Algorithmic Bias: The Amplification of COVID-19 Related Racism Online

Andrew Takais

April 30th, 2024

Introduction

Expanding into the 21st century, human medical advancement continues to present novel solutions to age-old health paradigms. From new cancer therapies to nanoparticle drug delivery systems, a field of endless improvements targets the variable human condition. Yet, in this age of extreme advancement, Coronavirus undermined global medical mastery and plummeted the world into a modern pandemic that has claimed the lives of over 1.18 million people in The United States alone (CDC, 2024). The effectiveness and precision in which COVID-19 spread is, on one side, attributed to the novelty of the virus. With no existing safeguards in place (i.e. vaccines), a highly contagious disease spreads like wildfire in a global community. While this of course presents part of the story, researchers now look back upon the pandemic with new insight: the social construction of COVID-19 worsened the severity and lengthened the duration of the experienced pandemic. Medicine alone cannot prevent the spread. Legislative bodies and their constituents must also make a joint effort, utilizing resources and information provided via credible health organizations such as the CDC. However, the divisive nature of politics, especially within the United States, polarized the issue at hand, leading to the spread of scientific misinformation and racism against Asian Americans that only served to deter anti-pandemic efforts. One significant issue stemming from COVID-19 was the drastic increase in Sinophobic sentiment at the start of the COVID-19 pandemic.

Prejudice forms a major pillar of misinformation that not only puts marginalized groups at further risk of oppression but also perpetuates fallacies in regard to the disease. A scapegoat in this case creates solidarity against an 'opposing' human entity rather than the cause of the pandemic. Bias can both be, and lead to, misinformation due to groups selectively seeking out or interpreting information in a way that supports their own preconceived biases while ignoring contradictory perspectives. This creates a cycle of misinformation that can become extremely hard to destroy. Stuck in a loop, a group with mutual prejudice continues to reinforce each other's beliefs while growing in their resistance to opposing rhetoric (Muhammad & Matthew, 2022). The effects in regard to COVID-19 can reach even further than oppression against a marginalized group, becoming harmful to society as a whole when the information incorrectly discusses vaccine efficacy, COVID-19 treatments, and the polarized politics surrounding the disease.

In a similar vein of technological advancement, social media platforms and their underlying algorithms have revolutionized the way information is disseminated and consumed, significantly impacting public health discourse during the COVID-19 pandemic. These platforms, guided by complex algorithms, often prioritize content that maximizes user

engagement, which can inadvertently amplify sensationalist and xenophobic narratives (Frenkel, S, & Myers, S. L., 2023). This algorithmic bias, while designed to enhance user experience, has the unintended consequence of promoting content that fuels misinformation and prejudice during critical times. As users interact with and share this biased content, the algorithms learn to promote similar messages, creating a feedback loop that can exacerbate public health crises by undermining trust in scientific authority and fostering division. Understanding the role of social media algorithms in shaping public discourse is crucial for developing more resilient public health strategies and combating the spread of harmful misinformation. Moreover, the lack of effective moderation on these platforms further complicates the challenge, as insufficient oversight allows hate speech and misinformation to proliferate unchecked. This inability to accurately police content not only perpetuates harmful narratives but also hinders efforts to foster a safer and more informed online community.

Diving into the COVID-19 domino effect illuminates new ways in which we can educate the public about global events, navigating around an automatic implicit bias. The goal of this research is to provide exploratory analysis into the increases in Sinophobic sentiment due to COVID-19's initial spread online, while quantifying the effect that the progression of time has on the frequency of hateful tweets. Additionally, it will explore how these existing forms of racial bias can be associated with the creation and perpetuation of additional forms of misinformation in regard to the pandemic, such as poor vaccine efficacy. And it will also acknowledge quantifiable limitations that stem from the nature of inference methodology.

COVID-19 & Sinophobia

Sinophobia is defined as the fear or hatred of China or Chinese people. And after COVID-19's identification in Wuhan, China, anti-Chinese sentiments heightened and spread throughout the United States. Hate crimes against Asian Americans increased 77% from 2019 to 2020, with a total of 9,000 violent crimes reported from March 2020 to June 2021 alone (Findling, 2022). While Sinophobia is not necessarily a novel concept, it is possible that the pandemic contributed to an increase in prejudice sentiment due to its discovery in China fueling existing bias.

Several studies have performed an initial analysis of the effects of COVID-19 on the perception of Chinese people and China as a country. A notable analysis can be found within Tahmasb's research at Boston University, where his team utilized alt-right social media like 4Chan to study Sinophobic sentiment spikes from around December 2019 up until the beginning of the pandemic in late March 2020. The results showed drastic upticks–as much as 250x increases–in the use of slurs against Chinese people beginning in January of 2020. With words such as "Chinese flu" gaining further traction around the same time that WHO declared Coronavirus a global pandemic. But the breadth of these existing analyses does not seem to proceed past what is considered the start of the pandemic in the western world.

Health is a critical topic that remains prevalent in every American's mind, and to put it at risk due to Coronavirus ambiguity, invoked national anxiety. While the magnitude of such fears differs from person to person, the prevalence of fear in regard to COVID-19 uncertainty was the highest level of national anxiety observed since the Cold War (Quadros et al., 2021). Additional research shows that a surplus in pandemic related information only serves to reinforce fears and uncertainties (Traczyk et al., 2018). A study led by Thomas Ramsey provides a crucial basis for social commentary with respect to COVID-19. His team investigated the effects of chronic or acute anxiety in a social context, and how feelings of fear can lead to or provoke implicit and explicit bias. The results demonstrate heightened risk perception after experiencing media meant to provoke feelings of uneasiness, with implicit bias against certain racial groups. Such a correlation further solidifies the assumption that the emergence of COVID-19 led to increased negative sentiment toward Chinese people, or more generally, against those of Asian descent.

The Role of Social Media Algorithms

How Do These Algorithms Work?

Social media algorithms are pivotal in determining the kind of content that surfaces on our feeds, profoundly shaping our online experiences. At the core of these algorithms are various signals indicating user preferences, such as likes, shares, comments, and the duration of content interaction. These engagements serve as critical inputs, telling the algorithm that the user enjoys a specific type of content, prompting it to prioritize and present more of the same. This feedback loop is designed to enhance user satisfaction by tailoring the content to perceived preferences, thereby increasing the time spent on the platform.

However, the exact workings of these algorithms remain opaque due to their proprietary protection. Social media companies guard these algorithms closely, primarily because they are integral to the platform's ability to maximize engagement and, by extension, revenue. Despite the secrecy, it is understood that these algorithms are fundamentally rooted in machine learning principles, utilizing various forms of regression analysis to process and predict user behavior. The simplest forms of these analyses include multiple variable and logistic regression.

Multiple variable regression in social media algorithms involves creating a complex equation that accounts for various significant variables—such as the frequency of interactions with certain types of content—and how they correlate with a continuous output, like time spent on the app. These equations can both express various relationships with the output (linear, polynomial, etc.) allowing extreme flexibility in the models ability to fit complex data that does not necessarily fit simplistic relationships.

$$\text{Time On App} = \beta_0 + \beta_{Likes} \times X_{Likes} + \beta_{Comments} \times X_{Comments} + \beta_3 \times X_{Shares} + \epsilon$$

In the above equation, β_i represents a coefficient for a specific variable, quantifying the impact X_i has on the dependent variable. The X_i in this case, is either a categorical or quantitative input to the equation that measures some aspect of online engagement (i.e. how frequently do you like posts on Instagram?)

Whereas typical linear regression assumes continuous outputs, logistic regression is used to represent categorical outcomes, maximizing the probability of a user's actions being correctly assigned to a discrete classification. In the binary sense, we can represent the output of the model as a probability of the input data being classified as one of the outcomes (i.e. user clicks on ad or does not click on ad). Each user interaction, from a like to a share, is assigned a weight (denoted as X_i in statistical models), reflecting its significance in influencing future content recommendations. Logistic regression uses a regression equation, mapping its output to a probability using a sigmoid function $f: \mathbb{R} \to (0,1)$:

$$f(Z) = \frac{1}{1 + e^{-Z}}$$

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The Bias-Variance Trade-Off

These regression models are not static; they are continuously refined through ongoing processing to better interpret the relationships between significant variables and user responses. The ultimate goal of these machine learning algorithms is predictive accuracy—they are constantly adjusted in response to new data to align the predicted outcomes as closely as possible to the actual observed behavior. This process of tweaking and testing is essential for the algorithms' ability to deliver increasingly personalized content to users, thus keeping them engaged and connected to the platform, as well as assigning them ads that are most correlated to their interests.

In the realm of machine learning and statistical science, the bias-variance trade-off is a critical challenge that developers must navigate. Within this concept, bias refers to the error introduced by approximating a real-world problem with a simplified, mathematical model, whereas variance refers to the error that arises from small fluctuations in the data used to train the model. Essentially, a model with high bias oversimplifies the relationship between predictors and the desired outcomes, whereas a model with high variance pays too much attention to training data, leading to low predictive accuracy.

In the context of social media algorithms, the allure of including more variables and richer data sets is clear: such models boast increased complexity and flexibility, which can dramatically reduce bias, allowing for a more nuanced representation of real-world phenomena, and human decisions, in mathematical form. This robustness in capturing underlying patterns means that the model's assumptions align more closely with the

intricate realities they aim to replicate. However, this comes at the cost of increased variance, referring to a model's sensitivity to fluctuations in the training data. A model with high variance may perform exceptionally well on training data but fail to generalize to unseen data, resulting in poor predictive performance.

While balancing the bias-variance trade-off, the extensive data and large number of variables that fuel these sophisticated models can lead to outputs that are challenging to interpret. This lack of interpretability presents substantial obstacles for engineers tasked with understanding and fine-tuning the models' decision-making processes. Without clear insight into how these algorithms arrive at specific recommendations or content prioritization, identifying and correcting biases within the model becomes a formidable task. Consequently, while striving to construct algorithms that faithfully reflect the complexity of their tasks, there's a constant battle to maintain the transparency and interpretability essential for mitigating problematic outputs, particularly in content recommendations that have far-reaching social implications.

The sophistication of these algorithms lies in their capacity to learn and evolve from vast quantities of data. However, this capability also presents challenges, particularly in moderating content and policing hate speech. The reliance on user engagement metrics can inadvertently lead to the amplification of sensationalist or divisive content, which tends to generate significant interaction but is harmful to society at large. The ongoing challenge for social media platforms is to balance algorithmic efficiency with the ethical implications of content distribution, ensuring that they contribute positively to public discourse without compromising the quality or safety of the online environment.

Social Media's Role in Amplifying Racism in the United states

Social media algorithms are powerful tools in determining the visibility and dissemination of content. The Pew Research Center states that around 53% of U.S. adults get their news from social media platforms. In the vast digital landscape of platforms like Facebook, X/Twitter, and YouTube, these algorithms often dictate what is seen and what fades into the background. Unfortunately, these algorithms may have played a role in worsening xenophobia in the United States by highlighting and promoting content laced with bias, misinformation, and racism. Given the sheer volume of data generated on these platforms daily, relying solely on human moderation to monitor and mitigate harmful content is impractical, if not impossible. Consequently, AI has stepped into the role of content moderation. AI systems are programmed to detect and filter out hate speech and misinformation; however, their capacity to do so accurately is not without significant limitations (Darbinyan, 2022).

An article published on Unite AI by Alex McFarland notes that the central issue lies in the nuanced nature of language and the complexity of social discourse. Social media feeds are filled with irony, sarcasm, cultural references, and idiomatic expressions—subtleties that current AI models, despite their advancement, often struggle to interpret correctly. Efforts to refine these AI-driven hate speech detection algorithms are continual, with researchers aiming to create adaptive models that can keep pace with evolving online discourse (McFarland, 2022). Despite their progress, these algorithms face the formidable task of

balancing moderation with the protection of free speech, sometimes erring on the side of caution and allowing harmful speech to slip through or, conversely, over-censoring and suppressing legitimate expression.

The pandemic's role in this scenario has highlighted extreme issues with automatic moderation. Misinformation related to COVID-19 has proliferated at an alarming rate, propagating racist sentiments and fostering an environment conducive to xenophobia. As sensational and emotionally charged content typically generates more engagement, the algorithms may inadvertently prioritize and amplify such content (Zenone, et al., 2022).

Case Study: Anti-Semiticism in Online Spaces

Before diving into a more quantitative investigation of COVID-19, we can see a similar issue unfolding in online spaces with the rise of anti-Muslim and anti-semitic sentiments.

Following the October 7th, 2023, Hamas attack, there was an alarming surge in online antisemitic and Islamophobic content, with social media platforms like X (formerly Twitter) experiencing significant increases in hate speech. According to a *New York Times* article titled "Antisemitic and Anti-Muslim Hate Speech Surges Across the Internet," the hashtag #HitlerWasRight was used in over 46,000 posts in just one month, a stark increase from less than 5,000 mentions in previous months. This spike is part of a broader 919% increase in antisemitic content on X during the same period. The complexities of these issues are deepened as platforms struggle to distinguish between anti-Zionism and antisemitism, often blurring the lines in moderation policies.

On the other hand, the same NYT article notes that the hashtag #LevelGaza also saw a dramatic increase in usage on X, with 3,000 mentions, up from less than a dozen the previous month. There was also a significant presence of hashtags like #MuslimPig and #KillMuslims. Platforms like TikTok and Facebook acted to remove flagged content, with TikTok removing over 730,000 videos violating its hate speech policies within a week. Some of the content, however, arose as difficult to detect with traditional methods, as users included veiled hate speech references, such as referring to Adolf Hitler as an "Austrian painter" rather than his name, circumnavigating automatic detection systems. Platforms like Tiktok and Meta provided a public statement after the proliferation of hate on their platforms was brought to their attention by the *New York Times*; however, X/Twitter refused to comment or reveal if any further moderation was conducted.

According to a study by Malena Dailey, in just the latter half of 2020 (July-December), over 6 billion posts were removed from major social media platforms, such as Instagram, Facebook, and X. Only about 4.5 million of these posts were found on X whereas 5.7 billion were on Facebook and 65 million on Instagram. Even accounting for the differing sizes of their userbase, Instagram moderated and removed up to five times as much content, with Facebook much higher than both. Additionally, Instagram, Facebook, and TikTok boast 90% preemptive removal rates—or the removal of problematic content before any user engagement—whereas Twitter claims around a 70% preemptive removal rate.

Quantitative Analysis of COVID-19 Related Content on X

Goals

As discussed above, social media can become a breeding ground for hate and misinformation. And through past events, X/Twitter has failed to publically state additional efforts to curve hate speech on their platform or reveal current content moderation algorithms. Additionally, since Twitter was purchased by Elon Musk in 2022, content moderation policies have been further relaxed (Ingram, 2022). The following analysis will be primarily concerned with content on X, and COVID-19 related tweets on the platform from 2020 to 2022.

Datasets

The data is primarily concerned with tweet data and text analysis. Tweets were pulled from IEEEDataPort which provided the Tweet IDs for all COVID-19 related tweets from March 20, 2020, to March 31, 2022. Due to Twitter API constraints, the CSV download only contained associated IDs which were then hydrated with the tweet core data via Hydrator software. The breadth of the data presented upwards to 2 billion tweets, but the collection process faced the rate limitations of the Twitter API. Tweet information was only collected on a monthly basis for three months out of the year (March, July, and November) for both 2020 and 2021, as well as an additional set for March 2022. The main variables of interest are the tweet's time stamp, the user's approximate location, and the text of the tweet. There are a total of 99,321 tweets in the data set.

Lastly, a data set from Pew Research was employed in order to obtain the political leanings of U.S. states. This included each state and its political leaning on the American political binary as a decimal (-1 being fully Democrat and +1 being fully Republican).

Hypothesis

Following the stated information, the following hypothesis are assumed:

1). As time progresses within the pandemic, the frequency of prejudiced tweets will decrease as a general trend and the degree of prejudice within tweets (I will call this its sentiment score) will decrease as well. There will be minor upticks during moments of high infection rates (i.e. Omicron) but will overall remain on a decline.

As discussed, fear surrounding the pandemic may invoke or allow pre-existing biases to become rationalized in the heads of their conceptionists. But there exists a sort of burn out that can occur in parallel to the progression of the pandemic. Research by Stevens et al. displays that people become increasingly desensitized to COVID-19 related news as time progresses. As the death tolls increased, it was found that participants became less and less concerned. This desensitization effect could justify why the initial biases associated with the fear of the unknown slowly decrease over time.

2). There will be sub-trends in which geographical locations tweets are sent from. States associated with more Republican views will possess a higher frequency of biased tweets

and a, on average, higher sentiment score, indicating more hateful language within the tweets.

Throughout the pandemic, COVID-19 has become increasingly politically polarized (Jungkunz, 2021). With a history of extreme nationalism and an 'America first' rhetoric, the Republican party has made its voice known to strictly oppose further government lock downs as to preserve one's quality of life. Additionally, research shows that some politicians actively engaged in passing the blame towards China while also vetoing pandemic deterring legislation. So this may imply that states and locations that have higher rates of Republican ideology will express more bias due to its normalization among the state's citizens and leaders (Silver, et al., 2020).

3). Geographical location (as gauged by Republican or Democrat state leanings) and time period can validly and significantly predict the degree of prejudice in a COVID-19 related tweet as per the previous justifications.

Methods

Prior to investigating further the legitimacy of the stated hypothesis, the research needs a dictionary of derogatory values that are associated with Asian hate in order to classify the degree of harm within a given tweet. This will require text analysis according to a collected vector of words that contain what this paper deems derogatory and prejudice. These words were obtained via a literature review and were compiled from the mentioned study by Tahmasbi and a *New York Times* Article by Kathy Hong.

The following, adapted figure shows a snapshot of word choice due to their drastic uptick once COVID-19 was declared a pandemic by WHO:

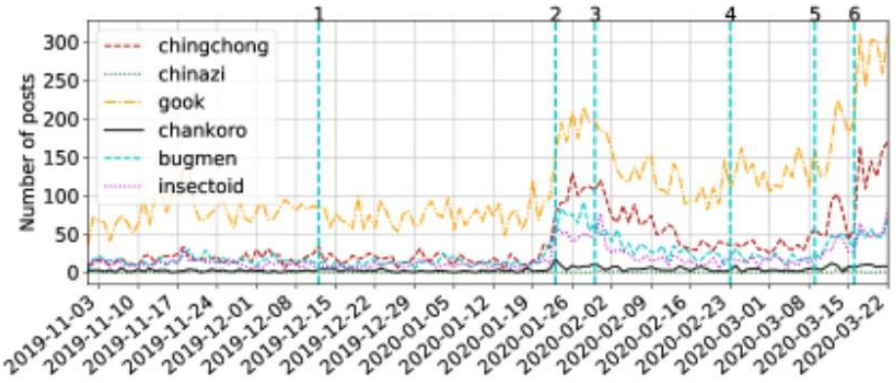


Figure Adapted from Tahmasbi et al.

With the following dictionary, text analysis will be conducted on the data set of extracted tweets between March 2020 and March 2022. These tweets will be screened for the mentioned words and given a sentiment score. For each word below contained in the tweet, the score will be incremented by one. This counts duplicate uses of the same

```
##    Derogatory_Words
## 1              chink
## 2            chinese
## 3              china
## 4            kungflu
## 5          chinavirus
## 6            oriental
## 7                gook
## 8             asshoe
## 9          chingchong
## 10             bugmen
## 11          insectoid
## 12          chankroo
## 13            chinazi
```

word. Only tweets deemed 'derogatory' (sentiment score >= 1) will be included to assess the magnitude of tweets overtime. With this, two visualizations were produced. One counts the number of biased tweets within each respective month/year grouping to assess the true frequency of biased tweets while the alternate plot displays the average of the sentiment scores for those same groupings. The latter represents an attempt to quantify the degree of prejudice within each tweet itself and to show this over the same time period.

The subsequent analysis took into account the geographical location of the tweets. Once again, this will look into both the number of prejudiced tweets within a given grouping and the average of the sentiment score within that same category The grouping here, however, is by state in the United States. Once these variables were measured, additional data was obtained from Pew Research that linked each state to their political leanings. Geographic visualizations were displayed to show biased tweet frequency and average sentiment score on a state-by-state basis. An associated graph with political leanings was also produced to draw side-by-side conclusions between the prejudiced tweets at hand and possible political associations. Additionally, locations provided via the Twitter data were standardized in terms of geographical location. To facilitate the process, location was restricted to only the United States, and prejudice tweets with a sentiment score of at least one were included. Names of states and cities were then standardized. This also allows the incorporation of a political orientation assessment that can add on to the ideas of bias being studied.

Lastly, a multivariate linear model was used to assess the significance in the relationship between the assessed variables and the degree of bias within Tweets. The only variables used in the model are political party (an extension on geographical location) and time period (Month/Year). The significance of each variable was assessed and explored. The model was only made to predict degree of bias within Tweets as this was the least

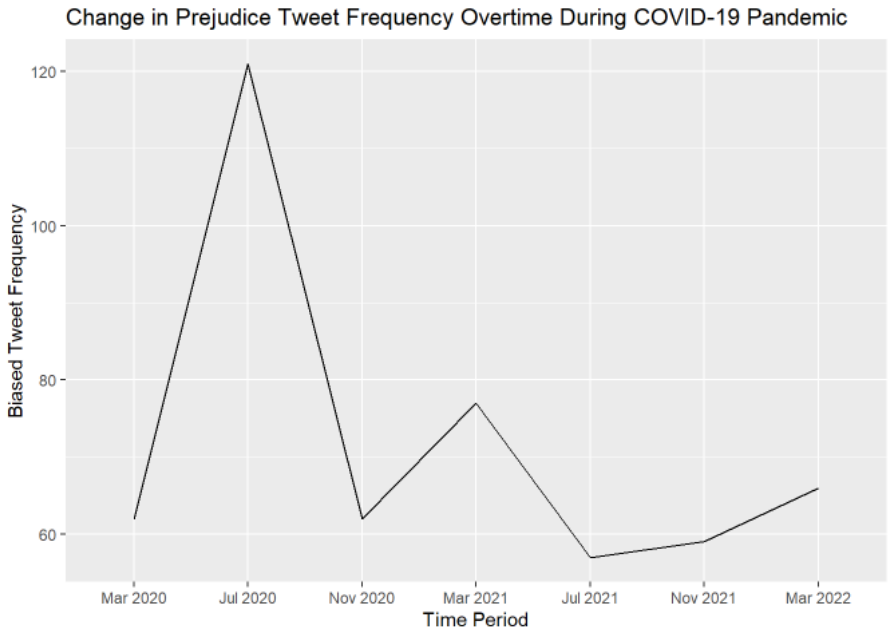
consistent variable in both the geographical and time period analysis. This will be used to make the final claim on validity of hypothesis one.

Within this analysis, a majority of the generalizations will only apply to the United States. As the research only considers tweets with a baseline of bias, and a majority of these tweets had American geo-tags.

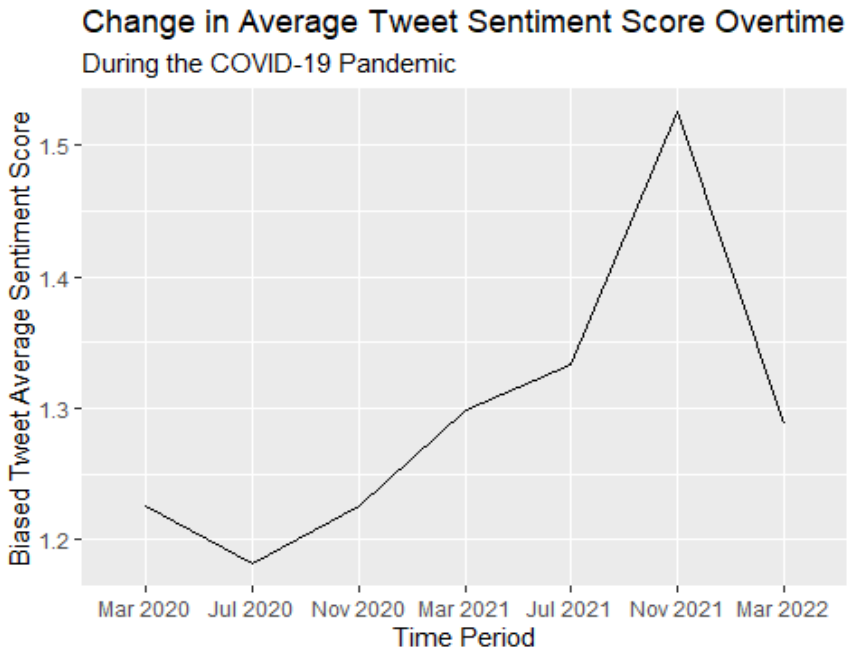
Results & Discussion

Tweet Frequency and Sentiment Score Results

The results for the first analysis, and for part of the first hypothesis, seem to be split in terms of conclusions. In terms of prejudiced tweet frequency, there definitely exists a sharp incline in Tweets employing what this study has deemed 'prejudiced' language. This incline correlates with the start of the pandemic in March 2020. This now expands beyond Boston University's inequity study and shows that the trend their team noticed beginning in March 2020 continues for several months into the pandemic. The magnitude of tweets nearly halves between July 2020 and November 2020 (122 to 62), with this lower trend continuing throughout the remainder of the data. There do exist additional spikes in prejudice Tweet frequency. This can be seen starting in November 2020 and continuing to March 2021, as well as a slight upwards trend beginning around November 2021. These spikes may be related to the state of the pandemic during those time periods, as COVID-19 numbers fluctuate depending on infection rates. Also, early 2021 saw the gradual rollout of Coronavirus vaccinations which were met with hesitancy due to fears associated with their expedited approval (Larson, et al., 2021). These sentiments, as well as rising infection rates, could lead to a reinvocation of bias similar to the start of the pandemic, but to a lesser magnitude. A similar story could be told for the transition from July 2021 to November 2021, where after large vaccination participation, the Omicron COVID-19 variant began to seemingly revert current progress in the eyes of the public (CDC, 2022).



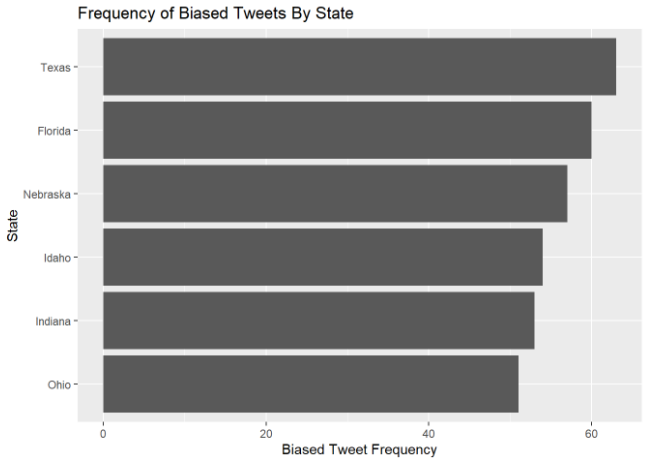
The results for average sentiment score across these same time periods tells a different story. Instead of a gradual decrease in the degree of prejudice per Tweet, there seems to be a gradual increase up until around the beginning of 2022. This implies that while the overall number of biased Tweets may decrease over the course of the pandemic due to reevaluation and further education of the disease, the biased tweets that remain can be, on average, even more biased. This hints at the existence of the mentioned cycle of misinformation that may reinforce problematic opinions in an echo chamber, such as those created by social media platforms. There is also a similar upwards trend that exists from July 2021 to the beginning of the year 2022 that most likely represents the mass infection rates produced by Omicron and the resultant reinvocation of health anxiety.



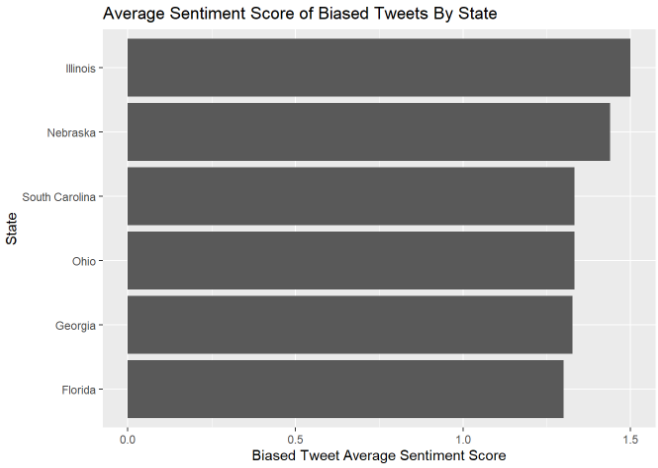
So while the former provides support for part of the first hypothesis (overall decrease in bias frequency), the second result contradicts the latter half of the first hypothesis that believed the degree of bias within tweets (sentiment score) will decrease with time as well.

Geographical Results

With the geographical data of the tweets, one can see trends between United States location and magnitude and degree of prejudice within tweets. The top five states with the highest frequency of tweets with biased language are Texas, Florida, Nebraska, Indiana, and Ohio, each with at least 40 tweets considered biased in the utilized snapshot of COVID-19



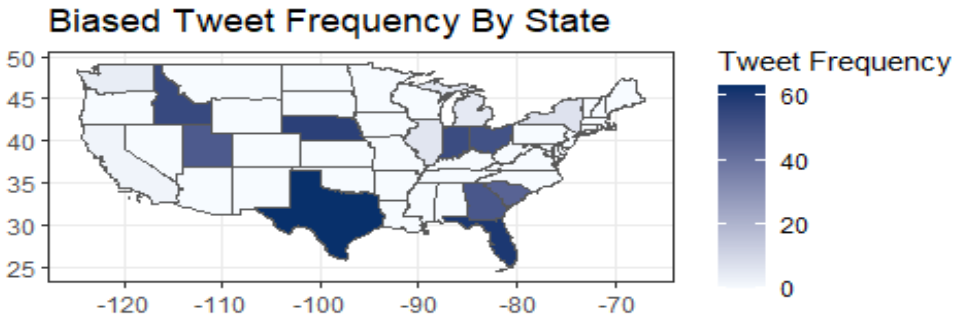
data. Additionally, the top five states with the highest average sentiment score are Illinois, Nebraska, South Carolina, Georgia, and Florida. One concern for this type of analysis is the differing population values among U.S. states. For example, Florida has a much higher population than Nebraska, which could lead to more tweets, and more biased tweets. But we can see from doing both the magnitude analysis and degree of bias analysis that an intersection exists that suggests that political leanings may outweigh the possible effects of population distribution.



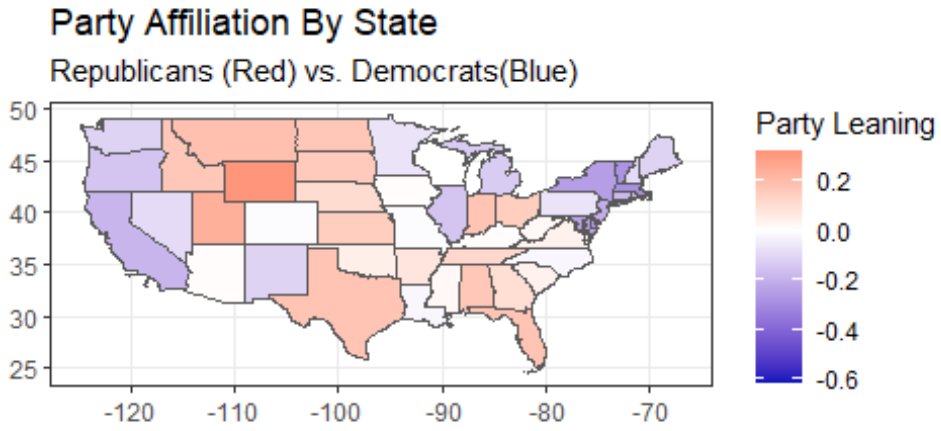
Note: a frequency of 0 does not necessarily imply that some states truly have no prejudice Tweets. As stated, there are limitations in the pulling of tweet data, and smaller periods of time were randomly accessed to attempt to see the breadth of public sentiment. States with a sentiment score of 0 may just not have been recorded extensively within the days collected for this study.

Geographical Results (Continued)

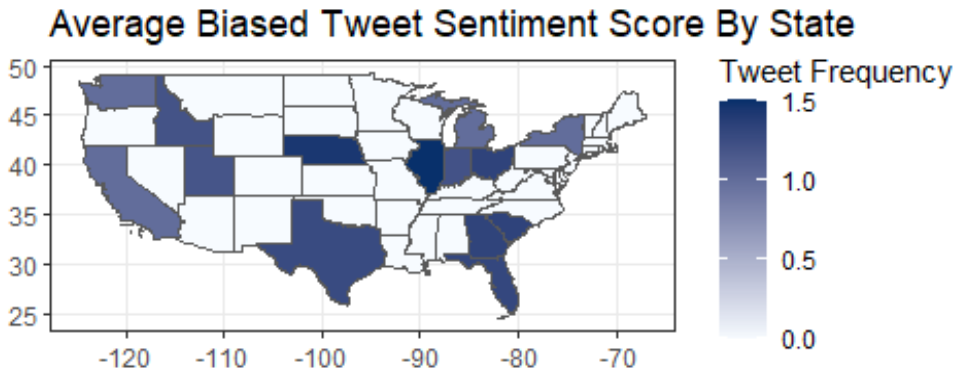
The first visualization below displays the geographical frequency of biased tweets on a state-by-state basis. The darker the fill, the higher the frequency of tweets recorded within the two-year pandemic period. This information can be compared to a map that represents political leanings of states in the U.S.



Upon first glance, one can see that many of the more darkly shaded states are associated with being more Republican leaning states. This can imply that those states with a higher frequency of prejudiced tweets are associated with more conservative ideology. These states not only typically vote Republican but possess Republican leaders and representatives. During the pandemic, many Republicans expressed widespread cynicism towards the pandemic, vaccinations, and China (Knox, 2022). Additionally, the extreme nationalism echoed within the party itself has been shown to lead to and exude xenophobic attitudes that may be perpetuated in states with non-diversified political opinions (Britannica). This implies that states that are more Republican leaning, will express more negative bias towards Chinese people and Asian-Americans based on their tweets during the pandemic.



Similar to the first set of analysis, the average sentiment score of each state does not completely follow this same trend of political party association.



When considering the average, the results are a little more inconclusive when it comes to political leaning. Additional 'blue' states (like California) are now highlighted, and they possess a relatively high sentiment score average. One can see that even though the averages of Democratic states are on par with those of Republican states, the frequencies do not reflect this. This implies there exists possible outlier tweets for these geographical locations that possess higher sentiment scores. So while it may be rarer to find an individual in California that exudes racism against Chinese individuals because of COVID-19, the ones that do exist may be just as biased as those in other states.

From these results, it seems that political leanings of a U.S. state may only imply possible tweet frequency, with the average sentiment score being a little inconclusive when calculated using geographical data. This supports the second hypothesis but contradicts the belief that Republican states will possess both the highest frequencies and highest degree of prejudice, as they only comply with the former. So again, only part of the second hypothesis holds after analysis.

Predicting Degree Bias With A Linear Model to Assess Hypothesis One Uncertainties

Due to the inconclusive nature of the analysis of sentiment score for the second hypothesis, the following model will only predict sentiment scores within Tweets based on political leanings (an extension of geographic location) and a time period (Month/Year). The results show that the variables are not significant, with p-values much greater than the alpha level 0.05. The only significant predictor of sentiment score is whether the tweet was published in November 2021. This once again most likely relates to Omicron and the subsequent infection rate spike that occurred at the end of the year 2021. This mostly likely results from the re-emergence of drastic pandemic policies as well as the increased publication of COVID-19 knowledge and information. This loops back to the study led by Traczyk that highlighted the correlation between fear and surplus pandemic info. This fear could then lead to more bias, as shown throughout this study.

```
##
## Call:
## lm(formula = total ~ month_year + party, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53078 -0.30895 -0.23090 -0.07811  2.76401
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.03046    0.12233   8.424 3.97e-16 ***
## month_yearJul 2021   0.14776    0.08902   1.660 0.097573 .
## month_yearMar 2020   0.04765    0.08655   0.551 0.582170
## month_yearMar 2021   0.12061    0.08079   1.493 0.136086
## month_yearMar 2022   0.11150    0.08485   1.314 0.189421
## month_yearNov 2020   0.04256    0.08651   0.492 0.622983
## month_yearNov 2021   0.34243    0.08795   3.894 0.000112 ***
```

```
## partyRepublican     0.15788    0.11629   1.358 0.175211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5539 on 496 degrees of freedom
## Multiple R-squared:  0.03697,    Adjusted R-squared:  0.02338
## F-statistic:  2.72 on 7 and 496 DF,  p-value: 0.008907
```

Conclusion

The research and results deviated from expectations and usefulness. While the data partially satisfies part of hypothesis 1 and 2, hypothesis 4 remains inconclusive and incorrect.

The frequency of biased Tweets throughout the pandemic was the most consistent variable among all tests and visualizations. It showed a definite spike at the start of the pandemic in March 2020, with a drastic drop and slight fluctuations for the remainder of the two-year period. As mentioned, these spikes most likely correspond to rises in infection rates due to new COVID-19 variants, such as Omicron. When considering the desensitization hypothesis that reinforced the justification for hypothesis 1, it appears that desensitization is a very reasonable explanation for this drop in prejudice tweets. This trend however is inverted when considering the degree of bias within each respective Tweet. This proposes the idea that while overall bias is decreasing, the extent of individual bias may be only reinforcing itself.

The topic of algorithms and moderation policies comes into sharp focus here. The findings suggest that the current moderation frameworks and algorithmic configurations may inadvertently prioritize sensationalist and trending content that, while increasing user engagement, also promotes harmful stereotypes and misinformation. This calls for a more nuanced approach to content moderation, one that can differentiate between harmful content and legitimate discourse without suppressing free expression. This highlights the dangers of misinformation and prejudice, as it is a self-reinforcing cycle that may only serve to enhance incorrect opinions about marginalized populations, and these perspective have no trouble reinforcing in online spaces with personalized content tailored to the user. And while desensitization has been looked at throughout the COVID-19 pandemic, it makes one wonder whether an individual can possess a resistance to desensitization due to their own confirmation bias. If true, this could impact how people approach global, or even personal, crises.

Additionally, while the results suggest that political affiliation also plays a part in the magnitude of biased tweets, it is important to look at all the data. Yes, Republicans and their representatives throughout the pandemic have expressed problematic ideology that enhanced prejudice towards Asian Americans. The mapping provided above displays this in a new light utilizing X as a medium for analysis. But this alone does not include all biases invoked due to the pandemic. One can see that sentiment/bias scores also exist in some of the most Democratic states (i.e. California). But while this may seem less of a concern due

to the drastically lower frequency, it exists as bias, nonetheless. Issues have to be resolved in all edge cases and discrimination must be held accountable in online spaces. Hypothesis 2 displaying some truth provides a starting point for how to address a core of the anti-Asian COVID-19 prejudice, but it does not justify dealing with just one side of the issue. Additionally, as seen with the linear model, party alone cannot significantly predict sentiment scores, implying that other factors and those of many backgrounds can publish and spread hateful content on Twitter.

Takeaways

COVID-19 represents the epitome of a modern medical crisis. On a global scale, the pandemic completely reshaped the lives of every person on Earth. But it represents only one example of how human health has been used to justify prejudice and bias. The research discussed here continues and expands upon research regarding the initial racial effects of COVID-19, and the results display that the prejudice invoked by the pandemic has long-lasting and far-reaching effects. Other diseases like Ebola have also shown localized racial issues when it comes to health. Back in 2014 at the height of the Ebola epidemic in Africa, discrimination and prejudice towards black people drastically elevated. The presence of Ebola increased racial profiling and fed into the already existing biases of Black people in the western world. They experienced being denied admittance to events and institutions on the basis of their skin or accent, but this time it is 'justified' due to the disease. Ebola became completely associated with blackness to the extent of dismissal and discrimination (BBC, 2014). Marginalized and oppressed groups already fight racism that is only enhanced by these instances of disease outbreaks.

Another recent example of a health crisis was the global outbreak of Monkeypox. With most cases being concentrated in gay men, the narrative almost immediately spun into anti-LGBT rhetoric. And while highlighting the groups most at risk is important, attributing and misclassifying an illness based on discriminating against sexual orientation creates further divides and puts such communities at higher risks of hate crimes (Findling, 2022). An extremely harmful aspect of this discrimination was the public misconception that Monkeypox was a sexually transmitted infection. Such incorrect information puts even more people at risk and could lead to a pandemic level threat. An article by Aoife Gallagher shows a spike on X on July 7th, 2022, where there were roughly 2,000 mentions of Tweets intertwining gay men and Monkeypox, as well as citing the gay man pedophile narrative for why children started contracting the disease.

This recurring theme in public health emergencies, where fear and misinformation lead to discriminatory behavior, emphasizes the need for a more conscientious approach to handling information dissemination and public reaction during such times. The role of social media and algorithmic bias in amplifying these prejudices during COVID-19, as detailed in this study, calls for an urgent reassessment of how information is managed and moderated across platforms. By examining the influence of algorithms that prioritize user engagement over factual accuracy, we can begin to understand how digital environments may exacerbate societal divisions in times of crisis.

Moving Towards the Future

An article by Greene and Gullo discusses how recent legislation seeks to tackle the issue of problematic moderation policies. A California bill titled A.B. 587 seeks to hold social media platforms more accountable by requiring extensive documentation on how exactly users can report other accounts, as well as clearly defining their community guidelines. It also requires " large social media companies to semiannually report to the state attorney general detailed information about the content moderation decisions they make," especially regarding "hot button issues like hate speech or racism, extremism or radicalization" (Greene & Gullo, 2024). This includes producing information regarding automated moderation algorithms and how decisions are effectively made in removing, or not removing, content. This law was challenged in court by Elon Musk on the basis of violating the First Amendment Right of platforms to cultivate free speech, but this challenge was dismissed, and the law upheld.

Returning to the actual AI models at play, a paper titled "AI Content Moderation, Racism and (de)Coloniality" presented by Eugenia Siapera critically examines the application of artificial intelligence in content moderation within digital platforms, highlighting a significant gap in the inclusion and consideration of racialized communities in the moderation process. It delves into the ways in which these communities are not only sidelined in defining what constitutes racist hate speech but are also exploited for their labor without compensation to train AI systems, perpetuating existing racial biases and inequalities. By leveraging Anibal Quijano's theory of the coloniality of power, the author argues that current AI moderation practices reflect and reinforce colonial power dynamics, treating racialized groups as both subjects and resources in the digital ecosystem without acknowledging their agency or compensating their contributions. The discussion brings to light the urgent need to rethink how AI models are trained and applied in content moderation. By proposing a shift towards a more inclusive, community-centered approach that values the voices and experiences of those most affected by online hate speech, the author states they could not only reduce biases but also foster systems that support healing and educational outcomes, thereby transforming content moderation into a tool for social justice.

Ethics & Limitations

The methods and data science principles employed within this project do exude a degree of hypocrisy in relation to the criticisms of existing social media algorithms, as there does exist a significant degree of assumption present within language processing. The research here does not confirm any relationships or intent of social media platforms in encouraging or cultivating racism and xenophobia. It is solely meant to highlight possible issues within existing moderation models and to encourage further testing. In order to better explore any relationships, more data must be employed, utilizing the billions of Tweets present within the IEEE database. Also, various inference methods that differ from novice strategies I have personally learned in introductory data science coursework, as well as those that shed

more light on the true relationship between the investigated variables, are necessary to remove 'black box' objectivity of social media algorithms.

In terms of the actual research, the dataset employed, as well as the API utilized to collect said data, of course has its limitations. The COVID-19 tweet data set collected from IEEE possessed nearly 2 billion COVID-19 related tweets, starting from October 2019 to December 2022. While in theory, the totality of this problem set could provide amazing insight, the sheer amount of data points made it unrealistic to use every tweet. Additionally, rate limits with Tweet hydration led to only using around 100,000 tweets from the same randomly selected months (March, July, November). Picking only 90 out of 365 days in both 2020 and 2021, and only 30 days in 2022, creates only a snapshot of possible data that could lead to incorrect and ineffective conclusions. With more time and resources, a true randomization on a daily basis could be much better to capture the true breadth of the tweet content. But this requires more data and more precision when it comes to randomness.

There also existed limitations when looking into geographical data on Twitter. The IEEE has a larger database that contains all COVID-19 related Tweets since October 2019, but as not all geographical data is available for every Tweet, there exists a smaller database with a smaller number of Tweets that this research employed. This ensured that a large majority of the data used had an attached geographical tag. Also, a lot of the geo-tags of biased tweets only related to locations within the United States. So that portion of the research was restricted to the U.S.A. Additionally, there existed a lack of standardization among geo-tags. With some Tweets including just a city, some just a name, some just the country, etc. This was dealt with as precisely as possible, targeting those that at least had state information. These were then standardized for analysis. So some of the state data may also be missing data points.

Lastly, the relatively small bank of Tweets in which analysis was performed did not include a large number of 'biased' tweets. So research was conducted in that small subset so as to not drastically depress values. Due to the nature of text analysis, there is no true way to distinguish between hateful tweets and tweets participating in racial discourse. So in this case, tweets that were not meant as derogatory, but contained the flagged terms, may have been grouped into biased tweets.

References

Centers for Disease Control and Prevention. (2024). United States Covid-19 Cases and Deaths by State over Time. *CDC*. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data.

Dailey, M. (2021). Content Moderation By The Numbers. *NetChoice*. https://netchoice.org/wp-content/uploads/2021/11/Content-Moderation-By-The-Numbers-v5.pdf

Darbinyan, R. (2022). The Growing Role of AI in Content Moderation. Forbes Tech Council. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/?sh=372559a74a17

Findling, M. (2022). COVID-19 Has Driven Racism And Violence Against Asian Americans: Perspectives From 12 National Polls. *Health Affairs*. https://www.healthaffairs.org/do/10.1377/forefront.20220411.655787/

Frenkel, S., & Myers, S. L. (2023). Antisemitic and Anti-Muslim Hate Speech Surges Across the Internet. *New York Times*. https://www.nytimes.com/2023/11/15/technology/hate-speech-israel-gaza-internet.html

Gallagher, A. (2022). Monkeypox and 'Groomers': How Twitter Facilitated a Hate-Riddled Public Health Disinformation Campaign*. Institute for Strategic Dialogue*. https://www.isdglobal.org

Greene, D., & Gullo, K. (2024). EFF Urges Ninth Circuit to Reinstate X's Legal Challenge to Unconstitutional California Law. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2024/02/eff-urges-ninth-circuit-reinstate-xs-legal-challenge-unconstitutional-california

He, Q., & Xie, Y. (2023) The moral filter of patriotic prejudice: How Americans view Chinese in the COVID-19 Era. *PNAS*. https://www.pnas.org/doi/10.1073/pnas.2212183119

Ingram, D. (2023). Fewer People Using Elon Musk's X as Platform Struggles to Keep Users. *NBC News*. https://www.nbcnews.com/tech/tech-news/fewer-people-using-elon-musks-x-struggles-keep-users-rcna144115

Jungkunz, S. (2021). Political Polarization During the COVID-19 Pandemic. *Frontiers*. https://www.frontiersin.org/articles/10.3389/fpos.2021.622512/full

Knox, O., & Anders, C. (2022). Analysis | There's bipartisan potential on China, if the GOP wants it. *The Washington Post.* https://www.washingtonpost.com/politics/2022/12/15/theres-bipartisan-potential-china-if-gop-wants-it/

Lamsal, R. (2019). Coronavirus Geo-Tagged Tweets. *IEEEDataPort*. https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset

Mazer, B. (2022). COVID Science Is Moving Backwards. *The Atlantic*. https://www.theatlantic.com/science/archive/2022/12/covid-science-data-bivalent-vaccines-paxlovid/672378/

McFarland, A. (2022). New Study Attempts to Improve Hate Speech Detection Algorithms. *Unite AI*. https://www.unite.ai/new-study-attempts-to-improve-hate-speech-detection-algorithms/

Party affiliation by state - Religion in America: U.S. Religious Data, Demographics and Statistics. (2021). *Pew Research Center*. https://www.pewresearch.org/religion/religious-landscape-study/compare/party-affiliation/by/state/

Powell, A. (2022). Is pandemic finally over? We asked the experts*. Harvard Gazette*. https://news.harvard.edu/gazette/story/2022/10/is-pandemic-finally-over-we-asked-the-experts/

Ramsey, T. (2012). Fear bias. *RUcore*. https://doi.org/doi:10.7282/T30V8BQ9

Shearer, E., & Mitchell, A. (2021). News Use Across Social Media Platforms in 2020. *Pew Research Center*. https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/

Siapera, E. (2021). AI Content Moderation, Racism and (de)Coloniality. *Springer Link.* https://link.springer.com/article/10.1007/s42380-021-00105-7

Stevens, Hannah R, et al. (2020). "Desensitization to Fear-Inducing COVID-19 Health News on Twitter: Observational Study." JMIR Infodemiology, JMIR Publications Inc., Toronto, Canada. https://infodemiology.jmir.org/2021/1/e26876.

Silver, L., & Delvin, K. (2020). US views of China more negative among Republicans than Democrats in mid-2020. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2020/07/30/republicans-see-china-more-negatively-than-democrats-even-as-criticism-rises-in-both-parties/

Traczyk, Jakub, et al. (2023). "Does Fear Increase Search Effort in More Numerate People? an Experimental Study Investigating Information Acquisition in a Decision from Experience Task." *Frontiers*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01203/full.

Zenone, M., et al. (2022). Title of the Article. *National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10125226/

Zurcher, A. (2014). Ebola, race and fear. *BBC*. https://www.bbc.com/news/blogs-echochambers-29714657