

Exploring the Causal Effects of Hate Speech on Social Media Users During the COVID-19 Pandemic

A Bayesian Structural Time-Series Analysis

Fatemeh Pourgholamali*, Akram Alam

Department of Computer Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran; f.pourgholamali@vru.ac.ir, akramalam05@gmail.com

ABSTRACT

Social media platforms are vital repositories of user-generated content, reflecting a range of emotions, interests, and discussions. Among these interactions, hate speech has emerged as a significant issue, influencing user behavior. While prior studies have attempted to analyze user characteristics to understand hate attitudes, they often rely on simple statistical comparisons and lack robust methods for causal effect estimation. This study investigates the causal effects of hate speech on user behavior on Twitter (now known as X) during the COVID-19 pandemic, characterized by heightened online discourse and harmful rhetoric. We focus on users who broadcast hate speech to determine how such expressions affect emotional responses. Using a Bayesian structural time-series modeling approach, we isolate the effects of hate speech from confounding factors, providing a solid framework for causal inference. Our findings indicate a significant shift in user emotions following instances of hate speech, demonstrating a measurable impact on user dynamics. We also analyze hashtag usage during this period, emphasizing their role in shaping online discourse. This study enhances understanding of the relationship between hate speech and user behavior, offering insights crucial for researchers, policymakers, and social media platforms in developing strategies to mitigate the adverse effects of online hate speech.

Keywords— Hate-Speech, Social Media, Time Series Analysis, COVID-19, Causal Effect, Sentiment Analysis.


1. Introduction

Software technologies are developing rapidly, giving rise to online social networks. In addition to the benefits of social media's widespread use, one negative aspect of this development is the increase in hate speech on these platforms, which is one of the most pervasive and potent threats globally [1]. The COVID-19 pandemic has significantly accelerated online discourse, including the proliferation of harmful rhetoric, highlighting the urgent need to understand the implications of hate speech during this critical period. Notably, the primary targets of such hate speech have been associated with the origins of the pandemic, particularly focusing on Asia and China, which led to numerous studies dedicated

to examining anti-Asian hate speech during the COVID-19 era [2].

According to [3, 4], hate speech is any form of abusive, intimidating, harassing, or hateful expression in online discussions that targets people because they are part of a social group. Henri Tajfel's social identity theory [5] sustains that individuals who assign themselves to groups look for a good social identity by comparing themselves to other groups. Therefore, it would seem that the need to validate a group's identity by disparaging others serves as one driving force behind hate speech.

The literature on the effects of hate speech on users is extensive and multifaceted, exploring various dimensions such as propagation dynamics, user characteristics, and societal impacts. Many researches have been

 <http://dx.doi.org/10.22133/ijwr.2025.488637.1248>

Citation F. Pourgholamali, A. Alam, "Exploring the Causal Effects of Hate Speech on Social Media Users During the COVID-19 Pandemic", *International Journal of Web Research*, vol. 8, no. 1, pp. 13-23, 2025, doi: <http://dx.doi.org/10.22133/ijwr.2025.488637.1248>.

*Corresponding Author

Article History: Received: 18 September 2024; Revised: 26 November 2024; Accepted: 3 December 2024.

Copyright © 2025 University of Science and Culture. Published by University of Science and Culture. This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International license (<https://creativecommons.org/licenses/by-nc/4.0/>). Noncommercial uses of the work are permitted, provided the original work is properly cited.

conducted to measure the effects of hate speech on society [4, 6]. Research indicates that hate speech negatively impacts societal attitudes towards targeted groups [4], particularly harming disabled individuals in Norway by affecting their psychological well-being and social solidarity [7].

Among researches that explore the hate speech diffusion across social media, the study presented in [8] demonstrates how specific semiotic indicators can facilitate the identification of patterns that elucidate the mechanisms of toxic content dissemination on platforms such as Twitter (X). The authors state that this platform is increasingly recognized as a primary vector for the propagation of hate speech and misinformation. Authors in [9] examine the rapid initiation and diffusion of online hate speech on Twitter through retweeting networks, demonstrating that hateful tweets rapidly gain the majority of their retweets and susceptible users but then experience a stall, whereas non-hateful tweets tend to sustain their spread over a longer duration, albeit at a slower rate.

Hate speech can have consequences not only for society or network, but also for the speaker. Many studies look into the impact of hate speech on those who promoted hate speech and compare it to their pre-hate speech. Existing literature has made attempts to analyze user characteristics and dispositions to understand the development of hate attitudes; however, many of these studies rely on simple statistical comparisons and lack causal effect estimations that are essential for assessing the impact of hate speech on user behavior. For instance, prior research has predominantly focused on the propagation dynamics of hate speech, examining how it spreads through social networks and identifying the characteristics of users who engage in such behavior [10]. The basis of this work is that two groups are initially sampled from the population; the first group is the treatment group, consisting of individuals who have used hate speech, and the second group is the control group, made up of individuals who oppose hate speech or are neutral. They then compare various characteristics between the two groups. While these studies provide valuable insights, they often fall short of establishing direct causal relationships and can only show correlation or difference between groups. Moreover, these methods ignore temporal dynamics and potential auto-correlation within time series data.

This study aims to fill this gap by exploring the causal effects of hate speech on user behavior on Twitter (now known as X) during the COVID-19 pandemic. Specifically, we investigate the behavior of users who engage in broadcasting hate speech, seeking to determine how such expressions impact emotional experiences. We use time series data, allowing for analysis of changes over time within and between groups. This approach models the counterfactual—what would have happened to the treatment group without the intervention—, and uses pre-intervention data to estimate this counterfactual. The difference between the actual and counterfactual outcomes is attributed to the intervention. This model directly addresses causality by attempting to isolate the effect of the intervention, and provides a more robust framework for causal inference than simple comparisons.

Our findings reveal a notable shift in user emotions and the prevalence of specific hashtags following instances of hate speech, indicating a measurable impact on social media dynamics during this critical time. The insights gained from this research are crucial for researchers, policymakers, and social media platforms in developing effective strategies to mitigate the adverse effects of online hate speech and promote healthier online interactions. Our contributions include:

- We proposed a method to investigate the causal impact of hate speech on user behavior using a structural Bayesian time series analysis.
- We collected a dataset of COVID-19-related tweets, containing both hate and non-hate users, exploiting a keyword-based strategy.
- We develop a matching strategy to identify specific control users that are used in the estimation counterfactuals.

2. Related Work

In recent years, the state of disorders of users in social networks has been widely studied [11, 12]. Investigation of pro-eating disorders [13], depression [14], and bipolar disorders [15] are among these studies. Authors in [16] show that social media platforms contribute to the rise of eating disorders among adolescents by promoting unrealistic beauty standards, which can lead to unhealthy dieting and excessive exercise. They emphasize that social media can also serve as a double-edged sword, with the potential to be used positively for support and intervention in mental health, provided that appropriate measures and campaigns are implemented.

As the most influential pandemic in recent years, the coronavirus has been extensively studied to investigate its effects on the mental health of people [17-20]. These approaches frequently rely more on self-report than on observational research, which has problems with validity and reliability [21, 22]. On the other hand, the social behavior of healthy people when faced with a phenomenon or event in social networks has been studied less [23]. On the other hand, online hate speech analysis is an important and active area of research within the field of social network analysis (SNA). The literature on the effects of hate speech on users is extensive and multifaceted, exploring various dimensions such as propagation dynamics, user characteristics, and societal impacts.

Authors in [24] pointed out that being the target of hate speech could seriously affect someone's career or personal life. They studied the society of journalists and found that some female journalists chose to completely withdraw from the public as a countermeasure to the effects of hate speech, while others responded by making the hateful remarks public. Schafer et al. [4] investigated whether hate speech affects society's attitude towards the attacked group and their perception of society. The results of this research indicate that hate speech causes negative effects on social solidarity. Vedeler et al. [7] have investigated the effect of hate speech on disabled people in society and have concluded that hate speech attacks harm disabled people more. The research reported a wide range of consequences of hate speech related to psychological, social and societal issues of disabled people in Norway. Bozhidarova et al. [25] explore the relationship between online discourse and physical hate crimes, using natural language processing to analyze sentiment and polarizing tweets, which correlate with hate crimes against marginalized groups.

Some researchers investigate the impact of hate speech through the social network. Authors in [8] address the rising issue of online hate speech on Twitter, focusing on its initiation and diffusion through the retweeting network. They showed that hate speech has a fast spread in the network through retweeting and slows down afterwards, whereas non-hateful posts tend to sustain their spread over a longer duration, albeit at a slower rate. Mathew et al. [26] conduct a temporal analysis of hate speech on Gab.com, revealing a steady increase in hate speech and a faster rate of new users adopting hateful ideologies over time. It employs the DeGroot model to assign hate intensity scores to users and examines the linguistic and network characteristics of hateful versus non-hateful users. The findings indicate that the language of the Gab community increasingly aligns with that of hateful users, suggesting a concerning shift in community dynamics.

While many studies attempt to investigate the impact of hate speech on the overall network, some works focus on the effect of hate speech on the users individually. Kuřík et al. [27] focus on the lived experiences of hate speech victims, documenting its pervasive nature across online and offline contexts. The study introduces the concept of "cumulative desensitization," which exacerbates the long-term effects of hate speech. The paper presented in [28] tackles online hate towards Asians by focusing on users, particularly focusing on the risk indicators of hate towards Asians and analyzing user behavior and language prior to the COVID-19 pandemic. 'Reference users' are randomly selected from a large collection of tweets related to COVID-19, ensuring a diverse sample that reflects various perspectives on the topic. The selection process involves excluding users who have previously used anti-Asian slurs before the COVID-19 pandemic, which helps to maintain the integrity of the reference group by focusing on users who have not engaged in hate speech. Hateful users reveal a tendency to use strong negative words and terms related to harm and degradation and share a higher volume of URLs from news media. Moreover, they showed an increased sharing and liking of tweets. With a similar approach, [10] showed that hateful users tend to exhibit a communication style characterized by self-revealing content, often expressing anger and negative emotions. These users often display more polarized and extreme behaviors, suggesting that engagement in hate speech can lead to a cycle of increasing radicalization and hostility.

To investigate the impact of hate speech, most of these approaches first separate two groups of users: hate group and users that are not hated as control group. Then the effect assessment is performed by comparing features and statistics across two groups without performing any time series analysis. In this paper, we aim to investigate the impact of hate speech through a more reliable and fundamental Casual Impact approach which is based on structural Bayesian time series analysis.

3. Approach Overviews

The goal of our research is to investigate the impact of hate speech on the speaker's emotions and sentiments over the social media. The approach overview, which is illustrated in Figure 1, consists of the following steps:

Data Gathering: Using Twitter API we gather users' posts related to COVID19 considering both hate and counter speech contents.

Distinguishing Between Hate and Non-hate Users: We gather posts from two user groups: hate and non-hate users, taking into account appropriate keywords.

Preprocessing: After that, we take a few actions to prepare our data for the following stages. Preprocessing tasks include deleting users with few posts, tokenizing, removing unrelated hashtags, and removing stop words from posts.

User's Post Modeling: Next, we model every post based on the sentiments and emotions that are expressed in it.

Matching Hate and Non-hate Users: For every treatment/hate user, we need one or a small number of control/non-hate users to conduct our analysis using causal impact analysis. Therefore, to have such assignments, we must carry out a matching process.

Estimating the Causal Impact of Hate Speech on the Features: To have a reliable comparison between hate users' pre- and post-hate behavior, we employ the causal impact estimation on our time series data. The details of this process are described in Section 4.2.

4. Methodology

4.1. Dataset

With the spread of the COVID-19 pandemic, hate speech has formed in many forums against some targets, especially Asian countries. As a text-based social network, Twitter is a great source of hate speech. We tracked the spread of hate speech related to the COVID-19 pandemic on Twitter between January 27, 2020 and October 15, 2021. We used Twitter's official APIs to gather pertinent COVID-19 tweets using a keyword-based strategy. As in [29], we employed a set of three sets of keywords and hashtags:

1. COVID-19 keywords are terms related to COVID-19 that are used to gather tweets about the pandemic;
2. Hate keywords are hashtags and keywords that indicate hate towards Asians in the context of COVID-19;
3. Counterspeech keywords are hashtags and keywords used to coordinate campaigns against hate speech and to support Asians.

In total, we used 42 keywords as shown in Table 1.

With these elaborations, our dataset is categorized into two parts according to two groups of users:

Hate Users: in this group, 3000 users that have used hate speech in their tweets are gathered. To get more confidence, we assume a user is a hate user if she has left at least three hate tweets in her timeline. Since we aim to compare the user behavior before and after becoming hateful, we consider the date of the first hate tweet as the time that the user becomes hateful (hate occurrence moment). As a result, the

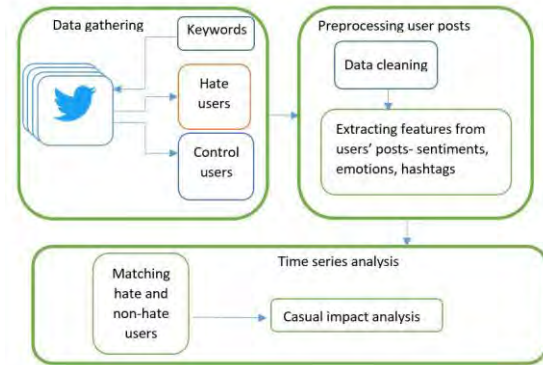


Figure. 1. An overview framework of this study.

Table 1. List of keywords and hashtags used for data collection [29].

Category	Keywords
COVID-19	coronavirus, covid 19, covid-19, covid19, corona virus
Hate keywords	#CCPVirus, #ChinaDidThis, #ChinaLied, #PeopleDied, #ChinaVirus, #Chinese Virus, Chinese virus, #ChineseBioterrorism, #FuckChina, #KungFlu, #MakeChinaPay, #wuhanflu, #wuhavirus, wuhan virus, chink, chinky, chonky, churka, cina, cokin, communistvirus, coolie, dink, niakoue', pastel de flango, slant, slant eye, slopehead, ting tong, yokel
Counterspeech keywords	#IAmNotAVirus, #WashTheHate, #Racis mIsAVirus, #IAmNotCovid19, #Be Cool2Asians, #StopAAPIHate, #Act ToChange, #HateIsAVirus

dataset of hate users is divided into BEFORE and AFTER according to the hate occurrence moment.

Non-hate Users: Similarly, we consider 3000 counterspeech users. These are users that defend the target of the hate-speech. We consider the time stamp the user has her counter-hate tweet as the counter-hate occurrence moment. Similar to the hate dataset, the non-hate dataset also is divided into BEFORE and AFTER according to the counter-hate occurrence moment.

After the data was collected, we performed a series of tasks to clean and prepare the dataset such as aligning the format of posts, omitting users with few posts, and date standardization.

4.2. Casual Impact Estimation

Policymakers frequently encounter difficulties when attempting to evaluate the impact of an intervention, such as a policy change, on a specific outcome. For instance, a website owner may seek to determine the effect of a webinar program on website

views by analyzing data collected before and after the implementation of the policy. It is essential to assess both the intended and unintended consequences of policies and interventions, whether positive or negative, to determine their overall effectiveness.

To address this need, researchers have developed various methodologies tailored to different temporal contexts for estimating intervention effects. These methodologies include: (1) time-invariant intervention effects, (2) time-varying intervention effects, and (3) dynamic regimes.

An intervention is considered time-invariant or fixed when it is implemented at a specific point in time, similar to a single-dose medication. In cases that multiple sequential interventions that vary over time are applied we have time-varying intervention effects, and for personalized intervention recommendations, dynamic regimes methodology is employed [30].

The most suitable methodology for this study is the estimation of time-invariant intervention effects. As mentioned, this approach is particularly appropriate when the intervention is applied at a specific point in time (hate speech occurrence), allowing for a clear evaluation of its impact on the outcome of interest. By focusing on time-invariant effects, we can effectively analyze the intervention influence without the complexities introduced by varying intervention conditions over time. So, this methodology will provide a robust framework for assessing the causal effects of the hate speech within the context of our study.

Formally speaking, let y_t represent the time series outcome recorded at times $t=1,2,\dots,n$ for treatment group, and x_t represent the time series for the control group. For the sake of simplicity we summarized the main notations of the formalization in Table 2.

At time T , $1 < T < n$, an intervention occurs. Our primary interest lies in modeling the potential outcome of the treatment group at $t > T$ had it not received the treatment, referred to as the counterfactual outcome i.e., $y_{t>T}^c$. The counterfactual outcome indicates what would have transpired within the treatment group had the policy not been implemented. The treatment effect can be estimated by calculating the difference between the observed values for the treatment group $y_{t>T}$ and the predicted counterfactual outcome (the unobserved values) $y_{t>T}^c$.

A method mostly used to capture the causal effect of the time series data before and after the intervention is difference-in-differences (DiD) [31].

This method goes by an assumption called common trends [32] that uses the change in the outcome of the control group as a counterfactual for the treatment group in the absence of the intervention. Figure 2 illustrates a hypothetical example of DiD on a website visit before and after the hosting of a webinar program. The aim is to see whether this program affected the number of daily viewers. In this example, users are divided into two groups: the treatment group and the control group. The treatment group receives the intervention, which consists of a specific webinar announcement made in March. The counterfactual analysis indicates what would have occurred in the treatment group if it had followed the same trajectory as the control group. The treatment effect is represented by the difference between the green dotted line and the orange line after the intervention. The results indicate that the specific webinar led to an increase in website traffic.

Given the T as the time of the intervention, $t < T$ and $t > T$ denote the pre- and post- treatment periods,

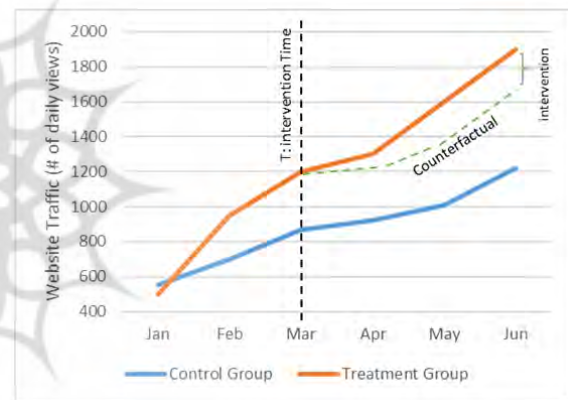


Figure. 2. An example of the difference-in-differences (DiD) method

Table 2. Main notations used in equations

Notation	Description
U	Set of hate users
U'	Set of non-hate users
y_t	Treated/hate user time series
x_t	Control/non-hate user time series
T	Hate occurrence moment in y
T'	Counter-hate occurrence moment in x
y_t^c	Counterfactual time series for y

respectively. We can calculate the DiD measure using the ATT metric as Equation (1):

$$DiD = \{E[y_{t>T} - y_{t<T}]\} - E\{[x_{t>T} - x_{t<T}]\} \quad (1)$$

A common method for modeling the causal effect of an intervention using a Difference-in-Differences (DiD) approach is a linear regression model.

While the difference-in-differences (DiD) design incorporates a temporal component, it has been noted that when applied to highly autocorrelated data, the model may underestimate the effect of the intervention [33]. To address this issue, Brodersen et al. [34] introduced a method called *Causal Impact*, which is widely utilized across various applications, including the effects of vaccines, the environmental consequences of aircraft emissions and aviation fuel taxes, as well as the relationship between mobile phone use and brain cancer [35-37]. This method expands upon the DiD framework and structural time-series models to assess the causal impact of discrete interventions. Causal Impact analyzes the relationship between the treatment and control groups prior to any intervention and forecasts the counterfactual series following the treatment. This approach is based on state-space models, as detailed below (Equation (2) to (4)) [38]:

$$y_t = \mu_t + \beta x_t + v_t \quad (2)$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + w_t \quad (3)$$

$$\delta_t = \delta_{t-1} + u_t \quad (4)$$

where y_t represents the observed outcome in the treatment group at time t . x_t represents the observed outcome in the control group at time t . μ_t represents the underlying trend in the treatment group. β is a coefficient representing the relationship between the treatment and control groups. δ_t represents the treatment effect at time t . v_t , u_t , and w_t are zero-mean noise terms. Equation (2) links the observed outcome in the treatment group to the underlying trend, the control group's outcome, and noise. Equations (3) and (4) define how the trend and treatment effect evolve over time, incorporating additional noise terms. The model is fitted to the observed data for the treatment and control groups up to a time point T , and then uses the learned parameters to predict counterfactuals for the treatment group beyond T . The difference between the observed and predicted values constitutes the estimated causal effect. The causal effect is estimated by subtracting the predicted from the observed treated time series,

which captures the semi-parametric Bayesian posterior distribution.

4.3. Matching Treatment Users With Control Users

The intervention time for the treatment group corresponds to the time of the user's first hate speech engagement i.e. T . To conduct more precise calculations and select appropriate control users, we adopt a strategy to identify the most suitable control users for the treatment group. We consider the time of expressing counterspeech among control users as the baseline time i.e. T' and select control users who have a maximum of 5 days between their counterspeech time T' , and the hate speech time T of the treatment user. Formally, given user $u \in U$ with time series y_t that has the hate occurrence time T , we aim to find some users $u \in U'$ with time series x_t that has T' as the counterspeech occurrence time with a less distance to T' . This is formulated in Equation (5) as follows:

$$|T - T'| < 5 \quad (5)$$

After the first step of filtering the control users, we select a user who has the greatest similarity to the treatment user in terms of their activity level on the social network. Ultimately, to achieve better results, we chose users with a balanced level of activity, meaning neither too high nor too low. We empirically selected users whose post volume is approximately 800 posts, with 500 posts made before T and T' time and 300 posts made after this time for the treatment and control users, respectively.

For the implementation we used the Casual Impact Package¹.

4.4. Features

Modeling User Posts: Our main goal is to trace the feelings of users before and after hate occurs. So, for each user, we extract some features from her posts.

We exploited two common libraries in Python TextBlob and Text2Emotion [39]. Sentiments i.e. positive, negative and neutral are annotated by TextBlob, while emotions are annotated using the Text2 Emotion model. We selected 10 features and modelled each post with these features. In addition to Positive, Negative and Neutral sentiments, we consider Subjectivity, Happiness, Anger, Surprise, Sadness and Fear as emotions. While most of these features are clear in this area, it is worth mentioning the definition of the Subjectivity feature. Subjectivity measures how much of the text is made up of factual

¹ <https://google.github.io/CausalImpact/CausalImpact.html>

information or subjective opinions. The higher Subjectivity means that the text contains personal opinion rather than factual information.

The pairwise Pearson Correlation Coefficients between different features are displayed in Figure 3. The strength of the relationship between the two variables increases with the correlation's absolute value. With reference to these values, we observe that two Positive and Negative features have a substantial negative correlation with the Neutral feature. Therefore, for the remainder of the investigation, we do not include the Neutral feature.

5. Results

Hashtags Analysis: As an intuitive illustration, we displayed the most frequent hashtags of users in different periods in Figure 4. We divided the posts into four groups. These groups of posts are selected from the beginning of 2020 until the time the hate speech occurs. The first part is related to the beginning of the pandemic and the last part is related to after becoming hateful.

We selected 20 top hashtags in each group to display. The first group, Figure 4a, which is related to the beginning of the spread of COVID-19, coincides with the 2019-2020 Hong Kong protests [40]. According to the picture, we can see that most of the hashtags are related to the Hong Kong protests. On the other hand, COVID-19 is still unknown and no significant COVID-19-related hashtag is seen. However, several COVID-19-related hashtags, such as “Wuhan” and “Coronavirus,” are shown in Figure 4b. Twitter mentions of the protests in Hong Kong started to decline at this time.

In Figure 4c, we are facing a significant increase in COVID-19-related hashtags, in such a way that there are only four non-related hashtags. Even though the existing demonstration in Hong Kong was the largest series of demonstrations in the history of this city [41], [42] and was therefore regarded as a significant event, it's interesting to note that the conversation about the COVID pandemic and related issues has taken the place of the demonstration on Twitter.

In Figure 4d, we can see users' hashtags after hate speech. We can see that the COVID-related hashtags have decreased and instead, the hashtags supporting the Hong Kong protests have increased significantly. This issue may have arisen because, at the time, many felt that Chinese people were victims of two things: first, that COVID-19 was first discovered in China; and second, that the Chinese government was implementing incorrect policies. As a result, people at the time expressed more empathy and support for the Chinese people, especially during the Hong Kong Protests.



Figure 3. Correlation Matrix between various user features



Figure 4. Top hashtags in four groups of tweets across different periods during the COVID-19 pandemic. Panel a) illustrates posts from the early phase of this period, while panels b) and c) present posts from subsequent times. Panel d) highlights posts that emerged during the occurrence of hate speech.

The shift in hashtag usage during this time from COVID-related topics to those supporting the Hong Kong protests highlights a significant social dynamic that warrants further exploration.

- **Contextualizing the Shift:** The decrease in COVID-related hashtags alongside the increase in support for the Hong Kong protests suggests a complex interplay between public sentiment and socio-political events. As the pandemic progressed, many individuals may have begun to associate the Chinese government with the challenges posed by COVID-19, leading to a nuanced perspective that recognized the plight of Chinese citizens in the face of both the pandemic and governmental policies.
- **Empathy and Solidarity:** The rise in hashtags supporting the Hong Kong protests indicates a shift in empathy towards the Chinese populace, reflecting a broader understanding of their struggles. This may suggest that, while some individuals expressed negative sentiments towards China due to the origins of

the virus, others recognized the distinct issues faced by the people in Hong Kong, fostering a sense of solidarity.

- **Implications for Hate Speech:** This trend also raises important questions about the nature of hate speech and its impact on public discourse. As users navigate their feelings about the pandemic and related issues, the evolution of hashtag usage may reveal underlying attitudes that could influence future discussions around race, nationality, and governmental accountability.

Causal Effect Estimation: In our study, the causal effect estimation was performed to check whether hate speech can cause to observation of a behavior in users or not. The time series for the various features is shown in Figures 5 and 6. The results of casual effect estimation provide 3 plots for each feature. The above chart displays time series corresponding to the observed values of the feature of hate and non-hate users in the 800-posts interval, along with the predicted values for the counterfactual time series represented as a dashed red line. The second plot shows the effect of the intervention on the treatment group i.e. the difference between the predicted and observed time series for the hate users. The third plot shows the Cumulative effect of the intervention.

Figure 5 shows the time series for Anger, Fear, Positive and Sadness features. Figure 5a denotes that after hate speech, the Anger emotion has increased in users.

This observation is compatible with our expectations since hate and anger are correlated emotions. However we can see from Figure 5b when users become hateful, the degree of the Fear feature decreases considerably. This observation is interesting since it was not as obvious as the previous observation was. The next Figure 5c shows a decrease in Positive feature after hate speech. But this decrease is not considerable. Figure 5d reports a considerable increase in the Sadness feature.

The results of causal effects for features Surprise, Happiness, Negative and subjectivity are shown in Figure 6. We observe that the Surprise emotion has increased considerably after hate speech. An increase in Negativity and a decrease in Happiness and Subjectivity after hate speech is also observable. While the increase in Negativity and decrease in Happiness are rather obvious, it is interesting that we observe users tend to subjective opinions rather than factual information after hate speech. Increasing an excited emotion after hate speech i.e. Surprise is interesting as well. The Causal Impact results are in line with the correlation Coefficients shown in Figure 3. For example, we can observe that hate speech has caused Sadness and Fear features to increase and

decrease respectively. On the other hand, Figure 3 shows that the correlation between these two features is -0.21 , suggesting a negative association. It has also been observed that hatred reduces positivity and happiness. This discovery aligns with the Happiness and Positive features' 0.28 correlation value.

We summarize our findings from the casual impact estimation in Figure 7. We found that:

Hate speech leads to an increase in Anger, Surprise, Sadness and Negativity.

Hate speech leads to a decrease in Fear, Happiness, Subjectivity and Positive.

Hate speech has the highest effect on Anger, Sadness and Surprise emotions and the lowest impact on Positive and Negative sentiments

Hate speech could change emotions of users significantly while its impact on sentiments is not considerable.

Comparing with Similar Work: In this section, we aim to conduct a qualitative comparison of our findings with two recent studies that investigate the impact of hate speech on user behavior.

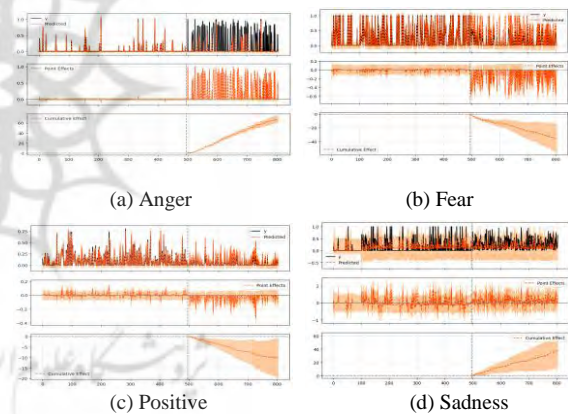


Figure 5. Casual Impact Estimation output for the Anger, Fear, Positive and Sadness features.

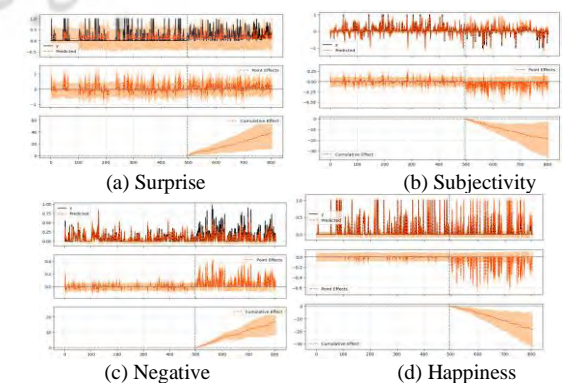


Figure 6. Casual Impact Estimation output for the Surprise, Subjectivity, Negative and Happiness features.

The study referenced in [10] demonstrates that users who engage in hate speech frequently express anger and negative emotions while tending to provide less factual information, which aligns with the results of our work. However, they did not report significant differences in means for the emotions of fear, sadness, and surprise between the treatment and control groups. In contrast, we found a significant causal effect for these emotions following exposure to hate speech. This comparison is summarized in Table 3.

While it is intuitively reasonable to discuss the expectation that emotions such as sadness should be affected after exposure to hate speech, we aim to explore this issue more fundamentally. Employing the difference in means method may yield inaccurate results that underestimate causal effects over time series. Frazier et al. [45] have noted that using the difference in means to estimate causal effects can lead to biased results due to inexact matching. This bias occurs when individuals in the treatment and control groups differ in ways that are not accounted for, potentially skewing the estimation of the average treatment effect (ATE).

In contrast, we have implemented a concise matching strategy (Section 4.3) to align control users with treatment users as closely as possible to obtain more reliable results. Consequently, we believe that there may be causal effects for the emotions of surprise, fear, and sadness that the difference in means method fails to detect, which can be revealed through structural Bayesian causal effect estimation.

6. Conclusion

This study provides an examination of the effects of hate speech on user behavior on Twitter (now known as X) during the COVID-19 pandemic. By employing a structural Bayesian time series analysis, we were able to capture the dynamic changes in user emotions. We also performed an evolution hashtag analysis across different periods during the COVID-19 pandemic. Our findings reveal significant shifts in emotional responses and social media discourse, underscoring the profound impact of hate speech on online interactions.

One of the notable outcomes of our analysis is the observed decrease in COVID-related hashtags, accompanied by a marked increase in hashtags supporting the Hong Kong protests. This shift suggests that the public sentiment evolved in response to the pandemic's context, as many individuals expressed empathy towards the Chinese populace, recognizing them as victims of both the virus's origins and the actions of the Chinese government. Such findings highlight the complex interplay between social issues and the discourse surrounding hate speech, revealing how external events can shape public attitudes and responses.

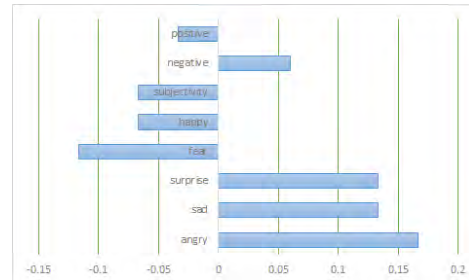


Figure 7. Relative comparison of the intensity of the causal effect of hatred on emotions.

Table 3. Comparison of Our Work with State-of-the-Art (SOTA) Findings from [10]

Common Feature	Comparison	Our Finding	Findings by [10]
Subjectivity (equivalent to complement of "fact orientation" feature in [10])	Aligned	Causally decreased	Decreased
Anger	Aligned	Causally increased	Increased
Negative	Aligned	Causally increased	Increased
Fear	Not-aligned	Causally decreased	No significant difference
Sadness	Not-aligned	Causally increased	No significant difference
Surprise	Not-aligned	Causally increased	No significant difference

Moreover, our experiments on causal effect estimation into user emotions following hate speech indicate a significant increase in feelings of surprise, alongside a rise in negativity and a decline in fear and subjectivity. The tendency for users to express more subjective opinions rather than factual information post-hate speech raises critical questions about the nature of online discourse and the potential for misinformation to proliferate in emotionally charged environments. The correlation analysis further supports these findings, demonstrating that hate speech is associated with increased sadness and decreased positivity, thus reinforcing the detrimental effects of hate speech on emotional well-being.

In summary, our study contributes to the growing body of literature on hate speech by providing a robust causal framework for understanding its effects on user behavior. The methodological advancements we introduced, including our matching strategy for identifying control users and our focus on temporal dynamics, offer valuable tools for future research in this area.

7. Future work

Future research can build upon our findings by exploring several key areas: 1) deeper exploration of the user post such as topics expressed by users to examine the specific topics users discuss in relation to hate speech. By identifying prevalent themes within user-generated content, we can gain deeper insights into the contexts in which hate speech arises and how it intersects with broader societal discussions.

2) Expanding our analysis to include multiple social media platforms will provide a more comprehensive understanding of hate speech dynamics. Different platforms may exhibit varying user behaviors and emotional responses, influenced by their unique cultures and community guidelines. Comparative studies can reveal how platform-specific factors shape the propagation and impact of hate speech.

3) Investigating how users engage with counter-speech in response to hate speech can provide insights into the effectiveness of different forms of resistance. Analyzing the characteristics of counter-speech and its impact on hate speech propagation will inform strategies for fostering positive discourse and resilience against hate.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

The authors' contribution of Fatemeh Pourgholamali and Akram Alam is as follows: FP: Study design, acquisition of data, conducting experiments, interpretation of the results, statistical analysis, drafting the manuscript;

AA: Study design, conducting experiments, drafting the manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

- [1] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression and Violent Behavior*, vol. 40, pp. 108–118, 2018. <https://doi.org/10.1016/j.avb.2018.05.003>
- [2] B. Lantz, M. R. Wenger, and J. M. Mills, "Fear, political legitimization, and racism: Examining anti-Asian xenophobia during the COVID-19 pandemic," *Race and Justice*, vol. 13, no. 1, pp. 80–104, 2023. <https://doi.org/10.1177/21533687221125817>
- [3] K. Erjavec and M. P. Kovačić, "You don't understand, this is a new war! Analysis of hate speech in news websites' comments," *Mass Communication and Society*, vol. 15, no. 6, pp. 899–920, 2012. <https://doi.org/10.1080/15205436.2011.619679>
- [4] S. Schäfer, M. Sülfow, and L. Reiners, "Hate speech as an indicator for the state of society," *Journal of Media Psychology*, vol. 30, no. 4, pp. 1–10, 2020. <https://doi.org/10.1027/1864-1105/a000294>
- [5] H. E. Tajfel, *Differentiation between social groups: Studies in the social psychology of intergroup relations*, Academic Press, 1978.
- [6] C. Calvert, "Hate speech and its harms: A communication theory perspective," *Journal of Communication*, vol. 47, no. 1, pp. 4–19, 1997. <https://doi.org/10.1111/j.1460-2466.1997.tb02690.x>
- [7] J. S. Vedeler, T. Olsen, and J. Eriksen, "Hate speech harms: A social justice discussion of disabled Norwegians' experiences," *Disability & Society*, vol. 34, no. 3, pp. 368–383, 2019. <https://doi.org/10.1080/09687599.2018.1515723>
- [8] M. Römer-Pieretti, E. Said-Hung, and J. Montero-Díaz, "Semiotic analysis of hate discourse in Spanish digital news media: Biden's inauguration case study," *Social Inclusion*, vol. 13, 2025. <https://doi.org/10.17645/si.9295>
- [9] S. Masud *et al.*, "Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on Twitter," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 504–515. <https://doi.org/10.1109/ICDE51399.2021.00050>
- [10] Z. Noorian, A. Ghenai, H. Moradisani, F. Zarrinkalam, and S. Zamani Alavijeh, "User-centric modeling of online hate through the lens of psycholinguistic patterns and behaviors in social media," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 4354–4366, June 2024. <https://doi.org/10.1109/TCSS.2024.3359010>
- [11] H. Yazdavar *et al.*, "Multimodal mental health analysis in social media," *PLoS ONE*, vol. 15, no. 4, e0226248, 2020. <https://doi.org/10.1371/journal.pone.0226248>
- [12] L. Sinnenberg *et al.*, "Twitter as a tool for health research: A systematic review," *American Journal of Public Health*, vol. 107, no. 1, pp. e1–e8, 2017. <https://doi.org/10.2105/AJPH.2016.303512>
- [13] S. Chancellor *et al.*, "Quantifying and predicting mental illness severity in online pro-eating disorder communities," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2016, pp. 1171–1184. <https://doi.org/10.1145/2818048.2819973>
- [14] M. De Choudhury *et al.*, "Predicting depression via social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 2013, pp. 128–137. <https://doi.org/10.1609/icwsm.v7i1.14432>
- [15] E. Kadkhoda *et al.*, "Bipolar disorder detection over social media," *Informatics in Medicine Unlocked*, vol. 32, 2022, 101042. <https://doi.org/10.1016/j.imu.2022.101042>
- [16] F. A. Nawaz *et al.*, "Social media use among adolescents with eating disorders: A double-edged sword," *Frontiers in Psychiatry*, vol. 15, 2024, 1300182. <https://doi.org/10.3389/fpsy.2024.1300182>
- [17] Kumar and K. R. Nayar, "COVID-19 and its mental health consequences," *Journal of Mental Health*, vol. 30, no. 1, pp. 1–2, 2021. <https://doi.org/10.1080/09638237.2020.1757052>
- [18] Pfefferbaum and C. S. North, "Mental health and the COVID-19 pandemic," *New England Journal of Medicine*, vol. 383, no. 6, pp. 510–512, 2020. <https://doi.org/10.1056/NEJMp2008017>
- [19] Talevi *et al.*, "Mental health outcomes of the COVID-19 pandemic," *Rivista di Psichiatria*, vol. 55, no. 3, pp. 137–144, 2020. <http://doi.org/10.1708/3382.33569>

- [20] X. Wang *et al.*, "Investigating mental health of US college students during the COVID-19 pandemic: Cross-sectional survey study," *Journal of Medical Internet Research*, vol. 22, no. 9, e22817, 2020. <https://doi.org/10.2196/22817>
- [21] N. Schwarz, "Self-reports: How the questions shape the answers," *American Psychologist*, vol. 54, no. 2, pp. 93-105, 1999. <https://doi.org/10.1037/0003-066X.54.2.93>
- [22] X. Fan *et al.*, "An exploratory study about inaccuracy and invalidity in adolescent self-report surveys," *Field Methods*, vol. 18, no. 3, pp. 223-244, 2006. <https://doi.org/10.1177/152822X06289161>
- [23] Y. Mejova and V. Suarez-Lledó, "Impact of online health awareness campaign: Case of national eating disorders association," in *International Conference on Social Informatics*, Springer, 2020, pp. 192-205. https://doi.org/10.1007/978-3-030-60975-7_15
- [24] K. Sarikakis *et al.*, "My haters and I: Personal and political responses to hate speech against female journalists in Austria," *Feminist Media Studies*, vol. 23, no. 1, pp. 67-82, 2023. <https://doi.org/10.1080/14680777.2021.1979068>
- [25] M. Bozhidarova *et al.*, "Hate speech and hate crimes: A data-driven study of evolving discourse around marginalized groups," in *IEEE International Conference on Big Data (BigData)*, Sorrento, Italy, 2023, pp. 3107-3116. <https://doi.org/10.1109/BigData59044.2023.10386312>
- [26] B. Mathew *et al.*, "Hate begets hate: A temporal study of hate speech." *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, 2020, pp. 1-24. <https://doi.org/10.1145/3415163>
- [27] B. Kuřík, M. Heřmanová, and J. Charvát, "Living hated: Everyday experiences of hate speech across online and offline contexts," *Communications*, vol. 49, no. 3, pp. 378-399, 2024. <https://doi.org/10.1515/commun-2023-0110>
- [28] J. An, H., Kwak, C. S., Lee, B., Jun, and Y. Y. Ahn, "Predicting anti-Asian hateful users on Twitter during COVID-19," *arXiv preprint arXiv:2109.07296*, 2021. <https://doi.org/10.48550/arXiv.2109.07296>
- [29] B. He *et al.*, "Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 90-94. <https://doi.org/10.1145/3487351.3488324>
- [30] R. Moraffah *et al.*, "Causal inference for time series analysis: Problems, methods and evaluation," *Knowledge and Information Systems*, vol. 63, pp. 3041-3085, 2021. <https://doi.org/10.1007/s10115-021-01621-0>
- [31] S. Athey and G. W. Imbens, "The state of applied econometrics: Causality and policy evaluation," *Economic Perspectives*, vol. 31, no. 2, pp. 3-32, 2017. <https://doi.org/10.1257/jep.31.2.3>
- [32] J. D. Angrist and J. Steen Pischke, *Mastering 'metrics: The path from cause to effect*, Princeton University Press, 2014.
- [33] M. Bertrand, E. Duo, and S. Mullainathan, "How much should we trust differences-in-differences estimates," *The Quarterly Journal of Economics*, vol. 119, no. 1, pp. 249-275, 2004. <https://doi.org/10.1162/003355304772839588>
- [34] K. H. Brodersen *et al.*, "Inferring causal impact using Bayesian structural time-series models," *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247-274, 2015. <https://doi.org/10.1214/14-AOAS788>
- [35] R. González and E. B. Hosoda, "Environmental impact of aircraft emissions and aviation fuel tax in Japan," *Journal of Air Transport Management*, vol. 57, pp. 234-240, 2016. <https://doi.org/10.1016/j.jairtraman.2016.08.006>
- [36] F. de Vocht, "Inferring the 1985-2014 impact of mobile phone use on selected brain cancer subtypes using Bayesian structural time series and synthetic controls," *Environment International*, vol. 97, pp. 100-107, 2016. <https://doi.org/10.1016/j.envint.2016.10.019>
- [37] F. de Vocht *et al.*, "The intervention effect of local alcohol licensing policies on hospital admission and crime: A natural experiment using a novel Bayesian synthetic time-series method," *Journal of Epidemiology and Community Health*, vol. 71, no. 9, pp. 912-918, 2017. <https://doi.org/10.1136/jech-2017-208931>
- [38] P. Samartsidis, S. R. Seaman, A. M. Presanis, M. Hickman and D. De Angelis, "Assessing the causal effect of binary interventions from observational panel data with few treated units," *Statistical Science*, vol. 34, no. 3, pp. 486-503, 2019. <https://doi.org/10.1214/19-STS713>
- [39] N. Aslam, F. Rustam, E. Lee, P. B. Washington and I. Ashraf, "Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model," *IEEE Access*, vol. 10, pp. 39313-39324, 2022. <https://doi.org/10.1109/ACCESS.2022.3165621>
- [40] H. Lee, "2019-20 Hong Kong protests: Storytelling through the best and worst times," *Eyeopener*. [Online]. Available: <https://theeyeopener.com/2020/10/2019-20-hong-kong-protests-storytelling-through-the-best-and-worst-times/>.
- [41] T. Y. Wang, "Hong Kong and the 2019 anti-extradition bill movement," *Journal of Asian and African Studies*, vol. 58, pp. 3-7, 2023. <https://doi.org/10.1177/00219096221124983>
- [42] T. Staff, "Hong Kong's political crisis deepens after the worst day of violence in decades," *TIME*. [Online]. Available: <https://time.com/5690681/hong-kong-crisis-unrest-protests/>
- [43] Frazier, S. Heng, and W. Zhou, "Bias Reduction in Matched Observational Studies with Continuous Treatments: Calipered Non-Bipartite Matching and Bias-Corrected Estimation and Inference," *arXiv preprint arXiv:2409.11701*, 2024. <https://doi.org/10.48550/arXiv.2409.11701>



Fatemeh Pourgholamali received a BS degree in computer engineering from the Shahid Bahonar University of Kerman, Iran, and MS and PhD degrees in computer engineering, software, from Ferdowsi University of Mashhad. Currently, she is an Assistant Professor at Vali-e-Asr University of Rafsanjan. Her research interests are mainly in Social network analysis, text mining, machine learning and Natural Language Processing



Akram Alam received the BS and MS degrees in computer engineering, software from the Misagh University of Rafsanjan. Her research interests include Social network analysis, text mining and machine learning.