#STOPASIANHATE COUNTERSPEECH ON TWITTER: EFFECTIVENESS OF COUNTERSPEECH STRATEGIES AND GEOSPATIAL ANALYSIS

Md Enamul Kabir

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2023

Committee:

Louisa Ha, Committee Chair

William Sawaya,
Graduate Faculty Representative

Lisa Hanasono

Yanqin Lu

ABSTRACT

Louisa Ha, Committee Chair

This dissertation investigates the effectiveness of counterspeech strategies employed on Twitter in response to anti-Asian hate during the COVID-19 pandemic. This research delves into the communicative strategies, emotional tones, and geospatial distribution of counterspeech, specifically focusing on its effectiveness in the United States. A supervised machine learning was employed to classify counterspeech tweets and counterspeech strategies based on empirical typology. By analyzing 106,388 tweets associated with the hashtag #StopAsianHate collected from November 2021 to May 2022, this research provides insights into the varied effectiveness of counterspeech strategies. The analysis revealed that though counterspeakers were using more negative tones in counterspeech tweets, the tweets with visual media and positive emotional tone received more engagement on Twitter through retweets and favorites compared to those with a negative or neutral tone. This study also breaks new ground by recognizing that higher level of racial diversity does not facilitate higher level of counterspeech against hate speech and hate crime. Additionally, this study highlights the varying degrees of participation in counterspeech across different ethnic groups within Asian American community and underscores the importance of tailored strategies in addressing hate speech. Recognizing this distinction proved essential in crafting evidence-based guidance for community and individual interventions while fostering support from allies of diverse racial backgrounds.

Dedicated to the resilience of the Asian Americans who braved the hatred, and the allies who stood in solidarity by their side.

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Louisa Ha. Without her exceptional mentorship, dedication, and encouragement, this dissertation would have remained a distant dream. Her belief in my abilities have been the driving force behind my work. Dr. Ha's support wasn't just academic; it was the wind beneath my research wings. Her guidance didn't merely light the path; it set the entire journey aglow. For all the times I was lost in a sea of data, her wisdom was my compass. So, here's to Dr. Ha, the maestro of my academic symphony, turning complex research notes into harmonious melodies.

I cannot express enough thanks to the members of my dissertation committee, Dr. Lisa Hanasono, Dr. Yanqin Lu, and Dr. William Sawaya, for the invaluable guidance and support throughout this academic journey. Dr. Hanasono’s expertise in exploring Asian American identity has been a beacon of light illuminating the path of this research. Dr. Lu introduced me to the world of computational method, and continued his guidance with his profound knowledge in the instrument of my research. Dr. William Sawaya’s meticulous attention to detail has uncovered hidden gems within this study. His thoughtful feedback has polished this work to where it is today.

I’d like to thank Dr. Brandon Boatwright, Assistant professor and the director of Social Media Listening Center at Clemson University for generously providing the data for this research. His cooperation and support were invaluable in conducting this research.

I would also like to extend my appreciation to my parents, whose unwavering support and encouragement have been a constant source of inspiration. Their sacrifices and belief in my educational pursuits have been the foundation of my achievements. To my spiritual teachers, Humayun Kabir and Ezaharul Kabir, whose wisdom and teachings have provided me with

valuable lessons, I am deeply thankful. Their guidance extends beyond the academic realm and has contributed to my personal growth. My heartfelt gratitude also goes out to my wife, Jannat Sarker, whose patience and understanding have sustained me during this demanding journey.

To all the individuals who have played a part in my academic and personal growth, I am truly grateful. This achievement is not mine alone, but the result of the collective support and encouragement from these remarkable individuals.

What are we without the people we have?

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1: INTRODUCTION

Hate speech is a growing global concern due to its detrimental effects on targeted groups, including severe mental distress and negative impacts on their collective self-esteem (Boeckmann & Liew, 2002), and is believed to contribute to acts of actual violence acts such as hate crimes (Brown, 2018). Past research showed that those who are the targets of hate offline are also more likely to be targeted in online social communities (Costello et al., 2016). This trend is particularly evident with hate speech directed toward Asians, which has increased exponentially since COVID-19 made news around the world. In the United States, the impact of hate speech toward Asian Americans has been exceptionally shocking and alarming. Along with the increase of COVID-19, a subsequent increase in racial slurs, threats, and physical assaults directed toward Asian Americans across the country was observed (Gilbert, 2020; Haynes, 2020; Capron, 2020). Hate speech directed at Asian Americans, such as "Go back to China," flooded online.

Similar to the transmission of COVID-19, discussions on social media about China's role in the global spread of the virus have intensified xenophobic narratives. The phenomenon was fanned in part by the former president of the United States and media outlets who used the "Chinese Virus" rhetoric incessantly throughout the pandemic. Over a three-month period spanning from March to June 2020, the Asia Pacific Policy and Planning Council and Chinese for Affirmative Action released a report indicating that there were over 2,100 anti-Asian American hate incidents related to COVID-19 that were reported across the country (Donaghue, 2020). The hatred rose to such a high level that the United States Congress took action by passing the COVID-19 hate crimes Act. This legislation aims to enhance efforts for preventing and addressing these hate crimes by increasing oversight and providing more resources. It is important to note that, while this legislation aims to address hate crimes, hate speech is still protected under the First Amendment in the United States (Li, 2022).

Anti-Asian Hate in the United States

Since the outset of the COVID-19 pandemic, Asian American adults have encountered a diverse range of hate incidents. These incidents span a wide spectrum, ranging from non-criminal actions, such as boycotting Asian restaurants, to verbal and physical attacks targeting Asian Americans in public spaces (Lantz & Wenger, 2023). Research has identified at least 82 incidents directly tied to the wearing of face masks. In these cases, Asian individuals faced attacks either for wearing a mask, which was misinterpreted as a sign of underlying illness, or for not wearing one (Ren & Feagin, 2021). This wave of animosity has also spilled over onto social media platforms, impacting the mental well-being, self-esteem, and emotional stability of Asian Americans (Chugh, 2022). Victims often grappled with heightened fear, anxiety, isolation, and disconnection from their own communities, accompanied by feelings of anger and frustration at the discrimination and violence they faced. A prevalent sentiment among many was a sense of powerlessness to make changes and a perception that their voices were not being heard (Perry & Alvi, 2012).

Instances of hate crimes targeting Asians have manifested as physical assaults, verbal harassment, and discriminatory treatment in public and work settings. The rise in such hate crimes can be traced back to the widespread dissemination of xenophobic and racist ideologies, often perpetuated by specific politicians and media outlets. These harmful stereotypes, linking Asians to disease and contagion, have only exacerbated the discrimination experienced by individuals of Asian descent.

The stigmatization of the Asian population in the United States is not a new phenomenon, as it finds its roots in a long-standing history of racist tropes associating Asians with diseases, such as the infamous "Yellow Peril" myth from the 19th century (Del Visco, 2019). This idea is based on the notion that Asians were an existential threat to white Western civilization because of their presumed inherent cultural and biological differences. There have been numerous occasions where Asians were singled out and discriminated against during times of disease outbreaks. The Chinese Exclusion Act of 1882, for example, was created in response to a smallpox outbreak in San Francisco, and enabled the forced vaccination of Chinese residents (Kil, 2012). This act was enforced under the guise of public health while intended to curb Chinese immigration and reinforce racist ideologies. Similarly, in the early 1900s, officials quarantined and burned down Chinatown neighborhoods in response to bubonic plague outbreaks (Barde, 2004). These actions were taken despite there being no scientific evidence to suggest that the disease was more prevalent in Chinatown than in other areas.

Such racist actions reveal the deep-seated prejudices that have historically existed against the Asian community in the United States. These actions have contributed to the formation of stereotypes and have reinforced the idea that Asians are dirty, unclean, and responsible for the spread of disease. In the current era of the COVID-19 pandemic, these historical prejudices have re-emerged in the form of hate crimes, discrimination, and stigmatization of Asian Americans. In responses to the hate speech on social media toward Asians, people post their reaction or opinion against these hate speeches and hate crime incidents, sometimes demonstrating sympathy for the victims. These posts are essentially counterspeech. However, little attention has been given to whether the counterspeech strategies they employed were effective in engaging audiences and mobilize support for Asians.

Purpose of the Study

The goal of this dissertation research was to examine the effectiveness of various counterspeech strategies employed on Twitter in response to anti-Asian hate. This study seeks to explore the communicative tactics and emotional tones used within counterspeech, assess the effectiveness of these response strategies, and examine their geospatial distribution across the United States in relation to the level of racial diversity. This research holds significant importance for several reasons. First, it is evident that xenophobia and hate speech are notably prevalent on social media platforms compared to other communication channels, largely due to the ease of posting and the absence of stringent gatekeeping processes (Barnidge et al., 2019). The COVID-19 pandemic has exacerbated this issue, as social media platforms emerged as prominent sources of unregulated and intense hate speech, particularly due to the reduction in face-to-face communication, which could otherwise help moderate the level of incivility (Papacharissi, 2004; Wilhelm et al., 2020). Moreover, prior research has underscored the profoundly negative impact of hate speech on minority groups, including Asian Americans and African Americans. This sums up the imperative need to identify effective strategies to combat such expressions of hatred (Boeckmann & Liew, 2002).

Significance and Justification of the Study

Past literature has consistently emphasized the importance of counterspeech as an effective initiative to counter hate speech on social media (Briggs & Feve, 2013). In line with this, Mathew et al. (2019) conducted a study focusing on the detection of counterspeech and reached a significant conclusion. They found that while blocking or removing hateful comments may offer immediate relief to the targeted individuals, such actions can have adverse consequences for the larger community. Although some academics have claimed that censorship might help minimize hate speech, there is also concern that it could violate citizens' rights and spread the problem rather than solve it (Strossen, 2018). In the United States, where freedom of speech is protected by law, this scenario takes on another complexity. Considering that censorship would limit freedom of speech, and once something has been blocked or deleted, it cannot be undone, Matthew et al. (2019) recommended counterspeech as the most effective and constitutional approach to counter hate speech.

Previous research has made significant strides in recognizing the significance of comprehending response strategies to combat hate speech on social media platforms (Benesch et al., 2016; Briggs & Eve, 2013; Cao et al., 2022; Garland et al., 2020; Matthew et al., 2019). However, the effectiveness of counterspeech strategies specifically targeting anti-Asian hate speech has not been thoroughly and empirically examined. Counterspeech strategies have been proposed as a means to tackle hate speech by promoting counter-narratives and alternative perspectives. However, there is limited research on the effectiveness of these strategies, particularly in the context of anti-Asian hate speech during COVID-19. Recent research uncovered that there are various strategies used by people online for counterspeech which are not all equally effective and some may reinforce hateful convictions and promote digital mob justice (Matthew et al., 2019). Assessing the effectiveness of various counterspeech strategies against Asian hate speech can fill the gap in research and offer evidence-based guidance for community or individual interventions, as well as proper approaches to allyship from individuals of different races. This research sheds light on those that have proven less effective, and therefore, may be avoided. To further unravel the dynamics of ally support, this research offers a better understanding of how various racial and ethnic groups use the counterspeech strategies.

The recent direction in counterspeech research has used mostly computational methods such as LDA topic models, and BERT techniques to focus on the identification of counterspeech keywords and the study of networks of counter-speakers (Garland et al., 2020; He et al., 2022). For methodological and theoretical complicacy, there has been little research on specific counterspeech strategies (Garland et al., 2020; Kennedy et al., 2017). Garland et al. (2020), in particular, addressed the problem of subjectivity in the automated identification of counter speech which led numerous researchers to follow manual coding of the text. This research, however, aims to develop a supervised machine learning model with counterspeech typologies that have been already discovered by empirical qualitative research (Benesch et al., 2016; Cao et al., 2022) because qualitative studies yield richer and more nuanced findings than topic models (Rossman & Rallis, 2016).

Finally, this research aims to explore the internalized identities of Asian Americans, investigating how different identities and experiences within Asian communities intersect with counterspeech and may vary. The study provides insights into the complexity of Asian American identities from within and how they relate to hate and counter speech. Additionally, legislative attempts to curb hate have been criticized as insufficient. Take the COVID-19 Hate Crime Act, for example—it's seen more as a symbolic gesture than a substantial change. Though it was a much-needed wake-up call, it was insufficient to bring the progress Asian communities require and to mitigate the deeply ingrained hate (Li, 2022). This research can provide valuable insights and recommendations for policymakers to develop effective strategies and interventions of counterspeech that are practically effective. In addition, the geographical comparisons of the counterspeech can also inform the legislation and law enforcement agencies about the occurrences of response and resistance in different locations against anti-Asian hate.

Conceptualization of Hate and Counterspeech

Hate Speech

Hate speech has been conceptualized as an expression that is "abusive, insulting, intimidating, harassing and/or inciting violence, hatred or discrimination." (Nemes, 2002, p. 196). In particular, hate speech includes verbal, non-verbal, and symbolic expressions (Strossen, 2018), acts of bullying, etc. (Chetty & Alathur, 2018). Such hate speech results in fear, intimidation, harassment, abuse, and discrimination and may be traced back to its functioning components, including its emitters, receivers, messages, channels, interactions, impacts, and interpretations, as stated by Paz et al. (2020).

Legal and Scholarly Definition of Hate Speech

There is no established international legal definition of "hate speech," and the description of what is "hateful" is contentious and debatable (Bromell, 2022). Rather than stressing "hate speech" per se, international law prohibits the incitement of prejudice, hatred, and violence (U.N. Office on Genocide Prevention and the Responsibility to Protect, 2019). Legal definitions focus on establishing criteria to identify the hate element. This involves proving the presence of a hate motive or the discriminatory selection of victims (Vergani et al., 2022). For example, the definition of the European Court of Human Rights is notably detailed with potential hate elements, "all forms of expression which spread, incite, promote, or justify racial hatred, xenophobia, anti-Semitism, or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination, and hostility towards minorities, migrants, and people of immigrant origin" (Council of Europe, 1997, p. 107). In a broader sense, legal definitions are geared towards punishment. On the contrary, scholarly definitions of hate speech are diverse and context-specific. It varies based on the specific context of scholars' areas of inquiry. For example, there is gendered hate speech (targeting women or transgender people, etc.), religious hate speech (e.g., targeting certain styles of clothing, practices, hijab, etc.), racist hate speech, hate speech targeting disability, etc. (Chetty & Alathur, 2018). A recent definition by Bromell (2022) framed "hate speech" as "public communication that intends or is imminently likely to incite discrimination, active hostility, or violence against individual persons on the basis of their actual or supposed membership in a social group with a protected characteristic such as nationality, race, or religion" (p. 150). Some scholars, however, emphasize that hate speech can be non-verbal as well. In fact, it may take the form of either vocal or non-verbal expressions of hatred, such as gestures, phrases, or symbols, such as the burning of crosses or representations of members of minority groups as beasts, among other examples (Strossen, 2018; Vergani et al., 2022). Overall, the scholarly definition of hate speech centers on both verbal and nonverbal expressions of hate directed towards a specific person or group on the basis of their identity (ethnicity, religion, sexual orientation, disability, gender, etc.) or perceived identity. Communication scholars stress the importance of analyzing hate speech in both traditional and new media. This helps grasp its nature, origin, its power to gather supporters, and how the audience interprets it (Paz, 2020).

Hate Speech vs Hate Crime

Hate crime, also known as a bias crime, is defined as "a criminal offense committed against a person, property, or society that is motivated, in whole or in part, by the offender's bias against a race, religion, disability, sexual orientation, or ethnicity or national origin" (Lee et al., 2007, p. 275). Hate crime has different features than hate speech. However, both concepts share a common source of being hate-motivated, despite varying the degree of intensity. Whereas hate crime centers around the hatred-influenced action causing physical harm to the victims, hate speech constitutes language, verbal and non-verbal communication, etc. Besides, not all hateful communications lead to real hate crimes, but hate crimes seldom occur without preceding stigmatization and dehumanization of targeted groups and incitement to hate events fueled by religious or racial prejudice, as stated by Rita Izsák (2015), the U.N. Special Rapporteur on minority problems. Thus, the correlation between hate speech and hate crime lies in the form of prejudice. In another way, hate speech is just a gateway offense to hate crimes. Crimes motivated by hatred go well beyond insulting words or non-verbal behaviors and include anything from vandalism and arson to physical violence and even murder. However, hate speech has distinct features, especially in the age of social media. For example, hate speech can be anonymous, and evidence suggests that people are more likely to express hatred and controversy on social media than in person (Brown, 2018). The convenience of instantaneousness makes hate speech unique as well. As compared to hate actions and offenses that have legal concerns attached, with the internet advantage, any hate expression can be instantaneously published on social media and other digital platforms (Brown, 2018). The fact that a hate crime is already legally considered a criminal act, but hate speech involves public communication that, in and of itself, is often not independently illegal, is also a crucial distinction. As a result, those who break the hate crime laws can be punished after the event. If these crimes are committed online, the communications (or content) have reached and the harm done long before the case reaches a court of law.

It’s important to note that the increase in hate crimes often triggers a corresponding rise in counterspeech. Counterspeech encompasses a range of reactions, not limited to denouncing hate speech itself but also condemning hate crimes (Tong et al., 2022). Particularly when a hate crime gains national attention or goes viral on social media, the level of counterspeech may surge as well. Individuals and communities feel compelled to voice their opposition, raise awareness, and promote unity in the face of discriminatory acts. Counterspeech serves as a powerful tool to foster dialogue, solidarity, and resilience against hate, contributing to a collective effort to combat hate crimes and promote a more inclusive society.

Regulating, Censoring, or Criminalizing Hate Speech

Different national legal systems have varying positions on regulating hate speech. In countries like Canada (Nemes, 2002) and Germany (Simpson, 2008), for instance, laws prohibiting hate speech have largely shielded marginalized communities from experiencing trauma as a result of their use of the internet. Germany was an early adopter of legal measures to curb hate speech. Other European countries followed very quickly. For example, a "hate speech" statute was introduced in Russia in 2017, only two weeks after Germany passed the NetzDG hate speech law, and it made direct reference to German legislation. Dozens more nations, including Malaysia and the Philippines, France, Turkey, and Venezuela, have since enacted laws along these lines (Bromell, 2022).

In the United States, on the contrary, the situation is entirely different. Freedom of expression, which includes the right to free speech, is protected under the First Amendment (U.S. Const. amen. I). This legislation complicates any efforts to regulate hate material. At this point, countries like the United States have yet to formulate legal provisions for regulating, censoring, or criminalizing hate speech. It's a common reaction for people to feel that a legislative or legal prohibition is necessary when anything emotionally distressing happens repeatedly. However, some scholars argue that it’s not the government’s responsibility to ban an expression or emotion, and that is not possible in the US due to the protection of the First Amendment. Thus, the debates about censoring hate speech mostly center around the harm and discrimination that hate speech causes.

Egalitarian scholars and ideologists propose their arguments to hold the government responsible for the regulation. This is part of the maximalist approach that focuses on the discrimination that is caused by hate speech and stresses that hate speech is itself discriminatory, or an expressive form of inequality, and, as such, should be legally regulated like other types of discriminatory activity (Elbahtimy, 2021). One of the advocates of such an approach, a critic of the First Amendment, MacKinnon (1993), argued that in silencing, delegitimizing, and marginalizing its targets, hate speech breaches the equal rights of those who are the subject of the speech. As a result, whereas the Fourteenth Amendment was intended to eliminate inequalities and discrimination, the First Amendment has been actively fostering them. In the hate speech literature, this argument is framed as the "silencing effect" (Elbahtimy, 2021). Further, in response to those who claim that protecting minorities from hate speech is a slippery slope to misuse of the law, proponents of hate speech regulation argue that the same danger of misuse applies to any limits on liberties and thus does not justify the tolerance and acceptance of hate speech (Brown, 2018; MacKinnon, 1993). Some egalitarian scholars blame social media companies for not being able to regulate the hate content on their platforms. For example, Brown (2018) argued that if media outlets are at fault for encouraging hatred because they make it easier to distribute written and televised materials, then governments should be held accountable for the conditions in which these outlets function.

A large body of libertarian scholars, however, argue that censoring hate speech is not a viable solution because hate speech laws can be susceptible to abuse of power and are inefficient. It may establish a tendency to restrict speech in ways that are detrimental to the autonomy of speakers and detrimental to democracy (Strossen, 2018). The libertarian argument believes that the potential damage done by acts of expression is "minimal," meaning it poses less of a threat right now and is much more hypothetical than the damage done by hate crimes. One concern that is typically raised in relation to laws censoring hate speech that have been enacted by governments is that laws may be vague and, as a result, chill numerous forms of valuable speech, including, but not limited to, political speech or public discourse that is construed more broadly (Baker, 2012). Brown (2012) argues that if a person of ordinary intellect cannot identify whether or not his or her speech can be considered hate speech, then that person should choose to say nothing at all that may be considered contentious, critical, or provocative to avoid the possibility of undesirable legal repercussions.

A uniquely unorthodox argument is observed in the individualist approach. The supporters of a more individualistic view about rights are generally opposed to any kind of group libel or defamation law, and they reject any attempt to draw parallels between the two and emphasize that incurring culpability under individual libel laws does not apply to cases of collective defamation (Boonin, 2012). Boonin's argument was neither liberal nor conservative, but it places obligations for rights solely on individuals. The communitarian approach, however, opposes individualist viewpoints and argues that individuals have social personalities because they are communally entrenched, meaning that their identities and the identities of the communities in which they participate are complimentary and intricately connected (Elbahtimy, 2021; Freeman, 1995). Negating the idea that only individuals have human rights, Freeman (1995) proposed a theory of collective human rights that suggests free speech should protect the community's most cherished moral and cultural values.

Conceptualization of Counterspeech

Counterspeech, which is an effort to counter online hate speech generated by citizens, aims to prevent the spread of hate speech and change the attitudes of those who engage in it

(Garland et al., 2020). Programs such as Social Media Helpline empower internet users to speak out against online hate by helping them recognize different types of hate speech and respond appropriately. The use of counterspeech is considered to be an effective method for responding to online hate and promoting civil and constructive online interactions. In this paper, counterspeech on Twitter is defined as a direct tweet that counters hateful or harmful speech hashtag related to anti-Asian hate. Concerning the nature of counterspeech in the digital sphere, Mathew et al. (2019) observed that while governments and organizations rarely engage in it, the vast majority of it is generated by users themselves. Recent studies predominantly recommended counterspeech as a response strategy against online hate speech due to its efficiency as opposed to blocking, deletion, and regulation (Benesch et al., 2016; Mathew et al., 2019).

Counterspeech as Response Strategy

The response and resistance to hate speech grew as the pandemic hit with severity. People reacted to anti-Asian American hate by speaking out against it on social media and intervening with open condemnation when they saw it in action (Tong et al., 2022). In particular, Twitter users, as well as numerous public and private organizations, have voiced their support for, solidarity with, or defense of an Asian entity on Twitter (He et al., 2021). Studies have indicated that counterspeech is the most viable and efficient response against hate speech online (Benesch et al., 2022). For instance, Cao et al. (2022) uncovered how people are using counterspeech by addressing the harm of racism online and encouraging them to become involved in the StopAsianHate movement. Benesch et al.'s (2022) qualitative study illustrated eight types of counterspeech strategies people employed, including pointing out the hypocrisy of hate posters and warning of possible offline and online consequences of hate speech, etc. Therefore, counterspeech, which generally refers to people's responses to hostile speech to halt it, limit its repercussions, and discourage it, became a potential technique to combat hate speech without resorting to open censorship. However, despite the prevalence of counterspeech as a method for combating hate speech, there is currently limited quantitative research about the effectiveness of its strategies. Thus, it remains unknown which counterspeech strategy is better received and more successful in lowering hate speech occurrences. (Gagliardone et al., 2015)

Twitter Engagement Metric as Effectiveness Measure

Twitter defines engagement as the accumulation of various interactions a user has with a tweet, encompassing clicks, retweets, replies, follows, likes, links, cards, hashtags, embedded media, usernames, profile photos, or tweet expansion (Twitter, n.d.). Twitter engagement metrics, however, have been used differently in different areas of scholarship. In recent Twitter engagement studies, engagement was typically quantified by considering user-mentions, favorites, retweets, replies, clicks, or detail expansions (Park et al. 2016). For example, a recent investigation characterized engagement metrics in two-way communications as encompassing the count of both incoming and outgoing mentions between users, with mentions including direct mentions, retweets, or replies (Rabarison et al., 2017). Bartlett and Krasodomski-Jones's (2015) study suggested the use of only likes, retweets, and comments to measure the effectiveness of a message. This metric helps gauge the extent of engagement with the content on Twitter.
This type of two-way engagement reflects active participation and interaction between users, which can be a valuable indicator of the effectiveness of counterspeech in fostering dialogue and response. Replies, in particular, signify deeper engagement, as they involve extended conversations. Analyzing reply threads can provide insights into the depth and quality of conversations and whether counterspeech strategies lead to meaningful dialogue. User-mentions can also contribute to the formation of supportive online communities. When users engage with

counterspeech content through mentions and replies, they can be seen as cooperative allies and contribute to community-building efforts (Linvill et al. 2022). However, user-mentions can be a risky measure as they may include spam and trolling behavior, where users may engage negatively or insincerely with counterspeech content (Inuwa-Dutse et al., 2018). These engagements can distort the effectiveness of counterspeech strategies and create noise in the data.

On the other hand, favorites and retweets can be extremely valuable as an effectiveness measure when assessing counterspeech strategies on the platform. It measures the extent of audience involvement and interaction with the content. Higher levels of retweets, likes, and replies signify that the message has successfully captured the attention of users and stimulated their interest or response. In this sense, the effectiveness of a tweet can be inferred from the degree to which it engages the target audience. Benesch and Jones (2019) suggested that by expressing approval through 'likes', counterspeech posts ascend in the relevance ranking, positioning them at the forefront and ideally overshadowing the hateful comments. Additionally, this engagement can be seen as an indicator of message reach and dissemination. When users retweet or share a tweet, they are essentially amplifying its reach by exposing it to their followers. This broader dissemination contributes to the effectiveness of the communication strategy, as the message spreads further and potentially reaches a larger and more diverse audience.

In this research, "effectiveness" is referred to the ability of counterspeech strategies to achieve an appreciable amount of engagement through retweets and favorites, as was used in Bartlett and Krasodomski-Jones's (2015) study. Each of these engagement metrics has its own meaning and can provide insights into the level of interest or engagement with the content of the tweet. For example, “favorites” is an indication that they appreciate or agree with the content of the tweet. On the other hand, when a user comments or retweets a tweet, they are providing their thoughts or feedback on the content of the tweet.

Among the three metrics under consideration, the number of likes stands out as a paramount success indicator. This is primarily attributed to the fact that a high number of likes often signifies a positive reception of the tweet's content by the audience. Moreover, likes hold a unique position as a public metric, accessible to anyone on the platform. To control for the effect of follower size on number of likes, this study employed a weighted measures for likes, replies and retweets which divides the number of likes and retweets and the number of followers of the user. This visibility not only offers insight into the tweet's popularity but also hints at its reach and impact. This practice of weighting likes and retweets has been used by Tohill and Ha (in press) and shows higher engagement level of smaller accounts with fewer followers than those with large number of followers (in press). Twitter's API allows for the extraction of this interactive data in conjunction with the tweets themselves.

What Constitutes an Asian American?

This section aims to delve into navigating the Asian American identity, considering the backdrop of anti-Asian hate incidents. To understand the complexities of Asian American identity, a two-part examination is presented. The first part entails exploring the scholarly and legal definitions that shape Asian American identity, while the second part delves into the cultural heterogeneity within the Asian American community. This exploration seeks to unravel the underlying politics surrounding the inclusion and exclusion of Asian American identity.

Scholarly Definition of Asian Americans

The government-enforced definition of Asian American identity was framed in 1997 as a "person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam" (U.S. Office of Management and Budget, 1997). However, as Lowe (1996) claimed, "the boundaries and definitions of Asian Americans are continually shifting and being contested from pressures both "inside" and "outside" the Asian-origin community" (p. 66), the diverse ethnic groups within Asian American identity are culturally different and distanced. The classification was an attempt to impose homogeneity and bridging the gap between distinct cultures, languages, and experiences among Asians. The necessity of a definition to stock East and South Asians in a homogenous pile has been a political one (Lowe, 1996). Ultimately, these identities can be seen as products of negotiation within the pluralistic fabric of US society, rather than fixed essences (Hall, 1990). In short, an explicit distinction between the definition of Asian and Asian American is evident. Asian American identity can be understood as a "socially constructed unity, a situationally specific position, assumed for political reasons" that facilitates the alliance of diverse groups within the United States (Lowe, 1996, p. 82). The following sections shed light on the perceived identity of Asian Americans through the lens of the 'government-enforced inclusion and exclusion' in this classification.

In a historical turn, the U.S. Supreme Court made a significant decision in 1923 that excluded individuals of Asian descent from being classified as White. It marked an important milestone in categorizing Asian identity. Ironically, during that very same year, the federal government classified immigrants from Iran, as well as other Middle Eastern and Central Asian countries, as White, a classification that was endorsed by the court. The government's role in defining Asian American identity is crucial to consider in this discussion, as the historic decision to classify Asians had significant repercussions and caused confusion and frustration for generations. This classification was primarily a diplomatic maneuver to prevent East and South Asians from being granted naturalization, as only White immigrants were permitted to become US citizens. Asians were excluded from being considered White, thus depriving them of U.S. citizenship for almost two centuries. The official racial classification of Asian Americans was based on geography to a certain extent, but it was a limited and incomplete representation of their diverse identities (Lee & Ramakrishnan, 2020). This led to a profound confusion to how the Southwest Asians and Middle Eastern Americans should identify themselves as they have been classified as White by the US census for a century but were stripped of their perceived whiteness recently (Aziz, 2022). To reiterate, despite being geographically situated in Asia, Southwest Asians, and Middle Easterns were not included in the Asian American identity. Thus, in addition to the scholarly classification of Asian American identity, we also need to take into account how people perceive themselves, as in, self-identification of Asian-ness. The scholarly idea of Asian American exclusion and inclusion has also been impacted heavily by government-enforced classification. Both East Asians (including but not limited to Chinese, Koreans, etc.) and South Asians (e.g., Indians, Sri Lankans, Bangladeshis, Pakistanis, etc.) are considered part of the Asian American identity according to the geographical foundation of scholarly understanding. Recent research, however, has shown that many South Asians experience a sense of exclusion from the Asian identity (Lee & Ramakrishnan, 2020).

In essence, there exists a noticeable disjuncture in how scholars define Asian Americans and how Asians themselves perceive their identity within the United States. This disjuncture can have an impact on the actions and behaviors of Asian Americans. This study addresses this gap or disjuncture to fully comprehend the context and to build research hypotheses regarding their responses to anti-Asian hate speech and hate crimes.

Background of Anti-Asian Hate in the US

Asian American Stereotypes

Research on Asian Americans has identified four common Asian American stereotypes in the U.S.: hypersexuality and submissiveness, model minority, forever foreigner, and yellow peril.

Hypersexuality and Submissiveness. White Americans have a long history of seeing East Asian women as sexual objects. It is a reflection of the belief that Asian American women have little agency in sexual relationships since their bodies are considered venues for male sexual pleasure regardless of the discomfort or lack of fulfillment that the women themselves feel (Wong & McCullough, 2021). They were portrayed as submissive, willing to please sexually, and uninterested in satisfying their sexual demands. Especially, in a pornographic representation study by Mayall and Russell (1993), women of Asian descent were found depicted as hypersexual and eroticized as slaves, submissives, and objects of bondage. That's in part due to the stigma attached to American troops and occupation in the Philippines (post-Philippine-American war), Thailand, and Vietnam (post-Vietnam war), and during the Japanese occupation in Japan in the late nineteenth and early twentieth centuries.

The intersectionality of identities plays a key role in this particular stereotype. Most importantly because it is directed specifically at women of Asian descent and reflects on sexualized racism (Mukkamala & Suyemoto, 2018). Wong and McCullough (2021) claimed that, in a broader picture, East Asian women in America were perceived as hyper-prototypical in terms of sexual representation. A prototypicality is defined as “the extent to which a person or subgroup (e.g., Asian Americans) is perceived to have attributes that best represent a larger group” (p. 88). In line with the definition, East Asian women, in particular, are deemed as overly sexual compared to the prototypical characteristics of Asian Americans.

Model Minority. Asian Americans often face the stereotype of being the “model minority” because of their supposedly high levels of professional success in academia. This is based on the myth that portrays Asian Americans as hard-working, industrious, and technically proficient, which is frequently regarded to be a positive stereotype (Kim et al., 2021). This notion has its origins in anti-Black racism, as it falsely places Asian Americans close to Whiteness. The model minority stereotype has been utilized to sustain the White supremacist narrative by denying the presence of institutional racism, showing that racial inequity in American society is due to individual underperformance, and setting Black Americans and Asian Americans against each other (Yi et al., 2022). The idea of portraying Asian Americans as a model minority was also a political and diplomatic strategy. This narrative was created to bolster US efforts to rally non-western countries away from the influence of communism. Consequently, since the Cold War, Asian Americans started to be more recognized. However, the Asian identity was not promoted in that integration process (Cheng, 2013). Rather, only the American part of their lives was promoted and featured. For example, Sammy Lee, one of the first Asian Americans to join the American Army, was only featured with his hometown in Fresno, California. The idea that Asian Americans are a “model minority” is often viewed as a positive portrayal of this group. However, it fails to take into account the wide range of experiences and identities within the Asian American community. This stereotype can lead to misunderstandings and conflicts between Asian Americans and other minority groups, and it can also create unrealistic expectations for Asian Americans to always excel and be successful (Hanasono et al., 2019). Furthermore, it overlooks the significant obstacles that Asian Americans face when trying to advance to leadership roles in the workplace. This phenomenon is commonly referred to as the “bamboo ceiling” – a term that evokes the idea of a strong and resilient barrier, much like bamboo, that is difficult to break through. The stereotype of the “model minority” paints Asian Americans as successful and high-achieving, but also passive and submissive, which can make it challenging for them to be recognized as leaders and contribute to the persistence of the bamboo ceiling.

Besides, being labeled as a “model minority,” which portrays that Asian Americans have accomplished enough to such extent that they almost reached the white status and are thus immune to racial discrimination in the United States; in reality, Asian Americans continue to face numerous types of racial biases in their daily lives, particularly in the workplace. Roughly one-third of Asian Americans in the professional workforce claimed to have encountered discrimination because of their race in a nationwide study (Kim et al., 2021).

Forever/Perpetual Foreigner. Numerous additional studies have shown that Asian Americans are still seen as “forever foreigners,” in contrast to other prejudices that have become far more subtle in modern society (Li & Nicholson, 2021). In contrast to those of European heritage, persons of Asian descent are seen to be incapable of assimilation, and their allegiance to the United States is often questioned (Ancheta, 2006). The most important trope of this stereotype includes an unsociable image of Asian Americans. Precisely, Asian Americans are seen as awkward, socially inept, emotionally cold, and unattractive, they are disproportionately likely to be socially excluded, ignored by peers, and least likely to be established relationships with (Li & Nicholson, 2021). It is quite ironic the same community was portrayed as a ‘model minority’ in the attempt of assimilating them into the US after the long history of exclusion. It's important to note that, while being extremely diverse, Asian Americans such as Chinese, Indians, Vietnamese, etc. have commonly faced the ‘forever foreigner’ in a broader context. Edward Said’s (2003) classic piece ‘orientalism’ comprehensively blames western ethnocentrism for this matter. Said argues that the West always presents Orientals as “them” as a danger to the well-being of “us” Westerners because of how it positions itself concerning an “exotic” but inferior “Orient” (Said, 2003). No matter how long they have been in the United States, immigrants from Asian nations are portrayed as lesser humans and an eternal danger to white people, hence they and their children are considered “forever foreigners.”

Yellow Peril. In contrast to the other stereotypes that have been highlighted up to this point, the yellow peril represents the most negative and overtly racist stereotype. In this context, Asians are stigmatized as dishonest intruders who spread illness, are seen as culturally and politically inferior to white people and are portrayed as posing a significant risk to white people (Del Visco, 2019). Yellow peril has been used by media to verbally and visually ‘outcast’ Asian Americans for centuries. The historic atrocities against Asian Americans including the Chinese exclusion act of 1882, the denials of their US citizenship, the prohibition on intermarriage between Whites and “Mongolians” by California’s Civil Code of 1905, the killing of Chinese families in Los Angeles, the killing of a Chinese man in Detroit of 1982 have been somewhat influenced by the yellow peril stereotype (Yoo, 2021).

This stereotype was marked by many scholars as a more direct manifestation of a West-constructed idea of Orientalism (Li & Nicholson, 2021; Said, 2003). While the impact of the orientalist attitude is undeniable, as we delve into the origins of the stereotypes, over centuries, we notice a common pattern of politicians and media using these stereotypes as a tool for political gains. For American media and politics, the yellow fear was twofold, centered on i) the “awakened” China as a newly empowered nation and ii) the impact of Chinese immigration to America. Throughout the end of the 19th century to the early 20th century, imagining a Chinese invasion of the United States in US entertainment media was quite prevalent (Lyman, 2000). It’s the same stereotype that made people believe that Asian Americans were collectively into espionage. During World War II, the anti-Japanese tales were reimagined into accusing Japanese people living on the Pacific coast of deliberately placing themselves in that location for the convenience of supplying intelligence about U.S. military operations to Japan (Lyman, 2000). Unfortunately, since the beginning of COVID-19, the model minority depiction has begun to dissipate and the threat of the ancient yellow peril has been resurrected again in American politics and media in the form of a disease-breeder stereotype (Wu & Nguyen, 2022).

Major Events in the COVID-19 Asian Hate Timeline

Throughout the COVID-19 pandemic, there has been a distressing surge in hate crimes and discriminatory incidents targeting Asian communities across the globe. The following timeline in Table 1 highlights notable events that illustrate the extent of Asian hate during this period:

January 2020: The first case of COVID-19 is reported in Wuhan, China.

March 2020: Former President Donald Trump begins using terms like "Chinese virus" and "kung flu" to refer to COVID-19, both on social media and in public statements.

March 16, 2020: In Midland, Texas, a tragic incident occurs where a Burmese family, including two young children, is mistakenly stabbed due to the false belief that they were Chinese and spreading COVID-19.

March 19, 2020: An Asian American woman faces verbal and physical assault on a New York City subway, as an individual wrongfully blames her for the spread of COVID-19.

March 31, 2020: The FBI issues a warning regarding the potential rise in hate crimes against Asian Americans amidst the pandemic.

April 2020: The Asian Pacific Policy and Planning Council establishes a website to document incidents of anti-Asian hate and discrimination related to COVID-19.

June 12, 2020: A Chinese American man becomes a victim of a stabbing incident in New York City's Chinatown.

September 8, 2020: A Filipino American man endures a racially motivated attack with a box cutter on a New York City subway, accompanied by racist slurs and accusations of spreading COVID-19.

February 3, 2021: A 91-year-old man is pushed to the ground in Oakland's Chinatown and tragically passes away a few days later due to his injuries.

March 16, 2021: A shooting spree takes place at three spas in the Atlanta area, resulting in the tragic deaths of eight individuals, including six Asian women. The perpetrator denies racially motivated intentions, claiming to target sex workers instead.

March 30, 2021: In New York City, an Asian woman endures multiple assaults and kicks while bystanders fail to intervene.

April 4, 2021: An Asian American family encounters verbal and physical assault during their vacation in Texas, as the attackers shout racist slurs and demand they go back to China.

May 19, 2021: President Joe Biden signs the COVID-19 Hate Crimes Act into law, aiming to address hate crimes against Asian Americans and Pacific Islanders.

May 2021: A man attacks an Asian family in Texas, killing four individuals, including a six-year-old boy.

June 2021: Anti-Asian hate crimes continue to rise, with several reports of physical assaults, verbal abuse, and vandalism across the country.

July 2021: A group of young people harass and attack an elderly Asian couple at a park in Los Angeles.

August 2021: Anti-Asian violence persists, with incidents reported in various cities, including New York City and San Francisco.

September 2021: Reports emerge of an increase in anti-Asian bullying and harassment in schools as students return to in-person learning.

October 2021: Organizations and community groups continue to raise awareness about anti-Asian hate and advocate for support and solidarity.

November 2021: The Asian American community and allies hold events and memorials to remember victims of anti-Asian hate crimes.

January 2022 - Atlanta, Georgia: In January, a college student in Atlanta, Georgia was targeted and physically assaulted in a hate crime. The incident was widely covered in the media and highlighted growing concerns about anti-Asian violence.

February 2022 - Vandalism and Threats in San Francisco: Multiple reports emerged of vandalism targeting Asian-owned businesses and threatening graffiti in San Francisco. Community members and local authorities condemned these acts.

March 2022 - Congressional Hearings: The U.S. Congress held hearings to address the rise in anti-Asian hate and violence. Several lawmakers, activists, and community leaders testified, shedding light on the urgent need for action.

April 2022 - New York City Subway Attack: A disturbing incident occurred in which a man physically assaulted an Asian woman in the New York City subway. The incident was captured on video and sparked outrage and calls for better safety measures.

May 2022 - Online Hate Speech: Anti-Asian hate continued to proliferate on social media platforms. Activists and organizations raised concerns about the prevalence of online hate speech targeting Asian Americans.

These incidents only scratch the surface of the numerous acts of hate and discrimination experienced by Asian communities throughout the COVID-19 pandemic.

Organization of the Dissertation

This dissertation is organized into five chapters, with the first chapter providing background on the issue of counterspeech and hate speech against Asian Americans in the US during the COVID-19 pandemic. Chapter 1 also explains the significance and purpose of the study. Chapter 2 delves into the previous research on co-cultural theory, social identity theory, and counterspeech strategies to hate both in the US, as well as outlining the rationale for the study's questions and hypotheses. Chapter 3 explains the methodology and approach, including the computational method that was used, the study's variables and how they were measured. The results of the data collection and analysis was presented in Chapter 4, with Chapter 5 offering discussions of the results, limitations, a conclusion, and recommendations for future research.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we embark on a critical examination of the conventional theories previously employed by researchers to comprehend hate and counterspeech dynamics, highlighting their limited applicability to our current investigation. Subsequently, a more fitting theoretical framework is presented that encompasses the social identity theory, Asian critical theory, and intergroup contact theory. These theories offer valuable insights into the multifaceted nature of the research problem at hand. Finally, we conclude the chapter by formulating hypotheses and posing questions that guides our exploration and shed light on the research problem from various angles.

Traditional Identity-Driven Theories

A critique of the conventionally applied theory in the studies of anti-Asian hate, such as co-cultural theory, demonstrates why it is inapplicable to the contemporary minority-minority hate against Asian Americans. Because Asian American hate was grounded in the course of history throughout the last one and a half centuries and observed many shifts and turns through multiple war effects, one single theory isn't sufficient to explain the root of hatred towards Asian Americans. For instance, the co-cultural theory, which has been a favorite of traditional scholars of racism and anti-Asian hate, based on the dominant vs. marginal structure, turned out not to be equipped well to explain the other racial minority groups' hate against Asians.

Co-cultural Theory

The co-cultural theory is one of the most researched theories of the communicative response of minority groups in prior research (Iwamoto & Liu, 2010; Jun et al., 2021; Orbe & Roberts, 2012; Rodriguez, 2012). The Muted Group Theory (Kramarae, 1981), standpoint theories, and cultural phenomenology are the theoretical foundations upon which co-cultural theory is constructed. The co-cultural theory posits that when communicating with members of the dominant group in a variety of settings, individuals who come from marginalized backgrounds create and pick a positioning strategy to use (Orbe, 1998). The presumption behind the co-cultural theory is that every society has a hierarchy that gives some groups more status than others. Members of dominant groups are more likely to hold positions of power because they have access to a wider range of privileges, which they may then use to construct and maintain communication networks that both mirror and bolster their respective domains of expertise (Orbe & Roberts, 2012). Contrarily, to deal with the repressive dominant institutions, members of non-dominant groups, such as racial and ethnic minorities, consciously adopt particular communication practices.

This theory was assumed to be consistent with the Asian American responses to the hate directed by the dominant group. The model in this case worked as the “dominant vs non-dominant” binary model, where white identity was captured as dominant, and Asian American identity was considered a non-dominant/marginal group (Ji & Chen, 2022; Jun et al., 2021). While this model works appropriately in the dominant vs. marginal structure, it is not without limitations. This model automatically puts Asian Americans in a marginalized group that is only affected by the white/dominant groups. This model also assumes that the hate directed towards Asian Americans is another power struggle similar to White vs Black struggles. The co-cultural theory has been applied to understand the experiences of Asian Americans during the COVID-19 pandemic who have been subjected to dual marginalization as a result of their Othered Chinese-ness (e.g., racialized immigrant Others and foreign Asians) and supposed contagiousness (e.g., suspicious, ill, and infectious) (Ji & Chen, 2023). The hatred resulted in an increased awareness of subtle forms of hostility and a preference for avoiding confrontation and assertiveness in the face of danger. Ji and Chen (2023) suggested that the present hate crises in health and racism exacerbated the already severe marginalization of immigrants Others, and Chinese immigrants in particular.

The co-cultural theory has also been applied in the studies of communicative responses of Asian Americans. For example, Jun et al. (2021) found that when reacting to COVID-19 racism, Asian Americans most often utilized the nonassertive strategy, followed by the assertive, and finally aggressive forms of communication. A more passive and less aggressive attitude was linked to having a more prominent ethnic identity and having experienced prejudice in the past. Their study revealed some gender-based insights, such as assertive responses being more common among males. Racial discrimination in the workplace was examined using the co-cultural theory lens, and it was found that the conversational connection between the employee and the employer comes to a standstill should a complaint be filed as part of the response to employee discrimination (Rodriguez, 2012). Besides, an individual’s future communication experiences in their present organizations are irrevocably altered for the worst if they have been subjected to racial prejudice in the past.

Whiteness of European Immigrants. Traditional theories also fail to account for the tension between Asian and European immigrants (Lowe, 1996). A huge influx of European immigrants was observed in the early nineteenth century. The fact that European and Asian immigrants entered around the same time but Asians were mistreated shows that any resemblance to the majority is more accepted. Europeans benefited from the perceived ‘white’ color in America, whereas Asians were stereotyped as “yellow peril” (Chan, 1990). However, studies show that there are causes of anti-Asian hate that co-cultural theory can’t explain. For example, Chan (1990) claimed that anti-Asian hate began at the early stage of Asian immigration, especially for Asians of post-colonial origin.

Post-colonial factors played a key role in the stereotypes and hatred against Asian immigrants. Even a century ago, early immigrants to the United States were treated differently based on their origins. Immigrants especially faced discrimination if their country of origin was a post-colonial country. Chan (1990) explained that the reason behind many Asians, such as Chinese and Filipinos, being mistreated was that they belonged to a post-colonial country that didn’t have a homeland government to defend their interests, in contrast to Europeans. However, such assumptions of co-cultural theory regarding white dominance could not explain the increased hate crime toward Asians. The FBI reports of hate crime statistics from 1990 to 2020 show that the offenders do not belong only to the dominant/white racial group (U.S. Department of Justice, 2021). In reality, many anti-Asian hate crimes are perpetrated by other minority groups. The fact that some of these hate crimes were committed by members of minority groups who are themselves, victims of racial prejudice (for example, Black or African American individuals), provides fertile ground for a theory that isn’t constructed in dominant-marginal group dyads. Thus, an alternative theory needs to be considered that explains the involvement of such non-dominant perpetrator groups. Most importantly, we need to consider the other potential factors, such as perceived socio-economic threat (by non-dominant perpetrators), etc. The FBI statistics on hate crimes are crucial to the understanding of the motivation of hate. In other words, hate crime is the final outcome on the scale of anti-Asian hate dynamics. When we study hate speech responses, strategies can be forged to prevent violent acts at its very doorstep (Izsák, 2015).

Theoretical Framework for This Study

Social Identity Theory

Due to the limitation of co-cultural theory in explaining hate speech and hate crime between minority groups, we may use social identity theory to analyze the responses to such speech and crime by various groups. According to Tajfel and Turner's (1986) social identification theory (SIT), an essential component of an individual's sense of self is derived from the groups to which they belong. To put it another way, individuals form judgments about themselves and others depending on the degree to which they feel a sense of belonging to certain groups and an emotional commitment to those groups. People are considered to be members of the in-group if their characteristics align with those of the group norms, whereas people whose characteristics deviate from those standards are considered to be members of the out-group (Hogg & Reid, 2006). According to the social identity theory (Turner & Oakes, 1986), this group identification is also responsible for influencing intergroup behavior. For instance, individuals who are considered to be part of the in-group will be viewed more positively than those who are considered to be part of the out-group (Tajfel & Turner, 1986). The perception of belonging to an in-group or an out-group may foster partiality in a variety of different ways. The relevance of SIT to the study of communication is in the perspective it offers on how people respond to messages based on their group membership. It's worth noting that people's affiliations inside subgroups of a larger group may become influential in and of itself. In this research, SIT is used to inquire into the many elements of Asian American identity, as well as any possible internalized identities among Asian Americans. Hence, the SIT provides a valuable framework for analyzing how group identification and affiliation impact responses to hate speech. By understanding the role of social identity in counterspeech against anti-Asian hate in the U.S., researchers and practitioners can work towards empowering affected communities, fostering solidarity, and combating hate speech effectively.

SIT has been applied in studies of intergroup conflicts. Ideas borrowed from social identity theory can be found reflected in many studies as well. Jun et al.'s (2021) findings on Asian American responses, for example, strengthen SIT's assumptions, such as Asian Americans with a higher attachment to their social identity place a higher importance on preserving the positive image of their whole ethnic group than they do on preserving their individual image. Supporting or engaging in conflict groups, such as the Irish Republican Army (IRA) or the Ulster Volunteer Force (UVF), is distinctively a manifestation of the social identification process with the group and its goals. As shown in Muldoon et al.'s (2008) research, social affiliation with the in-group is used by young people in Northern Ireland to justify their engagement in paramilitary activities. In the clinical setting, an application of SIT theory found that healthcare workers self-categorize into groups they perceive positively and others less favorably, especially when they believe that their group is at a disadvantage due to an imbalance of power (Bochatay et al., 2019).

Applicability of SIT Theory to Minority-Minority Hate

Multiple models were developed based on social identity theory that explains contemporary anti-Asian hate more appropriately. Among them, one worth noting is the minority-specific model which predicts that hate crimes directed at Asian Americans will have distinctive features in comparison to those directed at other minorities in the United States, such as African Americans and Hispanics. Asian Americans differ from other minorities in appearance, culture, and individual and communal accomplishment. The socio-economic position of Asian Americans is a key factor in this uniqueness. Unlike other minorities in the US, Asian Americans had the highest median household income of $77,166 in 2015, compared to $62,950 for whites, $36,898 for African Americans, and $45,148 for Hispanics). About 21.4% of Asian Americans had a postgraduate degree in 2015, compared to 13.4% of whites, 8.2% of African Americans, and 4.7% of Hispanics. Asian Americans' educational and economic achievements make them a model figure for other minorities by mainstream media, calling them collectively "model minority".

The concept of minority-minority hate, which encompasses hate crimes or tensions between different minority groups, often comes into play due to the complex dynamics of identity, hierarchy, and competition for limited resources. Many hate crimes against Asians are perpetrated not by the dominant group but by other minority groups. This phenomenon can be attributed to individuals from diverse racial backgrounds perceiving the success of Asian Americans as either a threat or a source of envy. As a result, the minority-specific model is particularly relevant when examining anti-Asian hatred. This model also explains the competition that arises among different racial groups during times of economic hardship, which contrasts with the idea of a racial binary of majority versus minority. Essentially, if one group believes that members of another group are attempting to take away their share of limited resources, it can lead to increased racial tensions and hate crimes (Zhang et al., 2022). Along the line, evidence suggest that Black individuals might experience economic rivalry with emerging immigrant communities, leading to a more widespread expression of anti-immigrant sentiment and racism (Demsas & Ramirez, 2021).

It's also crucial to recognize that minority groups within a larger society often do not identify themselves as part of a single, homogeneous minority group. Instead, they maintain distinct ethnic identities and histories. Asian Americans, for example, encompass a wide range of ethnic backgrounds, including Chinese, Japanese, Korean, Filipino, Vietnamese, and many others. Similarly, African Americans and Hispanics consist of diverse ethnic and cultural subgroups. In the United States, much like in other multicultural societies, a hierarchy can be observed among different minority groups. (Nteta, 2014). This hierarchy is not static and can change over time based on various factors, including socio-economic status, historical experiences, and media portrayal. Asian Americans have, at times, been positioned as a "model minority" due to their comparatively higher levels of educational and economic success. This perceived success can create tensions with other minority groups that may not enjoy the same level of positive representation or socio-economic advancement. Most importantly, this hierarchy historically places Black Americans at the bottom tier, experiencing systemic disadvantages and discrimination (Demsas & Ramirez, 2021).

This perspective views opportunities and resources as finite, believing that if one group gains, others must lose. This mindset further intensifies competition and can contribute to tensions between various communities as they vie for their share of the limited opportunities, they perceive to be available. In essence, the racial hierarchy that places Black Americans at a historical disadvantage plays a pivotal role in shaping the dynamics of competition among diverse racial and ethnic groups in the United States. It fosters a sense of inequity that fuels resentment and fuels a perception of a zero-sum game for the resources and opportunities within the society.

Asian Critical Theory

The Asian Critical Theory (AsianCrit) framework, which was anchored on critical race theory (CRT) and the experiences and voices of Asian Americans, has been proposed to counter the traditional white dominance over minority narratives and address the ethnic and historical background of Asians in the US (Iftikar & Museus, 2018; Museus & Iftikar, 2014). While conventional race scholarships and contemporary CRT have not adequately addressed the distinctions between the histories and experiences of Asian Americans and other racial groups, Asian Critical Theory sheds light on the racialization of Asian Americans as perpetual foreigners and the positioning of Asian Americans as "others" in the mainstream White culture of the United States (Huynh et al., 2011; Wu, 2002). This theory serves the need for a conceptual framework that emphasizes the racial realities of Asian Americans (Shin et al., 2022). AsianCrit theory can be applied to understand hate speech and motivation for a response within the Asian racial group. For this, we must delve into the formation and realities of Asian American identity. Tajfel and Turner's (1986) social identity theory allows recognizing the intricate interplay between group identity, racialization, and the broader sociocultural context. Such an integrative approach enables a deeper understanding of the complexities surrounding minority experiences and contributes to a more holistic framework for addressing issues of discrimination and marginalization.

The assumptions of AsianCrit and social identity theory acknowledges that Asian Americans have historically been subjected to racialization as perpetual foreigners. This historical context has had a profound impact on the formation of Asian American identity. The perception of Asian Americans as outsiders in the United States, despite their diverse and often deep-rooted histories in the country, has contributed to a shared sense of "otherness" among Asian Americans. This "othering" plays a pivotal role in shaping their collective identity and influencing how they perceive and respond to hate speech. A notable limitation in the contemporary hate speech discourse is the attempt to explain the response strategies with the assumption that Asian American as one entity or a homogenous racial group whereas the Asian

American identity is heterogeneous and most Asian ethnic groups don't share the same features (Lowe, 1996). Besides, existing literature showed that Asian Americans in general are very prone to identifying with their ethnicity instead of their racial identity (Iwamoto & Liu, 2010; Jun et al., 2021). For example, many South Asians, such as Indians, Pakistanis, and Sri Lankans, are more comfortable being identified with their nationalities. Many Filipino Americans also see themselves as apart from other East Asian communities because of their cultural, racial, and historical differences (Ocampo, 2016). To recap, AsianCrit theory underscores the significance of examining the intersectionality of Asian American identity. Asian Americans often possess multiple identities, such as gender, socioeconomic status, immigration status, and generational differences, which intersect with their racial identity. These intersections influence how Asian Americans experience and respond to hate speech, as their responses may be shaped by the interplay of these various identities.

Intergroup Contact Theory

Intergroup contact theory, proposed by Allport in 1954, posits that close contact and interactions between members of different groups can reduce prejudice and improve social relations. This theory emerged in the context of widespread racial segregation and discriminatory laws in the United States and has since become a widely recognized concept in social psychology. The theory asserts that with certain conditions, including equal status, cooperation, shared goals, and support from authorities, intergroup contact can reduce prejudice and promote positive social relations between different groups (Allport, 1954).

This theory quickly gained traction in the field of social science, inspiring a vast number of studies. It's worth noting that past research predominantly agreed that close contact between intergroup individuals renders positive relations between individuals. For instance, Boisjoly et al.’s (2006) research found that students become more empathetic with the social groups to which their roommates belong. However, subsequent research has called into question the theory's comprehensive effectiveness which led to recent advances in the theory that are fairly practical. Studies such as Dixon et al.'s (2007) and Jackman and Crane’s (1986) have shown that increased contact between groups does not necessarily result in support for policies addressing racial disparities, and may even decrease minority participation in efforts to address interracial tension. Alternative strategies have been encouraged for reducing prejudice and group conflict, as intergroup contact may not be a definitive solution. As Forbes (1997) noted, close contact can alleviate individual prejudice but not necessarily inter-group conflict. This theory has great applicability for explaining the rise of hate speech around COVID-19. Promoting good relationships and interactions between Asian Americans and members of other ethnic groups is proposed as a means of combating anti-Asian hate. One of the key premises of this theory is that more ethnic variety leads to less intergroup hatred. The low population density of Asian Americans in certain areas, according to Zhang (2016), makes intergroup contact theory especially important for understanding how the media might develop negative stereotypes about Asian Americans. In addition, intergroup contact theory is significant to this study since recent developments have made it applicable to contacts between various kinds of social groupings outside the majority-minority structure, such as those framed intergroup interaction based on social identity theory’s in-group vs out-group model (Wright et al., 1997).

The past literature explored the causes of racial hate from various theoretical angles. What's lacking is a framework that explains counterspeech as a response and evaluates how different forms of counterspeech strategies impact anti-Asian hate speech, considering the perspectives of both Asians and non-Asians.

Counterspeech Strategies to Anti-Asian Hate

Counterspeech Strategies of Asian Americans to the Hate Toward Them

There are several factors explaining the response behavior of Asian Americans to hate speech. Membership esteem, for instance, is an indicator of social identification strength that reflects how Asian Americans react to perceived threats to their groups. Boeckmann and Liew (2002) suggested that Asian Americans who value their Asian American identity responded more strongly than those who did not. This is in line with Tajfel and Tuner's (1986) claims of social identity theory, which maintains that just belonging to a group is insufficient to inspire members to act in concert, and the reaction of in-group members with strong membership esteem is usually emotionally stronger.

Several studies concluded that the individuals who were the targets of the hate speech including Asian Americans, Black Americans, Muslims, and homosexuals deemed the hate speech as a unique danger to their ethnic group rather than to themselves individually (Boeckmann & Liew, 2002; Leets, 2002). Asian Americans, in particular, consider hate speech targeting their identity as a more serious crime than theft (Boeckmann & Liew, 2002). However, the counterspeech of Asian Americans to hate speech depends on how they perceive hate speech and stereotypes. When it comes to hate speech using stereotypes, the response of Asian Americans depends on the kind of stereotype used. That is, to Asian Americans, inaccurate negative stereotyping is more offensive than accurate negative stereotyping; as a result, they respond with more sensitivity to the type of hate stereotype (Lee et al., 2007). Lee et al. explained that a negative but accurate stereotype may include the expression "Asians with slanted eyes".

Jun (2012) claimed that Asian Americans respond to hate speech in either non-assertive, assertive, or aggressive approaches. Asian Americans who choose the non-assertive response do so due to peer pressure, lack of experience, safety concerns, and deciding that an assertive response is not worth it. The study lends support to a previous finding that many Asian American victims choose to take a passive stance rather than taking any kind of offensive action (Leets, 2002). However, some Asian Americans also take assertive action “expressive communication about the incident, directly questioning the aggressor about his/her behaviors, making an official complaint, demanding an apology, clarifying the intention of the act, and telling the aggressor what was wrong without insulting the aggressor” (Jun, 2012, p. 342). Finally, verbal or nonverbal attacks, undermining the aggressor, and humiliating the aggressor are all examples of the aggressive response of Asian American that were found only in a few cases. Some Asian Americans who deem hate speech as being exceedingly insulting sought to respond collectively by taking actions such as tightening legislation against hate crimes, etc. (Lee et al., 2007). However, since there are so few Asian Americans in positions of power in the legislative process in the United States, these collective reactions have not yet produced positive results.

Asian Celebrities for Viral Circulation

In addition, Asian Americans now learn to use celebrities’ influence to promote the cause of anti-Asian hate. Many well-known Asians in the entertainment business have taken to Twitter to express their solidarity with the Asian community and call attention to the issues that Asians are confronted with in modern society. Because of the fame that celebrities have, their fans and followers that have provided them with a significant quantity of feedback on their postings. This eventually results in the viral circulation of their material as fans and followers repeat the first tweets in large numbers (Tong et al., 2021).

Non-Asian Americans' Strategies - Allyship

People in general responded to anti-Asian hate using verbal expression of critique on social media, and bystander intervention that support the victims or the anti-Asian hate movement. Many individuals have publicly denounced the racist behavior that was sparked by COVID-19, and they seek to put a stop to the hatred (Tong et al., 2022). Non-Asian Americans on Twitter have either (a) publicly supported, shown solidarity with, or defend an Asian entity, or (b) explicitly recognize racist, hateful, or violent behavior directed at an Asian entity and either call it out, critique, denounce, challenge, or oppose it (He et al., 2021). This is called the allyship strategy. The allyship strategy is shown in digital platforms' civic intervention and counterspeech of social media users.

Digital Platforms' Civic Intervention. Digital service providers and social media companies are taking action to counteract the spread of dangerous user-generated hate speech. In order to enforce their own community rules and terms of service, platforms depend largely on their users to step in and report information that is deemed inappropriate (Bromell, 2022). For example, in detecting hate speech, Facebook employees follow their determined framework of hate speech, such as "content that directly attacks people based on their race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender or gender identity or serious disabilities or diseases" (Facebook's Community Guidelines). However, these responses are often ill-timed and reliant on Artificial Intelligence in an excessive manner since they handle billions of data in the process. Another way many young women are responding to hate on social media is by volunteering as moderators on Instagram to clean up tagged photographs. These volunteer moderators can drown out digital vandalism, distressing photos, and harassment by flooding a user's tagged photos with pleasant images. This, in turn, drives graphic material and disrespectful memes to the bottom of an account, where they may be removed by official moderation procedures. (Farokhmanesh, 2019).

Counterspeech Through Allyship. Many organizations were seen demonstrating their support for Asian Americans, denouncing the hate. Most notably, public libraries across the US put out statements including strategies to combat the hate. Among the most insightful of such statements, as a recent study found, were a strengthened pledge to EDI (equity, diversity, and inclusion) and social justice, as well as expressions of solidarity with Asian American communities and condemnation of hate incidents and the dissemination of relevant historical context and informational resources (Chae, 2022).

As part of the counter-hate content, both Asian Americans and allies used counter hashtags as a response to anti-Asian hate speech on social media, He et al. (2021) found that the most popular counter hashtags that people used were #IAmNotAVirus, #WashTheHate, #RacismIsAVirus, etc. in response to the hate hashtags, such as #ChinaVirus, #ChineseVirus, Chinese virus, #FuckChina #ChineseBioterrorism, #KungFlu, #MakeChinaPay, #wuhanflu, #wuhanvirus, etc. The hashtag #StopAsianHate, however, remained one of the most prominent ones associated with counter speech tweets on Twitter and other social media platforms, particularly in the strategy of allyship in responding to anti-Asian hate speech (Lyu et al., 2021). Hashtags such as #StopAsianHate, #StopAAPIHate, etc. had reached a global hit through Twitter and people who were protesting against the discrimination towards Black people and police brutality merged their movement with #StopAsianHate to add support. Consequently, hashtags #StopAsianHate and #BlackLivesMatter joined in allyship (Chang et al., 2021). Especially on social media (e.g., Twitter, Facebook), people created visual posts of solidarity between the discrimination against Black people and Asian Americans. Tong et al. (2022) observed a similar pattern of allyship from protesters of the #BlackLivesMatter movement responding against anti-Asian hate. For example, one user expressed the pride of their allyship stating “I’m so proud of Houghton. There were more than 1000 people out in the streets today making their voices heard. More was supporting from their cars”. These tweets may either be direct responses to tweets that promote hatred, or they can be tweets that stand on their own. The following is an example of a tweet that fits under this category: “The virus did inherently come from China but you can’t just call it the Chinese virus because that’s racist. Or KungFlu because 1. It’s not a f*****g flu it is a Coronavirus which is a type of virus. And 2. That’s also racist.”

Effectiveness of Counterspeech Strategies

The effectiveness of counterspeech has been explored in several studies, particularly in cases where a counterspeaker directly engages with individuals who have posted hateful or dangerous messages, aiming to influence their opinions or behavior. For instance, Miškolci et al. (2018) discovered that responding directly to the original speaker did not effectively halt their behavior of posting hateful content. However, it did serve as an effective means to reach a broader audience and stimulate additional counterspeech. In this way, the act of countering hateful or dangerous messages through direct response extended beyond its immediate impact on the original speaker. It served as a mechanism to amplify the message and provoke a broader discourse within the online community. By capturing the attention and involvement of a larger audience, counterspeech had the potential to foster a collective response that addressed the harmful content and promoted alternative viewpoints (Buerger, 2021). This is particularly relevant to the issue of anti-Asian hate. In the case of anti-Asian hate speech on Twitter, expanding the reach of counterspeech is crucial for several reasons. First, it allows the message of tolerance, inclusion, and support for the Asian community to permeate beyond the confines of individual interactions, reaching a wider pool of users. By raising awareness and provoking discussions, counterspeakers can foster a collective response against hate speech, encouraging more individuals to join the conversation and take a stand against discrimination.

Dillion and Bushman (2015) ran an experiment to see if offline theories of bystander intervention apply to online contexts, specifically whether cyberbullying awareness predicts intervention. They discover that it does, and such intervention can be direct and indirect (68%) and can occur after the danger has gone. Some social media sites have advocated this practice as a means of eliminating online hatred, since, as Justice Louis Brandeis said, the antidote for bad speech is good speech. Especially because speech is protected by the first amendment in the United States, hate speech can't be regulated. This leaves counter-hate speech as only one viable option, and that is counterspeech or response by common people. Bartlett and Krasodomski-Jones (2015) emphasized counterspeech as a widely utilized, community-driven response to extremist or hate content. Such posts are frequently met with opposition, ridicule, and counter-initiatives. Past studies predominantly supported counterspeech over censorship or regulation to combat hate speech. Briggs and Feve's (2013) study, however, also supported the predominant narrative in support of counterspeech, emphasized that government strategic communication, alternative narratives, and counter-messaging should be used.

Cao et al.'s (2022) qualitative study analyzed #StopAsianHate in May 2021, during the timeframe when President Joe Biden signed the COVID-19 Hate Crimes Act. The study discovered the themes of counterspeech at that timeframe, such as 1) discourse that anti-Asian hatred is not a new phenomenon, recognizing and addressing the negative impacts of racial bias, 2) encouraging participation in counter hate movements, such as #StopAsianHate, promoting the recognition and appreciation of the Asian American and Pacific Islander community's culture, history, and contributions, and finally, awareness and visibility of the Asian American and Pacific Islander community. So far, however, the effectiveness of discourses such as counterspeech has not been empirically assessed. Benesch et al.'s (2016) study identified eight non-mutually exclusive strategies of counterspeech to hate speech, such as 1) presentation of facts to correct misstatements or misperceptions, 2) pointing out hypocrisy or contradictions, 3) warning of possible offline and online consequences of speech, 4) identification with the original speaker or target group, 5) denouncing speech as hateful or dangerous, 6) use of visual media, 7) use of humor, and 8) use of a particular tone, e.g., an empathetic one.

In the context of anti-Asian hate, social identity theory helps us understand how Asian Americans, and non-Asians, may engage in counterspeech. Asian Americans share a common social identity as a minority group, and their sense of self is influenced by this group membership. According to SIT, individuals derive a sense of belonging and emotional commitment from their in-group, which, in this case, refers to the Asian American community. When faced with anti-Asian hate speech, Asian Americans may feel a stronger emotional connection to their in-group and a heightened sense of solidarity. This can motivate them to respond through counterspeech, aiming to challenge and counteract the hate speech. By engaging in counterspeech, Asian Americans reaffirm their collective identity, demonstrate support for their community, and work towards minimizing the impact of hate speech. SIT also highlights that individuals have multiple social identities that intersect and influence their experiences and perceptions (Cao et al., 2022). People from other racial groups may have intersecting identities, such as being members of racial minority communities or having experienced their own forms of discrimination (Tajfel & Turner, 1986). This shared experience of navigating multiple identities can create a sense of commonality and encourage allyship towards Asians who are targeted by hate speech. Intersectional solidarity enhances collective psychosocial resilience in the face of anti-Asian racism during the COVID-19 pandemic (Cheng et al. 2021). By acknowledging and embracing the interconnectedness of diverse identities and experiences, individuals and communities can unite in their shared struggle against prejudice and hate in general. Therefore, counterspeech strategies, including 'allyship and affiliation' and 'denouncing the hate speech or act rather than the actor,' become relevant in this context.

However, in assessing the effectiveness of response strategies, the literature generally indicates that the counterspeech strategies are not all equally effective, and the language choice of users posting counterspeech is largely different from that of those posting non-counterspeech (Matthew et al., 2019). Matthew et al.'s (2019) study revealed that counterspeech comments receive more audience engagement, such as likes and replies, than non-counterspeech comments. However, which strategies of counterspeech are more effective remains unclear. Benesch et al. (2016) warned that some strategies of counterspeech, while they may have the intention of countering hate, without empirical evidence-driven guidance, may serve to strengthen hateful convictions and advocate mob justice in the digital realm. When it comes to countering anti-Asian hate, there may exist a research gap in understanding which specific counterspeech strategies are most effective in addressing this particular issue. Exploring this gap could shed light on developing impactful approaches to combating anti-Asian hate speech effectively. Hence, this research examines the effectiveness of the response strategies based on their audience engagement metrics (number of likes, retweets, and replies).

Additionally, while counterspeech encompasses a wide range of responses to hate speech, it is important to note that not all counterspeeches focus solely on denouncing hate speech. Many individuals and communities use counterspeech as a means to specifically condemn hate crimes. In these instances, the emphasis is on addressing the tangible acts of violence, discrimination, and harm inflicted on individuals or communities based on their race, religion, ethnicity, sexual orientation, disability, or other protected attributes. By denouncing hate crimes through counterspeech, individuals aim to draw attention to the seriousness of these acts and express their solidarity and support for the victims. This distinction highlights the multifaceted nature of counterspeech and its ability to address various dimensions of hate and intolerance in society. With the lack of research on the effectiveness of specific counterspeech strategies and the differentiation of hate speech and hate crimes, the following research questions are posed in this dissertation study:

RQ1: Which counterspeech strategies are more effective on Twitter?

RQ2a: Which counterspeech strategies are more associated with hate speech?

RQ2b: Which counterspeech strategies are more associated with hate crimes?

Emotional Tone of Counterspeech

Emotional tone is one of the eight strategies in the counterspeech taxonomy proposed by Benesch et al. (2016). Studies suggested that Asian American victims used several types of emotional responses (e.g., anger, outrage, sadness, fear, etc.) when exposed to hate speech (Barnes & Ephross, 1994). Both positive (such as pride and thankfulness) and negative (such as rage, grief, and terror) emotions and sentiments have been observed on Twitter in the counterspeech against hate speech. Tong et al. (2021) found that the majority of those who participated in response to anti-Asian hate speech were mostly sorrowful. People showed compassion for those who had been victimized by hate crimes and for the Asian community. For example, one user expressed their desire to put an end to the growing number of serious anti-Asian incidents by stating, “It breaks my heart and makes me sick to think of all the horrible things that have been inflicted upon Asian people.” This individual hoped to put an end to the increasing number of serious acts of anti-Asian hate online. Some users showed more assertiveness with rage in their response to the hate speech, for example, "If y’all followed the precautions, then the virus wouldn’t (as) be bad as now. Don’t blame Asians for it.” Studies in the past have pointed out that the tone of a counterspeech or response to a hate message determines whether the encounter has a detectable effect (Bartlett & Krasodomski-Jones, 2015; Frenett & Dow, 2009).

The emotional tone of counterspeech has also been emphasized in Benesch et al.’s (2016) and Mathew et al.’s (2019) research as an integral part of the counterspeech strategy. Mathew et al. (2019) measured the tone in two separate forms: positive tone (tweets that are empathic, kind, polite, or civil) and hostile tone (tweets that are abusive, hostile, or using obscene language). When it comes to the effectiveness of tones, Frenett and Dow (2009) discovered a strong correlation between the message's tone and the engagement rate. Aggressive messages, for instance, had a zero percent engagement rate. However, casual or heartfelt texts inspired 83% of recipients to react. Similarly, highlighting the negative repercussions of someone's hateful speech was less likely to result in prolonged participation than offering aid or sharing personal tales. Gruzd’s (2013) study also revealed that Twitter users naturally gravitate towards posting positive messages, with positive tweets being three times more likely to be shared compared to their negative counterparts. This robust empirical evidence underscores the persuasive impact of positive emotional expression in fostering engagement and dissemination within online communities. Zavattaro et al. (2015) suggested that a positive emotional tone can increase citizen engagement and participation on Twitter.

Positive emotional content also has a higher likelihood of going viral on Twitter (Berger & Milkman, 2013). Users who recognize this potential may share content in a tone that they believe has a higher chance of reaching a broader audience. Virality refers to the rapid spread and widespread sharing of content, such as articles, videos, or messages, among a large online audience. In the context of this research, achieving virality can be essential for the effective dissemination of counterspeech against hate speech. When counterspeech goes viral, it reaches a broader audience, raises awareness, and has a more substantial impact on combating hate speech and promoting positive messages. Virality can amplify the reach and influence of messages on social media platforms like Twitter, making it an important consideration for any communication strategy.

Furthermore, past studies suggest that the design of the Twitter user interface also plays a pivotal role in influencing these patterns. By discouraging the posting and sharing of negative messages, the platform inadvertently encourages a more positive online discourse (Gruzd, 2013). In essence, this highlights the role of the platform's design in reinforcing the effectiveness of a positive emotional tone in online communication.

Thus, as corroborated in past literature, it is evident that not all messages possess equal potential to achieve virality, increased participation, and resonance (Berger & Milkman, 2013; Frenett & Dow, 2009; Gruzd, 2013; Zavattaro et al. 2015). This study further reinforces the notion that the manner in which something is expressed (tone) on social media is as pivotal as the substance of the message itself. Particularly within the Twitter community, it becomes apparent that a positive tone employed in a campaign holds greater promise for broader reach compared to a negative tone. Thus, this study hypothesizes the positive effect of a positive tone in counterspeech,

H1: Counterspeech with a positive tone is more effective than a negative tone.

Counterspeech from South Asian Americans

Social identification theory (SIT) emphasizes that within larger groups, subgroups can also play a significant role. Asian Americans encompass diverse subgroups with distinct cultural backgrounds, experiences, and identities. These subgroups, such as Chinese Americans, Korean Americans, or South Asian Americans, have their own internal dynamics and affiliations within the larger Asian American community. Therefore, SIT directs towards a deeper understanding of the specific subgroup affiliations and how they may influence responses to anti-Asian hate speech. It’s important to note that South Asian immigrants are primarily newcomers to the United States, with their arrival occurring more recently compared to East Asian or other Asian American groups (Shankar & Srikanth, 1998). Their relatively recent migration suggests that they may not have fully integrated into the broader Asian American identity due to distinct cultural backgrounds and experiences (Shankar, 1998). Kibria’s (1996) commentary underscores the understanding of race, which varies among different Asian American groups, particularly among first-generation immigrants. It highlights the existence of unconscious biases that may influence perceptions of race and racial identity. This may imply that South Asians could be influenced by their unique cultural backgrounds when it comes to identifying as Asians. Shankar and Srikanth (1998) pointed out that, on the ground, there are multiple and interacting levels of ethnic identity formation. These understandings are often influenced by social location and are not uniform across different Asian American groups. The fact that different Asian American groups maintain their distinct understandings of race based on their social and cultural backgrounds can suggest that South Asians may also retain their unique ethnic identity that sets them apart from other Asian Americans.

The structural context in which more than thirty groups of Asian descent are lumped together as Asian American can contribute to differences in understanding race and identity. These differences may hinder bridge-building efforts among different Asian American groups, further suggesting that South Asians may not necessarily identify strongly as part of the broader Asian American identity. Morning's (2001) study indicates the reluctance of South Asians toward a generational divide. Unlike the first-generation South Asians' encounters with discrimination and deliberate avoidance of conventional American racial classifications, the second-generation South Asians who were born in the United States were more likely to choose White or Black as a racial category than Asian Americans (George, 1997). However, in both generations, the distinction reflects the South Asian community's reluctance to embrace the label "Asian." Kibria (1996) maintained that "I remain pessimistic about the meaningful inclusion of South Asians into the pan-Asian fold as long as the issue of race is avoided by members of the pan-Asian movement." (p. 84).

Existing literature further shows that Asian Americans in general are very prone to identifying with their ethnicity instead of their racial identity (Iwamoto & Liu, 2010; Jun et al., 2021). For example, many South Asians, such as Indians, Pakistanis, and Sri Lankans, are more comfortable being identified with their nationalities than their racial identity as Asian American. In fact, in the wake of the COVID-19 pandemic, many South Asians felt excluded from their racial identity (Lee & Ramakrishnan, 2020). Reports from South Asian Americans have highlighted their sense of exclusion from the broader Asian American identity due to distinctions in culture, religion, and racial/phenotypic traits. This exclusion has led to a noticeable absence of their representation in Asian American studies, narratives, and media depictions. In a similar vein, Southeast Asian Americans have expressed their experiences of being regarded as "other Asians" and being subjected to stereotypes that position them as inferior when compared to East Asian Americans (Nadal, 2019). The sense of exclusion was not only self-perceived by South Asian Americans but also observed in media publications. Ramakrishnan (2023) critiqued the New York Times for assuming that East Asians represent the majority of the Asian American community. In a 2016 opinion video about the victims of racial slurs, nearly all the Asian Americans featured were East Americans and no South Asian Americans.

Precisely, there is a disjuncture between the legal definition of Asian American and the perceived Asian American-ness that plays a key role in the counterspeech to anti-Asian hate. In the United States, racial self-identification (how an individual defines themselves) and perceived race (how the race of an individual is classified by others) are both very significant, and they do not always correspond with legal classification (Roth, 2018). In particular, in situations that place a strong emphasis on the distinctions between groups, social identity is most likely to come to the forefront and actively shape behavior. Thus, people who strongly identify with their social identity may respond in a more extreme fashion to hate speech directed at their group than individuals who are less invested in their group identity (Boeckmann & Liew, 2002). For instance, a South Asian clinical psychologist and co-founder and director of the Center for Cognitive Behavioral Therapy and Mindfulness, Dr. Suraji Wagage, explained, "I personally always feel uncomfortable checking 'Asian' when required to select my ethnicity on a form," “#StopAsianHate is a movement I stand in solidarity with, but am not central to” (Jagoo, 2022). Thus, as recent studies showed the perceived exclusion of South Asians of Asian identity (Lee & Ramakrishnan, 2020), the question remains, then, whether a Pakistani or Indian would similarly respond to the hate as an East Asian (e.g., Chinese, Taiwanese) would. To understand the salience of ethnic identity, social identity theory is more effectively fitted than co-cultural theory. To further expand AsianCrit theory's critique of Asian American realities, a clear difference is observed between the stereotypes directed at South Asian Americans and East Asian Americans. For instance, the most popular Asian stereotypes, such as disease breeder, model minority, yellow peril, etc., are only directed against East Asians, not South Asian Americans (Ibrahim et al., 1997; Oyserman & Sakamoto, 1997). The recent surge in COVID-related hate has predominantly targeted East Asians, primarily affecting individuals of Chinese, Japanese, and similar backgrounds. Considering how victims respond to hate, it leads us to wonder if all ethnicities within the Asian American category can engage in counterspeech at the same level. Given that the prevailing COVID-19 stereotypes disproportionately target East Asians and the evident disconnect South Asians feel from Asian identity, it is only plausible that many South Asians may not be as motivated to respond with the same degree of emotional intensity in their counter-messages, or they may opt not to respond at all.

However, this particular investigation becomes quite intricate due to the anonymity maintained by certain social media users concerning their ethnic identity. This anonymity introduces an additional layer of complexity, making it challenging to discern the ethnic backgrounds of individuals participating in online conversations about hate. On the other hand, influencers on Twitter (more than 10,000 followers) have more public information about their race and ethnicity, which can be a useful source to understand their counterspeech practices. The influential role played by social media influencers in driving and shaping public discourse is also a crucial factor to take into account. These influencers possess a substantial reach and impact within the digital sphere, which extends to issues like the #StopAsianHate movement. As they wield considerable influence over their followers and the broader online community, their actions and responses can significantly shape the tone and direction of discussions surrounding hate incidents. Therefore, the potential impact of these influencers must be thoroughly considered in understanding how responses to online hate manifest. Hence, this study posits the following hypothesis and research questions,

H2: South Asian American influencers are likely to exhibit lower participation in counterspeech compared to their East Asian American counterparts.

RQ3a: Is there a difference in the counterspeech strategies employed by Asian and non-Asian American influencers?

RQ3b: Is there a difference in the counterspeech strategies employed by South Asian and East Asian American influencers?

Geospatial Mapping of Counterspeech

There is a widespread belief that ethnically diverse cities are safer for minorities like Asian Americans, based on the assumption that there will be less hatred towards minorities and more resistance to hate. This assumption was predicated on majority-minority models for the most part and came from intergroup contact theory (Pettigrew, 1998). It might not be equally true in all cases. In reality, racially diverse spaces may not mean more resistance against anti-Asian hate, especially since many hate incidents against Asian Americans were perpetrated by non-whites (U.S. Department of Justice, 2021). Some of the perpetrators, in fact, were from other minority groups, which is the same reason why many traditional theories can’t explain the Asian American hate dynamics.

Studies also showed that a perceived increase in diversity makes many White Americans feel threatened and evokes expressions of explicit and implicit prejudice (Craig & Richeson, 2014). According to Outten et al. (2012), White Americans who were exposed to an article depicting a future racial diversity in which their racial group constitutes less than 50% of the national population demonstrated a heightened concern about a perceived threat to their societal status. This, in turn, led to stronger racial identification and more negative emotions towards the outgroup. Oliver and Wong (2003) found that more ethnic diversity at the metropolitan level was associated with higher levels of prejudice, but that the same diversity, when studied at the neighborhood level (a level more likely to enable positive interaction, such as intergroup friendships), was associated with lower levels of prejudice. The findings of Oliver and Wong's research suggest that the level of residential segregation of racial groups may play a role in shaping levels of hate. At the metropolitan level, where there is greater overall ethnic diversity but potentially less direct interaction between different racial and ethnic groups, hate tends to be higher. This may be due to limited opportunities for meaningful contact and understanding between diverse groups, leading to stereotypes, biases, and increased prejudice. On the other hand, at the neighborhood level, where individuals from different racial and ethnic backgrounds are more likely to have positive interactions, such as forming intergroup friendships, levels of hate tend to be lower. Thus, Asian Americans may find more allies and friends in smaller communities than at a large metropolitan level. In other words, Asian Americans who migrate to a big city with higher racial diversity, assuming that there may be more resistance to hate and more allyship and empathy, may find it quite the opposite. Thus, this research also examines the relationship between geographical location’s racial diversity and the respective use of counterspeech strategies following the research questions.

RQ4a: Is there a relationship between the racial diversity of the counterspeech location and the counterspeech occurences?

RQ4b: Is there a relationship between the racial diversity of the counterspeech location and the specific counterspeech strategies used?

Moreover, while there have been studies mapping hate crimes and hate speech geographically in recent years (Jendryke & McClure, 2019; Lingiardi et al., 2020), there is a notable gap in geospatial mapping based on counterspeech. This gap is significant in understanding the readiness of different counties and cities to promote a mindful community and address hate incidents effectively. By examining the prevalence of hate and the responses to it across various cities and states in the United States, a geospatial analysis can provide valuable insights. Geospatial tools like ArcGIS, combined with machine learning techniques, can be utilized to create an interactive map that visualizes and highlights the geographic distribution of counterspeech efforts (Jendrowski, 2019). This approach can enhance our understanding of the geographical dynamics of hate and the effectiveness of counterspeech strategies.

CHAPTER 3: METHODOLOGY

Two distinct methods were utilized in this study, tailored to address the research questions. Initially, this study employed computational methods, specifically supervised machine learning, to analyze the strategies in counterspeech tweets against anti-Asian hate during the COVID-19 pandemic. Subsequently, a quantitative content analysis was employed to identify the race and ethnicity of counterspeaker influencers.

Data Collection

For this study, the tweets associated with the hashtag #StopAsianHate during the COVID-19 pandemic were collected from January through May 2022. The selection of the #StopAsianHate hashtag for analysis is based on its prominence as one of the primary hashtags used to raise awareness about the urgent need to combat hate crimes against Asian Americans (Lyu et al., 2021). It is worth noting that a majority of the tweets utilizing this hashtag explicitly express support for solidarity movements aimed at Asian American communities (Lyu et al., 2021; Wongmith, 2022) to represent counterspeech. Moreover, the significance of the #StopAsianHate campaign extends beyond being just a hashtag; it has also spurred numerous offline protests and rallies while dominating the online space as a top trending hashtag during the COVID-19 pandemic (Fischer & Chen, 2021).

It is important to note that these data encompass the tweets posted during that specific time frame and do not represent the entirety of the pandemic. However, the selected timeframe of January to May 2022 for data collection in this study holds particular significance due to the emergence of the Twitter hashtag "#StopAsianHate" as a direct response to offensive terms such as "#ChineseVirus" and "#WuhanVirus." This context underscores the relevance and timeliness of examining the tweets associated with the #StopAsianHate hashtag during the specified data collection period. Moreover, 2022 has emerged as a critical period concerning hatred directed at Asian Americans. The Anti-Defamation League reports a significant increase in incidents of harassment against Asian Americans, rising from 21% in 2021 to 39% in 2022, aligning with the surge in offline anti-Asian hate crimes. This escalation in harassment follows a previous year marked by a notable surge in severe mistreatment towards Asian Americans, with a percentage increase from 11% in 2020 to 17% in 2021 (ADL, 2022).

Initially, the goal was to amass a collection of tweets associated with the hashtag #StopAsianHate, starting in late 2020 and continuing through 2022. Twitter's "Academic Research API" approach that was released with the "Twitter Academic Research API" would have been the best solution for processing such long-term archival data (Chen et al., 2021). With 'Academic Research Stream', or 'Academic Research Product Track', historic tweets could be accessible to academics upon request and with Twitter's permission, which unlocks the complete Twitter archive. Researchers get free access to the entirety of the Twitter archive, with a few restrictions, such as a limit of 10 million tweets a month (Ahmed, 2021). However, during the study, Twitter changed their data policy and imposed a whopping $42,000 price for the institutions to access historical tweets. The researcher's application for academic API was also not approved.

The dataset, comprising 106,388 tweets associated with the hashtag #StopAsianHate, was collected from January 2022 through May 2022 through the Sprinklr platform. Sprinklr is a social media analytics and data extraction tool that utilizes an API (application programming interface) to collect Twitter data. It offers various data collection parameters, including keywords, hashtags, user mentions, or specific Twitter accounts, which determine the scope of the data. By utilizing these parameters, Sprinklr interacts with the Twitter API, retrieving the relevant tweets and storing them in its own database for subsequent analysis and visualization. The tweets collected represent all tweets with the hashtag #StopAsianHate during the five-month time period.

Computational Method

This study employs computational method, specifically supervised machine learning, to analyze tweets that were posted in an effort to counter the anti-Asian hate during the COVID-19 pandemic. By utilizing machine learning algorithms that mimic human intelligence and learn from their environment, this approach proves highly effective in handling large datasets. The rapidly expanding realm of social media content provides a unique opportunity for conducting big data research (Garland et al., 2020; Kabir, 2022; Lyu et al., 2021; He et al., 2021; Mathew et al., 2019). In particular, the identification of online hate speech can now be automated without requiring input from victims or witnesses. Instead, machine learning algorithms can be deployed to detect instances of hate speech and counterspeech.

Machine learning is a unique subfield of artificial intelligence that focuses on the development of computational algorithms to simulate human intelligence by gathering information about their conditions. Several factors have led the researcher to conclude that machine learning is the best approach to analyzing the content of the tweets for this research. To start with, traditional manual content analysis is limited in its ability to analyze only a small portion of the extensive dataset. This approach often results in sampling errors, bias, and inconsistencies in coding. On the other hand, computational methods, especially those utilizing machine learning, offer the advantage of examining the entire dataset, free from coder bias, and reducing sampling errors. Second, textual data expressed in sentence forms can be classified by supervised machine learning thanks to its ability to learn annotating labels. More importantly, machine learning models can identify sarcasm (Pawar & Bhingarkar, 2020) and other emotional overtones in textual data (Kabir, 2022; Onan, 2022). Mathew et al. (2019) conducted a study that trained a machine learning model using over 9,000 hand-coded counterspeech tweets posted in response to hateful YouTube videos.

Hate speech online possesses distinct characteristics that make it an intriguing subject for examination using machine learning techniques. Researchers have been drawn to this area, seeking to automatically analyze people's opinions, sentiments, linguistic patterns, and network structures (Watanabe et al., 2018). Notably, studies have demonstrated that hate speech, offensive texts, and counterspeech can be detected by combining linguistic patterns with other variables (Bouazizi & Ohtsuki, 2016). Building upon this line of inquiry, Watanabe et al. (2018) discovered sentiment-based patterns in hate speech, including expressions of positive or negative emotions, as well as instances of hatefulness and offensiveness. For example, in the context of anti-Asian American hate speech, notable features encompass racism, sexism, prejudice against refugees, homophobia, and the use of "othering" language (Fortuna & Nunes, 2018).

Othering Language

Hate speech often employs the ideology of "othering," which involves comparing the differences between various groups through the lens of "Us against Them." It explains why "our" qualities are better than "their" ones, which are worthless and incompatible. This kind of attitude is shown by phrases like "send them home" (Fortuna & Nunes, 2018). As Asian Americans have been considered perpetual foreigners or forever foreigners, these features have been used to detect anti-Asian hate speech (Li & Nicholson, 2021).

Racial and Misogynistic Slurs

The use of sexist language in online conversations is commonplace on Twitter. The use of misogynistic rhetoric on Twitter has also been described in much research. Primary findings included a total of 100,000 occurrences of the term rape in UK-based Twitter accounts, of which around 12% looked to be threatening. Almost a third of the tweets on rape seemed to use the word lightly or metaphorically (Fortuna & Nunes, 2018). Aside from the rape threat hate speech, as the Hatebase database revealed, much misogynistic hate speech was sighted on social media, such as cunt (1225 sightings in 2022), cunter, bitch (over 4000 sightings in 2022), dyke (over 2,500 sightings in 2022), hoodrat (over 5,000 sightings in 2022), dagettes, etc. Hatebase is an international database of hate speech that is curated by its users and organized by location. The database of Hatebase is quite rich and has been cited multiple times in hate speech research (McIlroy-Young & Anderson, 2019). On the other hand, using a binary classifier for the labels "racist" and "nonracist," Kwok and Wang's study (2013) revealed that racist hate speeches contained painful historical allusions such as slavery and the presence of stereotypes or threats. Racist hate speech messages also included race-based slurs and stereotypical references such as "monkey" or "nigger", etc. (Fortuna & Nunes, 2018). Hatebase.com shared the most common ethnic slurs that have been posted on social media, such as mongoloid, rice-niggers, slanty eyes, dog eaters, gook-eyed, yellow invaders, etc. Among these, "whoriental" was the most common ethnic slur directed at Asian American women.

References to Stereotypes

Stereotypes are one of the most widely used characteristics of hate speech. Every form of preconceived stereotype has its own terminology, consisting of words, phrases, metaphors, and overall ideas. For instance, anti-Hispanic speech could refer to illegal immigration; anti-African-

American speech often refers to joblessness or growing up without both biological parents; and anti-Semitic rhetoric frequently refers to financial institutions, the media, and the Jewish community (Warner & Hirschberg, 2012). In addition to that, anti-Asian hate speeches contained stereotypes such as model minority (Kim et al., 2021), forever foreigner (Li & Nicholson, 2021), yellow peril (Kim et al., 2021), disease breeder (Wu & Nguyen, 2022), hypersexual (Wong & McCullough, 2021), etc.

Detecting Responses to Anti-Asian Hate Speech

Unlike hate speech, which can be detected using an established algorithm, responses to hate speeches can be best captured via Asian hate hashtags and counter hashtags because these hashtags have been observed flooding social media platforms since the beginning of COVID-19. The hate hashtags included #ChinaVirus, #ChineseVirus, Chinese virus, #ChineseBioterrorism, #FuckChina, #KungFlu, #MakeChinaPay, #wuhanflu, and #wuhanvirus (He et al., 2021; Nghiem & Morstatter, 2021). Consequently, as a form of resistance and response to the anti-Asian hate spreading on social media, users resorted to "counter speech" and "counter hashtags" such as #IAmNotAVirus, #WashTheHate, and #RacismIsAVirus, etc. The tweets in question are sometimes direct rebuttals to those who encourage hate or stand on their own. Besides, Asian Americans were found to employ nonassertive communication most often to react to COVID-19-related hate speech, followed by assertive and aggressive communication. Ethnic identification and previous discriminatory experience were connected with a nonassertive and less confrontational attitude. Men are usually more assertive (Jun et al., 2021).

Supervised Machine Learning

A supervised machine-learning algorithm was employed in this study. The term "supervised machine learning" refers to a method of teaching a computer model to automatically classify text by using a labeled dataset, where each document or text sample is assigned a predefined category or label by human annotators (Burscher et al., 2014). Considering the nature of the dataset in this study, which consists of social media text data, it is crucial to select an algorithm that is well-suited for such text-based analysis.

Random Forest Model

Random forest is a subset of the decision tree algorithm. Essentially, a decision tree is a systematic approach to classifying data by considering one feature, then another, until a final decision is reached. For example, to determine if an email is spam, we may first investigate if the email address appears strange and abrupt. If not, we can move on to the next criterion, such as the presence of suspicious web links. If there are any, we can classify the email as "spam," and if there are none, as "not spam." This decision-making process is best depicted using a graphical representation resembling a tree, which is how it got its name (Breiman, 2001). There are two beneficial aspects associated with the decision tree algorithm. To begin with, the explanation behind them is fairly simple. The model is comprehensible to even people who are not experts in machine learning. It makes this model particularly relevant since the potential readers of this research are from the media and communication fields and may not be familiar with machine learning techniques. This ensures accountability and transparency in social science research as well. Second, decision tree algorithms are applicable to almost all non-linear relationships (Van et al., 2022). However, this comes with a drawback. When a model is formulated as a set of yes-or-no questions, a lot of nuances are lost. Additionally, an incorrect choice made early in the tree (i.e., near its root node) cannot be undone later. Due to their rigidity, decision trees are vulnerable to overfitting, a situation in which the model fits the training data so well that it fails to adequately generalize to novel (test) data. Traditional decision trees are rarely employed in practical classification problems due to these limitations. Instead, an ensemble model, often known as random forests, is suggested by recent scholars (Onan et al., 2016). Several decision trees are estimated by drawing samples from the data, resulting in the "forest" metaphor. Then, the trees "vote" on which label to forecast to arrive at a final verdict, commonly known as a "majority vote." In addition to that, a decision tree is appropriate when there is not much data available (Van et al., 2022).

The random forest algorithm offers advantages for handling imbalanced data in machine learning classification. Imbalance occurs when one class has significantly fewer instances than the other, prompting various strategies for resolution, such as consensus cluster-based under-sampling and the balanced random forest model (Chen et al., 2004; Onan, 2019). This study focuses on analyzing unstructured and grammatically challenging Twitter data. Building on Kabir's (2022) findings, the Random Forest algorithm demonstrates superior performance when applied to such social media text data. This algorithm excels at processing high-dimensional, unstructured data like social media texts, which involve numerous features and intricate relationships. Consequently, we employed the random forest algorithm for computational analysis in this study. In the events where the training data revealed imbalance, the Balanced Random Forest (BRF) was employed to achieve superior performance.

Measurements

Computational Analysis

The following variable was coded using the computational technique.

Counterspeech Strategies. Initially, a team of two coders undertook the manual annotation of a total of 2000 tweets, aligning them with the established taxonomy of counterspeech strategies. This annotated dataset played a pivotal role in training a supervised machine learning algorithm known as 'random forest.' Subsequently, the algorithm was employed to classify the 'counterspeech strategies' as per the taxonomy outlined in the works of Benesch et al. (2016) and Cao et al. (2022). It is noteworthy that Benesch et al.'s (2016) taxonomy was also utilized in Mathew et al.'s (2019) research on counterspeech strategy detection which encompasses eight distinct categories, with seven of them being tested in this study. The first strategy, involving the presentation of facts to correct misstatements or misperceptions, was excluded from analysis since it pertains solely to replies or retweets, whereas our study focused exclusively on direct tweets. The second strategy, which revolves around acknowledging the enduring nature of anti-Asian hatred, recognizing its detrimental effects stemming from racial bias, resembles the strategy of warning of possible offline and online consequences of speech suggested by Benesch et al.'s (2016).

1) Presentation of facts to correct misstatements or misperceptions. Under this strategy, counter speakers will sometimes go to an extraordinary effort to convince outsiders that their knowledge or facts are incorrect, such as #COVID一19 is not the Flu and not SARS, pass it along https://t.co/0Gug6Typ3X . However, Benesch et al. (2016) argued that this approach is usually ineffective and these may arise as a reply to a misstatement or misinformation. Since this research is focused on counterspeech as direct tweets, but the correction of misstatement is mostly a response/reply to a tweet, this strategy was not examined in this study.

2) Pointing out hypocrisy or contradictions. For example, “Wait till the first white causality of Corona virus then they'll be crying its a bioweapon against white people despite the hundred thousands non white causalities. FACT”. After the manual labeling, very few instances were found under this category and they did not garner higher engagement metrics, and thus left

out of the machine learning analysis (Please see ‘the steps of machine learning analysis’ on page 46 for the labeling approach)

3) Warning of possible offline and online consequences of speech. Counterspeakers frequently employ the tactic of warning users of the potential repercussions of their nasty or harmful speech. Benesch et al., (2016), noted that warnings and threats were an effective strategy because they have been implemented in high-profile incidents, such as successful demands that individuals be dismissed from their employment for internet material. Example tweets under this category include “@RashidaTlaib The ban on travel was racist and the virus response is racist, the pattern here is that Ratshit is the racist clown. Everyone is hurting because of this Chinese virus response. It's because you ass clowns want the country on lock down to pander to these kinds of divisions.” The manually labeled dataset was used to train the machine learning algorithm for classifying the corpus. (Please see ‘the steps of machine learning analysis’ in page 46 for the labeling approach)

4) Affiliation. In certain instances, counterspeakers rely on a shared identity to assert that specific speech is undesirable for members of a particular group. Mathew et al. (2019) argued that people tend to evaluate in-group individuals as more trustworthy, honest, loyal, cooperative, and important to the group than outgroup members. “I would especially urge my fellow Jewish friends in the US to patronise their local Chinese restaurants/takeouts EXTRA atm. American Jews have a historic special bond with American Chinese food places. We need to help counteract this racist &amp; ignorant paranoia re: coronavirus.” The manually labeled dataset was used to train the machine learning algorithm for classifying the corpus. (Please see ‘the steps of machine learning analysis’ in page 46 for the labeling approach)

5) Denouncing speech as hateful or dangerous. Counterspeakers often identify speech and hashtags as hateful, or racist. In many cases, the content of hate speech is particularly denounced than the hate speaker. “@realDonaldTrump @WhiteHouse this is why it is NOT OK to call this a #ChineseVirus #RacismIsAVirus #racist #Covid_19”. The manually labeled dataset was used to train the machine learning algorithm for classifying the corpus. (Please see ‘the steps of machine learning analysis’ in page 46 for the labeling approach)

6) Use of visual media. Twitter allows users to include visuals (e.g., memes, graphics, photographs, animated gifs, and videos) in tweets, and counterspeakers frequently do so since images are more persuasive than words alone. For example, “Do your part to speak out against racist language and ideas that associate Asian people with COVID-19 stigma”. #WashTheHate #UtahWashTheHate https://t.co/Sp2DoWf1cG .

Sprinklr provided the information regarding the visual media use of each tweet which encompassed links, photos, videos, and GIFs. When searched the unique values using python programming language, a list was found where the type of visual media use is documented in the order it was posted on Twitter. For example, ‘PHOTO, PHOTO’ referred to the use of two photos, PHOTO, PHOTO, PHOTO referred to the use of three photos, and ‘VIDEO, PHOTO’ referred to the use of one video and one photo in the respective order, etc. (See appendix B). GIF stands for Graphics Interchange Format, which is an uploadable file format capable of displaying both still images and animated content. GIFs gained widespread popularity as a means of expressing reactions on social media, often without words. A number of tweets that used links were manually explored to see if any unique visual media can be found. The links were mostly for websites, and thus was coded in the category of no visual media. For the convenience of analysis, a variable called the 'use of visual media' was created in SPSS using the following code, link/no visual media=0, photo=1, video=2, GIF=3, multiple media use=4.

7) Use of humor. Humor is perceived as a linguistic or communicative performance, since humor appears to have a unique effect in counter-argument approach. It may alter the dynamics of communication, de-escalate tension, and attract far more attention to a message than it would ordinarily receive (Banesch et al., 2016). Example, “Can we not call it the Coronavirus or "Chinese Virus" and just call it the "Boomer Virus" and maybe people will take it seriously?” After the manual labeling, very few instances were found under this category and they did not garner higher engagement metrics, and thus left out of the machine learning analysis.

8) Emotional tone. Twitter counterspeech spans a wide range of tones and emotions, ranging from tweets that are just as offensive, angry, and cruel as the tweets to which they answer to tweets that are nice and respectful. However, Tweets can have a positive or negative impact depending on the tone used (Benesch et al., 2016). Thus, tone of counterspeech has been measured separately in this study. Studies in the past used both dictionary-based approach (tone dictionaries such as AFINN lexicon) and machine learning approach to determine positive, negative and neutral tone in the text. However, the efficiency of the dictionary-based approach is reduced as it fails to consider the surrounding context of the sentiment word (Nandwani & Verma, 2021). Whereas, the machine learning approach performs better once the algorithm is trained using manual labeling. Gamon’s (2004) study used a support vector machine to analyze 40,884 customer feedback responses that were gathered from surveys. The researchers tried several different combinations of features and were able to achieve an accuracy level of up to 85.47%. Ye et al. (2009) used multiple SVM, the N-gram model, and Naïve Bayes to analyze sentiment and reviews of seven popular destinations in Europe and the USA, which were

obtained from yahoo.com. The authors were able to achieve an accuracy of up to 87.17% using the n-gram model. In a recent study, Soumya and Pramod (2020) utilized machine learning techniques such as random forest and Naïve Bayes and found that the random forest approach, when used in conjunction with Unigram Sentiwordnet, yielded an accuracy of 95.6%.

For the detection of emotional tone in this study, Sprinklr's sentiment analysis tool was used which utilizes machine learning techniques to evaluate and classify sentiments expressed in various forms of text, such as blogs, reviews, social media, forums, news, and more, as either positive, negative, or neutral. Sprinklr claims to employ advanced deep learning methods to determine sentiment about 10 billion predictions per day with an accuracy level of over 80 percent (Sprinklr, 2023). Sprinklr's machine learning process begins with data collection, involving messages and unstructured data from over 25 social networking websites, 350 million web sources, internal data, surveys, and call transcripts. The next step is annotation, where a team of experts manually annotates and categorizes over a million messages spanning 20 industry verticals according to their associated sentiments. This comprehensive dataset serves as the foundation for training industry-specific models. During annotation, rigorous guidelines ensure consistency, and diverse examples are considered to encompass various word usages, including slang, jargon, and idioms. The model is then trained using this pre-labeled dataset and tested with new data to assess its performance and accuracy. Adjustments to parameters are made iteratively until the desired level of accuracy is achieved. Additionally, the platform allows users to provide feedback on message sentiment via the dashboard, enabling further model refinement. This feedback contributes to continuous enhancement of the sentiment analysis model.

Based on the sentiment analysis, the tonality of counterspeech was measured as positive tone (containing counterspeech tweets speech such as empathic, kind, polite, or civil), and hostile tone (containing abusive, hostile, or obscene language). Tweets with a neutral tone was coded as 0, positive tone was coded as 1, and a hostile tone was coded as -1.

9) Encouraging participation in counterhate. The last counterspeech strategy 'encouraging participation in counterhate' was found in Cao et al.'s (2022) study. Tweets under this category were observed promoting the recognition and appreciation of the Asian American and Pacific Islander community's culture, history, and contributions, and finally, awareness and visibility of the Asian American and Pacific Islander community. Examples include, "In May we celebrate #AAPIHeritageMonth. With anti-Asian hate and bigotry on the rise, we come together this month to uplift and celebrate AAPI voices while recommitting to #StopAsianHate", "#StopAsianHate like this man does", etc.

Manual Content Analysis

South Asian and East Asian Identity. The race and ethnicity section of this study was analyzed using the counterspeech tweets posted by influencers. "Influencers" are individuals or entities who have the ability to affect the opinions, behaviors, and decisions of their audience due to their expertise, authority, or popularity in a particular field or niche. Influencers are individuals who have built a substantial following on social media platforms and have gained the trust and attention of their audience. They use their platform to share content, recommendations, and opinions, which can influence their followers' attitudes and actions (Vodák et al., 2019). Individuals with 10,000 followers or more on social media platforms are considered "micro-influencers," which falls under the category of influencers overall (West, 2023). Influencers can be categorized into several types based on their followers count, such as micro-influencers:

10,000–100,000 followers, macro-influencers: 100,000–1 million followers; and mega- or celebrity influencers: 1 Million+ followers. Specifically, micro-influencers are known for their expertise in a specific field or their ability to create highly engaged audiences. In order to ascertain the South Asian and East Asian identities of the influencer counterspeaker, the racial or ethnic background of the tweeter was determined through an analysis of their Twitter profile. This involved manually exploring the profiles of each tweet to identify any explicit mention of racial or ethnic identity. Specifically, the focus of this study was influential and celebrity profiles, as they are more likely to provide information about their racial or ethnic background. During the process, the coders exercised caution and avoided making assumptions about the racial or ethnic identity of the tweet authors. This study adopted the threshold of more than 10,000 followers as an influencer (West, 2023). A total of 2514 tweets were found, and 500 of them were examined as part of this analysis. One of the coders who already participated in the manual labeling of the machine learning analysis was recruited for this coding procedure. The coder was trained to diligently collect ethnic information about each influencer, relying solely on publicly available data. The majority of influencer profiles belonged to celebrities, government officials, and politicians, with their ethnic information readily accessible through sources such as news magazines, Twitter bios, and official websites. The coder was trained to initiate research through tweet links in the dataset, leading to the examination of original tweets. Profile details, including usernames and bios, were carefully reviewed to extract any available ethnic information. In cases where such information was not present, further investigation took place on official websites and in news articles. After the coding was completed, a randomly selected 10 percent of the data was recoded by another coder to check for intercoder reliability. An online tool called 'Recal2' was used to calculate intercoder reliability. Recal2 is designed by Dr. Deen

Freelon, a professor of the Annenberg School for Communication at the University of Pennsylvania. The result showed that the intercoder agreement was 92.5%, Krippendorff's Alpha was .886, Cohen's Kappa was .885, and Scott's Pi was 0.885, which meets the general acceptance level of 0.7 (Burla et al., 2008). The South Asian influencers were coded as 0, and the East Asian influencers were coded as 1, White or European American influencers were coded as 3, Black or African American influencers were coded as 4, Hispanic Americans were coded as 5, Unsure race or ethnicity was coded as 6, and news or organization was coded as N/a or 7.

Effectiveness of Counterspeech Strategies. The effectiveness of the strategies was measured using Twitter audience engagement metrics, such as the number of favorites and retweets as was used in Bartlett and Krasodomski-Jones's (2015) study. Each of these engagement metrics has its meaning and can provide insights into the level of interest or engagement with the content of the tweet. For example, "favorites" is an indication that they appreciate or agree with the content of the tweet. On the other hand, when a user comments or retweets a tweet, they are providing their thoughts or feedback on the content of the tweet. It is important to note that certain users, particularly influencers, may have more followers than other users, resulting in a higher count of favorites and retweets. Therefore, to control for the effect of follower size, this study employed a weighted measures for favorites and retweets, which divided the number of favorites and retweets by the number of followers of the user. However, in order to check the variability in results, both weighted and unweighted engagement metrics were compared.

Racial Diversity. The diversity index obtained from the US census directory (census.gov) served as a metric to gauge the diversity of the location where a tweet was posted (McPhillips, 2020). State-level racial diversity was used as the location of tweets is easier to identify on a state-level to include both small towns and big cities. As outlined by the United States Census Bureau (2021), the diversity index ranges from 0 to 1. A value of 0 signifies a population where everyone shares the same racial characteristics, while a value approaching 1 indicates a highly diverse population with various racial and ethnic backgrounds.

The Steps of ML Computational Analysis

Manual Labeling

Machine learning analysis begins with annotating or labeling the tweets. The linguistic nuance present in such a text set necessitates the use of manual labeling as the procedure in this study (Cunha Lassance et al. 2019). When it comes to social science, manual labeling can be especially useful for validation; this helps to ensure that the analysis is founded on qualitative reasoning before moving on to the training of the algorithm and subsequent machine learning procedure (Chen et al. 2018). Estimating the precise number of labels needed is a complex task. Nonetheless, it can be reasonably asserted that for categorizing longer texts, typical sample sizes in social science research often fall within the range of 1,000 to 10,000 (Burscher et al., 2014; Van et al., 2022). Therefore, this study performed the manual labeling of 2,000 tweets. All nine variables that took part in the machine learning analysis were labeled. Two coders were recruited to participate in the manual labeling. The first coder was a researcher at a renowned non-government organization, and the second coder was an undergraduate-level student majoring in a humanities subject. The coders were trained following the procedure used in Vermeer's (2018) supervised machine learning study and suggested in Van et al.'s (2022) book. That is, the coders were trained using the definition, scope, measurement, and example tweets associated with each variable. A coding scheme was provided to the coders that included all this information. Then, a random sample of 150 from the data was assigned to the coders for manual labeling. Once the coders completed the 150 tweets for nine variables, the intercoder reliability was calculated using an online tool called ‘Recal2’. Recal2 is software to calculate the reliability of two coders designed by Dr. Deen Freelon, a professor of the Annenberg School for Communication at the University of Pennsylvania. The result showed the following scores for nine variables (see Appendix A),

Counterspeech: High agreement (96.7%) between coders. Moderate values for Scott's Pi (0.91), Cohen's Kappa (0.91), and Krippendorff's Alpha (0.91).

Against hate speech: Moderate agreement (82.7%) between coders. Values for Scott's Pi (0.56), Cohen's Kappa (0.57), and Krippendorff's Alpha (0.57) also moderate.

Against hate crime: High agreement (92%) between coders. Values for Scott's Pi (0.84), Cohen's Kappa (0.84), and Krippendorff's Alpha (0.84) are high as well.

Presentation of facts: High agreement (92.7%) between coders. Values for Scott's Pi (0.32), Cohen's Kappa (0.32), and Krippendorff's Alpha (0.32) are also low.

Pointing out hypocrisy: High agreement (94.7%) between coders. Values for Scott's Pi (0.40), Cohen's Kappa (0.40), and Krippendorff's Alpha (0.40) are moderate.

Consequences: Moderate agreement (85.4%) between coders. Values for Scott's Pi (0.69), Cohen's Kappa (0.69), and Krippendorff's Alpha (0.69) are moderate as well.

Affiliation: High agreement (92%) between coders. Values for Scott's Pi (0.69), Cohen's Kappa (0.64), and Krippendorff's Alpha (0.64) are moderate.

Denouncing hate: Moderate agreement (80%) between coders. Values for Scott's Pi (0.56), Cohen's Kappa (0.56), and Krippendorff's Alpha (0.56) are also moderate.

Encouraging participation: High agreement (94%) between coders. Values for Scott's Pi (0.86), Cohen's Kappa (0.86), and Krippendorff's Alpha (0.86) are high as well (See appendix A). The reliability agreement meets the general acceptance level of 0.7 (Burla et al., 2008).

Observing the considerable agreement, each coder was assigned 1000 tweets to label. In total, 2000 tweets were manually labeled for the machine learning procedure.

Pre-processing the Data

The collected text data underwent a series of preprocessing steps to ensure its suitability for analysis. This included removing punctuation, converting text to lowercase, eliminating stop words, and applying stemming or lemmatization techniques.

Word Embedding

Text data is unstructured, making it difficult for computers to directly process and learn from it. Word embedding method is used to convert the texts, its occurrences, and its values in the form that can be easily understood and processed by machine learning algorithms. Methods like Bag-of-Words (BOW), TF-IDF, Word2Vec, and GloVe are among the popular techniques utilized for text representation and feature extraction in the field of text analysis. These approaches enable the conversion of text into numerical representations, facilitating subsequent analysis and modeling tasks. Following is an overview of these techniques.

Bag-of-Words (BOW) is a simple and widely used method that represents text documents as a collection or "bag" of individual words, disregarding their order and structure. It creates a dictionary of all the unique words and a matrix where each row represents a document, and each column represents a unique word in the entire corpus. The matrix cells contain the frequency or presence of each word in each document. However, BOW does not capture the semantic meaning or relationships between words; it only focuses on their occurrences.

TF-IDF is a word-to-numeric conversion technique that takes into account both the term frequency (TF) and the inverse document frequency (IDF) of words. TF measures how frequently a word appears in a specific document, while IDF measures the rarity or importance of a word across the entire corpus. TF-IDF assigns higher weights to words that are more specific to individual documents and have lower occurrence across the corpus. TF-IDF captures the relevance of words within documents but still lacks the ability to capture the full semantic meaning and relationships between words (Ahuja et al. 2019).

Word embeddings, on the other hand, are powerful tools that help us capture the semantic relationships between words based on their contextual usage. Techniques like Word2Vec and GloVe create word embeddings by considering the surrounding words and their co-occurrence patterns in a large text corpus. These embeddings represent words as vectors in a high-dimensional space, where the distances and directions between vectors reflect the semantic relationships between words. It can be imagined as a map where the distances and directions between the words show us their connections in meaning. Word embeddings allow for more nuanced analysis, including similarity comparison, analogical reasoning, and capturing contextual meaning (Rodriguez & Spirling, 2022).

Word2Vec works by training a computer program to predict the words that are usually found near a target word. By doing this, it learns to assign each word a special code called a "vector" that represents its meaning. Precisely, it predicts the surrounding words given a target word, or predicting a target word based on its context (Karani, 2018). GloVe, on the other hand, focuses on how often words appear together in the text. It looks at the words that usually hang out together and builds a special matrix that keeps track of these relationships. Then it reduces the size of this matrix to create the word representations. This way, we can see how words are connected to each other in terms of meaning.

Comparing these methods, TF-IDF was found achieving the best performance and was used to extract features. After establishing a corpus containing all unique words within the document and assigning values to each word for calculation, the TF-IDF process generated a total of 4,149 features. Subsequently, these extracted features served as the input variables for training the machine learning algorithm. Following section illustrates the mechanism of feature extraction.

Feature Extraction

Feature extraction is a fundamental process in machine learning and data analysis, involving the identification and extraction of meaningful attributes from raw data (Ahuja et al., 2019). It plays a crucial role in transforming the data into a format that is suitable for modeling and analysis. Features represent specific characteristics or attributes of the data that carry relevant information and contribute to the performance of machine learning models (Zheng & Casari, 2018). The process of feature extraction can encompass several steps, such as feature selection and feature construction, etc. Let's understand each of these steps using examples from the context of social scientific research.

In feature selection, the most relevant variables or factors are chosen from a larger set of data that are likely to have a significant impact on the outcome variable being studied (Ahuja et al., 2019). For example, if we are investigating the factors influencing academic performance, we may select variables such as socioeconomic status, parental education, and study habits as important features to consider, while disregarding less relevant variables like favorite color or music preference.

Finally, feature construction involves creating new variables or features based on existing ones, which can capture more nuanced information and enhance our understanding of the phenomena under investigation. For instance, if we are studying social media usage and its impact on well-being, we may create a new feature representing the average daily screen time by combining variables related to specific social media platforms and usage patterns. These processes enable a machine learning model to identify important variables, transform data to address statistical assumptions, convert qualitative variables into numerical form, and generate new variables that provide a richer understanding of the phenomena we are studying. This study used TF-IDF method to convert the texts into numeric values and generated 4149 features through which the machine learning model was trained.

Training and Testing ML Model

Once the dataset labeling process was completed, the dataset was partitioned into two sets: a training dataset and a test dataset. The training dataset was utilized for the machine learning model to train and adjust its parameters, while the test dataset served as a means to evaluate the model's performance. Essentially, the test dataset allows us to gauge how effectively the model fulfills its intended task. Commonly, the training and test datasets are divided in ratios ranging from 50:50 to 80:20. Given the large size of our prediction dataset which is 106k, it was necessary that the machine learning model has maximum examples to learn from, which can help it generalize better. Therefore, in train-test data (manually labeled dataset), a larger portion should be allocated to the training set to ensure the model has a comprehensive learning experience. In this study, an 80:20 ratio of train-test split was chosen, allocating 80% of the data for training and 20% for testing. In past studies, researchers advised to increase the size of the training dataset slightly at the expense of reducing the size of the test dataset, if the labeled dataset is relatively smaller than the predicting data (Van et al., 2022).

Model Evaluation

In building machine learning model, the validation or evaluation process assumes a pivotal role. The widely adopted validation techniques, including the accuracy score, precision, recall, confusion matrix, and cross-validation through grid-search was employed to evaluate the model (Kabir, 2022; Van et al., 2022). Recall serves as a metric to assess the proportion of relevant instances that have been correctly retrieved from the dataset, relative to the total number of relevant instances. It offers insight into the algorithm's capability to identify all relevant items within the dataset. Conversely, precision quantifies the proportion of relevant instances among those retrieved by the algorithm, providing an estimation of the algorithm's accuracy in excluding irrelevant items (Van et al., 2022). Below is a brief description of the metrics with example. Precision measures the accuracy of the positive predictions made by the model. For example, in ‘Counterspeech or not?’ in the context of class 0 (denoted as "0"), the precision is 0.86. This means that when the model predicts class 0, it is correct 86% of the time. In the context of class 1 (denoted as "1"), the precision is 0.94. This means that when the model predicts class 1, it is correct 94% of the time. Recall (True Positive Rate) measures the ability of the model to identify all relevant instances in the dataset. For example, in ‘Counterspeech or not?’ in the context of class 0, the recall is 0.90. This means that the model correctly identifies 90% of the actual class 0 instances. In the context of class 1, the recall is 0.92. This means that the model correctly identifies 92% of the actual class 1 instances. F1-score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. For example, in ‘Counterspeech or not?’ In the context of class 0, the F1-score is

0.88. This indicates a balanced trade-off between precision and recall for class 0. In the context of class 1, the F1-score is 0.93. This indicates a balanced trade-off between precision and recall for class 1. Support is the number of actual occurrences of each class in the test dataset. For example, in ‘Counterspeech or not?’ For class 0, the support is 134, meaning there are 134 instances of class 0 in the test dataset. For class 1, the support is 237, indicating 237 instances of class 1 in the test dataset. Accuracy is the ratio of correctly predicted instances to the total instances in the dataset. For example, in ‘Counterspeech or not?’, the overall accuracy of the model is 0.91, meaning it correctly predicts the class labels for 91% of the instances in the test dataset.

Finally, confusion matrix is instrumental in analyzing the performance of a binary classifier. Its columns typically depict the number of instances predicted for a given class, while its rows indicate the number of instances belonging to the actual class. Through this matrix, we can ascertain the number of correct predictions, incorrect predictions, true positives, and false negatives (Van et al., 2022).

Counterspeech or not? (Counterspeech=1, Not counterspeech=0) Four algorithms were employed to train: Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and neural network (NN). RF performed the best (89%) (See table 1).

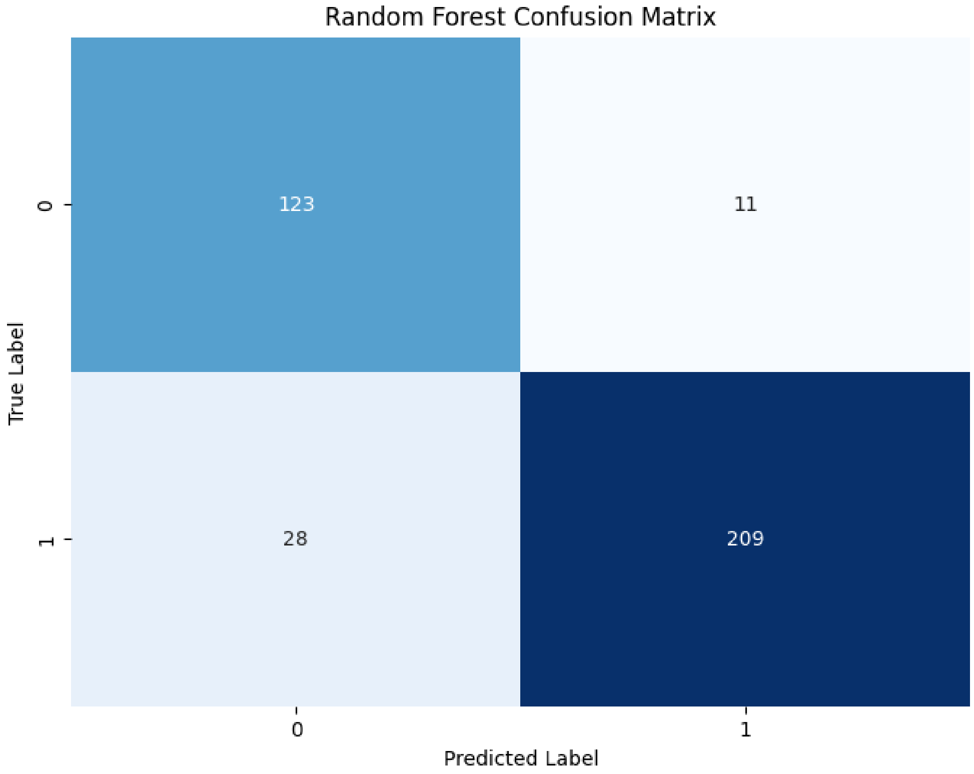
Table 1

Counterspeech Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.81	0.92	0.86	134	
1	0.95	0.88	0.91	237	0.89
			Total	371	

Figure 1

Confusion Matrix (Counterspeech)



True Negative (TN): The model correctly predicted the negative class for 121 instances. False Positive (FP): The model incorrectly predicted the positive class for 13 instances that actually belong to the negative class. False Negative (FN): The model incorrectly predicted the negative class for 19 instances that actually belong to the positive class. True Positive (TP): The model correctly predicted the positive class for 218 instances (see Figure 1).

Countering hate speech (present=1, absent=0). Four algorithms were employed to train: Random Forest (RF), Balanced Random Forest (BRF), Support Vector Machine (SVM), Naïve Bayes (NB), and neural network (NN). BRF performed the best (84%) (See Table 2).

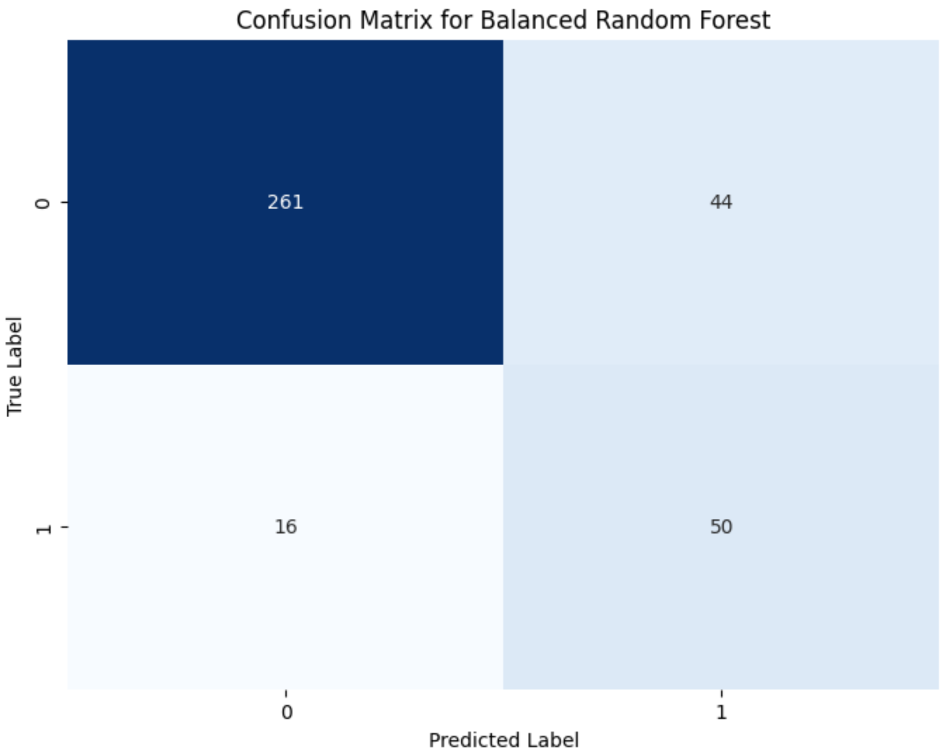
Table 2

Countering Hate Speech Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.94	0.86	0.90	305	
1	0.53	0.76	0.62	66	0.89
			Total	371	

Figure 2

Confusion Matrix (Countering Hate Speech)



Countering hate crime (present=1, absent=0). Four algorithms were employed to train: Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and neural network (NN). RF performed the best (93%) (See table 3).

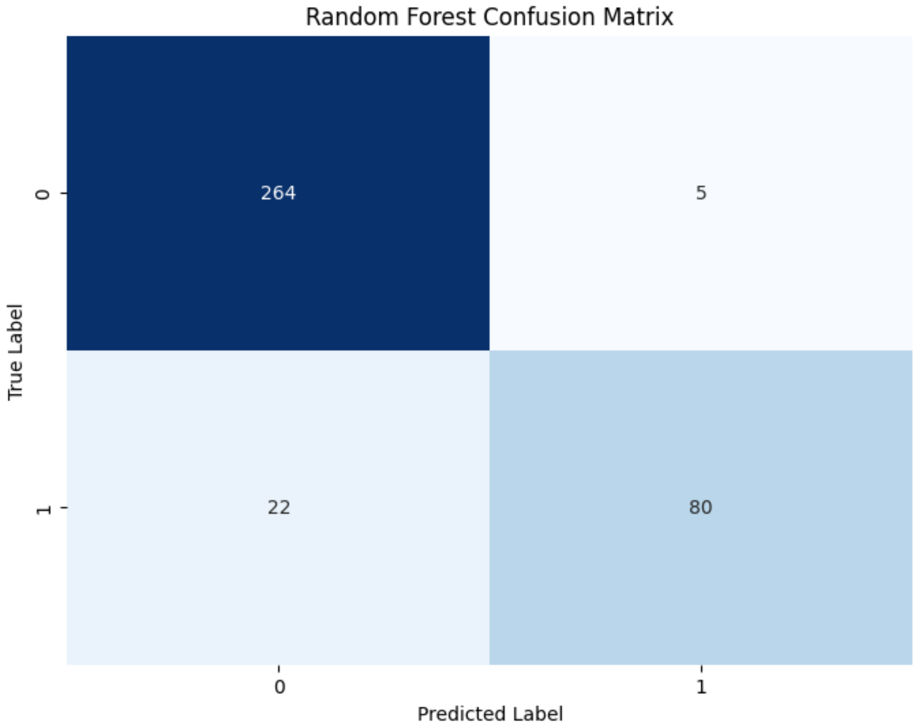
Table 3

Countering Hate Crime Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.92	0.98	0.95	269	
1	0.94	0.78	0.86	102	0.93
			Total	371	

Figure 3

Confusion Matrix (Countering Hate Crime)



Warning of possible offline and online consequences of speech. (present=1, absent=0). Four algorithms were employed to train: Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and neural network (NN). In addition, hyperparameter tuning was applied to improve the performance. RF performed the best (82%) (See Table 4).

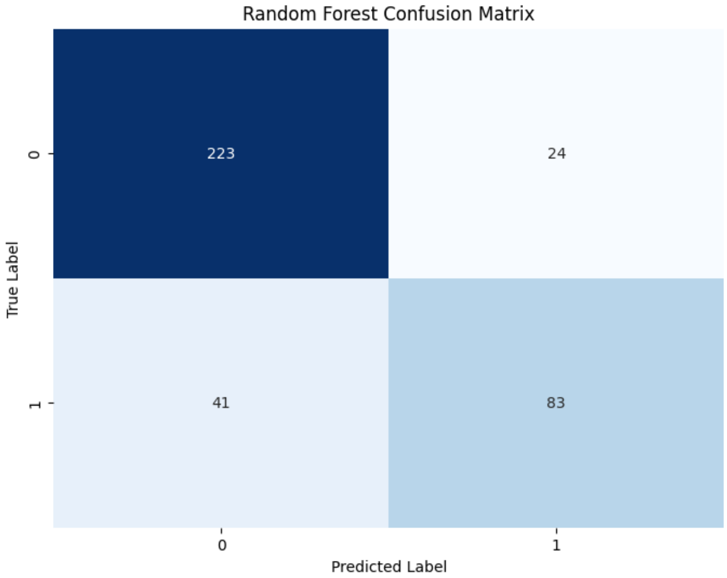
Table 4

Warning Consequences Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.84	0.90	0.87	247	
1	0.78	0.67	0.72	124	0.82
			Total	371	

Figure 4

Confusion Matrix (Warning Consequences)



Affiliation (Rejected) (present=1, absent=0). Four algorithms were employed to train. Random Forest (RF), BRF, Support Vector Machine (SVM), Naïve Bayes (NB), and neural network (NN). In addition, hyperparameter tuning, gradient boosting was applied to improve the performance. BRF performed the best (90%). But the confusion matrix revealed that only 9 instances were predicted correctly out of 21 for class 1. The class 1 was too small for the model to learn well. Therefore, this strategy wasn't analyzed in the study (See Table 5).

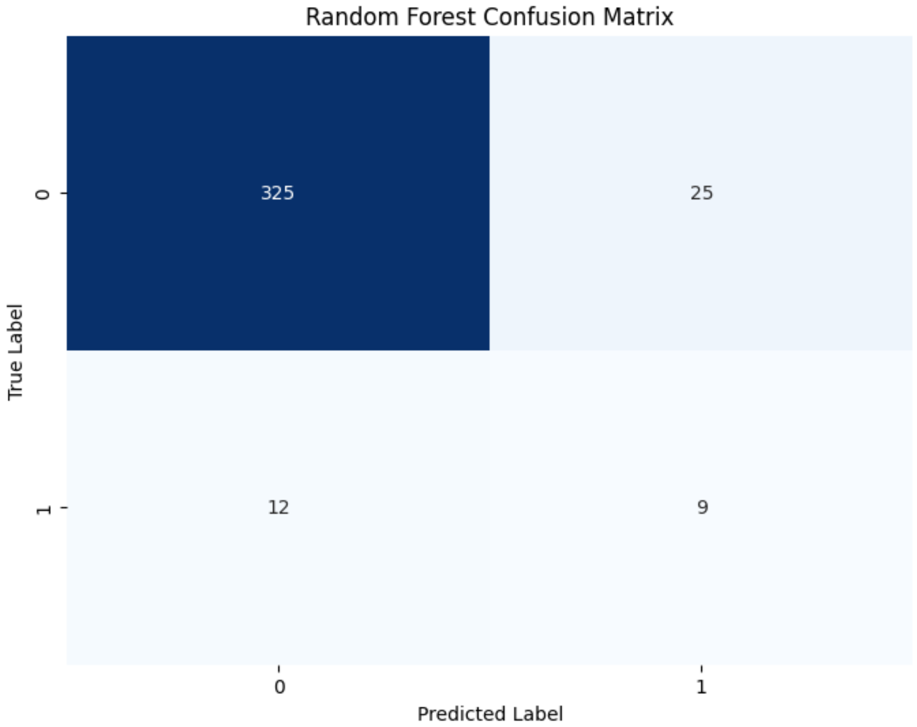
Table 5

Affiliation Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.96	0.93	0.95	350	
1	0.26	0.43	0.33	21	0.90
			Total	371	

Figure 5

Confusion Matrix (Affiliation)



Denouncing speech as hateful or dangerous. (present=1, absent=0). Five algorithms were employed to train. Random Forest (RF), BRF, Support Vector Machine (SVM), Naïve Bayes (NB), and neural network (NN). In addition, hyperparameter tuning, gradient boosting was applied to improve the performance. BRF performed the best (81%) (See table 6).

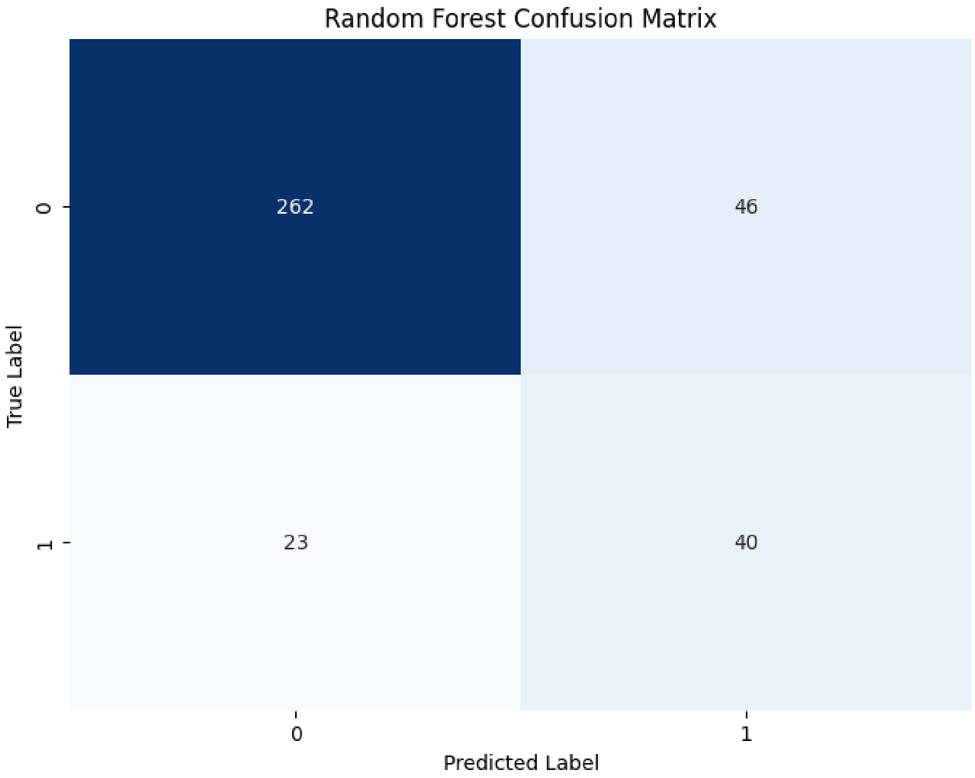
Table 6

Denouncement Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.92	0.85	0.88	308	
1	0.47	0.63	0.54	63	0.81
			Total	371	

Figure 6

Confusion Matrix (Denouncement)



Encouraging participation in counter hate (present=1, absent=0). Five algorithms were employed to train: Random Forest (RF), BRF, Support Vector Machine (SVM), Naïve Bayes

(NB), and neural network (NN). In addition, hyperparameter tuning, and gradient boosting was applied to improve the performance. BRF performed the best (84%) (See table 7).

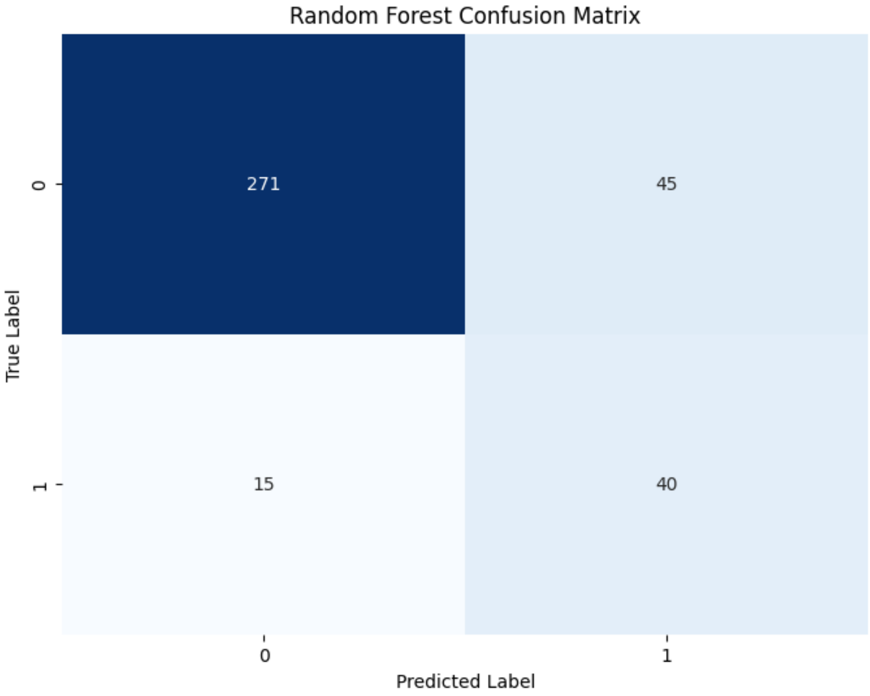
Table 7

Encouraging Participation Model Evaluation

	precision	recall	f1-score	support	Accuracy
0	0.95	0.86	0.90	316	
1	0.47	0.73	0.57	55	0.84
			Total	371	

Figure 7

Confusion Matrix (Encouraging Participation)



CHAPTER 4: RESULT

Profile of Anti-Asian Hate Tweets

The data was filtered using the hashtag #StopAsianHate, which may lead to an initial assumption that every tweet would pertain to counterspeech. However, such an assumption is problematic in light of the prevalent occurrence of 'hashtag hijacking' in contemporary social media (Mousavi & Ouyang, 2021). Hashtag hijacking is a social media marketing strategy employed by individuals or companies to increase the visibility of their own content by capitalizing on popular or trending hashtags. To mitigate the impact of such hijacked hashtags, the machine learning models built in this study were employed to ascertain whether a tweet constituted counterspeech or not. Among the total of 106,390 tweets collected, only those originating from the United States were analyzed, resulting in a dataset comprising 18,933 tweets. Next, the random forest-based model was employed in the dataset to classify the tweets based on the counterspeech strategies and counterspeech type. Within the subset, a total of 12,881 tweets were identified as instances of counterspeech, of which 2,170 instances were classified as counterspeech countering hate speech and 2,020 instances of countering hate crime, and the rest are neither countering hate speech nor hate crime (Figure 8 and Table 8). ‘Specified’ category is the percentage of counterspeech with either against hate crime or hate speech, whereas ‘not specified’ tweets are also counterspeech but don’t address hate speech or hate crime specifically. Additionally, for counterspeech strategies, there are 2,375 instances associated with consequences, 926 instances of denouncing, and 1,505 instances of encouraging behavior (Figure 9). These strategies were not as commonly used as expected.

Figure 8

Counterspeech Type

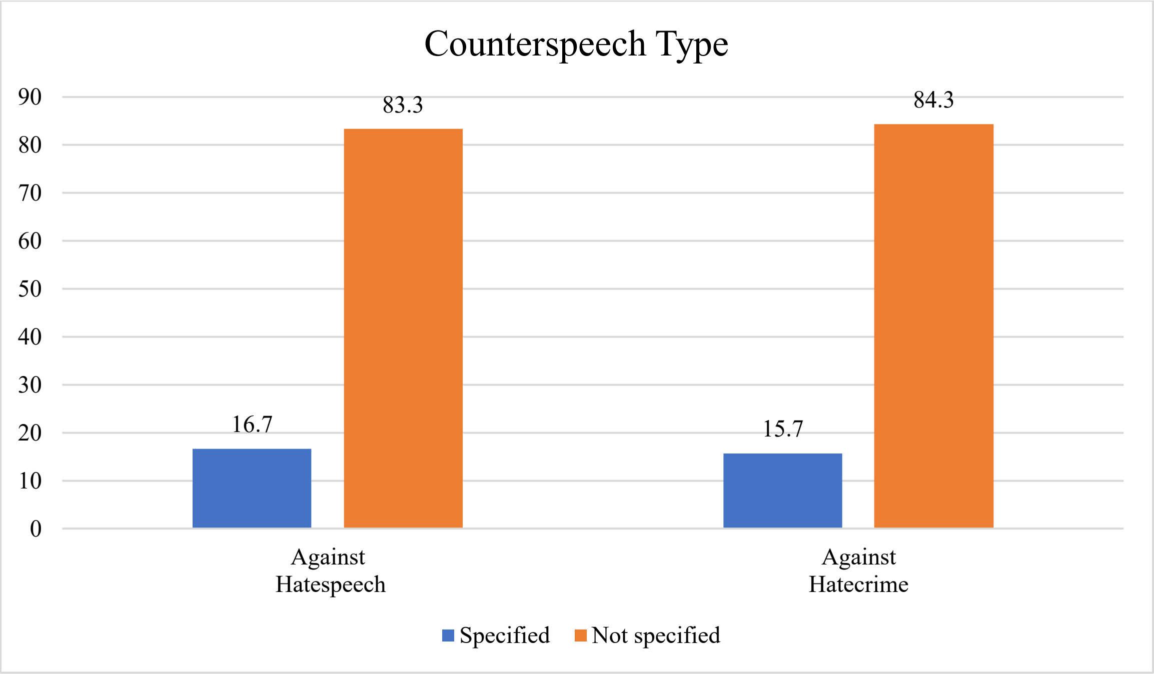


Table 8

Counterspeech Types and Strategies

	Counterspeech type	Counterspeech type	Counterspeech Strategies	Counterspeech Strategies	Counterspeech Strategies
	Against hate speech	Against hate crime	Warning Consequences	Denouncement	Encouraging participation
Total	2170	2020	2375	926	1505

Figure 9

Counterspeech Strategies Use by Percentages

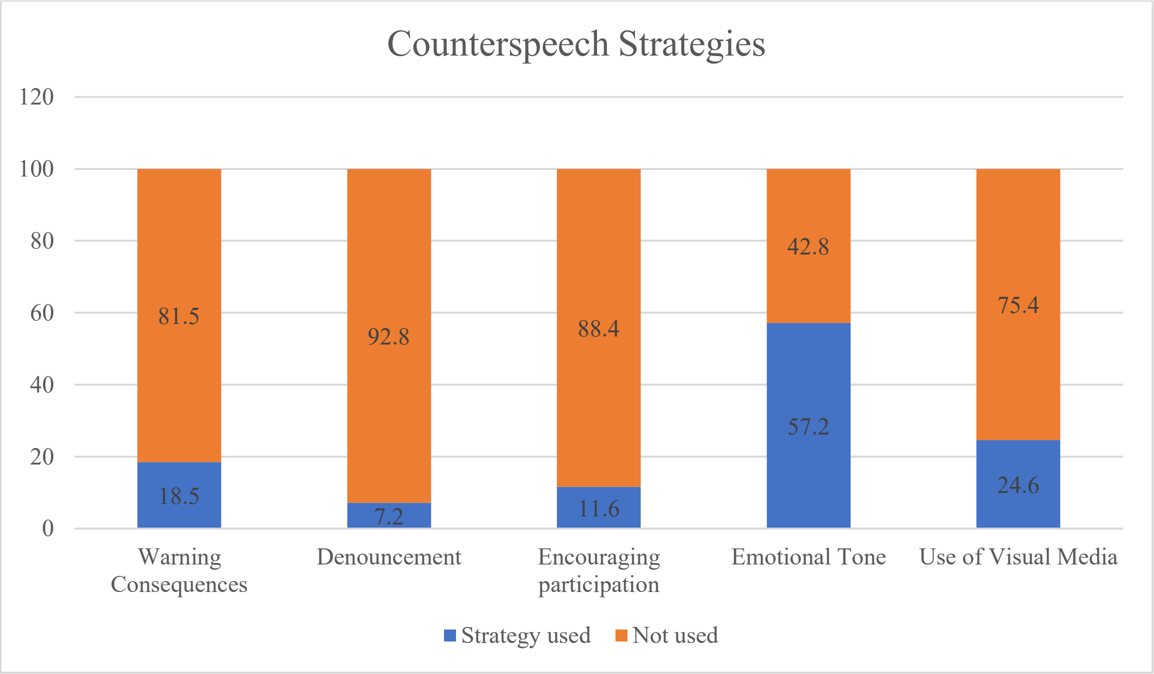


Table 9

Example Tweets for Counterspeech Type and Strategies

Classficications	Category	Example tweet
Counterspeech type	Addressing hate crime	'The Boys' and "Suicide Squad" actor Karen Fukuhara reveals she was assaulted in a hate crime attack outside of a cafe: â€œI was struck in the head by a man. This sh*t needs to stop. Us women, Asians and the elderly need your help #StopAsianH"te
	Addressing hate speech	We're being targeted because of the stereotypes about us, that we won't fight back, that we're submissive." This is disturbing and unacceptable. We all have a part in dismantling racism and discrimination. #StopAsianHate

Classficications	Category	Example tweet
	Counterspeech not addressing hate speech or hate crime	Everybody has an #AAPI-owned restaurant in their community or hopefully close enough. Most are run by families who really need our support now more than ever, so make your Sunday #VeryAsian & order some dumplings for dinner! #DumplingSunday #StopAsianHate #StopAAPIHate
Counterspeech Strategies	Consequences	The future of Asians in Virginia if Virginia Democrats don't abandon their racist Byrd Machine 2.0 "Massive Resistance" to equality under the law for Asian kids. #StopAsianHate in #unfairFax
	Denouncing	2/2 crimes will not be tolerated and we must to denounce this bigotry. Enough is enough! #StopAsianHate
	Encouraging	Happy Asian Pacific American Heritage Month! This month, we celebrate the rich history & culture of the AAPI communities and acknowledge the heightened discrimination they have faced since the beginning of the pandemic. Today & every day, let us all recommit to #StopAsianHate.
	Positive tone (Emotional tone)	The only good thing about that terrible call is it made me crave some proper dumplings & want to support my local #AAPI-owned restaurants even more than I already do (Jews love the Chinese food, but Korean is my fave & I could eat bibimbap every day) #StopAsianHate #StopAAPIHate

Classficications	Category	Example tweet
	Negative tone (Emotional tone)	Asian Americans face serious discrimination and increasing slurs especially related to the pandemic. We remember the tragedy in Georgia against the AAPI community, the lives lost, and reject these horrific actions and hate speech. #StopAsianHate
	Neutral tone (Non-emotional tone)	I hope #StopAsianHate is trending to raise awareness about the serious discrimination facing the Asian community and NOT because BTS didn't win a Grammy...

Data Analysis

Most Effective Counterspeech Strategy on Twitter

To answer the research question RQ1: Which counterspeech strategies are more effective on Twitter? An analysis of variance (ANOVA) was performed to examine whether the consequences, denouncing and encouraging had any effect on weighted favorites and weighted retweets. The results showed the strategy "Consequence" had a higher mean for both weighted retweets and weighted favorites compared to "Denounce" and "Encourage." Specifically, the mean for weighted retweets was 0.0024, and the mean for weighted favorites was 0.0064 in the "Consequence" strategy (Table 10).

Table 10

Mean Difference of Counterspeech Strategies and Engagement Metrics

		N	Mean	Std. Deviation	Std. Error
Weighted Retweets	Consequence	2371	.00244	.021298	.00044
	Denounce	319	.0023	.01405	.00079
	Encourage	473	.00258	.02645	.00121
	Total	3163	.00245	.02155	.00038
Weighted Favorites	Consequence	2371	.00644	.04968	.00102
	Denounce	319	.00359	.02037	.00114
	Encourage	473	.00533	.05896	.00271
	Total	3163	.00599	.04911	.0009

In the regression analysis predicting weighted favorites, the model's overall fit was statistically significant (F (5, 12843) = 6.644, p <.001), but it only explains a small proportion of the variance in weighted favorites (R2 =.003, adjusted R2 =.002) (see Table 11). Among the predictors, use of visual media exhibited a significant positive relationship with weighted favorites (β = 0.050, p <.001), indicating that an increase in the use of visual media was associated with higher weighted favorites. The other predictors, including emotional tone, consequences, denouncing, and encouraging, did not significantly contribute to the model's predictive power.

Table 11

Weighted Favorites and Counterspeech Strategies

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.005	.001		6.875	.000
emotional tone	.000	.001	.006	.615	.539
use of visual media	.005	.001	.050	5.530	<.0001
Consequences	-8.026E-5	.002	-.001	-.053	.957
Denouncing	-.001	.002	-.005	-.485	.627
Encouraging	.000	.002	-.001	-.145	.885

In the regression analysis, several predictors were examined concerning their influence on the dependent variable, Weighted Retweets. Among the predictor variables, emotional tone demonstrated a significant positive association with Weighted Retweets (B = 0.001, SE = 0.000, β = 0.022, t = 2.443, p = .015), while use of visual media also exhibited a significant positive relationship (B = 0.001, SE = 0.000, β = 0.025, t = 2.785, p = .005) (Table 12). Conversely, consequences showed a significant negative association with Weighted Retweets (B = -0.001, SE = 0.001, β = -0.019, t = -1.856, p = .063), although it approached significance. Denouncing and Encouraging did not display significant relationships with Weighted Retweets (B = -0.001, SE = 0.001, β = -0.005, t = -0.502, p = .615; B = 0.000, SE = 0.001, β = 0.002, t = 0.187, p = .851, respectively)

Table 12

Regression on Weighted Retweets and Counterspeech Strategies

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
	.039[a]	.002	.001	.026	
Coefficients[a]					
	B	Std. Error	Beta		
(Constant)	.004	.000		11.844	.000
emotional tone	.001	.000	.022	2.443	.015
use of visual media	.001	.000	.025	2.785	.005
consequences	-.001	.001	-.019	-1.856	.063
Denouncing	-.001	.001	-.005	-.502	.615
Encouraging	.000	.001	.002	.187	.851

Adjusted R^2=0.001

Counterspeech Strategy Most Associated with Hate Speech

To answer the research question, RQ2a: Which counterspeech strategies are more associated with hate speech? A chi-square test was performed (Table 13). The result showed that hate speech was significantly related to consequences ($\chi^2 = 540.447$, $p < .001$), with a Pearson's R value of .205, indicating a positive relationship. A significant association with hate speech was also found with denouncing ($\chi^2 = 1855.303$, $p < .001$), showing a Pearson's R value of .380, signifying a moderately positive association. Encouraging was also significantly related to hate speech ($\chi^2 = 1188.297$, $p < .001$), with a Pearson's R value of .304, indicating a positive relationship. Furthermore, hate speech was significantly related to emotional tone ($\chi^2 = 943.330$, $p < .001$), with a Pearson's R value of -.259, indicating a negative relationship, or a higher likelihood of negative tone. However, there was no significant association between hate speech and use of visual media ($\chi^2 = 6.858$, $p =.144$).

Table 13

Counterspeech Strategies Against Hate Speech

Counterspeech Strategies	χ^2 Value	df	p-Value	R
Consequences	540.447	1	< .001	.205
Denouncing	1855.303	1	< .001	.380
Encouraging	1188.297	1	< .001	.304
Use of Visual Media	6.858	4	.144	
Emotional Tone	943.330	2	< .001	-.259

Table 14

Counterspeech Against Hate Speech and Hate Crime Occurrence

Counterspeech Strategies	Total	Countering hate speech	Countering hate crime
Consequences	2371	776 (36.2%)	1225 (60.7%)
Denounce	924	624 (29.1%)	629 (31.2%)
Encouraging	1495	716 (33.4%)	854 (42.3%)
Use of Visual Media			
No visual media	9692	1612 (75.3%)	1304 (64.7%)
Photo	2636	445 (20.8%)	633 (31.4%)
Video	477	83 (3.9%)	78 (3.9%)
GIF	3	0 (0.0%)	0 (0.0%)
Multiple media	41	1 (0.0%)	2 (0.1%)
Emotional Tone			
Negative	5674	1580 (73.8%)	1199 (59.4%)
Neutral	5502	501 (23.4%)	720 (35.7%)
Positive	1673	60 (2.8%)	98 (4.9%)
Total	12849	2141	2017

Counterspeech Strategy Most Associated with Hate Crime

To answer RQ2b: Which counterspeech strategies are more associated with hate crimes, a chi-square was performed. The result revealed that hate crime was significantly related to consequences (χ^2 = 2842.399, p < .001), with a Pearson's R value of .470, indicating a strong positive relationship. A significant association was also found with denouncing (χ^2 = 2063.809, p < .001), showing a Pearson's R value of .401, signifying a moderately positive association. Encouraging was significantly related to hate crime (χ^2 = 2193.978, p < .001), with a Pearson's R value of .413, indicating a moderately positive relationship. Additionally, hate crime showed a significant association with Use of visual media (χ^2 = 178.947, p < .001). Furthermore, hate crime was significantly related to emotional tone (χ^2 = 277.358, p < .001), with a Pearson's R value of -.147, indicating a negative relationship (Table 15).

Table 15

Counterspeech Strategies Against Hate Crime

Strategies	χ^2	df	p-Value	R
Consequences	2842.399	1	< .001	.470
Denouncing	2063.809	1	< .001	.401
Encouraging	2193.978	1	< .001	.413
Use of Visual Media	178.947	4	< .001	
Emotional Tone	277.358	2	< .001	-0.147

Positive Tone More Effective Than Negative Tone

To examine hypothesis H1: Counterspeech with a positive tone is more effective than a negative tone, a one way ANOVA mean comparison was conducted to investigate the effects of emotional tone on weighted retweets and weighted favorites to see if there are differences after the follower size of the tweeter was controlled. For weighted retweets, the results revealed statistically significant differences among the groups (F(2, 12,846) = 10.062, p < 0.001). Concerning weighted favorites, a significant effect of the emotional tone was also observed (F(2, 12,846) = 3.206, p = 0.041) (Table 16). The descriptive statistics show that content with a positive emotional tone tends to receive more retweets and favorites on average compared to content with negative or neutral emotional tones. Therefore, hypothesis H1 is supported. Additionally, in order to check the variability in results, the study compared both weighted and unweighted engagement metrics for all strategies. The comparison results were the same for all strategies except the emotional tone. An ANOVA with emotional tone and raw (unweighted) favorites and retweets revealed a different result where emotional tone varied significantly between both categories of Twitter engagement metrics, (Favorites: F(2, 12846) = 11.648, p < .001; Retweets: F(2, 12846) = 10.045, p < .001). Counterspeech tweets with a negative emotional tone were more liked and retweeted than counterspeech tweets with a neutral or positive emotional tone.

Table 16

Effect of Emotional Tone on Weighted Favorites and Weighted Retweets

		Sum of Squares	df	Mean Square	F	Sig.
Weighted Favorites	Between Groups	.022	2	.011	3.206	.041
	Within Groups	43.298	12846	.003		
	Total	43.319	12848			
Weighted Retweets	Between Groups	.014	2	.007	10.062	.000
	Within Groups	8.839	12846	.001		
	Total	8.852	12848			

Table 17

Descriptive of Emotional Tone on Weighted Favorites and Weighted Retweets

	Descriptive	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Weighted Retweets	Negative	5674	0.0031	0.0302	0.0004	0.0023	0.0039	0	1
	Neutral	5502	0.0029	0.0246	0.0003	0.0022	0.0035	0	1
	Positive	1673	0.006	0.0145	0.0004	0.0053	0.0067	0	0.25
	Total	12849	0.0034	0.0262	0.0002	0.0029	0.0038	0	1
Weighted Favorites	Negative	5674	0.0057	0.0684	0.0009	0.0039	0.0075	0	4.12
	Neutral	5502	0.0049	0.0443	0.0006	0.0037	0.0061	0	1.89
	Positive	1673	0.009	0.0595	0.0015	0.0061	0.0118	0	1.52
	Total	12849	0.0058	0.0581	0.0005	0.0048	0.0068	0	4.12

South Asian and East Asian Counterspeech Occurrence Comparison

To examine the hypothesis, **H2:** South Asian American influencers are likely to exhibit lower engagement in counterspeech production compared to their East Asian American counterparts, a total of 500 tweets were manually coded. After selecting only the counterspeech tweets, the number came down to 396. A frequency distribution table showed that only 4.5 percent of the counterspeech in the US came from South Asian American influencers, whereas 29.8 percent of the counterspeech was produced by East Asian American influencers on Twitter (see table 18). This indicates that South Asian American influencers are likely to exhibit lower engagement in counterspeech production compared to East Asian American influencers. The results demonstrate notable variations in ethnic identity representation within the sample, which may have implications for the levels of engagement in counterspeech among different ethnic groups. Hypothesis H2, that South Asian American influencers are likely to exhibit lower participation in counterspeech production compared to their East Asian American counterparts, was also supported.

Table 18

Ethnic Identity of the Counterspeakers

Counterspeaker identity	Frequency	Percent	Valid Percent	Cumulative Percent
Black	41	10.4	10.4	10.4
East Asian	118	29.8	29.8	40.2
Hispanic	11	2.8	2.8	42.9
South Asian	18	4.5	4.5	71.5
Unsure	83	21.0	21.0	92.4
White	30	7.6	7.6	100.0
News, Organization, etc.	95	24.0	24.0	66.9
Total	396	100.0	100.0	

To examine the research question, RQ3a: Is there a difference in the counterspeech strategies employed by Asian and non-Asian American influencers? First, A chi-square test of independence was conducted to compare how Asian and non-Asian Americans were using emotional tone. The Pearson chi-square statistic was found to be χ^2 (4) = 29.652, p < .001, indicating a statistically significant association between the variables (see table 19).

Table 19

Crosstabulation of Asian_vs_Non Asian American and Emotional Tone

	Negative	Neutral	Positive	Total
Unsure	32 (19.51%)	29 (31.52%)	22 (64.71%)	83
Non Asian American	47 (28.66%)	22 (23.91%)	2 (5.88%)	71
Asian American	85 (51.83%)	41 (44.57%)	10 (29.41%)	136
Total	164	92	34	290

Table 20

Chi-square Test of Asian_vs_Non Asian American and Emotional Tone

	Value	Df	Asymptotic Significance (2-sided)
Pearson Chi-Square	29.652a	4	<.0001
Likelihood Ratio	28.527	4	<.0001
Linear-by-Linear Association	16.807	1	<.0001
N of Valid Cases	290		

Table 21

Crosstab of Asian_vs_Non Asian and Use of Visual Media

	None	Photo	Video	Total
Unsure Ethnic Identity	49 (28.65%)	11 (11.96%)	23 (40.31%)	83
Non Asian American	47 (27.33%)	20 (21.74%)	4 (7.02%)	71
Asian American	75 (43.86%)	27 (29.35%)	34 (59.65%)	136
Total	171	58	61	290

To answer the research question, RQ3b: Is there a difference in the counterspeech strategies employed by South Asian and East Asian American influencers? A series of chi-squared tests of independence were conducted to assess the relationship between the use of emotional tone by East Asian and South Asian Americans. The Pearson chi-square test indicated a difference in their use of emotional tone $\chi^2(2) = 7.495$, $p = .024$ (See table 24). The descriptive table further showed that both East Asian and South Asian influencers were using negative tones. A difference in the use of neutral and positive was also observed, but the sample was too small to draw conclusive result. The occurrence of counterspeech from South Asians, in general, was significantly fewer than that from East Asian Americans. Out of all the counterspeech tweets using negative emotional tone, 87.06% came from East Asian Americans, and 12.94% came from South Asian American influencers.

Table 22

Crosstab of East vs South Asian and Emotional Tone

	Negative	Neutral	Positive	Total
East Asian	74 (62.71%)	38 (32.20%)	6 (5.08%)	118 (100.00%)
South Asian	11 (61.11%)	3 (16.67%)	4 (22.22%)	18 (100.00%)
Total	85	41	10	136

A chi-square test was conducted to examine if there is any difference in the use of visual media strategy between East Asian and South Asian American influencers. The Pearson Chi-Square test yielded a statistically significant result ($\chi^2 = 10.268$, df = 2, $p = 0.006$, two-sided). Similarly, the Likelihood Ratio test also indicated a significant result ($\chi^2 = 14.297$, df = 2, $p = 0.001$, two-sided), as did the Linear-by-Linear Association test ($\chi^2 = 10.000$, df = 1, $p = 0.002$, two-sided) (Table 23). Surprisingly, very little use of visual media was observed in the counterspeech of South Asian influencers.

Table 23

Comparison Between East Asian and South Asian Influencers' Use of Visual Media

	None	Photo	Video	Total
East Asian	59 (78.67%)	25 (92.59%)	34 (100%)	118
South Asian	16 (21.33%)	2 (7.41%)	0	18
Total	75	27	34	136

To answer RQ4a: Is there a relationship between the racial diversity of the counterspeech location and the counterspeech occurrences, a bivariate regression analysis was performed to examine the effect of the racial diversity of the location on the counterspeech occurrence. The result showed that the 'racial diversity score', was statistically significant, F (1, 16910) = 538.096, p < .001. Specifically, the coefficient for 'racial diversity score' was estimated to be - 1.325 (SE = 0.057, t = -23.197, p < .001), suggesting that for every one-unit increase in 'racial diversity score', the counterspeech occurrences decreased by approximately 1.325 units (see table 30). This indicates that more diversity does not lead to higher counterspeech occurrences. Rather, higher racial diversity can be associated with a lower incidence of counterspeech. The regression analysis demonstrated that the model had statistical significance in explaining the variance in the occurrence of counterspeech. However, the overall model had a relatively low R-squared value (.03), indicating that only a small proportion of the variance in the counterspeech instances could be accounted for by the racial diversity index in the respective state. To answer the research question, RQ4b: Is there a relationship between the racial diversity of the counterspeech location and the specific counterspeech strategies used? A Pearson correlation was performed where the state's level of racial diversity(%) was examined in relation to counterspeech strategies. The results indicated that there was a statistically significant positive correlation between the State Diversity Index (%) and the ‘warning of consequences’ strategy (r = 0.026, p < 0.01), and a negative correlation with the use of visual media (r = -.079, p < 0.01) (see table 24).

Table 24

Relationship Between the Level of Racial Diversity and Counterspeech Strategies

		State_Diversity_Index(%)	Warning of consequences	Denouncement	Encouraging Participation	Use of Visual media
State Diversity Index(%)	Pearson Correlation	1	.026**	-.011	.009	-.079**
	Sig. (2-tailed)		.001	.166	.253	.000
	N	16912	16912	16912	16912	16912

**. Correlation is significant at the 0.01 level

The geospatial distribution further shows that most of the counterspeech tweets with consequences strategy were produced in the states with higher racial diversity, such as North Dakota (57.9%), New York (65.8%), Georgia (64.1%), etc., with the exception of South Dakota, whose racial diversity is only 35.6% (see Figure 11). Whereas, the use of visual media was higher in some of the states with lower racial diversity, such as, Maine (18.5%), Utah (40.7%), and Mississippi (40.5%) (see Figure 12). The numbers on the map of each state represent the racial diversity of that state in percentage. States with a higher Asian population, such as New York, seem to be more likely to use the consequence strategy and visual media. But California, with its high Asian population, was only moderate in its use of consequence strategy and visual media.

Figure 10

Geospatial Distribution of Counterspeech in the US States

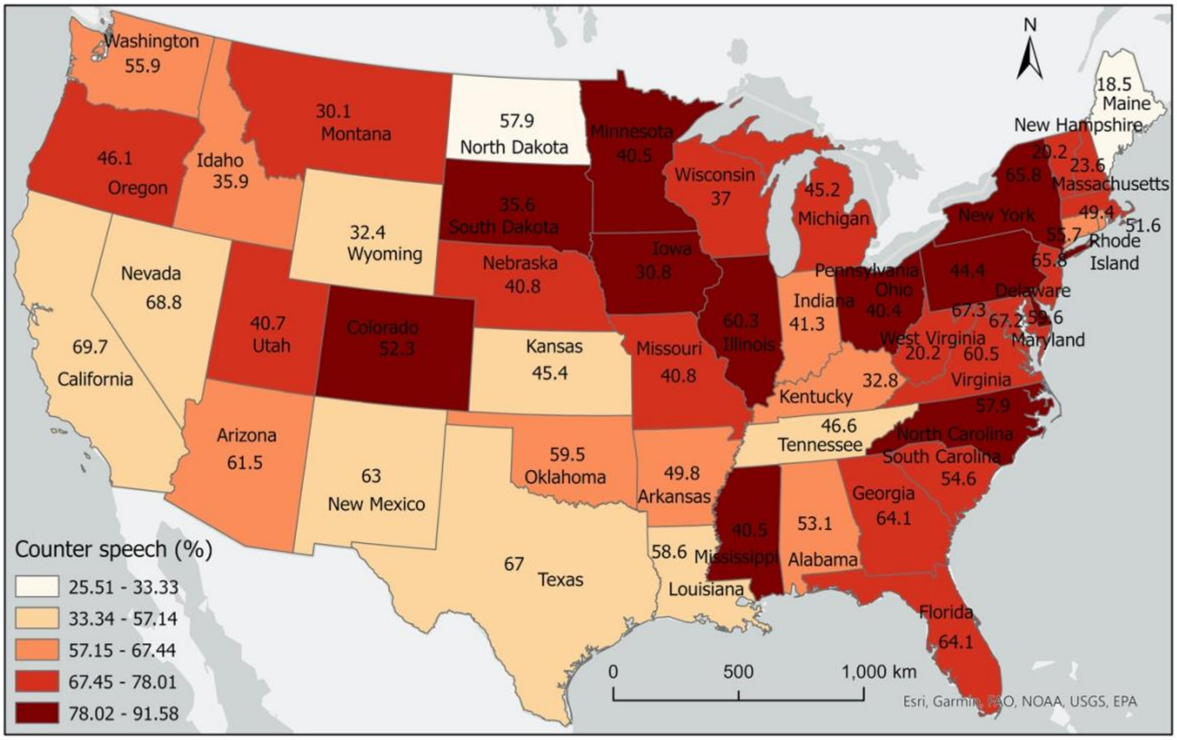


Figure 11

Geospatial Distribution of Consequences Strategy by States

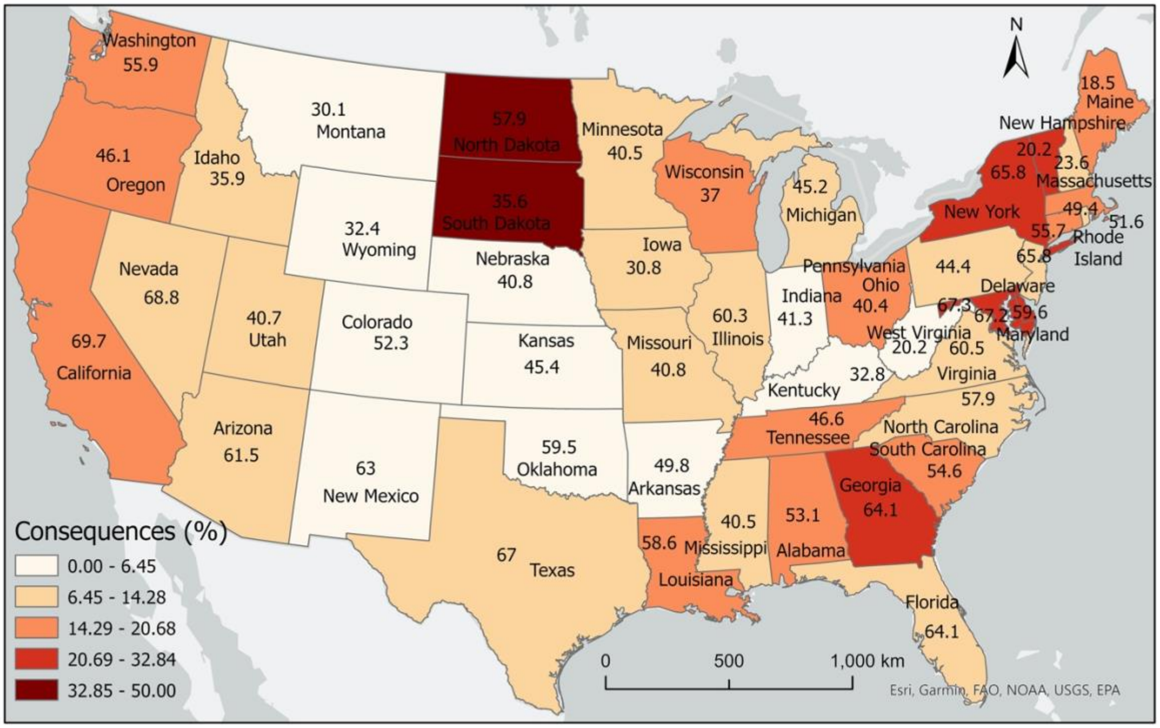
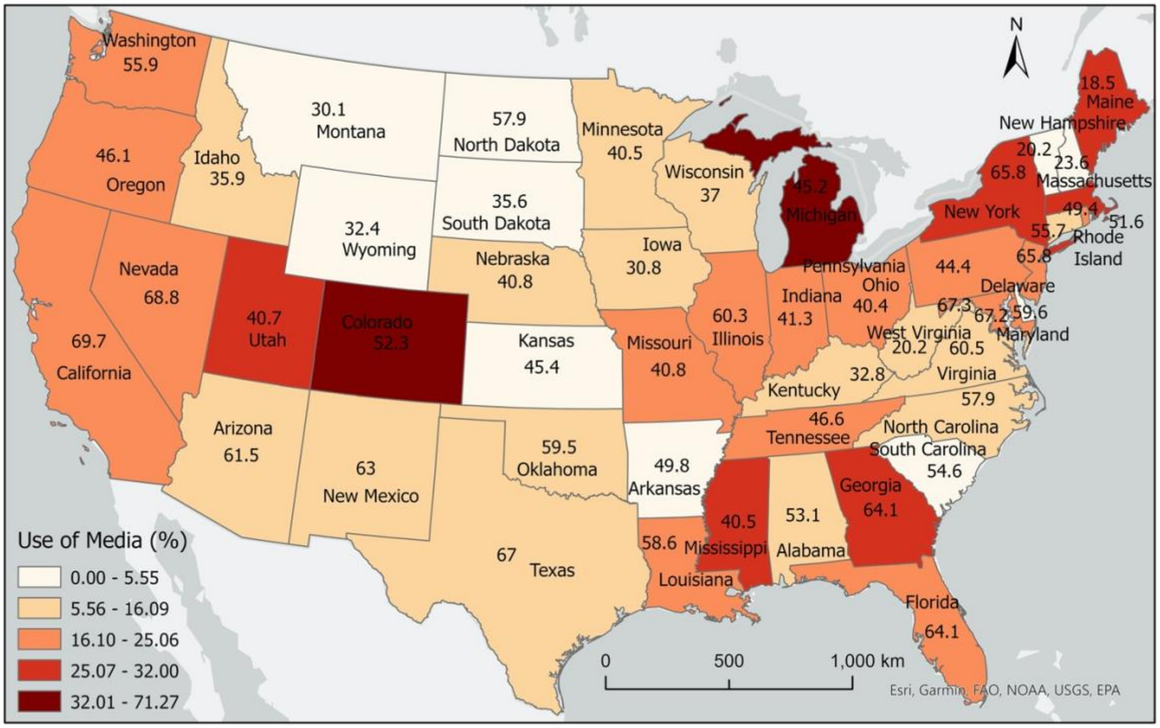


Figure 12

Geospatial Distribution of 'Use of Visual Media' by States



CHAPTER 5: DISCUSSION & CONCLUSION

The purpose of this study was to investigate the dynamics of counterspeech against anti-Asian hate on Twitter, particularly focusing on the strategies employed, the role of ethnic identity, and the impact of geographical diversity. In light of the research questions and hypotheses assessing the effectiveness of diverse counterspeech strategies on the Twitter platform, the data analysis revealed interesting findings. The following section discusses the effectiveness of the strategies found in the analysis.

Effectiveness of the Counterspeech Strategies

The findings suggest that the effectiveness of counterspeech strategies on Twitter varies depending on the specific engagement metric being considered. While the use of visual media appears to positively influence both favoriting and retweeting of counterspeech content, emotional tone was only associated with retweeting behavior. This indicates that emotional tones, which often evoke strong reactions and discussions, are more successful in terms of retweets. Retweets are a form of engagement that involves sharing content with one's followers, thereby amplifying its reach. Users may be more inclined to share emotionally charged content as it aligns with their sentiments or compels them to participate in ongoing conversations. On the other hand, the evidence that emotional tone doesn't significantly impact favorites or likes implies that these forms of engagement may be driven by other factors, that need empirical consideration in future research. Users may "like" or "favorite" content that they find agreeable or informative, even if it doesn't necessarily provoke strong emotions. These findings further suggest that, on Twitter, the use of visual media can be a particularly effective strategy for increasing user engagement, as indicated by higher weighted favorites and weighted retweets. This underscores the importance of incorporating visual elements, such as images and videos, into counterspeech content to maximize its impact. This aligns with Twitter’s business recommendation, indicating that videos and quizzes are likely to receive more outreach than other types of posts because they are quick-read and easy to digest (Demers, n.d.). In regards to the strategies ‘warning of consequences, denouncing hate speech and acts, and encouraging participation in anti-hate activities, this study didn’t find any significant effect on favorites and retweets. However, this finding is in line with past research, which concluded that not all types of counterspeech are equally effective (Benesch et al., 2016; Matthew et al., 2019). Within the broad spectrum of emotions encompassing both positive ones such as pride and thankfulness and negative ones such as rage, grief, and terror, the current study highlights the effectiveness of positive emotions. This finding further reinforces previous research that established a robust link between the tone of a message and its level of engagement. For example, past research showed that aggressive messages resulted in a 0% engagement rate, while casual or sincere texts garnered responses from 83% of recipients (Frenett & Dow, 2009). In this study, it becomes evident that a positive tone was more effective in terms of achieving broader outreach, particularly in the form of increased retweets. Although this study found more counterspeeches with a negative tone, they were not as effective as the ones with a positive tone. This means that when counterspeakers employed positive emotions in their responses to hate speech, they were not only successful in capturing the attention of their audience but also in compelling them to actively participate by sharing the content with their own followers. This heightened level of retweeting can be seen as a clear indicator of the content's resonance and influence within the online community, emphasizing the potential for positive and uplifting counterspeech to not only combat hate speech but also foster a sense of solidarity and support among Twitter users. Allyship is necessary to sustain and spread the message of a movement.

Additionally, the engagement metrics were weighted by the number of followers in this study to control the effect of follower size on number of likes and retweets. However, a comparison of weighted engagement measure and unweighted engagement measures showed that the results were the same for all the strategies except for emotional tone. An ANOVA with emotional tone and raw (unweighted) favorites and retweets revealed a different result when the effect of the number of followers is not controlled: negative emotional tone tweets were most liked and retweeted than neutral and positive emotional tone tweets, (favorites: F(2, 12846) = 11.648, p < .001; retweets: F(2, 12846) = 10.045, p < .001). A further investigation was conducted on only the counterspeech tweets posted by the influencers (more than 10,000 followers) to assess whether the result varies for users with a larger number of followers. An ANOVA on 1924 influencers' counterspeech tweets produced revealed that negative emotional tone is truly more effective for influencer counterspeakers in terms of both retweets (F(2, 1921) = 10.373, p < .001.) and favorites (F(2, 1921) = 12.775, p <.001) (see table 25).

Table 25

Descriptives of Emotional Tone on Favorites and Retweets (Unweighted)

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
						Lower Bound	Upper Bound
Retweets	-1.00	843	96.52	479.969	16.531	64.07	128.96
	.00	853	15.46	71.212	2.438	10.67	20.25
	1.00	228	51.31	514.225	34.055	-15.80	118.41
	Total	1924	55.22	368.485	8.401	38.75	71.70
Favorites	-1.00	843	306.65	1296.934	44.669	218.98	394.33
	.00	853	55.11	215.577	7.381	40.63	69.60
	1.00	228	182.21	1572.375	104.133	-22.98	387.40
	Total	1924	180.39	1030.963	23.504	134.29	226.48

This result paves the way for novel insights into influencer behavior and their audience on Twitter.

Using non-weighted likes and retweets as measures of engagement on Twitter carries significant implications. Firstly, it's essential to account for the number of followers when examining Twitter engagement metrics. Failing to make this adjustment can introduce bias into the data, favoring influencers with a substantial number of followers. This, in turn, may lead to an inaccurate representation of user behavior and engagement patterns.

Secondly, the notable difference in results provides a new perspective for two types of Twitter users. For individuals with a smaller following, the research findings suggest that incorporating a positive tone in their counterspeech can effectively increase engagement, aligning with hypothesis H1. However, influencers with a larger follower base can strategically leverage a negative tone to achieve similar engagement outcomes. This emphasizes the importance of adopting a nuanced approach to counterspeech strategy on the platform. The choice between a positive or negative tone should be contingent upon one's level of influence over the followers, ensuring a more tailored and effective engagement strategy.

Additionally, in the dataset of 12881 counterspeech tweets, only 18.5 percent were found as warning possible consequences of the hate, 7.2 percent were denouncing the hate act or speech, and 11.6 percent were encouraging participation in anti-hate activities. Such low instances of counterspeech strategy-use suggest a number of possible insights. The counterspeech in this study’s dataset might have relied on diverse approaches not previously studied. It's possible that other counterspeech strategies may have been employed that were not examined in this research. Besides, it is not surprising that certain strategies are more commonly used or perceived as more effective in countering hate issues on Twitter. Matthew et al.'s (2019) study found that one single strategy (hostility or negative emotion) encompassed 39 percent of the counterspeech, indicating that the majority of counterspeakers on Twitter rely on a limited number of strategies. The dominant strategies detected in this study, such as warning consequences, echoed the findings of past research. It also could be inferred that activists and individuals using emotional tone (57.2 percent) and visual media (24.6 percent) in their content may believe that their audience responds better to this type of strategy.

Counterspeech Strategies Against Hate Speech vs Hate Crime

Moving on to our second series of research questions, which explored the relationship between counterspeech strategies and counterspeech type, such as countering hate speech or hate crime. The analysis revealed significant correlations. Conversely, counterspeech exhibited a notable correlation with a negative emotional tone while speaking against hate speech, underscoring the fact that most individuals tend to employ a less positive tone when responding to hate speech. This observation aligns with previous research, particularly considering that a substantial portion of those engaging in responses to anti-Asian hate speech primarily expressed feelings of sorrow (Tong et al., 2021).

The findings pertaining to effective strategies reveal a noteworthy and actionable recommendation for activists combating hate speech. The analysis indicated that the utilization of visual media presents a substantial and positive correlation with heightened audience engagement. This signifies that incorporating visual elements such as images, videos, and GIFs into counterspeech efforts can be a powerful approach to enhancing the reach and impact of their message. However, despite the evident advantages associated with visual media, there appeared to be a deficiency in its use within counterspeech posts addressing hate speech and hate crime. The analysis revealed that there was no significant association between the use of visual media and counterspeech tweets against hate speech. This discrepancy suggests that those actively involved in responding to and countering hate speech might not be harnessing the full potential of visual media in their efforts. In light of this observation, a pertinent recommendation emerges: activists and advocates working to combat hate speech should consider integrating more visual media elements into their counterspeech strategies. By doing so, they can capture the opportunity to substantially boost audience engagement and, consequently, enhance the efficacy of their counter-narratives.

In the digital age, where attention spans are limited, visual content often garners more attention and has the potential to convey messages more powerfully. While text-based content undoubtedly plays a vital role in conveying information, visual media offers a dynamic and multi-sensory experience. It harnesses the innate human inclination to process images and videos swiftly, making it an effective vehicle for communication. When it comes to social media, visual elements can amplify the impact of counter-narratives, evoke empathy, and drive the message home with greater intensity (Nikolinakou & King, 2018).

Visual content's ability to transcend language barriers and cultural differences is another significant advantage. It speaks a universal language that resonates with diverse audiences (Buehner & Sommerfeldt, 2013), thereby expanding the reach of counterspeech efforts. Moreover, the immediate emotional response that images and videos can evoke often fosters a more profound connection between the message and the viewer. Therefore, this study recaps the importance of capitalizing on the engagement-boosting capabilities of visual media to effectively challenge and counter the spread of hate speech and hate crime.

Intergroup Relations, Racial Diversity and Counterspeech

Moving on to research questions RQ4a and RQ4b, which examined the racial diversity of the counterspeech location, illuminated several insights in this study. Contrary to the commonly held belief that ethnically diverse cities are inherently more resistant to hate speech, the analysis in this study suggests that such assumptions should be scrutinized more closely. While prior studies, rooted in intergroup contact theory, have posited that diversity fosters understanding and empathy among different racial groups (Pettigrew, 1998), the present findings demonstrate a more nuanced picture, indicating that more diversity does not lead to higher counterspeech occurrences. Rather, higher racial diversity can be associated with a lower incidence of counterspeech. In other words, the presence of racial diversity alone does not guarantee increased resistance to anti-Asian hate, which may involve a complex interplay of factors, such as the level of exposure to the ethnic group of the victims. This study's findings contribute to the theoretical discourse of racial diversity and resistance to Asian hate. This research prompts a reevaluation of longstanding assumptions within the field of intergroup relations and racial diversity. It challenges the conventional wisdom rooted in intergroup contact theory, which posits that greater racial diversity naturally leads to reduced prejudice and more active resistance against hatred. The findings indicate that this relationship is more nuanced, demonstrating that higher racial diversity does not necessarily correlate with increased counterspeech against anti-Asian hate. This suggests that while intergroup contact is essential, it may not be a one-size-fits-all solution. That is, close contact can reduce individual hate but might not resolve intergroup conflicts (Forbes, 1997).

The level of exposure to the targeted ethnic group and the perceptions of threat may be significant factors in the relationship between racial diversity and counterspeech. These variables are not typically emphasized in traditional intergroup contact theory but prove to be critical in this study. This addition underscores the need to consider the quality of intergroup interactions and the perceived threats faced by marginalized communities when assessing the impact of racial diversity. Therefore, to attain a more nuanced comprehension of intergroup exposure, future research endeavors may consider incorporating factors like racial and residential segregation when delving into issues related to intergroup hatred and its corresponding responses.

This study lends support to past studies that warned about the negative effects of the exposure to racial diversity (Craig & Richeson, 2014; Outten et al., 2012), Especially, highlighting the involvement of non-White individuals in perpetrating hate against Asian Americans, these findings shed light on the intricate dynamics surrounding the counterspeech to hate crime and hate speech. The observation gains significance when considering that a substantial number of hate incidents targeting Asian Americans were carried out by individuals from non-White racial backgrounds, as reported by the U.S. Department of Justice in 2021. This new perspective expands intergroup contact theory by acknowledging that hate is not limited to interactions between White and non-White individuals. It calls for a more comprehensive understanding of the role of racial identity and discrimination within marginalized communities highlighting the intricate dynamics within racial groups and the impact of intra-group relations.

In light of the findings, this study provides recommendations for policymakers, law enforcers, and researchers to caution against making broad assumptions about the positive impact of racial diversity on racial relations and urge a comprehensive examination of the underlying factors contributing to the complex landscape of hate and its counteractions. The complex relationship between diversity, perceptions of threat, and counterspeech suggests that a more comprehensive understanding is required to navigate the nuances of these dynamics.

Race and Ethnicity of Counterspeakers

The findings regarding the ethnicity of counterspeakers hold several important theoretical implications for understanding the dynamics of counterspeech among Asian and non-Asian American influencers on Twitter. Firstly, the observation that South Asian American influencers exhibit a lower incidence of counterspeech production compared to their East Asian American counterparts challenges existing assumptions about the uniformity of participation within the Asian American community. This finding marks a significant milestone in the discourse surrounding South Asian disconnects from the broader Asian identity. This disconnect appears to have permeated their counterspeech practices, as evidenced by the limited presence of South Asian influencers in American counterspeech efforts against anti-Asian hate. This further underscores the need to address this issue and foster a more inclusive conversation about the experiences of different Asian ethnicities. This also indicates that the inclination to engage in counterspeech may vary among other ethnic groups among Asian Americans, suggesting the need for a more nuanced understanding of how various factors influence counterspeech and allyship behavior within this diverse community.

Interestingly, a minimal use of visual media in the counterspeech efforts of South Asian influencers was observed. This discovery represents a noteworthy milestone in the ongoing discussion about the South Asian American influencers' counterspeech efforts. Specifically, East Asian Americans appeared to adopt a more assertive and proactive approach, utilizing multiple strategies to convey their messages effectively. In contrast, South Asian Americans, in addition to producing considerably fewer instances of counterspeech, seemed to employ fewer strategic elements in their messages. South Asian influencers post even fewer counterspeech on anti-Asian hate than other Black and White influencers. This observation raises theoretical implications for the discourse surrounding the ethnic experiences of Asian identity. Social identity theory posits that individuals derive part of their identity from their group affiliations, and this identification can lead to in-group favoritism and the distinction of "us" versus "them." Aligning with social identity theory’s assumption, it is possible that South Asian Americans have been experiencing a deficiency of belongingness to the Asian American in-group. It prompts us to reconsider the umbrella term 'Asian' and, at the very least, calls for a reevaluation of South Asians' perceived disconnect from the broader Asian identity. It is perhaps time to reexamine the root causes of this disconnect and engage in a more inclusive dialogue about the experiences of various Asian ethnicities.

The observation that Asian Americans employed a more negative tone in their counterspeech also aligns with social identity theory’s assumption of in-group sense, primarily through the lens of intergroup relations and identity dynamics. When examining how East Asian Americans and South Asian Americans employ different counterspeech strategies and emotional tones, it reflects the nuanced interplay of subgroup identities within the larger Asian American community. East Asian Americans and South Asian Americans may draw from their distinct subgroup affiliations and experiences when formulating their counterspeech to anti-Asian hate speech (Iwamoto & Liu, 2010; Jun et al., 2021).

The influencer counterspeakers in this study encompassed a diverse range of individuals, including politicians, government officials, mayors, actors, news presenters, and journalists, among others. As the coder examined the online information available about the counterspeakers through magazines, news media, their interviews, etc., many deeper insights unfolded. For example, the 'unsure' category mainly consisted of individuals with brighter skin tones, which made it challenging to determine their race or ethnicity without making assumptions. These influencers did not explicitly mention their race and ethnicity in their profiles. One influential counterspeaker, Chanel Rion, for instance, had a mixed racial background, with a Korean American mother and a White American father. It is problematic to determine her racial affiliation solely based on her phenotypic attributes, such as skin color, as she does not explicitly identify with a specific racial group. Nevertheless, she actively engaged in counterspeech against hatred towards Asian Americans, highlighting the complex nature of identity and solidarity. It is not known if her mixed Asian identity has anything to do with speaking up for Asians. But studying mixed-race influencers and their ethnic identity may help explain their activism and stances toward racial issues. This is a new avenue to expand race and social identity theory for mixed-race people, whose numbers are increasing in America.

In conclusion, this research has shed light on the nuances of counterspeech effectiveness, ethnic identity representation within counterspeakers, and the relationship between racial diversity and counterspeech strategies. These findings offer valuable insights for shaping more effective counterspeech efforts, understanding the complexities of Asian American identities, and addressing the challenges posed by anti-Asian hate speech. As we navigate the ever-evolving landscape of online discourse, we must continue to explore these dynamics to promote a more inclusive, empathetic, and informed digital society.

Limitations

While this dissertation has contributed valuable evidence and insights to the field of communication and hate speech, it is essential to acknowledge the limitations. First, this study primarily relied on Twitter data, which might not fully capture the entirety of hate speech and counterspeech occurrences across all online platforms. Different social media platforms may exhibit variations in content, engagement, and user behavior, which could influence the generalizability of the findings. The use of hashtags to identify counterspeech, while an efficient way to identify activists, will exclude those who use counterspeech but do not use hashtags or associate themselves with the anti-Asian hate movement. Hence, the 18,933 US tweets we found in the study did not cover people who do not use the #StopAsianHate hashtag. The large number of non-US #StopAsianHate tweets the study retrieved also warrants attention. While the researcher did not do a country-of-origin analysis of these tweets, they constitute a larger proportion of tweets than US tweets. The anti-racism movement, especially against Asians, has now gained attention from social media users worldwide. Further analysis of how the movement is gaining attention worldwide and how their counterspeech strategies differ from those of U.S. tweeters will enhance our understanding of the growth and expansion of the worldwide anti-hate movement.

Second, the research's focus on quantitative analysis, while valuable for uncovering patterns and associations, may not capture the richness of qualitative nuances present in hate speech and counterspeech. Qualitative insights, such as the specific language used and contextual factors, could provide a deeper understanding of the dynamics at play and should be considered in future studies. Third, the investigation into racial diversity and geographical context was primarily based on aggregate data at the state level. This approach overlooks potential variations within states or cities, which could be critical in understanding localized hate dynamics. Future research should consider more fine-grained geographic analyses. A focus on different racial diversity levels in comparable-sized cities may yield more insights on how racial diversity composition increases conflict levels and racial hatred. Yet the visual mapping attempts show that generally, racial diversity and counterspeech do not positively correspond to each other. For specific counterspeech strategies, certain states are more likely to employ strategies such as ‘warning consequences of hate’ and visual media.

Fourth, this research focuses on responses to anti-Asian hate through counterspeech. It could not delve deeply into the experiences and perspectives of individuals targeted by hate speech and their reactions toward these counterspeeches. Incorporating qualitative interviews or surveys with affected individuals could provide valuable insights into the psychological and emotional impact of these counterspeeches or allyships on them and inform more targeted counterspeech interventions.

The study's findings reveal a limitation in the existing counterspeech strategy taxonomies proposed by previous researchers, such as Benesch et al. (2016) and Mathew et al. (2019). These taxonomies, while valuable, were not comprehensive enough to fully characterize the rich landscape of counterspeech strategies employed in response to hate speech. Specifically, among the 18,933 tweets analyzed in the United States, only 18.5 percent were identified as warning against potential consequences of the hate, 7.2 percent were focused on denouncing the hate act or speech, and 11.6 percent were aimed at encouraging participation in anti-hate activities. These figures suggest that counterspeakers might have employed other, unclassified strategies that were not previously identified in the literature. To address this gap and uncover these additional counterspeech strategies, topic modeling can emerge as a promising avenue, as the dataset is too large for manual investigation. Latent Dirichlet Allocation (LDA) Topic modeling offers a systematic and data-driven approach to unveiling the hidden patterns and themes within the text (Albalawi et al., 2020; Guo et al., 2016; Kabir & Ha, 2023), providing a more comprehensive understanding of the strategies employed in response to the anti-Asian hate movement. The field of natural language processing and topic modeling is rapidly evolving as new methods and models have emerged over the last couple of years. Some novel approaches for topic modeling that can be employed for extracting new counterspeech strategies are LDA2Vec and NR-LDA. LDA2Vec combines Latent Dirichlet Allocation (LDA) and word embeddings to improve the interpretability of topics. It enhances the traditional LDA model by incorporating word vectors. NR-LDA (Non-Parametric Relational Topic Model) is an extension of the traditional LDA model that takes relationships between documents into account. It's useful for modeling topics in datasets with document-level relationships. A recently developed transformer-based language model by OpenAI is GPT-2, which can also be used for topic modeling by extracting topics from the generated text. However, a qualitative approach with human annotation can offer a nuanced and more accurate classification. Guo et al. (2016) suggested a novel avenue for annotating large amounts of data using crowdsourcing. Crowdsourcing is the practice of outsourcing data labeling and annotation tasks to a large group of online workers. This approach leverages the collective intelligence and labor of a diverse group of individuals, typically from around the world, to annotate or label data for training machine learning models, conducting research, and generating larger training datasets.

The nuances of relationships of counterspeech strategies and audience engagement warrant further investigation and consideration in the context of countering online hate speech and misinformation. This study did not assess the accuracy of the post and just how people react to different counterspeech strategies. Future research should explore these dynamics in depth to help promote constructive discourse on social media platforms.

Lastly, it is imperative to underscore that the comparative analysis of race and ethnicity of the counterspeakers presented within this dissertation pertains specifically to a cohort of social media influencers and not to the broader population of everyday Twitter users. Consequently, the findings should not be extrapolated to imply that non-celebrity counterspeakers, who constitute the majority of Twitter users, will necessarily exhibit similar behavioral patterns on social media platforms. Furthermore, it is essential to acknowledge that the quantity of counterspeech instances originating from non-Asian American counterspeakers (including individuals from Black, White, Hispanic, and other racial and ethnic backgrounds) as revealed in this study cannot be construed as representative of the entire spectrum of non-Asian counterspeakers across the United States. Rather, this research serves to elucidate distinctions in the use of counterspeech strategies by non-Asian counterspeakers, encompassing individuals of diverse racial and ethnic backgrounds such as Black, White, Hispanic Americans, among others. To recap, while these limitations provide valuable insights for refining future research endeavors, they do not diminish the significance of the findings presented in this study. Instead, they highlight opportunities for further exploration and improvement in the study of hate speech, counterspeech, and their complex dynamics in diverse online environments.

Recommendation for Future Research

Future research can extend its exploration of counterspeech practices to encompass a broader spectrum of common social media users by conducting surveys with a larger and more diverse sample. Leveraging the Twitter API's capability to access users' profile information linked to their tweets can facilitate the recruitment of survey participants. This approach not only allows for the collection of data on respondents' racial or ethnic identification but also can offer an opportunity to gain insights into the strategies employed by counterspeakers and how these strategies intersect with their ethnic experiences. Expanding the research to include a wider range of participants can enrich our understanding of counterspeech practices and their nuances within the broader social media landscape. Precisely, future research should undertake an extensive examination of the effectiveness of various counterspeech strategies, with a specific focus on strategies such as "warning of consequences" and the "use of visual media." This entails a nuanced investigation into how these strategies impact engagement and how their effectiveness varies in different online and offline contexts. By dissecting these strategies and understanding their underlying mechanisms, future researchers can provide critical insights into the design of more targeted and impactful counterspeech interventions. Given the intricate nature of hate speech within ethnically diverse communities, it is imperative for future studies to delve deeper into how distinct ethnic and racial groups within these communities interact and respond to hate speech. In addition to the quantitative engagement measures used in this study, future research can focus on the content analysis of replies to different types of counterspeech strategies. This inquiry may necessitate a comprehensive examination of intergroup relations, potential alliances, and conflicts that emerge in the process of countering hate. By unraveling these complexities, researchers can contribute to a more nuanced understanding of how hate manifests in diverse environments and inform strategies to foster unity and resilience.

As there are a large number of non-US tweets on the topic of anti-Asian hatred, further research can compare these tweets with US tweets to see the differences and similarities in solidarity and support for the cause, the use of counterspeech, and the success of the different counterspeech strategies.

Building upon the existing body of research on geographical location and racial diversity, future investigations should seek to uncover the intricacies of hate speech and counterspeech dynamics across diverse regions, cities, and neighborhoods. Examining the micro-level dynamics within urban environments can provide a granular understanding of how hate and counter-hate manifest and how counterspeech strategies evolve in response to localized factors. Such research is instrumental in tailoring interventions and strategies to address the unique challenges faced by communities in different geographic settings.

Conclusion

The goal of this dissertation was to investigate counterspeech strategies employed on Twitter in response to anti-Asian hate. It delved into the use of communicative tactics, emotional tones, and visual media and the effectiveness of those tactics, while also examining their geographical distribution across the United States in relation to racial diversity. This research holds significant importance, primarily due to the heightened prevalence of xenophobia and hate speech on social media platforms, especially during the COVID-19 pandemic. The negative impact of hate speech on minority groups, such as Asian Americans, underscored the urgent need to identify effective strategies for combating such expressions of hatred. As we conclude this dissertation, the crossroads of discovery and action are found. In the pursuit of delving into the realm of counterspeech strategies on Twitter, particularly in response to the disturbing rise of anti-Asian hate, an uncharted territory was ventured into. The intricate dynamics of communicative tactics, emotional tones, and the effectiveness of these responses were sought to be uncovered. Along the way, a geographical mapping was embarked upon, where the distribution of these counterspeech across the diverse landscapes of the United States, where the level of racial diversity intersects, was examined. The importance of this research becomes even clearer considering this age where xenophobia and hate speech find fertile ground on social media. The COVID-19 pandemic had not only exacerbated this issue but had thrust it into the forefront of our digital lives. The reduction in face-to-face communication left us vulnerable to unregulated torrents of hatred, and it became evident that the online realm was where the battle lines were drawn.

The groundwork had been laid by past studies, extolling the virtues of counterspeech as a potent tool against hate speech on social media. But a path less traveled was sought in this research, focusing on the unique challenges posed by anti-Asian hate speech during the pandemic. It was discerned that not all counterspeech strategies are created equal, and in some cases, they might inadvertently fan the flames of hatred. Yet, not merely one of discovery but also of innovation was a part of this dissertation. New pathways were forged in research methodology, crafting a machine learning model rooted in the wisdom of qualitative exploration. The intricate terrain of subjectivity was navigated, seeking to provide richer, more nuanced insights in training the machine learning models. Most importantly, peering into the internalized identities of Asian Americans, an understanding was reached of how their multifaceted experiences intersect with hate speech, unraveling a tapestry of identity and resilience. The legislative efforts to combat hate may have been critiqued as insufficient, but the findings in this study stand as a beacon of hope, offering practical recommendations for policymakers and law enforcement agencies. As the sun sets on this chapter of the dissertation, it is not an ending but a new beginning—an opportunity to shape a world where hate finds no refuge, and counterspeech emerges as a beacon of hope in the digital wilderness. The road ahead is challenging, but armed with knowledge, insight, and determination, we are prepared to forge ahead, making our world a better place, one counterspeech at a time.

REFERENCES

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine Learning Applications based on SVM Classification A Review. *Qubahan Academic Journal, 1*(2), 81-90.

ADL (2022). Online Hate and Harassment: *The American Experience 2022.* Retrieved from https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2022.

Aggarwal, C. C. (2018). *Machine learning for text.* Cham: Springer.

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science, 152*, 341-348.

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence, 3*, 42-42. https://doi.org/10.3389/frai.2020.00042

Allport, G. W. (1954). *The nature of prejudice.* Cambridge/Reading, MA: Addison-Wesley.

American Communities Project. (2020). *Diversity and Disparities.* Retrieved from https://s4.ad.brown.edu/Projects/Diversity/data/data.htm

Ancheta, A. N. (2006*). Race, rights, and the Asian American experience.* Rutgers University Press

Aziz, S. F. (2022). *The racial Muslim: When racism quashes religious freedom*. University of California Press.

Ahmed, W. (2021). Using Twitter as a data source an overview of social media research tools. *Impact of Social Sciences Blog.*

Bartlett, J., & Krasodomski-Jones, A. (2015). Counterspeech examining content that challenges extremism online. *DEMOS.*

Barnidge, M., Kim, B., Sherrill, L. A., Luknar, Ž., & Zhang, J. (2019). Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics & Informatics, 44*, N.PAG. https://doi-org.ezproxy.bgsu.edu/10.1016/j.tele.2019.101263

Barde, R. (2004). Plague in San Francisco: An essay review. *Journal of the History of Medicine and Allied Sciences, 59*(3), 463–470. https://doi.org/10.1093/jhmas/jrh104

Benesch, S., Ruths, D., Kelly P. Dillon, Haji Mohammad Saleem, and Lucas Wright. (2016). "Counterspeech on Twitter: A Field Study." Dangerous Speech Project.

Benesch S. & Jones D. (2019, August 13) Combating Hate Speech through Counterspeech. Retrieved from https://cyber.harvard.edu/story/2019-08/combating-hate-speech-through-counterspeech.

Berger, J., & Milkman, K. L. (2013). Emotion and virality: what makes online content go viral?. NIM Marketing Intelligence Review, 5(1), 18-23.

Bromell, D. (2022). *Regulating free speech in a digital age: Hate, harm and the limits of censorship.* Springer.

Buerger, C. (2021). Counterspeech: a literature review. *Available at SSRN 4066882*.

Boeckmann, R. J., & Liew, J. (2002). Hate speech: Asian american students' justice judgments and psychological responses. *Journal of Social Issues, 58*(2), 363-381. https://doi.org/10.1111/1540-4560.00265

Boisjoly, J., Duncan, G. J., Kremer, M., Levy, D. M., & Eccles, J. (2006). Empathy or antipathy? The impact of diversity. American Economic Review, 96(5), 1890-1905.

Briggs, R., & Feve, S. (2013). Review of programs to counter narratives of violent extremism.

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54–61. https://doi.org/10.1145/3209581

Breiman, Leo. (2001). Random forests. *Machine learning,* 45(1), 5-32.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190-206.

Buehner, T. M., & Sommerfeldt, E. J. (2013). Visual communication in the public sphere. *American Communication Journal, 15*(3), 1-13.

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From Text to Codings: Intercoder Reliability Assessment in Qualitative Content Analysis. *Nursing Research, 57*(2), 113-117. https://doi.org/10.1097/01.NNR.0000313482.33917.7d

Bochatay, N., Bajwa, N. M., Blondon, K. S., Junod Perron, N., Cullati, S., & Nendaz, M. R. (2019). Exploring group boundaries and conflicts: A social identity theory perspective. Medical Education, 53(8), 799-807. https://doi.org/10.1111/medu.13881

Baker. C. E. (2012) Hate speech. In: Herz M, Molnar P (eds) *The Content and Context of Hate Speech: Rethinking Regulation and Responses.* Cambridge.

Boonin, D. (2012). *Should race matter?: Unusual answers to the usual questions.* Cambridge University Press. https://doi.org/10.1017/CBO9781139003650

Brown, A. (2018). What is so special about online (as compared to offline) hate speech?. *Ethnicities*, *18*(3), 297-326.

Bouazizi, M., & Ohtsuki, T. O. (2016). A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, *4*, 5477-5488.

Brettschneider, C. L. (2012). *When the state speaks, what should it say?: How democracies can protect expression and promote equality.* Princeton University Press.

Chan, S. (1990). European and Asian Immigration into the United States in Comparative Perspective, 1820s to 1920s. In Yans-McLaughlin, V. (Ed.). *Immigration reconsidered: History, sociology, and politics*. Oxford University Press.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. Aggression and Violent Behavior, 40, 108-118. https://doi.org/10.1016/j.avb.2018.05.003

Cheng, C. I. (2013). Asian American firsts and the progress toward racial integration. In *Citizens of Asian America: Democracy and race during the cold war*. New York University Press.

Chugh, S. (2022). The effects of covid 19-related social media hate crime on Asian and Asian Americans' elf-esteem.

Craig, M. A., & Richeson, J. A. (2014). More diverse yet less tolerant? How the increasingly diverse racial landscape affects white Americans' racial attitudes. *Personality and Social Psychology Bulletin, 40(*6), 750-761.

Cao, J., Lee, C., Sun, W., & De Gagne, J. C. (2022). The# StopAsianHate movement on Twitter: a qualitative descriptive study. International journal of environmental research and public health, 19(7), 3757.

Capron, M. (2020, MAY 25). *'They bring it here!' Man threatens, harasses, spits on Asians, Washington cops say.* The News Tribune. https://www.thenewstribune.com/news/coronavirus/article242979446.html

Chae, A. (2022). # StopAsianHate: A Content Analysis of Public Library Statements Released in Response to Anti-Asian Hate. Public Library Quarterly, 1-28.

Chang, S., Chang, T., Nguyen, T., & Prasad, N. (2021). Impacts of the COVID-19 Pandemic to the Asian and Asian American Communities: Persistent History, Collective Resistance, and Intersectional Solidarity. *The Journal of Purdue Undergraduate Research*, *11*(1), 5.

Cheng, H. L., Kim, H. Y., Tsong, Y., & Joel Wong, Y. (2021). COVID-19 anti-Asian racism: A tripartite model of collective psychosocial resilience. *American Psychologist, 76*(4), 627

Chen, K., Duan, Z., & Yang, S. (2022). Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, *41*(1), 114-130.

Chen, N. C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(2), 1-20.

Cunha, A. A. L., Costa, M. C., & Pacheco, M. A. C. (2019, June). Sentiment analysis of youtube video comments using deep neural networks. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 561-570). Springer, Cham.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*(1-12), 24.

Chen, E., Deb, A., & Ferrara, E. (2022). # Election2020: The first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science*, *5*(1), 1-18.

Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access, 9,* 114381-114391.

Del Visco, S. (2019). Yellow peril, red scare: Race and communism in National Review. *Ethnic and Racial Studies*, 42(4), 626– 644. https://doi-org.ezproxy.bgsu.edu/10.1080/01419870.2017.1409900

Demers, J. (n.d.). The Psychology of Shareable Content. *Twitter Business.* Retrieved from https://business.twitter.com/en/blog/psychology-of-shareable-content.html

Demsas, J., & Ramirez, R. (2021, March 16). The history of tensions — and solidarity — between Black and Asian American communities, explained. *Vox*. Retrieved from https://www.vox.com/22321234/black-asian-american-tensions-solidarity-history

Denton, N. (2013). Interpreting U.S. segregation trends: Two perspectives. *City & Community, 12*(2), 156–159. https://doi.org/10.1111/cico.12019

Dey, L., Chakraborty, S., Biswas, A., Bose, B. & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier. *International Journal of Information Engineering and Electronic Business. 8*. 54-62 DOI:10.5815/ijieeb.2016.04.07

Dillon, K. P., & Bushman, B. J. (2015). "Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context." *Computers in Human Behavior , 45*, 144–150.

Dixon, J., Durrheim, K., & Tredoux, C. (2007). Intergroup contact and attitudes toward the principle and practice of racial equality. *Psychological Science, 18*, 867–872.

Donaghue, E. (2020, July 2). 2,120 hate incidents against Asian Americans reported during coronavirus pandemic. *CBSnews.* Retrieved from https://www.cbsnews.com/news/hate-incidents-against-asian-americans-reported-during-coronavirus-pandemic/

Elbahtimy, M. (2021). *The right to protection from incitement to hatred: An unsettled right.* Cambridge University Press.

Farokhmanesh, M. (2019, July 19). Instagram "tag cleaners" are fighting against digital vandalism. *The Verge.* Retrieved Sep 15, 2022, from https://www.theverge.com/2019/7/19/206 98192/instagram-moderation-tag-cleaners-digital-vandalism-gore-harassment-images-biancadevins

Facebook's Community Guidelines. Retrieved on Sep 15, 2022 from https://www.facebook.com/communitystandards#hate-speech

Freeman, M. (1995). Are there collective human rights?. *Political studies*, *43*(1), 25-40.

Farokhmanesh, M. (2019, July 19). Instagram "tag cleaners" are fighting against digital vandalism. *The Verge.* Retrieved Sep 15, 2022, from https://www.theverge.com/2019/7/19/206 98192/instagram-moderation-tag-cleaners-digital-vandalism-gore-harassment-images-biancadevins

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1-30.

Fischer, S. & Chen, S. (2021, Mar 23). #StopAsianHate hashtag goes viral following deadly attacks. *Axios*. Retrieved from https://www.axios.com/2021/03/23/stopasianhate-hashtag-viral

Forbes, H. D. (1997). *Ethnic conflict: Commerce, culture, and the contact hypothesis.* New Haven: CT.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* Unesco Publishing.

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. arXiv preprint arXiv:2006.01974.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on computational linguistics, 841–847

Gilbert, D. (2020, March 27). Anti-Chinese hate speech online has skyrocketed since the coronavirus crisis began. Vice news. https://www.vice.com/en/article/n7jywd/anti-chinese-hate-speech-online-has-skyrocketed-since-the-coronavirus-crisis-began

Gruzd, A. (2016). Netlytic: Software for Automated Text and Social Network Analysis. Available at http://Netlytic.org.

Gruzd, A. (2013). Emotions in the twitterverse and implications for user interface design. *AIS Transactions on Human-Computer Interaction, 5*(1), 42-56.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. *Probation Journal, 93*(2), 398–415. https://doi.org/10.1177/0264550519880595

Hall, S. (1990). Cultural Identity and Diaspora. In *Identity: community, culture, difference*. Lawrence & Wishart.

Hanasono, L. K., Matuga, J. M., & Yacobucci, M. M. (2019). Breaking the bamboo and glass ceilings: Challenges and opportunities for Asian and Asian American women faculty leaders. In *Asian Women Leadership*, 28-46.

Haynes, S. (2020, March 6). As *coronavirus spreads, so does xenophobia and anti-Asian racism,* Time. https://time.com/5797836/coronavirus-racism-stereotypes-attacks/

Ha, L., Ray, R., Matanji, F., & Yang, Y. (2022). How News Media Content and Fake News about the Trade War Are Shared on Twitter: A Topic Modeling and Content Analysis. In L. Ha & L.

Willnat (Eds.), *The U.S.–China Trade War: Global News Framing and Public Opinion in the Digital Age* (pp. 125–144). Michigan State University Press. https://doi.org/10.14321/j.ctv29z1h4p.9

Hatbase. Retrived from https://hatebase.org/

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In P. Melville, J. Leskovec, & F. Provost (Eds.), *Proceedings of the first workshop on social media analytics*, 80-88. New York, NY: ACM

Huynh, Q.-L., Devos, T., & Smalarz, L. (2011). Perpetual foreigner in one's own land: Potential implications for identity and psychological adjustment. *Journal of Social and Clinical Psychology, 30*(2), 133–162. https://doi.org/10.1521/jscp.2011.30.2.133

He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021, Nov). 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* (pp. 90-94).

Hoenkamp, E. (2011, September). Trading spaces: on the lore and limitations of latent semantic analysis. In *Conference on the Theory of Information Retrieval* (pp. 40-51). Springer, Berlin, Heidelberg.

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. Communication Theory, 16(1), 7–30. https://doiorg.ezproxy.mnsu.edu/10.1111/j.1468-2885.2006.00003.x

Ibrahim, F., Ohnishi, H., & Sandhu, D. S. (1997). Asian American identity development: A culture specific model for South Asian Americans. *Journal of Multicultural counseling and development*, *25*(1), 34-50.

Ibrahim, F., Ohnishi, H., & Sandhu, D. S. (1997). Asian American identity development: A culture specific model for South Asian Americans. *Journal of Multicultural counseling and development*, *25*(1), 34-50.

Iftikar, J. S., & Museus, S. D. (2018). On the utility of Asian critical (AsianCrit) theory in the field of education. *International Journal of Qualitative Studies in Education, 31*(10), 935–949. https://doi.org/10.1080/09518398.2018.1522008

Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on Twitter. *Neurocomputing, 315*, 496-511.

Izsák, R. (2015, January 5). Report of the Special Rapporteur on minority issues, Rita Izsák. U.N. General Assembly. UN Docs A/HRC/28/64. Retrieved Sep 15, 2022, from https://undocs.org/A/HRC/28/64

Iwamoto, D. K., & Liu, W. M. (2010). The impact of racial identity, ethnic identity, Asian values and race-related stress on Asian Americans and Asian international college students' psychological well-being. *Journal of Counseling Psychology, 57*(1), 79–91. https://doi.org/10.1037/a0017393Ji, Y., & Chen, Y. W. (2022). "Spat On and Coughed At": Co-Cultural Understanding of Chinese International Students' Experiences with Stigmatization during the COVID-19 Pandemic. *Health Communication*, 1-9.

Jackman, M.R., & Crane, M. (1986). "Some of my best friends are black...": interracial friendship and whites' racial attitudes. *Public Opinion Quarterly, 50,* pp. 459–86

Jagoo, K. (2022, May 17). South Asians Are Asians, Too. *Race and Social Justice.* Retrieved from https://www.verywellmind.com/south-asians-are-asian-too-5271761

Jang, S. H., Youm, S., & Yi, Y. J. (2023). Anti-Asian discourse in quora: Comparison of before and during the COVID-19 pandemic with machine-and deep-learning approaches. *Race and Justice*, *13*(1), 55-79.

Jendryke, M., & McClure, S. C. (2019). Mapping crime–Hate crimes and hate groups in the USA: A spatial analysis with gridded data. *Applied geography, 111*, 102072.

Jendrowski, J. (2019). Networks of Incivility on Twitter: The Changing Geography of Hate Speech in a New Social Media Landscape (Doctoral dissertation, State University of New York at Buffalo).

Ji, Y., & Chen, Y. W. (2023). "Spat On and Coughed At": Co-Cultural Understanding of Chinese International Students' Experiences with Stigmatization during the COVID-19 Pandemic. *Health communication, 38*(9), 1964–1972. https://doi.org/10.1080/10410236.2022.2045069

Jürgens, P., & Jungherr, A. (2016). A tutorial for using Twitter data in the social sciences: Data collection, preparation, and analysis. *Preparation, and Analysis (January 5, 2016)*.

Jun, J., Woo, B., Kim, J. K., Kim, P. D., & Zhang, N. (2021). Asian Americans' Communicative Responses to COVID-19 Discrimination in Application of Co-Cultural Theory. *Howard Journal of Communications*, *32*(3), 309-327.

Jun, J. (2012). Why are Asian Americans silent? Asian Americans' negotiation strategies for communicative discriminations. *Journal of International and Intercultural Communication, 5*, 329-348. https://doi.org/10.1080/17513057.2012.720700

Kabir, M. (2022). Topic and sentiment analysis of responses to Muslim clerics' misinformation correction about COVID-19 vaccine: Comparison of three machine learning models. *Online Media and Global Communication*. https://doi.org/10.1515/omgc-2022-0042

Kabir, M. E., & Ha, L. (2023). 8 How Mobile Users Differ from Non-Mobile Users in# IndiaFightsCorona on Twitter. In *Mobile Communication in Asian Society and Culture: Continuity and Changes across Private, Organizational, and Public Spheres*. Routledge.

Karani, D. (2018). Introduction to word embedding and word2vec. *Towards Data Sci.* Retrived from https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

Kibria, N. (1996). Not Asian, Black or White? Reflections on South Asian American Racial Identity, *Amerasia Journal, 22*(2), 77-86, DOI: 10.17953/amer.22.2.m363851655m22432

Kil, S. H. (2012). Fearing yellow, imagining white: Media analysis of the Chinese exclusion act of 1882. Social Identities, 18(6), 663–677. https://doi.org/10.1080/13504630.2012.708995

Kim, J. Y., Block, C. J., & Yu, H. (2021). Debunking the 'model minority' myth: How positive attitudes toward Asian Americans influence perceptions of racial microaggressions. *Journal of Vocational Behavior*, 131, 103648

Kramarae , C. (1981). *Women and men speaking.* Rowley, MA: Newbury House.

Kwok, I., & Wang, Y. (2013, June). Locate the hate: Detecting tweets against Blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C., & Sahay, S. (2017, August). Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online,* pp. 73-77.

Kulshrestha, R. (2019, July 19). A Beginner's Guide to Latent Dirichlet Allocation (LDA). Toward Data Science Retrieved from https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Lantz, B., & Wenger, M. R. (2023). Anti-Asian xenophobia, hate crime victimization, and fear of victimization during the COVID-19 pandemic. *Journal of interpersonal violence*, 38(1-2), NP1088-NP1116.

Leets, L., & Giles, H. (1999). Harmful speech in intergroup encounters: An organizational framework for communication research. In M. Roloff (Ed.), *Communication yearbook, 22*, 91–137. Thousand Oaks, CA: Sage Publications.

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology, 39*(7), 711-721.

Lee, J., & Ramakrishnan, K. (2020). Who counts as Asian. *Ethnic and Racial Studies, 43*(10), 1733-1756. https://doi.org/10.1080/01419870.2019.1671600

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Li, Y., & Nicholson, H. L. (2021). When "model minorities" become "yellow peril"—Othering and the racialization of asian americans in the COVID-19 pandemic. *Sociology Compass, 15*(2), e12849-n/a. https://doi.org/10.1111/soc4.12849

Li, C. (2022, June). The COVID-19 Hate Crime Act: Anti-Chinese Sentiment and Xenophobia in Times of Austerity. In *2022 8th International Conference on Humanities and Social Science Research (ICHSSR 2022)* (448-454). Atlantis Press.

Linvill, D. L., Warren, P. L., & Moore, A. E. (2022). Talking to Trolls—How Users Respond to a Coordinated Information Operation and Why They're So Supportive. *Journal of Computer-Mediated Communication, 27*(1), zmab022.

Lowe, L. (1996). *Immigrant acts: On Asian American cultural politics.* Duke University Press.

Lyman, S. M. (2000). The "yellow peril" mystique: Origins and vicissitudes of a racist discourse. *International Journal of Politics, Culture, and Society, 13*(4), 683-747. https://doi.org/10.1023/A:1022931309651

Lyu, H., Fan, Y., Xiong, Z., Komisarchik, M., & Luo, J. (2021). State-level Racially Motivated Hate Crimes Contrast Public Opinion on the# StopAsianHate and# StopAAPIHate Movement. arXiv preprint arXiv:2104.14536.

Lee, Y., Vue, S., Seklecki, R., & Ma, Y. (2007). How did Asian Americans respond to negative stereotypes and hate crimes? *The American Behavioral Scientist (Beverly Hills), 51*(2), 271-293. doi:10.1177/0002764207306059

Lowe, L. (1996). *Immigrant acts: On Asian American cultural politics.* Duke University Press.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P. & Mukherjee, A. (2019, July). Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media* (Vol. 13, pp. 369-380).

Mayall, A., & Russell, D. E. (1993). Racism in pornography. *Feminism & Psychology*, *3*(2), 275-281.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review, 38*(2), 128–146. https://doi.org/10.1177/0894439318791786

Mousavi, P., & Ouyang, J. (2021, August). Detecting hashtag hijacking for hashtag activism. In Proceedings of the 1st Workshop on *NLP for Positive Impact* (pp. 82-92).

Mukkamala, S., & Suyemoto, K. L. (2018). Racialized sexism/sexualized racism: A multimethod study of intersectional experiences of discrimination for Asian American women. *Asian American journal of psychology*, *9*(1), 32.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629-649.

Mayall, A., & Russell, D. E. (1993). Racism in pornography. *Feminism & Psychology*, *3*(2), 275-281.

Morning, A. (2001). The racial self-identification of South Asians in the United States. *Journal of Ethnic and Migration Studies, 27*(1), 61-79.

Mazumdar, S. (1989). Race and Racism: South Asians in the United States. *Frontiers of Asian American studies: writing, research, and commentary,* 25-38.

Muldoon, O. T., McLaughlin, K., Rougier, N., & Trew, K. (2008). Adolescents explanations of paramilitary involvement. *Journal of Peace Research, 45*(5), 681–695. doi:10.1177/0022343308094330.

Museus, S., & Iftikar, J. (2014). Asian Critical Theory (AsianCrit). In M. Y. Danico & A. C. Ocampo (Eds.), Asian American society: An encyclopedia, 95–98. SAGE.

MacKinnon, C. A. (1993). *Only words.* Harvard University Press.

McIlroy-Young, R., & Anderson, A. (2019, July). From “welcome new gabbers” to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 651-654).

McPhillips, D. (2020, Jan 22). How Racially and Ethnically Diverse Is Your City? https://www.usnews.com/news/cities/articles/2020-01-22/measuring-racial-and-ethnic-diversity-in-americas-cities

Nadal, K. L. (2019). The brown Asian American movement: Advocating for South Asian, Southeast Asian, and Filipino American communities. *Studies, 9*(10), 11.

Nadeau, C., & Bengio, Y. (1999). Inference for the generalization error. *Advances in neural information processing systems*, *12*.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social network analysis and mining, 11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

Nemes, I. (2002). Regulating hate speech in cyberspace: Issues of desirability and efficacy. *Information & Communications Technology Law, 11*(3), 193–220. https://doi-org.ezproxy.bgsu.edu/10.1080/1360083022000031902.

Nikolinakou, A., & King, K. W. (2018). Viral video ads: Emotional triggers and social media virality. *Psychology & marketing, 35*(10), 715-726.

Nghiem, H., & Morstatter, F. (2021). " Stop Asian Hate!": Refining Detection of Anti-Asian Hate Speech During the COVID-19 Pandemic. *arXiv preprint arXiv:2112.02265*.

Nteta, T. M. (2014). The Past Is Prologue: African American Opinion toward Undocumented Immigration. *Social Science History, 38*(3–4), 389–410. http://www.jstor.org/stable/90017041

Ocampo, A. C. (2016). *The Latinos of Asia: How Filipino Americans break the rules of race.* Stanford University Press.

Orbe, M. P. (1996). Laying the foundation for co-cultural communication theory: An inductive approach to studying “non-dominant” communication strategies and the factors that influence them. *Communication Studies*, *47*(3), 157-176.

Orbe, M. P., & Roberts, T. L. (2012). Co-cultural theorizing: Foundations, applications & extensions. *Howard Journal of Communications*, *23*(4), 293-311.

Oyserman, D., & Sakamoto, I. (1997). Being Asian American: Identity, cultural constructs, and stereotype perception. *The Journal of applied behavioral science*, *33*(4), 435-453.

Oliver, J. E., & Wong, J. (2003). Intergroup prejudice in multiethnic settings. *American Journal of Political Science, 47,* 567–582.

Onan, A. (2019). Consensus clustering-based undersampling approach to imbalanced learning. *Scientific Programming*, *2019*.

Onan, A. (2022). Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. Journal of King Saud University. *Computer and Information Sciences, 34*(5), 2098-2117. https://doi.org/10.1016/j.jksuci.2022.02.025

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232-247.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Outten, H. R., Schmitt, M. T., Miller, D. A., & Garcia, A. L.(2012). Feeling threatened about the future: Whites' emotional reactions to anticipated ethnic demographic changes. *Personality and Social Psychology Bulletin, 38*, 14–25

Oyserman, D., & Sakamoto, I. (1997). Being Asian American: Identity, cultural constructs, and stereotype perception. *The Journal of applied behavioral science*, *33*(4), 435-453.

Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Park, H., Reber, B. H., & Chon, M. (2016). Tweeting as health communication: Health organizations' use of Twitter for health promotion and public engagement. *Journal of Health Communication, 21*(2), 188-198

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate Speech: A Systematized Review. *SAGE Open*. https://doi.org/10.1177/2158244020973022

Pruitt, D. G., & Carnevale, P. J. (1993). Negotiation in social conflict. Pacific Grove, CA: Brooks/Cole

Pawar N. & Bhingarkar, S. (2020). Machine Learning based Sarcasm Detection on Twitter Data. *5th International Conference on Communication and Electronics Systems (ICCES),* pp. 957-961, doi: 10.1109/ICCES48766.2020.9137924.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology,* 49, 65–85.

Perry, B., & Alvi, S. (2012). 'We are all vulnerable' The in terrorem effects of hate crimes. *International Review of Victimology, 18*(1), 57-71.

Rahim, M. A. (1983). A measure of styles of handling interpersonal conflict. *Academy of Management Journal, 26*, 368-376.

Rabarison, K., Croston, M., Englar, N., Bish, C., Flynn, S., & Johnson, C. (2017). Measuring audience engagement for public health Twitter chats: Insights from #LiveFitNOLA. *JMIR Public Health and Surveillance, 3*(2), e34. DOI: 10.2196/publichealth.7181

Ramakrishnan. (2023, Feb 24). Key Facts on South Asians in America. *AAPI Data.* Retreived from https://aapidata.com/blog/facts-south-asians-2023/

Ren, J., & Feagin, J. (2021). Face mask symbolism in anti-Asian hate crimes. *Ethnic and Racial Studies, 44*(5), 746–758. https://doi.org/10.1080/01419870.2020.1826553

Rodriguez, L. Y. (2012). *Employee racial discrimination complaints: Exploring power through co-cultural theory*

Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics, 84*(1), 101-115.

Roth, W. D. (2018). Unsettled Identities Amid Settled Classifications? Toward a Sociology of Racial Appraisals. *Ethnic and Racial Studies, 41*(6), 1093–1112. doi: 10.1080/01419870.2018.1417616

Roth, W. D. (2018). Unsettled Identities Amid Settled Classifications? Toward a Sociology of Racial Appraisals. *Ethnic and Racial Studies, 41*(6), 1093–1112. doi: 10.1080/01419870.2018.1417616

Rossman, G. B., & Rallis, S. F. (2016). *An introduction to qualitative research: Learning in the field*. Sage Publications.

Ruiz, N. G., Edwards, K. & Lopez, M. H. (2021, April 21). One-third of Asian Americans fear threats, physical attacks and most say violence against them is rising. *Pew Research*

Center. https://www.pewresearch.org/fact-tank/2021/04/21/one-third-of-asian-americans-fear-threats-physical-attacks-and-most-say-violence-against-them-is-rising/

Shankar, R (1998). In Shankar, L. D., & Srikanth, R. *A part, yet apart: South Asians in Asian America*. Temple University Press.

Shankar, L. D., & Srikanth, R. (1998). *A part, yet apart: South Asians in Asian America*. Temple University Press.

Said, E. W. (2003). *Orientalism*. Penguin.

Snyder, R. M. (2015). An introduction to topic modeling as an unsupervised machine learning way to organize text information. *Association Supporting Computer Users in Education (ASCUE).*

Strossen, N. (2018). *Hate: Why we should resist it with free speech, not censorship. Inalienable rights series.* Oxford University Press.

Shin, R., Bae, J., Gu, M., Hsieh, K., Koo, A., Lee, O., & Lim, M. (2022). Asian Critical Theory and Counternarratives of Asian American Art Educators in US Higher Education. *Studies in Art Education, 63*(4), 313-329.

Soumya, S., & Pramod, K. V. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express, 6*(4), 300-305.

Sprinklr. (2023, May). Detect the sentiment present in customer messages accurately. Retrived from https://www.sprinklr.com/help/articles/ai-enrichments/detect-the-sentiment-present-in-customer-messages-accurately/645b5278e66f2e36b45187ac

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.*), Psychology of intergroup relations,* 7–12. Chicago, Nelson Hall.

Ting-Toomey, S., Stella, T. & Oetzel, J. (2003). Face Concerns in Interpersonal Conflict: A Cross-Cultural Empirical Test of the Face Negotiation Theory, *Communication Research, 30*(6). 599-624.

Ting-Toomey, S. (1988). Intercultural conflict styles: A face-negotiation theory. In Y. Y. Kim & W. Gudykunst (Eds.), *Theories in intercultural communication,* 213-235. Newbury Park, CA: Sage.

Tohill, L. & Ha, L. (in press). Election interference strategies among foreign news outlets, and audience engagement on Facebook, Twitter and YouTube during the U.S. 2020 Election. In Chattopadhyay, D. (ed.) *Global Journalism in Comparative Perspective: Case Studies on Journalistic Practice.* NY: Routledge.

Tong, X., Li, Y., Li, J., Bei, R., & Zhang, L. (2022). What are people talking about in #BlackLivesMatter and #StopAsianHate? exploring and categorizing twitter topics emerging in online social movements through the latent dirichlet allocation model. https://doi.org/10.1145/3514094.3534202

Twitter. (n.d.). Using the post activity dashboard. Twitter Help Center. Retrieved from. https://help.twitter.com/en/managing-your-account/using-the-post-activity-dashboard

U.S. Office of Management and Budget. 1997. *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*. Washington, DC: Executive Office of the President.

U.S. Department of Justice—Federal Bureau of Investigation. (2021). *Hate Crime Statistics, 2020.*

U.N. Office on Genocide Prevention and the Responsibility to Protect. (2019, June). United Nations strategy and plan of action on hate speech. Synopsis. Retrieved Sep 15, 2022, from https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_ plan_on_hate_speech_EN.pdf

United States Census Bureau. (2021, Aug 12). Racial and Ethnic Diversity in the United States: 2010 Census and 2020 Census. Retrieved from https://www.census.gov/library/visualizations/interactive/racial-and-ethnic-diversity-in-the-united-states-2010-and-2020-census.html

Vergani, M., Perry, B., Freilich, J., Chermak, S., Scrivens, R., & Link, R. (2022). PROTOCOL: Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents. *Campbell Systematic Review, 18*(2), https://doi.org/10.1002/cl2.1228

Van Atteveldt, W., Trilling, D., & Calderon, C. A. (2022). *Computational Analysis of Communication*. John Wiley & Sons.

Vodák, J., Novysedlák, M., Čakanová, L., & Pekár, M. (2019). Who is Influencer and How to Choose the Right One to Improve Brand Reputation?. Managing Global Transitions: *International Research Journal, 17*(2).

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access, 6, 13825-13835. https://doi.org/10.1109/ACCESS.2018.2806394

Wu, L., & Nguyen, N. (2022). From yellow peril to model minority and back to yellow peril. AERA Open, 8, 233285842110677. https://doi.org/10.1177/23328584211067796

Wong, Y. J., & McCullough, K. M. (2021). The intersectional prototypicality model: Understanding the discriminatory experiences of Asian American women and men. *Asian American Journal of Psychology*, *12*(2), 87.

Warner, W., & Hirschberg, J. (2012, June). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19-26).

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access, 6, 13825-13835. https://doi.org/10.1109/ACCESS.2018.2806394

Wong, Y. J., & McCullough, K. M. (2021). The intersectional prototypicality model: Understanding the discriminatory experiences of Asian American women and men. *Asian American Journal of Psychology, 12*(2), 87-99. https://doi.org/10.1037/aap0000208

Wu, F. (2002). *Yellow: Race in America beyond Black and White*. Basic Books.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, Neutralization strategies, and users' moral orientation. *Communication Research, 47*(6), 921–944. https://doi-org.ezproxy.bgsu.edu/10.1177/0093650219855330

West, C. (2023, Jan 9). Micro-influencer marketing: What you need to know. *Sprout Social.* Retrieved from. https://sproutsocial.com/insights/microinfluencer-marketing/

Wongmith, N. (2022). The Psychological Empowerment Impact of Twitter Microblogging: The Case of# stopasianhate During Covid-19 Pandemic (Doctoral dissertation, Syracuse University).

Wright, S. C., Aron, A., McLaughlin-Volpe, T., & Ropp, S. A. (1997). The extended contact effect: Knowledge of cross-group friendships and prejudice. *Journal of Personality and Social psychology, 73*(1), 73.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications, 36*(3), 6527-6535.

Yi, S. S., Kwon, S. C., Suss, R., Đoàn, L. N., John, I., Islam, N. S., & Trinh-Shevrin, C. (2022). The Mutually Reinforcing Cycle Of Poor Data Quality And Racialized Stereotypes That Shapes Asian American Health: Study examines poor data quality and racialized stereotypes that shape Asian American health. *Health Affairs*, *41*(2), 296-303.

Yoo, P. (2021). *From a whisper to a rallying cry: The killing of Vincent Chin and the trial that galvanized the Asian American Movement.* WW Norton.

Yu, T. (2006). Challenging the politics of the "model minority" stereotype: A case for educational equality. *Equity & Excellence in Education, 39*(4), 325–333.

Zavattaro, S. M., French, P. E., & Mohanty, S. D. (2015). A sentiment analysis of US local government tweets: The connection between tone and citizen involvement. *Government information quarterly, 32*(3), 333-341.

Zhang, Y., Zhang, L., & Benton, F. (2022). Hate crimes against Asian Americans. *American Journal of Criminal Justice, 47*(3), 441-461. https://doi.org/10.1007/s12103-020-09602-9

Zhang, Q. (2016). The mitigating effects of intergroup contact on negative stereotypes, perceived threats, and harmful discriminatory behavior toward Asian Americans. *Communication Research Reports, 33*(1), 1-8.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists.* O'Reilly Media, Inc.

APPENDIX A: INTERCODER AGREEMENT

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha	N Agree-ments	N Disagree-ments	N Cases
Counterspeech	96.7	0.91	0.91	0.91	145	5	150
Against hatespeech	82.7	0.56	0.57	0.57	124	26	150
Against hatecrime	92	0.84	0.84	0.84	138	12	150
Presentation of facts	92.7	0.32	0.32	0.32	139	11	150
Pointing out hypocrisy	94.7	0.40	0.40	0.40	142	8	150
Consequences	85.4	0.69	0.69	0.69	128	22	150
Affiliation	92	0.69	0.64	0.64	138	12	150
Denouncing hate	80	0.56	0.56	0.56	120	30	150
Encouraging participation	94	0.86	0.86	0.86	141	9	150

APPENDIX B: USE OF VISUAL MEDIA

Use of Visual Media
LINK
PHOTO
VIDEO
VIDEO,LINK
PHOTO,PHOTO,LINK
PHOTO,PHOTO
PHOTO,LINK
PHOTO,PHOTO,PHOTO,PHOTO,LINK
PHOTO,PHOTO,PHOTO
PHOTO,PHOTO,PHOTO,LINK
PHOTO,PHOTO,PHOTO,PHOTO
PHOTO,VIDEO
VIDEO,PHOTO,LINK
VIDEO,PHOTO
PHOTO,GIF
GIF
PHOTO,VIDEO,LINK

APPENDIX C: CODING SCHEME

Category	Code	Description	Example Tweets
Counterspeech/Not counterspeech	Counterspeech (1) Irrelevant (0)	Counterspeech is community-driven and crowd-sourced response to extremist or hate content with challenging the hate narratives. (Bartlett and Krasodomski-Jones, 2015). The various ways people may post counterspeech include 1) presentation of facts to correct misstatements or misperceptions, 2) pointing out hypocrisy or contradictions, 3) warning of possible offline and online consequences of speech, 4) identification with the original speaker or target group, 5) denouncing speech as hateful or dangerous, etc.	"It\'s not lost on me how many people said they would #StopAsianHate and are now actively defending a SF School Board Commissioner who made anti-Asian tweets." [Counterspeech] "The debut album by the Linda Lindas is adolescent angst done right! #thelindalindas #lindalindas #growinguplp #stopasianhate #punkmusic https://t.co/aHmjFBImzB" [Not Counterspeech]
Presentation of facts to correct misstatements or misperceptions	(present=1, absent=0)	Under this strategy, counter speakers will sometimes go to extraordinary efforts to convince outsiders that their knowledge or facts are incorrect.	*#COVID一19 is not the Flu and not SARS, pass it along https://t.co/0Gug6Typ3X*
Pointing out hypocrisy or contradictions.	(present=1, absent=0)	This strategy involves a counterspeaker who identifies instances of hypocrisy or inconsistency in the hate statements made by an individual. The primary aim of this tactic is to discredit the allegations, and the person in question may choose to offer justifications and rationalizations for their past behavior. Alternatively, if the individual is receptive to influence, they may commit to refraining from engaging in the dissonant conduct in the future. Attempting to discredit an opponent's position by pointing out their contradictory behavior or hypocritical stance.	"Wait till the first white causality of Corona virus then they'll be crying its a bioweapon against white people despite the hundred thousands non white causalities. FACT". "It\'s not lost on me how many people said they would #StopAsianHate and are now actively defending a SF School Board Commissioner who made anti-Asian tweets."

Counterspeech Strategies	Code	Description	Example Tweets
Warning of possible offline and online consequences of hate speech or hate crime	(present=1, absent=0)	Warning users of the potential repercussions of their nasty or harmful speech. Warnings and threats were an effective strategy because they have been implemented in high-profile incidents, such as successful demands that individuals be dismissed from their employment for internet material*. This includes tweeting news, facts, events, and incidents that may be the consequences of the anti-Asian hate crime or hate speech.	“*@RashidaTlaib The ban on travel was racist and the virus response is racist, the pattern here is that Ratshit is the racist clown. Everyone is hurting because of this Chinese virus response. It's because you ass clowns want the country on lock down to pander to these kinds of divisions.”*
Affiliation	(present=1, absent=0)	Expressing a shared identity to assert that specific speech is undesirable for members of a particular group. People tend to evaluate in-group individuals as more trustworthy, honest, loyal, cooperative, and important to the group than outgroup members*. Expressing solidarity and allyship.	*“I would especially urge my fellow Jewish friends in the US to patronise their local Chinese restaurants/takeouts EXTRA atm. American Jews have a historic special bond with American Chinese food places. We need to help counteract this racist &amp; ignorant paranoia re: coronavirus.”*
Denouncing speech or act as hateful or dangerous	(present=1, absent=0)	Identifying a speech, hashtags or any act as hateful, or racist. In this case, the content of hate speech is particularly denounced than the hate speaker.	“@realDonaldTrump @WhiteHouse this is why it is NOT OK to call this a #ChineseVirus #RacismIsAVirus #racist #Covid_19”.
Encouraging participation in counterhate	(present=1, absent=0)	Encouraging participation in counter hate movements, such as #StopAsianHate. Promoting the recognition and appreciation of the Asian American and Pacific Islander community's culture, history, and contributions, and finally, awareness and visibility of the Asian Americans.	Bravo, @SesameStreet! üíï Such an important step for representation and a bright spot of love in the effort to #StopAsianHate! https://t.co/NSa3YT68G9

Counterspeech Strategies	Code	Description	Example Tweets
Counterspeech to Hatespeech	(Present=1, Absent =0)	Whether the speaker addressing a hate crime or hate speech online. Hate speech is “an expression that is abusive, insulting, intimidating, harassing and/or inciting violence, hatred or discrimination." Hate speech includes verbal, non-verbal, and symbolic expressions.	[Hate speech] 20 years later, the trauma of seeing #Abercrombie\'s anti-Asian t-shirts still remains with many AAPIs like me. Many thanks to @Evan_Low, @angryasianman, @conniewang &amp; @kmiversen for making this story possible! https://t.co/8vatcwc0Do #netflix #WhiteHot #StopAsianHate
Counterspeech to hatecrime	(Present=1) /Absent =0)	Hate crime is "a criminal offense committed against a person, property, or society that is motivated, in whole or in part, by the offender’s bias against a race, religion, disability, sexual orientation, or ethnicity or national origin"	[Hate crime] COVID brought a pandemic of disease and of racism ‚Äî especially hate crimes against members of the AAPI community. Madam Chair Doris @Matsui4Congress shared her own powerful journey from an internment camp to Congress and urged us all to #StopAsianHate. https://t.co/lPMxTxOOD9 [Hate crime] 7 Asian women punched in their faces, elbowed, shoved by the same suspect within 2 hours on Sunday in Midtown Manhattan. All unprovoked. Suspect is in 20s, 5\'10", 190lbs, short blonde hair. @NYPDHateCrimes is investigating. 1/2 #StopAsianHate https://t.co/IhH0dfBpAm
Use of Humor	(present=1, absent=0)	Humor is perceived as a linguistic or communicative performance, since humor appears to have a unique effect in counter-argument approach. It may alter the dynamics of communication, de-escalate tension, and attract far more attention to a message than it would ordinarily receive (Banesch et al., 2016).	*“Can we not call it the Coronavirus or "Chinese Virus" and just call it the "Boomer Virus" and maybe people will take it seriously?*

APPENDIX D: MODELS & PYTHON CODES

https://github.com/mkabirdu/RF/blame/fd2bbaf6772375832be2fb02ece6db1d05ea1f82/%23SAH%20ML%20Model%20Building%20_2023-11-06.zip