

# Navigating Hate Speech: Bridging LLMs and Human Expertise in Public Officials' Online Communication

Nitheesha Nakka<sup>1,2</sup>, Isaac M. Pollert<sup>1,2</sup>, Lingyu Fuca<sup>1,2</sup>, Yuehong Cassandra Tai<sup>1,2</sup>

<sup>1</sup>Pennsylvania State University, Department of Political Science, University Park, PA, USA

<sup>2</sup>Pennsylvania State University, Center for Social Data Analytics, University Park, PA, USA  
nvn5240@psu.edu, imp5243@psu.edu, lingyu.fuca@psu.edu, yhcassai@psu.edu

## Abstract

The rise in hate speech targeting minority communities underscores the urgent need for effective tools to detect and address harmful content in digital communication. We examine over 3 million tweets posted by state legislators between 2020 and 2021, focusing on messages directed at Asian communities. To address the nuanced nature of hate speech, we develop three comprehensive definitions for identifying hate speech. With a human-in-the-loop approach, our fine-tuned BERT-NLI model achieved improved classification performance. We find that while anti-Asian tweets comprise only a small portion of legislators' total tweets, there are distinct geographic patterns across states. In addition, the frequency of posting pro-Asian or anti-Asian tweets is significantly influenced by legislators' demographics. Women and Democrats post more pro-Asian content while men and Republicans post more anti-Asian content. By combining advanced computational methods with human oversight, this study advances efforts to address sensitive issues in politicians' digital discourse with greater precision and accountability.

## Introduction

In March of 2021, eight people were murdered in a spa in Atlanta, Georgia by a lone White gunman. Of the eight killed, six were Asian women. After the COVID-19 pandemic, a surge in hate speech and discriminatory acts against Asians set the stage for this devastating massacre. This incident, among others, led to the establishment of the #StopAsian-Hate movement, and was emblematic of broader trends in the American public sphere. The Atlanta Spa murders were not just outlier events. The Center for the Study of Hate and Extremism (CSUSB) reported that hate crimes targeted at Asians were up 149% in 2020, despite overall crime rates dropping by 7 points. This type of targeted harassment also persists into the online space. The CSUSB also reports that sinophobic slurs on X (formerly Twitter, and hereafter, Twitter) more than doubled in March of 2020 alone.

The spread of false information and offensive language on social networking platforms poses a significant threat to democracy in the United States (Osmundsen et al. 2021;

Siegel et al. 2021; Tai, Buma, and Desmarais 2023). However, tackling hate speech is challenging due to detection difficulties, especially in text where hate-related keywords are not explicitly included (Davidson et al. 2017; Chung et al. 2019; Ziems et al. 2024). This challenge is further compounded when public officials engage in hate speech, as their rhetoric is more nuanced.

Our study seeks to better understand how state lawmakers' social media activity overlaps with hate speech. State legislators constitute a substantial yet overlooked group among elected officials despite their crucial functions in the American political system. Their various responsibilities include formulating election administration regulations (Hajnal, Kuk, and Lajevardi 2018; Shino, Suttman-Lea, and Smith 2022), and mandating public health strategies (Diekema 2014) or overseeing abortion regulations (Kreitzer 2015; Pollert and Mooney 2022). While the significant impact of political elites on public opinion has been extensively studied (Berinsky 2017; Slothuus and Bisgaard 2021), our understanding of the online communication of state legislators remains limited, apart from a few recent investigations (Kim et al. 2021; Cook 2017; Payson et al. 2022; Tai, Buma, and Desmarais 2023; Butler, Kousser, and Oklobdzija 2023; Gopal et al. 2024). To bridge this gap, we examine the language utilized by state legislators on Twitter, focusing specifically on their communication regarding the Asian community. Specifically, we address the following research questions:

**RQ1:** What is the prevalence of pro-Asian and anti-Asian speech in state legislators' tweets?

**RQ2:** Do pro-Asian tweets or anti-Asian tweets receive more engagement online?

**RQ3:** How do legislators' demographics, including party, race and gender, influence their sharing of pro-Asian and anti-Asian speech?

To answer these questions, we explore a more comprehensive definition of online hate speech. In doing so we identify three parameters by which we measure online sentiment directed towards Asian minority populations during a particularly volatile and racially charged time in American history, they are as follows: anti-Asian, pro-Asian, and neutral. Based on 3,345,232 tweets collected from 3,580 state legislators across 50 states in the United States between January

1, 2020, and December 31, 2021, we first filtered out 25,102 Asian-related tweets posted by 2,004 lawmakers. The fluctuation of Asian-related tweets followed salient issues, like the introduction of the term “Chinese Virus” (ABC News 2021), the House committee hearing on the increase in anti-Asian American crimes (CNBC 2021), and the tragic Atlanta shootings (KXAN News 2021).

Using our hand-coded samples, we then fine-tuned a large language model and incorporated a human-in-the-loop approach, which improved classification performance. With this model, we detected 5,964 pro-Asian tweets and 1,305 anti-Asian tweets, revealing distinct patterns by party, gender, race, and across all 50 U.S. states. Our analysis of engagement metrics highlights an important distinction between content re-sharing and visibility. The difference in Retweets and Likes between pro- and anti-Asian speech suggests that, while supportive content may resonate more positively with users, harmful content tends to spread more widely. These divergent patterns imply that antagonistic messages can achieve disproportionate reach, even when there is strong normative support for marginalized groups. This finding highlights the need for interventions that not only address harmful speech but also encourage positive support behaviors, such as amplifying pro-Asian content. Our statistical results further reveal that women and Democratic legislators post more pro-Asian tweets, while men and Republican legislators post more anti-Asian tweets. Asian American legislators, in particular, share more pro-Asian content than any other racial group and, along with other non-White legislators, are more likely to use their platforms to counter anti-Asian rhetoric.

Our study is the first to analyze hate (i.e., anti-Asian) and counter-hate (i.e., pro-Asian) speech by state legislators, who represent a large population of elected officials and occupy more prominent positions than average citizens in the U.S. Our research opens a new avenue for studying online harmful speech and lay the groundwork for future research on mechanisms and intervention strategies. Methodologically, we suggest that for high-stakes tasks, like detecting hate speech by politicians, incorporating human oversight is essential. Accurate detection not only supports downstream analysis but also deepens our understanding of the complex dynamics between identity, ideology, and online discourse.

The remainder of the paper is organized as follows. First, we provide background on online political messaging and prior work on detecting hate speech. We then present our data and methods, followed by the results for each of our three research questions. Based on our findings, we also provide an additional descriptive analysis of speech patterns by party. We conclude with a discussion of our results and limitations.

## Background

### Legislators and Online Messaging

Twitter is a key platform for legislators to voice their beliefs, including stances on Asian minorities, and to market themselves effectively. Tweets enable lawmakers to broadcast their views to large audiences, fostering a feedback loop be-

tween constituents and legislators. In recent years, lawmakers are especially reactive to racial issues, specifically during the COVID-19 pandemic (Abascal, Makovi, and Xu 2023). Influenced by the President’s rhetoric during this time, many politicians have adopted more aggressive language online (Russell 2018). Importantly, hate speech—particularly from public officials—has been shown to correlate with increases in hate crimes (Müller and Schwarz 2020). The increased racial tensions from 2016 to 2020 amplified long-standing issues of resentment (Enders and Thornton 2022). Scholars have examined the evolving role of Asians in America’s racial hierarchy, where the “model minority” trope contrasts with perceptions of foreignness (Hua and Junn 2021; Zou and Cheryan 2017). Polarization exacerbates this dynamic, with explicit hostility from one party pushing many Asians to align with the left (Pei and Mehta 2020; Abascal, Makovi, and Xu 2023; Chan, Kim, and Leung 2022).

Social media use is a hallmark of modern political careers. Twitter, in particular, is a popular platform among state legislators who do not have access to more professionalized media coverage from major news outlets. Part of Twitter’s success as a campaign tool is attributed to the communication style that is unique to social media (Kruikemeier 2014; Vergeer 2015). Twitter rhetoric tends to be more personable and allows politicians to appear more approachable, which is a proven successful campaign strategy (Gaynor and Gimpel 2023). Twitter also gives legislators greater control over their branding. Alongside touting the party bylines, politicians must also frame their rhetoric in a way that resonates with the beliefs of their constituents. State legislators can receive direct public feedback, making the platform particularly conducive to this type of responsive communication.

### Related Works

Given the significance of political rhetoric and its societal impact, detecting hate speech by public officials is both critical and challenging. Despite advances in traditional and deep learning techniques, hate speech remains a challenging topic to detect (Davidson et al. 2017; Founta et al. 2018; Ziems et al. 2024). This difficulty originates from a lack of consensus on definitions (Toliat et al. 2022; Siegel et al. 2021), low intercoder reliability (He et al. 2021; Vidgen et al. 2020), and the substandard writing that is prevalent on social media (Li and Ning 2022). While state-of-the-art deep learning models like BERT outperform traditional machine learning algorithms in hate speech detection, they frequently fail to capture the contextual nuances of conversational dynamics and events (Govers et al. 2023). Generative AI shows potential in advancing hate speech detection (Li et al. 2024), but these models still struggle with accuracy and low to medium agreements with human coders. For example, a comparative evaluation of 14 large language models for hate speech detection found that the highest macro-F1 score was just 32.3, with an intercoder reliability of 0.2 (Ziems et al. 2024).

These challenges are further compounded in political texts authored by public officials, where rhetoric tends to be more nuanced, sophisticated, and context-sensitive. To address this gap, we fine-tuned a classifier tailored to identify

subtle patterns of hate speech in political discourse.

## Data

We constructed our dataset based on original data collected and presented by Tai et al (Tai et al. 2024). They collect 3,345,232 tweets from every state in the United States between January 1, 2020, and December 31, 2021, utilizing the Academic Twitter API with the R package `academictwitter` (Barrie and Ho 2021). Out of the 7,992 state representatives and senators serving during the period of data collection, 4,641 lawmakers had X accounts and 3,580 made posts. These 3,580 legislators are included in our multilevel regression analysis. This includes 1884 Democrats, 16 Independents and 1680 Republicans. It is important to acknowledge that a significant number of Twitter accounts may have been inactive during the collection period or inaccessible after legislators left office.

Since our interest is tweets targeting Asian minority groups, we used a dictionary of 191 which includes the following: 1) words naming Asian countries and subgroups of Asian peoples; 2) racially motivated insults and slurs 3) as well as racially motivated activism in support of the Asian communities- to filter the tweets data (see Appendix for examples of keywords). Specifically, this dictionary combines a list of hate-related keywords (He et al. 2021) and a list of negative language patterns used against Asian people included in The Weaponized Word dataset. The Weaponized Word lexicon is one of the largest and most dynamic online harassment-related dictionaries available, which we accessed through their API (<https://weaponizedword.org/>). This combination allows us to confidently capture more implicit or dog-whistle forms of hate speech against Asians alongside Asian related rhetoric generally. Filtering tweets based on keywords rendered 25,102 Asian related tweets.

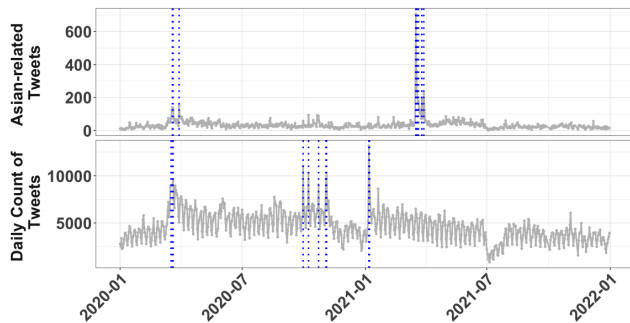


Figure 1: Trends of Tweets and Asian-Related Tweets by Day

To gain a deeper insight into the patterns of overall tweets and tweets related to Asia and Asian peoples, we first illustrate the daily number of total tweets and Asian-related tweets over a 2-year time frame in Figure 1. The bottom panel of Figure 1 reveals that the highest tweet volumes were observed in March 2020, November 2020, and January 2021, coinciding with the occurrences of the COVID-19 pandemic, the 2020 presidential election, and the Jan-

uary 6th insurrection. The top panel illustrates the variations in the volume of tweets related to Asia and Asian peoples, which generally stayed consistent but experienced occasional spikes in March 2020 when the President first began using the term “Chinese Virus” in his tweets (ABC News 2021), in March 2021 during a House committee hearing on the increase in crimes targeting Asian American (CNBC 2021), and after the tragic Atlanta Spa shootings in which six Asian American women were murdered (KXAN News 2021).

## Methods

### Defining Online Hate Speech

The definition of hate speech is a topic of ongoing debate. Scholars typically offer their own interpretations of hate speech (Toliat et al. 2022; Siegel et al. 2021). In addition, the distinctive language patterns used in social media- characterized by brevity and often substandard writing- add complexities to the task of grasping semantics and identifying distinguishing elements of hate speech. This difficulty persists despite the availability of supplementary resources like sentiment hashtags aimed at enhancing the effectiveness of hate speech detection (Li and Ning 2022). Since hate speech lacks a singular definition in a legal or academic sense, we incorporate three popular hate speech definitions taken from the United Nations, the American Library Association, and the Meta social media company. We develop our own definition of hate speech targeted toward Asian minorities, which we label as “anti-Asian” speech. We are also interested in speech in support of Asian minority groups and speech that is neutral toward Asian minority groups. We label these kinds of speech as “pro-Asian” and “neutral” respectively. Our definitions are as follows:

- **Anti-Asian:** Expressions with intent to harm that involve pejorative or discriminatory remarks targeting Asian populations based on identity factors such as race, ethnicity, gender, or culture. This includes spreading conspiratorial, false, or misleading information aimed at undermining or stigmatizing Asian groups, countries of origin or political systems.
- **Pro-Asian:** Expressions with intent to support and actively defend Asian communities against discrimination, debunk conspiracies, combat mis/disinformation with factual content, and raise awareness of issues affecting Asian populations. This language seeks to positively recognize and empower Asian groups in the face of anti-Asian rhetoric.
- **Neutral:** Expressions with no affective stance toward Asian populations, neither harmful nor explicitly supportive. This includes objective reporting on issues relevant to the Asian community, critiques of societal or political harms, or criticisms of governments without targeting Asian people or cultures.

## Natural Language Inference Model on Classification Task and Human-in-loop

To classify whether a tweet is anti-Asian, pro-Asian, or neutral, we fine-tuned a BERT-NLI model pre-trained on classification task developed by (Laurer et al. 2023). Taking advantage of deep transfer learning and “prior knowledge” accumulated in language models, this BERT-NLI model performs well on a universal task and adapts effectively to novel classification tasks (Laurer et al. 2024). Since we specified definitions for pro- and anti-Asian, and neutral speech in our study, we human-coded a sample of tweets to prepare the training data in the natural language inference task format with premise and hypothesis pairs.

In our human-coded sample, tweets classified as 1 are neutral towards Asians, and do not make an affectively charged stance in either direction or are a news article reporting on a related issue. Tweets coded 0 are those that contain anti-Asian speech and 2 are pro-Asian tweets. We trained two research assistants to annotate a random sample tweets according to our coding scheme. Over 12 weeks, the annotation phase involved three rounds of refinement to solidify the reliability of our coding scheme. Ultimately, we manually coded 1,586 tweets, identifying 134 as anti-Asian, 369 as pro-Asian, and 1,083 as neutral. The final round of coding achieved an intercoder reliability score (Krippendorff’s alpha) of 0.593. The diverse backgrounds and perspectives of annotators often lead to low intercoder reliability (Vidgen et al. 2020; Ross et al. 2017), particularly in cases involving subjects with ambiguous boundaries. In hate speech annotation, achieving only moderate agreement, typically around 0.6, is common, as demonstrated in prior studies (He et al. 2021). Given the inherent difficulty in detecting hate speech, we trained our classifiers on the annotated data while adopting a conservative classification approach by setting a higher probability threshold of 0.7. Specifically, we only assigned a labeled class if the classifier’s predicted probability exceeded 0.7. For cases where the predicted probability fell below this threshold, we classified them as neutral.

Table 1: Performance Metric

Model	Precision	Recall	F1 score	Weighted Accuracy	MCC
Zero-Shot	0.49	0.50	0.50	0.68	0.35
Fine-tuned	0.78	0.77	0.78	0.85	0.69
Human-in-loop	0.78	0.77	0.77	0.88	0.75

Table 2: Performance Metric of Individual Classes in Human-in-loop Model

Category	Precision	Recall	F1 score
Anti-Asian Speech	0.57	0.48	0.52
Pro-Asian Speech	0.86	0.90	0.88
Neutral Speech	0.92	0.93	0.92

To address the imbalance in our data, we used all 134

anti-Asian tweets and randomly sampled 134 pro-Asian and neutral tweets during the training process. To mitigate concerns about the moderate intercoder reliability, we adopted a human-in-the-loop approach after the first round of training. From the coded data derived from the first round fine-tuning, we sampled 150 instances across all probability ranges, focusing on those with probabilities between 0.45 and 0.65. A researcher with subject area expertise then recoded these samples to optimize model performance. This new sample remained imbalanced, with 10 anti-Asian, 14 pro-Asian, and 126 neutral tweets. To increase diversity in the data, we prompted GPT-4o to generate synthetic anti- and pro-Asian tweets via a few-shot learning approach based on our samples. Each generated tweet was manually reviewed to ensure it matched the intended sentiment. After filtering out tweets that misinterpreted or missed key components, we had 40 anti-Asian and 43 pro-Asian synthetic tweets, which were added to our training data to further fine-tune the model.

For evaluation, we used the Matthews Correlation Coefficient (MCC) as our primary performance metric, as it provides a balanced assessment by considering all elements of the confusion matrix and is less sensitive to class imbalance. Table 1 shows that our fine-tuned model and human-in-the-loop approach significantly improved performance. The large performance gap between the pre-trained BERT-NLI model and our fine-tuned models is likely due to two factors. First, our definition of anti-Asian tweets excludes criticisms of Asian governments but includes disinformation-based criticisms and harmful conspiracies, improving classification precision. Second, our dataset focuses on text from politicians, whose language tends to be more nuanced and complex than that of typical X users, potentially making such narratives underrepresented in the training dataset used by (Laurer et al. 2023). With an MCC of 0.75, our human-in-the-loop model demonstrates strong predictive capability, and we used this further fine-tuned model to classify the remaining tweets in our dataset.

As shown in Table 2, the classifier struggles to detect anti-Asian speech. To address the concern, we manually reviewed all tweets labeled as anti-Asian. Through our manual scrutiny in 1,305 anti-Asian labeled posts, we had 816 confirmed anti-Asian tweets, 40 pro-Asian tweets, and 449 neutral tweets. We re-categorized tweets that targeted the Asian government and were originally given the anti-Asian label as neutral. Sarcastic tweets were also often mislabeled as anti-Asian tweets, which presents a challenge in stance detection tasks (Chen et al. 2024). For example, a tweet posted by Senator Marko Liias from Washington State reads: “Oooo, the China virus has become the China plague! If only they spent more time saving lives than renaming #COVID19.” This was mistakenly labeled as anti-Asian speech despite its sarcastic tone. Keeping track of these nuances in our hand-validated anti-Asian labels enhanced the reliability of our data.

## Results

### Pro- and Anti-Asian Tweets Distributions—RQ1

Through our fine-tuned model and hand-coded verified anti-Asian tweets, we identified 816 anti-Asian tweets (3.2% of

Asian-related tweets, 0.02% of total tweets) and 6,004 pro-Asian tweets (23.9% of Asian-related tweets, 0.18% of total tweets). Despite the relatively small number of tweets, there are distinct geographic patterns as shown in Figures 2 and 3. When comparing the distributions of pro-Asian and anti-Asian tweets with the distribution of Asian populations (seen in Figure 4) we see that regions with historically large Asian populations, like California, show high pro-Asian sentiment and lower anti-Asian sentiment. Conversely, areas with smaller Asian populations (e.g., Arizona, Arkansas) display the opposite trend. Notably, Texas presents a more nuanced picture. Despite its sizable Asian population, the prevalence of anti-Asian sentiment remains evident. This can be attributed, in part, to the state's political landscape, as a Republican stronghold. As subsequent analysis will show, political affiliation plays a significant role in shaping Asian sentiment.

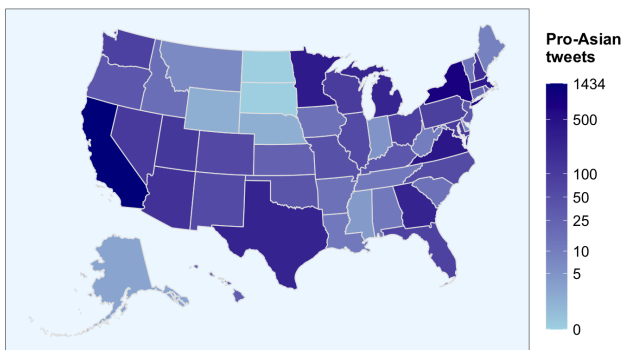


Figure 2: Number of Pro-Asian Tweets per State

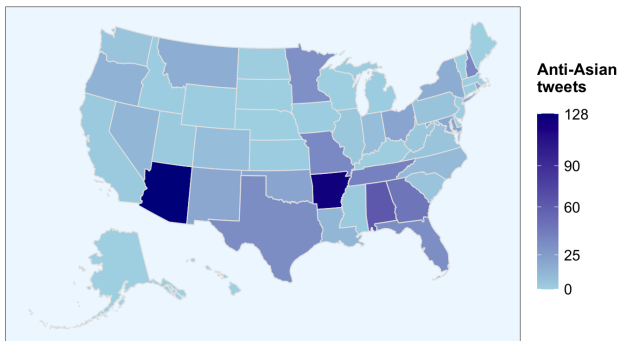


Figure 3: Number of Anti-Asian Tweets per State

### Engagement Comparison–RQ2

We then examined online interactions with anti-Asian, pro-Asian, and neutral speech-types, specifically comparing the number of Likes and Retweets. First, we used pairwise Wilcoxon rank-sum tests to compare the distribution of Likes across each speech-type, given the strong skew and presence of outliers in the data. P-values were adjusted using the Bonferroni correction to account for multiple com-

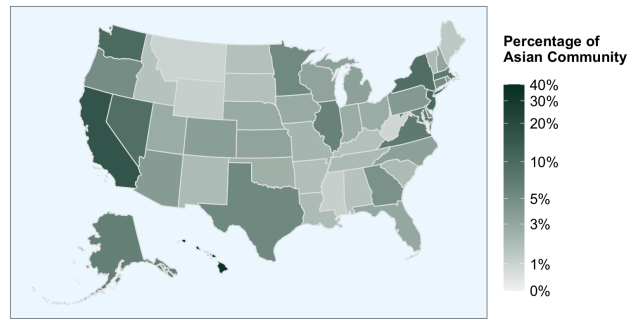


Figure 4: Number of Asian Communities across 50 States

parisons. Anti-Asian speech (median = 0, Q3 = 1) received significantly fewer Likes in the distribution than neutral (median = 0, Q3 = 3,  $p < 0.001$ ) or pro-Asian (median = 1, Q3 = 13,  $p < 0.0001$ ). Pro-Asian speech received more Likes in the distribution than neutral speech. Figure 5 highlights these differences. Pro-Asian tweets receive the highest overall number of Likes, reflecting stronger approval or support for pro-Asian content compared to anti-Asian or neutral tweets. Anti-Asian tweets have a low median number of Likes, similar to neutral tweets, which shows minimal engagement overall.

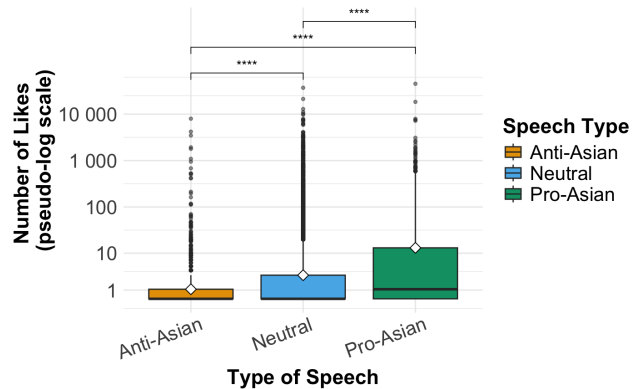


Figure 5: Distribution of Likes by Speech Type

Using the same pairwise Wilcoxon rank-sum test with Bonferroni correction, the distribution of Retweets shows the opposite pattern. Anti-Asian speech (median = 77, Q3 = 3,252) received significantly more Retweets in the distribution than neutral (median = 6, Q3 = 187,  $p < 0.001$ ) or pro-Asian (median = 6, Q3 = 21,  $p < 0.0001$ ). Figure 6 showcases this further. A legislator might Retweet a post containing anti-Asian messaging to disparage the original tweet or promote it. But, regardless of the reasoning behind the Retweet, the significantly higher number of Retweets for anti-Asian content indicates the wide distribution of this harmful content. This finding emphasizes the necessity for accurately detecting harmful speech in political discourse since Retweets can drastically increase the visibility of abusive posts (Biswas et al. 2024).

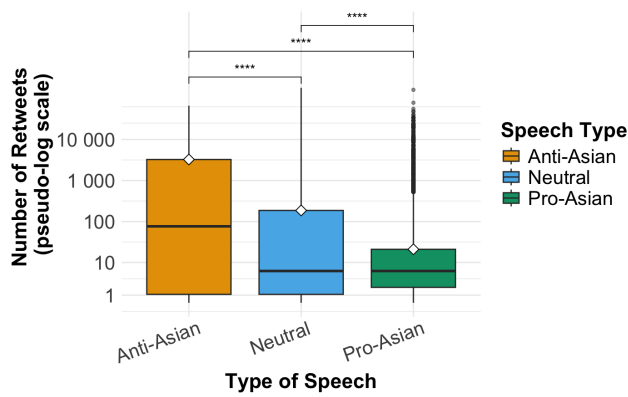


Figure 6: Distribution of Retweets by Speech Type

When examining engagement patterns overall, neutral speech consistently attracts the least engagement across both metrics. Both anti-Asian and pro-Asian speech show broader variability in engagement, with significant statistical differences in Likes and Retweets, highlighting polarized audience responses. When comparing the distribution of Retweets to the distribution of Likes anti-Asian speech tends to garner more Retweets than Likes, indicating higher amplification through sharing. Though the motivation behind sharing an anti-Asian post can vary, the resulting widespread distribution of anti-Asian content remains. It is illuminating that pro-Asian messaging does not receive the same level of distribution (via Retweets) on social media. This is in line with prior research that finds that negative political messaging and harmful attacks receive more Retweets that messaging without attacks (Fine and Hunt 2023). Pro-Asian speech receives more Likes overall. The meaning behind a Like is more definitive, suggesting stronger approval or support of pro-Asian content. Taken together, these findings suggest that although supportive content may resonate more positively with users, harmful content tends to spread more widely—potentially driven by mechanisms such as algorithmic amplification, or polarizing appeal. These divergent patterns imply a critical distinction between endorsement and visibility: content that aligns with normative support may be appreciated, but antagonistic or harmful messages can still achieve disproportionate reach, posing broader risks for public discourse. Ultimately, we find that speech type influences how users interact with content on social media, with anti-Asian speech being amplified more through Retweets and pro-Asian speech receiving more approval through likes.

### Multilevel Regression Results—RQ3

To systematically investigate these patterns, we conducted a multilevel negative binomial regression to account for the over-dispersed nature of the dataset and the prevalence of zeros. We included a random intercept at the state level to control for unobserved heterogeneity across states and controlled for the total number of Asian-related tweets and legislators' overall tweet counts. Figure 7 presents the coeffi-

cient estimates for two models predicting the count of pro-Asian speech (i.e., green circles) and anti-Asian speech (i.e., blue, triangles) among legislators. The y-axis depicts variables representing demographics (party, race, gender) and tweet behavior (number of posts about Asian-related topics and total tweets). The reference categories for the models are White (for race), Male (for gender), and Democrat (for party), meaning all results should be interpreted in comparison to these groups. The x-axis represents the coefficient estimates, with positive values indicating a greater likelihood of engaging in each respective speech type, and negative values indicate a lower likelihood.

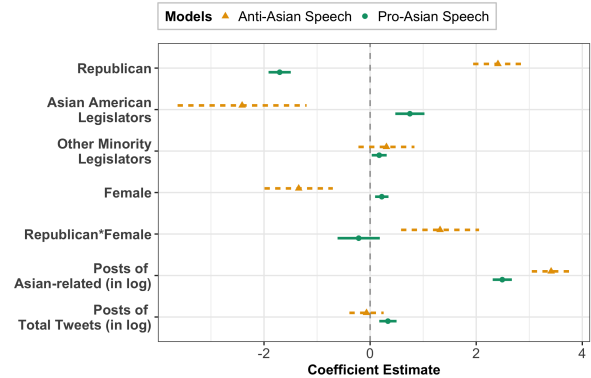


Figure 7: Predicting Pro and Anti Speech by Party, Gender and Race

**Party** A clear partisan divide emerges in the results. Compared to Democrats, Republican legislators are significantly more likely than Democrats to engage in anti-Asian speech, as indicated by their strong positive coefficient. Conversely, Republicans are less likely to engage in pro-Asian speech. We do not find a statistically significant effect from Independent legislators on sharing either pro or anti Asian posts. The results for Independent legislators contained large standard errors and is not presented in Figure 7 due in part to the small sample size ( $n = 16$ ) and the fact that none of the Independent legislators posted anti-Asian tweets.

These results are in line with expectations given the polarized nature of the subject and the salience bump it received from the President's direct role in framing Asian speech during this period. Both parties have been shown to selectively assign blame along partisan lines (Graham and Singh 2024), which would suggest the diametrically opposed effects that we see here. Republican narratives focused on China as the locus of hate, which had downstream effects on general Asian sentiment, intentional or not. This targeted approach to hate is commonplace within these sub circles, even more so when they occur in online spaces (Hobbs et al. 2024).

**Race** Racial identity also plays an important role in shaping speech patterns. Asian American legislators are less likely to engage in anti-Asian speech and more likely to engage in pro-Asian speech, compared to their White colleagues. Other Minority lawmakers also show a higher likelihood of pro-Asian speech than White legislators. However,

Minority Legislators do not exhibit a statistically significant difference in the likelihood of sharing anti-Asian speech.

These results suggest that non-White legislators, particularly Asian Americans, are more likely to use their platforms to counter anti-Asian rhetoric and advocate for their communities. However, the mixed effects regarding anti-Asian speech do raise questions related to the relationship between lawmaker race and policy positions when we move beyond the black-white dimension that tends to frame these discussions.

**Gender** Gender differences are also statistically significant. The female variable has a  $-1.35$  coefficient in anti-Asian speech models, and the pro-Asian speech models for the same variable show a  $.22$  coefficient. Compared to male lawmakers, female legislators are less likely to engage in anti-Asian speech. We discuss these results in the context of the literature below. Female participation in pro-Asian speech is positive and statistically significant. A clear pattern emerges in the underlying logic of salient Asian speech during the pandemic period and how it relates to the gender of lawmakers on social media. It is important to note that of the legislators included in this analysis, none identified as non-binary at the time of this study, and to the best of our knowledge.

The interaction of race and gender creates novel experiences in politics (Brown 2014), and this relationship is not additive. This results in potentially unexplored arenas of legislative behavior in a typically heterogeneous pool of officials. Female lawmakers, specifically non-white female lawmakers, are also often leaders in race-gender issues (Reingold, Widner, and Harmon 2020). This may explain the differences in how female and non-Asian, non-White, male legislators approach these issues. Both types of lawmakers fall below the top rungs of the traditional dominance hierarchy that, here, forwards anti-Asian speech, though they sit at different levels. Something about gender, evidently, makes this more prominent. This could also arise from other hypotheses about the above average capacity of female lawmakers over their male counterparts (Anzia and Berry 2011). The innate femaleness of the lawmaker increases their odds of legislating effectively, which in turn would suggest they would speak in defense of their constituents, even if only indirectly as a surrogate representative (Mansbridge 2003; Angevine 2017). Their male contemporaries here are less likely to behave in a similar manner. The positive results for pro-Asian speech suggest some may see a role for themselves in speaking in defense of constituents, surrogate or otherwise. The lack of effect for anti-Asian speech requires an explanation that is less clear. Whatever it is that makes male lawmakers more susceptible to anti-Asian speech may require evidence beyond what our work can demonstrate.

**Gender X Party** A third identity could moderate the above relationships. Partisanship, here used as a group signifier (Ruckelshaus 2022; West and Iyengar 2022), could also vary digital behavior based on gender. Our results here also demonstrate this, as Republican females have a positive and statistically significant likelihood of tweeting anti-Asian speech. Additionally, the positive odds for tweeting

pro-Asian speech predicted by the main effect of gender are not repeated. Republican female pro-Asian speech odds are negative without statistical significance. It would appear that some element of Republicanism is more dominant than gender in these interactions. Other work dealing with elite and conservative attitudes during this era (Broockman and Skovron 2018; Hopkins and Noel 2022; Filindra and Harbridge-Yong 2022) echoes this sentiment.

**Tweet Count** Finally, a legislator's tweeting behavior appears to shape their speech patterns. Those who tweet more frequently about Asian-related topics are likely to engage in both pro-Asian and anti-Asian speech, with a slightly higher likelihood to post pro-Asian tweets. This suggests that greater engagement with Asian issues is linked to both supportive and harmful rhetoric. The total number of tweets a legislator posts is also positively associated with pro-Asian speech, probably because Democratic legislators usually share much more tweets.

Overall, these findings highlight that partisanship, race, gender, and social media behavior all play a role in shaping how legislators engage with anti-Asian and pro-Asian rhetoric. These patterns pave the way for further research into the geographic, ideological, and demographic factors that drive political elites' engagement with pro- or anti-Asian rhetoric. The code needed to replicate this study can be found here: <https://osf.io/wxmd6/>.

## Party and Speech Type

Considering the striking partisan differences in anti-Asian and pro-Asian content we conduct additional descriptive analyses on the partisan speech patterns. Figures 8 and 9 display the terms that differentiate the content of tweets posted by Republican versus Democratic Legislators for anti-Asian and pro-Asian speech, respectively. To do this, we employ the "Fightin' Words" approach (Monroe, Colaresi, and Quinn 2008). The x-axis for both Figures 8 and 9 represents how frequently a term appears within a group, and the y-axis reflects the degree to which a term is associated with a group.

Regarding anti-Asian speech, Democrats who have tweeted anti-Asian content frequently use words like "virus" alongside "China". This can be seen in the following example tweet which reads "Good morning to everyone except FANG, China, whomever killed Malcom X, those of you that eat oatmeal and most Congressional Republicans." The anti-Asian rhetoric among Republicans is even more definitive with keywords like "threat" and "media" reflecting the larger sentiment that views China as a political and economic rival. This is seen in the following example of an anti-Asian Republican tweet that Republican legislators Retweeted to their own account, "RT @SenTomCotton: China is our adversary. China has been ripping us off, stealing our jobs, taking our factories, and threatening our allies". This is popular sentiment among Republicans tweeting about China in relation to the term "threat". These sentiments perpetuate an aggravated sense of danger around China and the Chinese people and are often not supported with any evidence.

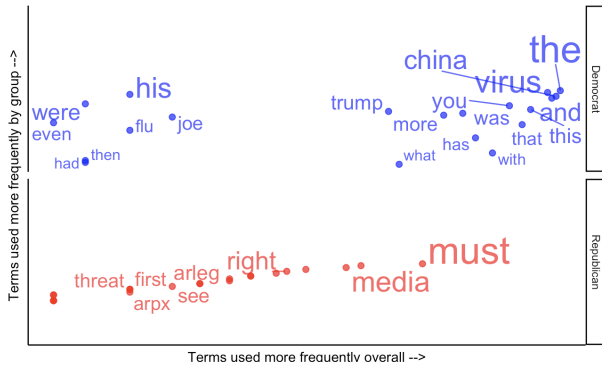


Figure 8: Anti-Asian Discourse Comparison: Democrats vs Republicans

There are also distinctions in pro-Asian speech between Democrats and Republicans. In Figure 9 we see that terms like “AAPI”, “march” and “justice” are highly prevalent in Democrats pro-Asian content. This coincides with the Stop Asian Hate movement that took place during this time. This sentiment is also seen in examples like the following tweet posted by a Democratic legislator that reads “Then stop calling it Chinese Virus @realDonaldTrump”. This tweet, along with many of the pro-Asian tweets posted by Democrats, admonishes the use of the term “China virus” for the Covid-19 virus. Among Republicans, the most frequent terms include “Chinese”, “racist”, “asianamerican” and “antiasian”. Similar to Democrats, Republican legislators posting pro-Asian content also discouraged the use of racialized slurs popularized during the pandemic as seen in the following tweet “@moodyforelpasso Blaming the Chinese people for Covid-19 is wrong and dangerous. Excusing the continued, intentional lies of Communist Chinese Govt is also wrong and dangerous. #txlege”.

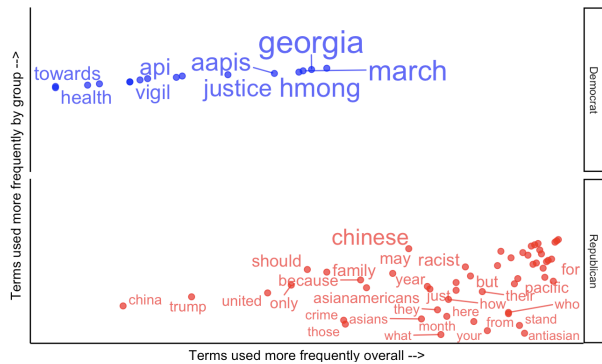


Figure 9: Pro-Asian Discourse Comparison: Democrats vs Republicans

## Discussion

This study uses a dataset of three million tweets posted by state legislators and a fine-tuned large language model to explore legislators’ pro and anti-Asian rhetoric. We find that

pro and anti-Asian speech constitutes only a small fraction of total tweets, but there are notable spatial variations across states. Asian-related rhetoric also varies across gender, race, and party affiliation. Women, Democrats, and Asian Americans are more likely to post pro-Asian tweets. The racial and partisan differences can partly be explained by the higher representation of Asian Americans on the ideological left. On the other hand, anti-Asian rhetoric is more prevalent among Republicans and men, reflecting conservative sentiment during the COVID-19 pandemic, partly fueled by derogatory language against Asian Americans utilized by high-profile political elites.

A major contribution of this research is its examination of sub-national political elite’s speech patterns during the Covid-19 pandemic. State legislators play a critical role in standardizing political rhetoric. Therefore, identifying patterns in their promotion or denunciation of hate speech online is an important puzzle piece in understanding American political dialogues. Our methodological recommendation emphasizes the need for a human-integrated approach to detect the nuances of hate speech. To this end, another key contribution of this study includes three comprehensive definitions of hate speech, counter-hate speech and neutral speech. These definitions are specific to the Asian identity for the purposes of this study but are adaptable for detecting other types of identity-based discrimination.

## Limitations

Our study has two key limitations. First, due to the time frame of our data and focus of our analysis, we are cautious about generalizing our findings. The pandemic period at large was atypical. As a catalyst for hate, however, this period is not unique. Hate speech victims, particularly those targeted by right wing extremist groups, tend to cycle in and out of focus (Hobbs et al. 2024). The targets of these attacks, both digital and otherwise, move based on social dynamics of the hate speech users (Soral, Bilewicz, and Winiewski 2018) and elite cues (Piazza 2020). The link between Covid-19 and the salience of Asians as a target group has been well documented (Chen et al. 2021; Abascal, Makovi, and Xu 2023), and is in many ways a return to more traditional forms of Asian discrimination (Hua and Junn 2021). While there is good justification for caution in Pandemic era studies, we do believe this particular element is not a detriment to our study.

Second, while we examined factors influencing Asian-related speech, the underlying motivations for sharing pro- and anti-Asian speech may differ significantly. Retweets and likes, additionally, appear to demonstrate potentially significant relationships. Our overall understanding of those metrics and how they relate to human behavior and opinion, however, requires more work. Future research should investigate these potential effects and their causes. Third, although we manually reviewed the synthetic data generated by ChatGPT-4o using a few-shot learning approach, we acknowledge that it may not fully capture every aspect of nuanced language, rhetorical strategies, and topic diversity present in real legislator speech. Future research should

consider expanding to other target groups and broader time periods to diversify data sources.

## Conclusion

In conclusion, we present a comprehensive framework for defining hate speech, neutral speech, and counter-hate speech, specifically in the context of Asian American communities. This refined framework is crucial for advancing the accuracy and consistency of hate speech research. By incorporating a human-in-the-loop approach, our model achieved improved classification performance, demonstrating the indispensable role of human expertise in navigating the nuances of hate speech detection. We hope this study serves as a foundation for future efforts to combine advanced computational methods with expert oversight, enabling a more precise and context-aware analysis of hate speech patterns in politically and socially charged environments.

## References

- Abascal, M.; Makovi, K.; and Xu, Y. 2023. Politics, not vulnerability: Republicans discriminated against chinese-born americans throughout the covid-19 pandemic. *Journal of Race, Ethnicity, and Politics* 8(1):83–104.
- ABC News. 2021. Trump's 'Chinese virus' tweet helped lead to rise in racist anti-Asian hashtags, study finds. *ABC News*. Accessed: Insert Date.
- Angevine, S. 2017. Representing all women: An analysis of congress, foreign policy, and the boundaries of women's surrogate representation. *Political Research Quarterly* 70(1):98–110.
- Anzia, S. F., and Berry, C. R. 2011. The jackie (and jill) robinson effect: Why do congresswomen outperform congressmen? *American Journal of Political Science* 55(3):478–493.
- Barrie, C., and Ho, J. C.-t. 2021. academictwitter: an r package to access the twitter academic research product track v2 api endpoint. *Journal of Open Source Software* 6(62):3272.
- Berinsky, A. J. 2017. Rumors and health care reform: Experiments in political misinformation. *British journal of political science* 47(2):241–262.
- Biswas, A.; Lin, Y.-R.; Tai, Y. C.; and Desmarais, B. A. 2024. Political elites in the attention economy: Visibility over civility and credibility? *arXiv preprint arXiv:2407.16014*.
- Broockman, D. E., and Skovron, C. 2018. Bias in perceptions of public opinion among political elites. *American Political Science Review* 112(3):542–563.
- Brown, N. E. 2014. *Sisters in the statehouse: Black women and legislative decision making*. Oxford University Press.
- Butler, D. M.; Kousser, T.; and Oklobdzija, S. 2023. Do male and female legislators have different twitter communication styles? *State Politics & Policy Quarterly* 23(2):117–139.
- Chan, N. K. M.; Kim, J. Y.; and Leung, V. 2022. COVID-19 and Asian Americans: How Elite Messaging and Social Exclusion Shape Partisan Attitudes. *Perspectives on Politics* 20(2):618–634. Num Pages: 618-634 Place: Cambridge, United Kingdom Publisher: Cambridge University Press Section: Special Issue Articles: Pandemic Politics.
- Chen, A.; Nyhan, B.; Reifler, J.; Robertson, R.; and Wilson, C. 2021. Exposure to Alternative & Extremist Content on YouTube. Technical report, Anti-Defamation League.
- Chen, W.; Lin, F.; Li, G.; and Liu, B. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing* 578:127428.
- Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.
- CNBC. 2021. House to hold hearing on rise of anti-Asian violence during COVID. <https://www.cnn.com/2021/03/17/house-to-hold-hearing-on-rise-of-anti-asian-violence-during-covid.html>. Accessed: Insert Date.
- Cook, J. M. 2017. Twitter adoption and activity in us legislatures: A 50-state study. *American Behavioral Scientist* 61(7):724–740.
- Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 512–515.
- Diekema, D. S. 2014. Personal belief exemptions from school vaccination requirements. *Annual Review of Public Health* 35:275–292.
- Enders, A. M., and Thornton, J. R. 2022. Racial Resentment, Electoral Loss, and Satisfaction with Democracy Among Whites in the United States: 2004–2016. *Political Behavior* 44(1):389–410.
- Filindra, A., and Harbridge-Yong, L. 2022. How Do Partisans Navigate Intra-Group Conflict? A Theory of Leadership-Driven Motivated Reasoning. *Political Behavior* 44(3):1437–1458.
- Fine, J. A., and Hunt, M. F. 2023. Negativity and elite message diffusion on social media. *Political Behavior* 45(3):955–973.
- Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.
- Gaynor, S. W., and Gimpel, J. G. 2023. Building support through the personalization of twitter messages in a permanent campaign. *American Politics Research* 1532673X231184434.
- Gopal, I.; Kim, T.; Nakka, N.; Boehmke, F. J.; Harden, J. J.; and Desmarais, B. A. 2024. The national network of us state legislators on twitter. *Political science research and methods* 1–13.
- Govers, J.; Feldman, P.; Dant, A.; and Patros, P. 2023. Down the rabbit hole: Detecting online extremism, radicali-

- sation, and politicised hate speech. *ACM Computing Surveys* 55(14s):1–35.
- Graham, M. H., and Singh, S. 2024. An outbreak of selective attribution: Partisanship and blame in the covid-19 pandemic. *American Political Science Review* 118(1):423–441.
- Hajnal, Z.; Kuk, J.; and Lajevardi, N. 2018. We all agree: Strict voter id laws disproportionately burden minorities. *The Journal of Politics* 80(3):1052–1059.
- He, B.; Ziemis, C.; Soni, S.; Ramakrishnan, N.; Yang, D.; and Kumar, S. 2021. 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*, 90–94.
- Hobbs, W.; Lajevardi, N.; Li, X.; and Lucas, C. 2024. From anti-muslim to anti-jewish: target substitution on fringe social media platforms and the persistence of online and offline hate. *Political Behavior* 46(3):1747–1769.
- Hopkins, D. J., and Noel, H. 2022. Trump and the Shifting Meaning of “Conservative”: Using Activists’ Pairwise Comparisons to Measure Politicians’ Perceived Ideologies. *American Political Science Review* 116(3):1133–1140. Publisher: Cambridge University Press.
- Hua, W., and Junn, J. 2021. Amidst pandemic and racial upheaval: Where asian americans fit. *Journal of Race, Ethnicity, and Politics* 6(1):16–32.
- Kim, T.; Nakka, N.; Gopal, I.; Desmarais, B. A.; Mancinelli, A.; Harden, J. J.; Ko, H.; and Boehmke, F. J. 2021. Attention to the covid-19 pandemic on twitter: Partisan differences among us state legislators. *Legislative studies quarterly*.
- Kreitzer, R. J. 2015. Politics and morality in state abortion policy. *State Politics & Policy Quarterly* 15(1):41–66.
- Kruikemeier, S. 2014. How political candidates use twitter and the impact on votes. *Computers in human behavior* 34:131–139.
- KXAN News. 2021. Atlanta shootings put spotlight on surging anti-Asian sentiment in America. <https://www.kxan.com/news/atlanta-shootings-put-spotlight-on-surging-anti-asian-sentiment-in-america/>. Accessed: Insert Date.
- Laurer, M.; van Atteveldt, W.; Casas, A.; and Welbers, K. 2023. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.
- Laurer, M.; Van Atteveldt, W.; Casas, A.; and Welbers, K. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis* 32(1):84–100.
- Li, J., and Ning, Y. 2022. Anti-asian hate speech detection via data augmented semantic relation inference. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 607–617.
- Li, L.; Fan, L.; Atreja, S.; and Hemphill, L. 2024. “hot” chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web* 18(2):1–36.
- Mansbridge, J. 2003. Rethinking representation. *American political science review* 97(4):515–528.
- Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4):372–403.
- Müller, K., and Schwarz, C. 2020. Fanning the Flames of Hate: Social Media and Hate Crime.
- Osmundsen, M.; Bor, A.; Vahlstrup, P. B.; Bechmann, A.; and Petersen, M. B. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review* 115(3):999–1015.
- Payson, J.; Casas, A.; Nagler, J.; Bonneau, R.; and Tucker, J. A. 2022. Using social media data to reveal patterns of policy engagement in state legislatures. *State Politics & Policy Quarterly* 22(4):371–395.
- Pei, X., and Mehta, D. 2020. # coronavirus or# chinese-virus?!: Understanding the negative sentiment reflected in tweets with racist hashtags across the development of covid-19. *arXiv preprint arXiv:2005.08224*.
- Piazza, J. A. 2020. Politician hate speech and domestic terrorism. *International Interactions* 46(3):431–453.
- Pollert, I., and Mooney, C. Z. 2022. Substantive and political learning among the us states: Abortion policy diffusion, 1993–2016. *State Politics & Policy Quarterly* 22(3):320–343.
- Reingold, B.; Widner, K.; and Harmon, R. 2020. Legislating at the Intersections: Race, Gender, and Representation. *Political Research Quarterly* 73(4):819–833. Publisher: SAGE Publications Inc.
- Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.
- Ruckelshaus, J. 2022. What kind of identity is partisan identity? “social” versus “political” partisanship in divided democracies. *American Political Science Review* 116(4):1477–1489.
- Russell, A. 2018. U.S. Senators on Twitter: Asymmetric Party Rhetoric in 140 Characters. *American Politics Research* 46(4):695–723. Publisher: SAGE Publications Inc.
- Shino, E.; Suttman-Lea, M.; and Smith, D. A. 2022. Determinants of rejected mail ballots in georgia’s 2018 general election. *Political Research Quarterly* 75(1):231–243.
- Siegel, A. A.; Nikitin, E.; Barberá, P.; Sterling, J.; Pullen, B.; Bonneau, R.; Nagler, J.; Tucker, J. A.; et al. 2021. Trumping hate on twitter? online hate speech in the 2016 us election campaign and its aftermath. *Quarterly Journal of Political Science* 16(1):71–104.
- Slothuus, R., and Bisgaard, M. 2021. How political parties shape public opinion in the real world. *American Journal of Political Science* 65(4):896–911.

Soral, W.; Bilewicz, M.; and Winiewski, M. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior* 44(2):136–146.

Tai, Y. C.; Nakka, N.; Patni, K. N.; Rajtmajer, S.; Munger, K.; Lin, Y.-R.; and Desmarais, B. 2024. The digitally accountable public representation database: Measuring online communication by federal, state, and local officials.

Tai, Y. C.; Buma, R.; and Desmarais, B. A. 2023. Official yet questionable: examining misinformation in us state legislators' tweets. *Journal of Information Technology & Politics* 1–13.

Toliat, A.; Levitan, S. I.; Peng, Z.; and Etemadpour, R. 2022. Asian hate speech detection on twitter during covid-19. *Frontiers in Artificial Intelligence* 5:932381.

Vergeer, M. 2015. Twitter and political campaigning. *Sociology compass* 9(9):745–760.

Vidgen, B.; Botelho, A.; Broniatowski, D.; Guest, E.; Hall, M.; Margetts, H.; Tromble, R.; Waseem, Z.; and Hale, S. 2020. Detecting east asian prejudice on social media. *arXiv preprint arXiv:2005.03909*.

West, E. A., and Iyengar, S. 2022. Partisanship as a social identity: Implications for polarization. *Political Behavior* 44(2):807–838.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics* 50(1):237–291.

Zou, L. X., and Cheryan, S. 2017. Two axes of subordination: A new model of racial position. *Journal of personality and social psychology* 112(5):696.

## Ethics Checklist

### Ethics Statement

Due to the sensitive nature of this study detecting anti-Asian content and the number of harmful slurs that are prevalent in this study, we took extra care to prepare hand-coders for hate speech detection tasks.

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, and the tweets used in this collection are from public officials (state lawmakers), so they are not covered under IRB protections that a standard research subject would have. Even still, this would not encroach on any of these issues even for standard cases. Data is approached in the aggregate, and the only descriptive figures with names just deal with total tweets - all publicly available data.**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, and we**

**devote an appropriate amount of time for this in our methods section**

- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **NA, Our data consists entirely of tweet metadata. Everything should be accounted for.**
- (e) Did you describe the limitations of your work? **Yes, and we address these in the final section of the main body.**
- (f) Did you discuss any potential negative societal impacts of your work? **NA, Our work examines lawmaker digital speech, attempting to make sense of what causes hate speech and counter hate speech.**
- (g) Did you discuss any potential misuse of your work? **NA, It is not immediately clear what elements of this could be used for nefarious purposes. Those whose speech we are using are public officials and are (or should be given their exalted role in society) aware of the inherently public nature of their speech.**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, and our work follows standard protocol in terms replication and the availability of our data for other researchers. Raw tweets are not included in this, though again, they are not protected given that we are looking at the speech of public officials.**
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, and we believe that our work meets the guidelines.**

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **Yes, and these are discussed in the intro section and the results.**
- (b) Have you provided justifications for all theoretical results? **Yes, and we discuss this in our results section as well as our conclusion/limitations.**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes, and these are discussed in the first half of the paper as we outline our process and expectations.**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Yes, and our model is robust and we account for all of the standard variables in state lawmaker behavior.**
- (e) Did you address potential biases or limitations in your theoretical framework? **Yes, and we address model limitations in the appropriate sections, and our conclusion also devotes time to the limits of this study.**
- (f) Have you related your theoretical results to the existing literature in social science? **Yes, and our background section and our conclusion tie this back to more general theory on state digital politics.**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the so-

- cial science domain? [Yes, and we give more detail on this in our conclusion.](#)
3. Additionally, if you are including theoretical proofs...
    - (a) Did you state the full set of assumptions of all theoretical results? *NA*
    - (b) Did you include complete proofs of all theoretical results? *NA*
  4. Additionally, if you ran machine learning experiments...
    - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes, we provide a link to the code needed to reproduce our results at the end of the Results section. However we do not provide Twitter data in accordance with Twitter privacy policies.](#)
    - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes, and this is featured in our "Method" section.](#)
    - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes, and this is featured in our "Method" section.](#)
    - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? *NA*
    - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes, and we provide model comparisons and justification for our final model in Table 1.](#)
    - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? [Yes, and this is featured in our "Methods" section.](#)
  5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
    - (a) If your work uses existing assets, did you cite the creators? *NA*
    - (b) Did you mention the license of the assets? *NA*
    - (c) Did you include any new assets in the supplemental material or as a URL? *NA*
    - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? *NA*
    - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? *NA*
    - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? [No, we are not releasing tweets data at the time of this study, please see item 4a.](#)
    - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? *NA*
  6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
    - (a) Did you include the full text of instructions given to participants and screenshots? *NA*

- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? *NA*
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? *NA*
- (d) Did you discuss how data is stored, shared, and deidentified? *NA*

## Appendix

### Asian Related Keyword Examples

"Asia", "Asian", "Cambodia", "China", "Chinese", "Filipino", "Hmong", "India", "AAPI", "Uyghurs", "China Virus", "Chinese Virus", "Kung flu", "bamboo coon", "chigger", "chiggers", "chinese wetbacks", "ching chong", "ching chongs", "slant eyes", "sideways pussies", "sideways pussy", "sideways vagina", "bat eater", "boycottchina", "MakeChinaPay", "FreeTibet", "StandwithHK", "StopAsianHateCrimes", "AsianAmericans", "PROTECTASIANLIVES", "IamNotAVirus", "IAmNotCovid19", "BeCool2Asians", "StopAAPIHate"

### Hand Labeling Procedure

We have three rounds of hand-coding tweets as pro-Asian, anti-Asian and neutral. We do this to ensure the reliability of our hand-labeled data and improve model training outcomes. Two researcher initially hand-coded data into three categories: pro-Asian, anti-Asian and neutral. To ensure label accuracy and consistency, our team implemented a structured hand-coding workflow grounded in a consensus on anti-Asian, pro-Asian, and neutral definitions. The entire process is detailed below.

First, we sampled 1,000 tweets from the pool of Asian-related tweets. During the first round of hand-coding, two coders individually labeled the same 1,000 tweets sample. Coders then checked the number of intercoder agreements and discussed any discrepancies to further refine the coding scheme. Based on these discussions coders further specified additional rules that were instructive for subsequent rounds of coding:

1. References to governments or political parties (e.g., "CCP," "party," "Chinese Communist Party") should be classified as neutral rather than anti-Asian. The criticism directed at a political entity is considered a political attitude rather than harmful targeting of individuals, cultures, or racial-identities.
 

Example: "If this is on Facebook, it won't be there long #Censorship And, YES! This would be the best outcome for the people in communist oppression in China #WestTaiwan #InstapunditStillRules #GoGlennReynolds <https://t.co/4RfIOSFsL4>"
2. Not all anti-racism tweets are regarded as pro-Asian, unless the tweets specifically positions themselves as disavowing hate toward Asian communities
 

Example: "Democracy won't win this unless we end racism. We can't use democracy to deny one another - as voter suppression law in Georgia just did - and be a

free people. We are either all free or we are an autocracy by another name - and the Chinese government wins <https://t.co/8QJyLy0MIW>”

3. News reports are treated as neutral, regardless of the reporters’ personal political ideology

Example: “RT @sfchronicle: Distressed by the rise in xenophobia and racism during the coronavirus pandemic, a coalition of Asian American groups base”

After the first-round of hand-labeling and establishing a more refined coding-scheme for pro, anti and neutral subsequent rounds of hand-coding included additional samples of tweets resulting in 1,586 total hand-coded tweets. Table 3 features the Krippendorff’s alpha intercoder agreement scores between coders for each round of hand-labeling. By the third round the agreement rate began plateauing, therefore; we proceeded to input our hand labeled data to train our model.

Table 3: Krippendorff’s Alpha Intercoder Reliability Scores

Hand-Coding Rounds	Score
Round 1	0.49
Round 2	0.58
Round 3	0.59