Information Processing and Management 62 (2025) 104079

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users

Amira Ghenai [a,*], Zeinab Noorian [a], Hadiseh Moradisani [b], Parya Abadeh [b], Caroline Erentzen [c], Fattane Zarrinkalam [b]

[a] Ted Rogers School of Information Management, Toronto Metropolitan University, Toronto, Canada
[b] School of Engineering, University of Guelph, Guelph, Canada
[c] Department of Psychology, Toronto Metropolitan University, Toronto, Canada

ARTICLE INFO

*Keywords:*
Hate speech
Causal inference
Linguistic analysis
Emotional dynamics
Social media behavior

ABSTRACT

Online hate speech has become a critical issue, particularly during the COVID-19 pandemic, when anti-Asian sentiment surged across social media platforms. However, the causal mechanisms driving emotional and behavioral shifts in users posting hateful content remain understudied. This study investigates the causal relationship between engaging in hateful content and changes in linguistic and emotional expression on social media. Using a dataset of 6,002 Twitter/X users, we employ causal inference techniques, including propensity score matching, and advanced topic modeling to compare users posting hateful content with a matched group of non-hateful users. Our main findings can be summarized as follows: (a) Users who post hateful content show significantly higher levels of anger, anxiety, and negative emotions, along with increased third-person pronoun usage. (b) Moral outrage and profanity levels peak during hateful posts but decline over time, while remaining elevated compared to non-hateful posts. (c) Hateful posts are more interconnected, cover more diverse topics, and are more similar to one another, revealing lower cohesion within individual posts but higher cohesion across posts. These findings contribute to understanding the causal effects of online hate speech on user behavior, offering actionable insights for social media platforms to mitigate the spread of hateful content and its broader societal impact.

## 1. Introduction

In recent years, social media platforms have become central to the ways in which individuals exchange information, shape public discourse, and influence behavior. While platforms like Twitter/X facilitate the rapid dissemination of content, they also enable the spread of harmful behaviors, including hate speech and misinformation, which can have significant societal consequences (Cinelli et al., 2020; Kietzmann, Hermkens, McCarthy, & Silvestre, 2011; Mathew et al., 2019; Stieglitz & Dang-Xuan, 2013; Wang, Zhang, Fan, & Zhao, 2022). This is particularly evident during crises, such as the COVID-19 pandemic, when East Asians were frequently targeted due to the virus's origins in China (Chetty & Alathur, 2018; Gover, Harper, & Langton, 2020; Mathew et al., 2019; Tong & DeAndrea, 2023).

Social media platforms create ideal environments for the dissemination of hateful content by connecting like-minded users and fostering echo chambers, where political beliefs are reinforced and polarized (Boutyline & Willer, 2017; Williams, McMurray, Kurz, & Lambert, 2015). Moreover, social media users often misjudge the diversity of their audience, leading to "context collapse", which

* Corresponding author.
*E-mail address:* aghenai@torontomu.ca (A. Ghenai).

https://doi.org/10.1016/j.ipm.2025.104079
Received 8 October 2024; Received in revised form 24 December 2024; Accepted 19 January 2025
Available online 5 February 2025

further amplifies digital hate, cyber-racism, and extremist perspectives (Bührer, Koban, & Matthes, 2024; Chris Hale, 2012; League, 2021; Marwick & Boyd, 2011). Hate speech is frequently driven by antisocial traits such as sadism and psychopathy (Frischlich, Schatto-Eckrodt, Boberg, & Wintterlin, 2021; Lumsden & Morgan, 2017), with some users motivated by a desire for social approval from their peers (Walther, 2022).

Each social media platform has unique features that influence how hate speech spreads. Insights from Facebook studies (Kalsnes & Ihlebæk, 2021; Leonhard, Rueß, Obermaier, & Reinemann, 2018) or controlled experiments (Álvarez-Benjumea & Winter, 2018; Bautista-Ortuño, Castro-Toledo, Perea-García, & Rodríguez-Gómez, 2018) may not fully apply to Twitter/X, which has a global, diverse audience that amplifies hate speech more rapidly. This can cause greater harm to targeted communities and increase the risk of offline violence (Casula, Anupam, & Parvin, 2021; Chetty & Alathur, 2018; Gallacher, Heerdink, & Hewstone, 2021; Soral, Bilewicz, & Winiewski, 2018; Watanabe, Bouazizi, & Ohtsuki, 2018). These dynamics make Twitter/X a crucial platform for studying hate speech.

While substantial progress has been made in detecting online hate speech using a variety of computational methods, including traditional approaches like TF-IDF (Aziz, Maarof, & Zainal, 2021; Burnap & Williams, 2016; Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016; Ombui, Muchemi, & Wagacha, 2019; Waseem & Hovy, 2016a) and lexicon-based techniques (Basile, 2019; Bauwelinck, Jacobs, Hoste, & Lefever, 2019; Capozzi et al., 2019; Orts, 2019; Perelló, Tomás, Garcia-Garcia, Garcia-Rodriguez, & Camacho-Collados, 2019; Ribeiro & Silva, 2019; Sanguinetti, Poletto, Bosco, Patti, & Stranisci, 2018; Tellez, Moctezuma, Miranda-Jiménez, & Graff, 2018; Vega, Reyes-Magaña, Gómez-Adorno, & Bel-Enguix, 2019), as well as more advanced methods such as deep learning (Arango, Pérez, & Poblete, 2019; Chakraborty et al., 2018; Djuric et al., 2015; Frenda et al., 2020; Pavlopoulos, Sorensen, Androutsopoulos, & Dixon, 2018), critical gaps and limitations persist within the research landscape.

These limitations can be categorized as follows: **(1)** Most studies rely on static detection models that analyze hate speech using keywords or abusive language features (e.g., Davidson, Warmsley, Macy, and Weber (2017) and Mathew, Saha, Yimam et al. (2019)). While these models are effective for identifying isolated instances of online hate, they fail to capture the longitudinal and evolutionary dynamics of hate speech engagement. Specifically, they do not examine how hateful posts evolves over time or the cumulative effects of sustained engagement on user behavior. **(2)** Existing research often overlooks the causal relationship between hate speech and user behavior. Specifically, it does not investigate how hateful content engagement affects linguistic, emotional, and social dynamics over time or how such behavior interacts with broader psychosocial factors. **(3)** Hate speech narratives are rarely studied as cohesive systems. Prior work has not adequately explored the structural interconnectedness and thematic specificity of hateful content, leaving significant gaps in understanding how narratives sustain and propagate harmful ideologies (Lewandowsky, Cook & Lloyd, 2018; Zannettou, ElSherief, Belding, Nilizadeh, & Stringhini, 2020). **(4)** Research often examines online hate speech in isolation, focusing on the content rather than the individuals who post it. This approach overlooks critical distinctions between *hateful users*–defined as individuals who frequently post hate speech–and *non-hateful users*, who engage in non-hateful discourse. This oversight limits insights into the linguistic, emotional, and cognitive traits that distinguish *hateful* users from their counterparts (e.g., Chiril, Pamungkas, Benamara, Moriceau, & Patti, 2022; Jahan & Oussalah, 2023; Watanabe et al., 2018). **(5)** Existing public hate speech datasets, while valuable, are primarily focused on identifying hateful content or analyzing social network structures, without providing comprehensive user timelines or content unrelated to hate speech (Repository, 2020; Ribeiro & Benevenuto, 2022). For example, the dataset by Ribeiro and Benevenuto (2022) contains lists of hateful users and their network connections but does not include the actual text of user timelines, which is essential for linguistic and psycholinguistic analysis.

This study seeks to overcome these limitations by adopting a dynamic and causal approach to understanding hateful posts in social media. Our dataset addresses the existing research gap by collecting the complete timelines (up to 3200 tweets per user) for both hateful and non-hateful users, enabling a longitudinal analysis of user behavior and content beyond hateful posts instances. This unique approach allows us to explore how users' linguistic, emotional, and cognitive behaviors evolve over time, offering insights that extend beyond the static analysis of hateful posts. Further, we employ causal inference techniques and advanced topic modeling to analyze the emotional, linguistic, and thematic characteristics of hateful users compared to a matched group of non-hateful users.

The implications of these limitations are profound. Hate speech contributes to societal harm, including discrimination, polarization, and offline violence (Chetty & Alathur, 2018; Watanabe et al., 2018). Addressing these gaps is essential for designing more effective mitigation strategies that target not only the content of hate speech but also the underlying behaviors and structures that enable its propagation. By integrating a longitudinal perspective, this research provides actionable insights into how hate posts evolves and sustains harmful ideologies. Our findings offer theoretical contributions to the fields of information behavior and hate speech dynamics while providing practical implications for policymakers and social media platforms aiming to mitigate the societal impact of hateful content.

The research questions (RQs) driving this study are as follows:

**RQ1**: *How does posting hateful content on social media influence the linguistic and cognitive characteristics of users compared to those who do not post hateful content?*

**RQ2**: *How do the thematic patterns and specificity (degree of focus and clarity) in hateful posts on social media differ from those in non-hateful posts, and what do these differences reveal about the nature and dynamics of hateful content?*

The remainder of this paper is organized as follows. Section 2 lays the theoretical foundation by examining linguistic markers, moral outrage, and thematic coherence in hate speech. Section 3 outlines our methodology, integrating theoretical perspectives with computational techniques to analyze social media data during the COVID-19 pandemic. Section 4 presents the results, while Section 5 discusses related work. Finally, Section 6 provides a discussion of implications and limitations, and Section 7 concludes the paper.

## 2. Theoretical foundation

### 2.1. Linguistic and emotional markers in information behavior

#### 2.1.1. Emotional and dehumanizing language

Dehumanization occurs when an outgroup is seen as not fully human, being less evolved, less moral, and undeserving of dignity and respect (Haslam, 2006; Kteily & Landry, 2022). It is strongly associated with outgroup prejudice and a strong predictor of bias, hostility, and hate (Kteily & Bruneau, 2017). Dehumanization has also been associated with expressions of anger, aggression, fear, and anxiety towards the outgroup (Giner-Sorolla & Russell, 2019; Haybron, 2002; Matsumoto, Hwang, & Frank, 2016; Sell, Tooby, & Cosmides, 2009). Previous research has demonstrated that dehumanization is a crucial component of hate speech, often conveyed through negative emotions such as anger and fear (Alorainy, Burnap, Liu, Javed, & Williams, 2018; Davidson et al., 2017; ElSherief, Kulkarni, Nguyen, Wang, & Belding, 2018; Martins, Gomes, Almeida, Novais, & Henriques, 2018; Zannettou et al., 2020).

These emotions contribute to dehumanizing attitudes, portraying targeted groups as deserving of hostility and aggression (Giner-Sorolla & Russell, 2019; Sell et al., 2009) or depicting them as some sort of monstrous evil (Giner-Sorolla & Russell, 2019; Haybron, 2002). Those who engage in dehumanization tend to use more negative emotional language and less positive emotional language when discussing an outgroup (Markowitz & Slovic, 2020). Similarly, hate speech has been associated with negative emotions and anger (Mathew, Kumar, Goyal, Mukherjee, et al., 2018). Based on these insights, we hypothesized:

**Hypothesis 1a (*H1a*).** *Hateful users would exhibit higher levels of negative emotions, such as anger, anxiety, and sadness, compared to non-hateful users.*

Power has also been identified in dehumanization processes, wherein those with less power are more likely to be dehumanized by those with more power (Gwinn, Judd, & Park, 2013; Haslam & Loughnan, 2014). Dehumanization may facilitate the maintenance of power imbalances, oppression, violence, and the belief that those with more power are somehow more deserving of systemic benefits (e.g., Jost, Banaji, & Nosek, 2004; Pratto, Sidanius, Stallworth, & Malle, 1994). For example, Markowitz and Slovic (2020) observed that those who engaged in dehumanizing language tended to use more power-related words when describing an illegal immigrant crossing into the United States.

Perceptions of threat and risk have consistently been identified as contributing to intergroup prejudice, particularly where one does not feel they have the ability to control that threat Esses, Medianu, and Lawson (2013), Pavetich and Stathi (2021), Schaller and Neuberg (2012) and Stephan, Diaz-Loving, and Duran (2000). Moreover, during times of crisis and increased threat, such as COVID, there can be a tendency to blame outgroups (Kim, Sherman, & Updegraff, 2016; Tennen & Affleck, 1990). The dehumanization process also enables comfort with the elimination of the despised outgroup, including their death, killing, execution, and calls for violence (Goff, Eberhardt, Williams, & Jackson, 2008; Paasch-Colberg, Strippel, Trebbe, & Emmer, 2021). Generalized hate speech (directed at a group vs. an individual target) was found to include death-related words such as kill, murder, and terminate (ElSherief et al., 2018). Thus, we hypothesized that:

**Hypothesis 1b (*H1b*).** *Hateful users would show a greater frequency of language pertaining to power, risk, and death compared to non-hateful users.*

Hate speech is by definition focused on an outgroup target, and it has been observed to use a greater number of third-person pronouns compared to non-hate speech (e.g., ElSherief et al., 2018; Faulkner & Bliuc, 2018; Zannettou et al., 2020). Third-person pronouns can indicate a sense of detachment, focus on outgroup members, and homogenization of the outgroup as a monolithic and abstract collective identity (Perdue, Dovidio, Gurtman, & Tyler, 1990). Indeed, the use of "them" pronouns is associated with negative evaluations (Perdue et al., 1990), reduced outgroup empathy (Shih, Stotzer, & Gutiérrez, 2013), and divisive political speech (Matos & Miller, 2023). Thus, it was hypothesized that:

**Hypothesis 1c (*H1c*).** *Hateful users would use third-person pronouns more frequently than non-hateful users, emphasizing detachment and the perception of others as abstract entities rather than individuals.*

#### 2.1.2. Profanity and dehumanization

Ethnophaulisms are ethnic slurs used to derogate and dehumanize a racial outgroup, and they are often found in hate speech (Carter, 1944; Leader, Mullen, & Rice, 2009). Slurs and profanity are not limited to ethnic identity and may be directed towards groups based on sexual orientation, religion, gender, age, disability status, and any other social identity one wishes to disparage (e.g., Bartlett, Reffin, Rumball, & Williamson, 2014; Bilewicz & Soral, 2020). Such language and profanity are typically intended to be offensive and condescending (Jeshion, 2013; Thurlow, 2001) and meant to convey negative stereotypes and connotations (Anderson & Lepore, 2013; Vallée, 2014). Where accompanied by aggression, it can be a particularly potent amplifier of hate speech. Zahrah, Nurse, and Goldsmith (2022) studied online discussions about COVID-19 on four social media platforms and noted that the subset that targeted China and East Asian persons was associated with a higher level of profanity and slurs compared to other topics, such as mask-wearing. The presence of profanity in a tweet does not mean that it is hateful (Davidson et al., 2017), although hate speech is generally associated with greater use of profanity (Malmasi & Zampieri, 2017; Mathew et al., 2018). Thus, it was hypothesized that:

**Hypothesis 1d (*H1d*).** *Hateful users would exhibit a higher use of profanity compared to non-hateful users.*

*2.1.3. Moral outrage and the spread of hate speech*

Moral outrage language is typically seen where there is a perceived violation of one's moral code, and typically there is some person, group, or target that is identified for punishment (Brady, McLoughlin, Doan, & Crockett, 2021; Crockett, 2017; Salerno & Peter-Hagene, 2013). In many cases, this may be motivated by a desire to obtain social approval and status via "moral grandstanding" or "virtue signaling" (Grubbs, Warmke, Tosi, James, & Campbell, 2019). Moral outrage has been associated with political polarization and the spread of misinformation (Young & Young, 2020) Social media that uses moral outrage language is more likely to be shared (Brady & Crockett, 2024; Brady, Wills, Jost, Tucker, & Van Bavel, 2017), with a 20% increase in transmission for each moral-emotional word (Brady et al., 2017). Moreover, this increased interaction with one's posts can increase subsequent moral outrage language, suggestive of a reinforcement effect via emotional contagion (Brady, Crockett, & Van Bavel, 2020; Brady et al., 2021). Faulkner and Bliuc (2018) found that online racist speech was more likely than non-racist speech to use morality-based language related to purity, authority, and religion. Similarly, morality-based language occurs more frequently in hateful content compared to non-hateful content (Solovev & Pröllochs, 2023). As a result, it was hypothesized that:

**Hypothesis 1e (*H1e*).** *Hateful users would demonstrate higher levels of moral outrage in their language compared to non-hateful users.*

*2.2. Thematic coherence and complexity in hate speech narratives*

*2.2.1. Conceptual coherence*

Expanding on the emotional and dehumanizing language present in hate speech, it is vital to understand the specific topics and how they interrelate within social media discourse. There is evidence that hate speech shows both more heterogeneous topics that are connected in a less cohesive way (Papcunová et al., 2023). That is, hate speech narratives tend to exhibit lower specificity and less cohesion as they connect various unrelated ideas to justify discrimination and perpetuate a hateful world view. Those who engage in hate speech may have a particular narrow world view, endorsing more populist beliefs and outgroup hatred (Papcunová et al., 2023; Salmela & Von Scheve, 2017).

Interestingly, conspiracy theories often intertwine with online hate speech. As noted by Wood, Douglas, and Sutton (2012), the conspiracist belief system is driven by higher-order beliefs about the world rather than by the consistency of the individual belief systems. Belief in conspiracy theories is associated with the tendency to detect threats, assign agency, and see patterns in random things (Van Prooijen & Van Vugt, 2018). Those who believe in one conspiracy theory tend to believe in other conspiracy theories even if they are contradictory (Goertzel, 1994; Swami, Chamorro-Premuzic, & Furnham, 2010; Wood et al., 2012), and some conspiracy theories may be relied on to explain the lack of support/evidence for another theory (Boudry & Braeckman, 2012). As a result, conspiracy theorists often make connections between heterogeneous topics (Goertzel, 1994; Lewandowsky, Cook & Lloyd, 2018).

Conspiracy texts typically have more tightly interconnected topics but are less coherent within text, suggesting a tendency to connect scattered topics together but to explore them in relatively shallow depth (Lewandowsky, Cook & Lloyd, 2018; Miani, Hills, & Bangerter, 2022; Wood et al., 2012). Interestingly, hateful users and conspiracy theorists share similar characteristics, including low agreeableness, a sense of persecution to the self/ingroup, and poor cognitive reasoning skills (Douglas, Sutton, & Cichocka, 2017; Markowitz et al., 2021; Swami, Voracek, Stieger, Tran, & Furnham, 2014) and a tendency to hold populist views (van Prooijen et al., 2022). Moreover, many conspiracy theories are racist or hateful in origin (Pollard, 2016; Swami, Barron, Weis, & Furnham, 2018), including those pertaining to COVID-19 (Baider, 2022; Douglas, 2021). Indeed, Markowitz et al. (2021) found that those who believed in COVID-19 conspiracy theories engaged in more dehumanization towards East Asians. We thus proposed the following hypotheses:

**Hypothesis 2a (*H2a*).** *Hateful posts will exhibit a more tightly connected network of related topics compared to non-hateful posts, as the hateful worldview forces unrelated ideas to support the belief that certain groups are inferior or dangerous.*

**Hypothesis 2b (*H2b*).** *As a consequence of the tightly interconnected topic structure described in Hypothesis 2a, hateful posts will exhibit higher global cohesion compared to non-hateful posts, reflecting the tendency of hateful posts to revolve around interconnected and thematically unified topics.*

*2.2.2. Integrative complexity*

In addition to the interconnectedness of topics, one may consider how in-depth any one topic is discussed. Rather than looking at what is discussed in a text, integrative complexity looks at the way it is discussed (Suedfeld & Tetlock, 1977). It considers both the differentiation of a topic into dimensions/issues as well as the integration of those dimensions (Suedfeld, Tetlock, & Streufert, 1992). Integratively simple statements incorporate reductive black-and-white thinking and refusal to entertain opposing points of view, whereas integratively complex thinking recognizes the multiplicity of positions and dimensions to an issue and integrates them with superordinate connecting principles (Baker-Brown et al., 1992; Suedfeld et al., 1992). Integrative complexity is lower in online speech directed at like-minded audiences (Jakob, Dobbrick, & Wessler, 2023), racist online speech (Faulkner & Bliuc, 2018; Gregory & Piff, 2021), and violent extremist rhetoric (Fearon & Boyd-MacMillan, 2016). Communications that show a reduction in integrative complexity are more likely to result in conflict and violence; increases in integrative complexity lead to cooperation and peace (Suedfeld, Leighton, & Conway III, 2006). This suggests that racist and hate speech may be more simplistic in nature, lacking nuance or integration of ideas. This is in line with general findings that cognitive reasoning skills are lower among those who endorse racist beliefs (e.g., Dhont & Hodson, 2014; Hodson & Busseri, 2012). Thus, it was hypothesized that:

**Hypothesis 2c (*H2c*).** *Hateful posts will display lower specificity, meaning they are less focused on addressing individual topics in detail. Instead, such posts tend to rely on broad, generalized statements that lack nuanced or integratively complex reasoning and language*

## 3. Methodology

In this section, we first describe the dataset utilized in our research, including its source and characteristics. We then outline the methodology employed to test our hypothesis, detailing the our analytical approach.

### 3.1. Data collection

For this study, we leverage Twitter/X timeline data of individuals who posted hateful content, focusing on anti-Asian hate as our case study. We begin with the dataset curated by Noorian, Ghenai, Moradisani, Zarrinkalam, and Alavijeh (2024), which comprises 3001 hateful users and 3001 non-hateful users,[1] with a total of 5,417,041 tweets in their timelines prior to posting content about the topic related to anti-Asian hate and neutral discussions during January 15, 2020 and April 17, 2020. Their curated dataset went through a rigorous filtering process to ensure only genuine and non-automated accounts were selected. Additionally, to ensure true labeling for users in the hateful group or non-hateful group, they filtered out users with fewer than three tweets classified as hateful or non-hateful related to the anti-Asian community.[2]

The focus of this study is to analyze the impact of sharing hateful content on the linguistic and behavioral characteristics of hateful users after posting hateful content. To do this, we extend the original dataset by collecting the timelines of hateful users for up to 120 days after their initial post of the hateful content, and similarly for users in the non-hateful group. Following prior studies that analyze social media behavior over short-to-medium observation periods (Brady et al., 2017; Kramer, Guillory, & Hancock, 2014), the decision to focus on a 120-day timeline was guided by both practical and empirical considerations. The Twitter/X API imposes a strict limit of 3200 tweets per user, which naturally constrains the total historical data that can be collected. Given the observed tweet rate in our dataset – approximately 5 tweets per day on average – 120 days allows for the collection of around 600 tweets per user to enrich the originally collected dataset (Noorian et al., 2024). Extending the timeline beyond 120 days would disproportionately exclude users with lower tweet volumes or inconsistent activity patterns, introducing bias and reducing the overall sample size. This process resulted in a final dataset consisting of a group of 3001 hateful users who posted 1,590,331 tweets in their timelines, and a corresponding group of 3001 non-hateful users with a total of 1,857,907 tweets collected for a period of January 16, 2020 and September 14, 2020.

### 3.2. Propensity score analysis

#### 3.2.1. Design and rationale

To effectively investigate the impact of hateful posts on the linguistic and cognitive characteristics of individuals who publish such content, we employ a robust causal inference framework grounded in matching techniques. This method is well-supported by literature in statistics and computational social media studies and has been successfully applied in previous quantitative social media studies (Kiciman, Counts, & Gasser, 2018; Saha et al., 2019; Verma, Bhardwaj, Aledavood, De Choudhury, & Kumar, 2022). Specifically, the approach simulates a Randomized Controlled Trial (RCT) setting by controlling for as many covariates as possible (Imbens & Rubin, 2015). Our methodology is grounded in the potential outcome framework (Imbens & Rubin, 2015; Rubin, 2005), where we investigate whether an outcome is caused by a treatment. In our study, the outcome is the set of psycholinguistic features reflected in the Twitter/X posts of users, and the treatment is the act of posting hateful content on Twitter/X.

Formally, we compare two potential outcomes: $Y_i(T = 1)$, representing the outcome when users are exposed to the treatment (posting hateful content), and $Y_i(T = 0)$ representing the outcome when they are not exposed. Since it is impossible to observe both outcomes for the same individual, we utilize a potential outcome framework. This framework estimates the counterfactual outcome for an individual by comparing their observed outcomes to those of matched individuals from the control group, ensuring similar baseline covariate distributions. To achieve this, we employ the stratified propensity score matching (PSM) method (Olteanu, Varol, & Kiciman, 2017) to match users in *Treatment* groups (hateful users posting hateful content) with those in the *Control* group (non-hateful users generally talking about the topic and posting non-hateful content) based on several behavioral attributes such as Twitter/X interaction and Linguistic cues.

The choice to use stratified PSM was driven by its robustness and suitability for datasets characterized by high variability in user behavior, such as those found on social media platforms. Stratified PSM divides users into strata based on their propensity scores, ensuring that treatment and control users within each stratum are highly comparable in terms of their baseline behavioral attributes. This approach minimizes bias by creating balance across key covariates such as linguistic cues and interaction patterns on Twitter/X. Compared to nearest-neighbor matching (Cui, Marder, Click, Hoekstra, & Bruce, 2022), which often discards unmatched users and reduces statistical power, stratified PSM retains more data while still achieving a high degree of balance within each stratum. Another key advantage of stratified PSM is its ability to avoid the challenges associated with inverse probability weighting

[1] The hateful posts' labels were annotated in He et al. (2021). These labels were generated through a rigorous annotation process involving annotators of Indian and Chinese backgrounds who were trained and supervised to detect hateful posts accurately while accounting for linguistic and cultural nuances.

[2] Please refer to He et al. (2021)'s work for the full list of keywords used for the data collection.

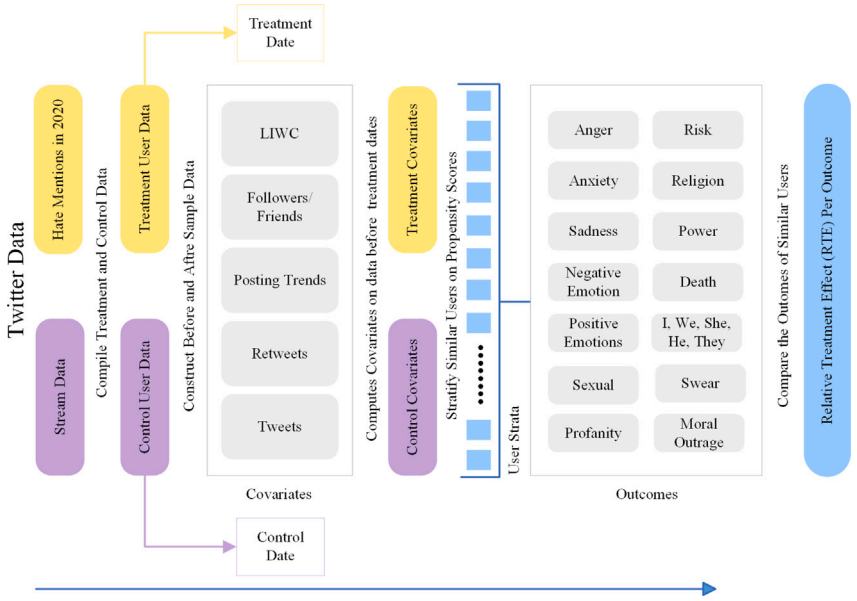


**Fig. 1.** The propensity score analysis process.

(IPW) (Bettega, Mendelson, Leyrat, & Bailly, 2024; Emine, Elif, Çelik, & Kadir, 2024). In IPW, extreme propensity scores can lead to highly unstable weights, which may distort the analysis unless additional stabilization techniques are applied. Stratified PSM, by contrast, does not rely on weighting and is therefore less prone to such instability. Moreover, stratified PSM allows for subgroup analyses within individual strata, providing deeper insights into how treatment effects may vary across different levels of baseline characteristics. This granular approach aligns well with the objectives of our study, which seeks to understand various psycholinguistic and behavioral differences between users posting hateful content compared with those who do not (Caliendo & Kopeinig, 2008; Ou, Zhao, Zuo, & Wu, 2024)

To estimate counterfactual outcomes, we leverage the matched control group within each stratum. This approach assumes that, within a given stratum, the outcomes observed in the control group closely approximate what would have occurred for the treatment group in the absence of the treatment. By ensuring that treatment and control users in each stratum are comparable, this method provides a robust framework for causal inference. Fig. 1 illustrates an overview of the proposed PSM process.

### *3.2.2. Constructing the before and after samples*

As we aim to measure the changes following the posting of hateful content, we divide our dataset into before and after samples relative to the treatment date (the date of posting hateful content). For each treatment user, the date of their first hate tweet is designated as their treatment date. In the control group, we assign a *placebo date* that matches the distribution of hate dates, with a 5-day interval to account for temporal confounding effects. This step ensured that observed differences in outcomes were not influenced by time-dependent factors. Consequently, we split our treatment and control groups into *Before* and *After* samples.

### *3.2.3. Matching for causal inference*

**Matching Covariates:** The covariates used in this study are carefully selected to capture the critical psycholinguistic, behavioral, and social dimensions of user activity on social media. These covariates fall into three primary categories: Linguistic Inquiry and Word Count (LIWC) features, user activity metrics, and network metrics, chosen for their demonstrated effectiveness in prior research (Kim, Razi, Alsoubai, Wisniewski, & De Choudhury, 2024; Saha et al., 2019). Together, these covariates provide a comprehensive representation of user behavior and reduce the potential for confounding in causal inference.

We utilize the LIWC classification framework (Pennebaker, Boyd, Jordan, & Blackburn, 2015), which has been widely validated in the literature for analyzing psychological and linguistic dimensions. From the 72 LIWC categories, we refine our selection to 11 high-level categories based on their relevance to capturing the emotional and linguistic nuances of user behavior. These LIWC features include: *affective processes*, which capture emotional expressions such as joy, anger, and sadness, *cognitive processes*, reflecting words related to causation, insight, and tentativeness, *biological processes*, encompassing references to health, body, and sensory experiences, *drives*, including words signaling power, affiliation, and reward, *time orientation*, capturing temporal focus through references to the past, present, or future, *psychological processes*, such as language reflecting anxiety, sadness, or self-awareness, and *social processes*, include pronoun usage and references to interpersonal interactions. The second set of covariates is **user activity metrics**, such as tweet frequency, retweet activity, and temporal posting patterns (posting interval in seconds between two consecutive tweets), which capture users' engagement levels and behavioral trends. The third set of covariates are **network metrics**, such as the number

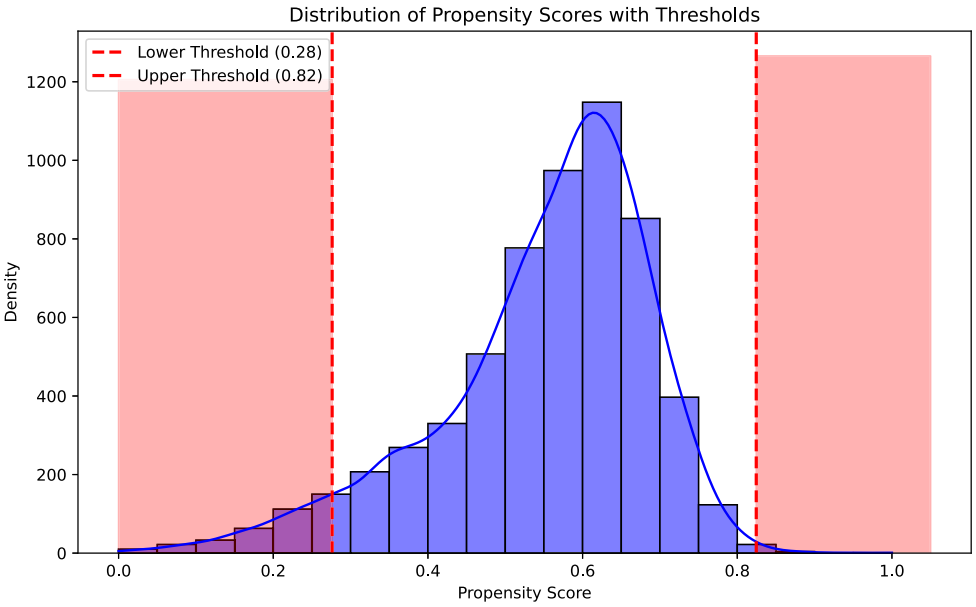


**Fig. 2.** Propensity score distribution (shaded region represents those dropped in our analysis).

of followers and friends which provide insights into users' connectivity and social interaction. By combining these covariates, we ensure a robust and multidimensional approach to matching, enabling meaningful comparisons and reducing potential biases in our analysis.

**Propensity Score Analysis:** To determine the propensity scores, we develop a logistic regression model predicting a user's likelihood of sharing hateful content based on their characteristics. We eliminate outliers with propensity scores outside 2 standard deviations from the mean. The remaining scores are divided into 10 equal-width strata. To ensure a robust causal analysis, we exclude strata with insufficient numbers of treatment or control users, as recommended in causal inference research (De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016). By setting a minimum of 50 users per group per stratum, we establish 10 strata, encompassing 2966 treatment and 2719 control users (Fig. 2).

**Quality of Matching:** To evaluate the quality of matching between treatment and control groups after propensity score matching, we employed the Standardized Mean Difference (SMD) as the primary metric. SMD is a widely accepted and robust measure for assessing covariate balance in causal inference studies (Saha et al., 2019; Verma et al., 2022). It quantifies the difference in means of a covariate between treatment and control groups, scaled by their pooled standard deviation. This property ensures that the assessment remains consistent regardless of variations in group sizes or distributions. In this study, we adhere to the established threshold of SMD values below 0.25, as recommended in the literature (Saha et al., 2019; Verma et al., 2022). This threshold ensures that covariates are sufficiently balanced across treatment and control groups, thereby reducing potential bias. Following propensity score matching, we calculate the SMD for all covariates included in the analysis, such as linguistic features, activity metrics, and network characteristics. The results indicate that all covariates achieved SMD values below the threshold, confirming a high degree of balance between the groups.

**Temporal Variations of Covariates Over Time:** To ensure clarity and transparency regarding the covariates used in our propensity score matching (PSM) process, we examine their temporal variations over the observed timeline for both hate and control groups. While these covariates were primarily used to match the two groups before the reference point (posting hateful content), analyzing their trends across the timeline offers additional insights into behavioral dynamics leading up to and following the reference point.

For instance, for *psychological processes*, users who post hateful content show a decline from an average of 3.3 before the reference point to approximately 3.1 after, while control users exhibit a smaller decrease from about 2.8 to 2.7. In *linguistic dimensions*, the hateful group demonstrates a sharp drop from 11.8 to 10.6 after the reference point, whereas the control group increases slightly from 11.1 to 11.3 before gradually declining to 11.0. Finally, *informal language* shows a noticeable increase in both groups post-reference, with the hateful group rising from 1.0 to 1.2, compared to a smaller increase in the control group from 0.9 to 1.0. Fig. B.7 in Appendix B shows the visualizations the temporal changes in selected covariates over all the studied period.

We further examine the temporal trends of retweets, followers, and friends between the hate and control groups. The results reveal a distinct decline in retweet activity for the hateful group after the reference point, suggesting a reduction in engagement following the posting of hateful content, whereas the control group exhibits relative stability in this measure. By contrast, the number of followers and friends remains largely stable across both groups, with minimal variations observed over the timeline. Fig. B.8 in Appendix B shows the visualizations of these measures over all the time period studied. This analysis shows that the covariates,

though stable enough for matching, exhibit meaningful shifts over time that align with the broader behavioral patterns analyzed in this study.

To further assess the consistency of these trends, we analyze the temporal variations of covariates within the largest strata (stratum 5 and stratum 6). The results confirm that the patterns observed at the aggregate level – such as declines in psychological processes, and linguistic dimensions post-reference – are consistent within these subgroups. This reinforces that the observed behavioral shifts in the hateful group are systematic and not driven by a specific subgroup. Visualizations for these strata are provided in Fig. C.9 and Fig. C.10 in Appendix C.

#### 3.2.4. Defining and measuring outcomes

We analyze the impact of hateful content on the linguistic and cognitive traits of the individuals who create and share it. Following is the list of the outcome measures to address the hypothesis **H1a**, **H1b**, **H1c**, **H1d**, and **H1e**:

*Dehumanization*: To measure signs of dehumanization, we follow methodologies for affective and cognitive outcomes commonly used in social media studies (Ernala, Rizvi, Birnbaum, Kane, & De Choudhury, 2017; Saha, Weber, & De Choudhury, 2018). Specifically, we quantify psycholinguistic shifts in effect and cognition by examining changes in the normalized occurrences of words using the well-validated LIWC lexicon (Tausczik & Pennebaker, 2010). The LIWC categories we use to measure dehumanization include emotions – such as *anger*, *anxiety*, *negative emotions*, *positive emotions*, and *sadness* – and risk-related terms, including *death*, *risk*, and *power*.

*Pronouns*: To assess the focus on outgroup members and distant individuals through pronoun usage, we follow methodologies used in studies that analyze linguistic patterns with the LIWC lexicon (Boyd & Pennebaker, 2017; Tausczik & Pennebaker, 2010). Specifically, we measure changes in pronoun usage by analyzing variations in the normalized frequency of words. The LIWC categories we use to measure pronoun usage are personal pronouns—such as *I*, *we*, *she/he*, and *they*.

*Profanity*: To measure signs of profanity in tweets, we utilize the LIWC lexicon to quantify the normalized occurrences of words in the categories of *sexual* and *swear*. Additionally, we use the profanity dictionary from the Surge AI (2023), which contains over 1600 popular English profanities and their variations. We categorize the occurrences of each specific profane word into the following categories based on Teh and Cheng (2020)'s work: Behavior (conduct towards others), Disability (attacks on a person's disability), Gender (profane words referring to gender or body parts), Physical (attacks on physical appearance), Religion (profane words related to religion), Sexual orientation (attacks on sexual identity), Social class (discrimination based on social or economic status), and Others (profane words not classified in any of the above categories). Additionally, we redefine the Ethnicity category (attacks on cultural or national social groups) to include words related to the COVID-19 pandemic, specifically 'chinavirus' and 'china virus,' expanding its scope to reflect the evolving language of hate during this period.

*Moral Outrage*: Inspired by existing literature on moral outrage (e.g., Brady et al., 2017; Faulkner & Bliuc, 2016), we sought to measure signs of moral outrage in individuals spreading hateful content. Twitter/X is particularly appropriate for measuring signs of moral outrage due to the occurrence of regular high-profile, rapid swells of outrage on this platform (Ronson, 2016). To quantify moral outrage, we utilize the moral outrage classifier developed by Brady et al. (2021). This supervised machine learning model, trained on 26,000 annotated tweets from episodes of public outrage, detects expressions of moral indignation, and condemnation, and calls for justice or punishment.

**Measuring changes in the outcomes:** To assess the impact of online hateful content manifested in **H1a**, **H1b**, **H1c**, **H1d**, and **H1e**, we utilize the Relative Treatment Effect (RTE), a method widely used in prior studies similar to ours (Kiciman et al., 2018; Saha et al., 2019). The RTE quantifies the ratio of treatment effects between the treatment and control groups for a given variable. It provides an intuitive measure of how the treatment impacts the variable under study relative to the control group.

We first calculate the RTE in each stratum for each outcome by comparing the average outcomes between the treatment and control groups (Kiciman et al., 2018; Saha et al., 2019). We then determine the mean RTE for each outcome using a weighted average across the strata. An RTE value greater than 1 ($RTE > 1$) signifies that the treatment group exhibits a higher relative effect on the variable compared to the control group, while an RTE value less than 1 ($RTE < 1$) indicates that the treatment group exhibits a lower relative effect. The interpretation of RTE is linear, meaning that changes in RTE correspond directly to proportional changes in the observed effect.

To address **H2**, we apply the BERTopic method to extract topics from the tweets by (1) representing tweets as semantic embedding vectors, (2) reducing dimensionality, (3) and clustering into topics (Grootendorst, 2022). In our implementation, we use the all-MiniLM-L6-v2[3] sentence embedding model to create text embeddings. This model was chosen for its proven ability to generate high-quality sentence embeddings while requiring relatively low computational resources, as demonstrated in previous studies (Kloo, Cruickshank, & Carley, 2024; Yin & Zhang, 2024). These attributes are particularly advantageous for analyzing large-scale social media datasets, such as our collection of tweets, which contain short and informal texts. Additionally, the lightweight architecture of all-MiniLM-L6-v2 enables faster processing, making it an ideal choice for applications where efficiency is a priority. The authors tested the all-MiniLM-L12-v2[4] model as well which produced similar results. all-MiniLM-L6-v2 was faster than the other model on our hardware, but future work should evaluate several models to balance performance (i.e., consistency with other models) and speed.

For dimensionality reduction, we use UMAP as suggested by the BERTopic documentation (Grootendorst, 2022). We then apply the HDBSCAN clustering algorithm, which does not require hyperparameter tuning. Once the topic clusters are determined, we

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

perform the labeling process by following the methodology detailed by Kloo et al. (2024). To effectively label the topics, we utilize the GPT-4 model (Ye et al., 2023). We apply GPT-4 with the top 10 most representative documents for each topic and request a concise description of each topic in 5 words or fewer. The use of GPT-4 provides human-readable and contextually rich labels that align with the underlying themes of the topics, offering a significant advantage over traditional methods like TF-IDF. To validate the generated labels, given the manageable number of topics in our dataset, two authors conducted a manual human evaluation process. They assessed the clarity, accuracy, and thematic alignment of the labels with the associated tweets. This process ensures that each label meaningfully represent the topics and is usable for subsequent analyses. No refinements were necessary, as the labels were consistently accurate and interpretable. Inspired by the work of Miani et al. (2022), we measure the following:

*Topic Interconnectedness*: Interconnectedness refers to how tweets are connected to each other through topics. We assess interconnectedness by examining the networks resulting from the co-occurrences of topics. Specifically, we measure edge connectivity, which is defined as the number of nodes interconnected to each other. We measure the interconnectedness for both hateful and non-hateful posts to identify and compare the differences between these groups. To compute interconnectedness, we create network objects from the topic matrices generated by BERTopic. We convert these matrices into graph objects using the NetworkX Python package,[5] where nodes represent topics and edges represent their co-occurrences. We compute the edge connectivity, which is the number of edges associated with each node.

For both hateful and non-hateful topic networks, we begin by computing the between-topic correlation matrix and extracting the Pearson correlation coefficient (r) for each topic pair. To convert these correlation matrices into co-occurrence matrices, we apply a threshold to the correlation values. Without a threshold, all topics would co-occur, resulting in maximum connectivity, while a high threshold would result in no topic co-occurrence. To determine an appropriate threshold, we explore different correlation values and select a threshold that provides a meaningful degree of connectivity where we aim to balance retaining significant relationships between topics while excluding weaker, less informative connections. Specifically, following prior studies (Miani et al., 2022), we calculate the mean of the absolute correlation coefficients $|r|$ for both the hateful and control groups. Then, we set the threshold at the overall mean value of $|r| = 0.01025$, representing a balance point where topics exhibit meaningful co-occurrence without overloading the network with noise or redundant connections. This choice reflects both the statistical properties of the data and practical considerations for analyzing topic interconnectivity within the hateful and control groups. For further context, see the connectivity results and visualizations in Appendix A, which demonstrate the effects of varying the threshold and provide empirical support for our selection.

Further, we measure (1) entropy using the entropy function from the Python package Scipy (Bressert, 2012), which indicates the extent to which nodes in a network are interconnected in a random, nonsystematic way; (2) clustering coefficient using the average-clustering function from the NetworkX package (Hagberg, Swart, & Schult, 2008), which calculates the average clustering coefficient for all nodes in the graph, providing a measure of the overall tendency of nodes to form tightly-knit groups; (3) distance using the average-shortest-path-length function from the NetworkX package, which extracts the average shortest path length through nodes; and (4) density using the edge-density function from the NetworkX package, which computes the ratio of the number of edges to the number of possible edges.

*Global Cohesion*: Global cohesion measurement evaluates the similarity between tweets in either treatment or control groups. To achieve this, we represent each tweet in these groups using TF-IDF vectors. We then calculate the lexical overlap between documents within each group by computing cosine similarity (CS) scores. The CS output for each tweet is a vector indicating its similarity to other tweets. We average this vector to obtain a single value for each tweet.

*Topic Specificity*: Topic specificity refers to how focused or dispersed the content of a tweet is across various topics. To measure topic specificity in tweets, we utilize the probability score vector that indicates the likelihood of each tweet belonging to different topics. This specificity is quantified by assessing the extent to which each tweet contained a varying number of topics.

To evaluate this metric, we calculate the inequality of topic distribution within each tweet using the Gini coefficient, an established measure of inequality or dispersion (Cowell, 2011). The Gini coefficient ranges from 0 to 1, where a higher value signifies greater inequality among topic probabilities, indicating that a document is well represented by fewer, dominant topics. Conversely, a lower Gini coefficient suggests a more equal distribution across multiple topics, signifying that the tweet is represented by a broader range of topics. Hence, tweets with higher Gini coefficients are predominantly characterized by a single topic, reflecting more focused content, while those with lower coefficients are more evenly distributed across various topics, indicating more diverse content. Here's the Gini coefficient formula:

$$\text{Gini} = \frac{2\sum_{i=1}^{n} i p_i}{n\sum_{i=1}^{n} p_i} - \frac{n+1}{n} \tag{1}$$

where $P_i$ represents the sorted values of the probabilities and $n$ is the number of topics.

**Statistical analyses:** To statistically test **H1a**, **H1b**, **H1c**, **H1d**, and **H1e**, we run an independent sample t-test to examine the difference in RTE across all stratums. As a measure of effect size for t-tests, we use Cohen's d (Cohen, 2013) to examine changes in outcomes between treatment and control users, per outcome, and strata.

To test **H2a**, we first extract the degree of interconnectedness by counting the number of edges for each node in the network. To statistically test whether interconnectedness is higher in hateful compared to non-hateful networks, we run linear mixed-effects models using the Statsmodels packages (Seabold & Perktold, 2010). In each model, we predict the number of edges by the subcorpus

[5] https://networkx.org/.

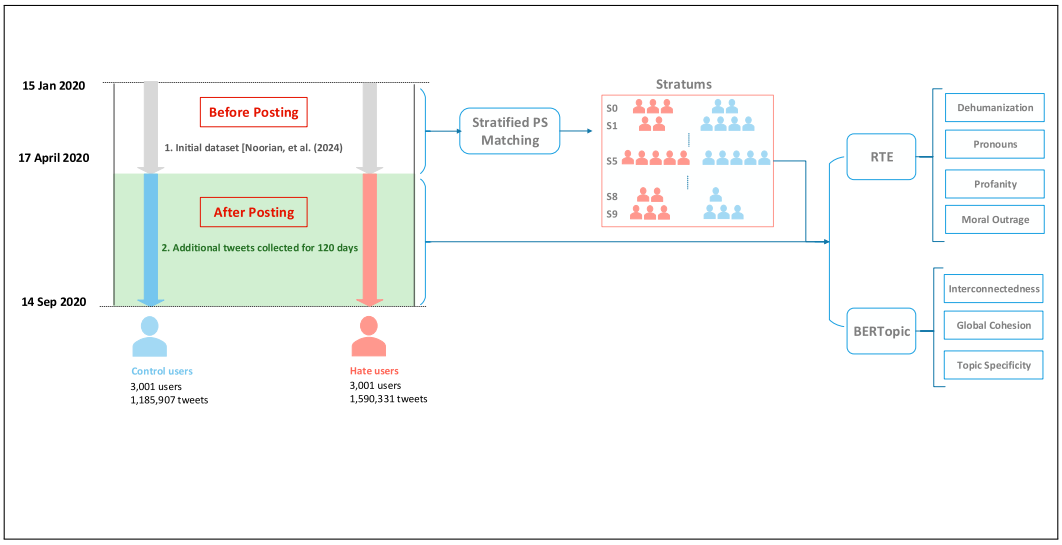


**Fig. 3.** Overview of the methodological framework, highlighting key steps including data collection and propensity score matching, as well as the use of RTE analysis and topic modeling to compute and evaluate the study's outcomes.

(i.e., hate and control tweet), clustering observations within nodes. The network-type is the independent variable, and node-id is a random effect to account for clustering within nodes.

To test **H2b** and **H2c**, we run a series of linear mixed-effects models for each dependent variable (topic specificity and cosine similarity) using the Statsmodels packages. In each model, we specify whether the tweet was hateful as a fixed effect and include tweet word count as a covariate. We also specify the user ID as a random intercept. Following the work suggested by Brady et al. (2021), we include word count as a covariate in our analyses. Finally, we report the standardized regression coefficients beta ($\beta$) for predictors of interest and measures of fit such as $R^2$. Specifically, we report both marginal $R^2(R^2m)$, which is associated with the variance explained by the fixed effects, and conditional $R^2$ ($R^2c$), which is associated with the variance explained by the entire model, including both fixed and random effects.

Fig. 3 provides an overview of the methodological steps detailed above, summarizing the process from dataset preparation to outcome analysis. It outlines the initial tweet collection, stratified propensity score matching, BERTopic modeling, and the evaluation of key metrics such as RTE for dehumanization, profanity, pronoun usage, and moral outrage, as well as interconnectedness, cohesion, and specificity of topics.

## 4. Results

### *4.1. RQ1. Observations about linguistic and cognitive outcomes*

In the following subsections, we report the results of our observations on the linguistic and cognitive outcomes manifested in *RQ1*, averaged across all users in different stratums.

#### *4.1.1. Emotions and dehumanization outcomes*

The analysis of RTE values for emotions reveals significant findings (Fig. 4(a)). Consistent with **H1a**, the RTE for anger consistently exceeds 1, with values around 1.73 to 1.75, indicating higher anger levels in hateful users. Anxiety shows an RTE value of approximately 1.18, suggesting moderately higher anxiety levels. Negative emotions also have an RTE above 1, between 1.34 and 1.35, indicating elevated negative emotions. Conversely, the RTE for positive emotions is below 1, showing lower positive emotions in hateful users. Sadness has an RTE slightly above 1, ranging from 1.05 to 1.12, indicating marginally higher sadness levels. Looking at the significance of these observations, most outcomes show consistent Cohen's d values exceeding 0.2, indicating substantial effects, while anxiety and sadness occasionally fall below the threshold. Further, our longitudinal analysis does not reveal significant changes in emotional levels over time. The RTE values for emotions such as anger, anxiety, negative emotions, positive emotions, and sadness remain relatively stable throughout the observed period.

The analysis of RTE values for other dehumanizing language categories reveals significant differences between hateful and control users (Fig. 4(b)). Consistent with **H1b**, the RTE for the category of 'death' consistently exceeds 1, starting at 1.35 and slightly decreasing to around 1.18 (after 60 days), indicating higher references to death among hateful users. Similarly, the RTE for 'risk' remains above 1, ranging from 1.14 to 1.17, suggesting elevated mentions of risk-related terms in hateful users. The 'power' category also exhibits increased values, starting at 1.10 and gradually increasing to 1.13 by day 60, reflecting a higher prevalence of power-related language in hateful users. A manual examination of a sample of these mentions reveals themes of blame towards

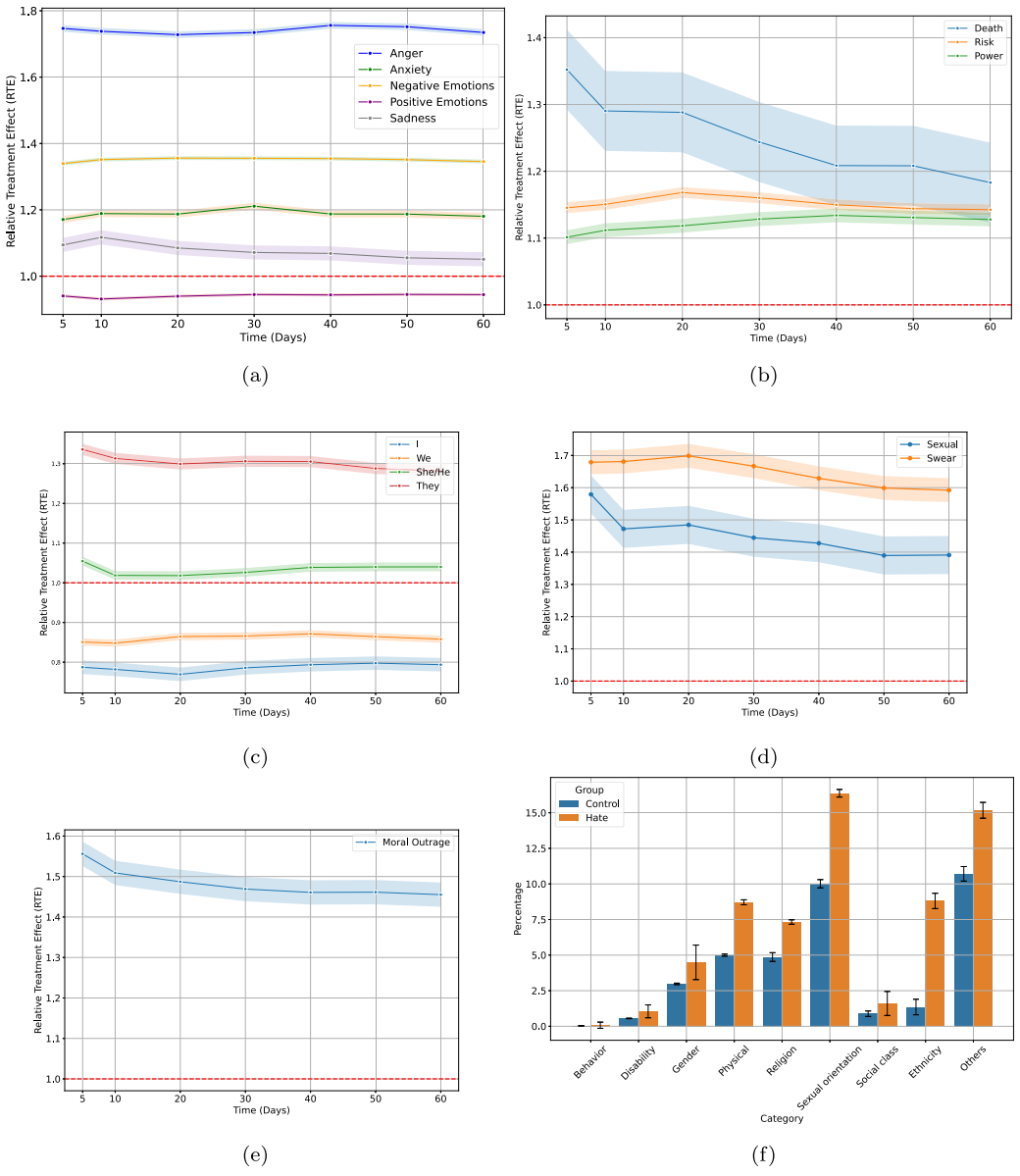


**Fig. 4.** The weighted average RTE over time for different outcomes ((a),(b),(c), (d) and (e)). An independent sample t-test confirmed these differences (t[−6.08, 2.81]; $p < 0.05$). The shaded areas represent the 95% confidence interval. (f) represents the distribution of profanity across different categories between two groups.

the Chinese people for COVID-19 deaths, fearmongering related to the pandemic, and criticism towards the Chinese Government intertwined with racism. Analyzing the significance of these observations, most outcomes consistently had Cohen's d values above 0.2, indicating substantial effects, whereas risk sometimes fell below this threshold. The longitudinal analysis shows that these elevated levels remained relatively stable throughout the observation period.

### *4.1.2. Pronouns usage patterns*

Our analysis of pronoun usage among hateful users compared to control users shows distinct trends (Fig. 4(c)). Consistent with **H1c**, hateful users exhibit a higher frequency of third-person pronouns 'she/he' and 'they', indicating detachment and viewing others as abstract entities. For 'she/he' pronouns, the RTE values remain consistently above 1, starting at 1.05 on day 5 and remaining at 1.04 by day 60. For 'they' pronouns, the RTE starts at 1.34 on day 5 and stabilizes around 1.28 by day 60. Conversely, first-person pronouns ('i' and 'we') show RTE values less than 1 for hateful users, indicating lesser self-referencing and group inclusion. Most pronouns consistently have Cohen's d values above 0.2, indicating substantial effects, while 'they' occasionally falls below this threshold.

#### 4.1.3. Profanity usage

We observe that the RTE for the LIWC profanity categories ('sexual' and 'swear') show an increase in the first 5 days after mentioning hateful content, with values of 1.58 and 1.68, respectively (Fig. 4(d)). This is followed by a gradual decrease over time, with the RTE values reaching 1.39 and 1.59 by day 60 ('swear' shows substantial effects with Cohen's values above 0.2, while 'sexual' remains below this threshold). This trend suggests that while the initial response to hateful posts involves heightened use of profanity, this effect diminishes over time.

Further, from the dictionary count of the Surge AI profanity repository, we find that tweets containing hateful content have a higher prevalence of profane words across all categories compared to control tweets. Here, we report the results averaged over the four largest strata out of ten (See Fig. 4(f)). We thus find strong support for hypothesis **H1d**, as there is more profane language among users posting hateful content, which gradually decreases over time.

#### 4.1.4. Moral outrage

There is a higher overall level of moral outrage among hateful users (Fig. 4(e)). Consistent with **H1e**, social media users who post hateful content maintain higher levels of moral outrage in their language compared to users who do not (with this level decreasing over time). Specifically, the RTE starts out at 1.56 on day 5 and shows a gradual decline over time, reaching 1.46 by day 60. Despite this decrease, the RTE consistently remains above 1, indicating that hateful users exhibit higher levels of moral outrage throughout the observed period. Given the significance of these findings, Cohen's d values for moral outrage are consistently above 0.2 after the 5-day period.

To sum up, in **RQ1**, we examine the linguistic and cognitive characteristics of social media users who post hateful content compared to those who do not, and our analyses show significant differences between these groups. There are elevated levels of anger, anxiety, negative emotions, and sadness among hateful users, but lower positive emotions. Additionally, the use of dehumanizing language is more prevalent among hateful users, evidenced by the higher frequency of third-person pronouns, indicating detachment and viewing others as abstract entities, and less use of first-person pronouns ('i' and 'we'), indicating less self-referencing and group inclusion. Both profanity and moral outrage start high and decrease over time, but the RTE values remain above 1 throughout the entire observed period. These findings confirm that hateful users significantly differs from control on linguistic and cognitive characteristics, with elevated negative emotions, dehumanizing language, profanity, and moral outrage, thereby answering **RQ1**. Additionally, we note that results are mostly consistent across all strata (see Fig. D.11 in Appendix D).

### 4.2. RQ2. Thematic patterns in hateful posts

To answer **RQ2**, we apply topic analysis on the largest four strata out of ten following the propensity score matching phase. Our analysis reveals similar results and trends in three out of the four selected strata. Due to space limitations, we present the detailed results only for Stratum 5, which is representative of the observed trends (additional details and results for the other two strata are available in Table F.3 in Appendix F). Stratum 5 includes 1095 users: 614 hateful users and 481 control users. They posted a total of 631,504 tweets, with hateful users contributing 327,187 tweets and control users posting 304,317 tweets.

#### 4.2.1. Topic analysis

The results of the BERTopic model revealed distinct patterns that highlight the differences in focus and sentiment between hateful users and control users. The 20 topics for each group are listed in Table E.1 and Table E.2 in Appendix E. Hateful users predominantly engage in negative and controversial topics. For instance, topics such as "Hong Kong protests" and "US Politics" are common among hateful users, highlighting a tendency to focus on controversial issues and blame certain groups, especially during the COVID-19 pandemic. In contrast, control users tend to discuss topics with more neutral or positive content. Examples include topics like "Good thank true like", which suggests a focus on gratitude and positivity, and "Job search resume help", highlighting practical and supportive conversations about career and personal development. Additionally, control users are more likely to discuss everyday life and hobbies, such as "Music radio listen stayhome" and "Drawing art enjoy kids", showcasing a broader range of interests and a more communal tone. These observations are consistent with our analysis of the RTA values of the LIWC emotion categories we discussed earlier. In the next section, we take a deeper look into the dynamics of these topics.

#### 4.2.2. Topic interconnectedness

The results of the topic interconnectedness analysis indicate that topics predominantly associated with hateful content, referred to as 'hateful topics', are more interconnected than those associated with non-hateful posts, referred to as 'non-hateful topics,' with Beta = 1.68, SE = 0.57, $t$-value = 2.93,P = 0.003, and $R^2$ (m/c) = 0.036/0.76 (See Fig. 5). Additionally, entropy is higher in the hateful topic networks, indicating a more evenly distributed connection pattern where all hateful topics are actively involved in the network. In addition, clustering coefficients are higher, meaning that hateful topics are more likely to form tightly knit groups or clusters. Moreover, the shortest paths are lower, meaning that hateful topics are more directly connected. Finally, density is greater in hate networks, which indicates that a higher proportion of possible connections between hateful topics are realized. Detailed properties of each network are provided in Appendix D. Thus, these metrics collectively support the hypothesis **H2a**, demonstrating that hateful posts exhibit a more tightly connected network of related topics compared to non-hateful posts.

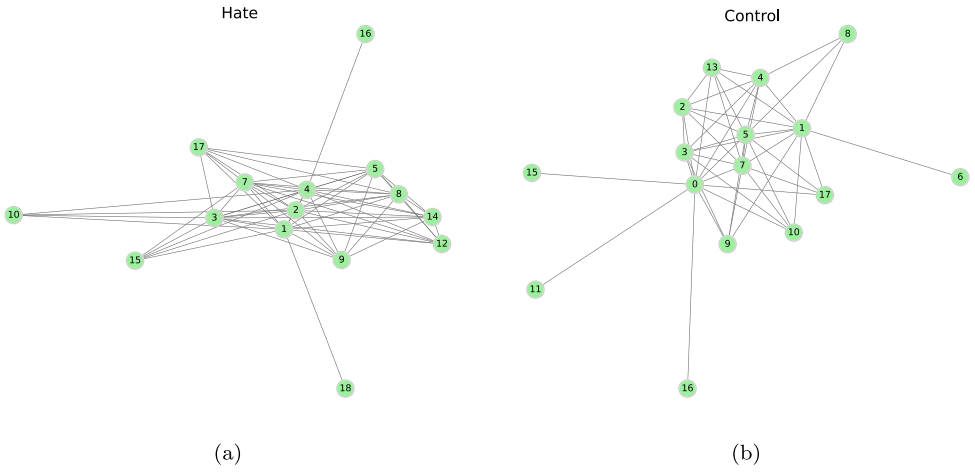


**Fig. 5.** Network interconnectedness of topics: Nodes symbolize topics, and edges represent the co-occurrence of topics within tweets. The thickness of each edge corresponds to the strength of the connection, specifically indicating the frequency of co-occurrence between topics. Network plots were derived from hate-related (a) and control (b) topics. The labeling for the actual topic numbers is provided in Appendix C[6].

#### *4.2.3. Global cohesion*

We calculate Pairwise CS between each tweet and the remaining tweets. Compared to control tweets, hateful tweets demonstrate greater similarity to each other: Beta = 0.001, SE < 0.0001, $t$-value = 39.06, $p$-value < 0.001, $R^2m/c$ = 0.05/0.26. These findings support hypothesis **H2b**, in that tweets containing hateful content show a higher degree of similarity to each other, more consistently referencing similar worldviews compared to non-hateful content.

#### *4.2.4. Topic specificity*

Looking at topic specificity, we evaluate the inequality of within-tweet topic distributions using the Gini coefficient. Results from the linear mixed-effects models, which controlled for word count and nesting within user IDs, show that hateful tweets have lower topic specificity compared to non-hateful tweets. The model results are as follows: Beta = −0.004, SE < 0.001, $t$-value = −12.33, $p$-value < 0.001. The $R^2m/c$ is 0.01/ 0.17. This negative Beta value suggests that hateful tweets are associated with lower Gini coefficients, indicating lower topic specificity compared to non-hateful tweets. This suggests that hateful content is more concentrated and less varied compared to non-hateful content. Thus, we find support for **H2c**, suggesting that hateful posts lack topic-specific detail and nuance compared to non-hateful content, often reflecting simplistic reasoning and a reliance on generalized statements. This lack of integrative complexity highlights a tendency to oversimplify issues, avoiding detailed exploration of individual topics while reinforcing consistent, overarching themes.

To summarize, in **RQ2**, we explore how thematic patterns and specificity of hateful posts differ from non-hateful posts. Our analyses indicate that hateful topics are more interconnected and exhibit higher entropy, clustering coefficients, and density, as well as lower distances between nodes. Hateful tweets also demonstrate greater similarity to each other, indicating higher global cohesion compared to control tweets. Furthermore, hateful tweets show lower topic specificity, reflecting a consistent and focused narrative.

## 5. Related work

Despite significant advancements in hate speech detection and analysis, several critical gaps persist, limiting our understanding of hate speech dynamics and their broader implications. This study addresses these gaps by focusing on previously underexplored dimensions, including temporal dynamics, causal inference, and narrative structures, as outlined below:

### *5.1. Static detection models vs. Dynamic user behavior analysis*

Most prior studies on hate speech detection rely on static datasets and keyword-based models, which analyze hate content through isolated instances without accounting for its temporal or longitudinal evolution. For example, studies such as Davidson et al. (2017) and Waseem and Hovy (2016b) employed lexicon-based or keyword-focused approaches to classify hate speech. These models effectively identify hate speech at a single point in time but fail to capture how hate speech evolves or influences user behavior over an extended period.

[6] Topic "0" represents outliers that do not strongly belong to any coherent topics, as automatically assigned by BERTopic. These outliers were included in the visualization for completeness but were not central to the analysis.

However, some studies have shifted focus towards temporal dynamics. Rajadesingan, Resnick, and Budak (2020) explored temporal patterns by analyzing the intensity of hate speech over time, providing insights into how hate intensity fluctuates in response to external events. Similarly, Pavlopoulos, Sorensen, Dixon, Thain, and Androutsopoulos (2020) investigated the temporal impact of hate speech mentions, examining the behavioral changes associated with specific instances of hate speech.

While these studies provide valuable insights into temporal trends, they primarily focus on surface-level metrics, such as intensity and frequency, or on short-term behavioral triggers. In contrast, our work uniquely examines how hate speech causally affects users' psycholinguistic and topical characteristics over time. By integrating linguistic, emotional, and thematic analyses with causal inference methods, we uncover deeper psychosocial impacts and explore how user behavior and narratives evolve longitudinally. This dynamic approach extends beyond detection to reveal the evolving emotional and cognitive markers associated with hate speech engagement, offering a more comprehensive understanding of its societal effects.

### *5.2. Neglect of narrative cohesion and structural dynamics*

Hateful posts are not isolated but function as interconnected systems, where thematic cohesion and structural dynamics play a crucial role in their propagation. Existing studies have explored these dynamics to varying extents. For example, Zannettou et al. (2018) analyzed the flow of hate speech across platforms, highlighting the interconnectedness of narratives and their migration patterns. Similarly, Lewandowsky, Cook and Ecker (2018) investigated the role of thematic networks in sustaining conspiracy theories, which often intersect with hate speech, emphasizing the structural properties that allow such narratives to persist. Ribeiro, Calais, Santos, Almeida, and Meira (2021) examined co-occurrence patterns of hateful topics using network analysis, uncovering clusters of recurring themes that reinforce hateful ideologies.

While these studies offer valuable insights, they primarily focus on structural properties or topic co-occurrences, often neglecting the psycholinguistic dimensions of hateful posts. In contrast, our work integrates structural analysis with BERTopic modeling to simultaneously examine linguistic, emotional, and thematic aspects of hate speech. This allows us to uncover not only how topics co-occur but also how they form cohesive yet narrow ideological frameworks. Our findings reveal that hateful topics exhibit high global cohesion and low specificity, enabling them to sustain harmful narratives effectively. By combining structural and psycholinguistic analyses, our study provides a holistic understanding of the mechanisms driving the persistence and spread of hateful content.

### *5.3. Limited behavioral comparisons between hateful and non-hateful users*

Existing research has often analyzed hateful users in isolation, focusing on their linguistic, emotional, or behavioral traits without systematically comparing them to non-hate users. For instance, Ribeiro et al. (2021) characterized hateful users on Twitter/X by analyzing their network activity, linguistic markers, and emotional tendencies, but their work did not directly compare these users with non-hateful counterparts. Similarly, Mathew, Saha, Yimam et al. (2019) investigated hateful users' engagement patterns and language use, highlighting their higher activity levels and prevalence of negative emotional expressions.

Some studies have addressed this gap by contrasting hateful and non-hateful users. For example, Chatzakou et al. (2017) conducted a comparative analysis of abusive and non-abusive users on Twitter/X, revealing differences in posting frequency, linguistic style, and sentiment. Similarly, De Gibert, Perez, Garcia-Pablos, and Cuadros (2018) explored linguistic and contextual differences between hateful and non-hateful content in text corpora, identifying distinctive patterns such as a higher use of pronouns and extreme sentiment in hate speech.

While these studies offer valuable insights, they lack a robust causal framework to uncover the underlying factors that distinguish users who post hateful content from those who do not. Our study builds on these efforts by employing propensity score matching to systematically compare hateful and non-hateful users, focusing on nuanced linguistic, cognitive, and emotional differences. This approach enables us to isolate the causal effects of hateful posts engagement, revealing unique patterns such as increased moral outrage, heightened negative emotions, and elevated use of dehumanizing language among hateful users.

### *5.4. Underexplored longitudinal and psychosocial effects*

While significant progress has been made in hate speech detection and analysis, many studies focus on static snapshots of hate speech or its immediate impacts, neglecting its long-term effects on user behavior and mental health. For instance, De Choudhury et al. (2016) explored the relationship between online language and mental health but focused primarily on descriptive correlations rather than longitudinal causal dynamics. Similarly, Mathew, Rethinam, Singh, and Mukherjee (2021) analyzed temporal patterns in hate speech but did not examine its sustained psychological impacts on users.

Some studies have attempted to bridge this gap. For example, Kiciman et al. (2018) employed causal inference techniques to study the effects of online behaviors on users' well-being, providing a framework for understanding longitudinal impacts. However, this work does not specifically address the psychosocial dynamics of hateful posts engagement. Other research, such as by Ribeiro, Santos, Almeida, and Meira (2020), examined the emotional evolution of users involved in toxic interactions, shedding light on behavioral changes but without integrating psycholinguistic markers into the analysis.

Our study builds on this foundation by incorporating both longitudinal and psychosocial dimensions into the analysis of hate speech. By leveraging causal inference methods such as propensity score matching, we systematically examine how hate speech engagement affects users' linguistic, emotional, and cognitive behaviors over time. Our findings reveal that hate speech users experience heightened negative emotions, sustained moral outrage, and persistent dehumanizing language patterns, which collectively indicate deeper psychosocial impacts.

### 5.5. Advancing methodological frameworks

Many existing studies on hate speech detection rely on traditional descriptive analyses or keyword-based classification methods, which are limited in their ability to uncover deeper causal relationships or thematic structures. For example, Davidson et al. (2017) employed lexicon-based approaches to classify hate speech, effectively identifying instances of harmful content but offering little insight into the broader contextual or behavioral patterns associated with such content. Similarly, Nobata et al. (2016) utilized linguistic features for hate speech detection but focused on static datasets, lacking the capacity to explore longitudinal dynamics or user behavior.

In recent years, some studies have moved towards more advanced methodologies. Kiciman et al. (2018) introduced causal inference techniques to study the impact of online behaviors on health outcomes, showcasing the potential for causal frameworks in social media research. Ribeiro et al. (2020) employed network-based approaches to analyze the interconnectedness of hateful topics, providing structural insights into how such narratives spread. However, these studies typically focus on a single dimension – such as causality, structure, or behavior – without integrating multiple perspectives.

Our study advances the methodological landscape by combining causal inference, psycholinguistic analysis, and BERTopic modeling to provide a comprehensive understanding of hate speech dynamics. Through propensity score matching, we establish a robust causal framework to isolate the effects of hate speech engagement. By integrating BERTopic, we examine the global cohesion and specificity of hateful posts, uncovering how thematic structures sustain harmful ideologies. Additionally, our inclusion of longitudinal data allows us to analyze how linguistic, emotional, and cognitive behaviors evolve over time. This multi-dimensional approach bridges the gap between descriptive and causal analyses, offering actionable insights for researchers and policymakers aiming to mitigate the spread and impact of hate speech.

## 6. Discussion

### 6.1. Summary of findings

Our findings reveal clear differences in the linguistic behaviors and cognitive traits of users who engage in hateful posts, demonstrating how information behavior in these users diverges from that of non-hateful users.

Our first research question explores cognitive and linguistic differences between hateful and non-hateful Twitter/X users. Hateful users exhibited higher levels of anger, anxiety, and negative emotions, consistent with prior work linking these emotions to dehumanizing attitudes and hateful posts (Alorainy et al., 2018; ElSherief et al., 2018; Giner-Sorolla & Russell, 2019; Haybron, 2002; Mathew et al., 2018; Matsumoto et al., 2016; Sell et al., 2009). They also used more language related to power, death, and risk, which are associated with prejudice and dehumanization (ElSherief et al., 2018; Goff et al., 2008; Markowitz & Slovic, 2020; Paasch-Colberg et al., 2021). The use of third-person pronouns suggests an outgroup focus, reinforcing detachment and hostility. These findings enhance information behavior theory by illustrating how hateful users disseminate harmful content through a dehumanizing, emotionally charged lens (ElSherief et al., 2018; Faulkner & Bliuc, 2018; Zannettou et al., 2020).

Profanity and offensive language are also notably higher among hateful users, consistent with the literature on linguistic markers of hate (Bartlett et al., 2014; Bilewicz & Soral, 2020; Malmasi & Zampieri, 2017; Mathew et al., 2018). This reinforces the role of language as a vehicle for hate speech dissemination, where profanity amplifies emotional aggression and alienates targeted groups (Davidson et al., 2017). Our study emphasizes the information transmission role of moral outrage in sustaining engagement with hateful posts.

These results support previous work showing that collective aggression is key to sustaining negative emotions in hateful content. Our findings align with (Strathern, Schoenfeld, Ghawi, & Pfeffer, 2020), who found that moral outrage and collective aggression towards specific targets amplified the spread of hate speech on Twitter/X. Similarly, Wang, Zhou, and Kinneer (2024) found that moral framing in #StopAsianHate campaigns, especially the use of virtue values, increased engagement. This suggests that moral outrage plays a significant role in driving the propagation of hate speech online.

Further, our findings indicate that hateful posts are tightly connected and cohesive. This finding is consistent with prior work on information coherence in conspiracy theories, suggesting that conspiracy theories are characterized by tightly interconnected information networks (Lewandowsky, Cook & Lloyd, 2018; Miani et al., 2022; Vergani, Martinez Arranz, Scrivens, & Orellana, 2022; Wood et al., 2012). This interconnectedness facilitates information retention and dissemination, contributing to the spread of harmful narratives within social media ecosystems. Understanding the coherence of hateful topics helps shed light on the broader structure of harmful information flows on platforms like Twitter/X.

### 6.2. Limitations

While Twitter/X is a popular platform for information sharing, our reliance on data from a single platform may limit the generalizability of our findings across other platforms. Platform-specific behaviors and content moderation policies can affect how hate speech spreads, meaning insights from Twitter/X may not fully apply to closed platforms like Facebook or ephemeral messaging apps (Vicente, 2023). Additionally, the focus on English-language hate speech introduces linguistic and cultural variability as a limitation, potentially overlooking how hate is expressed in other languages. These factors should be addressed in future research that seeks to generalize across platforms and cultures.

### 6.3. Ethical considerations

We prioritized user anonymity and privacy throughout this study. Personally identifiable information (PII) was removed during data collection and analysis, and the dataset was processed to ensure that no sensitive user information could be traced back to individuals. All analyses rely solely on aggregated metrics and group-level comparisons, avoiding any focus on individual users. These measures align with ethical standards for research involving public social media data, ensuring compliance with principles of transparency, privacy, and cultural sensitivity.

To further promote transparency and encourage replication, we are willing to share our dataset, up to 500,000 records, upon request, in accordance with Twitter/X's Terms of Service at the time of data collection. This ensures adherence to platform guidelines while enabling replication and extension of our analyses.

### 6.4. Implications: Practical and theoretical

The practical implications of this study are multifaceted, addressing both individual social media users and social media platform administrators and policymakers. Our findings suggest that content moderation strategies should focus on tracking interconnected hateful topics to predict future trends and prevent harm. Despite the current efforts to curb hate content, it remains a significant issue in social media platforms (Bührer et al., 2024; Chris Hale, 2012; League, 2021). The temporal dynamics of moral outrage observed in this study indicate that interventions targeting peak emotional engagement could significantly reduce the spread of hate speech.

For instance, platforms could implement real-time sentiment monitoring systems to detect surges in moral outrage based on linguistic markers (e.g., heightened profanity, strong emotional language). During these peak periods, platforms could employ time-sensitive intervention mechanisms, such as temporarily increasing moderation efforts, issuing proactive warnings to users engaging with inflammatory content, or reducing the visibility of content flagged as likely to incite moral outrage. A real-world example includes YouTube's approach of demonetizing or limiting the reach of videos during sensitive news cycles to prevent the escalation of harmful narratives, which could be adapted to target hate speech specifically. Similarly, platforms like Twitter/X could introduce temporary friction mechanisms (e.g., requiring a cooling-off period before posting replies) when moral outrage metrics spike, thereby mitigating the escalation and spread of hate speech during emotionally charged events. These interventions align with the findings of Pennycook et al. (2021), who demonstrated that shifting attention to accuracy can significantly reduce the spread of misinformation online.

For individual users, particularly those targeted by hate speech, our findings suggest several specific strategies to manage and respond to online hate. Unlike existing research that primarily focuses on identifying hate speech (e.g., Chiril et al., 2022; Jahan & Oussalah, 2023; Watanabe et al., 2018), this study centers on understanding the individuals spreading hate speech. Knowing that hate speech often involves dehumanizing language and heightened negative emotions (e.g., Markowitz & Slovic, 2020; Mathew et al., 2018), users can employ resilience-building techniques such as cognitive-behavioral strategies to counteract the emotional toll (Beck, 2020). For example, platforms could implement dynamic support systems during peak periods of moral outrage, such as promoting messages of empathy and compassion through automated nudges or interstitial content designed to encourage prosocial behavior. Evidence from interventions designed to foster empathy in online communication has shown promise in reducing hostile interactions by encouraging users to view situations from others' perspectives (Farrelly & Bennett, 2018). Supportive online communities and counter-hate initiatives (Garland, Ghazi-Zahedi, Young, Hébert-Dufresne, & Galesic, 2022; Mathew, Saha, Tharad et al., 2019) can also play an active role during these high-risk times by promoting positive content.

In addition to platform-level interventions, these findings have significant implications for policy development and large-scale social initiatives. Governments, NGOs, and organizations working to combat hate speech can leverage the insights from this study to design targeted educational campaigns that address the underlying emotional and cognitive patterns associated with hate speech. For instance, incorporating the understanding of heightened moral outrage and negative emotional states into public awareness campaigns could help de-escalate online hostility by promoting critical thinking and emotional regulation. Evidence from large-scale interventions, such as the European Union's Code of Conduct on Countering Illegal Hate Speech Online, highlights the effectiveness of collaborative approaches in reducing online hate speech through monitoring and reporting mechanisms (European Commission, 2016)

The theoretical contributions of this study lie in understanding how hate speech operates as information within online ecosystems (Chetty & Alathur, 2018; Watanabe et al., 2018). A critical gap in existing research is the limited exploration of how hate speech networks evolve longitudinally, particularly in terms of user behaviors and thematic patterns over time. This study addresses this gap by providing a novel analysis of the interconnectedness of topics within hate speech, moving beyond static detection models to examine dynamic processes. By analyzing the temporal dynamics and tightly cohesive structures of hate speech topics, our findings reveal insights into how harmful information networks form, persist, and adapt over time. Both conspiracy theorists and hateful users exhibit tightly cohesive topic structures, which suggest a common psychological architecture that enables the propagation of specific narratives. These cohesive structures shed light on how unrelated ideas are woven together to support shared beliefs and ideologies, extending network theory and conspiracy theory research into hate speech analysis (Lewandowsky, Cook & Lloyd, 2018; Papcunová et al., 2023; Zannettou et al., 2019).

Our findings that hateful users show elevated anger, anxiety, and negative emotions, with lower positive emotions, support information diffusion and social contagion theories, which suggest that negative emotions spread more virally in social networks, especially in hate speech contexts (Bakshy, Hofman, Mason, & Watts, 2011; Barberá, 2015). These results deepen our understanding

of how negative emotional content becomes entrenched, creating feedback loops that sustain harmful behaviors (Centola, 2010). This underscores the need for real-time monitoring to mitigate the spread and long-term impact of emotionally charged hateful posts on users and communities.

In summary, this study bridges multiple disciplines – psychology, sociology, computational linguistics, and policy research – by offering a comprehensive framework for understanding hate speech through cognitive, emotional, and structural lenses. The findings highlight actionable pathways for designing time-sensitive interventions, fostering resilience in individuals, and informing more effective content moderation policies. By addressing critical research gaps and providing a scalable framework, this study lays the groundwork for interdisciplinary collaborations aimed at mitigating the broader societal impacts of hate speech. Its significance extends beyond academic inquiry, offering practical and theoretical insights to drive meaningful social change.

## 7. Conclusion and future work

This study provides a comprehensive examination of the emotional, linguistic, and thematic dynamics of hate speech on social media, with a particular focus on its longitudinal and causal impacts on user behavior. By employing advanced methodologies, including propensity score matching and topic modeling, we demonstrate significant differences between hateful users and their non-hateful counterparts. These differences, such as elevated anger, anxiety, and moral outrage, as well as tightly interconnected and focused hateful posts, offer critical insights into how hate speech sustains harmful ideologies and affects social media ecosystems.

Our findings contribute both theoretically and practically, highlighting the need for real-time interventions and tailored content moderation strategies. These strategies could mitigate the spread of hate speech and its negative societal impacts by addressing the interconnected narratives and emotional triggers that drive its propagation.

While this study focuses on Twitter/X, the methodology is inherently flexible and can be adapted to other platforms, including emerging ones like Bluesky. The propensity score matching covariates can be customized to leverage platform-specific metrics, such as reposts, likes, replies, or unique interaction patterns, to measure user activities and engagement effectively. Additionally, the linguistic and cognitive outcomes as well as the thematic analysis components rely solely on user-generated text, ensuring the approach's applicability across different platforms, regardless of their structural or technical characteristics. Future work will refine these methodologies to address platform-specific nuances, enabling broader applicability and enhancing the framework's relevance in diverse social media contexts.

## CRediT authorship contribution statement

**Amira Ghenai:** Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Zeinab Noorian:** Supervision, Methodology, Funding acquisition, Data curation. **Hadiseh Moradisani:** Visualization, Investigation, Data curation. **Parya Abadeh:** Resources, Methodology, Investigation. **Caroline Erentzen:** Writing – review & editing, Writing – original draft, Formal analysis. **Fattane Zarrinkalam:** Writing – review & editing, Funding acquisition.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Amira Ghenai reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. Zeinab Noorian reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. Fattane Zarrinkalam reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grants RGPIN-2022-03809, RGPIN-2022-04197, and RGPIN-2022-05193.

## Appendix A. Edge connectivity

Edge connectivity (Y-axis) by $|r|$ thresholds (X-axis) for the BERTopic correlation matrices. Lines refer to hate users (red) and control users (green) networks. The threshold value refers to the mean of $|r|$ coefficients. If the threshold is too low ($|r| = 0$ on the X axis), all topics are connected to each other, resulting in the mean average connectivity being equal to k, the total number of topics (20 in our case). If the threshold is too high, topic co-occurrences are not detected, causing nodes in the network to be disconnected, resulting in an average degree of connectivity of 0. To determine a meaningful threshold, we adopted a method consistent with prior studies, such as the work by Miani et al. (2022), which employed a similar approach to defining thresholds in network analysis. Specifically, we calculated the mean of $|r|$ coefficients across the correlation matrices for both groups (hate and control), which were 0.0113 and 0.0092, respectively. Taking the average of these values, we set the threshold at **0.01025**, marked by the vertical dashed line in the figure. This threshold was chosen as it balances the connectivity across the networks, ensuring a meaningful

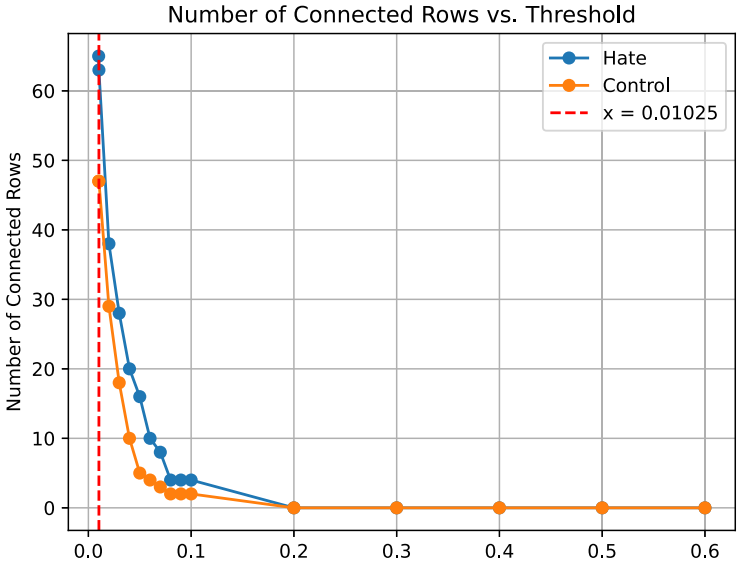


**Fig. A.6.** Edge connectivity as a function of $|r|$ threshold for BERTopic matrices.

degree of co-occurrence without over- or under- linking topics. The rationale for this threshold lies in ensuring statistical grounding for the network structure while avoiding arbitrary cutoff points. When the threshold is set lower than **0.01025**, it results in overly dense networks (at a threshold of 0.01, the number of connected rows increases to 65 for hate and 47 for control), which fail to reflect meaningful relationships between topics. Conversely, thresholds higher than this value disconnect most topics (at a threshold of 0.04, the number of connected rows reduces to 20 for hate and 10 for control), leading to sparse networks with limited analytical value. By selecting **0.01025**, we ensured a balance between these extremes. A sensitivity analysis of thresholds revealed that small variations (e.g., ±10%) around this threshold did not significantly alter the findings (see Fig. A.6).

## Appendix B. Temporal variations of covariates over time

See Figs. B.7 and B.8.

## Appendix C. Temporal variations of covariates over time across strata

See Figs. C.9 and C.10.

## Appendix D. Relative treatment effect

The heatmap illustrates the relative treatment effect across all users over a 10-day period, segmented by stratum and various categories such as emotions (anxiety, anger, sadness), pronouns (i, we, she/he, they), and other linguistic features (death, religion, sexual, swear). The consistent coloring across most strata indicates that the results are stable and do not show significant variations across different strata. This stability suggests that the observed patterns in hate speech and its linguistic and emotional characteristics are robust and not heavily influenced by the specific stratification of the users. Consequently, the consistency of the propensity scores across different strata strengthens the reliability of our findings regarding the impact of hate speech on users' language and emotional expressions (see Fig. D.11)

## Appendix E. Key topics in hate speech and control users' tweets

See Tables E.1 and E.2.

## Appendix F. Statistics for the hate speech and control networks

See Table F.3.

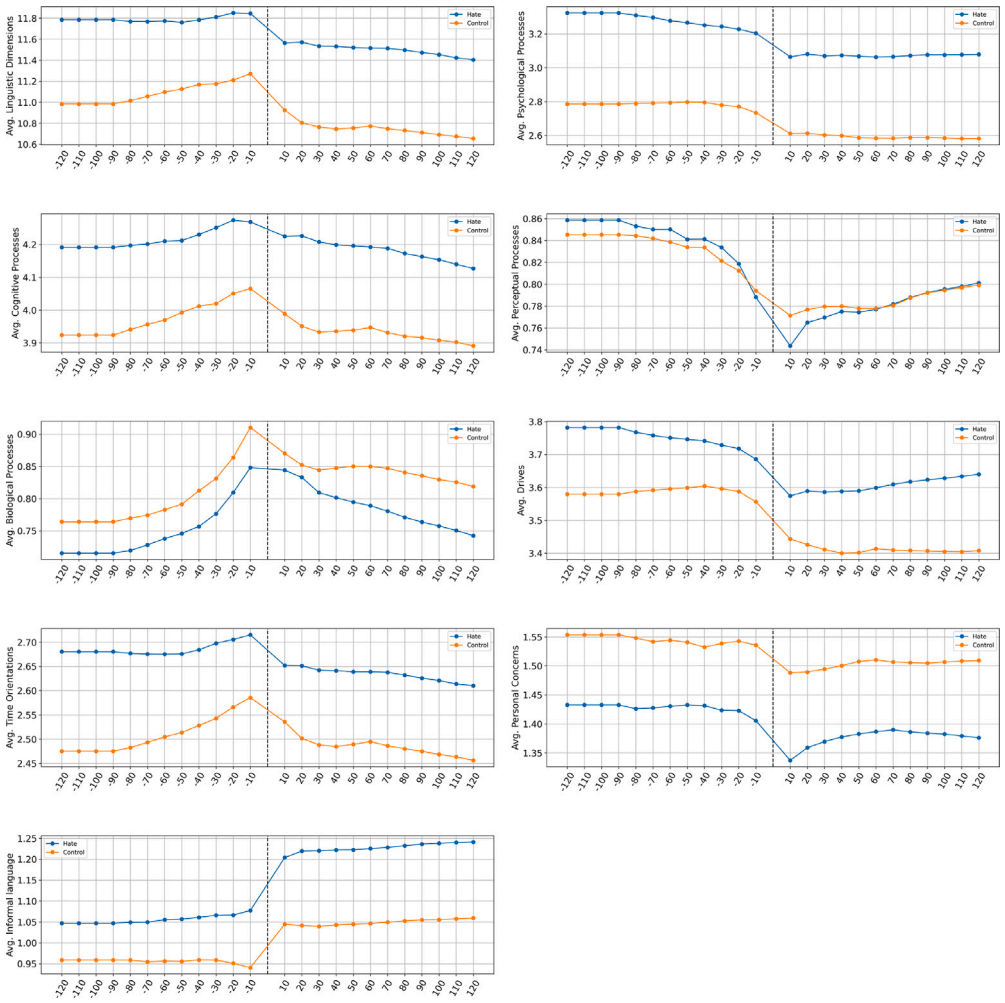


**Fig. B.7.** Temporal distributions of various LIWC categories over time. Each plot represents the average value of a specific LIWC category on the *y*-axis, calculated across all users, with respect to the number of days before (−) and after (+) posting hateful content (x-axis). The vertical dashed line indicates the transition point between the "before" and "after" periods. Comparisons are shown between the "Hate" group (blue) and the "Control" group (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

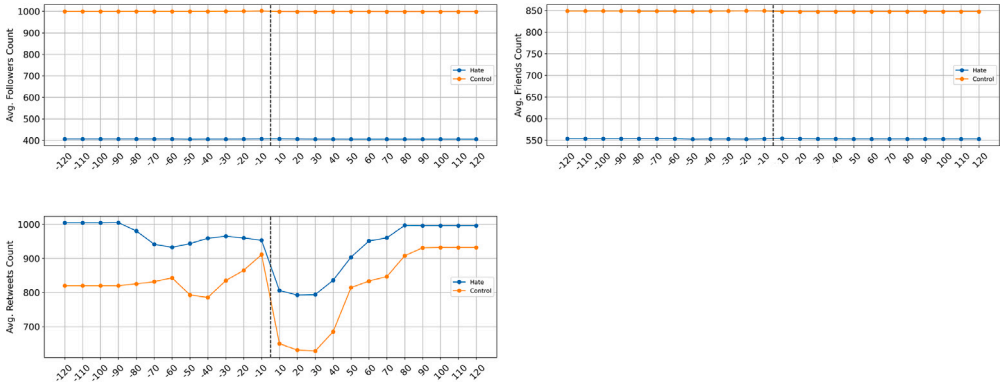


**Fig. B.8.** Temporal distributions of user engagement metrics (followers, friends, and retweets) over time. Each plot represents the average value of a specific measure category on the *y*-axis, calculated across all users, with respect to the number of days before (−) and after (+) posting hateful content (x-axis). The vertical dashed line indicates the transition point between the "before" and "after" periods. Comparisons are shown between the "Hate" group (blue) and the "Control" group (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

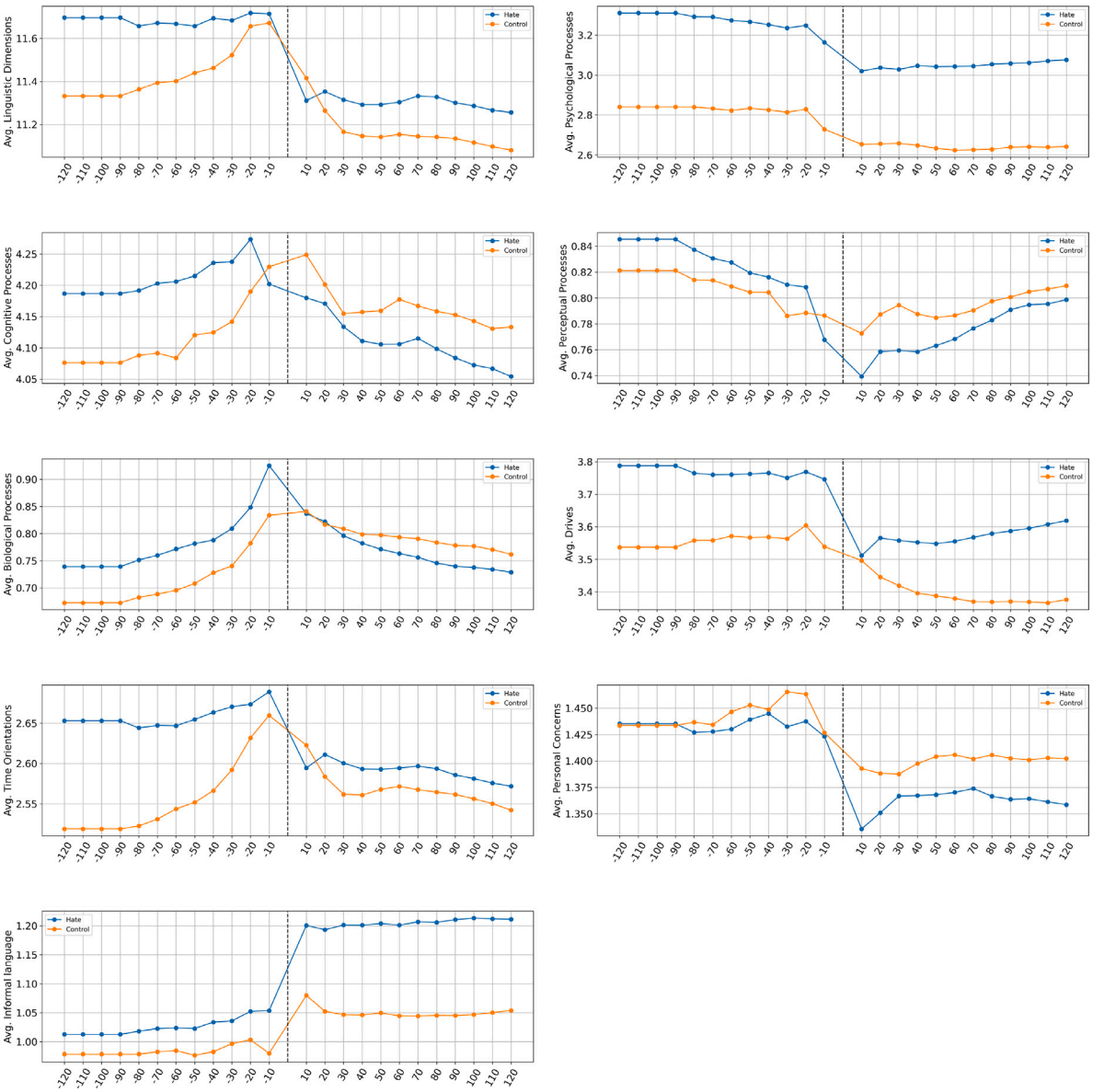


**Fig. C.9.** Temporal distributions of various LIWC categories over time for **stratum 5**. Each plot represents the average value of a specific LIWC category on the *y*-axis, calculated across all users, with respect to the number of days before (−) and after (+) posting hateful content (x-axis). The vertical dashed line indicates the transition point between the "before" and "after" periods. Comparisons are shown between the "Hate" group (blue) and the "Control" group (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

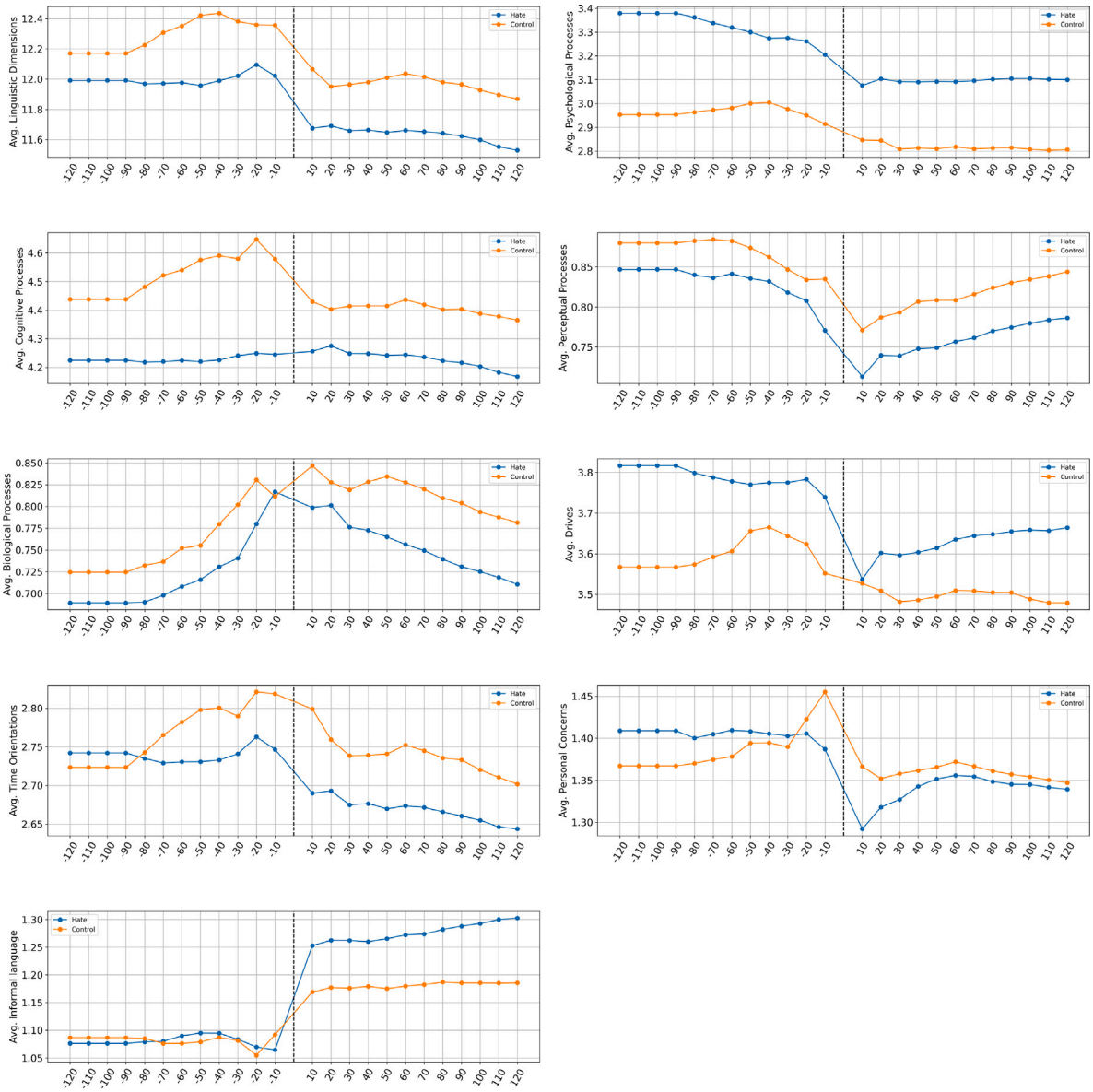


**Fig. C.10.** Temporal distributions of various LIWC categories over time for **stratum 6**. Each plot represents the average value of a specific LIWC category on the *y*-axis, calculated across all users, with respect to the number of days before (−) and after (+) posting hateful content (x-axis). The vertical dashed line indicates the transition point between the "before" and "after" periods. Comparisons are shown between the "Hate" group (blue) and the "Control" group (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

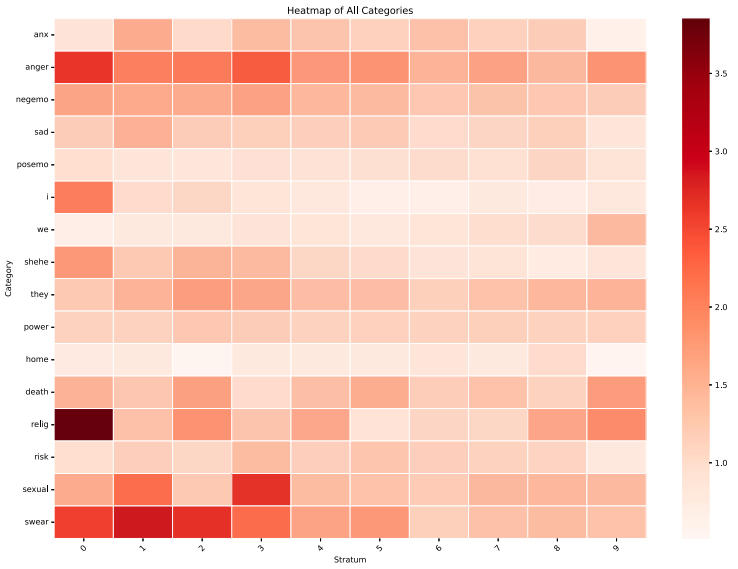


**Fig. D.11.** Relative treatment effect across all users in 10 days period.

**Table E.1**
Topics extracted from tweets of control users.

| | Topic | Representation |
|---|---|---|
| 1 | RT people COVID amp | ['rt', 'people', 'covid', 'amp', 'coronavirus', 'us', 'like', 'china', 'one', 'get', 'trump', 'time', 'new', 'would', 'need', 'know', 'dont', 'good', 'virus', 'world'] |
| 2 | RT COVID coronavirus amp | ['rt', 'covid', 'coronavirus', 'amp', 'china', 'people', 'us', 'cases', 'new', 'police', 'trump', 'one', 'like', 'hong', 'virus', 'kong', 'time', 'may', 'need', 'get'] |
| 3 | Baseball RT good like | ['nan', 'rt', 'baseball', 'dawgs', 'good', 'like', 'dirt', 'dirtdawgs', 'one', 'please', 'thank', 'day', 'god', 'posted', 'dirtdawgsports', 'photo', 'us', 'know', 'love', 'great'] |
| 4 | Masks face wear ventilators | ['masks', 'mask', 'face', 'wear', 'wearing', 'ventilators', 'ppe', 'dental', 'practices', 'dposillicocom', 'gowns', 'gloves', 'sanitizer', 'surgical', 'catalog', 'shields', 'kn', 'medical', 'covid', 'ir'] |
| 5 | Job search resume help | ['job', 'jvs', 'jcfs', 'hirepower', 'search', 'jobsearch', 'help', 'looking', 'services', 'resume', 'networking', 'ampm', 'leadership', 'career', 'digitaltransformation', 'whether', 'opportunity', 'recruiter', 'careerservices', 'rent'] |
| 6 | Food quicker help meals | ['food', 'fpuc', 'pua', 'quicker', 'additional', 'ohioans', 'midmay', 'help', 'acceptable', 'set', 'anything', 'well', 'restaurants', 'would', 'meals', 'need', 'like', 'dinner', 'said', 'coffee'] |
| 7 | Michigan reopen stay home | ['andover', 'home', 'michigan', 'reopen', 'jcfs', 'stay', 'michiganprotest', 'insurance', 'reopening', 'order', 'closed', 'people', 'keeping', 'chicago', 'roundtrip', 'illinois', 'michiganders', 'stayathome', 'specsavers', 'insured'] |
| 8 | Music radio listen stayhome | ['nothingbutgoodmusic', 'stayhome', 'playing', 'radio', 'listen', 'music', 'starting', 'song', 'online', 'live', 'mix', 'messdj', 'soul', 'gold', 'album', 'songs', 'legends', 'ace', 'love', 'dj'] |
| 9 | Social distancing mental health | ['fauci', 'social', 'distancing', 'dr', 'abortion', 'mental', 'depression', 'health', 'smoking', 'rt', 'anthony', 'loneliness', 'individuals', 'trump', 'hanks', 'dementia', 'people', 'distance', 'cncd', 'modelled'] |
| 10 | Drawing art enjoy kids | ['drawing', 'art', 'rtehomeschool', 'vids', 'enjoy', 'mondaymotivaton', 'creative', 'artistontwitter', 'artcompetition', 'picasso', 'happyathome', 'subscribe', 'drawingvideo', 'wereallinthistogether', 'draw', 'drawabcanimals', 'artacademy', 'kids', 'competition', 'potential'] |
| 11 | God bless airtel broadband | ['airtel', 'broadband', 'sir', 'bless', 'stayingconnected', 'broadbandheroes', 'expressvpn', 'abundantly', 'almighty', 'god', 'remain', 'may', 'connectivity', 'generosity', 'data', 'family', 'continue', 'use', 'servers', 'ur'] |
| 12 | Predictive analytics detect infection | ['code', 'ehrs', 'analytics', 'predictive', 'likelihood', 'detect', 'pred', 'nlp', 'uses', 'infection', 'statistics', 'complete', 'applying', 'tested', 'apply', 'wolff', 'proposes', 'date', 'nearly', 'mr'] |
| 13 | Eid stay home safe | ['eid', 'imam', 'dominiccummngs', 'driven', 'eidmubarak', 'stayalert', 'stay', 'changed', 'park', 'staysafe', 'advice', 'alert', 'eidulfitr', 'local', 'wear', 'break', 'miles', 'stayathome', 'prayer', 'long'] |

(*continued on next page*)

**Table E.1** (*continued*).

| | Topic | Representation |
|---|---|---|
| 14 | Automatically followed checked unfollowed | [‘automatically’, ‘checked’, ‘followed’, ‘unfollowed’, ‘people’, ‘person’, ‘one’, ‘unfollow’, ‘madre’, ‘dozens’, ‘mother’, ‘mothersday’, ‘sc’, ‘follow’, ‘techtrouble’, ‘hooyo’, ‘easter’, ‘mutter’, ‘mediapersons’, ‘cashback’] |
| 15 | Weight loss method fast | [‘weight’, ‘try’, ‘method’, ‘loads’, ‘shortened’, ‘really’, ‘fast’, ‘recently’, ‘saw’, ‘easier’, ‘lose’, ‘lost’, ‘link’, ‘staying’, ‘ive’, ‘keep’, ‘want’, ‘make’, ‘obesity’, ‘like’] |
| 16 | Tutoring supplemental reviews help | [‘tutoring’, ‘supplemental’, ‘reviews’, ‘paper’, ‘learn’, ‘ondemand’, ‘help’, ‘available’, ‘caldeira’, ‘shog’, ‘gravitational’, ‘inpatient’, ‘insanely’, ‘incomplete’, ‘remix’, ‘week’, ‘hurricane’, ‘waves’, ‘two’, ‘jealous’] |
| 17 | Court suspends constitution federal | [‘corona’, ‘coronatyranny’, ‘suspends’, ‘constitution’, ‘washington’, ‘court’, ‘federal’, ‘via’, ‘defiant’, ‘compliant’, ‘state’, ‘us’, ‘non’, ‘patient’, ‘leave’, ‘die’, ‘first’, ‘warriors’, ‘free’, ‘live’] |
| 18 | Studied eastern philosophy hind | [‘ghazwaehind’, ‘studied’, ‘philosophy’, ‘eastern’, ‘hind’, ‘mayanmar’, ‘drawbridge’, ‘design’, ‘found’, ‘poland’, ‘hungary’, ‘replaced’, ‘uiux’, ‘slovakia’, ‘appwebsite’, ‘banger’, ‘thailand’, ‘nz’, ‘europe’, ‘interface’] |
| 19 | US America Texas Alabama | [‘us’, ‘america’, ‘texas’, ‘alabama’, ‘whataburger’, ‘uso’, ‘texan’, ‘gucci’, ‘united’, ‘ostrich’, ‘south’, ‘dominican’, ‘texans’, ‘austin’, ‘states’, ‘latin’, ‘sacramento’, ‘rico’, ‘puerto’, ‘caribbean’] |
| 20 | Misidentified remains settlers swords | [‘remains’, ‘misidentified’, ‘swords’, ‘sorting’, ‘settlers’, ‘vikings’, ‘bones’, ‘approx’, ‘originally’, ‘buried’, ‘shields’, ‘discovered’, ‘male’, ‘researchers’, ‘female’, ‘often’, ‘half’, ‘society’, ‘finally’, ‘instead’] |

**Table E.2**
Topics extracted from tweets of hate speech users.

| | Topic | Representation |
|---|---|---|
| 1 | RT China people | [‘rt’, ‘china’, ‘people’, ‘amp’, ‘chinese’, ‘us’, ‘coronavirus’, ‘hong’, ‘world’, ‘kong’, ‘like’, ‘covid’, ‘one’, ‘ccp’, ‘virus’, ‘police’, ‘time’, ‘get’, ‘would’, ‘dont’] |
| 2 | Hong Kong protests | [‘hong’, ‘china’, ‘kong’, ‘rt’, ‘amp’, ‘ccp’, ‘chinese’, ‘hongkong’, ‘people’, ‘coronavirus’, ‘world’, ‘police’, ‘us’, ‘hk’, ‘covid’, ‘law’, ‘taiwan’, ‘one’, ‘new’, ‘beijing’] |
| 3 | Positive comments | [‘good’, ‘thank’, ‘true’, ‘like’, ‘well’, ‘rt’, ‘amen’, ‘happy’, ‘love’, ‘one’, ‘please’, ‘get’, ‘yes’, ‘right’, ‘time’, ‘thanks’, ‘dont’, ‘agree’, ‘fuck’, ‘know’] |
| 4 | US politics | [‘biden’, ‘pelosi’, ‘trump’, ‘democrat’, ‘president’, ‘rt’, ‘people’, ‘black’, ‘blame’, ‘presidential’, ‘white’, ‘joe’, ‘democrats’, ‘fauci’, ‘like’, ‘amp’, ‘impeach’, ‘nancy’, ‘dc’, ‘us’] |
| 5 | Twitter lockdowns | [‘rt’, ‘twitter’, ‘tweet’, ‘lockdown’, ‘tsla’, ‘video’, ‘nnevvy’, ‘sign’, ‘please’, ‘tweets’, ‘petition’, ‘hashtag’, ‘time’, ‘like’, ‘one’, ‘hourlywolves’, ‘retweet’, ‘get’, ‘presents’, ‘tslaq’] |
| 6 | Bill Gates money | [‘gates’, ‘bill’, ‘money’, ‘oil’, ‘toilet’, ‘get’, ‘market’, ‘tax’, ‘people’, ‘paper’, ‘pay’, ‘food’, ‘healthcare’, ‘rt’, ‘time’, ‘need’, ‘health’, ‘us’, ‘even’, ‘dont’] |
| 7 | UK bloggers | [‘england’, ‘bloggers’, ‘usa’, ‘via’, ‘twitter’, ‘democratic’, ‘boris’, ‘partiesbut’, ‘politically’, ‘johnson’, ‘bbc’, ‘independent’, ‘us’, ‘republican’, ‘british’, ‘ireland’, ‘american’, ‘untraditional’, ‘electability’, ‘judgement’] |
| 8 | Food and cooking | [‘nan’, ‘soy’, ‘rice’, ‘microwave’, ‘shanti’, ‘om’, ‘monkey’, ‘ying’, ‘banana’, ‘pandas’, ‘le’, ‘boy’, ‘boos’, ‘chimp’, ‘latte’, ‘chinks’, ‘pregnancy’, ‘bag’, ‘panda’, ‘soyou’] |
| 9 | Book promotion | [‘sanjay’, ‘book’, ‘thriller’, ‘mittal’, ‘read’, ‘story’, ‘written’, ‘life’, ‘battle’, ‘kindle’, ‘futurist’, ‘prophecy’, ‘romantic’, ‘writer’, ‘love’, ‘future’, ‘indian’, ‘link’, ‘geopolitics’, ‘tomorrow’] |
| 10 | Education | [‘school’, ‘schools’, ‘teachers’, ‘python’, ‘fees’, ‘java’, ‘harvard’, ‘find’, ‘class’, ‘enquiryform’, ‘parents’, ‘programming’, ‘furever’, ‘rt’, ‘confucius’, ‘students’, ‘institutes’, ‘chinasponsored’, ‘shutters’, ‘sweden’] |
| 11 | Growth and waves | [‘curve’, ‘growth’, ‘earthquake’, ‘wave’, ‘flattening’, ‘boroughs’, ‘bsmts’, ‘antennas’, ‘meters’, ‘glte’, ‘gs’, ‘lamppost’, ‘interact’, ‘counting’, ‘know’, ‘flatten’, ‘epicenter’, ‘nyc’, ‘previous’, ‘smart’] |
| 12 | Follow and unfollow | [‘unfollowed’, ‘automatically’, ‘checked’, ‘follow’, ‘followers’, ‘standing’, ‘person’, ‘followed’, ‘begining’, ‘snatched’, ‘defining’, ‘one’, ‘vague’, ‘walked’, ‘regulations’, ‘people’, ‘search’, ‘towards’, ‘alone’, ‘body’] |
| 13 | Australia port | [‘darwin’, ‘port’, ‘australia’, ‘triadcolluding’, ‘hatemongering’, ‘victoria’, ‘dna’, ‘permitted’, ‘settle’, ‘au’, ‘exchange’, ‘leased’, ‘scheme’, ‘australian’, ‘castle’, ‘kits’, ‘body’, ‘deal’, ‘pieces’, ‘andrews’] |
| 14 | CEO experiences | [‘ceo’, ‘carolinano’, ‘cruise’, ‘captain’, ‘experienceso’, ‘inexperienced’, ‘ships’, ‘corporation’, ‘hire’, ‘ship’, ‘expected’, ‘buttigieg’, ‘pete’, ‘cruises’, ‘zero’, ‘powerful’, ‘voters’, ‘company’, ‘major’, ‘head’] |
| 15 | American hero | [‘applauded’, ‘american’, ‘made’, ‘company’, ‘heroes’, ‘rosemary’, ‘gibson’, ‘repo’, ‘initiative’, ‘hero’, ‘courage’, ‘praised’, ‘yearold’, ‘loudly’, ‘reporter’, ‘crazy’, ‘child’, ‘view’, ‘arrested’, ‘panky’] |
| 16 | Welded doors | [‘door’, ‘doors’, ‘basement’, ‘welding’, ‘shut’, ‘room’, ‘apartment’, ‘bubbawallacehoax’, ‘building’, ‘weld’, ‘wall’, ‘car’, ‘welded’, ‘mom’, ‘rented’, ‘garage’, ‘noose’, ‘storage’, ‘video’, ‘build’] |
| 17 | Jogger incident | [‘jogger’, ‘joggers’, ‘daytona’, ‘wrestlemania’, ‘nascar’, ‘jogging’, ‘great’, ‘jog’, ‘deserved’, ‘sheeeeit’, ‘hamma’, ‘corvette’, ‘sheeeit’, ‘wwe’, ‘track’, ‘promo’, ‘arbery’, ‘trump’, ‘na’, ‘ahmaud’] |
| 18 | Temperature changes | [‘presaged’, ‘temperatures’, ‘temperature’, ‘sauna’, ‘flare’, ‘scuffles’, ‘recall’, ‘extradition’, ‘legco’, ‘among’, ‘lawmakers’, ‘infrared’, ‘mass’, ‘almost’, ‘exactly’, ‘bill’, ‘ago’, ‘protests’, ‘year’, ‘saunas’] |
| 19 | Unemployment rate | [‘depression’, ‘unemployment’, ‘rate’, ‘encompassed’, ‘great’, ‘brrrrrr’, ‘claims’, ‘quitting’, ‘autumn’, ‘fifty’, ‘depressed’, ‘jobless’, ‘activated’, ‘million’, ‘mid’, ‘filed’, ‘measure’, ‘isolated’, ‘sooner’, ‘haha’] |
| 20 | Redirects and links | [‘aylwards’, ‘redirects’, ‘bruce’, ‘original’, ‘page’, ‘link’, ‘longer’, ‘found’, ‘via’, ‘trumpgt’, ‘ltretweet’, ‘redirect’, ‘harmful’, ‘asset’, ‘relevant’, ‘sister’, ‘amen’, ‘rt’, ‘issues’, ‘’] |

**Table F.3**
Statistics for the hate and control networks obtained from the BERTopic matrices.

| Stratum | Type | Shortest path | Entropy | Clustering coefficient | Density |
|---|---|---|---|---|---|
| Stratum 4 | Hate | **0.023** | **2.307** | **0.85** | **0.592** |
| | Control | 0.025 | 1.383 | 0.814 | 0.53 |
| Stratum 5 | Hate | **0.033** | **2.246** | **0.753** | **0.591** |
| | Control | 0.04 | 2.166 | 0.564 | 0.392 |
| Stratum 7 | Hate | **0.029** | **2.426** | **0.726** | **0.5** |
| | Control | 0.031 | 1.885 | 0.569 | 0.449 |

## Data availability

Data will be made available on request.

## References


Alorainy, W., Burnap, P., Liu, H., Javed, A., & Williams, M. L. (2018). Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample. In *2018 international conference on machine learning and cybernetics* (pp. 581–586). IEEE.

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237.

Anderson, L., & Lepore, E. (2013). Slurring words. *Noûs*, *47*(1), 25–48.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. arXiv preprint URL: https://arxiv.org/pdf/1903.08983, arXiv:1903.08983.

Aziz, N. A. A., Maarof, M. A., & Zainal, A. (2021). Hate speech and offensive language detection: A new feature set with filter-embedded combining feature selection. In *Proceedings of the 2021 3rd international cyber resilience conference* (pp. 29–31). Online: http://dx.doi.org/10.1109/CRC51269.2021.9385951.

Baider, F. (2022). Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. *International Journal for the Semiotics of Law-Revue Internationale de Sémiotique Juridique*, *35*(6), 2347–2371.

Baker-Brown, G., Ballard, E. J., Bluck, S., De Vries, B., Suedfeld, P., & Tetlock, P. E. (1992). The conceptual/integrative complexity scoring manual. In *Motivation and personality: Handbook of thematic content analysis* (pp. 401–418).

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65–74). ACM.

Barberá, P. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Bartlett, J., Reffin, J., Rumball, N., & Williamson, S. (2014). Anti-social media. *Demos*, *2014*, 1–51.

Basile, V. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 54–63). Minneapolis, MN, USA.

Bautista-Ortuño, R., Castro-Toledo, F. J., Perea-García, J. O., & Rodríguez-Gómez, N. (2018). " may I offend you?" An experimental study on perceived offensiveness in online violent communication and hate speech. *International E-Journal of Criminal Sciences*, (12).

Bauwelinck, N., Jacobs, G., Hoste, V., & Lefever, E. (2019). LT3 at SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter (hatEval). In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 436–440). Minneapolis, MN, USA.

Beck, J. S. (2020). *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Bettega, F., Mendelson, M., Leyrat, C., & Bailly, S. (2024). Use and reporting of inverse-probability-of-treatment weighting (IPTW) for multi-category treatments in medical research: a systematic review. *Journal of Clinical Epidemiology*, Article 111338.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*, 3–33.

Boudry, M., & Braeckman, J. (2012). How convenient! The epistemic rationale of self-validating belief systems. *Philosophical Psychology*, *25*(3), 341–364.

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, *38*(3), 551–569.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68.

Brady, W. J., & Crockett, M. (2024). Norm psychology in the digital age: how social media shapes the cultural evolution of normativity. *Perspectives on Psychological Science*, *19*(1), 62–64.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Bressert, E. (2012). *SciPy and NumPy: an overview for developers*. " O'Reilly Media, Inc.".

Bührer, S., Koban, K., & Matthes, J. (2024). The WWW of digital hate perpetration: What, who, and why? A scoping review. *Computers in Human Behavior*, Article 108321.

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 1–15. http://dx.doi.org/10.1140/epjds/s13688-016-0072-6.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Capozzi, A., Lai, M., Basile, V., Poletto, F., Sanguinetti, M., Bosco, C., et al. (2019). Computational linguistics against hate: Hate speech detection and visualization on social media in the 'Contro L'Odio' project. In *CEUR workshop proceedings* (pp. 1–6).

Carter, W. A. (1944). Nicknames and minority groups. *Phylon (1940-1956)*, *5*(3), 241–245.

Casula, P., Anupam, A., & Parvin, N. (2021). "We found no violation!": Twitter's violent threats policy and toxicity in online discourse. In *Proceedings of the 10th international conference on communities & technologies-wicked problems in the age of tech* (pp. 151–159).

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197.

Chakraborty, A., Kumar, R., Gaonkar, B., Reddy, S. B., Ganguly, N., & Gummadi, K. (2018). Taboo or not? Detecting hate speech with the help of social contexts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, *2*(3), 88:1–88:22. http://dx.doi.org/10.1145/3232676, URL: https://dl.acm.org/doi/abs/10.1145/3232676.

Chatzakou, D., Kourtellis, N., Blackburn, J., Stringhini, G., De Cristofaro, E., & Cristofaro, E. D. (2017). Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the international AAAI conference on web and social media* (pp. 13–22).

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118.

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, 1–31.

Chris Hale, W. (2012). Extremism on the World Wide Web: A research review. *Criminal Justice Studies*, *25*(4), 343–356.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 1–10.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Cowell, F. A. (2011). *Measuring inequality*. Oxford University Press.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771.

Cui, Z., Marder, E. P., Click, E. S., Hoekstra, R. M., & Bruce, B. B. (2022). Nearest-neighbors matching for case–control study analyses: Better risk factor identification from a study of sporadic campylobacteriosis in the United States. *Epidemiology*, *33*(5), 633–641.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media* (pp. 512–515).

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2098–2110).

De Gibert, O., Perez, N., Garcia-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd workshop on abusive language online* (pp. 11–20).

Dhont, K., & Hodson, G. (2014). Does lower cognitive ability predict greater prejudice? *Current Directions in Psychological Science*, *23*(6), 454–459.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web* (pp. 29–30). Florence, Italy.

Douglas, K. M. (2021). COVID-19 conspiracy theories. *Group Processes & Intergroup Relations*, *24*(2), 270–275.

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science*, *26*(6), 538–542.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*.

Emine, Ö., Elif, O., Çelik, Y., & Kadir, U. (2024). Youth at work: Exploring the relationship between social status, a history of child labour, and health. *Social Science & Medicine*, Article 117579.

Ernala, S. K., Rizvi, A. F., Birnbaum, M. L., Kane, J. M., & De Choudhury, M. (2017). Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–27.

Esses, V. M., Medianu, S., & Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees. *Journal of Social Issues*, *69*(3), 518–536.

European Commission (2016). European commission: The EU code of conduct on countering illegal hate speech online. Available at coe.int.

Farrelly, D., & Bennett, M. (2018). Empathy leads to increased online charitable behaviour when time is the currency. *Journal of Community & Applied Social Psychology*, *28*(1), 42–46.

Faulkner, N., & Bliuc, A.-M. (2016). 'It's okay to be racist': moral disengagement in online discussions of racist incidents in Australia. *Ethnic and Racial Studies*, *39*(14), 2545–2563.

Faulkner, N., & Bliuc, A.-M. (2018). Breaking down the language of online racism: A comparison of the psychological dimensions of communication in racist, anti-racist, and non-activist groups. *Analyses of Social Issues and Public Policy*, *18*(1), 307–322.

Fearon, P. A., & Boyd-MacMillan, E. M. (2016). Complexity under stress: Integrative approaches to overdetermined vulnerabilities. *Journal of Strategic Security*, *9*(4), 11–31.

Frenda, S., Nicosia, M., Basile, V., Patti, V., Bosco, C., & Fernandez, R. (2020). Hate speech and its relationship with offensive language in social media. In *Lecture notes in computer science*: *vol. 12020*, *Proceedings of the 18th international conference on advances in social networks analysis and mining* (pp. 856–860). Barcelona, Spain: Springer, http://dx.doi.org/10.1007/978-3-030-36687-2_77, URL: https://link.springer.com/chapter/10.1007/978-3-030-36687-2_77.

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., & Wintterlin, F. (2021). Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication*, *9*(1), 195–208.

Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2021). Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media+ Society*, *7*(1), Article 2056305120984445.

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11*(1), 3.

Giner-Sorolla, R., & Russell, P. S. (2019). Not just disgust: Fear and anger also relate to intergroup dehumanization. *Collabra: Psychology*, *5*(1), 56.

Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, 731–742.

Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology*, *94*(2), 292.

Gover, A. R., Harper, S. B., & Langton, L. (2020). Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality. *American Journal of Criminal Justice*, *45*(4), 647–667.

Gregory, A. L., & Piff, P. K. (2021). Finding uncommon ground: Extremist online forum engagement predicts integrative complexity. *Plos One*, *16*(1), Article e0245651.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Grubbs, J. B., Warmke, B., Tosi, J., James, A. S., & Campbell, W. K. (2019). Moral grandstanding in public discourse: Status-seeking motives as a potential explanatory mechanism in predicting conflict. *PLoS One*, *14*(10), Article e0223749.

Gwinn, J. D., Judd, C. M., & Park, B. (2013). Less power=less human? Effects of power differentials on dehumanization. *Journal of Experimental Social Psychology*, *49*(3), 464–470.

Hagberg, A., Swart, P. J., & Schult, D. A. (2008). *Exploring network structure, dynamics, and function using NetworkX*: *Technical Report*, Los Alamos, NM (United States): Los Alamos National Laboratory (LANL).

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*(1), 399–423.

Haybron, D. M. (2002). Moral monsters and saints. *The Monist*, *85*(2), 260–284.

He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 90–94).

Hodson, G., & Busseri, M. A. (2012). Bright minds and dark attitudes: Lower cognitive ability predicts greater prejudice through right-wing ideology and low intergroup contact. *Psychological Science*, *23*(2), 187–195.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, Article 126232.

Jakob, J., Dobbrick, T., & Wessler, H. (2023). The integrative complexity of online user comments across different types of democracy and discussion arenas. *The International Journal of Press/Politics*, *28*(3), 580–600.

Jeshion, R. (2013). Expressivism and the offensiveness of slurs. *Philosophical Perspectives*, *27*(1), 231–259.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881–919.

Kalsnes, B., & Ihlebæk, K. A. (2021). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society*, *43*(2), 326–342.

Kiciman, E., Counts, S., & Gasser, M. (2018). Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Proceedings of the international AAAI conference on web and social media*.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251.

Kim, S., Razi, A., Alsoubai, A., Wisniewski, P. J., & De Choudhury, M. (2024). Assessing the impact of online harassment on youth mental health in private networked spaces. In *Proceedings of the international AAAI conference on web and social media* (pp. 826–838).

Kim, H. S., Sherman, D. K., & Updegraff, J. A. (2016). Fear of Ebola: The influence of collectivism on xenophobic threat responses. *Psychological Science*, *27*(7), 935–944.

Kloo, I., Cruickshank, I. J., & Carley, K. M. (2024). A cross-platform topic analysis of the Nazi narrative on Twitter and telegram during the 2022 Russian invasion of Ukraine. In *Proceedings of the international AAAI conference on web and social media* (pp. 839–850).

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Kteily, N. S., & Bruneau, E. (2017). Darker demons of our nature: The need to (re) focus attention on blatant forms of dehumanization. *Current Directions in Psychological Science*, *26*(6), 487–494.

Kteily, N. S., & Landry, A. P. (2022). Dehumanization: Trends, insights, and challenges. *Trends in Cognitive Sciences*, *26*(3), 222–240.

Leader, T., Mullen, B., & Rice, D. (2009). Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: What puts the" hate" into hate speech? *Journal of Personality and Social Psychology*, *96*(1), 170.

League, A.-D. (2021). *Online hate and harassment. The American experience 2021* (pp. 10–23). New York, NY, USA: Center for Technology and Society.

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Studies in Communication and Media*, *7*(4), 555–579.

Lewandowsky, S., Cook, J., & Ecker, U. K. (2018). Falsified data: A conceptual framework for understanding the persistence of conspiratorial and hate narratives. *Journal of Applied Research in Memory and Cognition*, *7*(4), 314–326.

Lewandowsky, S., Cook, J., & Lloyd, E. (2018). The 'Alice in Wonderland'mechanics of the rejection of (climate) science: simulating coherence by conspiracism. *Synthese*, *195*, 175–196.

Lumsden, K., & Morgan, H. (2017). Media framing of trolling and online abuse: silencing strategies, symbolic violence, and victim blaming. *Feminist Media Studies*, *17*(6), 926–940.

Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. arXiv preprint arXiv:1712.06427.

Markowitz, D. M., Shoots-Reinhard, B., Peters, E., Silverstein, M. C., Goodwin, R., & Bjälkebring, P. (2021). Dehumanization during the COVID-19 pandemic. *Frontiers in Psychology*, *12*, Article 634543.

Markowitz, D. M., & Slovic, P. (2020). Social, psychological, and demographic characteristics of dehumanization toward immigrants. *Proceedings of the National Academy of Sciences*, *117*(17), 9260–9269.

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian conference on intelligent systems* (pp. 61–66). IEEE.

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133.

Mathew, B., Kumar, N., Goyal, P., Mukherjee, A., et al. (2018). Analyzing the hate and counter speech accounts on twitter. arXiv preprint arXiv:1812.02712.

Mathew, B., Rethinam, V., Singh, A., & Mukherjee, A. (2021). Temporal analysis of hate speech on Twitter during crises. In *Proceedings of the international AAAI conference on web and social media*.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., et al. (2019). Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media* (pp. 369–380).

Mathew, B., Saha, P., Yimam, S., Biemann, C., Goyal, P., & Mukherjee, A. (2019). HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 45–54). ACM.

Matos, Y., & Miller, J. L. (2023). The politics of pronouns: how Trump framed the ingroup in the 2016 presidential election. *Politics, Groups, and Identities*, *11*(3), 507–525.

Matsumoto, D., Hwang, H. C., & Frank, M. G. (2016). The effects of incidental anger, contempt, and disgust on hostile language and implicit behaviors. *Journal of Applied Social Psychology*, *46*(8), 437–452.

Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in) coherence as a signature of conspiracy worldviews. *Science Advances*, *8*(43), eabq3668.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153). Montreal, Canada: http://dx.doi.org/10.1145/2872427.2883062.

Noorian, Z., Ghenai, A., Moradisani, H., Zarrinkalam, F., & Alavijeh, S. Z. (2024). User-centric modeling of online hate through the lens of psycholinguistic patterns and behaviors in social media. *IEEE Transactions on Computational Social Systems*.

Olteanu, A., Varol, O., & Kiciman, E. (2017). Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 370–386).

Ombui, E., Muchemi, L., & Wagacha, P. (2019). Hate speech detection in code-switched text messages. In *Proceedings of the 3rd international symposium on multidisciplinary studies and innovative technologies* (pp. 1–6). Ankara, Turkey: http://dx.doi.org/10.1109/ISMSIT.2019.8932841.

Orts, Ò. G. (2019). Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 460–463). Minneapolis, MN, USA.

Ou, G., Zhao, K., Zuo, R., & Wu, J. (2024). Effects of research funding on the academic impact and societal visibility of scientific research. *Journal of Informetrics*, *18*(4), Article 101592.

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180.

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., et al. (2023). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems*, *9*(3), 2827–2842.
Pavetich, M., & Stathi, S. (2021). Investigating antecedents of Islamophobia: The role of perceived control over terrorism, threat, meta-dehumanization, and dehumanization. *Journal of Community & Applied Social Psychology*, *31*(4), 369–382.
Pavlopoulos, J., Sorensen, J., Androutsopoulos, I., & Dixon, L. (2018). Improved abuse comment moderation with transfer learning. In *Lecture notes in computer science*: *vol. 11057*, *Proceedings of the 2nd workshop on abusive language online* (pp. 206–217). Santa Fe, NM, USA: Springer, http://dx.doi.org/10.1007/978-3-319-93417-4_48, URL: https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48.
Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? arXiv preprint arXiv:2006.00998.
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.
Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.
Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, *59*(3), 475.
Perelló, C., Tomás, D., Garcia-Garcia, A., Garcia-Rodriguez, J., & Camacho-Collados, J. (2019). UA at SemEval-2019 task 5: Setting a strong linear baseline for hate speech detection. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 508–513). Minneapolis, MN, USA.
Pollard, J. (2016). Skinhead culture: the ideologies, mythologies, religions and conspiracy theories of racist skinheads. *Patterns of Prejudice*, *50*(4–5), 398–419.
Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741.
van Prooijen, J.-W., Cohen Rodrigues, T., Bunzel, C., Georgescu, O., Komáromy, D., & Krouwel, A. P. (2022). Populist gullibility: Conspiracy theories, news credibility, bullshit receptivity, and paranormal belief. *Political Psychology*, *43*(6), 1061–1079.
Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the international AAAI conference on web and social media* (pp. 557–568).
Repository, H. S. D. (2020). Hate speech data: A repository of hate speech datasets. https://hatespeechdata.com. (Accessed 04 December 2024).
Ribeiro, M. H., & Benevenuto, F. (2022). Hateful users on Twitter. https://github.com/manoelhortaribeiro/HatefulUsersTwitter. (Accessed 04 December 2024).
Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V., & Meira, W., Jr. (2021). Like sheep among wolves: Characterizing hateful users on Twitter. In *Proceedings of the international AAAI conference on web and social media* (pp. 615–622).
Ribeiro, M. H., Santos, Y., Almeida, V., & Meira, W. (2020). Emotional dynamics in toxic online interactions. In *Proceedings of the international AAAI conference on web and social media*.
Ribeiro, A., & Silva, N. (2019). INF-HatEval at SemEval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on Twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 420–425). Minneapolis, MN, USA.
Ronson, J. (2016). *So you've been publicly shamed*. Riverhead Books.
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.
Saha, K., Sugar, B., Torous, J., Abrahao, B., Kıcıman, E., & De Choudhury, M. (2019). A social media study on the effects of psychiatric medication use. In *Proceedings of the international AAAI conference on web and social media* (pp. 440–451).
Saha, K., Weber, I., & De Choudhury, M. (2018). A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *Proceedings of the international AAAI conference on web and social media*.
Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science*, *24*(10), 2069–2078.
Salmela, M., & Von Scheve, C. (2017). Emotional roots of right-wing political populism. *Social Science Information*, *56*(4), 567–595.
Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation* (pp. 2798–2805). Miyazaki, Japan.
Schaller, M., & Neuberg, S. L. (2012). Danger, disease, and the nature of prejudice (s). In *Advances in experimental social psychology*: *vol. 46*, (pp. 1–54). Elsevier.
Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. *SciPy*, *7*(1).
Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073–15078.
Shih, M. J., Stotzer, R., & Gutiérrez, A. S. (2013). Perspective-taking and empathy: Generalizing the reduction of group bias towards Asian Americans to general outgroups. *Asian American Journal of Psychology*, *4*(2), 79.
Solovev, K., & Pröllochs, N. (2023). Moralized language predicts hate speech on social media. *PNAS Nexus*, *2*(1), pgac281.
Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146.
Stephan, W. G., Diaz-Loving, R., & Duran, A. (2000). Integrated threat theory and intercultural attitudes: Mexico and the United States. *Journal of Cross-Cultural Psychology*, *31*(2), 240–249.
Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining*, *3*(4), 1277–1291.
Strathern, W., Schoenfeld, M., Ghawi, R., & Pfeffer, J. (2020). Against the others! Detecting moral outrage in social media networks. In *2020 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 322–326). IEEE.
Suedfeld, P., Leighton, D. C., & Conway III, L. G. (2006). Integrative complexity and cognitive management in international confrontations: Research and potential applications.
Suedfeld, P., & Tetlock, P. (1977). Integrative complexity of communications in international crises. *Journal of Conflict Resolution*, *21*(1), 169–184.
Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity.
Surge AI (2023). Profanity detection dataset. URL: https://github.com/surge-ai/profanity. GitHub repository.
Swami, V., Barron, D., Weis, L., & Furnham, A. (2018). To B rexit or not to B rexit: The roles of I slamophobia, conspiracist beliefs, and integrated threat in voting intentions for the U nited K ingdom E uropean U nion membership referendum. *British Journal of Psychology*, *109*(1), 156–179.
Swami, V., Chamorro-Premuzic, T., & Furnham, A. (2010). Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Applied Cognitive Psychology*, *24*(6), 749–761.
Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition*, *133*(3), 572–585.
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.
Teh, P. L., & Cheng, C.-B. (2020). Profanity and hate speech detection. *International Journal of Information and Management Sciences*, *31*(3), 227–246.
Tellez, E., Moctezuma, D., Miranda-Jiménez, S., & Graff, M. (2018). An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems*, *149*, 110–123. http://dx.doi.org/10.1016/j.knosys.2018.03.013.
Tennen, H., & Affleck, G. (1990). Blaming others for threatening events. *Psychological Bulletin*, *108*(2), 209.
Thurlow, C. (2001). Naming the "outsider within": Homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. *Journal of Adolescence*, *24*(1), 25–38.
Tong, S. T., & DeAndrea, D. C. (2023). The effects of observer expectations on judgments of anti-Asian hate tweets and online activism response. *Social Media+ Society*, *9*(1), Article 20563051231157299.
Vallée, R. (2014). Slurring and common knowledge of ordinary language. *Journal of Pragmatics*, *61*, 78–90.

Van Prooijen, J.-W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on Psychological Science*, *13*(6), 770–788.

Vega, L., Reyes-Magaña, J., Gómez-Adorno, H., & Bel-Enguix, G. (2019). MineriaUNAM at SemEval-2019 task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 447–452). Minneapolis, MN, USA.

Vergani, M., Martinez Arranz, A., Scrivens, R., & Orellana, L. (2022). Hate speech in a telegram conspiracy channel during the first year of the COVID-19 pandemic. *Social Media+ Society*, *8*(4), Article 20563051221138758.

Verma, G., Bhardwaj, A., Aledavood, T., De Choudhury, M., & Kumar, S. (2022). Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports*, *12*(1), 8045.

Vicente, P. (2023). Sampling Twitter users for social science research: evidence from a systematic review of the literature. *Quality & Quantity*, *57*(6), 5449–5489.

Walther, J. B. (2022). Social media and online hate. *Current Opinion in Psychology*, *45*, Article 101298.

Wang, X., Zhang, M., Fan, W., & Zhao, K. (2022). Understanding the spread of COVID-19 misinformation on social media: The effects of topics and a political leader's nudge. *Journal of the Association for Information Science and Technology*, *73*(5), 726–737.

Wang, R., Zhou, A., & Kinneer, T. H. (2024). Moral framing and issue-based framing of# StopAsianHate campaigns on Twitter. *Chinese Journal of Communication*, *17*(1), 42–60.

Waseem, Z., & Hovy, D. (2016a). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93). San Diego, CA, USA: http://dx.doi.org/10.18653/v1/N16-2013.

Waseem, Z., & Hovy, D. (2016b). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93). Association for Computational Linguistics.

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.

Williams, H. T., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science*, *3*(6), 767–773.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., et al. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. arXiv preprint arXiv:2303.10420.

Yin, C., & Zhang, Z. (2024). A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *2024 2nd international conference on image, algorithms and artificial intelligence* (pp. 677–684). Atlantis Press.

Young, D. G., & Young, D. G. (2020). *Irony and outrage: The polarized landscape of rage, fear, and laughter in the United States*. USA: Oxford University Press.

Zahrah, F., Nurse, J. R., & Goldsmith, M. (2022). A comparison of online hate on reddit and 4chan: a case study of the 2020 us election. In *Proceedings of the 37th ACM/SIGAPP symposium on applied computing* (pp. 1797–1800).

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., et al. (2018). On the origins of memes by means of fringe web communities. In *Proceedings of the internet measurement conference* (pp. 188–202). ACM.

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the Web. In *Proceedings of the 2019 world wide web conference* (pp. 2181–2191). ACM.

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM conference on web science* (pp. 125–134).