Harm in Layers: Compositions of Misinformative Hate in Anti-Asian Speech and Impacts on Perceived Harmfulness

JIAWEI ZHOU, Georgia Institute of Technology, USA
GAURAV VERMA, Georgia Institute of Technology, USA
LEI ZHANG, Georgia Institute of Technology, USA
NICHOLAS CHANG, Georgia Institute of Technology, USA
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA

During times of crisis, heightened anxiety and fear create fertile ground for hate speech and misinformation, as people are more likely to fall for and be influenced by it. This paper looks into the interwoven relationship between anti-Asian hatred and COVID-19 misinformation amid the pandemic. By analyzing 785,798 Asian hate tweets and surveying 308 diverse participants, this empirical study explores how hateful content portrays the Asian community, whether it is based on truth, and what makes such portrayal harmful. We observed a high prevalence of **misinformative hate speech** that appeared to be lengthier, less emotional, and carried more pronounced motivational drives than general hate speech. Overall, we found that anti-Asian rhetoric was characterized by an antagonism and inferiority framing, with misinformative hate underscoring antagonism and general hate emphasizing calls for action. Among all entities being explicitly criticized, China and the Chinese were constantly named to assign blame, with misinformative hate more likely to finger-point than general hate. Our survey results indicated that hateful messages with misinformation, demographic targeting, or divisive references were perceived as significantly more damaging. Individuals who placed less importance on free speech, had personal encounters with hate speech, or believed in the natural origin of COVID-19 were more likely to perceive higher severity. Taken together, this work highlights the distinct compositions of hate within misinformative hate speech that influences perceived harmfulness and adds to the complexity of defining and moderating harmful content. We discuss the implications for designing more contextualized and culturally sensitive counter-strategies, as well as building more adaptive, explainable moderation approaches.

Content Warning: *Descriptions and quotes in this paper may be discriminative and inaccurate. We advise discretion as readers may find some expressions to be hateful, violent, or misleading.*

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Social media**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: misinformative hate, hate speech, misinformation, narrative, social media

ACM Reference Format:
Jiawei Zhou, Gaurav Verma, Lei Zhang, Nicholas Chang, and Munmun De Choudhury. 2024. Harm in Layers: Compositions of Misinformative Hate in Anti-Asian Speech and Impacts on Perceived Harmfulness. 1, 1 (November 2024), 23 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' addresses: Jiawei Zhou, Georgia Institute of Technology, Atlanta, GA, USA, j.zhou@gatech.edu; Gaurav Verma, Georgia Institute of Technology, Atlanta, GA, USA, gverma@gatech.edu; Lei Zhang, Georgia Institute of Technology, Atlanta, GA, USA, lzhang793@gatech.edu; Nicholas Chang, Georgia Institute of Technology, Atlanta, GA, USA, nickchang@gatech.edu; Munmun De Choudhury, Georgia Institute of Technology, Atlanta, GA, USA, munmund@gatech.edu.

XXXX-XXXX/2024/11-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Hate speech is offensive discourse targeting a group or an individual based on inherent characteristics such as race, religion, gender, or sexual orientation [57]. It has emerged as a pervasive societal concern with far-reaching consequences on individual mental well-being [66, 81, 83], real-world hate crimes [45, 82], and social cohesion [32]. What is more concerning is that the prominence of social media platforms has provided hate speech with an expansive outlet to disseminate discriminatory or prejudiced ideas. One evident case is the unprecedented rise in online and offline hate speech and crimes against the Asian population over the COVID-19 pandemic. Since 2020, there has been a disturbing 1662% rise in anti-Asian hate speech [80] and a 339% surge in anti-Asian crimes [82]. Fear of safety and the prevalence of microaggressions contribute to Asians' disproportionately higher stress, anxiety, and depression during this time [81, 83].

Unfortunately, the proliferation of anti-Asian hate is not entirely a surprise, given the interwoven relationship between hate speech and misinformation[1] during the pandemic [29, 43]. Misinformative hateful rhetoric takes advantage of people's sentiments [6] to make prejudiced views emotionally contagious and seemingly justified. In the case of COVID-19, misinformation has been constantly used to rationalize hate with frustrations of the pandemic, lockdowns, and economic influences. Prominent instances, such as claims of purposeful creation and release of the Coronavirus and assertions of Coronavirus as a biological and political attack or cover-up for 5G technology dangers [26], have constantly tended to blame the Asian community and provide fertile ground for hate speech to thrive. Despite the growing recognition of the dangers posed by both misinformation and hate speech, there remains a gap in understanding the overlaps and divergences in their constructions and how they may influence perceived severity.

As hate speech is characterized by its portrayal of specific communities, understanding the *compositions of hate* necessitates understanding how it represents the community – whether it is grounded in truth – and which entities it targets. This understanding allows us to further examine how these compositions influence the *perception of harm*. Based on these two facets, we attend to the hate speech problem by situating our work in the crisis-invoked rise of hatred and prejudice towards Asians — a community whose challenges are frequently downplayed by the "model minority" sterotype [47, 79] and is further marginalized by society's heightened emotional vulnerability during the crisis. Specifically, we pose the following research questions:

RQ1. How are misinformative and general hate speech composed, through what narratives and towards which targeted entities?

RQ2. How do content and individual factors in hate speech influence the perceived harmfulness?

To answer these questions, we first used an anti-Asian hateful tweets dataset `COVID-HATE` [33] and acquired the accessible portion of 785,798 hateful tweets. By extracting narratives and targeted entities, we identified a pervasive narrative of antagonism and inferiority framing where China and the Chinese were constantly directly targeted. Through misinformation classification, we discovered a high prevalence of misinformation in anti-Asian rhetoric. Our linguistic analysis found misinformative hate speech to be lengthier, less emotional, and more pronounced in motivational drives. It tended to influence readers through more information and motivational content. Misinformative hate also had a higher tendency to explicitly name entities to assign blame and highlight antagonism narratives. In contrast, general hate speech primarily focused on expressing emotions and calling for actions. Then, to empirically examine factors contributing to perceived harmfulness, we surveyed a diverse group of 308 participants to rate hateful tweets that varied in misinformation presence, hate narratives, and targeted entities. Our findings indicated that hate

[1] We refer to misinformation as a broad category that includes false or partially false information spread either unintentionally or intentionally [59].

speech containing misinformation, targeting demographics, or promoting divisive narratives was perceived as significantly more damaging. Additionally, individuals who placed less importance on free speech, had personal encounters with hate speech, and believed in the natural origin of COVID-19 were more likely to perceive a higher level of harm.

This paper makes several main contributions: **(1)** We offer a comprehensive understanding of misinformation-fueled hate speech's characteristics and perceived severity through comparisons with general hate speech. **(2)** We present empirical evidence on the prevalence and severity of misinformative hate speech and discuss potential avenues for improving hate speech mediation through user agency and contextualized moderation. **(3)** We offer a comprehensive understanding of how the Asian community is portrayed in hate speech and how content and individual factors contribute to higher harmfulness. These findings bear implications for developing more prioritized detection models and more cultural- and context-sensitive counterspeech.

Positionality and Ethics: All authors identify as Asian and majority are first-generation immigrants or foreign students in the U.S., with some authors being Chinese. Our shared, yet subtly different, identities helped us interpret the results while recognizing the cultural and societal dimensions of the issues. We acknowledge the granularity within the broad racial category of Asians in the limitation section. The survey study was approved by our Institutional Review Board, and we adopted best practices in managing and presenting public social media data, by removing personally identifiable information and paraphrasing content quoted as examples.

2 BACKGROUND AND RELATED WORK

In this work, we adhere to the widely-adopted definition of Asian from the U.S. Census Bureau, which describes Asian as a "person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent" [10].

2.1 Anti-Asian Hate and the Surge over the Pandemic

This work is motivated by the surge of anti-Asian hate speech and crimes during the COVID-19 pandemic. The unprecedented rise in hatred underscores the pressing need to address antagonism towards Asians, a community that remains relatively less studied in Human-Computer Interaction (HCI) and Computer-Supported Cooperative Work (CSCW).

Since the Gold Rush, the growing population of Chinese immigrants has been closely tied to discrimination that led to the 1882 Chinese Exclusion Act, marking the first and only legal prohibition of an entire ethnic group from immigration [42, 46]. This act not only fueled racism but also propagated a "Chinese invasion" rhetoric in the media [18], which gradually extended to other parts of Asia [46]. In the pervasive "alien invasions" framing of Asians, Chinese were denounced as coolies and slaves, Japanese as unscrupulous and aggressive, and Indians as dirty and diseased [46]. The othering of Asians was further sustained through governmental policies setting barriers to citizenship, education, and housing rights [29]. More recently, since the 1960s, Asian immigrants have been assigned with a new stereotype — a "model minority" [47, 79]. Despite appearing positive on the surface, this stereotype overlooks the diversity within the demographic and downplays any existing discrimination and obstacles, which also tends to further other Asians from other minority groups [47, 79].

The onset of the COVID-19 pandemic led to dramatic surges of Anti-Asian hate speech and crimes by 1662% and 339% [80, 82]. Former President Donald Trump's use of terms like "Chinese virus" or "China virus" in multiple tweets contributed to the rise in anti-Asian sentiment [29, 43]. Fueled by misinformation such as the intentional creation of COVID-19, discrimination and violence were justified under a "deservedness" rhetoric, taking advantage of people's emotions [6].

Consequently, Asians were found to experience disproportionate stress, anxiety, and depression [81, 83]. Dosono and Semaan [23, 24], in their work studying Asian American and Pacific Islander (AAPI) communities, have stressed the need for a *"narrative revision process"* to push back disparaging stereotypes and colonized framings. While prior studies have shown that Asians were portrayed as "dirty" or "sickly" during the pandemic and blamed for spreading germs [29], there is still a need for a more nuanced understanding of how the Asian community is framed and whether the representation is based on truth, so as to reshape prejudiced narratives. To this end, our work provides a more comprehensive understanding of the portrayal of the Asian community in misinformative hate speech and general hate speech, shedding light on future venues for developing more contextualized counterspeech and more prioritized reactions.

2.2 Hate Speech and Its Harm

Hate speech is a pervasive social issue with significant consequences on mental well-being [66, 74, 81, 83], violence and discrimination [45, 82], and social cohesion [32]. The contagion-like nature of harassment and conspiracy ideas on social media platforms can normalize the existence of hostile content [4] and breed hostility in exposed individuals [33, 62]. For example, Cheng et al. [17]'s work showed that experiencing both a negative mood and encountering troll posts from others substantially raises the likelihood of a user engaging in trolling behavior. What's worse, when hate rhetoric implies that targets have committed offenses, harmful comments are perceived as more justified [6]. Hate speech supported by misinformation exploits people's emotions and rationalizes hate through a sense of justice. The catalytic role of misinformation in discrimination and hatred is not unique to pandemic-induced anti-Asian hate. During crises, heightened anxiety and fear foster an environment conducive to the flourishing of hate speech and misinformation [50, 71]. The coexistence of hate and misinformation has been repetitively seen in xenophobia [41], stigmatization of health conditions [27], and the demonization of Muslim population [11].

To combat hate speech, a large body of prior research has demonstrated the capability to identify hate speech through technical means [21, 33], human labor [15], or combinations of the two [36]. In efforts to improve automatic detection, some work highlighted the importance of context learning when facing implicit forms of hate speech [25] or text modification attacks with typos and non-hate words [31]. Orthogonally, other studies have also cautioned the issues of fairness [52, 68] and effectiveness [13, 38] in detection models. The challenge of assessing hate speech lies in its subjective nature. Previous studies have emphasized the influence of various factors on people's judgments about the harmfulness of content, such as insider-outsider perspective differences [44, 48], cross-country values [39, 70], and political leanings [78].

Previous work indicates that individuals penalized in content moderation often perceive treatment as unnecessary or unfair [37, 38]. However, when explanations are provided, people are more likely to view moderation actions as fair and continue engagement with fewer removals in the future [37]. Accordingly, researchers have proposed frameworks to reason the types and severity of hate speech. Thomas et al. [74] conducted a literature review and summarized three types of criteria — audience, medium, capabilities — to differentiate types of hate speech. Scheuerman et al. [69] conducted interviews with experts and the general population to identify personal dimensions of severity, including perspective, intent, agency, experience, scale, urgency, vulnerability, medium, and sphere. Our work adds to these insights by providing empirical evidence on the influences of content and individual factors on perceived harm, as well as engaging a large and diverse group of participants while grounding our analysis in real-world data.

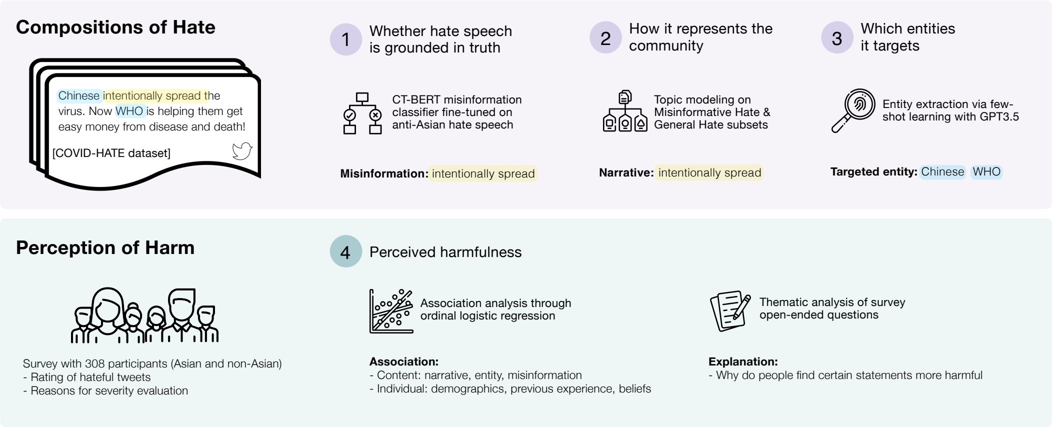

Compositions of Hate
Chinese intentionally spread the virus. Now WHO is helping them get easy money from disease and death!
[COVID-HATE dataset]
1 Whether hate speech is grounded in truth
CT-BERT misinformation classifier fine-tuned on anti-Asian hate speech
Misinformation: intentionally spread
2 How it represents the community
Topic modeling on Misinformative Hate & General Hate subsets
Narrative: intentionally spread
3 Which entities it targets
Entity extraction via few-shot learning with GPT3.5
Targeted entity: Chinese WHO
Perception of Harm
Survey with 308 participants (Asian and non-Asian)
- Rating of hateful tweets
- Reasons for severity evaluation
4 Perceived harmfulness
Association analysis through ordinal logistic regression
Association:
- Content: narrative, entity, misinformation
- Individual: demographics, previous experience, beliefs
Thematic analysis of survey open-ended questions
Explanation:
- Why do people find certain statements more harmful

Fig. 1. An overview figure summarizes the major parts of our study. We investigate **compositions of hate** (RQ1) through unpacking ① the presence of misinformation, ② the collective narrative, and ③ the targeted entities. Then, we study **perception of harm** (RQ2) by examining ④ the content and inidivual factors that precipitate higher perceived severity.

3 METHOD

Fig. 1 gives an overview of our study that aims to understand the constructions of misinformative and general hate speech and how they may influence perceived severity. First, we investigate the *compositions of hate* (RQ1) by exploring ① what grounds hate speech is based on, ② how it represents the affected community, and ③ which entities it targets. Then, we attend to the *perception of harm* (RQ2) by examining ④ the factors contributing to higher perceived severity.

3.1 Data

We used an existing dataset COVID-HATE [33] of anti-Asian hate speech on Twitter, which ranged between January 15, 2020 and March 26, 2021. This dataset contains real-time tweets (no retweets considered to focus on original expressions) collected through Twitter's official Streaming and Search APIs. Keywords used for data collection include COVID-19 (e.g., 'coronavirus', 'covid19') and hate keywords (e.g., 'kungflu', 'ch*nky') drawn from research literature and news articles. Then, the authors trained a BERT-based classifier with 3,555 manually annotated tweets to identify hateful tweets in the collected data (F1 score = 0.762). This dataset has been used extensively to develop and evaluate better hate speech detection tools [76, 86]. We acquired the accessible portion of hateful tweets, consisting of 785,798 hate speech instances. As social media companies, including Twitter, began to make data access for academic purposes challenging or unaffordable, this existing dataset aligns with our objective of understanding anti-Asian hate speech during the pandemic.

3.2 Misinformation Classification and Linguistic Comparison

While previous work has shown the interwoven relationship between hate speech and misinformation [43], there is limited work understanding the distinctions in linguistic style and harmfulness between hate speech that contains misinformation and one that does not. This exploration is crucial for developing more effective counterspeech and content moderation strategies.

As the first step of our analysis into the representation of the Asian community, we utilized and fine-tuned the highly-cited COVID-Twitter-BERT (CT-BERT) model [28] to detect misinformation in hate speech (① in Fig. 1). This model was developed by the winner among 166 participating teams

Table 1. Examples of hate speech as misinformative or not.

Example tweet	Mis	No
From day one, China hide the wuhancoronavirus information. WHO is an accomplice who helped them cheat the whole world. Don't let them escape!	✓	
Chinese intentionally spread the virus. Now they are getting easy money from disease and death!	✓	
Take down CCP. God bless the kind people.		✓

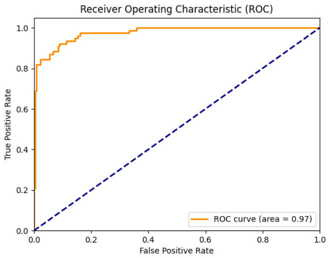


Fig. 2. ROC-AUC plot of the Misinformation Classifier.

in the AAAI 2021 shared task challenge of COVID-19 Fake News Detection [60]. We observed from manual annotation based evaluation that CT-BERT has a large number of false positive cases on hate speech data. To balance the sensitivity of the misinformation classifier facing hateful language, two researchers annotated 2300 hate speech tweets to be misinformative or not. We followed the principles of fact-checking [30] by focusing only on verifiable facts (e.g., "they are accountable for this creation of bioattack") rather than opinions (e.g., "they should be punished"). We then cross-referenced rulings from reputable fact-check organizations such as PolitiFact and the Washington Post's Fact Checker, as well as authoritative databases. Given that the collected tweets span from 2020 to 2021, it is highly likely that the stated facts have already been evaluated by professionals or are verifiable through data and research. If posts lacked professional rulings or direct evidence, we did not mark them as misinformation. The annotation process was conducted in an iterative approach. In the first round, two researchers double-annotated 200 samples and discussed initial differences. Then, we double-annotated another 200 samples to calculate the inter-rater agreement. To improve the relatively low interrater agreement (Cohen's κ) of 0.53, we added another round of double-annotation of 200 samples and achieved a κ of 0.63. We found this agreement acceptable in meeting fair to good standards [3], and split the remaining 1700 data points between the two researchers. 2k annotated data points were used to fine-tune CT-BERT model, and the model was validated on the remaining 300 self-annotated hate speech (Sensitivity 0.9559, Specificity 0.9483, Precision 0.8442, F1 0.8966). Fig. 2 shows the ROC-AUC plot of the classifier, and Table 1 provides example tweets for misinformative and general hate speech.

To analyze the linguistic differences between misinformative and general hate speech, we employed the Linguistic Inquiry and Word Count (LIWC) [73] as our analytic tool. LIWC is a validated psycholinguistic lexicon widely used in analyzing social media data and empirical studies of online harassment and misinformation [44, 85]. We specifically considered the following categories of psycholinguistic features:

- **Expressive features:** *Word count* calculates the total number of words. *Authentic speech* signals spontaneous and non-regulated language. *Tone* measures the direction and degree of emotions.
- **Sociocognitive expressions:** *Social referents* covers references to social relations like friends and families. *Cognitive process* represents human cognitive processing such as causation and differences. *Informal language* is common in daily conversations and *swear* or strong language considered vulgar or socially unacceptable.
- **Affective language:** *Affect* reflects expressions related to emotional status, covering *positive emotions* (e.g., "good, love") and *negative emotions* such as *anxious and anger.*
- **Motivational drives:** *Drives* represents people's urges or efforts to achieve specific goals, through expressions of needs for *affiliation, achievement, power, reward, or risk-avoidance.*

3.3 Narrative Extraction

To identify prevalent narratives in anti-Asian hate speech, we employed a topic modeling approach to extract latent topic clusters in both Misinformative Hate and General Hate subsets (② in Fig. 1). Specifically, we leveraged Latent Dirichlet Allocation (LDA) [7], an unsupervised machine learning algorithm extensively utilized for analyzing document corpora to uncover latent distributions of prevalent topics [85]. The data underwent the preprocessing and cleaning process, including tokenization, removal of stop words, and lemmatization. Including n-grams (n=1,2) with a frequency of appearance greater than 10, we transformed our dataset into a bag-of-words representation. Since LDA does not inherently determine the optimal number of topics, we used the coherence score as a metric to identify the most fitting number of topics for the model. The coherence metric assesses the tendency of words within a topic to co-occur, a measure that has been shown to correlate with expert evaluations of topic quality [56]. The number of topics (k) was iterated from 5 to 30 to calculate the coherence score. We observed the highest coherence scores at k= 13 and 10 for Misinformative Hate and General Hate respectively (Appendix A for all model performance).

Upon obtaining the best-fit LDA models' clusters and prevalent keywords, we employed thematic analysis to assign meaningful and human-interpretable labels to these topic clusters. This procedure involved three researchers, all self-identifying as Asians and possessing adequate knowledge of anti-Asian hate speech. The researchers independently coded the topics, relying on clusters of keywords and referencing back to sample posts associated with each topic. Subsequently, during collaborative sessions, the researchers convened to compare and discuss their individual codes, leading to the cohesive consolidation of codes into overarching thematic labels.

3.4 Targeted Entity Extraction

To extract targeted entities criticized in hate speech (③ in Fig. 1), we explored various methodologies involving GPT3.5 [9], LLaMA-2 [75], and open-source named entity recognition (NER) packages (SpaCy [35] and NLTK [51]). Following an initial test with 30 random examples, we observed better performance by GPT3.5 in identifying targeted entities. We attribute the relatively weaker performance of traditional NER packages to potential limitations in contextual awareness and language understanding [1, 19], especially in cases where entities like viruses and China are frequently mentioned in hateful tweets but not necessarily criticized. We adopted a few-shot learning approach using GPT3.5 [40, 87] for entity extraction and classification, categorizing entities into countries (e.g., China), demographics (e.g., Chinese), organizations (e.g., WHO), individuals (e.g., U.S. president), objects (e.g., vaccines), and cultures (e.g., gastronomy).

We formulated this task using the prompt of "Extract the targeted primary entity(s) that are criticized in tweets and label them as country, demographic, organization, individual, objects, and culture" followed by examples with expected outputs and additional instructions such as JSON output format and omission of non-criticized entities. We used the `gpt-3.5-turbo-0301` and set the temperature to 0. Several rounds of testing were conducted to refine the prompting approach and validate its performance, including a sanity check on representative documents from all topics extracted in Section 3.3. In the 65 Misinformative Hate samples, we identified nine false positive cases (13.8%, with five low-frequency entities appearing only once in the sanity test) and five false negative cases (7.7%). In the 50 General Hate samples, we found six false positive cases (12%, with two low-frequency entities appearing only once in the sanity test) and zero false negative cases. After the sanity check, we employed this approach on a random sample of 26,000 hate speech tweets across all topics, equally balanced between Misinformative Hate and General Hate.

After collecting results from GPT3.5, we calculated the frequency of all entities and retained only those with a frequency greater than 20. Subsequently, using a combination of manual grouping and automatic cleaning processes, we standardized different ways of referencing a single entity (e.g., “democratic party”, “democrats”, “dems” all pertaining to democratic political organization). We also specified the object category, classifying entities into viruses (including subcategories [2] of virus, virus with divisive labels, and virus as a weapon/terrorism), products, and medical supplies.

3.5 Survey on Perceived Harmfulness

Lastly, we conducted a survey to examine what and how content and individual factors affect perceived harmfulness (④ in Fig. 1). With approval from the Institutional Review Board at our institution, we recruited participants from the crowd-sourcing research platform Prolific[3]. We recruited 308 participants, giving us a 5.6% margin of error with a 95% confidence level for the target population of 56.9 million U.S. Twitter users [22]. We defined the inclusion criteria as adults who can read English and live in America. We oversampled the Asian population due to the focus of this study. Considering the sensitivity of our topic, participants were provided with a list of mental well-being resources and were reminded that they could stop participation at any time.

We conducted pilot testing to define our survey before launching the study in Prolific. We first ran multiple rounds of testing with the research team and through our institution’s student network. Then, the survey was tested with 10 respondents in Prolific. The final median completion time for our survey was 8 minutes 7 seconds, and each participant was compensated with an hourly rate of $12. Five participants did not complete the survey and were not included in the results. The survey was hosted on a website created for this study and consisted of four parts:

- **Demographic information:** Questions included age, gender, race/ethnicity, and history of international relocation (i.e., immigrated to or lived in another country for an extended time [12]).
- **Baseline attitudes and previous experience:** To understand participants’ pre-existing attitudes and experience of hate speech, we asked about how frequently they observed hate speech on social media and how harmful they found it to be, whether they have personally encountered hate speech that targeted them, and how they would prioritize the freedom of speech in relation to the potential harm it can cause. Additionally, to count for prevalent conspiracy theories in anti-Asian hate speech [43], we also asked about participants’ beliefs of COVID-19 origin.
- **Rating of hateful tweets:** Each participant rated ten hateful tweets randomly selected across all topics in `COVID-HATE` dataset. For each tweet and one attention-checker, participants were asked how harmful they found the given tweet to be (0 = Not harmful, 1 = Minimal harmful, 2 = Slightly harmful, 3 = Moderately harmful, 4 = Very harmful, 5 = Extremely harmful).
- **Perspectives on harmfulness:** We included two open-ended questions, appeared in the middle and at the end of the survey, about how they define harm in online content and why they find certain statements more harmful. Participants were encouraged to provide at least three factors.

All questions were required, but participants could choose not to disclose their gender, personal encounters with hate speech, and attitudes towards the origin of COVID-19. Table 2 and Appendix C describe demographics and baseline attitudes and experience. 51.95% and 21.43% participants have occasionally or frequently observed hate speech on social media, while 38.96% participants have personally encountered hate speech. On average, participants rated hate speech to be very harmful (mean = 3.70 on a 0-5 scale) and slightly preferred freedom of speech in relation to the protection of hate speech harm (mean = 2.35 on a 0-5 scale). 40.58% of participants believed COVID-19 originated naturally, and 22.73% believed it was created or modified by humans.

[2] Example entities under subcategories: virus - “covid”, virus with divisive labels - “Chinese virus”, virus as a weapon/terrorism - “bio weapon”

[3] https://www.prolific.com/

Table 2. Demographics of survey participants (N = 308). See Appendix C for baseline attitudes and experiences.

Age	18-24	56
	25-34	100
	35-44	70
	45-54	39
	55-64	25
	65 or older	18
Gender	Female	150
	Male	148
	Non-binary	7
	Other	2
	Prefer not to say	1
Race/Ethnicity	Asian	123
	Black or African American	31
	Hispanic or Latino or Spanish Origin	18
	Native Hawaiian or Pacific Islander	2
	White	127
	Other	7
International relocation history	Yes	64
	No	244

To examine the effects of content and individual factors on perceived harmfulness, we performed a multivariate ordinal logistic regression model [49, 64] to quantify the relationship between harmfulness and content and individual characteristics (Equation 1). The ordinal outcome variable of perceived harmfulness consists of six values (0-5). Content factors included misinformation presence, narrative types, and targeted entities. Individual factors included as adjusting covariates in the proportional odds modeling were demographics, international relocation history, prior experience with observing and encountering hate speech, baseline perception of hate speech harm, attitude of freedom of speech in relation to potential harm, and attitudes regarding the origin of COVID-19. We calculated the variance inflation factor to assess multicollinearity with continuous variables (i.e., baseline attitudes and experiences) centered by subtracting respective means [65].

$$\mathcal{H} \sim Misinfo + Narrative + TargetEntity + Demographics + BaselineAttitude + PrevExperience \tag{1}$$

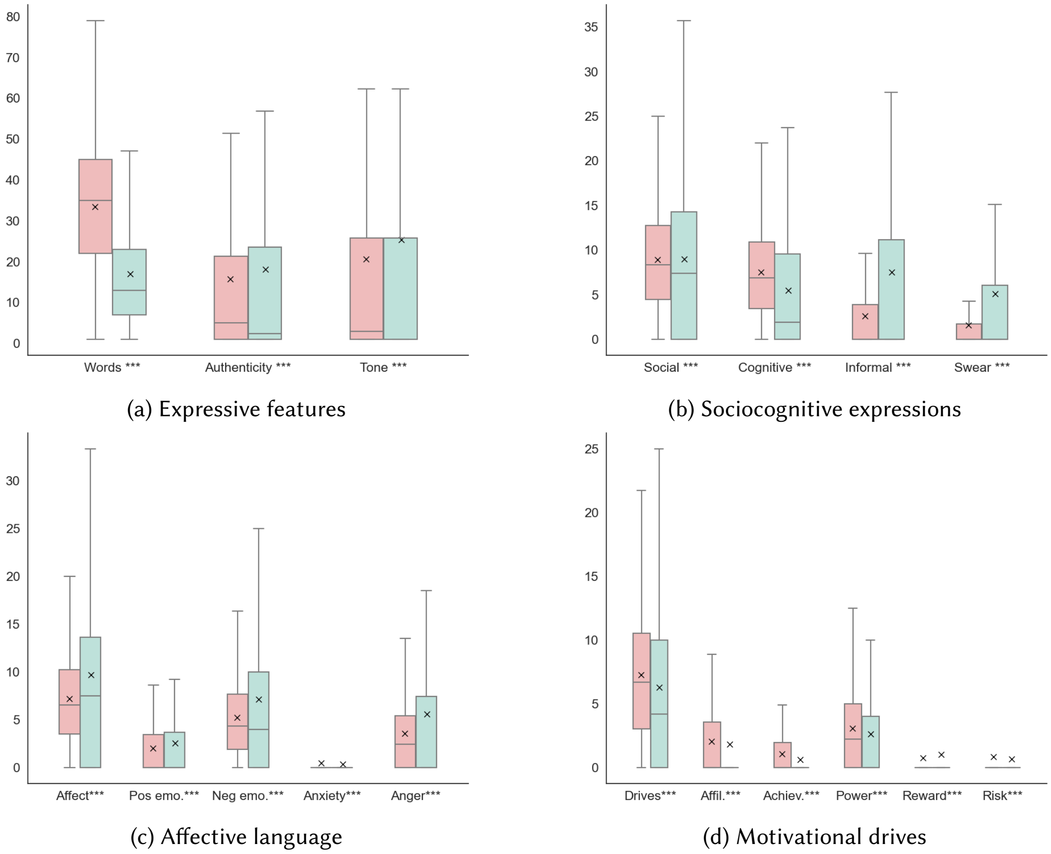
To further understand the factors contributing to the perception of higher severity, we conducted a thematic analysis to summarize factors participants considered when accessing harmfulness. Two researchers first went through all the data to establish familiarity and a general understanding. To inductively come up with categories, the two researchers independently annotated 50 participants' answers by identifying keywords or taking notes to summarize the answers. Then, they met and reviewed all the keywords and notes, and used thematic analysis to group lower-level annotations into higher-level categories. To validate the categorization, another round of coding was performed on an additional 50 participants' answers, ensuring no new categories emerged. Lastly, the first author used the themes to deductively re-code all the answers.

4 RESULTS

4.1 Compositions of Hate

4.1.1 Prevalence and Linguistic Style of Misinformative Hate (1 in Fig. 1). We observed a high prevalence of misinformation in hate speech, as well as significant linguistic differences between misinformative and general hate. Specifically, we identified that 28% of the hate speech instances contain misinformation (Misinformative Hate: 220,653 and General Hate: 565,145). A Mann–Whitney U test showed all linguistic features significantly different between the two types of hate speech.

As depicted in Fig. 3, overall, Misinformative Hate tended to use more words, a less informal and emotional tone, and more pronounced expressions of motivational drives. These findings



(a) Expressive features

(b) Sociocognitive expressions

(c) Affective language

(d) Motivational drives

Fig. 3. Box plots for LIWC categories between Misinformative Hate and General Hate (a higher number of "tone" below 50 suggests a negative emotional tone). Misinformative Hate appears to be lengthier, less emotional, and carries more pronounced motivational drives. All linguistic features between the two hate speech types are significantly different as per a Mann–Whitney U test (*** p<0.001, ** p<0.01, * p<0.05).

suggest that Misinformative Hate is not only about expressing hate but also involves manipulation through more information and motivational content, potentially aiming to persuade or influence the audience. In contrast, General Hate was more direct and fueled by negative emotions and swearing. We unpack these observations further in the paragraphs below.

Misinformative Hate appeared lengthier than General Hate. This can be interpreted together with the affective language feature where Misinformative Hate tended to combine emotion with actual statements or evidence to support the hate. For example, *"Because China failed on everything in 2019 [...] so they use this dirty lab generated virus to take revenge on all Chinese and the rest of the world..."* is an example of using rationale to motivate and justify the conspiracy theory of COVID-19 being lab-created. General Hate, on the other hand, focused more on expressing emotions and sometimes without real content, as seen in tweets like *"F*cking savages"* and *"Disgusting people"*.

We found that General Hate was more direct and casual with more informal and swearing expressions and more references to social relations. In contrast, Misinformative Hate included more descriptions of cognitive processing. We believe this aligned with the goal of Misinformative Hate to employ analytical presentations like comparisons and causations to rationalize the hate.

For example, *"Australia exports food to people need it, while China gives the world sh*t cheap socks and sh*t cheap TVs. [...]"* compared actions of different countries, and *"[...] this all started cuz some ch*nks want bat soup"* used causal reasoning.

Both Misinformative Hate and General Hate had a predominantly negative emotional tone, with General Hate displaying a more emotionally charged language style. It tended to be more emotional and negative, with increased expressions of anger and a broader range of affective expressions. Interestingly, expressions of anxiety and sadness were not commonly found in both types of hate speech, contrary to previous work suggesting that emotional appeals, especially fear and anxiety, are common in conspiracy theories and crisis times [58, 71, 77].

The expression of motivational drives was more apparent in Misinformative Hate through greater use of affiliation, achievement, and power words. We believe this tendency is in accordance with prevailing hate narratives during the pandemic, such as China's interference in intergovernmental organizations and foreign policies (topics MT_1, MT_5 discussed below in Sec 4.1.2), China benefiting from creating and spreading the virus (topics MT_8, MT_2, MT_0 in Sec 4.1.2), and the call for global unity against China (topic MT_{12} in Sec 4.1.2). Contrary to existing work that found fear and status enhancement were the more prominent motivators in prompting hate speech [61], expressions of reward and risk were not common in both Misinformative Hate and General Hate.

4.1.2 Collective Narratives (② in Fig. 1). We found shared portrayals in both Misinformative Hate and General Hate to be lying, terrorist-like, uncivilized, aggressive, and deserving of punishment. These narratives primarily surrounded ideas concerning suspicious virus origins and infection numbers, the lab-created virus as a bio-weapon, uncivilized culture and gastronomy, aggressive behaviors and policies, and the perceived need for punishing wrongdoing. Table 3 summarizes dominant topics, respective 13 and 10 topics in misinformative and general hate speech.

Misinformative Hate underscored antagonism and employed misinformation to prove and justify the portrayal of an enemy with manipulative actions and malicious intentions. For instance, one tweet writes *"China has killed 170,000 Americans with Chinese virus! Before this attack, China stole science & technology from US and use Kawakatsu to block Japan's linear technology."* These narratives together formed an "us and them" opposition, aligning with previous work that identifies othering as a common pattern in hateful discourse [2]. On the other hand, General Hate contained more action-based narratives especially in supporting discrimination and prompting divisive labels for COVID. Supporting discrimination was mainly done by reinforcing racial slurs, including direct slurs of "ch*nk" and "ch*ng ch*ng", indirect slang such as "ling ling", and appearance-attacking phrases like "slinty eyed". Prominent divisive COVID labels were "chinese virus" and "china virus".

When addressing the same narrative, Misinformative Hate incorporated explanations and justifications beyond simple emotional release. For instance, while General Hate used direct claims like *"China lied people died #chinavirus"* to present China as an untrustworthy nation, Misinformative Hate included logical reasoning and evidence to support such statement, such as *"Too bad we can't ask the Doctors and Nurses that have disappeared after they tried to tell the truth about the Virus. Why else would China want all test samples of the Virus testing to be destroyed? China doesn't want the World to find out how it started."* Similarly, for narratives about the suspicious origin and cure of COVID-19, General Hate mainly focused on reiterating and reinforcing divisive labels like "chinese virus" while Misinformative Hate used evidence and reasoning to establish the blame and suspicion.

4.1.3 Targeted Entities (③ in Fig. 1). We discovered the most commonly criticized entities were viruses, countries, organizations, and demographics, with Misinformative Hate more likely to name out specific targeted entities. Table 4 shows the frequencies of different entities.

We found more than half of the hateful tweets blamed the virus, and around 39% instances used divisive labels or weapon framing to refer to COVID-19, especially in misinformative hate

Table 3. Prevalent narratives derived from topic modeling and thematic analysis. Misinformative Hate underscored antagonism, while General Hate contained more action-based narratives.

Narratives		Topic	Perc	Keywords
Enemy	🔴	Threat to the world (MT_{12})	2.47%	china, world, virus, chinese, spread, wuhan, destroy, attacked_america, chinavirus
Manipulative	🔴	Interference in foreign countries (MT_1)	2.28%	china, chinese, virus, communist, usa, america, economy, country, want
	🔴	Manipulation on media storytelling and intergovernmental organization (MT_5)	1.73%	chinesevirus, china, globaltimesnews, wuhanvirus, chinavirus, drtedros, chinese, spread, shame
Malicious	🔴	Conspiracy theory on virus origins (MT_8)	3.80%	virus, chinese, china, call, people, blame, wuhan, say, come, spread
	🔴	Responsibility in starting and spreading the virus (MT_2)	1.11%	f*ck, https_co, coronavirus, covid, f*cking, f*ckchina, china, ch*nk, chinavirus, go
	🔴	Suspicious origin and cure of the virus (MT_0)	0.91%	virus, chinese, vaccine, go, one, god, cure, wuhan, covid, get
Untrustworthy	🔴	Untrustworthy nation with lies (MT_6)	4.52%	china, virus, lie, chinese, world, wuhan, know, spread, people, tell
	🟢	Untrustworthy nation with lies and different values (GT_1)	2.72%	lie, china, know, ccp, https_co, americans, way, people, communist, thank
Terrorist	🔴	Political party as a terrorist organization (MT_3)	1.18%	ccp, ccpvirus, world, chinese, communist_party, virus, https_co, take, terrorist, follow
	🔴	Weaponization of virus (MT_7)	1.84%	chinese, virus, kill, people, world, china, make, https_co, create, many
	🟢	Political party and country as a terrorist organization (GT_7)	5.22%	ccp, ccpvirus, china, terrorist, world, evil, https_co, chinavirus, chinazi, makechinapay
Uncivilized	🔴	Dehumanization of demographic and culture (MT_{11})	3.20%	chinese, virus, eat, sh*t, bat, https_co, china, people, animal, corona
	🟢	Dehumanization of demographic and culture (GT_8)	3.12%	kill, https_co, god, dog, ch*nky, eat, scum, f*cking, disgusting, animal
Aggressive	🔴	Fearmongering of aggressive and violent consequences (MT_9)	1.11%	kill, go, day, people, see, ch*nk, get, corona, ur, die
	🟢	Aggressive colonization and foreign policies (GT_6)	2.61%	china, country, https_co, kill_chinesevirus, modern_slave, pay, freetibet_freebalochistan, freehk_freehongkong, freebalochistan_chinabackoff, support_freehk
Call for punishment	🔴	Retributive responses and attribution of pandemic blame (MT_4)	2.59%	china, chinavirus, world, covid, make, wuhanvirus, virus, chinesevirus, pay, spread
	🔴	Boycotts as a punishment (MT_{10})	1.34%	chinese, virus, india, product, world, country, china, spread, use, kill
	🟢	Advocacy for global confrontation (GT_9)	14.27%	china, world, country, chinese, chinesevirus, india, people, make, time
Support discrimination	🟢	Supporting ethnic slurs (GT_3)	9.82%	f*ck, ch*nk, https_co, say, call, covid, coronavirus, shut, go, man
	🟢	Swearing insults and blaming behaviors (GT_5)	9.68%	get, sh*t, go, f*cking, *ss, https_co, b*tch, coronavirus, corona, man
	🟢	Appearance attacks and sexualized derogation on demographics (GT_4)	7.19%	d*nk, ch*nky, love, https_co, bastard, little, eye, s*ck, big, dunk
Promote divisive labels	🟢	Supporting national labels for the virus (GT_0)	9.04%	china, chinavirus, chinesevirus, wuhanvirus, covid, f*ckchina, https_co, die, world, make
	🟢	Supporting ethnic labels for the virus (GT_2)	8.26%	chinese, virus, china, people, call, come, wuhan, https_co, corona, say

speech. Nearly half of hateful tweets (44.97%) directly named out and criticized China. About 20% hate speech examples targeted demographics, particularly towards the Chinese. While fewer General Hate instances directly aimed at a population group, it was significantly more likely to use discriminative slurs in referring to the Chinese community. It also uniquely specified Chinese immigrants as an entity to blame. For example, one tweet stated *"Chinese immigrants should be globally banished! Unlike HKers, they immigrate like locusts instead of standing up against tyranny! They are savages with astronomical corruption, greed & no sign of civilization! They are source of inequality! They are VIRUS & survival of virus is criminal!!"*

Table 4. Occurrence frequencies of targeted entities in Misinformative Hate (N = 13000) and General Hate (N = 13000). A χ^2 test is performed to determine whether there was a significant difference between Misinformative Hate and General Hate (*** p<0.001, ** p<0.01, * p<0.05).

Entity	Total	Misinformative Hate	General Hate	p
Virus	50.48%	67.76%	33.19%	***
Virus with divisive labels	34.15%	40.51%	27.80%	***
Virus as weapon/terrorism	4.77%	9.08%	0.46%	***
Country	47.36%	50.63%	44.09%	***
China	44.97%	46.39%	43.54%	***
America	3.18%	5.75%	0.61%	***
Organization	21.03%	26.85%	15.22%	***
Chinese communist party	14.64%	18.78%	10.50%	***
Chinese-owned/founded companies	2.44%	4.88%	0.00%	***
Intergovernmental organization	2.82%	3.89%	1.75%	***
Democratic party	0.58%	1.16%	0.00%	***
Chinese government	1.11%	1.04%	1.18%	
Chinese media	0.67%	0.97%	0.37%	***
Demographic	20.52%	24.08%	16.96%	***
Chinese	16.47%	18.68%	14.26%	***
Chinese - discriminative slur	5.92%	1.58%	10.26%	***
Chinese - communists	0.50%	0.58%	0.42%	
Chinese immigrants	0.11%	0.00%	0.22%	***
Individual	6.66%	9.75%	3.57%	***
Xi	2.75%	3.61%	1.88%	***
Biden	1.74%	3.13%	0.35%	***
Trump	0.73%	1.04%	0.43%	***
Sanders	0.52%	1.03%	0.00%	***
Ghebreyesus	0.62%	1.02%	0.22%	***
Culture	4.07%	7.28%	0.85%	***
Object	3.49%	5.75%	1.24%	***
Chinese products	1.93%	2.62%	1.24%	***
Medical supplies	0.81%	1.62%	0.00%	***

Relatively fewer individuals were named, with most of them being politicians. Among them, one distinctive person was Dr. Tedros Adhanom Ghebreyesus, who was the Director-General of the World Health Organization (WHO). This can relate to our findings about the conspiracy theory of China's influences on intergovernmental organizations, as well as the evident presence of these organizations as a criticized entity.

4.2 Perception of Harm

From RQ2 (4 in Fig. 1), we found hate speech targeted at certain demographics, infused by misinformation, and prompt national or ethnic labels had the highest likelihood to be perceived as more harmful. Table 5 shows the relationship between perceived harmfulness and content and individual characteristics. Below we unpack our findings in detail.

On the individual level, participants with older age had a slightly higher chance (odds ratio: 1.082) of perceiving more harm than younger participants. Contrary to our assumption, we found our Asian participants did not have a higher likelihood of perceiving higher severity (odds ratio 0.990, CI 0.856-1.143). A potential explanation is provided by P136, who said *"I think part of me has become very desensitized to racism and anti-Asian rhetoric, so much so that it takes a lot to really offend me."*. No significant difference was found in participants' gender.

When considering people's preexisting attitudes and experiences, our results show that participants who have personally encountered hate speech were 1.128 times more likely to experience a higher intensity of harm. Similarly, people who believed in the natural origin of COVID-19 (odds ratio 1.330, CI 1.108-1.353) were more likely to find more elevated harmfulness. Participants who were negative or neutral about the natural origins of COVID claimed that they *"don't think it's harmful to question the narrative [...] as the circumstances of its creation are peculiar, to say the least."*

Table 5. Regression results between perceived harmfulness and content & individual attributes (*** p<0.001, ** p<0.01, * p<0.05). The odds ratio represents the likelihood of a higher harmful score in one group compared to another. The variance inflation factor (VIF) measures the severity of multicollinearity.

Variables	**Odds ratio**	p	**[CI, 0.025-0.975]**	**VIF**
Demographics				
Age	1.082	***	(1.031 - 1.135)	2.634
Gender (masculine-feminine)	1.028		(0.962 - 1.100)	2.112
Asian (yes)	0.990		(0.856 - 1.143)	1.985
International relocation history (yes)	1.031		(0.870 - 1.221)	1.487
Attitudes & Experiences				
Frequency of observing hate speech (low-high)	1.023		(0.935 - 1.120)	1.206
Baseline attitude of hate speech harm (low-high)	1.563	***	(1.458 - 1.677)	1.391
Personal encounter of hate speech (yes)	1.128	***	(1.051 - 1.210)	1.949
Prioritization of freedom of speech (high-low)	1.198	***	(1.140 - 1.259)	1.283
Attitude of COVID natural origin (neg-pos)	1.330	***	(1.218 - 1.452)	1.160
Targeted Entities				
Virus	0.797	*	(0.664 - 0.955)	1.405
Virus with divisive labels	1.191	*	(1.025 - 1.384)	1.852
Virus as weapon/terrorism	1.414	*	(1.013 - 1.974)	1.144
Country	1.453	***	(1.246 - 1.694)	2.442
Organization	0.882		(0.750 - 1.038)	1.872
Demographic	**1.907**	***	(1.618 - 2.248)	1.808
Individual	1.088		(0.877 - 1.350)	1.208
Culture	1.168		(0.898 - 1.520)	1.356
Object	0.883		(0.742 - 1.050)	1.325
Misinformation	**1.845**	***	(1.514 - 2.248)	5.148
Narratives				
Enemy	1.128		(0.770 - 1.652)	1.667
Manipulative	0.691	*	(0.502 - 0.952)	2.246
Malicious	0.660	**	(0.486 - 0.896)	2.318
Liar	0.791		(0.581 - 1.077)	1.642
Terrorist	0.914		(0.675 - 1.237)	2.254
Uncivilized	0.637	*	(0.446 - 0.909)	1.624
Call for punishment	1.012		(0.754 - 1.358)	2.103
Support discrimination	1.225		(0.875 - 1.716)	1.760
Promote divisive labels	**1.685**	**	(1.212 - 2.342)	1.508

(P11) or *"don't agree with weaponizing the term hate speech [...] (such as) social media censoring people for talking about Covid's lab origins, when all indications are that it did in fact come from a lab."* (P79). Consistent with our expectations, people who prioritized freedom of speech tended to place a lower severity on hateful tweets. P102 claimed that *"I acknowledge they have freedom of speech; so if they want to spout off hateful rhetoric, I won't try and stop them."*

Misinformative hate speech was rated significantly more harmful than non-misinformative hate tweets, with the second-highest adjusted odds ratio of 1.845 (CI 1.514-2.248). The main reason mentioned by our participants is that misinformation can justify hate and rationalize discrimination and threats. As explained by P104, *"people are more likely to agree with a statement that has 'facts' that back up its claim. "*. P72 further pointed out the long-term damage behind the rationalized hatred in that *"[f]alse information and hate speech can make it seem 'normal' to be a bigot towards minorities. When people feel more 'normal', they are more likely to act out against minorities verbally, physically, etc."* Misinformation becomes handy in creating or reinforcing stereotypes that are wrongful in nature or unfairly generalized from occasional or individual incidents, as said by P69, *"stereotyping not based in fact that paints a broad group of people in a negative light can easily spread online and after awhile be accepted as truth by others."* Participants also noted the secondary harm of misinformative hate in it being more persuasive with a confident tone and harder to identify or refute pseudo-facts, which is likely to erode the trust in authorities and science. When people buy into the falsehood and act according to it, risky behaviors – such as rejecting masks or vaccines because they could be manufactured by China – can result in health consequences. For instance, P85 said *"I know many people who did not survive Covid due to online content that discouraged safety measures and vaccines."*

In terms of targeted entities, hate speech targeting demographics were found to be the element with the highest adjusted odds ratio among all elements (1.981, CI: 1.684-2.330). As P194 pointed out, such expressions formed *"a clear 'us vs them' mentality with fingerpointing at exactly who is 'them'. This type of divisive mentality has been part of many atrocities in history, and it creates a blind devotion to the 'us' side with an equally blind hatred towards the 'them' side."* Tweets that targeted at counties were more likely to be seen as more harmful (odds 1.453, CI: 1.246-1.694). Participants thought such tweets acted as *"a dog whistle for attackers"* (P276) by *"paint(ing) China as a whole as the enemy and directed a lot of vitriol towards China - should someone who is riled up by these posts see someone who they think is related to China, they might act aggressively towards that person."* (P194). Hate speech that referred to COVID with divisive labels (odds: 1.191) or weapon framings (odds: 1.453) were more likely to be assigned with higher severity, while tweets diretly targeted virus tended to have lower harmful scores (odds: 0.797).

Among prevalent narratives, tweets promoting divisive labels had higher odds of being perceived as more harmful (odds: 1.685, CI: 1.212-2.342). Partcipants interested these labels were *"the catalysts for the spread of misinformation"* (P184), which could translate into real actions as *"certain people believed this and did violent things to Chinese people after seeing this type of thing online or because they kept hearing people like Trump call it the China virus"* (P290).

5 DISCUSSION

5.1 Tackling Harm in Layers with Context- and Culturally-Sensitive Counterspeech

Online hate speech is not a monolithic concept and demands customized counter-strategies that cater to the underlying problem to be tackled. Although misinformation-infused hate speech was evident in the case of pandemic-related anti-Asian rhetoric, this coexistence of misinformation and harassment is not a one-off event and has been repetitively observed in various other forms of discrimination, such as xenophobia [41] and the demonization of Muslims [11]. As revealed in our findings, misinformative hate was not only distinctive from general hate speech in linguistics, narratives, and likelihood of finger-pointing, but also perceived to be significantly more damaging. We found such misinformative hate tended to incorporate explanations and justifications beyond simple emotional release, which is concerning as it rationalizes discrimination by framing harmful views as a means of safeguarding identity or community under a seemingly just cause. Previous work has shown that online harassment presented with a sense of "deservedness" is seen as more deserved and justified [6]. This is especially true during crisis times when people are more vulnerable with heightened anxiety and fear [67, 71]. Therefore, addressing misinformative hate speech requires efforts to both mitigate antagonistic emotions and correct factual misconceptions. Counterspeech efforts should not just be about spreading the message of "love" online [53], but adapt to the underlying problems, whether moderating hate speech used as a vehicle to perpetuate misinformation or misinformation wielding discrimination to reinforce falsehood. If hate speech is a conduit to sensationalize misinformation or spread propaganda chastising a particular (usually minoritized) community, counterspeech also needs to consider fact-checking and approaches such as observational correction [8] in misinformation correction.

In utilizing counterspeech to mitigate harm, it is crucial to recognize the need for both refutation of hateful comments as well as public education regarding the discriminatory nature of certain expressions and stereotypes. In our survey, we noticed circumstances where people were unaware of anti-Asian discriminative slang. For instance, our participants expressed uncertainty about the nature of certain racial slang or references after seeing these words contained in multiple tweets or used in a more extreme manner. P103 , who rated the first tweet that contained the word 'ch*nk' as 'not at all harmful', admitted later in the open-ended question that *"I didn't realize ch*nk was an insult. It should have been rated slightly harmful."* Another participant, P13, said, *"I wasn't sure about the ch*nky rat one because that just sounds like [...] a cute way to say chunky as far as I understand."* Public education bears greater importance when the targeted community is not adequately or fairly depicted in traditional media and public knowledge — as seen in the case of Asians that either remain as opaque [16] or are generalized to certain stereotypes [5, 24]. In the absence of accessible or reliable information about a minority community or foreign culture, misinformation may fill the void. Within the narrative that portrays the Asian community as uncivilized, we found many posts not only overgeneralize certain uncommon or historical food cultures, but solely attribute virus outbreaks to them. Amplifying the perspectives or voices of the affected community can assist in correcting misconstrued cultural underpinnings of minoritized groups and address potential deficiencies in cultural awareness or community power.

5.2 Towards Adaptability and Explainability in Hate Speech Detection

One significant challenge of existing hate speech detection approaches lies in their static nature — data is labeled while algorithms learn from the existing dataset. However, online hate speech exists within social networks and communities where language is constantly evolving [14]. As new events unfold in the real world, new forms of hate expressions and types also percolate into the online sphere, posing a challenge for detection approaches to keep up. Therefore, our findings highlight the need for scalable solutions to go beyond static approaches and tackle hate speech. By offering an empirically grounded and nuanced understanding of the underlying hate narratives, as presented in this work, we envision the potential to enrich existing theories of hate speech that anchor on the relationship between social identities and power dynamics [84], by unveiling patterns that underlie different manifestations of online hate. This knowledge can, in turn, support the design of scalable technical solutions, such as those harnessing active learning, reinforcement learning, or online learning, to efficiently adapt to newer forms of online hate as they emerge.

An additional advantage of tailoring hate speech detection tasks with a focus on narratives and semantics is that it offers improved explainability by unpacking the relevance to certain narratives and highlighting phrases that semantically align with known hate speech patterns. State-of-the-art hate speech detection models, despite commendable performance in straightforward detection tasks, are shown to lack in explainability metrics [54], while contextual explanation is demonstrated to improve usefulness and trust [55]. Therefore, we posit our approach as a potential way to enhance the interpretability and trustworthiness of content moderation. For example, among tweets that prompt divisive labels for the virus (e.g., Chinese virus or China virus), many used historical examples such as the Spanish Flu to defend the normalization of ethnic labeling and dismiss the harm and hate it entails. In cases like this, narratives are useful in illustrating the broader impact where one comment with minimal intention of harm, when compounded with thousands like it, can escalate into unbearable consequences of division and harm. Providing explanations that contextualize hate speech within a broader environment of similar narratives can help explain the impact and intent behind such language and raise overall awareness.

5.3 User Agency and Potential Burden in Social Media Moderation Design

Social media users play a crucial role in moderation by actively flagging or reporting problematic content. Major platforms like Facebook, Twitter, Instagram, and TikTok currently utilize a single-class selection style, requiring users to specify *one* issue category, such as hate speech, false information, harassment, and violence. However, this arguably reductive approach dismisses the complex and multifaceted nature of problematic content [34], as our findings reveal the coexistence of these problems within hate narratives. In response, we propose a more granular selection design for social media moderation that utilizes user agency and contributes to more granular data, allowing for more flexible aggregation methods. Platforms can offer users more flexibility in expressing concerns about different dimensions of problematic speech. Users can choose multiple categorical types applicable to the information they wish to report and may further specify subcategories, such as intimidation, if desired. Platforms can then aggregate reports across multiple users and dimensions to identify content that poses significant risks for further review or actions. This approach can contribute to a more nuanced understanding of reported content, and possibly assist with more targeted counter-strategies in later stages.

Lastly, we reflect on the additional burden that may lay on the affected community and individuals. Research has emphasized how individuals derive their sense of self from group affiliations in tackling online hate [20] and how in-group membership is valuable in defining and evaluating online harassment [44, 48]. To some extent, their identities and lived experiences make them essential and well-equipped to identify and contextualize problematic speech. However, this can also place an unfair burden on them to constantly defend their community and educate others. In-group members or moderators often remain as a bottleneck, especially in community-centered platforms, leading to challenges of burnout and emotional burden [23, 72]. Additionally, we need to consider the preexisting and cultural psychological burden. Evidence shows that Asians have already faced disproportionate psychological effects during COVID-19 [81, 83]. Many – especially those in Confucian heritage cultures – value collectivism, where harmony is prized and individuals tend to avoid losing face or engaging in confrontations [63]. We posit that identifying and summarizing hate narratives has the potential to mitigate a portion of the physical and emotional labor by pre-informing moderators of potential stressors or assisting them in manual review prioritization.

5.4 Limitations and Future Work

In this work, we adopted the widely-used definition of the term “Asian” from the U.S. Census Bureau, which categorizes Asians as individuals with origins in the Far East, Southeast Asia, or the Indian subcontinent [10]. We recognize that this definition was historically politically motivated and does not distinguish the subgroups within the Asian population, despite great cultural diversity. Given that our survey participants were recruited from the U.S., our findings may be more representative of the U.S. context. Furthermore, this work is situated in the context of anti-Asian hate during the COVID-19 pandemic, and its findings are centered on the Asian community, which has not been sufficiently studied in HCI and CSCW. Nevertheless, we believe our methods are generalizable and could be reasonably applied to other crises, hate speech contexts, and communities. Our survey included demographic questions at the beginning to contextualize responses within the participants’ demographic and personal backgrounds, which could potentially bias participants into self-stereotyping in later responses. The misinformation annotation achieved a coefficient of 0.63. While indicating fair to good agreement according to conventional guidelines, the variability underscores the challenge of achieving consensus in identifying harmful content.

6 CONCLUSION

During crises, heightened emotional vulnerability can create an environment where hate speech and misinformation spread more easily, amplifying and reinforcing each other’s impact. This paper looks into the surge of anti-Asian rhetoric in the COVID-19 pandemic. Our computational findings from Twitter data and survey results characterized the compositions of hate and the dimensions of harm. Misinformative hate speech was not only prevalent and distinctive from general hate speech in linguistics, narratives, and direct blame, but also perceived as significantly more damaging. We discovered a pervasive antagonism and inferiority framing with misinformative hate underscoring antagonism and general hate emphasizing calls for action. Among all entities being explicitly criticized, China and the Chinese were constantly pointed out to assign blame with misinformative hate more likely to finger-point than general hate. Our participants found hate speech with misinformation, demographic targeting, or divisive narratives as significantly more harmful. On the individual level, people who placed less importance on free speech, had personal encounters with hate speech, or believed in the natural origin of COVID-19 were more likely to perceive an increased level of detriment. Collectively, our work seeks to spark conversation and future study in 1) designing more contextualized counterspeech that caters to the underlying problem, whether it involves factual misconceptions or unfamiliarity in subtle forms of prejudice, 2) creating more adaptive and interpretable detection approaches through narratives and semantics, 3) enhancing user agency and awareness of potential burden in moderation design.

REFERENCES

[1] Oshin Agarwal and Ani Nenkova. 2023. Named Entity Recognition in a Very Homogenous Domain. In *Findings of the Association for Computational Linguistics: EACL 2023*. 1805–1810.

[2] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. 2019. "The enemy among us" detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)* 13, 3 (2019), 1–26.

[3] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics* 27, 1 (1999), 3–23. https://doi.org/10.2307/3315487

[4] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[5] Tiffany Besana, Dalal Katsiaficas, and Aerika Brittian Loyd. 2019. Asian American media representation: A film analysis and implications for identity development. *Research in Human Development* 16, 3-4 (2019), 201–225.

[6] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When online harassment is perceived as justified. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[7] David M Blei, Andrew Y Ng, and Michael T. Jordan. 2002. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, Vol. 3. 993–1022.

[8] Leticia Bode, Emily K Vraga, and Melissa Tully. 2020. Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review* (2020).

[9] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei Openai. 2020. Language Models are Few-Shot Learners. (2020).

[10] U.S. Census Bureau. [n. d.]. About the topic of race. https://www.census.gov/topics/population/race/about.html

[11] Brian Robert Calfano, Paul A Djupe, Daniel Cox, and Robert Jones. 2016. Muslim mistrust: The resilience of negative public attitudes after complimentary information. *Journal of Media and Religion* 15, 1 (2016), 29–42.

[12] David Card, Christian Dustmann, and Ian Preston. 2005. Understanding attitudes to immigration: The migration and minority module of the first European Social Survey. (2005).

[13] Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. Thyghgapp: Instagram content moderation and lexical variation in Pro-Eating disorder communities. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* 27 (2016), 1201–1213. https://doi.org/10.1145/2818048.2819963

[14] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.

[15] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.

[16] Yuchen Chen, Alex Jiahong Lu, and Angela Xiao Wu. 2023. 'China'as a 'Black Box?'Rethinking methods through a sociotechnical perspective. *Information, Communication & Society* 26, 2 (2023), 253–269.

[17] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

[18] Stuart Chinn. 2016. Trump and Chinese Exclusion: Contemporary parallels with legislative debates over the Chinese Exclusion Act of 1882. *Tenn. L. Rev.* 84 (2016), 681.

[19] Zhendong Chu, Ruiyi Zhang, Tong Yu, Rajiv Jain, Vlad I Morariu, Jiuxiang Gu, and Ani Nenkova. 2023. Improving a Named Entity Recognizer Trained on Noisy Data with a Few Clean Instances. *arXiv preprint arXiv:2310.16790* (2023).

[20] Matthew Costello, James Hawdon, Colin Bernatzky, and Kelly Mendes. 2019. Social group identity and perceptions of online hate. *Sociological inquiry* 89, 3 (2019), 427–452.

[21] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*. 86–95.

[22] Stacy Jo Dixon. 2023. Number of Twitter users in the United States from 2017 to 2022. https://www.statista.com/statistics/232818/active-us-twitter-user-growth/

[23] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[24] Bryan Dosono and Bryan Semaan. 2020. Decolonizing tactics as collective resilience: Identity work of AAPI communities on Reddit. *Proceedings of the ACM on Human-Computer interaction* 4, CSCW1 (2020), 1–20.

[25] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322* (2021).

[26] Adam M Enders, Joseph E Uscinski, Casey Klofstad, and Justin Stoler. 2020. The different forms of COVID-19 misinformation and their consequences. *The Harvard Kennedy School Misinformation Review* (2020).

[27] Renee Garett and Sean D Young. 2022. The role of misinformation and stigma in opioid use disorder treatment uptake. *Substance use & misuse* 57, 8 (2022), 1332–1336.

[28] Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. (1 2021). https://link.springer.com/chapter/10.1007/978-3-030-73696-5_12

[29] Angela R Gover, Shannon B Harper, and Lynn Langton. 2020. Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice* 45 (2020), 647–667.

[30] Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique* 10, 3 (2017), 518–537.

[31] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

[32] Tal Orian Harel, Jessica Katz Jameson, and Ifat Maoz. 2020. The normalization of hatred: Identity, affective polarization, and dehumanization on Facebook in the context of intractable political conflict. *Social Media+ Society* 6, 2 (2020), 2056305120913983.

[33] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 90–94.

[34] Eric Heinze. 2018. Toward a Legal Concept of Hatred: democracy, Ontology, and the Limits of Deconstruction. In *Hate, Politics and Law: Critical Perspectives on Combating Hate*. Oxford University Press.

[35] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[36] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[37] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[38] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[39] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one* 16, 8 (2021), e0256762.

[40] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[41] Kayla Keener. 2017. Alternative facts and fake news: Digital mediation and the affective spread of hate in the era of Trump. *J. Hate Stud.* 14 (2017), 137.

[42] Sang Hea Kil. 2012. Fearing yellow, imagining white: Media analysis of the Chinese exclusion act of 1882. *Social Identities* 18, 6 (2012), 663–677.

[43] Jae Yeon Kim and Aniket Kesari. 2021. Misinformation and hate speech: The case of anti-Asian hate speech during the COVID-19 pandemic. *Journal of Online Trust and Safety* 1, 1 (2021).

[44] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 290–302.

[45] Monika Kopytowska and Fabienne Baider. 2017. From stereotypes and prejudice to verbal and physical violence: Hate speech in context. *Lodz Papers in Pragmatics* 13, 2 (2017), 133–152.

[46] Erika Lee. 2002. The Chinese exclusion example: Race, immigration, and American gatekeeping, 1882-1924. *Journal of American Ethnic History* 21, 3 (2002), 36–62.

[47] Stacy J Lee. 2015. *Unraveling the" model minority" stereotype: Listening to Asian American youth.* Teachers College Press.

[48] Laura Leets and Howard Giles. 1997. Words as weapons—when do they wound? Investigations of harmful speech. *Human Communication Research* 24, 2 (1997), 260–301.

[49] Tim Futing Liao. 1994. *Interpreting probability models: Logit, probit, and other generalized linear models.* Number 101. Sage.

[50] Brooke Fisher Liu, Logen Bartz, and Noreen Duke. 2016. Communicating crisis uncertainty: A review of the knowledge gaps. *Public Relations Review* 42, 3 (9 2016), 479–487. https://doi.org/10.1016/J.PUBREV.2016.03.003

[51] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.* 63–70.

[52] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How Do Users Experience Moderation?: A Systematic Literature Review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–30.

[53] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 369–380.

[54] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.

[55] Christian Meske and Enrico Bunde. 2023. Design principles for user interfaces in AI-Based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers* 25, 2 (2023), 743–773.

[56] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference.* Association for Computational Linguistics, 262–272.

[57] United Nations. [n. d.]. What is hate speech? https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[58] Taylor Nelson, Nicole Kagan, Claire Critchlow, Alan Hillard, and Albert Hsu. 2020. The Danger of Misinformation in the COVID-19 Crisis. *Missouri Medicine* 117, 6 (2020), 510–512. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7721433/

[59] Cailin O'Connor and James Owen Weatherall. 2019. *The misinformation age: How false beliefs spread.* Yale University Press.

[60] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an Infodemic: COVID-19 Fake News Dataset.

[61] Shruti Phadke and Tanushree Mitra. 2020. Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–13.

[62] Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. What makes people join conspiracy communities? role of social factors in conspiracy engagement. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–30.

[63] Nguyen Phuong-Mai, Cees Terlouw, and Albert Pilot. 2005. Cooperative learning vs Confucian heritage culture's collectivism: confrontation to reveal some cultural conflicts and mismatch. *Asia Europe Journal* 3 (2005), 403–419.

[64] Daniel Powers and Yu Xie. 2008. *Statistical methods for categorical data analysis.* Emerald Group Publishing.

[65] Cecil Robinson and Randall E Schumacker. [n. d.]. Interaction effects: centering, variance inflation factor, and interpretation issues. *Multiple linear regression viewpoints* 35, 1 ([n. d.]), 6–11.

[66] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science.* 255–264.

[67] Koustuv Saha, John Torous, Eric D Caine, and Munmun De Choudhury. 2020. Psychosocial effects of the COVID-19 pandemic: large-scale quasi-experimental study on social media. *Journal of medical internet research* 22, 11 (2020), e22600.

[68] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[69] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[70] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[71] Michel Setbon and Jocelyn Raude. 2010. Factors in vaccination intention against the pandemic influenza A/H1N1. *European Journal of Public Health* 20, 5 (10 2010), 490–494. https://doi.org/10.1093/EURPUB/CKQ054

[72] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[73] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[74] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 247–267.

[75] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[76] Turki Turki and Sanjiban Sekhar Roy. 2022. Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer. *Applied Sciences* 12, 13 (2022), 6611.

[77] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (2022), 1–9.

[78] Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur Khudabukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: unifying human and machine disagreement on what is offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 11648–11668.

[79] Frieda Wong and Richard Halgin. 2006. The "model minority": Bane or blessing for Asian Americans? *Journal of Multicultural Counseling and Development* 34, 1 (2006), 38–49.

[80] Emma Woollacott. 2022. Anti-Asian hate speech rocketed 1,662% last year. https://www.forbes.com/sites/emmawoollacott/2021/11/15/anti-asian-hate-speech-rocketed-1662-last-year/?sh=1e0631e959f1

[81] Cary Wu, Yue Qian, and Rima Wilkes. 2021. Anti-Asian discrimination and the Asian-white mental health gap during COVID-19. *Ethnic and Racial Studies* 44, 5 (2021), 819–835.

[82] Kimmy Yam. 2022. Anti-asian hate crimes increased 339 percent nationwide last year, report says. https://www.nbcnews.com/news/asian-america/anti-asian-hate-crimes-increased-339-percent-nationwide-last-year-repo-rcna14282

[83] Xiaodi Yan, Yi Zhu, Syed Ali Hussain, and Mary Bresnahan. 2022. Anti-Asian microaggressions in the time of COVID-19: Impact on coping, stress, and well-being. *Asian American Journal of Psychology* (2022).

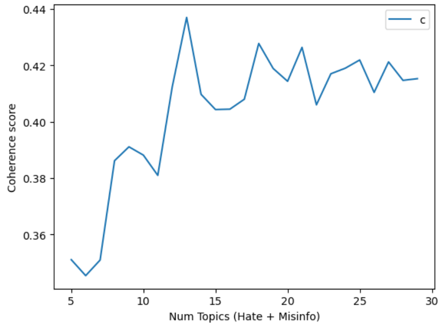
[84] Chenghui Zhang. 2023. Perceiving racial hate crimes: a power-relation perspective. *Journal of experimental criminology* 19, 3 (2023), 663–689.

[85] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581318

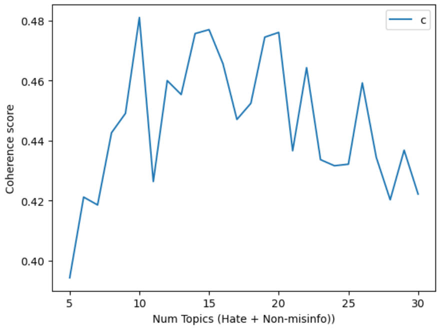
[86] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145* (2023).

[87] Caleb Ziems, Omar Shaikh, Zhehao Zhang, William Held, Jiaao Chen, and Diyi Yang. 2023. Can large language models transform computational social science? *Computational Linguistics* (2023), 1–53.

A COHERENCE SCORES FOR LDA MODELS

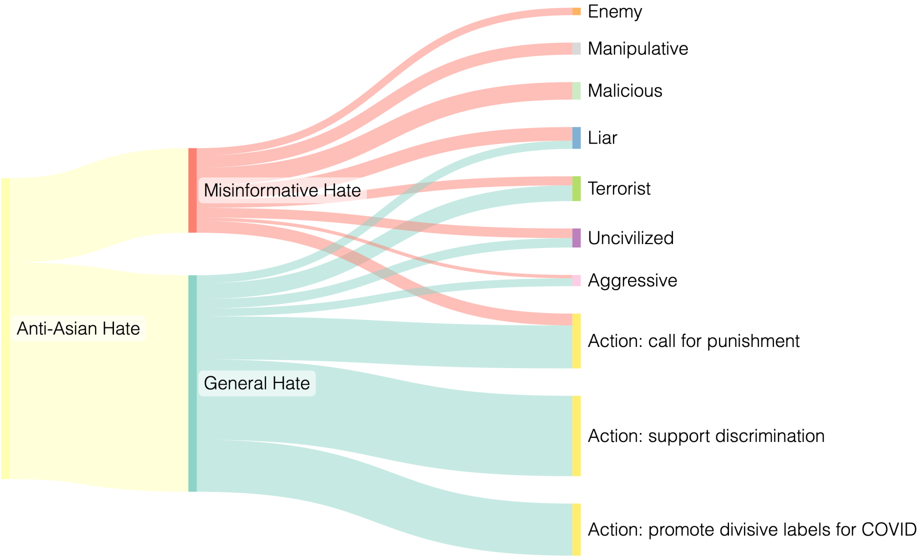


(a) Misinformative Hate (highest coherence score observed at k=13)



(b) General Hate (highest coherence score observed at k=10)

B DISTRIBUTION OF PREVALENT NARRATIVES.



C BASELINES OF SURVEY PARTICIPANTS (N = 308)

Frequency in observing hate speech	Never	15
	Rarely	67
	Occasionally	160
	Frequently	66
Personal encounters with hate speech	Yes	120
	No	176
	Prefer not to say	12
Attitude of hate speech harm	Not harmful	3
	Minimal harmful	10
	Slightly harmful	30
	Moderately harmful	72
	Very harmful	110
	Extremely harmful	83
Attitude of freedom of speech	Strongly Prefer Freedom of Speech	33
	Prefer Freedom of Speech	71
	Slightly Prefer Freedom of Speech	65
	Slightly Prefer Protection from Harm	55
	Prefer Protection from Harm	61
	Strongly Prefer Protection from Harm	23
Attitude of COVID origin	Created or modified by humans	70
	No strong opinion either way	110
	Originated naturally	125
	Prefer not to say	3