The Effects of Government Persecutions for Social Media Content: Evidence from Russia

Georgii Zherebilov

Einaudi Institute for Economics and Finance (EIEF) & Libera Università Internazionale degli Studi Sociali "Guido Carli" (LUISS)

Abstract

Media censorship is an integral part of authoritarian regimes. While most economic and political research focuses on traditional media censorship that impedes access to information, in this paper I explore censorship that instead mitigates social media activity through selective criminal cases, thus threatening the population. Specifically, I look at the effect of Russian online extremism criminal cases on Twitter activity. Using time variation of cases in Russian regions in an event-study framework, I find that an extra online-extremism case reduces the average number of tweets in a region by 0.17-0.28 standard deviations and increases the average number of political tweets in a region by 0.06-0.11 standard deviations. These results indicate that the government indeed is able to mitigate online activity by harsh legislation-based interventions, but such policy comes at the cost of fostering politics-related discussions.

KEYWORDS: Social media, censorship, extremism, event study

Introduction

Media censorship is a common mechanism for supporting authoritarian regimes. Such regimes spend enormous resources to restrict access to politically sensitive information in order to stay in power (Fedasiuk, 2020). Vast economic literature on media and politics reports that media censorship effectively alters political beliefs in favor of the authorities, not only by directly restricting access to politically sensitive information but also by suppressing the demand for such information (Chen and Yang, 2019). This effect may be even more severe in economies with weak democratic institutions (which is usually the case for autocratic regimes) where political competition is low and media coverage is essential for providing versatile information about the government and the opposition (Enikolopov et al., 2011). Nonetheless, while most papers define media censorship as an action focused on restricting access to politically sensitive information, alternative censorship practices may instead restrict the ability of the population to share and discuss such information. For instance, an authoritarian government may impede access to social media in order to prevent horizontal information exchange. In this case, by restricting the horizontal flows of information, government aggravates the collective action problem, i.e., deteriorates the ability of a population to make a collective effort in order to achieve socially beneficial political outcomes (Enikolopov et al, 2020). One remarkable example of how social media can alleviate collective action problem and lead to beneficial political outcomes is Arab Spring - a famous series of protests throughout the Arab world in the early 2010s which was heavily inspired by social media posts and led to some authoritarian leaders resigning their posts (Acemoglu et al., 2018).

One potential strategy of impeding horizontal flows of information is simply to restrict access to the horizontal-information-exchange technologies, such as social media (e.g. Facebook or Twitter). This is in line with traditional censorship practices and is used fairly often in autocratic regimes throughout the world. For example, during the Belarusian protests of 2020, the government practiced mass Internet shutdowns in order to impede protest coordination (Bloomberg, 2020). The same action was performed by

the Iranian government during the protests of November 2019 (The Washington Post, 2019).

Another strategy avoids direct access restrictions that may be quite expensive and hard to implement. Instead, it creates or uses existing legislation base to legally persecute those who spread politically sensitive information online, such as information about protests and criticism of authorities. Such strategy relies on standard mechanisms of crime and punishment models which operate with the probability of being punished for illegal action and losses imposed for this action (Becker, 1986). In this scope, every person who posts politically sensitive information online faces some probability of being accused of spreading illegal information and some potential punishment for this action (e.g., fees or incarceration). As long as the government can affect both the probability of being accused (by altering the extent of monitoring and legislation) and the severity of punishment, it can potentially create an effective system where the expected losses of posting and sharing politically sensitive information are higher than the benefits. In this case, the government technically imposes a self-censorship mechanism without resorting to standard (and reputationally costly) restriction practices. Moreover, by increasing the severity of the punishment, it can avoid costly overarching monitoring while keeping the expected losses of posting politically sensitive content high enough. Therefore, potentially, such a way of impeding horizontal flows of information may be as effective as standard censorship practices while being significantly cheaper.

In this paper, I explore such a strategy of censorship in the setting of current Russian legislation, that is severely misused to restrict freedom of speech. Specifically, I explore the empirical effects of anti-extremism legislation (the major source of internet-related cases) on Twitter activity in Russian regions. As a treatment, I use the timing of the criminal cases on extremism actions done via the Internet, available online by region. As an outcome, I use the total and the average number of tweets posted weekly in every region. Using time variation of cases in different regions I employ an event study to see

the dynamics of tweets posted around a time window of every case available. I find that after another case in a region the average number of tweets drops by 0.17-0.28 standard deviations and the average number of politics-related tweets rises by 0.06-0.11 standard deviations. This finding implies that, while overall the anti-extremism legislation has some censorship effects, it can also backfire by inciting active discussions between politically active users of Twitter.

This finding contributes to the large body of research on the political economy of social media (Zhuravskaya, Petrova, Enikolopov, ARE, 2020). First, it brings novel empirical evidence of how mechanisms of horizontal information flows on social media cope with government-induced shocks interrupting these flows. I show that harsh interventions by the government indeed have a significant effect on Twitter activity. Nonetheless, the effect on politically related content is positive, therefore Twitter users are able to confront the misuse of legislation by actively discussing it online. A piece of anecdotal evidence for this interpretation is the alleviation of anti-extremism legislation in October of 2018 provoked by hot discussions about legislation misuse on the Internet (Vedomosti, 2018). Second, past evidence suggests that in regimes with weak political institutes media outlets can significantly affect political outcomes, e.g. elections, by enriching prior knowledge of voters about parties (Enikolopov et al., 2011). This paper suggests that government can significantly affect political outcomes through harsh media interventions. The benefits of such interventions, therefore, depend on the magnitude and the sign of the censorship effects. For example, if another anti-extremism case incites political debates online, it can affect current authorities negatively by spreading the knowledge of the government's misuse of legislation.

In what follows, I provide a brief outline of current anti-extremism legislation in Russia and known cases of its misuse in Section I. In Section II I describe the data used. Section III describes the empirical model of the study. Section IV describes the results of the empirical analysis. Section V discusses the robustness of the results and future potential

advancements. Section VI concludes.

1 Anti-Extremism Legislation in Russia and the Cases of its Misuse

In 2017 a doctor from Khabarovsk Region became a suspect in a criminal case of extremism (Article 282 of the Criminal Code of the Russian Federation). His lawyer said that the case was open because the suspect liked the picture condemning the participation of Russians in the war in Eastern Ukraine on Odnoklassniki – a Russian social network (Meduza, 2018). In 2018, a resident of St. Petersburg, Eduard Nikitin, was accused under the same article for reposts made in 2015. The case was based on two messages: in the first case, Nikitin posted a joke about elections, in the second – a caricature of Russian patriots (Meduza, 2018).

These cases are not unique: in the peak years of 2017-2018, there were more than 500 cases of internet extremism reported in Russia[1]. The most commonly used article of the Russian criminal code for internet extremism is article 282 – incitement of hatred or enmity, as well as humiliation of human dignity. Other articles used less frequently are 280 (public calls for extremist) and 280.1 (Public calls for the implementation of actions aimed at violating the territorial integrity of the Russian Federation). The dynamics of the cases based on these articles throughout years is presented in Figure 1.

Many of these cases are considered to be a misuse of anti-extremism legislation. For example, in 2019 Russian center of extremism studies Sova reports that out of 97 sentences for hate speech only 27 are legitimate (Sova, 2019). This evidence suggests that anti-extremism legislation is commonly used by authorities to restrict freedom of speech.

In 2018, after many cases of anti-extremism legislation misuse had received widespread publicity in mass media, the legislation was alleviated: some divisions of anti-extremism

[1] https://beta.dostoevsky.io/en-GB/

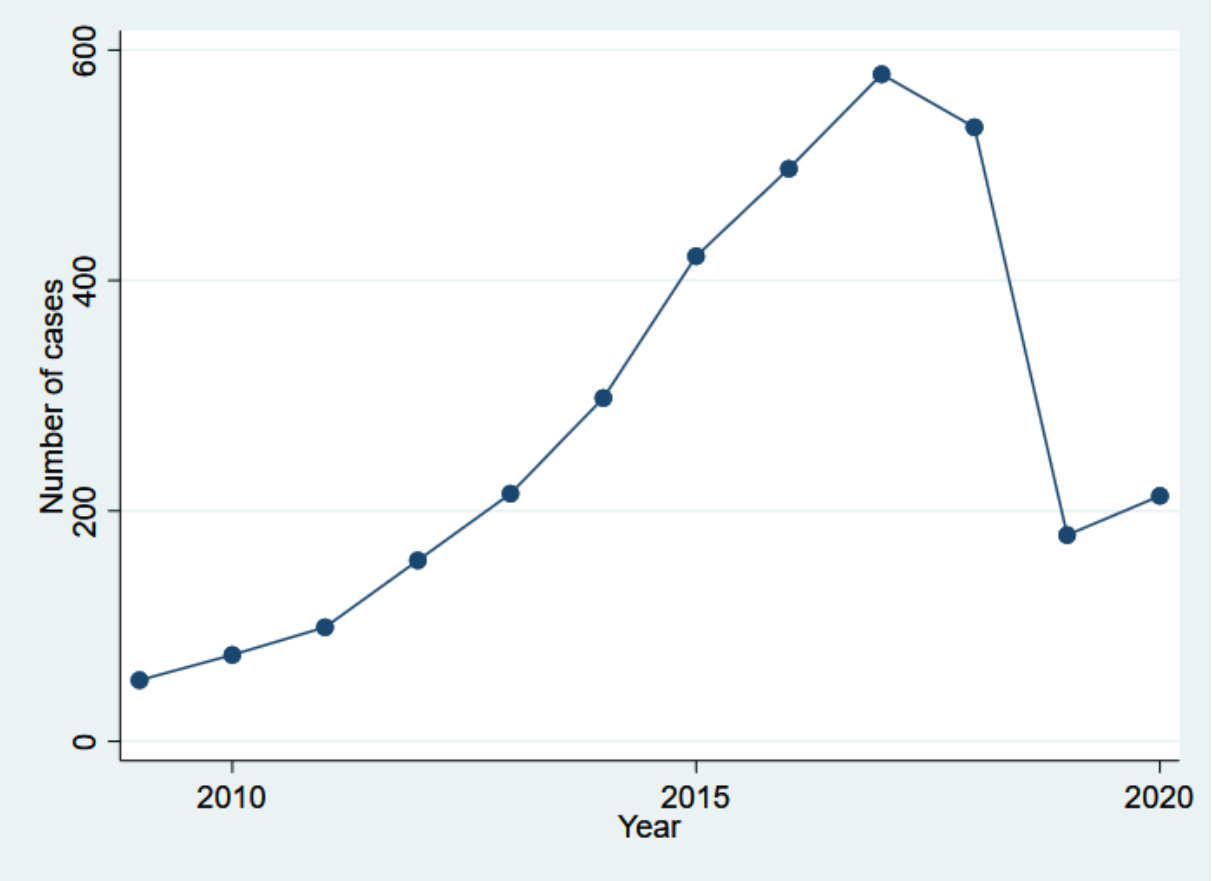


Figure 1: Total number of internet extremism cases: Dostoevsky

articles have been decriminalized and moved from the criminal code to the administrative code. After this, there was a significant drop in the number of cases (Figure 1). Nonetheless, as Sova reports, legislation misuse is still a common practice (Sova, 2019).

2 Data

I use several sources of data for this exercise. To measure Twitter activity across regions, I calculate the average and the total number of weekly tweets of a random sample of Twitter users by region. The sample is obtained as follows. I randomly scatter 10000 points on a map of Russia (Figure 2). Then I obtain tweets posted around these points and users who posted these tweets. Nonetheless, the geolocation of a tweet does not uniquely identify the permanent residence of a user. Therefore, I leave only accounts with user locations stated in a profile. This method is unable to retrieve accounts with no user location and with no tweets with geolocation. Therefore, I have to assume that the decision to state the

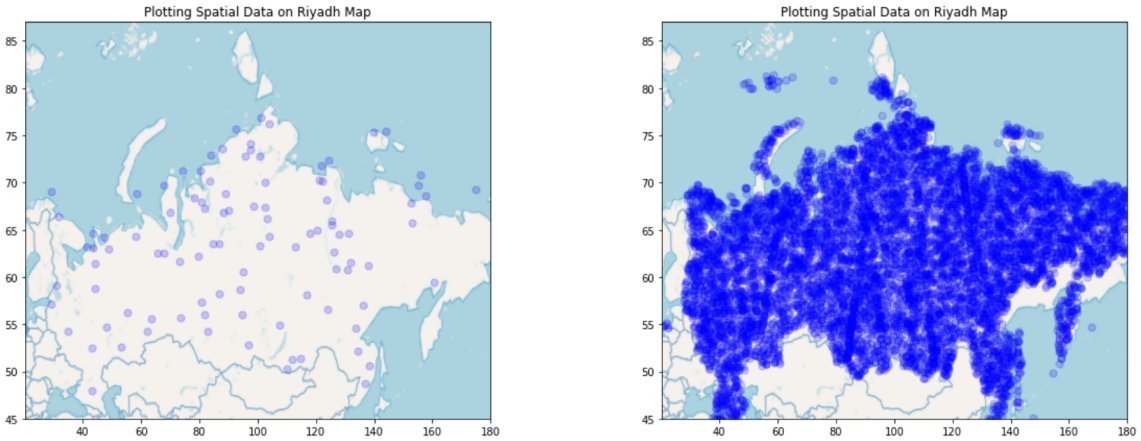
location and to use geolocation of tweets is not related to the posting activity of a user. In the end, 10000 randomly scattered points result in approximately 6000 unique Twitter accounts. Out of them, 2642 accounts from 55 regions constitute the final sample of open accounts with user location available (Moscow, Saint-Petersburg, and regions with no extremism cases are omitted). The number of accounts in every region ranges from 3 to 130.

For every user, I then download all the tweets available, count the weekly number of all tweets and politics-related tweets. I flag a tweet as political if it contains at least one keyword from a list of words related to politics with high probability like "Putin", "Navalny", "protest", etc. I also include in the list words that are likely to be related to major political events, like "poisoning"[2]. The list of the words used is provided in Appendix I.

Twitter API allows downloading only the last 3200 tweets for every account. That means that for some users I do not observe data before some date when the cap of 3200 tweets was exceeded. Technically, this fact can bias our estimates as the truncation is not random: there's more omitted data for the accounts that post a lot of content. In the final sample 38% of accounts exceed this cap. Nevertheless, the problem is likely to be not very severe because most of the accounts that exceed the cap severely seem to be either bots or news feeds, that are not of the interest for the research question. Therefore, I try two specifications of the model: either I keep the accounts that exceed the cap and make sure that they do not disappear in the time window of any event, or I drop them.

For the treatment I use data provided by the online system "Pravosudie" that provides data on the federal courts' performance, including the registry of all criminal cases by region. I focus on the cases of public calls for extremism and hate crimes done via the Internet (a full list of articles is provided in Appendix II). The exact date of a case

[2]On 20 August 2020, Russian opposition figure and anti-corruption activist Alexei Navalny was poisoned with a Novichok nerve agent and was hospitalized in serious condition. This case was broadly discussed in social networks.



(a) 100 points (b) 10000 points

Figure 2: Randomly scattered points on the map of Russia

is therefore the date of treatment. Figure 3 shows the dynamics of the cases in Russia throughout the years.

I use only the cases with sentences released[3]. In addition, I omit cases with intersecting time windows. If there are multiple cases happening in a single week, I use the number of cases as treatment. Therefore, the treatment is a categorical variable, showing the number of cases in a single week. The final sample has 55 regions, 51 of them having at least one case, not intersecting with others. In my analysis I use only data for years 2020-2021 for several reasons. First, in this case less users exceed the cap of 3200 tweets. Second, it is easier to flag political content related to some political events. For example, if a tweet contains the word "poisoning" in 2020, it is most likely related to the poisoning of Navalny. Lastly, as I mentioned before, in 2018 anti-extremism legislation was alleviated. Therefore, it would be misleading to equalize anti-extremism cases before and after 2018.

3 Empirical Model

The main hypothesis of the study is that internet extremism cases significantly affect posting activity in social media, either for all content or only for political content. The

[3]In section 5 I briefly discuss the specification with all the cases included.

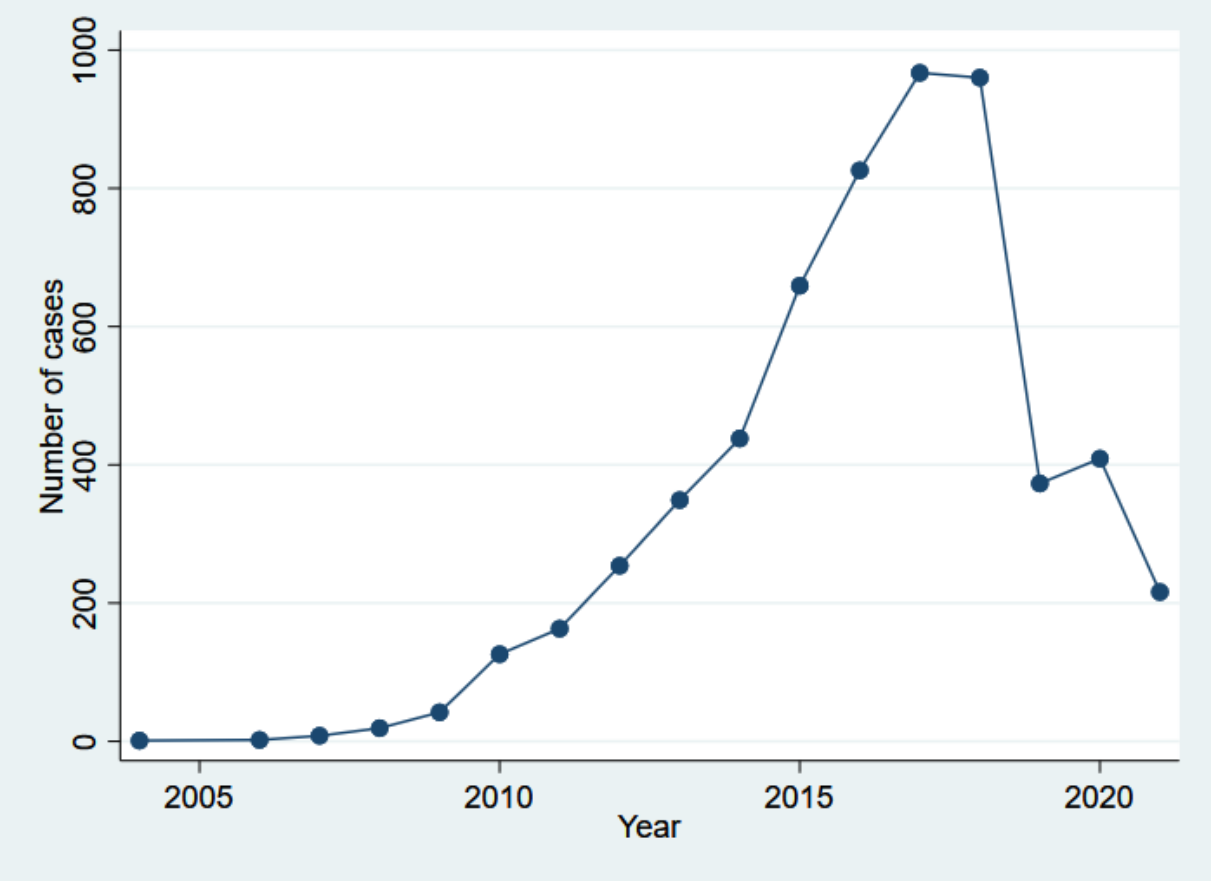


Figure 3: Total number of internet extremism cases: Scraping

sign of the effect can potentially be both positive and negative: either people can mitigate their activity out of fear of being persecuted, or they can try to discuss cases more actively in order to give them publicity.

OLS estimates of this model may be biased because there are potential omitted region characteristics that may affect both online activity and number of internet extremism cases, e.g. technological development. To circumvent this problem I employ time variation in the timing of extremism cases across regions in an event-study framework. This strategy uses the fact that extremism cases in different regions happen at different times and allows us to use regions with no cases in a time window as controls. The main assumption of this approach is that the timing of the case is exogenous to the evolution of the outcome within a certain relatively narrow time window. In the framework of this study, it means that new cases must not be based on the region's online activity. If this is the

case, controlling for time and region fixed effects, we can claim that the effect on the online activity after a new case is driven solely by this case and not by unobserved confounders. This assumption may fail if there is some monitoring performed by authorities. If this is the case, higher activity in a region may induce authorities to increase monitoring, and this will result in more cases being opened. Despite the Russian government having had plans for unfolding country-wide monitoring systems (Meduza, 2018), these projects have not been realized. Usually, the cases are based on third-party denunciations: anybody can file a complaint about a post on social media, and the authorities have to proceed with it (Meduza, 2018). Therefore, it does not seem that new cases are based on regional activity. It is more likely that every new case is a result of local region idiosyncrasies. Therefore, the event study framework seems to be a plausible way to proceed.

I use the following specification:

$$y_{it} = \eta_i + \gamma_t + \sum_{k=-3}^{4} \beta_k * I(K_{it} = k) + \epsilon_{it} \tag{1}$$

The primary outcome of interest y_{it} is a posting activity measure in the region i in week t. For activity measures, I use the average and the total number of tweets in a region in a specific week. I control for region fixed effects η_i and for year-week fixed effects γ_t.

Let E_i be the week in which there was a new case opened in region i. Then K_{it} shows time relative to this event, that is $K_{it} = t - E_i$. Therefore, coefficient β_k represents the effect of another internet extremism case that happens k weeks before week t if $k > 0$ or k weeks after week t if $k < 0$.

It is needed to say that most often there are multiple cases in a region. This fact brings the trade off related to the time window selected. A wider time window mitigates the problem of omitted variable bias, as with a too narrow time window we can mistakenly attribute the effects of further periods to the earlier periods. In this case, we underestimate the long-run effects of the cases and overestimate the short-run effects. Meanwhile,

when making the time window wider, we have to exclude more cases from the analysis because there must be no intersections between time windows of different cases. The main specification defines the time window as (-3, 4) with the coefficient of $t = -1$ being omitted. It is assumed that the effect of a case after 4 weeks is significantly negligible. The robustness section provides specifications with the regions having only one case in years 2020-2021 and a fully saturated event study model.

4 Results

Figure 4 provides the results for the main specification of the model. It shows the effect of internet extremism cases on the average number of all tweets by region for all users (including those who exceed the cap). Year-week and region fixed effects are included. Time window of an event is defined on a scale from 3 weeks before the case to 4 weeks after the case[4].

The pre-case coefficients are negative but not statistically significant. That is, prior to a new case, there are no significant differences in Twitter activity in the regions that would face a case later and those that wouldn't. That is, there is no significantly different Twitter activity trend in the regions prior to a new case. This indicates that the timing of a new case is not related to the preceding Twitter activity. All the post-case coefficients are negative and significant at the 95-99% level. The instant effect on the week when a new case is open is -4.20 (0.17 standard deviations). That is, on the week of a new case Twitter activity drops by 4.20 tweets on average. The effect is increasing in time, up to a drop of 6.79 tweets (0.28 standard deviations) on average by week 4 after a new case is open. The results suggest that there is a significant lasting (at least 4 weeks) negative effect of internet extremism cases on Twitter activity.

Table 1 provides a number of alternative specifications for the subset of the users who

[4]Standard errors are clustered on the regional level in all specifications.

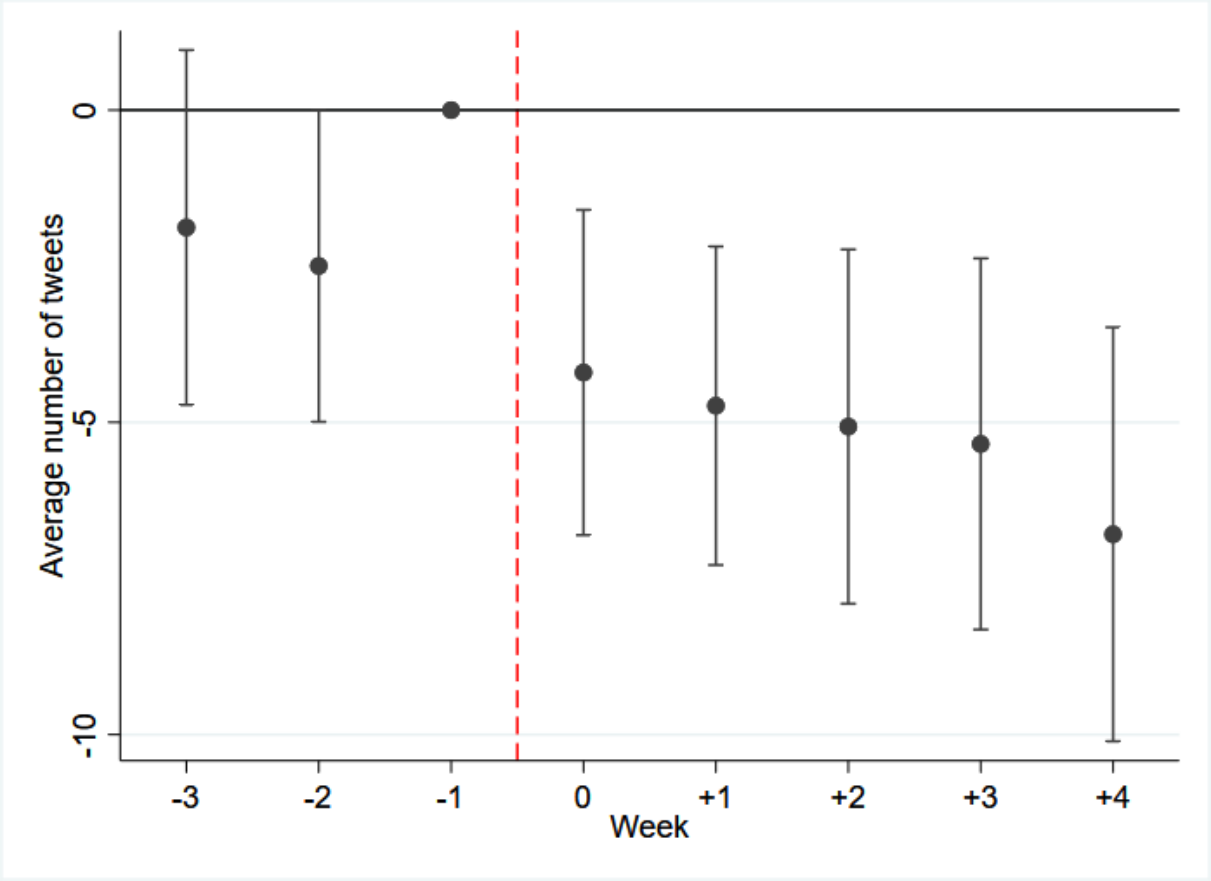


Figure 4: The effect of internet criminal cases on Twitter activity: all tweets

do not exceed the 3200 tweets cap[5]. In these alternative specifications, I use the total number of tweets as a measure of Twitter activity in a region as well as the natural logarithms of both the average and the total number of tweets[6]. In addition, one specification uses data at the individual instead of the regional level. Nonetheless, potentially due to no variation in the cases within the region, the results for this specification are insignificant. The analysis that follows relates to the specifications at the regional level.

There are a couple of important features present in all of the specifications. First, in all specifications, post-case coefficients are negative and mostly 95-99% significant. There are specifications where there is no significant instant effect, but it becomes significant starting from week 2. It can be explained by the fact that as a treatment date I use the date when the case was just open and it takes time to proceed with the decision and for

[5]All the tables report specifications for the subset of users who do not exceed the cap.
[6]One is added to every variable when calculating logs.

Table 1: Twitter activity: All tweets

	(1) Average	(2) Total	(3) ln(Average)	(4) ln(Total)	(5) Individual
3 weeks before	-1.265	-93.44*	-0.0460	-0.0575	-0.384
	(-1.11)	(-2.66)	(-1.10)	(-1.12)	(-0.97)
2 weeks before	-0.861	-86.37*	-0.0173	-0.0288	0.0674
	(-0.87)	(-2.26)	(-0.47)	(-0.64)	(0.16)
Week of treatment	-1.720	-114.2**	-0.0560	-0.0716	-0.897
	(-1.61)	(-2.79)	(-1.32)	(-1.42)	(-1.78)
1 week after	-1.694	-119.8**	-0.0694	-0.0876	-0.679
	(-1.36)	(-3.01)	(-1.63)	(-1.75)	(-1.59)
2 week after	-1.598	-116.7**	-0.0464	-0.0641	-0.148
	(-1.40)	(-2.94)	(-1.15)	(-1.34)	(-0.33)
3 week after	-2.077*	-137.4**	-0.0846*	-0.108*	-0.284
	(-2.02)	(-2.88)	(-2.22)	(-2.31)	(-0.76)
4 week after	-2.320*	-137.2**	-0.0747	-0.0987*	-0.391
	(-2.30)	(-2.87)	(-1.94)	(-2.21)	(-1.02)
Observations	2991	2991	2991	2991	85463
Region FE	YES	YES	YES	YES	YES
Year-week FE	YES	YES	YES	YES	YES
Individual FE	NO	NO	NO	NO	YES

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

the news to spread in the population. This as well can be the reason for the increasing negative trend of the effect. Meanwhile, pre-case coefficients are closer to zero and less significant in all specifications. Nonetheless, they are still negative and significant at a 95% level in the specification with the total number of tweets used as a measure of Twitter activity.

This disturbance can be potentially caused by many reasons. First, there are possible spillovers between regions: it can be that a new case in one region also affects another regions' Twitter activity. Suppose there was a case at time t in a region A and a case at time $t+1$. If there are spillovers present, the effect from region A will transfer to region B at time t, that is, one week before the case in the region B. In this case, we will observe a drop in activity in the region B before an actual case in this region. This can be the reason why pre-case coefficients are negative and sometimes significant. Another potential reason for negative pre-case coefficients is the ability of users to delete tweets. If a user learns about a new case in her region, she can not only mitigate her activity but also delete tweets from the past that she finds potentially sensitive. If users tend to delete sensitive tweets in the past, we can see a negative effect spreading into the past as well. Finally, it can be the case that people learn about new cases before they are officially open. For example, a user can inform others that authorities have filed a complaint about her tweets before the case officially proceeds. If users learned about the cases before they are open, it could create a significantly negative effect before the date of the treatment. In the next section, I discuss potential strategies to test these hypotheses.

Figure 5 provides the results for the main specification of the model for the politics-related tweets. It shows the effect of internet extremism cases on the average number of political tweets by region for all users (including those who exceed the cap). Year-week and region fixed effects are included. The time window of an event is defined on a scale from 3 weeks before the case to 4 weeks after the case.

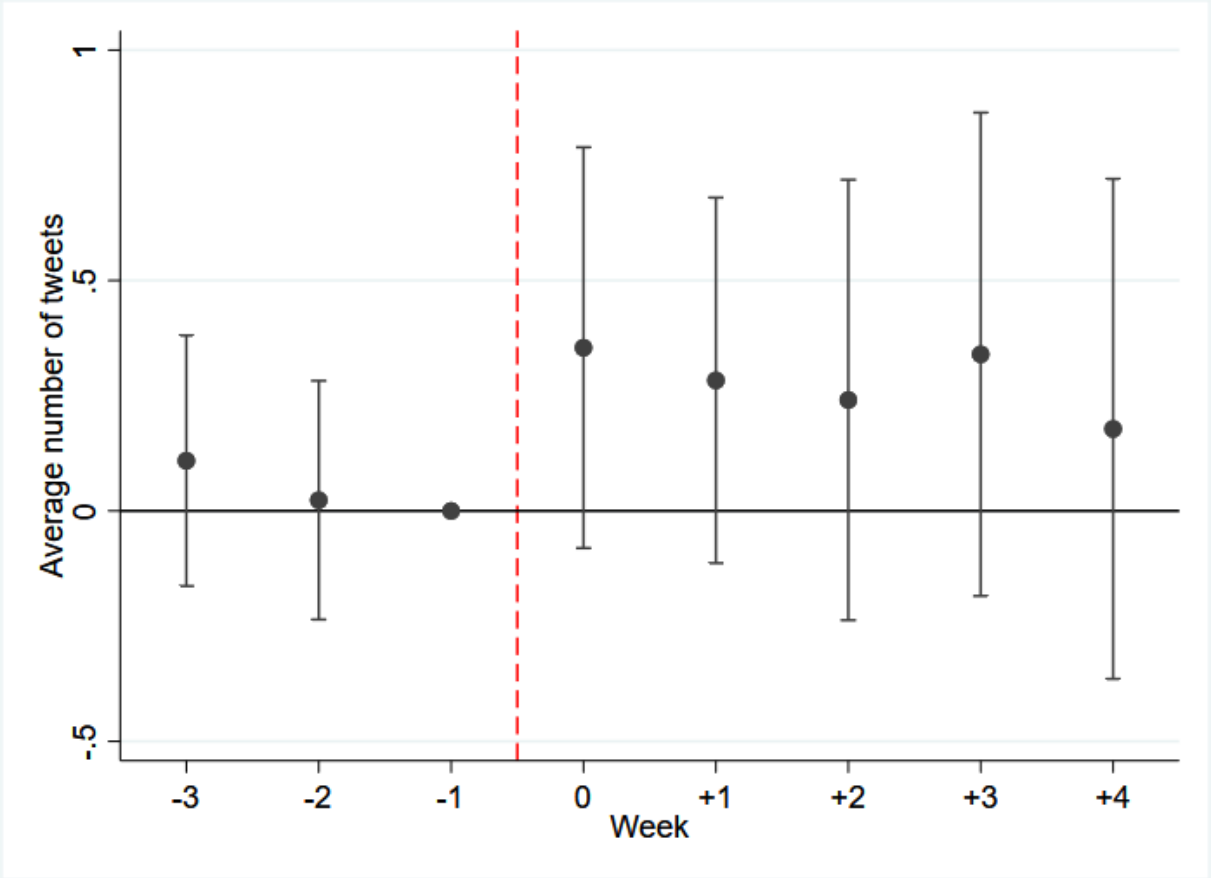


Figure 5: The effect of internet criminal cases on Twitter activity: political tweets

It can be clearly seen that post-case coefficients are all positive albeit not statistically significant. The instant effect on the week of a new case is equal to +0,35 (0.11 standard deviations). That is, on the week of a new internet extremism case Twitter users post on average 0.35 tweets more. This effect is persistent throughout time from 1 week after a case to 4 weeks after a case and ranges between +0.17 (0.057 standard deviations) at week 4 to +0.34 (0.11 standard deviations) at week 3.

Lack of significance can be provoked by some reasons. First, users do not post much about politics. It results in a distribution of the average number of tweets being extremely skewed towards zero and very low variance between regions. The standard deviation of the average number of political tweets is equal to 3,13 while for the full sample of tweets it is equal to 24,63. Figure 6 shows the empirical distribution of the average number of political tweets by region. Therefore, there is probably not enough variation in the data

for the results to be significant enough.

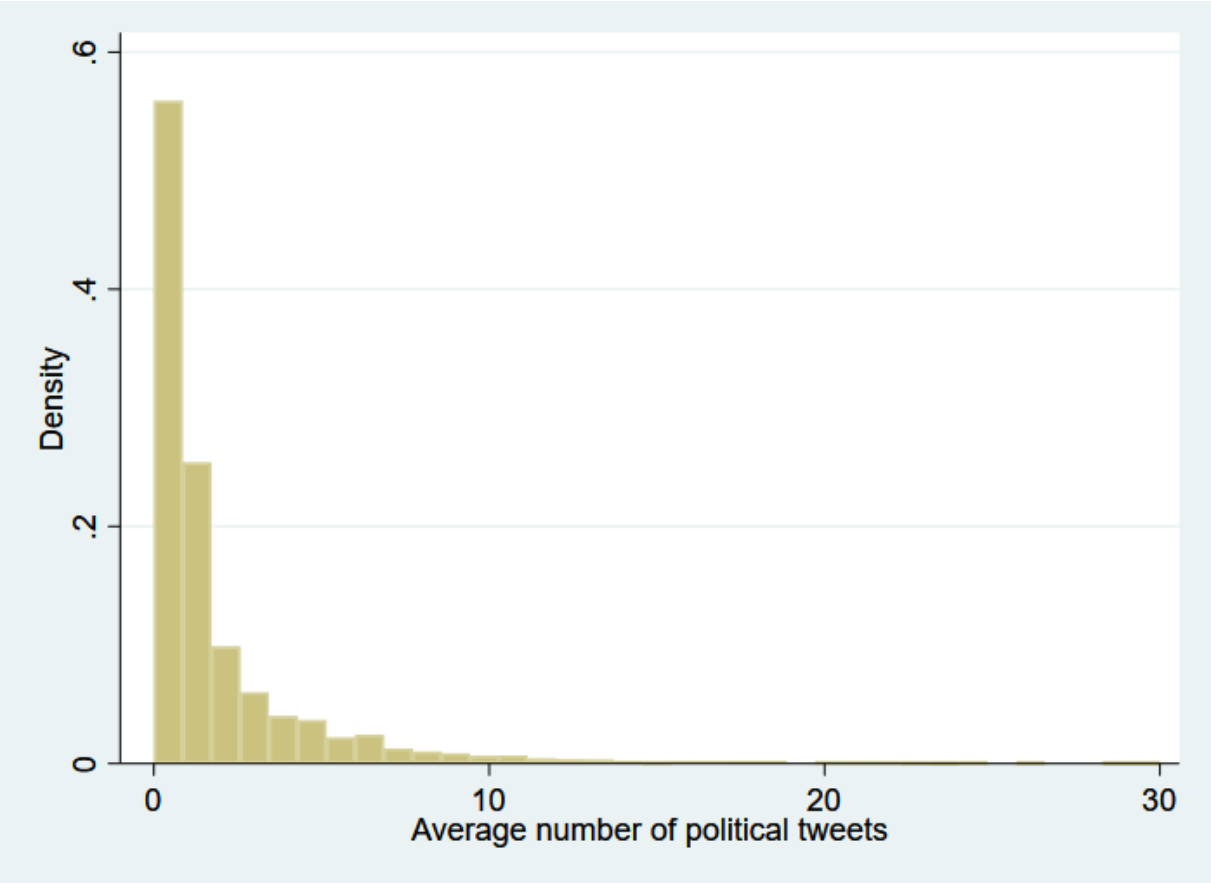


Figure 6: Political tweets distribution

Another reason might be the poor definition of politically related tweets. I flag a tweet as political if it contains a word that is likely to be related to politics. Obviously, this procedure is not very precise. For instance, a tweet containing the word "Tsar" can contain a critique of the last-longing presidency of Vladimir Putin. On another hand, a tweet with the same word can be purely historical, providing information about past Russian emperors. In addition, this strategy does not separate tweets based on their emotional and critical extent. For example, a tweet containing the word "Putin" can either support or criticize the president of Russia. It is more likely that there will be no effect of an internet extremism case for the tweets containing positive judgments about the current regime. Nonetheless, the results are still telling. Table 2 provides more specifications for the model with political tweets. Again, every specification shows an important feature:

post-case coefficients are positive, higher, and more significant than the pre-case coefficients. This evidence shows that there is a positive but statistically insignificant effect of internet extremism cases on Twitter political activity. This fact suggests that Twitter users post slightly more political tweets when there is a new case open in a region. Nonetheless, the data and the model used cannot reject the hypothesis of this effect being zero.

Table 2: Twitter activity: Political tweets

	(1) Average	(2) Total	(3) ln(Average)	(4) ln(Total)	(5) Individual
3 weeks before	0.109	2.433	0.0279	0.0702	0.114
	(0.80)	(0.45)	(0.81)	(1.04)	(1.04)
2 weeks before	0.0233	-0.110	-0.0148	-0.0538	0.0524
	(0.18)	(-0.02)	(-0.46)	(-0.81)	(0.41)
Week of treatment	0.354	9.763	0.0545	0.0889	0.295
	(1.63)	(1.00)	(1.43)	(1.23)	(1.60)
1 week after	0.283	13.60	0.0598	0.120	0.383
	(1.43)	(0.96)	(1.64)	(1.69)	(1.52)
2 week after	0.241	16.00	0.0482	0.114	0.468
	(1.01)	(1.02)	(1.47)	(1.83)	(1.58)
3 week after	0.340	14.01	0.0404	0.0710	0.428
	(1.30)	(0.88)	(1.12)	(1.05)	(1.49)
4 week after	0.178	7.145	0.0127	0.0206	0.293
	(0.66)	(0.51)	(0.28)	(0.27)	(1.15)
Observations	2991	2991	2991	2991	115068
Region FE	YES	YES	YES	YES	YES
Year-week FE	YES	YES	YES	YES	YES
Individual FE	NO	NO	NO	NO	YES

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Overall, our model suggests that in general Twitter users post less content when they face another internet extremism case in their region. That is, the extremism legislation misuse by Russian authorities has some self-censorship effect for Twitter activity. Nonetheless, what is peculiar, the number of political tweets increases after another case

is open. In general, people learn about potential prosecutions and become less active on Twitter, while there still exists some politically active sub-population of Twitter users that takes the risk of publicly discussing the decisions of the government.

5 Robustness and Future Potential Advancements

In this section, I provide different alternative specifications of the model to assess its robustness. First, I estimate the main specification from section 4 using all the cases available, not only those with a sentence released. The aim of this exercise is to explore the heterogeneity of the effect of cases with different outcomes. Potentially, users react differently on cases with sentences released (specifically if punishment is applied) and on cases that still proceed or, for example, have been canceled. Figure 7 shows the effect of all internet extremism cases on the average weekly amount of tweets by region.

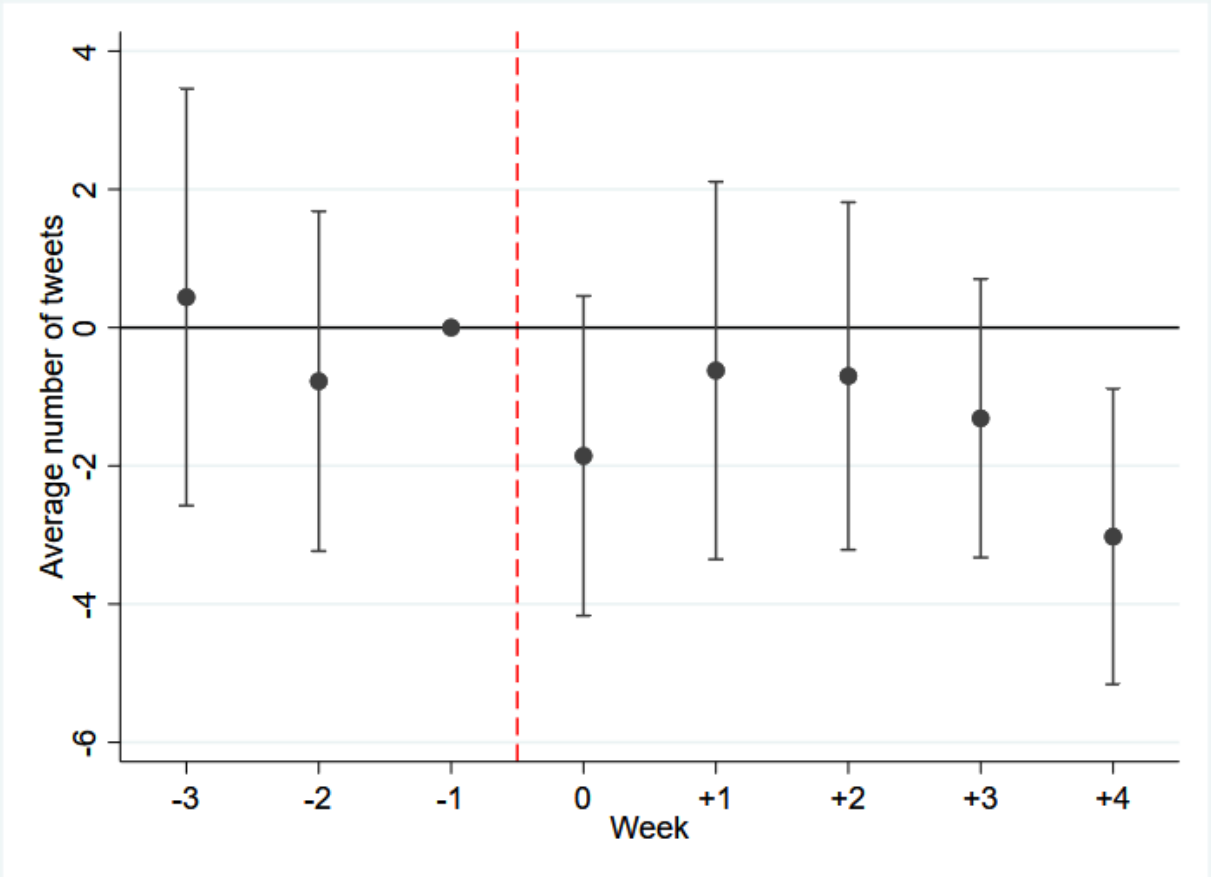


Figure 7: The effect of internet criminal cases on Twitter activity: all tweets, all cases

The effect is less significant than the specification with the cases having sentences released, leaving only the coefficient for week 4 after a case to be significantly negative. It suggests that there is indeed heterogeneity in the effects caused by cases with different outcomes. Specifically, users do not react too strongly to the cases with no sentence released. Figure 8 shows the effect of all internet extremism cases on the average weekly amount of political tweets by region. Oppositely to the specification with all the tweets included, including all cases in the specification with political tweets only increases the magnitude and significance of the effect. Another internet extremism case increases the average number of political tweets by 0.23 standard deviations in week 1 compared to the baseline effect of 0.09 standard deviations tweets in week 1. This effect is also significant at a 95% level. Nevertheless, the coefficients for weeks 2-4 are insignificant. All other specifications also provide insignificant results. Obtaining higher effects while including all cases seems counterintuitive. It means that the effect of the cases with no outcome is stronger than the effect of the cases with sentences released for political tweets. Potentially the reason may be reversed causality. The treatment date is the date when the case is open, and it is unlikely that a case can be finalized in one week. Therefore, it can be that wide public discussion can lead to the closure of a case. Nevertheless, such weak magnitude (less than 1 tweet increase on average) is unlikely to affect the decisions of authorities, therefore more research is needed in order to explain this counterintuitive effect.

In the following exercise, I leave out only the regions with one case in the years 2020-2021. By doing so I can estimate a fully saturated event study model, as far as there are no intersections of time windows of different cases. The benefit of this model is that we do not omit weeks far before or after the case. In such a manner we can estimate the effects more precisely as there is no omitted variable bias emerging in the case of a narrow time window. The goal is to estimate the following specification:

$$y_{it} = \eta_i + \gamma_t + \sum_{k=-\infty}^{+\infty} \beta_k * I(K_{it} = k) + \epsilon_{it} \tag{2}$$

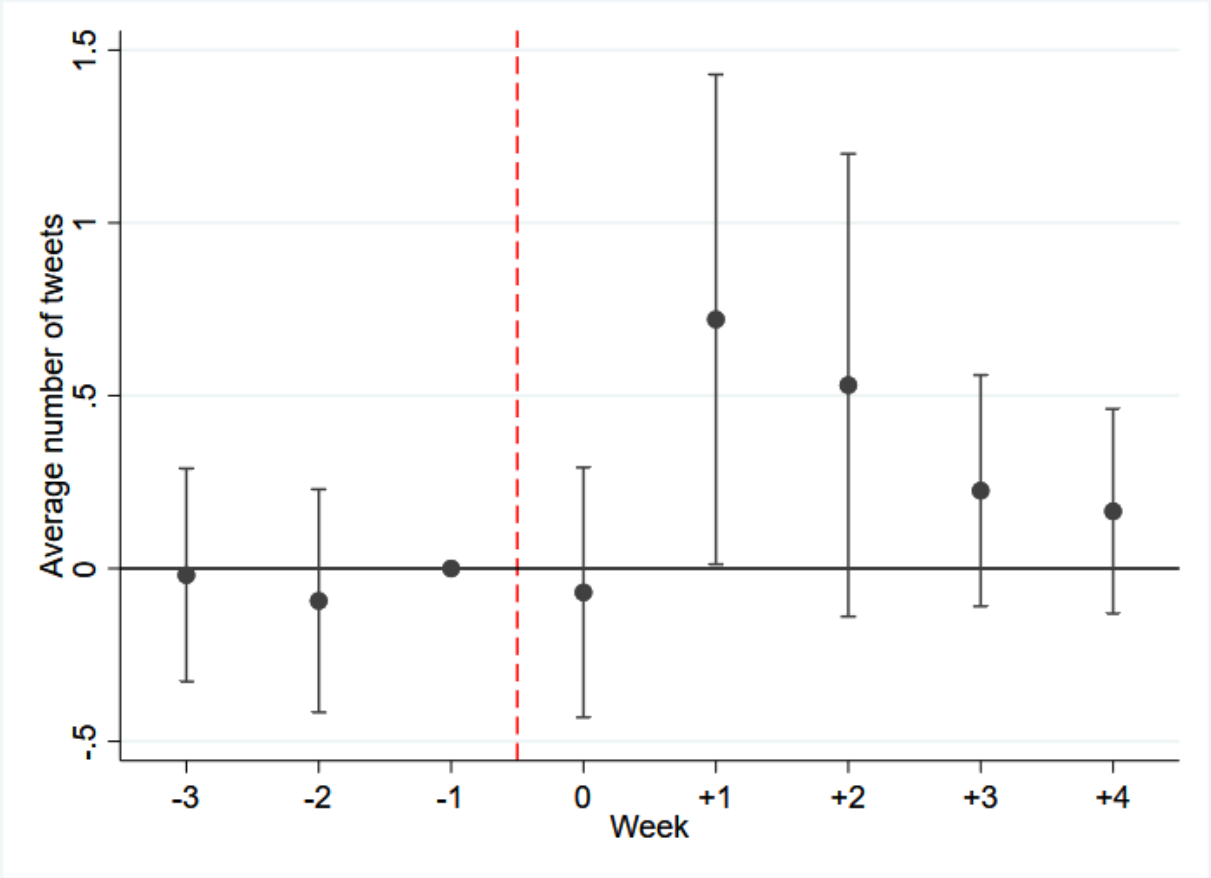


Figure 8: The effect of internet criminal cases on Twitter activity: political tweets, all cases

Nonetheless, Borusyak and Jaravel (2017) report that in such a model the coefficients of interest cannot be identified due to multicollinearity coming from the fact that the time fixed effects and event time are linearly dependent. The solution is to bin the effects of some weeks. Specifically, I use the following specification:

$$y_{it} = \eta_i + \gamma_t + \beta_{-4} * I(K_{it} < -3) + \sum_{k=-3}^{+4} \beta_k * I(K_{it} = k) + \beta_5 * I(K_{it} > 4) + \epsilon_{it} \qquad (3)$$

That is, I bin coefficients that are more than 3 weeks before and more than 4 weeks after a case. Figure 8 provides the estimates for this model for all tweets. Figure 9 provides the estimates for political tweets. In both cases, the magnitude of the effects is smaller and the effects are not significant. Although considering the fact that there are only 11 out of 55 regions left, one cannot be sure whether the effects are insignificant because we use a broader time window or because the altered sample is too small.

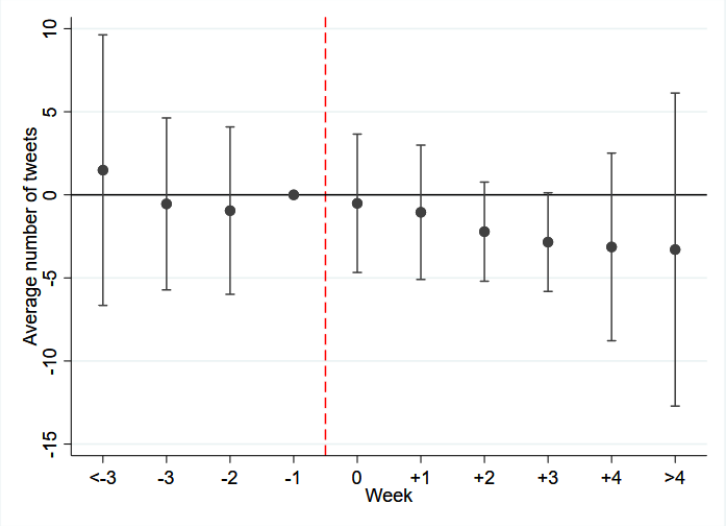


Figure 9: The effect of internet criminal cases on Twitter activity: political tweets, fully saturated model

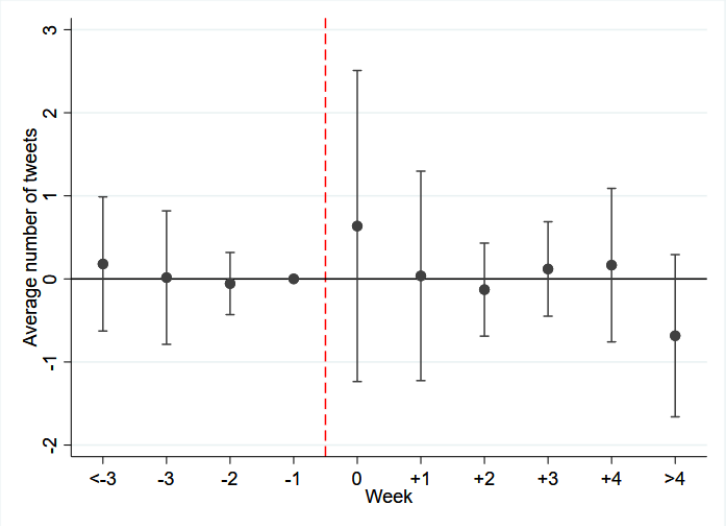


Figure 10: The effect of internet criminal cases on Twitter activity: political tweets, fully saturated model

The rest of this section is devoted to the discussion of future potential advancements. One issue of this paper is that the pre-case coefficients are negative and significant in some specifications. As discussed above, there are some potential reasons for this, specifically, between-region spillover and tweet deletions. It could be useful to test these two hypotheses.

The spillover hypotheses can be tested by including coefficients for the cases from different regions. One potential problem is that this would greatly increase the number of time window intersections, therefore a lot of cases would be dropped. Ideally, it would be useful to assess how widely known every case is, for example, parsing news feeds. To test the deletion hypothesis one can estimate a model with more coefficients for the weeks before a case. If some of them were negative, it could mean that users retroactively delete past tweets once they learn about an internet extremism case in their region.

Another issue is a poor definition of political tweets. Flagging tweets as political based only on the inclusion of keywords may be quite imprecise. For example, when people learn that some content can be perceived as potentially extremist, they may start disguising their messages. One famous example of such behavior is Chinese people using pictures of Winnie The Pooh to depict the president of China Xi Jinping (The Guardian, 2018). Obviously, such content wouldn't be flagged as political by the mechanism I use. Moreover, tweets containing the same exact words related to politics may have completely different meanings and opposite opinions. It is more likely that a tweet criticizing the government would be deemed to be extremist rather than the one praising the government. It would be of great use to apply a more precise mechanism of flagging political tweets, for example, with the help of machine learning.

6 Conclusion

Social media is crucial for modern politics for its ability to facilitate horizontal information flows and mitigate the collective action problem. Unsurprisingly, authoritarian regimes and other actors experiment with various ways of censoring and suppressing the flow of politically sensitive information. In this paper, I study the effect of internet extremism legislation misuse on Twitter activity. I find evidence this effect is present and statistically significant. That is, on the week when there's a new case open in a region, the average number of tweets decreases by 0.17 standard deviations. The effect is persistent and in-

creasing in time for at least 1 month: the drop in the average number of tweets increases from 0.19 standard deviations on week 1 after a case to 0.28 standard deviations on week 4 after a case. This evidence suggests that the Russian government can successfully mitigate online activity by harsh legislation-based interventions.

What is peculiar, the effect is significantly negative for all tweets, not for political ones. That can indirectly acknowledge the ambiguity of Russian anti-extremism legislation: users are not sure what content can be perceived as extremist and post less content in general. What is even more peculiar, the model shows that the effect of internet extremism cases is positive for political tweets. On the week of a new case, the average number of political tweets increases by 0.11 standard deviations and this effect stays persistent for at least one month. This finding suggests that, although the government is able to mitigate online activity in general, these actions come at the price of fostering political debates online. Nonetheless, the effect on political content is statistically insignificant, therefore more research with a more precise political content definition is needed.

The potential mechanism for this self-censorship effect can be explained by the model of Becker, 1986. In the frame of online extremism, this model suggests that users bear the punishment costs of posting sensitive content online and the probability of being punished for this content. When learning about new cases, users can reevaluate the probability of being punished and decide to post less content to decrease this probability. Moreover, by making the potential punishment more severe, the government can cut the costs of monitoring the content online, achieving effective censorship without resorting to traditional costly methods of restricting access to the internet. Therefore, the findings of this paper provide ground for further research on censorship mechanisms, highlighting the trade off between the severity of punishments, costs of monitoring, and costs of fostering political discussions.

7 References

A. Fedasiuk, R. (2021). Buying Silence: The Price of Internet Censorship in China. China Brief, 21(1).

B. Chen, Y., & Yang, D. Y. (2019). The impact of media censorship: 1984 or brave new world?. American Economic Review, 109(6), 2294-2332.

C. Enikolopov, R., Petrova, M.,& Zhuravskaya, E. (2011). Media and political persuasion: Evidence from Russia. American Economic Review, 101(7), 3253-85.

D. Enikolopov, R., Makarin, A., & Petrova, M. (2020). Social media and protest participation: Evidence from Russia. Econometrica, 88(4), 1479-1514.

E. Acemoglu, D., Hassan, T. A., & Tahoun, A. (2018). The power of the street: Evidence from Egypt's Arab Spring. The Review of Financial Studies, 31(1), 1-42.

F. https://www.bloomberg.com/news/articles/2020-08-28/belarusian-officials-shut-down-internet-with-technology-made-by-u-s-firm

G. https://www.washingtonpost.com/politics/2019/11/27/iran-shut-down-internet-stop-protests-how-long/

H. Becker, G. S. (1968). Crime and punishment: An economic approach. In The economic dimensions of crime (pp. 13-68). Palgrave Macmillan, London.

I. Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. Annual Review of Economics, 12, 415-438.

J. https://www.vedomosti.ru/politics/articles/2018/10/03/782645-putin-o-dekriminalizatsii

K. https://meduza.io/news/2018/07/13/v-habarovskom-krae-na-vracha-zaveli-ugolovnoe-delo-za-klass-v-odnoklassnikah

L. https://meduza.io/news/2018/08/20/zhitelya-peterburga-obvinili-v-ekstremizme-za-reposty-anekdota-pro-vybory-i-karikatury-na-vatnikov

M. https://www.sova-center.ru/en/misuse/reports-analyses/2020/04/d42333/#_Toc38274133

N. https://meduza.io/feature/2018/10/16/politsiya-po-vsey-rossii-pokupaet-sistemy-monitoringa-sotssetey-oni-pomogayut-iskat-ekstremizm-ne-vyhodya-iz-rabochego-kabineta

O. https://meduza.io/feature/2018/07/30/dvuh-zhiteley-barnaula-obvinyayut-v-ekstremizme-iz-za-memov-vo-vkontakte-zayavleniya-na-nih-podali-odni-i-te-zhe-studentki-yurfaka

P. Borusyak, K., & Jaravel, X. (2017). Revisiting event study designs. Available at SSRN 2826228.

Q. https://www.theguardian.com/world/2018/aug/07/china-bans-winnie-the-pooh-film-to-stop-comparisons-to-president-xi

Appendix

A.1: List of words for flagging political content

Путин, Навальн, государств, власт, президент, правительств , госдум, митинг, протест, страна, страны, страну, стране , страной, Росси, свобод , аквадискотек, дворец, дворца, дворце, царь, царя, царю , отрав, трус, гульфик, полиц, ФСБ.
Approximate English translation: Putin, Navalny, government, power, president, administration, State Duma, rally, protest, country, Russia, freedom, aquadiscoteque[7], palace, tsar, poison, underwear, codpiece, police, FSB.

A.2: List of criminal articles used

(1) **280.2**: Public calls to extremist activities done with the use of the mass media or information and telecommunication networks, including the Internet

(2) **280.1.2**: Public calls for the implementation of actions aimed at violating the territorial integrity of the Russian Federation committed using the mass media or electronic or information and telecommunication networks (including the Internet)

[7]the list contains words related to the most famous political events in Russia in 2020-2021: investigation about the unregistered Putin's palace and the poisoning of Navalny

(3) **282.1**: Actions aimed at inciting hatred or enmity, as well as humiliating the dignity of a person or a group of persons on the basis of gender, race, nationality, language, origin, attitude to religion, as well as belonging to any social group, committed in public or using means mass media or information and telecommunication networks, including the Internet

(4) **282.2**: Same actions done:

   (a) with the use of violence or with the threat of its use

   (b) by a person using his official affiliation

   (c) by an organized group