Fighting Abuse while Promoting Free Speech: Policies to Reduce Opinion Manipulation in Online Platforms

Dahae Jeong
Arizona State Univ.
dahae.jeong@asu.edu

Sang-Pil Han
Arizona State Univ.
shan73@asu.edu

Sungho Park
Arizona State Univ.
spark104@asu.edu

Seok Kee Lee
Hansung Univ.
seelee@hansung.ac.kr

Abstract

With the rise of misinformation epidemic, this study aims to empirically investigate the consequences of an online commenting platform's activity-capping policy on abusers' and regular users' activities. Utilizing a quasi-experimental setting, we find that restrictive policies not only curtail the activity of the abusers but also promote the activity of regular users. Results show that the policy has an asymmetric effect on abusers and regular users— while it effectively reduces the actions of the malicious users by 1.8%, it promotes the activities of the regular users by 2.2%. To better understand the behavioral change of the regular users, we draw from the rational economic perspective of voting decisions and provide initial evidence that such policy measures reinforce the subjective probability of being influential on the outcome. This study will provide valuable implications to managers and policymakers to estimate the consequences of and to combat against malicious behaviors and to promote free speech in online platforms.

1. Introduction

With the emergence of online communicating platforms, people share and gather information from the web. While it has allowed the social actors to directly exchange their opinions at a little to no cost [20], at the same time, it has also allowed the malicious actors to spread false information quickly to sway the public's views on specific topics [23]. False information and politically-biased misinformation on online platforms have engendered unprecedented economic and political problems throughout the world [12].

As a mean to control the quality of the contents and to dilute the spread of misinformation, many online platforms have introduced a way to aggregate opinion – with upvoting and downvoting on posts and comments [22]. For example, Reddit offers its users to upvote/downvote on posts and comments, and its relative votes determine the visibility of the posts. Such a system improves "digital democracy" [22], but at the same time, it provides opportunities for malicious users to disseminate information without much friction by utilizing upvote/downvote bots and sockpuppets. As Muchnik et al. (2013) found, prior rating on comments/posts significantly affects how people perceive and rate [18]. Thus, attempts to manipulate public opinion using upvotes/downvotes have become a severe issue to online platforms and their users.

One straightforward solution for the online platforms to fight the misinformation by manipulating the popularity of postings is to place restrictions on how frequently, easily, and quickly a user influences the popularity of the information. Some platforms have already adopted such restrictive policy measures that do not discriminate between abuser and regular users. For instance, WhatsApp introduced a message-forwarding cap, that users can only share a message up to five times, to curb the overly fast distribution of misinformation [8]. Recent scholarly work demonstrates that setting a limit on political URLs shared in Twitter effectively reduces the fake news spread [7].

Although an activity cap may help online platforms regain social accountability, it may affect the activity and traffic levels accrued by users and thus the profit of the platforms. As the capping policy symmetrically affects both abusers and regular users, it is plausible that the capping policy not only reduces the malicious actions but also reduces the activity levels of regular users. As the activity and traffic levels accrued by users are one of the essential metrics for the platform's profit, it is vital to understand the full ramifications of policy interventions by examining its impact on the malicious users and the regular users. That is, we aim to investigate the efficacy of such policy measures in restraining malicious activities as well as in suppressing, or even bolstering, the regular users' activities.

While scholarly work in online opinion manipulation is growing, current literature focuses on the spread of misinformation and discusses ways to

URI: https://hdl.handle.net/10125/64228
978-0-9981331-3-3

restrain malicious activities. Only a handful of studies show the platform's policy on regular users. For example, Ma and Agarwal (2007) found that identity verification, a form of platform's abuser-restraining measure, increases users' satisfaction and knowledge sharing [16]. However, their work is based on the survey data and did not delve into its potential impact on regular users. Also, Grinberg et al. (2019) showed the effectiveness of setting a limit on political URLs in Twitter on fake news spread, but it was based on a simulation [7]. To our best knowledge, no study to date has empirically examined the consequences of an online commenting platform's capping policy on regular users' activities. This study aims to fill that void by attributing the efficacy of the online platform's policy measure to meaningful behavioral changes of regular users.

We empirically investigate the consequences of a deterrence policy on regular users by utilizing a natural quasi-experiment setting. To this end, we collaborated with one of the largest online platforms in East Asia. The platform implemented a deterrence policy during our sample period by restricting the number of comments and the count of upvotes and downvotes a user can cast. We exploit this institutional variation to make inferences about the effects of the deterrence policy on both abusers and regular users, utilizing individual-level log data before and after the policy implementation.

Results show some interesting findings. First, policy implementation positively affects the overall activities of the voting activities. Second, the policy has an asymmetric effect on abusive users and regular users – while it effectively reduces the actions of the malicious users by 1.8%, it promotes the activities of the general users by 2.2%. Third, the policy increases the activities of the regular users, regardless of their previous voting activities. While the policy shows stronger effect on increasing the users with previous voting activities, it also effectively attracts users without any voting activities to engage in the voting behavior.

Further, to better understand underlying mechanism driving regular users' behavioral changes, we draw upon the rational economic perspective of voting decisions by Riker and Ordeshook (1968) [19], and posit that the utility of expressing their liking or disliking on a given opinion (i.e., comment) as a function of the subjective probability of being able to influence the popularity of that opinion by up-voting or down-voting it, respectively. Based on the fact that a preemptive measure against malicious actors employed by a platform is likely reinforce, if not strengthen, the belief that the voice of regular users will be heard and represented accurately due to the expected decrease of abusive behaviors in the platform. We support our hypothesis by finding some initial evidence that the capping policy reinforces the subjective probability of being influential on the outcome in the upvote actions.

Our results contribute to the extant literature of fake information and online opinion manipulation and our understanding of policy measures to restrain the manipulative activities. Reports from the social network platforms and popular press have documented the potential consequences of such a restrictive policy on the dissemination of the fake information. Our study focuses on another yet important outcome of such a restrictive policy on the regular users' activity levels. Our results indicate that the platform's simple input capping policy suppresses only the activities of the abusers and instead increases the activities of the regular users by strengthening the belief that the voice of regular users will be heard and represented accurately due to the expected decrease of abusive behaviors in the platform. Further, our research provides meaningful managerial insights into the online platforms.

2. Opinion manipulation in online platform

Online manipulation has received significant attention from researchers since the emergence of online communicating platforms. With the proliferation of the online platforms which enable sharing of the user-generated contents, concerns on malicious action and its impact on the people's behavior have brought scholarly attention in the field. The extant literature on online manipulation focuses on product reviews by investigating notable characteristics of fake reviews [1], detecting fake reviews [3, 9, 11], and examining the consequences of the malicious review activities and devising potential solutions to suppress deceptive reviews [14 ,17]. For example, Anderson and Simester (2014) find that fake reviews are more negative and contain less explanation about the products [1]. Fake review detection research focuses on the formulation of detection algorithms, leveraging reviewer information, review quality and product attributes [9], reviewer characteristics and interactions among them [11], and linguistic cues in the review contents [3]. Lastly, a few studies have documented potential solution to combat deceptive reviews. Mayzlin et al. (2014) found that significantly less malicious reviews exist under the platform with a verified reviewer feature [17]. Also, Lappas et al. (2016) investigates the impact of malicious reviews on the ranking of the business shown in the platform and provides potential response strategies for the attacked business owners [14].

Another stream of literature on online manipulation focuses on the spread of misinformation in the social media and commenting platforms and its consequences. A rich set of research focus on the malicious actions of social and political bots in social media [5, 10], the

impact of the manipulative information [21, 23], and how people react to the mal-information [2, 18]. Kollanyi et al. (2016) analyze tweets generated during the 2016 US presidential election and find that automated activities accounts reached up to 27% at its peak [10]. Forelle et al. (2014) investigated activities of political bots in Venezuela and finds that only a small number of bots generate a high volume of retweets [5]. Vosoughi et al. (2018) examine the spread of false news and find that news with misinformation spreads further and faster than the true news [23]. Shao et al. (2018) find that automated bots are the leading cause of false information spread on Twitter [21]. And lastly, some papers look into the consequences of manipulated information and automated activities. In their 2013 paper, Muchnik et al. find that positively biased opinions generate positive herding effect while negatively biased opinions get people to correct the bias [18]. Also, Badawy et al. (2018) find evidence that users are vulnerable to malicious activities. People helped to share tweets from Russian trolls during the 2016 election [2].

While the negative consequences of online manipulation have been well noted by previous literature, little research has addressed the strategies to mitigate this negative impact. As noted previously, a few studies discuss the effectiveness of user verification on malicious activities [17] and deterrence policies [7]. However, the literature focuses on the impact of such policy measures on abuser activities and misinformation spread. Only a handful of studies examined the consequences of deterring mechanisms on regular users' behavior. Ma and Agarwal (2007) find that identity verification of the content generator increases users' satisfaction and knowledge sharing [16]. However, their work is based on the survey data from the users from the online communities. To our knowledge, no study to date has empirically examined the consequences of such policies on regular users. As more platforms have adopted strategies to combat malicious actions, practitioners need to understand the impact of these strategies on non-malicious users as the volume generated by the users is directly linked to the firm profit. Thus, more scholarly attention to this topic is needed.

3. Conceptual framework

In this study, we aim to empirically investigate the effect of the online commenting platform's deterring mechanism on regular user activities. Expressly, our focal policy targets to limit the volume and the frequency of activities from malicious accounts by imposing a restriction of the total number of activities per account and adds ten-second intermittent pauses or delays between activities (e.g., commenting, up-voting, down-voting). The pause implemented would effectively reduce the activities of the automated bots, and the restriction on the total number of activities would force attackers to put extra effort and time to make more accounts if they want to engage in the malicious activities of high volume. Hence, our focal policy is expected to ramp up the friction to malicious users thus deterring their activities, but at the same time it can give rise to either the same level of suppression of activity restriction towards regular users or potential encouragement to regular users to more engage in the platform for the increased authority provided under the manipulation deterrence initiative. Thus the multifaceted role of the coercive focal policy remains still unanswered.

3.1. The calculus of voting

According to the voting decision model developed by Riker and Ordeshook (1968), an individual's decision to vote is determined by the following trade-off involving four factors:

$$P \times B + D > C$$

where P stands for the subjective probability of casting a vote that is pivotal to the election outcome, B means perceived benefit one would get if the supporting candidate wins, D as the personal satisfaction a voter would gain from participating in an election, and C representing the cost and effort associated with voting. In the context of random utility maximization framework, an individual's decision on whether or not to engage in voting depends on the expected benefit the individual would get from the action and the cost associated with it. Accordingly, a voter would decide to vote only if the expected utility (probability x benefit + satisfaction from participation) is higher than the cost associated with voting [19].

We apply this theoretical framework to our research context. Similar to the election turnout, users in our focal platform determine whether or not to participate in the voting activity based on the four factors above. As our goal of this research is to see the impact of the platform policy on the user activities, we assume that B and D are fixed – the perceived benefit of promoting/demoting a comment and the benefit from participating in the commenting system would not change before and after the platform policy implementation. We assume that the benefits one would get from promoting/demoting a comment roots from the comment itself, such as contents and quality of the comment. Similarly, satisfaction from participation depends on the personal characteristics of the regular user. As we have little to believe that the policy change is affecting these root causes of both factors, we focus

primarily on the remaining two factors – the perceived probability of casting a decisive vote and the cost associated with voting to develop our hypothesis. We first examine the cost posited from the policy intervention and then look at the change in the subjective probability of casting a decisive vote.

3.2. Cost of the policy intervention

As described before, our focal policy puts a restriction on the user activity by limiting the frequency and volume of the activities. Restricting the volume of the activities may cause an increase in the opportunity cost of voting, as it puts votes to become scarce [15]. Also, a ten-second pause increases regular users' time cost as they have to spend extra time to participate in the voting activities. Thus, we first hypothesize that a platform's policy to control for the input would increase the cost of the regular users voting activities.

3.3. Perceived probability of being decisive

The policy may also change the perceived probability of being decisive based on the expected effectiveness of the policy. Although the extant literature on election turnout assumes P as ignorable [6, 13], we focus on the perceived probability of being pivotal for two reasons: first, the online platform displays the number of upvotes and downvotes in real-time, so that users can predict the importance of their vote on the outcome. Also, the total number of voters in our commenting platform is smaller than the number of voters in political elections, and therefore, we cannot say that the probability of being decisive is ignorable.

We hypothesize that the policy intervention would increase the regular users' expectation on the commenting system efficacy and thus would enhance the perceived probability of a vote being decisive. The policy against malicious users will strengthen the belief that the voice of regular users will be heard and represented accurately due to the expected decrease of abusive behaviors in the platform. For this reason, regular users modify their beliefs on the commenting system efficacy and thus alter P. If an individual expects that the policy would effectively reduce the malicious action, it will increase the perceived probability of achieving the goal as the total number of votes on a comment would decrease. Hence, we presume that policy intervention would effectively increase the regular users' perceived probability of their vote being pivotal. Also, we note that the change in belief would be heterogeneous across users based on their current beliefs, past engagement in voting activities, and other individual characteristics.

In sum, we expect that a restrictive policy measure increases the cost of voting, but also boost the perceived probability of being decisive. The contrary impact of the policy on two factors of the voting decision model makes the regular users' behavioral change caused by the policy intervention an empirical question.

4. Research context and data

4.1. Research context

This paper utilizes the data provided by one of the largest online platforms in East Asia. The platform is similar to Yahoo! Inc. as it offers various online services including search engine, news, email, entertainment, finance, shopping, blogs, and online forums. In its news platform, users can react with comment posting, and upvote or downvote on comments other users had posted after reading an article. The platform displays comments based on the number of netvotes (number of upvotes – number of downvotes). Comments with the top five highest netvotes are displayed right below the news article with a mobile device, and the comments with the top ten highest netvotes are displayed with a PC. A screenshot of the commenting system is shown in Figure 1.

In the focal platform, the most prominent and critical issue is the political manipulation by clicking "upvote" or "downvote" buttons on the comments. An ongoing investigation of the political scandal in Korea revealed that there had been automated attacks on clicking upvotes and downvotes in favor of a particular political camp. Thus, to reduce the manipulative actions of clicking upvotes and downvotes, the platform implemented an input control policy on April 25, 2018, at 11 AM. This policy has two components. First, it limits the total number of upvotes and downvotes a user account can perform to 50 per day, and the total number of comments a user account can write to 3 per article. Second, it poses a 10-second pause in between clicking upvotes and downvotes. As the policy aims to reduce the political manipulation activities, it is only valid for political news articles.

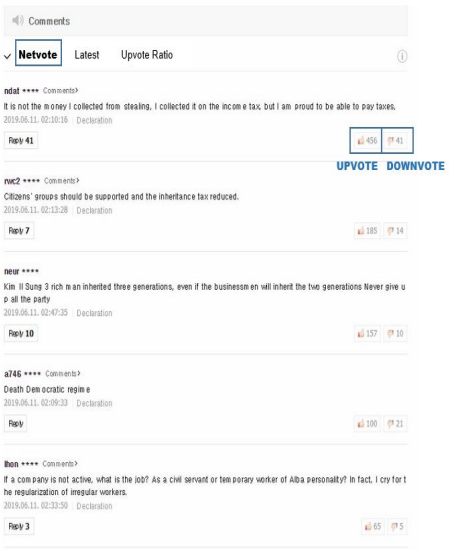

Figure 1. A Screenshot of Comments Displayed in the Platform

The platform had provided multiple announcements on their policy change. The platform posted an announcement about this deterrence policy at the top of the commenting system, as shown in Figure 2A. Also, if a user tried to vote multiple times with the pause in between votes less than 10 seconds or vote over 50 times, a popup notice about the deterrence policy is shown (Figure 2B). Thus, through these announcements, we assume that users were aware of the policy implementation.

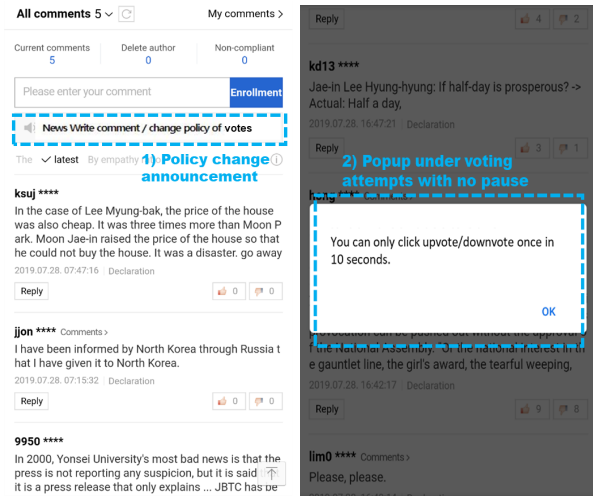

Figure 2. Announcements on capping policy

4.2. Data

For this study, we focus on five political news articles and five non-political news articles that were posted in the platform's online news portal websites on April 25, 2018. All the articles were posted before the implementation of the platform's policy. The focal articles were selected based on the popularity – the top 5 most viewed political articles and top 5 most viewed entertainment articles. We create the following two datasets: The first dataset contains the information about the comments recorded on the focal articles. After reading a news article, users can react with comment posting, and upvote or downvote on comments other users had posted. A user can upvote or downvote only once for each comment. The comment data records fields such as pseudo-id of the comment writer, time the comment was written, the content of the comment, and information about people who responded to the comment including their pseudo-ids, timestamps and the types of the activities (upvoting or downvoting). 19,541 comments and 308,767 upvote/downvote clicks were recorded for our ten focal articles. A more detailed description of the data is presented in Table 1.

Table 1. Descriptive Statistics of Focal Articles

	Posting time	Comments	Total Votes
Political 1	4/25/18 9:00	7,795	127,008
Political 2	4/25/18 10:15	476	7,573
Political 3	4/25/18 8:50	2,178	33,811
Political 4	4/25/18 9:46	793	15,939
Political 5	4/25/18 8:47	3,401	44,786
Entertainment 1	4/25/18 10:15	1,549	31,982
Entertainment 2	4/25/18 8:34	1,538	11,234
Entertainment 3	4/25/18 9:05	1,441	30,264
Entertainment 4	4/25/18 9:19	267	5,743
Entertainment 5	4/25/18 9:40	103	427

The second dataset contains the complete server log files for users who visited the first political article or the first entertainment article and spans six hours from 8 AM to 2 PM on April 25, 2018. This dataset records every activity and request the client (user) makes within site, including timestamps, URLs the user visited, referring URLs, types of requests, and browser and device used. The data contain about 53 million clickstream records of 680,968 users.

One interesting aspect of our data is that it enables us to distinguish abuser accounts from regular users. The platform identified the accounts that were used for manipulation activities based on its own detecting mechanism. It flagged user IDs that had clicked upvote/downvote more than five times in five minutes on a comment made under the same IP and cookie information as malicious accounts. Although these malicious accounts may not be the full set of the total abuser accounts, we argue that the detecting mechanism reasonably locates the accounts that had been involved in automated actions. Thus, we use the abuser accounts flagged by the platform to estimate the effect of the policy in this paper.

Using the first dataset, we construct an individual-minute level panel data set that contains individual voting activities on our ten focal articles. Also, for the individuals who had visited the first political article and the first entertainment article (Political 1 and Entertainment 1), we utilize the second dataset to create individual-level voting activities before entering the focal articles.

As mentioned before, the setting of the policy implementation brings a quasi-experimental design and allows us to causally inference the effect of policy implementation on the user activities. To estimate the impact of platform's input control policy on the abuser and user activities, we use difference-in-difference estimation technique [4, 25]. In order to estimate the effect of an intervention using the difference-in-difference technique, we need observations from a unit (in this case, individual) in both pre- and post-policy periods. Therefore, we select users who visited both pre- and post-policy. We also limit the time bandwidth to 30 minutes before and after the policy implementation to screen out the unobservables that could affect the voting behavior. We exclude individuals who visited both the focal political and entertainment articles from our sample, as they may have been exposed to the platform's policy and thus their behavior afterward may be affected by the policy even in the entertainment articles.

In sum, our final sample includes 63,697 individual-minute level observations from 5,781 individuals. Among the individuals in the sample, about 5.2% of the accounts were identified as abusers. Summary statistics of the data is presented in Table 2.

Table 2. Summary Statistics

Panel A – Number of Individuals in Each Group		
	Control	Treatment
User	1,973	3,506
Abuser	26	276

Panel B - Summary Statistics				
	Mean	S. Dev	Min	Max
Ind(Treatment)	0.6377	0.4807	0	1
Ind(Post-Policy)	0.5123	0.4999	0	1
Ind(Abuser)	0.0779	0.2680	0	1
Log(activity+1)	0.1065	0.4160	0	3.9512
# Obs	63,697			

5. Empirical analysis and results

5.1. Difference-in-difference-in-difference

5.1.1. Difference-in-difference-in-difference. We begin our empirical analysis by aiming to examine whether the impact of the policy varies on abusers and regular users. In order to estimate the effect, we use the difference-in-difference-in-difference (DDD) specification. The DDD estimate of policy intervention is estimated by:

(1) $Y_{it} = \alpha_i + \lambda_t + \delta_1 Pol_t \times T_i + \delta_2 Pol_t \times T_i \times Ab_i + \theta X_{it} + \varepsilon_{it}$,

where α_i is an individual fixed effect and λ_t is a time fixed effect in minute level. Same as our previous equation, our dependent variable Y_{it} is log(number of total votes+1) an individual i performed at minute t. Pol_t is a post-policy indicator. It is equal to 1 if t is greater than the policy implementation. T_i is an indicator for the individuals in the treatment group - people who had visited political articles. X_{it} controls for the focal articles each individual i was active on at minute t. Ab_i is an abuser indicator - 1 if an individual i is marked as an abuser by the platform. From the above specification, our difference-in-difference coefficient δ_1 captures the effect of the platform's policy on user activities in the focal ten articles while $\delta_1 + \delta_2$ captures the effect of the platform's policy on abuser activities in the focal ten articles.

One potential issue with the above specification in equation (1) is that it does not account for the difference in duration of the focal article visits each individual has. Users visit the focal article at a different time and the length of their stay in the focal article is also different. If there exist time trends based on the duration of the users, the previous model would not capture it. To handle this issue, we include individual-specific linear time trends:

(2) $Y_{it} = \alpha_i + \lambda_t + \delta_1 Pol_t \times T_i + \delta_2 Pol_t \times T_i \times Ab_i + \beta X_{it} + \theta_i \tau + \varepsilon_{it}$

where τ equals 1 if an individual i entered into the focal article and linearly increases by each minute. Here, θ_i controls the linear individual time trend. In other

words, we allow each individual to have different linear time trend after controlling for the individual and time fixed effects.

Results from equation (1) are presented in Table 3 column (1) and the results from equation (2) are presented in Table 3 column (2). As can be seen from both specifications, the policy implementation is significantly and positively correlated with the total number of voting for the regular users but negatively correlated with the voting activities of the abusers. The results under equation (2) suggest that the policy increases the regular user voting activities by 2.2% while decreasing the voting activities of the abusers by 1.8%. This result implies that 1) the capping policy effectively restricts the behavior of the abusers, 2) expected utility under the policy outweighs the cost imposed by the policy for the regular users. As explained in our theoretical framework, it may occur from the increase in the perceived probability that user's voting activities would better represent their opinion in the commenting system. This change in the perceived probability mainly comes from the expected decline in the abusive actions under the policy.

Table 3. Results from DDD models

Dependent Variable	(1) Ln(vote+1)	(2) Ln(vote+1)
$Pol_t \times T_i$	0.0156 *** (0.0059)	0.0219 ** (0.0099)
$Pol_t \times T_i \times Ab_i$	-.0590 *** (0.0111)	-0.0405 ** (0.0186)
Controls	YES	YES
Observations	63,697	63,697
R-Squared	0.4939	0.5803
Individual FE	YES	YES
Minute FE	YES	YES
Individual Trend	NO	LINEAR

Note: Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

5.1.2. Heterogeneous effect by previous voting activities. Our results indicate that the change in expected utility under the policy outweighs the cost imposed by the policy for the regular users. It may occur from the change in the perceived probability that voting activities would better promote their opinion without the abusers. We examine further by dividing the regular users into two, one with previous voting activities and the other without previous voting activities. We argue that users who have been engaged in the voting activities have a better understanding of the magnitude of the malicious actions and have stronger beliefs on the accountability of the commenting system. Thus, we expect that the restrictive policy will reinforce the activities of regular users with previous voting activities.

To test our hypothesis, we create an indicator variable based on the voting activities before they visit our focal articles. As we only have information about the previous activities for the individuals who had visited the first political article and the first entertainment article (Political 1 and Entertainment 1), we limit our sample to these individuals. Also, as our goal for the analysis is to see the heterogeneous policy effect among the regular users, we exclude the individuals marked as abusers. Thus, our new sample for the analysis contains 34,108 individual-minute level observations from 3,689 regular users.

We first estimate the effect of policy on the users separately based on the previous activities using difference-in-difference:

(3) $Y_{it} = \alpha_i + \lambda_t + \delta_1 Pol_t \times T_i + \varepsilon_{it}$,

where α_i is an individual fixed effect and λ_t is a time fixed effect in minute level. Again, our dependent variable Y_{it} is log (number of total votes+1) an individual i performed at minute t. Pol_t is a post-policy indicator. It is equal to 1 if t is greater than the policy implementation. T_i is an indicator for the individuals in the treatment group. We divide our sample into two, one containing observations from the users without any previous activities and the other with observations from users with previous activities, and estimate using equation (3).

We also estimate the heterogeneous effect by users' previous activities using difference-in-difference-in-difference estimation similar to equation (1):

(4) $Y_{it} = \alpha_i + \lambda_t + \delta_1 Pol_t \times T_i + \delta_2 Pol_t \times T_i \times Active_i + \varepsilon_{it}$,

where α_i is an individual fixed effect and λ_t is a time fixed effect in minute level. Same as our previous equation, our dependent variable Y_{it} is log (number of total votes+1) an individual i performed at minute t. Pol_t is a post-policy indicator. It is equal to 1 if t is greater than the policy implementation. T_i is an indicator for the individuals in the treatment group, that is, people that had visited political articles. $Active_i$_i indicates for the users' previous activities, where the indicator equals 1 if an individual i had been engaged in voting behavior before he/she visited our focal articles and 0 if an individual i had no previous activities.

Table 4 shows the result from equation (3) and (4). Results on Table 4 column (1) and (2) show that the policy intervention increases voting activities of the overall users, regardless of their previous voting behavior. Results on Table 4 column (3) directly compare the impact of the policy intervention on users with previous voting experience and without previous voting experience. We can see that users who had been

engaged in voting activities before the focal article show greater increase in the number of voting after the treatment, and thus the results support our hypothesis that more attentive users are more likely to have a greater perception change in probability of voting a decisive vote.

Table 4. Results of heterogeneous policy effects

Sample	(1) W/O activity	(2) With activity	(3) All sample
$Pol_t \times T_i$	0.0095 ** (0.0047)	0.0645 *** (0.0247)	0.0081 (0.0063)
$Pol_t \times T_i \times Active_i$	-	-	0.0578 *** (0.0099)
Controls	YES	YES	YES
Obs.	27,828	6,280	34,108
R-Squared	0.4106	0.4227	0.4451
Individual FE	YES	YES	YES
Minute FE	YES	YES	YES

Note: Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

5.2. Random Utility Model

Results from DDD models show strong evidence that policy intervention induces regular users' behavioral changes. However, it has two limitations: the sample used in the DDD analysis is limited as we only include limited bandwidth of the time and individuals who had been engaged in voting activities during pre- and post-treatment periods. Also, it is still uncertain of what drove the behavioral changes of regular users. To overcome this issue, we propose a random utility model based on our conceptual framework described in Section 3.

5.2.1. Revisiting calculus of voting. To empirically examine the impact of restrictive policy on the regular users' voting decisions, we revisit the rational voting model mentioned in the previous section:

(5) $U_{ic} = P_{ic} \times B_{ic} + D_i - C_{ic}$,

where *i* indicates individual and *c* indicates comment. Again, an individual's utility for voting a comment comes from the probability of one's vote being decisive P_{ic}, benefit from the comment being promoted/demoted B_{ic}, satisfaction from voting D_i and the cost of voting C_{ic}. Unlike traditional elections, users in our context can vote only once per comment (upvote or downvote) and can vote across multiple comments. Thus, we assume that an individual has a different level of utility per comment.

In our model, *P* is defined as a function of closeness, following the operationalization of Riker and Ordeshook (1968). We develop a closeness measure as the difference in netvotes between a comment and a comment ranked next [19]. As the platform displays comments based on the number of netvotes, this measure reflects the closeness of winning (moving to the upper ranking). When a user clicks a news article, the comments and the number of netvotes displayed below the article do not change unless the user refreshes the webpage. Thus, this variable varies by the individual user's time of arrival to the focal article.

We also include a policy indicator in our model. We code 1 if an individual *i* engaged in a voting activity on comment *c* under the policy intervention and 0 otherwise. And we include an interaction term of policy indicator and closeness to empirically examine the impact of the capping policy on the perceived probability of being influential on the outcome.

5.2.2. Random utility model. Drawing upon a random utility model, we rewrite the previous equation as:

(6) $U_{ic} = \beta X_i + \alpha Z_c + \delta T_{ic} + \varepsilon_{ic}$,

where U_{ic} is the i[th] individual's expected utility of voting on comment *c*, X_i represents individual-level variables, Z_c represents comment-level variables, and T_{ic} represents the variable of interest: closeness and policy indicator.

Our observed outcome, y_{ic}, represents the voting activity of an individual *i* on comment *c* and takes either one or zero. If a user had voted to the comment, y_{ic} takes a value of 1, and 0 of one had not voted to the comment. Assuming a latent regression determines the observed outcome variable, we have a probability model of voting:

(7) $\Pr(vote = 1|X_i, Z_c, T_{ic}) = \Pr(\varepsilon_{ic} > -(\beta X_i + \alpha Z_c + \delta T_{ic}))$

Using a simple linear probability model, we estimate the above equation and provide initial evidence for the validity of our conceptual model. We employ comment-level fixed effects in terms of Z_c and individual-level fixed effects and individual's arrival time dummies (in 30 minute interval) for X_i. In our context, users can either upvote or downvote a comment but both, so we estimate the propensity of up-voting and down-voting, respectively. Further, we argue that ranking of the comment has an inevitable effect on user's decision to vote, due to the platform's design factor. Therefore, we also estimate the above equations separately by the ranking of each comment. We utilize the netvote distance from the *i*[th] ranked comment to the *i*-1[th] ranked comment as the closeness in the upvote estimation, and the netvote distance from the *i*[th] ranked comment to the *i*+1[th] ranked comment as the closeness in the downvote estimation. In the case of first-ranked comment, we

utilize the netvote distance to the second-ranked comment for both upvote and downvote analysis. We also include inverse ranking as a control variable in the full-sample analysis, and user's previous voting counts and their device as controls in the analysis by ranking.

To estimate the equation (7), we create a sample based on our second dataset. We include all the regular users who have visited political 1 article. As explained before, users with PC see 10 comments below the article and users with mobile devices see 5 comments below the article. Therefore, we include 5 individual-comment level observations for the mobile device users and 10 individual-comment level observations for the PC users. The sample includes 2,282,672 observations from 438,889 users.

Table 5. Results from RUM

Panel A : DV=Upvote			
	Closeness	Policy	Closeness× Policy
All	-.0000003 (0.000000)	0.59701*** (0.004154)	
All	.00000008 (0.0000004)	0.59718*** (0.00415)	-.0000007* (0.0000004)
Rank 1	-.000002 (0.0000049)	0.77219*** (0.03407)	-0.000074 (0.0000661)
Rank 2	.0000051* (0.0000026)	0.62057*** (0.01368)	-.000128*** (0.000018)
Rank 3	.000005** (0.0000024)	0.75085*** (0.01452)	-.000227*** (0.00002)
Rank 4	-.000022*** (0.000007)	0.69746*** (0.01874)	-.000258*** (0.000079)
Rank 5	-.0000036 (0.000002)	0.51897*** (0.01415)	.000204*** (0.000032)
Panel B : DV=Downvote			
	Closeness	Policy	Closeness× Policy
All	.0000007*** (0.000000)	0.24573*** (0.00248)	
All	.0000003 (0.0000002)	0.24551*** (0.002483)	.0000006*** (0.0000002)
Rank 1	-.0000052** (0.000002)	0.17130*** (0.01472)	-.000072** (0.000029)
Rank 2	.0000014 (0.0000017)	0.29541*** (0.007994)	-.0000086 (0.000013)
Rank 3	-.0000009 (0.0000014)	0.13142*** (0.00638)	0.00007** (0.000027)
Rank 4	.00000093 (0.000001)	0.14457*** (0.00648)	-.000057*** (0.000014)
Rank 5	-.0000008 (0.0000008)	0.21524*** (0.00653)	0.000011*** (0.000004)

Note: Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 5 reports results from 7 separate analysis based on equation (7). As mentioned before, panel A uses upvote as the dependent variable, and panel B uses downvote as the dependent variable. In upvoting decisions, we can see from column (3) that as netvote distance from the focal comment to the comment ranked above gets smaller, regular users are more likely to vote 'like' after the policy implementation in the most analysis. Also, by comparing the estimates of the main effect and the interaction effect, we support our theoretical model that users are more likely to vote when the distance between the focal comment and the previous comment gets smaller, that is, one's vote becomes more decisive. In the analysis of ranking 5, we find a contrary effect. Although more investigation is needed, we suspect that it is due to the platform's commenting system design, that only the top five comments are seen in the news article page without any clicks.

In the downvote analysis, the results are not easily explainable. Some of the interaction estimates are positive, while the others are negative. There may be several explanations: First, the underlying mechanism of the voting decision may be different in the 'disliking' situation as the traditional voting decision framework we employ focuses on the voting decision in the elections – more similar to 'liking' the candidate than 'disliking'. Second, in this initial study, we do not control for the valence of the comments and the users, and these may have a more significant effect on downvotes than the upvotes.

6. Conclusion and Implications

With the growing attention on online manipulation, many online platforms have imposed or considering to impose the policies to control malicious behavior. This paper plans to empirically examine the efficacy of the online platform's policy measure to meaningful behavioral changes of abusers and regular users, respectively. Our results show that the policy has an asymmetric effect on abusive users and regular users – while it effectively reduces the actions of the abusers by 1.8%, it promotes the activities of the regular users by 2.2%. We also find that the policy increases the activities of the regular users, regardless of their previous voting activities. We also draw from rational voting decision theory and show some initial evidence that the capping policy reinforces the perceived probability of being influential on the outcome in the upvote actions.

Our results add to the current literature of fake information and online opinion manipulation and the effectiveness of the platform's input policies to restrain

the manipulative activities on the regular user. To the extent, reports from the social network platforms and popular press have documented the potential negative consequences of such a restrictive policy on the users' activities. However, our results and findings indicate that the platform's simple input control policy suppresses only the activities of the abusers and instead increases the activities of the regular users. Further, our research applies the voting decision framework to the online user behavior by empirically investigating the voting activities in the online political news platform. Online voting has an interesting feature—real-time display of closeness—and thus this study will provide new insights on the impact of closeness on voting decisions on online platforms where a voter can observe their actual closeness measure at the time of their voting decision.

This research also provides meaningful managerial implications. Our results show that an activity-capping policy may be an alternative to abuser detection and suspension for many online communicating platforms, review sites and news platforms utilizing users votes on the popularity of the contents, Capping policy reduces abusers' influence on public opinion by inducing more participation from the regular users. This study shows that a simple, easily implementable deterrence policy can mitigate abusive actions while promoting user discussion and opinion exchanges on the online news platforms.

7. References

[1] Anderson, E. T., and Simester, D. I. 2014. "Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception," Journal of Marketing Research (51:3), pp. 249-269.

[2] Badawy, A., Ferrara, E., and Lerman, C. 2018. "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain: pp. 258-265.

[3] Banerjee, S., and Chua, A. Y. K. 2014. "A study of manipulative and authentic negative reviews," Proceedings of the International Conference on Ubiquitous Information Management and Communication (8), pp. 1-6.

[4] Bertrand, M., Duflo, E., and Mullainathan, S. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" The Quarterly Journal of Economics, 119(1), 249-275.

[5] Forelle, M., Howard., P., Monroy-Hernandez., A., and Savage, S. 2014. "Political Bots and the Manipulation of Public Opinion in Venezuela," Working paper.

[6] Gelman, A., King, G., and Boscardin, J. 1998. "Estimating the Probability of Events that Have Never Occurred: When Is Your Vote Decisive?" Journal of the American Statistical Association, 93, pp. 1–9.

[7] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. 2019. "Fake News on Twitter During the 2016 U.S. Presidential Election," Science, 363(6425), 374-378.

[8] Hern, A., and Safi,M. (2019, January 21). "WhatsApp puts limit on message forwarding to fight fake news," The Guardian.

[9] Jindal, N., and Liu, B. 2008. "Opinion spam and analysis," Proceedings of the International Conference on Web Search and Data Mining, pp. 219-230.

[10] Kollanyi, B., Howard, P. N., and Woolley, S. C. 2016. "Bots and Automation over Twitter during the First US Presidential Debate," Project on Computational Propaganda.

[11] Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," Journal of Management Information Systems (35:1), pp. 350-380.

[12] Kumar, S., and Shah, N. 2018. "False Information on Web and Social Media: A Survey," Social Media Analytics: Advances and Applications, CRC press.

[13] Lacy, D., and Burden, B. 1999. "The Vote-Stealing and Turnout Effects of Ross Perot in the 1992 U.S. Presidential Election," American Journal of Political Science, 43(1), 233-255. doi:10.2307/2991792

[14] Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry," Information Systems Research (27:4), pp. 940-961.

[15] Lynn, M. 1991. Scarcity effects on value: A quantitative review of the commodity theory literature[Electronic version]. Retrieved from Cornell University, School of Hospitality

[16] Ma, M., and Agarwal, R. 2007. "Through a glass darkly: information technology design, identity verification, and knowledge contribution in online communities," Information Systems Research (18:1), pp. 42-67.

[17] Mayzlin, D., Dover, Y., and Chevalier, J. 2014. "Promotional reviews: An empirical investigation of online review manipulation," American Economic Review (104:8), pp. 2421-2455.

[18] Muchnik, L., Aral, S., and Taylor, S. J. 2013. "Social influence bias: A randomized experiment," Science, 341(6146), 647-651.

[19] Riker, W. H., and Ordeshook, P. 1968. "A Theory of the Calculus of Voting," American Political Science Review (62:25).

[20] Shanahan, M. K. 2017. Journalism, Online Comments, and the Future of Public Discourse, Routledge.

[21] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., and Menczer, F. 2018. "The spread of low-credibility content by social bots," Nature communications, 9(1), 4787.

[22] Stoddard, G. 2015. Popularity and quality in social news aggregators: A study of reddit and hacker news. In WWW.

[23] Vosoughi, S., Roy, D., and Aral, S. 2018. "The Spread of True and False News Online," Sience (MAR 2018), pp. 1146-1151.

[24] Wolfinger R., and Rosenstone, S. 1980. Who Votes? Yale University Press.

[25] Wooldridge, J. (2007). What's New in Econometrics? Lecture 10 Difference-in-Differences Estimation.