



The impact of politicians' behaviors on hate speech spread: hate speech adoption threshold on Twitter in Japan

Taehee Kim¹ · Yuki Ogawa²

Received: 22 September 2023 / Accepted: 29 February 2024 / Published online: 6 April 2024
© The Author(s) 2024

Abstract

In this paper, we hypothesized that a leading politician's behavior signaling support for the ideas of hate groups will increase hate speech by shifting social norms. To examine the hypotheses, in particular, to quantify the social norms shift, the study utilized an adoption threshold measure based on Twitter retweet networks. Our empirical study focused on the hate speech spread effect of an announcement by Yuriko Koike, Governor of Tokyo, declining to participate in a memorial ceremony for Korean massacre victims. The results support the hypothesis: after Koike's announcement, adoption thresholds of hate speech users were significantly lowered, while the number of hate speech users increased. Further, this study compared the effect of Koike's announcement to the effect of a North Korean missile launch, a national security threat to Japan. The average hate speech adoption threshold was lower after Koike's announcement than after the missile launch, suggesting that a leading politician's behavior could have a greater impact on shifting norms of prejudice than even a nationally threatening event.

Keywords Hate speech · Complex contagion · Social media · Social norms · Japan · Prejudice

✉ Taehee Kim
taehee.kim@uni-konstanz.de
<https://scholar.google.com/citations?user=VvE0QbcAAAAJ&hl=en>

Yuki Ogawa
ogawayk@tcu.ac.jp
https://scholar.google.com/citations?hl=en&user=CQiM0KAAAAAJ&view_op=list_works&sortby=pubdate

¹ Department of Politics and Public Administration, University of Konstanz, Universitätsstraße 10, Konstanz 78464, Germany

² Design and Data Science, Tokyo City University, Tamazutsumi 1-28-1, Setagaya-ku, Tokyo 158-8557, Japan

Introduction

In recent years, the expression of hostile prejudice has increased in major online spaces, in particular social media such as Twitter. What is the mechanism behind the quick spread of online hate speech in this short period? It has been suggested that social norms, or shared rules of behavior, are one of the prime factors that affect individuals' prejudiced attitudes and behavior [2, 7], and this has subsequently been confirmed by empirical studies in various settings [9, 31, 39].

Previous studies have reported that “social referents,” who gain comparatively more attention from other individuals, are particularly influential over social norms [35]. A recent series of studies focused on the effect of Donald Trump's discourse on increases in prejudiced expressions and behavior. Trump's political rise and his eventual victory in the 2016 presidential election in the United States increased the number of hate crimes and online xenophobic posts in the United States [11, 30] as well as racially biased attitudes in European countries [13].

The most compelling explanation of this phenomenon is a shift in social norms caused by Trump's behavior, toward further acceptance of prejudiced behavior. Trump made prejudiced and violent expressions toward a range of groups during the election campaign. His victory signaled that a considerable number of individuals followed his ideas, which made him a stronger social referent. Despite the multiple empirical studies devoted to how politicians can impact prejudiced behavior by shifting social norms, the suggested mechanism is still under-investigated. Most studies demonstrate only a part of the mechanism: the connection of politicians' behavior either with social norms [8, 32] or with individuals' behavior [19]. As an exception, using survey experiments, [4] showed how politicians' behavior changes both social norms and their behavior. However, their results are limited to an experimental setting, and their external validity may be questionable. Further, most studies in the area have focused on Donald Trump, while it is still unknown how the mechanism functions in other contexts. In addition, given there exists such an effect, its effect size is not well discussed in previous studies.

Given these research gaps, this study aims to provide evidence of the whole suggested mechanism using real behavior trace data. More specifically, this study demonstrates that a leading politician's behavior signaling support for hate ideas increases hate speech on Twitter by shifting social norms. To quantify the social norm shift, this study utilizes each user's adoption thresholds for expressing hate speech, that is, a minimum number of sources required to express hate publicly. In the context of Twitter, the number of sources would be the number of hate posts published by users who are connected to the focal user on a Twitter network. The idea of the adoption threshold is based on the characteristics of prejudiced behavior, which spreads through complex contagions requiring multiple sources of contact. Prejudiced behavior generally contrasts social norms; thus, it is accompanied by social costs such as disrespect, contempt, and repudiation from others. To adopt these types of behavior, individuals may require independent affirmation or reinforcement from multiple sources [6]. Since the level of required sources would depend on the degree of social norms, the change in the

required sources, that is, the adoption thresholds can be used as a direct measure to infer social norm shift.

To provide empirical evidence, this study focuses on hate speech on Twitter in Japan toward Koreans. In Japan, nativist hate groups targeting Koreans have emerged in the late 2000s [15], and they have actively utilized social media to spread hate speech [40].¹ We show that the growing increase of anti-Korean posts on Twitter was magnified by a statement by a leading politician, Yuriko Koike, which supported the claims of hate groups in Japan.

Further, the study also compares the effect size of Koike's announcement with that of a national security threat, a North Korean missile launch. Incidents concerning national security are another well-known factor eliciting prejudiced attitudes and behavior [5, 10, 17, 27]. This study takes advantage of the timing of these two events, which occurred in the same week. This situation enables us to measure their impacts in the same framework and compare them with each other.

The remainder of this paper proceeds as follows: The next section discusses the related literature on the expression of prejudice, based on which we elaborate a model of hate speech spread and derives a hypothesis. The third section introduces the two target events, Koike's statement and North Korea's missile launch, and the research strategies, including design and data collection. In particular, detection of hate speech posts on Twitter is crucial for this research; therefore, the fourth section begins by describing our method of detecting hate speech in detail. After presenting the analysis results, the last section will conclude this paper with some discussion.

Theory and hypothesis

Hate speech can be defined as speech intended to promote hatred on the basis of who the targets are, for example, based on their religion, ethnicity, national origin, or other identity factors [36, 41]. It has been suggested that social norms are one of the prime factors that alter attitudes and behavior. Accordingly, providing group norms makes individuals' prejudiced attitudes draw closer to the group norms [39], while group norm promotion and social sanctioning reduce prejudiced behavior [31].

Recent studies have also focused on individual leading politicians' behavior as a factor for prejudiced attitudes and behavior [4, 8, 13, 19, 32]. The primary mechanism is the capacity of a leading politician to shift social norms of prejudice. This is mainly because people with high status in a group tend to have more capability to affect group norms than other members do [31, 35]. These expectations align with political communication and public opinion studies, suggesting that elites are influential in affecting how people think about politics and that minorities tend to refrain from expressing their opinion [34, 42].

One of the most well-studied examples of this mechanism is Donald Trump's US presidential election win in 2016, analyzed in multiple survey experiments: [4] demonstrated that Trump's perceived popularity and his victory in the election increased individuals' willingness to engage in public xenophobic behavior and the social

¹ In this study, "Koreans" refer to people who have ethnic roots on the Korean Peninsula.

acceptability of such behavior. Similarly, [8] demonstrated that Trump's aggression increased respondents' acceptability of prejudice toward the targeted groups after the 2016 presidential election.

However, the suggested mechanism has not yet been fully tested by previous studies. Most studies have focused on change in perception level but not in actual behavior. Although [4] showed change in behavior, the result was limited to a certain experimental setting, and we do not know whether it can be generalized to a more realistic setting.

This study aims to provide evidence for the suggested mechanism of hate speech on Twitter, a real social network. In particular, this study shows that a leading politician's behavior signaling support for hate shifts social norms so that acceptance of prejudiced behavior increases and facilitates posting hate speech on Twitter. Regarding spreading behavior on social networks, it is well known that weak ties are more influential than close ties to spread information [14]. However, this weak ties theory was suggested based on the spread of information, such as job information. By contrast, single "weak ties" are presumably insufficient to spread prejudiced behavior because it is a socially undesirable act. It has been suggested that when collective behaviors are "costly, risky, or controversial, the willingness to participate may require independent affirmation or reinforcement from multiple sources" [6, p.703]; therefore, prejudiced behavior is more likely to spread in complex contagions with affirmation from multiple sources. In other words, it is reasonable to assume that the level of needed affirmation for hate speech directly reflects individuals' perceptions about social norms on that matter. Thus, change in the level of needed affirmation for posting hate speech can be interpreted as a shifting social norm concerning prejudice.

To quantify needed affirmation from multiple sources, this study utilizes adoption thresholds, that is, the minimum number of sources required to publish hate speech. If we assume that prominent public figures' gestures supporting hate speech should weaken the corresponding social norms, they should lead to lower adoption thresholds.

Hypothesis: A leading politician's behavior that signals support for hate groups' ideas will increase the number of hate speech by decreasing the adoption thresholds of hate speech users.

In addition to testing the existence of the effect above, this study will also substantiate the effect sizes of interest by comparing the impact of a national security threat. Such a threatening event can also be considered to trigger hate speech toward out-groups perceived as responsible for the event. For example, the terrorist attacks of September 11, 2001, in the United States, triggered prejudiced behavior toward Muslims in Western countries [1, 20], created negative attitudes toward immigrants [17], and increased hate crimes against minorities [25]. Additionally, studies confirm that terror attacks increase segregation [10, 12] and impact attitudes of people in third countries [27]. Apart from terror attacks, diverse types of threat or perceived threat can affect attitudes: an increasing number of immigrants activates threat perception and opposition to immigrants [33]; territorial threat negatively impacts

political tolerance [21]; security threat predicts exclusionist political attitudes toward minority groups [5].

Given these findings, we can expect that a threatening event related to national security may profoundly increase prejudiced behavior by lowering the adoption thresholds. The corresponding effect can provide a reference with which we can evaluate the effect size of our hypothesis above. To make use of this affordance, however, we must investigate a social situation in which both types of triggering factors can be observed at almost the same time.

Research design

This study is primarily interested in the effect of political leaders' behavior on social norms concerning racist hate speech in the online sphere. To compare this effect with the effect of a threatening event, which has been well documented in previous research, we need a case where both incidents take place at almost the same time. Japan during Summer 2017 provides such a case.

While there are numerous online spheres in which hate speech can appear, this study will focus on Twitter, for the following reasons: In Japan, Twitter is one of the most used SNS communication tools, along with Facebook, and hate groups actively utilize Twitter to spread their racist beliefs to the larger public [40].² Another merit of Twitter is that we can observe user relationships in several ways, such as by who follows or retweets whom, which allows us to measure adoption thresholds of racist hate speech.

Hate speech in Japan, Summer 2017

Japan, during the Summer of 2017, experienced at least two events that affected hate speech behavior of Twitter users. The first event was provoked by Yuriko Koike, the Governor of Tokyo. On 24 August, media started to report Koike's announcement that she would not send a memorial address to the annual memorial ceremony for the victims of a massacre in 1923, targeting Korean residents right after the Great Kanto Earthquake. Her decision was unexpected and quite surprising as prior governors of Tokyo had sent memorial addresses to the ceremony every year since 1973, when the ceremony had started, and Koike herself had sent one in a previous year.³

With regard to her decision, Koike only stated that she would refrain from sending a memorial address as she found it appropriate to pay respect to all victims of the

² According to the Japanese national survey conducted by Ministry of Internal Affairs and Communications in 2015, *Line*, *Facebook*, and *Twitter* are the most used SNS tools in Japan. Given the fact that the main function of *Line* is private chat, *Facebook* and *Twitter* are the first and second communication tools for public comment, given that 31% and 35% of respondents answered that they had used Twitter and Facebook in the previous year, respectively. For more details, see <http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc242220.html> (Japanese).

³ "Tokyo Gov. Skips Eulogy to Koreans Massacred after 1923 Quake," *The Mainichi* August 25, 2017. <https://mainichi.jp/english/articles/20170825/p2a/00m/0na/015000c> (accessed Feb 9, 2023).

earthquake, not only Korean victims.⁴ Although Koike did not state any further reasons why she canceled the address, it is highly plausible that hate groups perceived her decision as supporting their ideas. The hate groups had claimed that the number of victims of the disaster was exaggerated or that even the tragedy was fictional. In addition, given the meeting record at the Tokyo Metropolitan Assembly, it seems clear that there were pressures from radical right-wingers⁵ behind her decision. At an assembly meeting in March 2017, five months prior to Koike's announcement, a radical right-wing assembly member, Toshiaki Koga, stated that the number of Korean residents massacred in the disaster advertised by the ceremony organizer was baseless and that there were other opinions regarding the number which were far smaller. He claimed that the governor should reconsider sending a memorial address to the ceremony in future, otherwise the governor would be complicit in "distorting" historical facts. Koike responded that it was metropolitan bureaucrats who had sent out the address in the previous year as a customary procedure and that she would make an appropriate decision about sending the address after checking it carefully.⁶ Five months later, she announced that she would not send a memorial address. On September 1, the day of the memorial ceremony, Koike indeed did not send a memorial address. This time lag between the announcement and the memorial ceremony helps us differentiate the effects of the two events.

The second event was a missile launch by North Korea over Japan on August 29. Although North Korea had launched many missiles before, this one was exceptional in terms of the threat level: It was the first ballistic weapon fired over Japan without any notification, and it fell close to the sea route connecting Japan and the US.⁷ The Japanese prime minister at that time, Shinzo Abe, called this incident an "unprecedented threat" to Japan and called for an emergency meeting of the U.N. Security Council.⁸ This incident also had a strong impact on Japanese residents. The residents living in 12 prefectures received alert messages via diverse channels, for example, cell phone messages, loudspeakers, and television, notifying them of the incident and telling them to seek a safe place.⁹ Major media also reported this incident as a serious threat to Japan. Since the missile launched in the early morning, at about 6 am, this news became the top news of the day. Given these circumstances, it

⁴ "Tokyo Governor Rapped for failing to Send Eulogy to 1923 Korean Massacre Victims," *The Mainichi* September 1, 2017. <https://mainichi.jp/english/articles/20170901/p2a/00m/0na/017000c> (accessed Feb 9, 2023).

⁵ Even though Japan's radical right-wingers and its hate groups are not directly associated, they share a core ideology and penchant for historical revisionism [16], and online hate groups have shown strong support for radical right-wingers [37].

⁶ This conversation is documented in the 2017 record of proceedings, No. 4 (<https://www.gikai.metro.tokyo.jp/record/proceedings/2017-1/04.html>).

⁷ North Korea also launched a missile over Japan without any notification in 1998; later, they announced that it was for satellite launch vehicles. See "North Korea Fires Missile," *BBC News* August 29, 2017.

⁸ "North Korea Fires Missile over Japan in 'Unprecedented Threat'," *BBC News* August 29, 2017. <https://www.bbc.co.uk/news/world-asia-41078187> (accessed Feb 9, 2023).

⁹ "Northern Japan Wakes up to 'J-Alert' as N. Korean Missile Files over Japan," *The Mainichi* August 29, 2017. <https://mainichi.jp/english/articles/20170829/p2a/00m/0na/014000c> (accessed Feb 9, 2023).

is highly plausible that this incident increased negative attitudes toward Koreans in Japan.¹⁰

Measuring Adoption Threshold in Twitter

Here, an important task is to measure our dependent variable: the adoption threshold of hate speech users, that is, the minimum number of sources (i.e., hate user alters) that facilitate one to publish a hate speech post.

To measure a user's adoption threshold at a specific period, we need to know about the user's network beforehand. The study utilizes retweet relationships observed prior to a target event to capture the user's network (hereafter, the base network). In particular, retweet networks for 7 days prior to the event (hereafter, the base period) are collected. This is a directed network:

$$G_{retweet} = (V, E)$$

$$E = \{(i, j) : i \neq j, i \text{ retweet } j \text{ in the base period}\}$$

where V is the set of users in the collected data, and E is the set of edges. We do not differentiate between multiple retweets and one retweet from the same user.

In using retweet history to capture a user's network, the study assumes that a user retweeting another user is also likely to follow or read the other user's posts. This assumption may seem to be restrictive since a user can retweet others' posts via a third user. Nevertheless, we stick to this assumption because the main interest of the study is not the absolute level of the threshold but the difference in the threshold level between days with and without contributing events. Even though the measure may be biased because of over-counting, the study expects the volume of bias to be more or less similar over the overall period and to not affect the temporal comparison.

The adoption threshold is measured by counting a hate speech user's alters who published hate speech prior to the target user in the target period. Suppose there is a user A who published a hate post in the target period. Among A 's alters, who are detected by retweet relations in the base period, two have already published hate posts in the target period. In that case, A 's adoption threshold in this period is 2. We count each hate speech user's adoption threshold when they publish a hate post for the first time in the period; in other words, we count the adoption threshold of each user only once even if they publish multiple hate posts. For a total adoption threshold for the target period, we take an average of all hate users' threshold (hereafter, ADT).

The above measurement can be formalized in the following way. As described earlier, the base network is generated by using retweet relationships during the 7 days before the focusing events: Let V_T be a set of hate users detected in the target period. Let $U = (u \in \{V \cap V_T\}, \leq)$ be an ordered set that contains those who appear

¹⁰ One might wonder why the missile launch would increase hate toward Koreans in general, including South Korean nationals, rather than only North Koreans. This is because haters targeted Koreans as an ethnic group, which include nationals of both Koreas.

in the base network and are detected as hate users in the target period, where ordering is based on the time when their first hate posts are published. ADT is an average of each user's adoption threshold in the period; therefore, ADT is,

$$\frac{1}{|U|} \sum_{u \in U} |\{(u, w) \in E : u \neq w, w \in C\}|,$$

where $C = \{c \in U : c \leq u\}$.¹¹

Data Collection

As described above, we focus on two events, respectively occurring events on August 25 and 29, 2017. To cover these events, the study collected Twitter data from August 14 to September 1; among them, 7 days of retweet history, from August 14 to August 20, will be used to generate the base network.

To obtain a retweet network of racist hate speech, the study first crawled tweets from Twitter's standard search API using 15 keywords.¹² To obtain relevant tweets to the degree possible, we crawled tweets at two days after they were published.¹³

Table 1 shows 15 keywords used to collect all tweets associated with racist harassment against Koreans. Keyword selection is based on [40] who investigated Japanese racism toward Koreans. The 15 keywords can be classified into four categories: *General*, *Harassment*, *Diplomatic issues*, and *Anti-left wing*. *Harassment* refers to words which are directly associated with racist harassment. Other keywords in three categories are chosen to gather all relevant tweets to this discourse.

Two general words used to refer to Korea (*Chosen*) and Koreans who live in Japan (*Zainichi*) were selected as keywords because they are often used with derogatory nuance in the online space.¹⁴ *Anti-left wing* words are used to criticize people who "defend" Korea(ns) (who are basically considered to be left-wingers), and *Diplomatic issues* covers the main actual topics of bilateral discussion between South Korea and Japan. By using these wide-scope keywords, it is likely that most tweets

¹¹ C is a set that contains hate speech users who published prior to a user $u \in U$.

¹² Twitter's standard search API allows you to collect recent tweets published in the past 7 days. (for more detail, see: <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview>). The reason why this study relied on the standard search API instead of the streaming API, which allows collecting real-time tweets, is that keyword tracking of the streaming API is not supported for "[n]on-space separated languages" such as Japanese (for more detail, see: <https://dev.twitter.com/streaming/overview>).

¹³ There is a possibility that some tweets were already deleted before the data were gathered, because there is a time lag of 2 days between publishing and crawling. However, it has been said that Twitter did not tightly control harassment tweets in Japan at that time; thus, most harassment tweets likely remained. Thus, the study presumes that attained data were not seriously biased even if some of them were deleted by users themselves. Java and Python were used to collect tweets. For the Python scripts, we adopted some of the scripts provided by [23, 24].

¹⁴ For example, *Zainichi* is a general word referring to foreigners who live in Japan, and has been used in internet discourse as a derogatory expression for Korean residents who obtained permanent residency after World War II. Similarly, *Chosen* generally refers to the Korean peninsula; however, it can also have a derogatory nuance depending on the context. There are other general terms indicating South Korea(n) and North Korea(n), *Kankoku* and *Kita-Chosen*, that do not have any derogatory nuance.

related to the racist harassment during the target period will have been collected. In total, over 19 days, 2,016,982 tweets were gathered from 257,698 distinct users.

The base period

The base network was generated based on the retweet history of 7 days, from August 14–20. This period covers two major events, occurring on August 14 and 15 respectively, which led to diplomatic friction between Japan and South Korea.

August 14 is the South Korean memorial day for the comfort women, who were forced to serve as sex slaves for the Imperial Japanese Army during World War II. South Korean and Japanese governments have different viewpoints concerning comfort women, and that has led to diplomatic conflicts multiple times. In particular, Japanese extremists have protested against the views and claims of South Korea.

The second event was the anniversary of the end of World War II, 15th August, which has caused diplomatic friction between Japan and the countries it formerly colonized, such as North and South Korea and China. Friction is mainly because of some cabinet ministers' visits to the Imperial Shrine of Yasukuni. Yasukuni Shrine enshrines the war dead, including people who were convicted as war criminals in World War II. South and North Korea and China have criticized visits to the shrine by Japanese prime ministers and cabinet members since they considered the visits to be an affirmation of the righteousness of the past war. The right-wing extremists tend to strongly support visiting Yasukuni Shrine and to protest against criticism. In 2017, although no cabinet members visited the shrine out of consideration for the diplomatic relationships with Korea and China, Prime Minister Abe sent money to the shrine, and some major politicians, such as Tomomi Inada, who had resigned the office of the Ministry of Defence a month previously, visited the shrine.

Given a spike in the number of tweets around the events, they are likely to have stimulated communication between users with negative attitudes toward Korea (see Table 2 and Fig. 1). This is useful for creating a base network since it allows us to observe more retweet relationships among users. Additionally, the base period covers a week; thus, it captures the relationships between users who are only active on particular days of the week.

The Target Period

The target period, during which adoption thresholds are measured, is from August 21 to September 1 and covers Koike's announcement and North Korea's missile launch. Estimation of the effect of these events premises that Twitter users are aware of the events. It is difficult to pinpoint when these effects start and end, as we cannot measure the awareness of each Twitter user. Therefore, we focus on the period during which these issues are most salient, as it is likely that the effect will be most pronounced during this time.

Although the media began reporting Koike's announcement on the 24th, it appears that there was a time lag between the start date of reporting and its uptake on Twitter. Based on the keyword trends presented in Fig. 2, the event appears to

Table 1 Keywords

Keywords	
General	<i>Chosen</i> , Korea including both North and South <i>Zainichi</i> , Koreans living in Japan
Harassment	<i>Chon</i> , a derogatory term for Koreans <i>Danko</i> , breaking-off of diplomatic relations <i>Nansen</i> , a derogatory term for South Koreans <i>Senjin</i> , a derogatory term for Koreans <i>Kenkan</i> , anti-Korean sentiment
Diplomatic issues	<i>Takeshima</i> , disputed island between Japan and Korea <i>Ianfu</i> , Comfort women <i>Yasukuni</i> , Yasukuni Shrine
Anti-left wing	<i>Han-nichi</i> , anti-Japan <i>Baikoku</i> , unpatriotic <i>Sayoku</i> , left-wing <i>Payoku</i> , a derogatory term for the left

have become a popular topic around the 26th and 27th, at which the frequency of the keyword “Koike” is at its highest level. Although the corresponding frequency was also at a higher level on the 24th, it was only for a short period. Top daily retweets also follow this trend. Out of the top 10 most retweeted tweets each day, 7 and 9 of them were related to Koike’s announcement on the 26th and 27th, respectively. In contrast, only 1, 0 and 4 were related to it on the 24th, 25th and 28th.¹⁵

There are two possible explanations for this time lag. First, while some media outlets began reporting the news on the 24th, it only made it to the major morning newspapers on the following day (the 25th), and it was not a top news item.¹⁶ Second, as of the 24th, Koike’s intentions and the context behind her decision were not clear. While there were speculations, the reporting mainly focused on the fact that she declined to send the address. Her intentions became more evident during the press conference on the 25th, where she addressed her decision and the reasons for that. After this press conference, on the 26th, major media reported Koike’s decision with more detailed context and interpretation, which is likely to have sparked controversies and made the topic more visible on Twitter.¹⁷ The topmost retweeted tweets also illustrate the development of the issue. Initially, there was only one tweet ranked among the top 10 most retweeted tweets on the 24th, which only introduced

¹⁵ For more detail of the top retweets, see Appendix C.

¹⁶ For example, on the 25th, the news was featured in the morning edition of the *Asahi Shimbun* one of the major national newspapers, though not on the front page, but in the society section. For more detail information about the newspaper’s pages, see https://www.asahi.com/shimen/20170825/index_tokyo_list.html (accessed Feb 9, 2023).

¹⁷ For instance, see the morning newspaper of the *Asahi Shimbun* from August 26th here: https://www.asahi.com/shimen/20170826/index_tokyo_list.html (Japanese, accessed Feb 9, 2023). Additionally, the press conference is available here: <https://www.youtube.com/watch?v=VVI4s7mcu9E> (accessed Feb 9, 2023).

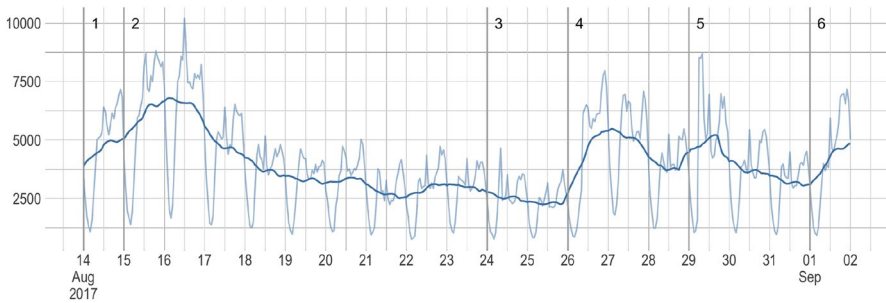


Fig. 1 Total number of tweets by hour. The lighter color shows number of tweets each hour; The darker color shows moving averages calculated every hour using a 24-hour time window. Numbers in the figure correspond to the numbers representing major events listed in Table 2

the media coverage reporting Koike’s decision to decline sending a memorial address. In contrast, the top retweeted tweets from the 26th encompassed not only Koike’s decision but also its interpretation and users’ own opinions about the issue.

Given the trend of the keyword and top retweeted tweets, as well as the context of how the issue developed, we consider the 26th as the point when Koike’s announcement became salient on Twitter. Starting from the 26th, there are three peaks in the keyword trend for “Koike.” First, it heightened on the 26th at 11 a.m. and reached the second highest peak on the 26th at 9 p.m. It shows a tendency to fall after the third peak on the 27th at 9 a.m. Thus, the study focuses on the time between the 26th at 11 a.m. and the 27th at 9 a.m. to examine the effect of Koike’s announcement.

By contrast, there was almost no time lag with the missile event on the 29th. Two hours after the missile launch, 29th at 8 am, the keyword “missile” heightened, and it reaches its highest peak at noon. The trend changed to fall after the third peak at 7 p.m. at the same day. The top retweets also follow this trend. Out of the top 10 most retweeted tweets, 8 of them were related to the missile launch on the 29th, while 0 and 4 were related on the 28th and 30th, respectively (Appendix C). Thus, we focus on the 29th between 8 a.m. and 7 p.m. to test the effect of the missile launch.

Analysis

In this section, after detecting hate speech posts using a machine learning algorithm, we evaluate the effect of the two targeted events on hate speech dissemination, using the average adoption threshold.

Table 2 Number of tweets and relevant events during the investigated period. The italic numbers corresponds to the numbers appeared in Fig. 1 and 2, which indicates the dates of the events

Day	Tweets	Events
The base period for generating base network: from 14th August to 20th August		
2017-08-14	113,510	Comfort women memorial day (South Korea) – 1
2017-08-15	144,608	Anniversary of the end of World War II (Japan), Visiting Yasukuni (Japanese politicians) – 2
2017-08-16	158,481	
2017-08-17	111,683	
2017-08-18	87,933	
2017-08-19	78,201	
2017-08-20	79,961	
The target period: from August 21 to September 1		
2017-08-21	64,385	
2017-08-22	69,985	
2017-08-23	71,822	
2017-08-24	60,312	Media started to report that Koike would skip sending a memorial address – 3
2017-08-25	54,675	At the press conference, Koike addressed her decision not to send a memorial address
2017-08-26	110,996	Koike's comment become popular on Twitter – 4
2017-08-27	122,996	Koike's comment become popular on Twitter
2017-08-28	89,287	
2017-08-29	120,392	North Korea launched the missile – 5
2017-08-30	87,204	
2017-08-31	75,790	
2017-09-01	102,194	Memorial ceremony commemorating the Korean victims of the massacre that followed the 1923 Great Kanto Earthquake – 6

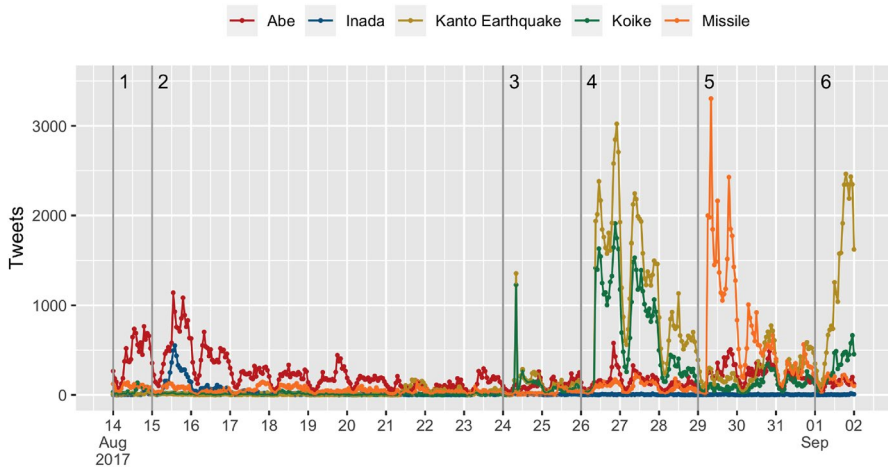


Fig. 2 The number of tweets including keywords associated with political events (hourly). *Abe* refers to Shinzo Abe, the prime minister, and *Inada* to Tomomi Inada, former Ministry of Defense. *Koike* refers to Yuriko Koike, the Governor of Tokyo. Numbers in the figure correspond to the numbers representing major events listed in Table 2

Detecting Hate Speech using a Support Vector Machine

In this section, we detect hate speech posts using a supervised machine learning algorithm, the support vector machine (SVM).¹⁸ SVM is a popular machine learning algorithm that has demonstrated good performance in classification tasks. In particular, SVM can handle data with lots of features, such as text data. Further, by utilizing different Kernel functions, it allows flexible fit for diverse data structures.¹⁹

This study utilizes linear Kernel since most text classification problems are linearly separable and it shows the best performance for data with many features [18, 22]. To improve performance, term frequency-inverse document frequency (tf-idf) values are used as weights, and Grid Search is used to find the best parameters (those which produce the highest performance). Before the machine learning process, each text is segmented into words, and the words are lemmatized to improve machine learning performance.²⁰

¹⁸ The scikit-learn libraries in Python are used to build the models for SVM. The algorithm code can be provided by the author upon request.

¹⁹ The study also tested a Naive Bayes algorithm, as these are known for good performance in text classification [28, 29]. The study adopts SVM because it outperformed the tested Naive Bayes algorithm. Using the same training set, the multinomial Naive Bayes model predicts 85% of tweets correctly. Precision and recall were 0.80 and 0.56, respectively.

²⁰ For both process, MeCab (<http://taku910.github.io/mecab/>) which provides part-of-speech and morphological analysis for Japanese, is used. Specifically, the project utilizes the mecab-ipadic-NEologd dictionary (<https://github.com/neologd/mecab-ipadic-neologd>), because it recognizes new words which are frequently used in the online space. The algorithm code (written in Python) can be provided by the author upon request.

For the training algorithm, 2062 tweets were randomly chosen from the datasets and manually classified into two categories: hate speech (1) or not (0).²¹ Tweet posts were classified as hate posts when they expressed harassment toward Koreans, that is, attacks, threats, or insults. Among 2062 tweets, 553 were identified as hate posts (26.8%).

K-fold cross-validation ($k=10$) was used to examine classification performance. The algorithm predicts 88% of tweets correctly. Precision and recall values provide information regarding false detection: 79% of the cases detected as hate posts were actual hate posts (precision) and 73% of actual hate posts were predicted to be hate posts (recall).²²

According to a manual inspection, most false-positive cases were in the gray zone between hate and non-hate speech. As we described before, only cases which expressed clear hate toward Korean were labeled as hate speech in the training set. By contrast, multiple tweets in the analysis were classified as hate posts because they contained offensive words, even though they did not apparently express hate toward Koreans. This type of false-positive group should not cause serious problems for the estimation of an adoption threshold. Even though they are in the false-positive group because of the strict coding standards of this study, it is also true that these posts are fairly close to hate speech. Furthermore, it can be assumed that this type of case behaves similarly to hate speech posts.

There are also cases classifying hate posts as non-hate posts (false-negatives). The current model often fails to classify posts when they express hate but do not use offensive words. Such cases are not desirable since we would not possibly observe all hate speech users. However, we can reasonably assume that these errors appear randomly over time. Therefore, bias in our inference about temporal dynamics of adoption threshold may be ignorable, even if the inference may be less efficient.

Measuring Average Adoption Thresholds

To examine the hypotheses concerning the effect of political leaders' behavior on adoption thresholds and evaluate its effect size, we compare the average adoption threshold (hereafter, ADT) measures at each hour during the target period. Figure 3 shows the estimated ADT with 95% confidence interval. Each ADT is calculated by using a 24-hour time window starting at the corresponding time. For example, the ADT of August 21 at 0 a.m. is the average of adoption thresholds of those Twitter users who published their first hate tweets between August 21 0 a.m. and August 22 0 a.m. The reason we adopted 24 h time window is that it is natural to assume

²¹ Two coders, fluent in Japanese and understanding the Japanese context, manually classified the tweets. The matching rate and the Cohen's κ coefficient between them were 85.5% and 0.615, respectively. The level of agreement is substantial [26]; thus, the study uses one of the sets of coding results for the training set.

²² Precision = True Positive / (True Positive + False Positive); Recall = True Positive / (True Positive + False Negative). The reason for the high accuracy compared with other measurements is the unbalanced sample size of the two classes. The proportion of non-hate posts is much higher than that of hate posts (73% of posts in the training set are non-hate posts), and precision and recall rate of non-hate posts are higher than those of hate posts.

that the mechanism investigated here, posting tweets in response to others' postings, occurs in 24-hours-rhythms. Since such circadian rhythms can vary according to individuals in terms of when they are active and inactive, we measure ADTs in each hour so that we can capture general tendency.

As discussed above, we focus on the period between the 26th at 11 a.m. and 27th at 9 a.m. as well as that between the 29th at 8 a.m. and at 7 p.m. The former is subject to the effect of Koike's announcement and the latter to the effect of the missile event. However, we cannot directly translate these periods to the one in the ADTs, since the 24 h windows are used to measure ADTs. For example, the time window to measure the ADT at 9 a.m. on the 26th covers 22 h of the period subject to the Koike's effect since the latter period begins at 11 a.m. of the same day. Thus, we consider the ADTs based on the period overlapped for more than 12 h with the event period as those which are subject to the corresponding event. In Fig. 3, those ADTs colored red are considered to be affected by the Koike's event and those colored blue by the missile event. The other ADTs are considered to be a control period.

According to Fig. 3, both targeted events have lowered the adoption threshold level, which aligns with the hypothesis. Concerning the effect of Koike's announcement, ADTs under its effect are between 3.5 and 4.0, while those in the previous period are mostly between 3.8 and 4.4. A closer look at the period under Koike's effect reveals that ADTs generally fall on the 26th and remain at the lower level during the 27th and that they show less variance. By contrast, the other periods rather show the users' 24-hour rhythms.

The ADTs under the missile event also seem to be lower than other control periods, but their pattern is different from those under the Koike's event. While under the Koike's effect it took almost a day to reach the lowest ADT level (26th 10 p.m.), the missile event hits the lowest level faster, in about half a day. This possibly reflects the speed of the news spreading to the public through the different media channels. While the missile news spread to almost all Japanese citizens in the couple of hours after the missile launched, Koike's announcement spread to only some citizens, and it took some time to spread.²³

Additionally, the ADTs show more fluctuation under the missile event than Koike's event. While ADTs under the missile event increased after hitting the lowest value, ADTs under Koike's event remained in the lower range for a while after hitting the lowest value. The speed of the news spreading to Twitter users is a possible reason for this pattern. Since Koike's announcement spread more slowly compared to the missile event, some users reacted later than others, which possibly contributed to this pattern. Another possible explanation is the distinct nature of these events. Unlike the missile launch, Koike's announcement is more directly related to hate, as it could be interpreted as supporting the claims of hate groups. The effect of such a different nature of events should be carefully examined in future research.

²³ Since the missile news was widely spread, it brought more number of non-habitual hate users. Appendix A shows that the missile event introduced more new haters, who published a hate post in the target period but not in the base period.

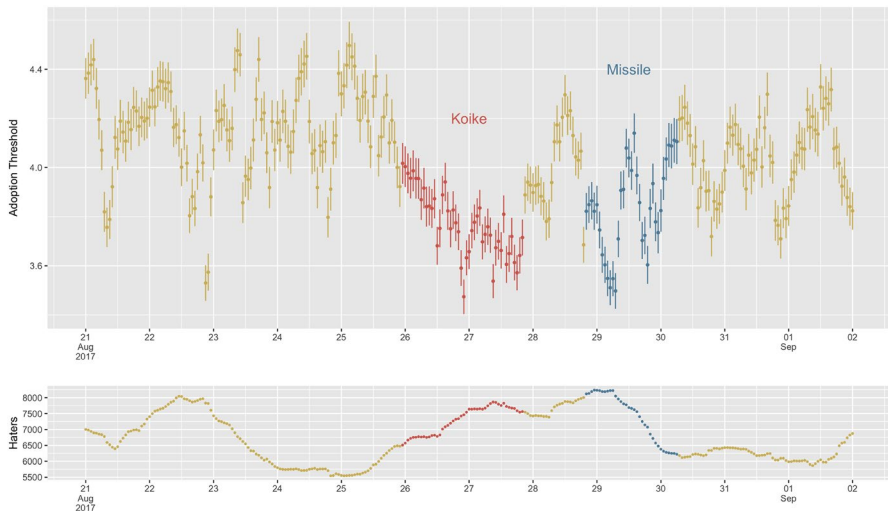


Fig. 3 The ADTs (average adoption thresholds) and the number of the users expressing hate speech. Each data point is calculated using a 24-hour time window, starting with the data point time. For example, the ADT of August 21st at 0 a.m. is the average of adoption thresholds of haters appeared from August 21 at 0 a.m. to August 22 0 a.m. In the ADT plot, dots and vertical lines indicate the point estimates of ADT and their 95% confidence intervals. The data points subject to the effect of events are colored in red (Koike) and blue (Missile)

Furthermore, we can observe a general tendency in that the number of hate users increases as the ADT decreases. The number of haters were heightened at both Koike and the missile event period.²⁴

Regression analysis

Having visually inspected the temporal development of the ADTs in Fig. 3, this section provides additional statistical evidence for the effect of leading politicians on hate speech behavior.

The dependent variable is ADTs measured each hour. The main independent variables are whether the corresponding ADT is subject to the targeted events or not. We included two event dummy variables, correspondingly: 1 indicates a corresponding ADT which is subject to the respective event, otherwise 0. Since each ADT was measured using a 24-h time window and was calculated each hour, there are 23 h time frame overlap between two consecutive ADTs, which causes auto-correlation. To remove this auto-correlation, a lagged dependent variable, ADTs at $t - 1$, is

²⁴ Besides the periods corresponding to the two targeted events, another irregular pattern on August 22 can be detected by visual inspection; on that day, some of the ADTs placed in the lowest range and a high number of active hate users could be seen, which is comparable to the Koike and the missile event. Based on hate tweets contents published in the corresponding date, this can be triggered by news reporting a violent criminal act committed by a Korean national.

included. Additionally, dummy variables for each hour are included to control the circadian trend.

According to Table 3, Koike's announcement had a significantly negative effect on ADTs at 5% level, which corresponds to the hypothesis as well as the above visual inspection. One may wonder about the small effect size of -0.08 . Here, we have to keep in mind that the effect size refers to only differences after one hour, whereas Koike's announcement had an effect over almost two whole days, that is, 46 h. The accumulated effect for this period based on the regression results is -0.34 , which results in increases of 2,029 hate posts and 522 hate users. This accounts for 5% of total hate posts and 4% of total hate users that appeared in the corresponding period.²⁵

The missile event also shows a significantly negative effect at the 5% level, while the effect size is almost half of that of the Koike's event. The accumulated effect during the whole missile event is -0.25 , which leads to the increase of 1,279 hate posts and 329 haters. This accounts for 3% of total hate posts and total hate users appeared in the missile event period. This smaller effect size might suggest that the Koike's event had a stronger impact on lowering hate users ADTs than the missile event. However, the difference in the effect sizes is not significant (see Appendix B.2).

One may wonder whether Koike's effect lasts even after August 27; if so, it will also affect our estimate of the missile event's effect. While this possibility cannot be excluded, the extent of the Koike's effect here should be small since Koike's announcement did not appear in major Twitter topic on the 29th,²⁶ and the level of ADTs came back at a similar level to the pre-Koike period, before the missile event took place. Even if there was a small amount of effect of the Koike's announcement that lasted, it should have resulted in a larger effect of the missile event. In other words, the result in Table 3 can be interpreted as a conservative result for comparison of two effect sizes.

Discussion

In this study, we hypothesized that a leading politician's behavior signaling support for racist harassers will increase hate speech by shifting social norms. To examine the hypothesis and to quantify the social norm shift, the study utilized an adoption threshold measure using a Twitter retweet network. Our empirical study focused on the effect on hate speech spread of an announcement by Yuriko Koike, Governor of Tokyo. The results support the hypothesis: after her announcement concerning the memorial ceremony for Korean massacre victims, ADTs of hate speech users were significantly lowered. This led to increases of hate posts and hate users; Koike's

²⁵ The increased numbers of hate posts and hate users were predicted by using the regression results of ADTs on the number of hate posts and hate users. For more detail of the regression results, see Appendix B.3.

²⁶ In the missile period, the ratio of the tweets including the keyword *Koike* to *Missile* is 4%, and additionally, there is no tweet mentioning Koike's event among the top 50 retweeted tweets on the 29th.

announcement was responsible for about 5% (2,029) of hate posts and 4% (522) of hate users appeared during the period.

Further, this study compared the effect of Koike's announcement to that of North Korea's missile launch, a national security threat. Among habitual users, who appeared both in the base and the target period, the average level of ADTs and the number of hate users were comparable in size. Meanwhile, the estimated effect of Koike's announcement in the regression analysis was larger than that of the missile launch. This is a surprising result considering the characteristics of the two events. There were large differences in the number of people who accessed the news; while the missile event was headline news on the day, Koike's announcement was not. This suggests that powerful people can have considerable capacity to guide other people's behavior especially when social norms matter. Additionally, it is worth emphasizing that Koike's announcement was not a direct attack on Koreans—rather, it signaled support for the radical right-wing discourse. It is reasonable to conjecture that politicians' behavior which more directly attacks minorities could introduce wider hate speech spread than that demonstrated in this study.

However, there are also several limitations that should be acknowledged. Our threshold measurement, defined as the minimum number of sources required to motivate an individual to publish a hate speech post, operates under the assumption that users observe a certain number of hate posts from their alters before taking action themselves. This assumption implies a dependent relationship between prior hate speech posts from alters and subsequent posts by the observer. However, it is conceivable that some users do not adhere to this assumption. For example, some users may post hate speech independently, without seeing prior posts within their network. Consequently, our threshold measurement may incorporate some level of error due to these potential independent actions. Future studies addressing this issue will be necessary to more accurately measure adoption thresholds.

This study is designed to gauge aggregate-level shifts in social norms triggered by external events. With our primary focus on changes in social norms, we have assumed that the influence exerted by social networks remains constant throughout the period being studied, thereby not delving into the effects of networks. Nonetheless, a thorough investigation into how individual's network influences their perception of social norms and their subsequent behaviors could significantly enhance our understanding of prejudiced behavior on social media. As important area for such study would involve analyzing the proportion of hate alters within an individual's network and investigating the potential disproportionate influence these hate alters may exert.

This study posits that users' psychological inclination to post hate speech remained constant during the investigated period. However, one may wonder whether the first event had influenced users' psychological inclination, thereby affecting the estimated results in the second event. The most likely scenario is that experiencing the first event made users more comfortable with posting hate speech due to desensitization to hate speech. Desensitization is a psychological process that refers to a decrease in emotional or psychological sensitivity to a particular stimulus over time. Previous studies suggested that frequent exposure to hate speech leads to a lack of sensitivity to the usage of hate speech in everyday conversation and

Table 3 The effects of two targeted events on the adoption threshold. Bold values indicate significance at the 5% level. The estimation result of the 23 h dummies are omitted. The complete results are available in Appendix B.1

	Estimate	Std. Error
(Intercept)	0.99	0.16
Events		
Koike	– 0.08	0.02
Missile	– 0.05	0.02
Controls		
Lagged dependent variable	0.76	0.04
Hour dummies 01 - 23	<i>Omitted</i>	
N	288	
R-squared	0.777	

in online behavior through a process of desensitization [3, 38]. Although our core findings would remain unchanged even if such an effect existed, as it should have resulted even in a larger effect of the second event (i.e., the missile launch), it is worth considering in future research.

One may wonder whether the results of this study are specific to the Japanese case. One possible argument is that the Japanese group-centered culture is responsible for the results above. Such arguments seem to be unrealistic, for the following reasons: First, the absolute level of the adoption threshold is not at stake in the study. Japanese Twitter users may have a higher absolute adoption threshold because of their culture in comparison with those in other countries; however, the results above are about the relative change in the threshold level caused by certain events, which is less likely to be affected by such cultural factors. Second, our theoretical consideration is based on more general psychological mechanisms rather than culturally specific ones. There is no reason that these psychological mechanisms should operate only in Japan. Third, similar phenomena are also empirically observed around the world. As introduced before, it has been reported that hate speech, racist discourse, and hate crime strongly increased in the United States after Donald Trump took over the presidency.

Meanwhile, varied possible events in different contexts do not have to show the same size of effects on spread of hate speech behavior. The effects will be differently conditioned by their context. To obtain more general effects, further studies examining other, diverse cases will be needed.

Appendices

A The number of haters during the target period

See Fig. 4.

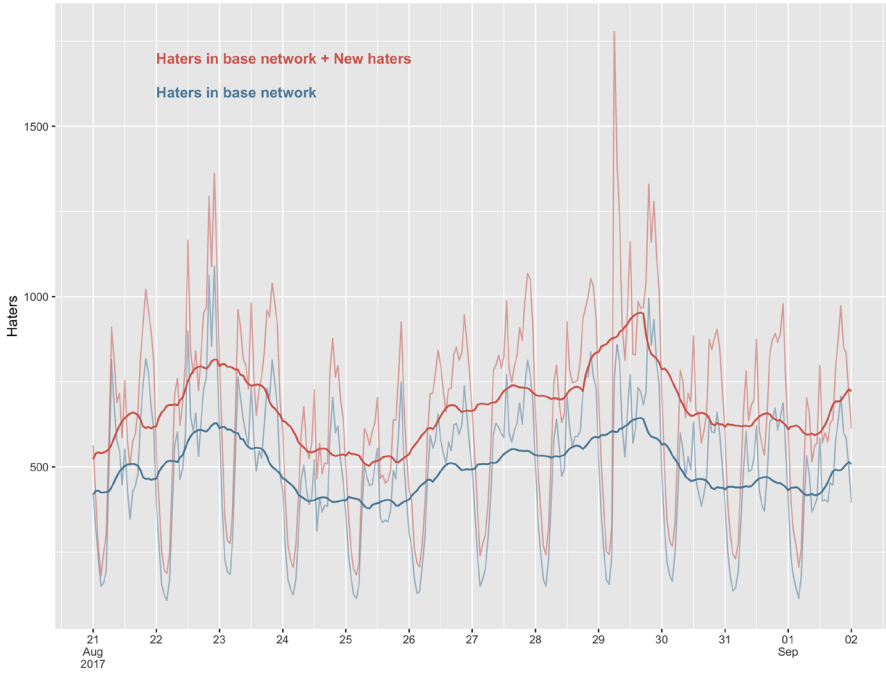


Fig. 4 Number of haters during the target period. The lighter colors show number of haters each hour; The darker colors show moving averages calculated every hour using a 24-hour time window. Blue color indicates hate users appeared in both base network and corresponding time; Red color indicates total number of haters appearing at corresponding time

B Regression results

B.1 The effect of two targeted events on the adoption threshold (full results of Table 3)

See Table 4.

Table 4 Bold values indicate significance at the 5% level. Detailed explanation of the variables are stated in the section 4.3

	Estimate	Std. Error
(Intercept)	0.99	0.16
Events		
Koike	– 0.08	0.02
Missile	– 0.05	0.02
Controls		
Lagged dependent variable	0.76	0.04
hours 01	0.03	0.05
hours 02	0.01	0.05
hours 03	0.01	0.05
hours 04	– 0.01	0.05
hours 05	–0.06	0.05
hours 06	–0.04	0.05
hours 07	–0.01	0.05
hours 08	0.01	0.05
hours 09	0.01	0.05
hours 10	–0.01	0.05
hours 11	–0.03	0.05
hours 12	–0.03	0.05
hours 13	0.02	0.05
hours 14	–0.07	0.05
hours 15	–0.04	0.05
hours 16	0.03	0.05
hours 17	–0.08	0.05
hours 18	–0.04	0.05
hours 19	– 0.12	0.05
hours 20	–0.04	0.05
hours 21	–0.09	0.05
hours 22	–0.05	0.05
hours 23	0	0.05
N	288	
R-squared	0.777	

B.2 Examining the difference of the two event's effect

See Table 5.

Table 5 *Koike* is a dummy variable with 1 indicating the ADTs which are subject to Koike's event. *Koike + Missile* is a dummy variable with 1 indicating the ADTs which are subject to both Koike and the missile event. Control variables are the same as the ones included in the regression model Table 3. Bold values indicates significance at the 5% level

	Estimate	Std. Error
(Intercept)	0.99	0.16
Events		
Koike	-0.03	0.03
Koike + Missile	- 0.05	0.02
Controls		
Lagged dependent variable	0.76	0.04
hours 01	0.03	0.05
hours 02	0.01	0.05
hours 03	0.01	0.05
hours 04	-0.01	0.05
hours 05	-0.06	0.05
hours 06	-0.04	0.05
hours 07	-0.01	0.05
hours 08	0.01	0.05
hours 09	0.01	0.05
hours 10	-0.01	0.05
hours 11	-0.03	0.05
hours 12	-0.03	0.05
hours 13	0.02	0.05
hours 14	-0.07	0.05
hours 15	-0.04	0.05
hours 16	0.03	0.05
hours 17	-0.08	0.05
hours 18	-0.04	0.05
hours 19	- 0.12	0.05
hours 20	-0.04	0.05
hours 21	-0.09	0.05
hours 22	-0.05	0.05
hours 23	0	0.05
N	288	
R-squared	0.777	

B.3 The effect of ADT on the number of hate posts and hate users

See Table 6.

Table 6 Bold values indicate significance at the 5% level. The independent variable is ADTs measured every hour using 24-h time window. The dependent variables are the number of hate posts and hate users which corresponds to the ADTs

	Hate posts		Hate users	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	46,854.22	3663.51	12,942.24	751.52
ADT	- 5946.64	910.49	- 1528.68	186.77
N	289		289	
R-squared	0.129		0.189	

C Top retweets

Table 7 displays the count of tweets related to our targeted events, Koike's announcement and the missile launch, among the top 10 most retweeted tweets. To construct the table, we initially identified the 10 most retweeted tweets for each day and manually categorized each of these tweets based on their relevance to the targeted events. Tweets were classified as related to Koike's announcement if they directly mentioned the announcement or expressed opinions about it. Analogously, tweets were categorized as related to the missile launch if they mentioned the launch itself or other aspects such as missile alarms.

Table 7 Number of Tweets related to the events among the 10 most retweeted tweets

Day	Koike	Missile
2017-08-21	0	0
2017-08-22	0	0
2017-08-23	0	0
2017-08-24	1	0
2017-08-25	0	1
2017-08-26	7	1
2017-08-27	9	1
2017-08-28	4	0
2017-08-29	0	8
2017-08-30	0	4
2017-08-31	0	2
2017-09-01	1	0

Acknowledgements The authors thank Ulrik Brandes, Karsten Donnay, Susumu Shikano, David Garcia and his working group's reading seminar participants, along with the participants of the "Evolving Legal Norms Regarding Hate Speech in Japan" panel at the 2020 APSA Annual Meeting, for their helpful comments and feedback. We are grateful to the University of Oldenburg for financial support.

Author Contributions T.K. developed the theoretical framework, designed the study, analyzed the data, interpreted the results, and wrote the manuscript. Both T.K. and Y.O. contributed to data collection and machine learning.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability statement The tweet IDs used for this study, including their labels, are available at https://osf.io/rg7em/?view_only=f8a9ef048ec147cc8caf7bf76aa78a7a.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Allen, C., & Nielsen, J. S. (2002). *Summary report on islamophobia in the eu after 11 september 2001*. European Monitoring Centre on Racism and Xenophobia. https://fra.europa.eu/sites/default/files/fra_uploads/199-Synthesis-report_en.pdf
2. Alvarez, J. E. (1999). Crimes of states/crimes of hate: Lessons from Rwanda. *Yale Journal of International Law*, 24, 365.
3. Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41, 3–33.
4. Bursztyjn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11), 3522–3548.
5. Canetti-Nisim, D., Ariely, G., & Halperin, E. (2008). Life, pocketbook, or culture: The role of perceived security threats in promoting exclusionist political attitudes toward minorities in Israel. *Political Research Quarterly*, 61(1), 90–103.
6. Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3), 702–734.
7. Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, 82(3), 359.
8. Crandall, C. S., Miller, J. M., & White, M. H. (2018). Changing norms following the 2016 US presidential election: The trump effect on prejudice. *Social Psychological and Personality Science*, 9(2), 186–192.
9. Dovidio, J. F., & Gaertner, S. L. (1999). Reducing prejudice. *Current Directions in Psychological Science*, 8(4), 101–105. <https://doi.org/10.1111/1467-8721.00024>
10. Edling, C., Rydgren, J., & Sandell, R. (2016). Terrorism, belief formation, and residential integration: Population dynamics in the aftermath of the 2004 Madrid terror bombings. *American Behavioral Scientist*, 60(10), 1215–1231.

11. Edwards, G.S., & S. Rushin. (2018). The effect of President Trump's election on hate crimes. Available at SSRN: <https://ssrn.com/abstract=3102652>.
12. Gautier, P. A., Siegmann, A., & Van Vuuren, A. (2009). Terrorism and attitudes towards minorities: The effect of the Theo van Gogh murder on house prices in Amsterdam. *Journal of Urban Economics*, 65(2), 113–126.
13. Giani, M., & Méon, P.G. (2019). Global racist contagion following Donald Trump's election. *British Journal of Political Science*, 51(3), 1332–1339.
14. Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78, 1360–1380.
15. Higuchi, N. (2018). The radical right in Japan. In J. Rydgren (Ed.), *The Oxford handbook of the radical right*. Oxford University Press.
16. Higuchi, N. (2021). Japan's postcolonial hate speech, In S. Higaki, & Y. Nasu (Eds.), *Hate speech in Japan: The possibility of a non-regulatory approach*. Cambridge University Press.
17. Hopkins, D.J. (2010). Politicized places: Explaining where and when immigrants provoke local opposition. *American Political Science Review*, 104(1), 40–60.
18. Hsu, C.W., C.C. Chang, & C.J. Lin. 2016. A practical guide to support vector classification. <http://www.datasceinceassn.org/sites/default/files/Practical%20Guide%20to%20Support%20Vector%20Classification.pdf>.
19. Huang, J., & Low, C. (2017). Trumping norms: Lab evidence on aggressive communication before and after the 2016 US presidential election. *American Economic Review*, 107(5), 120–124.
20. Human Rights Watch. (2002). We are not the enemy: Hate crimes against Arabs, Muslims, and those perceived to be Arab or Muslim after September 11. *United States Country Report*, 6(14).
21. Hutchison, M. L., & Gibler, D. M. (2007). Political tolerance and territorial threat: A cross-national study. *The Journal of Politics*, 69(1), 128–142.
22. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec, & C. Rouveirol (Eds.), *Machine learning*. Berlin Heidelberg: Springer.
23. Jürgens, P., & Jungherr, A. (2016). A tutorial for using Twitter data in the Social Sciences: Data collection, preparation, and analysis. Available at SSRN: <http://ssrn.com/abstract=2710146>
24. Jürgens, P., & Jungherr, A. (2016). twitterresearch [Computer software]. Retrieved from <https://github.com/trifle/twitterresearch>
25. King, R. D., & Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology*, 51(4), 871–894.
26. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
27. Legewie, J. (2013). Terrorist events and attitudes toward immigrants: A natural experiment. *American Journal of Sociology*, 118(5), 1199–1245.
28. McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI'98 Workshop on Learning for Text Categorization* (pp. 41–48). Wisconsin: Madison.
29. Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive bayes-which naive bayes? *CEAS* (Vol. 17, pp. 28–69). CA: Mountain View.
30. Müller, K., & Schwarz, C. (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. Available at SSRN: <https://ssrn.com/abstract=3149103>
31. Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39, 629–649. <https://doi.org/10.1007/s11109-016-9373-5>
32. Newman, B., Merolla, J. L., Shah, S., Lemi, D. C., Collingwood, L., & Ramakrishnan, S. K. (2021). The Trump Effect: An experimental investigation of the emboldening effect of racially inflammatory elite communication. *British Journal of Political Science*, 51(3), 1138–1159. <https://doi.org/10.1017/S0007123419000590>
33. Newman, B. J. (2013). Acculturating contexts and Anglo opposition to immigration in the United States. *American Journal of Political Science*, 57(2), 374–390.
34. Noelle-Neumann, E. (1984). *The spiral of silence: Public opinion-Our social skin*. Chicago: The University of Chicago Press.
35. Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3), 566–571. <https://doi.org/10.1073/pnas.1514483113>
36. Rosenfeld, M. (2002). Hate speech in constitutional jurisprudence: A comparative analysis. *Cardozo L. Rev.*, 24, 1523.

37. Schäfer, F., Evert, S., & Heinrich, P. (2017). Japan's 2014 general election: Political bots, right-wing internet activism, and Prime Minister Shinzō Abe's hidden nationalist agenda. *Big Data*, 5(4), 294–309.
38. Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44(2), 136–146.
39. Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, 27(4), 486–496. <https://doi.org/10.1177/0146167201274009>
40. Taka, F. (2015). *Dissecting racism in Japan (in Japanese)*. Keisoshobo.
41. United Nations. (2019). *United Nations strategy and plan of action on hate speech*. United Nations. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf
42. Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge: Cambridge University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.