, D., & Kwon, K. H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, *51*, 363–372. https://doi.org/10.1016/j.chb.2015.04.046

Coleman, S., & Freelon, D. (2016). *Handbook of digital politics*. Elgar Publishing Limited, Edward.

Dollbaum, J. M., Lallouet, M., & Noble, B. (2021). *Navalny: Putin's nemesis, russia's future?* Oxford University Press. https://doi.org/10.1093/oso/9780197611708.001.0001

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2016). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35–52. https://doi.org/10.1111/jcc4.12180

Glazunova, S. (2020). "four populisms" of alexey navalny: An analysis of russian non-systemic opposition discourse on YouTube. *Media and Communication*, *8*(4), 121–132. https://doi.org/10.17645/mac.v8i4.3169

Glazunova, S. (2022). *Digital activism in russia*. Springer International Publishing. https://doi.org/10.1007/978-3-030-93503-0

Grimmer, J., Stewart, B. M., & Roberts, M. E. (2022). *Text as data a new framework for machine learning and the social sciences: A new framework for machine learning and the social sciences*. Princeton University Press.

Guriev, S., & Treisman, D. (2022). *Spin dictators: The changing face of tyranny in the 21st century*. Princeton University Press.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? exploring the affordances of facebook and YouTube for political expression. *Computers in Human Behavior*, *29*(3), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

Harel, T. O., Jameson, J. K., & Maoz, I. (2020). The normalization of hatred: Identity, affective polarization, and dehumanization on facebook in the context of intractable political conflict. *Social Media & Society*, *6*(2), 205630512091398. https://doi.org/10.1177/2056305120913983

Harteveld, E. (2021). Ticking all the boxes? a comparative study of social sorting and affective polarization. *Electoral Studies*, *72*, 102337. https://doi.org/10.1016/j.electstud.2021.102337

Huang, C., & Kuo, T.-c. (2022). Actual and perceived polarization on independence-unification views in taiwan. *Asian Journal of Communication*, *32*(2), 75–92. https://doi.org/10.1080/01292986.2021.2022174

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, *58*(4), 621–633. https://doi.org/10.1080/08838151.2014.966365

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Iyengar, S., & Westwood, S. J. (2014). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Kazun, A. (2019). To cover or not to cover: Alexei navalny in russian media. *International Area Studies Review*, *22*(4), 312–326. https://doi.org/10.1177/2233865919846727

Kazun, A., & Semykina, K. (2019). Presidential elections 2018: The struggle of putin and navalny for a media agenda. *Problems of Post-Communism*, *67*(6), 455–466. https://doi.org/10.1080/10758216.2019.1685893

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, *61*(4), 971–988. https://doi.org/10.1111/ajps.12291

Kosmidis, S., & Theocharis, Y. (2020). Can social media incivility induce enthusiasm? *Public Opinion Quarterly*, *84*(S1), 284–308. https://doi.org/10.1093/poq/nfaa014

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? examining public vs interpersonal swearing in response to donald trump's YouTube campaign videos. *Internet Research*, *27*(4), 991–1010. https://doi.org/10.1108/intr-02-2017-0072

Laebens, M. G., & Öztürk, A. (2020). Partisanship and autocratization: Polarization, power asymmetry, and partisan social identities in turkey. *Comparative Political Studies*, *54*(2), 245–279. https://doi.org/10.1177/0010414020926199

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). A new generation of perspective API: Efficient multilingual character-level transformers. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539147

Li, S., Feng, B., Li, N., & Tan, X. (2015). How social context cues in online support-seeking influence self-disclosure in support provision. *Communication Quarterly*, *63*(5), 586–602. https://doi.org/10.1080/01463373.2015.1078389

Litvinenko, A. (2021). YouTube as alternative television in russia: Political videos during the presidential election campaign 2018. *Social Media & Society*, *7*(1), 205630512098445. https://doi.org/10.1177/2056305120984455

Marchal, N. (2021). Be nice or leave me alone: An intergroup perspective on affective polarization in online political discussions. *Communication Research*, *49*(3), 376–398. https://doi.org/10.1177/00936502211042516

Mateo, E. (2022). "all of belarus has come out onto the streets": Exploring nationwide protest and the role of pre-existing social networks. *Post-Soviet Affairs*, *38*(1-2), 26–42. https://doi.org/10.1080/1060586x.2022.2026127

McCoy, J., & Somer, M. (2018). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The ANNALS*

of the American Academy of Political and Social Science, *681*(1), 234–271. https://doi.org/10.1177/0002716218818782

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2019). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, *38*(5), 616–632. https://doi.org/10.1177/0894439319828012

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, *24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Nugent, E. R. (2020). The psychology of repression and polarization. *World Politics*, *72*(2), 291–334. https://doi.org/10.1017/s0043887120000015

Papacharissi, Z. (2014). *Affective publics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199999736.001.0001

Pashakhin, S. (2021). Public agenda fragmentation beyond established democracies: The case of russian online publics in 2017. *Russian Journal of Communication*, *13*(3), 305–324. https://doi.org/10.1080/19409419.2021.1995277

Perry, P. O., & Benoit, K. (2017). Scaling text with the class affinity model. https://doi.org/10.48550/ARXIV.1710.08963

Rajadesingan A., B. C., Resnick P. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 557–568. https://ojs.aaai.org/index.php/ICWSM/article/view/7323

Rega, R., & Marchetti, R. (2021). The strategic use of incivility in contemporary politics. the case of the 2018 italian general election on facebook. *The Communication Review*, *24*(2), 107–132. https://doi.org/10.1080/10714421.2021.1938464

Rieder, B. (2015). Youtube data tools (version 1.22) [software]. https://tools.digitalmethods.net/netvizz/youtube/

Röchert, D., Neubaum, G., Ross, B., Brachten, F., & Stieglitz, S. (2020). Opinion-based homogeneity on YouTube. *Computational Communication Research*, *2*(1), 81–108. https://doi.org/10.5117/ccr2020.1.004.roch

Rojas, H. (2010). "corrective" actions in the public sphere: How perceptions of media and media effects shape political behaviors. *International Journal of Public Opinion Research*, *22*(3), 343–363. https://doi.org/10.1093/ijpor/edq018

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *49*(3), 399–425. https://doi.org/10.1177/0093650220921314

Rozenas, A., & Stukal, D. (2019). How autocrats manipulate economic news: Evidence from russia's state-controlled television. *The Journal of Politics*, *81*(3), 982–996. https://doi.org/10.1086/703208

Sanovich, S., Stukal, D., & Tucker, J. A. (2018). Turning the virtual tables: Government strategies for addressing online opposition with an application to russia. *Comparative Politics*, *50*(3), 435–482. https://doi.org/10.5129/001041518822704890

Shugars, S., & Beauchamp, N. (2019). Why keep arguing? predicting engagement in political conversations online. *SAGE Open*, *9*(1), 215824401982885. https://doi.org/10.1177/2158244019828850

Stromer-Galley, J. (2006). Diversity of political conversation on the internet: Users' perspectives. *Journal of Computer-Mediated Communication*, *8*(3), 0–0. https://doi.org/10.1111/j.1083-6101.2003.tb00215.x

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting bots on russian political twitter. *Big Data*, *5*(4), 310–324. https://doi.org/10.1089/big.2017.0038

Stukal, D., Sanovich, S., Tucker, J. A., & Bonneau, R. (2019). For whom the bot tolls: A neural networks approach to measuring political orientation of twitter bots in russia. *SAGE Open*, *9*(2), 215824401982771. https://doi.org/10.1177/2158244019827715

Stukal, D. K., Akhremenko, A. S., & Petrov, A. P. (2022). Affective political polarization and hate speech: Made for each other? *RUDN Journal of Political Science*, *24*(3), 480–498. https://doi.org/10.22363/2313-1438-2022-24-3-480-498

Suhay, E., Bello-Pardo, E., & Maurer, B. (2017). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, *23*(1), 95–115. https://doi.org/10.1177/1940161217740697

Sun, Q., Wojcieszak, M., & Davidson, S. (2021). Over-time trends in incivility on social media: Evidence from political, non-political, and mixed sub-reddits over eleven years. *Frontiers in Political Science*, *3*. https://doi.org/10.3389/fpos.2021.741605

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on twitter. *SAGE Open*, *10*(2), 215824402091944. https://doi.org/10.1177/2158244020919447

Titov, A. (2017). The timing is just right for navalny to challenge putin's regime. https://blogs.lse.ac.uk/europpblog/2017/06/14/navalny-challenge-to-putin-regime/

Toepfl, F. (2020). Comparing authoritarian publics: The benefits and risks of three types of publics for autocrats. *Communication Theory*, *30*(2), 105–125. https://doi.org/10.1093/ct/qtz015

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, *119*(42). https://doi.org/10.1073/pnas.2207159119

Törnberg, P., & Uitermark, J. (2021). Tweeting ourselves to death: The cultural logic of digital capitalism. *Media, Culture & Society*, *44*(3), 574–590. https://doi.org/10.1177/01634437211053766

Turchenko, M., & Golosov, G. V. (2020). Smart enough to make a difference? an empirical test of the efficacy of strategic voting in russia's authoritarian elections. *Post-Soviet Affairs*, *37*(1), 65–79. https://doi.org/10.1080/1060586x.2020.1796386

Unkel, J., & Kümpel, A. S. (2022). Patterns of incivility on u.s. congress members' social media accounts: A comprehensive analysis of the influence of platform, post, and person characteristics. *Frontiers in Political Science*, *4*. https://doi.org/10.3389/fpos.2022.809805

Urman, A. (2019). News consumption of russian vkontakte users: Polarization and-news avoidance. *International Journal Of Communication*, *13*(25). https://ijoc.org/index.php/ijoc/article/view/11161/2838

Vargo, C. J., & Hopp, T. (2020). Fear, anger, and political advertisement engagement: A computational case study of russian-linked facebook and instagram content. *Journalism & Mass Communication Quarterly*, *97*(3), 743–761. https://doi.org/10.1177/1077699020911884

Voggeser, B. J., Singh, R. K., & Göritz, A. S. (2018). Self-control in online discussions: Disinhibited online behavior as a failure to recognize social cues. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02372

Wojcieszak, M., & Mutz, D. (2009). Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement? *Journal of Communication*, *59*(1), 40–56. https://doi.org/10.1111/j.1460-2466.2008.01403.x

Wu, T.-Y., & Atkin, D. J. (2018). To comment or not to comment: Examining the influences of anonymity and social support on one's willingness to express in online news discussions. *New Media & Society*, *20*(12), 4512–4532. https://doi.org/10.1177/1461444818776629

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, *38*(1-2), 98–139. https://doi.org/10.1080/10584609.2020.1785067