Computational Approaches to Enhance the Integrity of Social Media: From Detection to Intervention

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht von
Dominik Bär

München, 28.05.2025

Computational Approaches to Enhance the Integrity of Social Media: From Detection to Intervention

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht von
Dominik Bär

München, 28.05.2025

1. Berichterstatter:	Prof. Dr. Stefan Feuerriegel
2. Berichterstatter:	Prof. Dr. Nicolas Pröllochs
3. Berichterstatter:	Prof. Dr. Niklas Kühl
Tag der mündlichen Prüfung:	28.07.2025

Acknowledgments

This dissertation seeks to advance our understanding of integrity and accountability in digital platforms—a topic that sits at the intersection of computer science, social science, and public interest. Tackling such complex questions would not have been possible without the support and collaboration of many people to whom I am deeply grateful.

First and foremost, I thank Stefan Feuerriegel, my doctoral advisor, for his outstanding mentorship and unwavering support throughout this journey. His guidance and trust gave me the freedom to explore ambitious questions, while his insightful feedback pushed me to refine my thinking and sharpen my contributions.

I am also sincerely grateful to Nicolas Pröllochs, a trusted collaborator and co-author, for the many enriching discussions and his generous engagement with my work. I thank Niklas Kühl for serving as a referee and for his thoughtful evaluation of this dissertation. I further thank Sven Mayer for chairing my defense and Barbara Plank for kindly joining the committee as a substitute examiner.

My thanks extend to my colleagues and co-authors at LMU and MCML, especially at the Institute of AI in Management, for their camaraderie, thoughtful feedback, and daily inspiration. I am equally grateful to my collaborators at other institutions, whose openness and insight have broadened this work and made collaboration a true pleasure. In particular, I would like to thank Abdurahman Maarouf, Francesco Pierri, and Gianmarco De Francisci Morales, with whom I collaborated on essential parts of this dissertation.

I also acknowledge the academic mentors during my Bachelor's and Master's studies, who encouraged me early on and shaped my path toward research.

Finally, my heartfelt thanks go to my family. To my parents, Maria and Michael, for their love and constant encouragement. To my siblings, Matthias and Isabel, and to Kathrin, for your constant encouragement and support throughout this journey. And to my friends and extended family—thank you for your belief in me and your support through every stage of my academic journey.

Abstract

Social media has profoundly transformed society. However, the proliferation of harmful content — such as hate speech, misinformation, and conspiracy theories — challenges platform integrity and poses a significant threat to society. Enhancing the integrity of social media is thus of utmost importance. In this dissertation, we propose a computational approach to improve the integrity of social media along three key dimensions: (1) the detection and understanding of harmful content, (2) audits of social media platforms, and (3) interventions to counter abusive behavior. To achieve this, we combine state-of-the-art methods from computer science with insights from the social sciences and explore each dimension through distinct case studies in three parts.

The first part focuses on detecting and understanding harmful content. Specifically, we employ a machine learning approach to identify QAnon conspiracy theorists on Parler and profile their characteristics compared to other users. We then statistically analyze the diffusion dynamics of online rumors to gain deeper insights into how harmful content spreads on social media.

The second part shifts attention to auditing social media platforms. Here, we examine how proprietary algorithms, often beyond societal control, can perpetuate biases. As a case study, we analyze the delivery of political ads on Meta during the 2021 German Federal Election.

The third part explores interventions aimed at countering harmful content and fostering civil behavior on social media. This includes results from a large-scale, pre-registered field experiment evaluating the effectiveness of AI-generated counterspeech, employing large language models, in reducing online hate speech.

By combining computational methodologies with interdisciplinary perspectives, this dissertation advances the understanding of social media vulnerabilities and delivers actionable solutions for cultivating safer digital environments. Scientifically, it demonstrates the practical application of machine learning, natural language processing, and social media analytics in detecting harmful content, auditing opaque algorithms, and evaluating scalable interventions. These contributions extend computational social science literature and inform strategies for algorithmic accountability and fair content delivery. On a societal level, the results emphasize the importance of transparency in social media platforms and highlight both the potential and risks of automated interventions, such as AI-generated counterspeech, in curbing online harms. Overall, this dissertation provides a roadmap for platform providers and policymakers committed to promoting equity, inclusivity, and democratic values within social media ecosystems.

Zusammenfassung

Soziale Medien haben unsere Gesellschaft tiefgreifend verändert. Die Verbreitung schädlicher Inhalte — wie Hassrede, Fehlinformationen und Verschwörungstheorien — stellt jedoch eine große Herausforderung für die Integrität von Plattformen dar und bedroht unsere Gesellschaft. Die Verbesserung der Integrität sozialer Medien ist daher von höchster Bedeutung. In dieser Dissertation schlagen wir einen computergestützten Ansatz vor, um die Integrität sozialer Medien in drei wesentlichen Bereichen zu stärken: (1) die Erkennung und das Verständnis schädlicher Inhalte, (2) Audits von Social-Media-Plattformen und (3) Interventionen zur Bekämpfung von schädlichem Verhalten. Um dies zu erreichen, kombinieren wir modernste Methoden der Informatik mit Erkenntnissen aus den Sozialwissenschaften und untersuchen jeden dieser Bereiche in separaten Fallstudien, die in drei Teilen dargestellt werden.

Der erste Teil konzentriert sich auf die Erkennung und das Verständnis schädlicher Inhalte. Konkret verwenden wir maschinelles Lernen, um QAnon-Verschwörungstheorethiker auf der Plattform Parler zu identifizieren und deren Nutzungsverhalten im Vergleich zu "normalen" Nutzern zu analysieren. Anschließend analysieren wir statistisch die Dynamik der Verbreitung von Online-Gerüchten, um tiefere Einblicke in die Verbreitung schädlicher Inhalte in sozialen Medien zu gewinnen.

Der zweite Teil beschäftigt sich mit Audits von Social-Media-Plattformen. Hier untersuchen wir, wie proprietäre Algorithmen, die sich oft der gesellschaftlichen Kontrolle entziehen, Verzerrungen verstärken können. Als Fallstudie analysieren wir die Verbreitung politischer Werbung auf Meta während der deutschen Bundestagswahl 2021.

Der dritte Teil befasst sich mit Interventionen, um schädlichen Inhalten entgegenzuwirken und respektvolles Verhalten in sozialen Medien zu fördern. Dazu erörtern wir die Ergebnisse eines groß angelegten, präregistrierten Feldexperiments, in dem die Wirksamkeit von KI-generierter Gegenrede (Counterspeech) unter Verwendung großer Sprachmodelle (Large Language Models) bei der Reduzierung von Online-Hassrede untersucht wird.

Durch die Kombination von Methoden aus der Informatik mit interdisziplinären Ansätzen erweitert diese Dissertation unser Verständnis für Schwachstellen von sozialen Medien und präsentiert Lösungen für die Schaffung eines sichereren digitalen Umfelds. Auf wissenschaftlicher Ebene demonstriert diese Arbeit die praktische Anwendung von maschinellem Lernen, natürlicher Sprachverarbeitung (Natural Language Processing) und Social-Media-Analysen zur Erkennung schädlicher Inhalte, der Prüfung von Algorithmen und der Bewertung skalierbarer Interventionen. Diese Beiträge erweitern die Computational Social Science Literatur und liefern wertvolle Erkenntnisse zum verantwortungsvollen Umgang mit Algorithmen und fairer Verbreitung von Inhalten auf Sozialen Medien. Auf gesellschaftlicher Ebene unterstreichen die Ergebnisse die Bedeutung von Transparenz auf Social-Media-Plattformen und verdeutlichen sowohl das Potenzial als auch die Risiken automatisierter Interventionen, wie z.B. KI-generierte Gegenrede, zur Eindämmung von Online-Hassrede. Insgesamt liefert diese Dissertation damit konkrete Vorschläge für Plattformbetreiber und politische Entscheidungsträger zur Förderung von Gerechtigkeit, Inklusivität und demokratischen Werten in sozialen Medien.

Contents

Introduction

1 Introduction

Social media has profoundly transformed society. Globally, there are around 5.2 billion social media users, which amounts to 64 % of the world's population [1]. These users spend, on average, more than two hours a day on platforms such as Facebook, Instagram, and X (formerly Twitter) [2]. Consequently, social media exerts a substantial influence on individuals and society at large [3–6].

Social media has brought numerous societal benefits. Social media has significantly enhanced global connectedness and enabled rapid communication across geographic and cultural boundaries [7]. These capabilities have fostered both social and political change, as digital platforms provide new channels for civic engagement and collective action [7]. For example, social media has played a pivotal role in organizing political protests [3, 8] and increasing political participation [9]. Moreover, social media has been leveraged to raise awareness about climate change [5], tackle gender inequality [4], and promote public health measures during the COVID-19 pandemic [6]. As such, social media has positively impacted the offline world across a wide range of domains.

However, there are also concerns about the negative impact of social media on society. In particular, the proliferation of harmful content, such as hate speech, misinformation, and conspiracy theories, challenges platform integrity and poses a significant threat to society [10]. On an individual level, exposure to harmful content may cause detrimental consequences for individuals' mental and physical well-being [11–16]. On a societal level, harmful content may undermine democratic institutions [10, 17, 18], increase polarization and hostility between groups [19–24], or even lead to real-world violence [21–25]. Overall, enhancing the integrity of social media is thus of utmost importance for society.

In response to the proliferation of harmful content, platform owners, NGOs, and policymakers have launched initiatives to enhance the integrity of social media. Platforms have tackled harmful content by combining manual and automated content moderation with features encouraging users to flag and report content [10, 26, 27]. NGOs support these efforts by promoting human rights on social media and supporting victims of harmful content [28, 29]. Lastly, policymakers have established regulatory frameworks that hold platforms accountable for hosting harmful content (e.g., the Digital Services Act in the European Union) [30–32].

In addition to platform- and policy-driven initiatives, academic research plays a vital role in enhancing the integrity of social media. Prior work has contributed by documenting the prevalence and consequences of harmful content [33–38], auditing opaque algorithms and moderation practices [39–46], and evaluating interventions to mitigate online harms

[21, 47–51]. Building on this foundation, this dissertation adopts a holistic computational approach to improve platform integrity across three key dimensions: (1) the detection and understanding of harmful content, (2) audits of social media platforms, and (3) interventions to counter harmful behavior.

This dissertation contributes to each of these dimensions by proposing a holistic computational approach to enhance the integrity of social media. Specifically, we adopt an interdisciplinary framework combining state-of-the-art methods from computer science, such as machine learning and natural language processing, with theoretical and empirical insights from the social sciences. This combination enables both scalable analysis of online behavior and a nuanced understanding of its societal implications. We demonstrate the utility of this approach across five case studies, each addressing a core challenge in the detection and understanding, auditing, and mitigation of harmful content on social media.

The first part of this dissertation focuses on the detection and understanding of harmful content. Accurate detection is a prerequisite for analyzing key dimensions of harmful content, such as its spreading dynamics or offline impact, and for developing effective countermeasures [52]. In this context, we review state-of-the-art computational methods that enable the detection of harmful content at scale. We also discuss how we can reliably identify users who have frequently shared harmful content (see Section 3). As a practical example, we present findings from a case study on detecting QAnon conspiracy theorists on the fringe social media platform Parler (see Section 7).

Once harmful content can be reliably identified, the next step is understanding its dynamics and consequences. We will thus continue with a review on how such content spreads on social media and discuss its individual and societal effects online and offline. In this context, we present two case studies: The first case study analyzes how emotional language affects the spread of online rumors (Section 8). The second case study examines differences in the diffusion patterns between true and false rumors (Section 9).

The second part of this dissertation turns to auditing social media platforms. As content delivery and moderation processes are largely opaque and governed by proprietary algorithms, biases can distort content exposure or even amplify harmful content [41–43, 53–55]. Auditing is, therefore, essential for promoting transparency, ensuring fair content delivery, and holding platforms accountable for their societal impact [44, 45, 56]. Although third-party access to platform data remains limited, increasing public and regulatory pressure is opening new avenues for independent research to hold platform owners accountable for algorithmic decisions that shape user experiences and public discourse (see Section 4). In this context, we present a case study examining how the algorithmic delivery of political advertisements on Meta during the 2021 German Federal Election may affect political campaigns, highlighting both the challenges and the necessity of auditing in democratic societies (see Section 10).

The third part addresses interventions to counter harmful behavior. These can be top-down (platform-enforced policies such as content bans or changes to feed algorithms) or bottom-up (community-based mechanisms such as user flagging and peer feedback) strategies to moderate user-generated content. We discuss how top-down strategies offer consistency and scalability but may provoke backlash or be perceived as censorship [57, 58]. Moreover, we analyze bottom-up strategies that promote user engagement and cultural sensitivity but may lack reliability and may even reinforce existing biases [57, 59] (see Section 5). As a case study, we evaluate the effectiveness of counterspeech generated by state-of-the-art large language models (LLMs) to reduce online hate speech (see Section 11).

Overall, this dissertation offers a holistic approach to enhancing the integrity of social media (see Fig. 1 for an overview). By combining computational methods with interdisciplinary perspectives, it advances our understanding of platform vulnerabilities and provides actionable solutions for creating safer digital environments. Scientifically, the work demonstrates the application of machine learning, natural language processing, and social media analytics in detecting harmful content, auditing opaque algorithms, and testing scalable interventions. These contributions extend the computational social science literature and inform strategies for algorithmic accountability and equitable content delivery. Societally, the findings underscore the importance of transparency and highlight both the potential and limitations of automated interventions in curbing online harm. Overall, this work provides a roadmap for platform owners and policymakers seeking to foster inclusive, democratic, and equitable social media ecosystems.

This thesis is structured around a general introduction (see Chapter 1) that outlines the broader scientific context and methodological approaches to enhancing the integrity of social media, followed by five case studies corresponding to each part of the proposed framework (see Chapter 2 to Chapter 4): detection and understanding (see Section 7 to Section 9), auditing (see Section 10), and intervention (see Section 11).

In the remainder of this introduction, we first review the concept of harmful content and its individual and societal impact (see Section 2). We then discuss computational approaches to detecting and understanding harmful content (see Section 3), followed by a review of research focused on auditing social media platforms (see Section 5) and developing interventions to mitigate online harms (see Section 4). Finally, we conclude with a discussion (see Section 6) of the dissertation's contributions, broader implications, and future directions for platform owners, policymakers, and researchers aiming to enhance the integrity of social media.

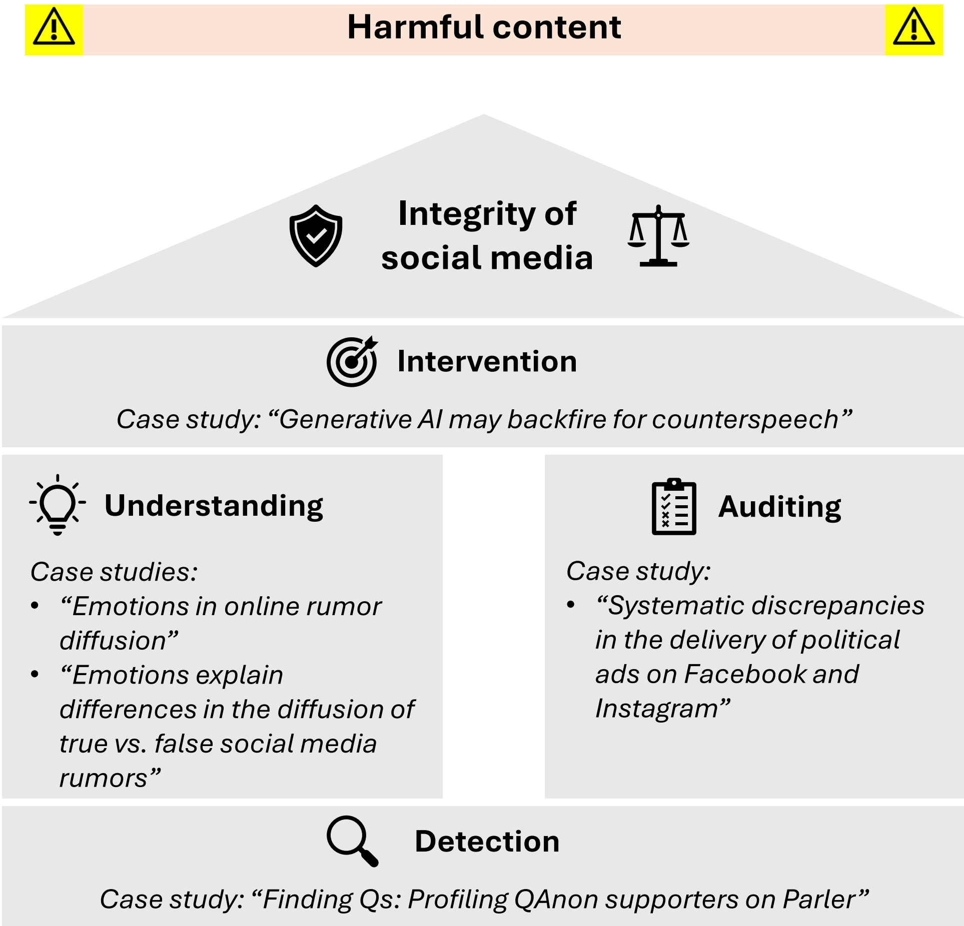

Figure 1: **Computational approaches to enhance the integrity of social media:** This dissertation presents a holistic computational approach aimed at improving the integrity of social media along three key dimensions: (1) detection and understanding of harmful content, (2) audits of social media platforms, and (3) interventions to counter harmful behavior. We illustrate each part with distinct case studies. Detecting harmful content serves as the foundation and prerequisite for subsequent analyses. Understanding harmful content dives into its dynamics and consequences, while auditing examines the role of social media platforms and their content delivery mechanisms. Insights from both understanding and auditing inform targeted interventions to mitigate harmful content. Together, this framework strengthens the integrity of social media platforms.

2 Harmful content on social media and its individual and societal effects

We define harmful content on social media as any type of content, including posts, replies, or media, that can cause harm to individuals or society as a whole. Prominent examples of harmful content include hate speech, misinformation, conspiracy theories, or extremist content [22, 30, 60–64]. In recent years, the spread of harmful content has increased tremendously [65–67], especially through social media platforms and algorithmic content delivery [39, 68–70]. This growing prevalence has led to heightened concerns regarding the individual and societal consequences of exposure to such content, prompting research, policy discussions, and public debates on how best to mitigate its impact [30–32, 60–65, 71].

Harmful content on social media can impact individuals in multiple ways. Initially, exposure to harmful content often triggers immediate psychological and emotional reactions, such as distress, anxiety, feelings of offense, or perceived discrimination [11, 12, 15]. Persistent exposure to harmful content can escalate these emotional responses into long-term mental health conditions, including depression, anxiety disorders, or even physical health consequences (e.g., incorrect medical decisions based on misinformation) [11–16, 72]. These severe health impacts frequently result in behavioral changes. For example, users may withdraw from online interactions, engage in self-censorship, or become susceptible to radicalization and extremism [13, 19, 25]. Addressing and mitigating harmful content is thus crucial, not only for individual well-being but also to preserve societal cohesion.

At the societal level, harmful content on social media undermines the stability of democratic societies [10, 19, 20, 22, 73–75]. Exposure to harmful content, such as misinformation, conspiracy theories, and propaganda, erodes trust in crucial societal institutions, including the media, scientific communities, and government bodies [10, 18, 64, 75, 76]. For example, the spread of misinformation during the COVID-19 pandemic played a central role in vaccine hesitancy and undermined compliance with public health measures [14, 16, 77]. Moreover, harmful content contributes to the increasing polarization of political discourse [23, 24, 78, 79], particularly by amplifying uncivil content and reinforcing echo chambers through algorithmic curation [38, 46, 69]. Such polarization impedes constructive debate and threatens the cohesion of democratic systems [20, 21, 74, 80]. In extreme cases, the consequences of harmful content on social media can even lead to real-world violence [19], as seen in the Christchurch mosque shooting in 2019 and the storming of the U.S. Capitol on January 6, 2021 [73, 81, 82]. Therefore, addressing harmful content is vital not only to protect public health and safety but also to preserve democratic integrity and societal cohesion.

In this dissertation, we contribute to the growing body of literature by focusing on three key categories of harmful content: (1) online hate speech, (2) misinformation, disinformation, and propaganda, and (3) conspiracy theories. Each of these content types presents unique challenges and implications for both individuals and society. In the following sections, we provide a detailed review of each category, exploring their individual and societal effects, and discussing efforts by policymakers, platform owners, and researchers to mitigate their impact.

2.1 Online hate speech

The United Nations defines hate speech as "any kind of communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are – in other words, based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity factor" [65]. The prevalence of online hate speech has risen considerably in recent years [65, 83]. Addressing and mitigating online hate speech is, therefore, a critical step toward enhancing the integrity of social media and preserving democratic values in digital spaces.

Online hate speech poses a serious threat to both individual well-being and broader societal cohesion [11–13, 15, 20, 21]. Individuals who are exposed to online hate speech frequently experience psychological consequences that adversely affect their mental and physical health [11–13, 15, 72]. For instance, encountering hate speech can foster feelings of insecurity [15]. Among adolescents, experiences of online discrimination have been associated with elevated levels of stress, as well as higher rates of depression and anxiety [11, 12]. Moreover, efforts to speak out against hate speech are often suppressed by both social media algorithms and user behavior, further exacerbating social isolation and impeding the formation of support networks [13]. Exposure to hate speech also affects interpersonal behavior by diminishing individuals' capacity for empathy and their ability to recognize and respond to others' suffering [72]. Collectively, these findings underscore the impact of online hate speech on individuals, which may eventually lead to societal consequences.

Online hate speech also poses substantial risks to societal cohesion. Deterioration of individuals' capacity for empathy, resulting from repeated exposure to hate speech, can negatively impact prosocial behavior and weaken solidarity [72]. More broadly, hate speech propagates additional hatred [84], fostering hostility between social groups [20, 21]. For instance, ongoing geopolitical conflicts, such as those in the Middle East, have been associated with sharp increases in antisemitic and anti-Muslim hate speech on social media [83, 85]. Consequently, online hate speech contributes to political and social polarization [21] and, in extreme cases, can even motivate real-world violence [20]. Examples include the 2018 Pittsburgh synagogue shooting [86], the 2019 Christchurch mosque shooting [81], and the role of Facebook in exacerbating violence during the 2017 Rohingya genocide in Myanmar [87]. Countering the spread of online hate speech is therefore essential for preserving social cohesion, maintaining democratic discourse, and preventing the escalation of online hostility into real-world violence.

Regulators, platform owners, and researchers have implemented various measures to counter online hate speech. Nevertheless, according to the United Nations, existing legislative frameworks for prosecuting hate speech often remain insufficient [31]. In addition, social media companies have shown only limited commitment to effectively addressing hate speech on their platforms [31]. In response, researchers have studied multiple aspects of online hate speech, including its virality [88], the characteristics of users disseminating hateful content [36, 89], and the development of automated detection systems [36, 90, 91]. Furthermore, considerable efforts have been devoted to identifying strategies to curb the spread of hate speech, with counterspeech emerging as a promising intervention [21, 47–49]. Counterspeech involves replying to hateful content to challenge harmful narratives and encourage more civil behavior among users [92]. Building on this concept, we later propose a scalable framework that generates contextualized counterspeech using large language models (LLMs) and evaluate its effectiveness through a large-scale, pre-registered field experiment (see Section 11).

2.2 Misinformation, disinformation, and propaganda on social media

Misinformation generally refers to the unintentional dissemination of false or misleading information [60, 66]. In contrast, disinformation involves the deliberate spread of false information with the intent to deceive or cause harm to individuals or society [60]. Thus, the distinction between the two lies in the element of intent. Propaganda is also typically intentional and is used to influence, persuade, or manipulate public opinion [93]. Unlike disinformation, propaganda does not necessarily rely on falsehoods but often presents information selectively or in a biased manner to promote a specific agenda [19].

Despite these conceptual differences, misinformation, disinformation, and propaganda are all considered harmful. In fact, the individual and societal consequences are often comparable, as negative consequences can arise regardless of the sender's motivation. Generally, misinformation, disinformation, and propaganda can threaten individuals' health, undermine public trust, distort democratic discourse, and contribute to social division [14, 16, 19, 34, 60, 66, 67, 88, 94]. Researchers also often use the term "online rumor" to emphasize the unverified and potentially misleading nature of claims disseminated on social media [37]. Safeguarding social media platforms against the spread of such content is therefore essential to protecting democratic values and maintaining a well-informed public.

Misinformation, disinformation, and propaganda can adversely affect individuals' health, beliefs, and behaviors. For instance, the World Health Organization (WHO) has emphasized that the widespread dissemination of misleading content can pose serious risks to individual health [14]. During the COVID-19 pandemic, misinformation was directly associated with non-compliance with public health measures and vaccine hesitancy, both of which contributed to preventable health complications and increased mortality [16]. Beyond health, propaganda can influence individuals' beliefs, attitudes, and behaviors by appealing to emotion, exploiting existing biases, and eroding trust in credible sources [19]. More broadly, repeated exposure to misleading or manipulative content distorts information processing and weakens the perceived credibility of factual information [95–97]. Even individuals with high levels of knowledge and media literacy may struggle to distinguish facts from falsehoods when repeatedly exposed to false or emotionally charged narratives

[98]. These dynamics illustrate how misinformation, disinformation, and propaganda can harm individual decision-making, with potentially far-reaching societal consequences when such harms scale across populations.

The societal effects of misinformation, disinformation, and propaganda are far-reaching and detrimental [10, 19, 75, 99]. These forms of harmful content affect critical domains such as sustainability [75, 97, 100], public health [10, 14, 16], and public safety [19, 25]. For example, climate misinformation undermines public understanding of environmental challenges and hampers collective action toward sustainability goals [97, 100, 101]. Similarly, misinformation related to public health reduces compliance with preventive measures and facilitates the spread of disease [16]. In the realm of public safety, misinformation and propaganda have been linked to radicalization and the incitement of violence [19, 25].

Moreover, misinformation, disinformation, and propaganda is frequently deployed to manipulate public opinion and interfere with democratic processes [62], such as during the 2016 U.S. presidential election [94] or in shaping public sentiment during the 2022 Russian invasion of Ukraine [34, 88]. These dynamics are especially concerning because structural inequalities not only facilitate the spread of misinformation but also amplify its disproportionate impact, often resulting in severe risks for disadvantaged communities [102]. Finally, the erosion of public trust and the amplification of polarization threaten the foundations of democracies [10]. These risks underscore the urgent need for coordinated efforts by policymakers, platform owners, and researchers to safeguard the integrity of social media.

A wide range of measures to counter misinformation, disinformation, and propaganda has been proposed and implemented by policymakers, social media platforms, and researchers. In the European Union, for example, the Digital Services Act mandates that large online platforms such as Facebook and Instagram take action to mitigate the spread of false or misleading content [32]. In response, platforms have implemented interventions such as human and automated content moderation, where problematic posts are downranked or deleted [26, 103]. Features that encourage users to flag or verify content complement these efforts [10, 26]. For instance, X (formerly Twitter) allows users to collaboratively add context to misleading posts through its "Community Notes" feature [27].

Researchers have extensively studied the effectiveness of various interventions against misinformation, disinformation, and propaganda. Prominent examples include (1) accuracy nudges, which nudge users to consider the reliability of information before posting [51, 104–108]; (2) media literacy tips, which aim to educate users about how to recognize misinformation [50, 107–109]; (3) debunking or rebuttal strategies, which correct false claims by providing factual information [76, 107, 109, 110]; and (4) deplatforming of users that spread misinformation [73, 111]. In addition, scholars have evaluated the prevalence and reach of misinformation and propaganda [34, 94, 112–116], as well as the mechanisms by which such content spreads on social media [37, 38, 46, 117, 118]. Understanding the diffusion dynamics is particularly important since it helps to develop effective countermeasures. This dissertation contributes to this literature by examining how emotions affect the spread of online rumors (see Section 8 and Section 9).

2.3 Conspiracy theories on social media

A conspiracy theory is defined as an explanation for societal events that alleges the involvement of a powerful and malevolent group acting in secret [23, 24]. Conspiracy theories are often propagated on social media [22, 119–125] and, unlike healthy skepticism toward governments and elites, belief in conspiracies can pose serious threats to democracy and public safety [23, 24]. For example, belief in conspiracy theories has been linked to negative public health outcomes [125, 126], political extremism [24], racism [23], and even violence [127]. Overall, this highlights the harmful nature of conspiracy theories and the importance of limiting their spread on social media.

Conspiracy theories can negatively affect individuals at psychological, behavioral, and social levels. Psychologically, belief in conspiracy theories is associated with heightened feelings of powerlessness and a diminished sense of control and autonomy over one's life [18]. Individuals who endorse conspiracy theories also tend to exhibit higher levels of distrust, not only in governmental institutions and science, but also in interpersonal relationships [17, 18]. These beliefs are often rooted in cognitive biases, such as the tendency to perceive agency in random events or to overestimate the likelihood of rare and extreme outcomes [18]. Behaviorally, conspiracy beliefs are linked to lower civic engagement and reduced participation in democratic processes, such as voting [18]. This disengagement often originates from a belief that institutional mechanisms are fundamentally corrupt or manipulated. Such beliefs can also have severe health risks: For example, conspiracy beliefs led to vaccine hesitancy during the COVID-19 pandemic and jeopardized individual health outcomes [77]. Socially, belief in conspiracy theories can foster alienation and social withdrawal. Individuals who endorse such beliefs may disengage from mainstream communities and face stigmatization, particularly when their views are perceived as extreme or irrational [18]. This social exclusion can, in turn, reinforce reliance on alternative communities and contribute to a cycle of isolation and radicalization [18, 22]. Overall, the psychological, behavioral, and social consequences of conspiracy beliefs are not only concerning at the individual level but may also scale up to contribute to broader societal challenges.

Conspiracy theories can have severe societal consequences. From a public health perspective, beliefs in conspiracy theories are often associated with non-compliance with collective health measures [126]. For example, conspiracy beliefs fueled vaccine hesitancy and thereby contributed to the spread of COVID-19 [126]. Similarly, exposure to conspiracy theories has been shown to undermine public support for sustainability policies [128]. From a democratic perspective, conspiracy beliefs reduce political participation [18, 128] and erode trust in democratic institutions [17, 129], thereby facilitating the rise of anti-democratic movements. Conspiracy theories also provide fertile ground for political extremism and foster societal polarization [23, 24, 78, 79]. In addition, they often promote hostility, racism, and radicalization [22–24]. In extreme cases, conspiracy theories may even incite real-world violence [22–24]. For example, the "replacement theory," which claims that a secret elite seeks to replace white Americans with non-white populations, was a motivating factor for the perpetrator in the 2022 Buffalo shooting, in which ten people were killed [22]. Similarly, conspiracy theories alleging widespread election fraud during the 2020 U.S. presidential election contributed to the storming of the U.S. Capitol on January 6, 2021, an unprecedented attack on American democracy [73, 82]. Given that social media often facilitates the spread of conspiracy theories [22, 119–125], it is crucial to understand how these platforms can be effectively safeguarded against the influence of conspiratorial content.

Mitigating conspiracy theories poses a particular challenge. In contrast to misinformation, conspiracy theories often embed false information within broader ideological narratives [130]. For example, the QAnon conspiracy theory, which alleges that a secret elite controls global politics, operates a child exploitation network, and seeks to undermine societal order, is considered a "meta-narrative" [130] or "super-conspiracy theory" [131], as it connects multiple conspiracy narratives [120, 132, 133].

Regulators generally address conspiracy theories in ways similar to misinformation and disinformation, i.e., by issuing public education campaigns [61] and holding platforms accountable for monitoring harmful content, for instance, under the Digital Services Act in the European Union [32]. Platforms also employ similar moderation tools, including automated detection and content removal. Additionally, many have enacted large-scale bans on users and groups associated with promoting conspiracy theories [22, 111, 120, 124, 127, 131].

Research has investigated various dimensions of conspiracy theories. One line of work focuses on defining and categorizing specific conspiracy narratives [130, 132, 134, 135]. Others have examined the evolution and spread of conspiracy theories across mainstream platforms such as YouTube [123, 136], Reddit [121], and X (formerly Twitter) [124], as well as fringe platforms like Gab [125, 127, 137], Voat [131], Parler [120], and messaging services such as Telegram [122]. Furthermore, recent work has explored how artificial intelligence in the form of large language models can be used to reduce conspiratorial beliefs or election myths [138, 139]. In this dissertation, we contribute to this literature by developing machine learning methods to detect conspiracy theorists and by analyzing how conspiracy theorists differ from other users in their online behavior (see Section 7).

3 Detecting and understanding harmful content

3.1 Detecting harmful content on social media

Detecting harmful content is a crucial first step in identifying and addressing its societal threat. Specifically, detection enables the quantification of online threats, such as the prevalence of hate speech, misinformation, or conspiracy theories, and, thereby, enables downstream analyses, including studies about diffusion dynamics and audits of social media platforms. Detection thus provides an empirical foundation for understanding harms that might otherwise remain anecdotal or under-researched [52]. As such, detection builds the foundation to enhance the integrity of social media.

Detection typically focuses on identifying single instances of harmful content. These may include hateful posts or comments, misleading images or videos, or multimodal content that combines text and visuals. Accordingly, computational methods for the detection of harmful content are usually tailored to the modality of the input, ranging from text-based to image-based and multimodal approaches. While the majority of prior research has focused on text [33, 35, 88, 90, 140–149], there is a growing body of work addressing visual or multimodal content, such as fake images [150] or memes that combine text with visual elements [151].

In the case of text-based detection, early methods relied on simple heuristics, such as identifying predefined lists of hateful terms [90]. These approaches were later extended with feature-based models incorporating lexical patterns (e.g., n-grams), stylistic indicators (e.g., punctuation, word length), user-level metadata (e.g., account age, follower count), and diffusion characteristics (e.g., how far and how fast a post spreads) [33, 35, 90, 140–142, 152]. Methodologically, these detection tasks are typically framed as supervised classification problems and addressed using machine learning models such as support vector machines, random forests, or gradient boosting [140, 152]. More recent work has employed deep learning architectures [35, 141–145], as well as probabilistic and statistical models [33]. Currently, the state of the art in the detection of harmful content relies on pre-trained large language models, which demonstrate strong performance in identifying subtle or context-dependent harmful content [88, 146–149].

Another stream of research focuses on detecting harmful accounts [36, 91, 153–157]. This includes accounts that systematically spread misinformation, hate speech, or other types of content that violate platform guidelines. Harmful accounts can be broadly categorized into two types: (1) social bots, i.e., automated accounts that algorithmically disseminate content [153], and (2) harmful users, i.e., accounts operated by humans who intentionally or repeatedly share harmful content.

Social bots may serve benign purposes (e.g., news aggregation or entertainment), but they are also frequently deployed to amplify hate speech, misinformation, or propaganda [34, 153, 158, 159]. For example, social bots played a considerable role in disseminating pro-Russian propaganda during the 2022 Russian invasion of Ukraine [34]. A substantial body of work has focused on bot detection [91, 153–156]. However, bots cannot be influenced through interventions such that downstream analyses typically focus on comparing bots to other users [91].

In contrast, harmful users are human-operated accounts that consistently violate platform rules. These users may engage in behaviors such as posting hate speech, inciting violence, or spreading misinformation [36, 91, 160]. The methodology used to detect these users is similar to that applied for identifying bots, but the downstream implications differ. Previous studies have developed methods to detect users who consistently propagate harmful content, such as hate speech or misinformation [36, 91, 161]. Unlike bots, harmful users can, in principle, respond to interventions, making their identification crucial for implementing interventions. Therefore, accurate detection of harmful users is essential for timely content moderation and for developing scalable, user-specific interventions.

To identify social bots and harmful users, researchers rely on a diverse set of features that characterize social media behavior. Specifically, five main feature groups are commonly used: (1) user features, (2) linguistic features, (3) network features, (4) content features, and (5) temporal features.

Each feature group is constructed to describe a specific characteristic of social media users: (1) User features include metadata such as account age, posting frequency, or properties of a user's friendship and follower networks. (2) Linguistic features describe patterns in a user's language use, including sentiment, syntactic complexity, hashtag frequency, or lexical diversity. (3) Network features capture the structural properties of a user's connections, such as centrality within a reshare or mention network. (4) Content features refer to the actual messages, posts, or media shared by the account. (5) Temporal features capture patterns of activity over time, such as heightened posting activity or engagement across specific hours or days.

These features have been employed in various machine learning pipelines to detect different types of accounts [36, 91, 153–157]. Methodologically, the task is typically framed as a supervised classification problem. Common models include logistic regression, random forests, support vector machines, and gradient boosting [36, 91, 153, 154, 157], ensembles of these classifiers [156], as well as deep learning models such as neural networks [155].

Evaluating methods for detecting harmful content is essential for assessing their ability to reliably distinguish between harmful and non-harmful content. Typically, evaluations are conducted using curated datasets annotated with ground-truth labels for classification tasks [88, 144, 162–167]. These labels are often generated through human annotation, which is considered the gold standard [88]. However, alternative approaches have been developed that infer so-called *weak labels* based on contextual cues, such as comments or user behavior [144]. To promote comparability, researchers also organize shared tasks and evaluation challenges where detection models are tested on benchmark datasets under standardized conditions [162, 168].

The performance of models to detect harmful content is commonly evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristics curve (ROC-AUC) [33, 36, 88, 154]. An overview of the most common metrics, including the formula to compute them as well as a short description, is in Table 1.

Table 1: **Performance Metrics for Model Evaluation:** Let TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. The table shows the classification metric (Metric), formula to compute the metric (Formula), and a description of the metric (Description).

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Measures the overall proportion of correct predictions. It can be misleading in imbalanced datasets where the majority class dominates.
Precision (Specificity)	$\frac{TP}{TP+FP}$	Indicates the proportion of predicted positive instances that are actually positive. High precision means fewer false positives, important when false accusations are costly.
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	Measures the proportion of actual positive instances correctly identified. High recall is crucial when missing harmful content is risky.
F1-score	$2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	The harmonic mean of precision and recall, providing a balanced metric useful in cases of class imbalance.
ROC-AUC	$\int_0^1 TP(FP)\,dFP$	The ROC-AUC measures the area under the receiver operating characteristic curve and provides a trade-off between the true positive rate and the false positive rate across thresholds.

The choice of metric depends on the application. For example, when detecting misinformation, minimizing false negatives (high recall) may be prioritized to prevent harmful content from spreading. In contrast, in hate speech detection, high precision may be emphasized to avoid censoring benign posts.

A key challenge in evaluating methods for detection is the issue of class imbalance, as harmful content typically represents only a small fraction of social media data [88, 167]. As a result, evaluation metrics must be chosen with care. For instance, accuracy is generally a poor measure in imbalanced settings, as it may yield high values by predominantly predicting the majority class [169]. To address this, researchers often rely on alternative metrics such as precision, recall, F1-score, and ROC-AUC, which provide a more nuanced view of classifier performance under imbalance [169]. In addition, many studies employ resampling techniques such as upsampling the minority class or downsampling the majority class [169]. Alternatively, one can apply cost-sensitive learning to penalize misclassification of harmful content more heavily [169]. However, these strategies may artificially alter class distributions and thereby distort the model's real-world performance [167, 169], where the actual share of harmful content remains low. Consequently, evaluations must be interpreted in light of both metric selection and class distribution to ensure reliable assessment of model effectiveness.

Additional challenges result from the need to continuously update detection models in response to linguistic or cultural shifts, such as the use of new expressions or hashtags [58, 167, 170]. Annotator bias is another concern, particularly in subjective classification tasks [171]. This is typically mitigated by employing multiple annotators per instance and measuring inter-annotator agreement [36, 88, 152]. Moreover, definitions of harmful content may vary across cultures, underscoring the need for culturally sensitive classifiers [167, 170]. A related limitation is the dominance of English-language datasets in current research, which mandates the development of detection models for additional languages to tackle harmful content across the globe [172].

Overall, detection constitutes the foundation for addressing the societal risks posed by harmful content on social media. By enabling the identification of harmful content and users, detection serves as the basis for downstream tasks such as understanding, auditing, and interventions. It also allows researchers and policymakers to quantify emerging threats and document harms that may otherwise remain anecdotal or invisible [52]. Yet, detection is not without challenges, ranging from dataset bias and class imbalance to evolving cultural norms and language diversity [88, 167, 169, 170]. Despite these limitations, advances in computational methods, especially those leveraging large language models, have significantly improved our ability to detect harmful content and users [88, 146–149].

We illustrate detection methods in our first case study, demonstrating how state-of-the-art machine learning methods can be applied to identify conspiracy theorists on social media platforms. Specifically, we profile QAnon supporters on the fringe platform Parler, thereby showcasing the use of machine learning for detection as a tool to enhance platform integrity and accountability. The complete case study is in Section 7.

3.2 Understanding harmful content on social media

To understand how harmful content poses a threat to society, it is essential to examine how such content operates on social media. Beyond detecting and quantifying the prevalence of hate speech, misinformation, or conspiracy theories, it is crucial to study the underlying characteristics of harmful content, how it spreads across networks, and how it influences both online interactions and offline behavior. This deeper understanding is fundamental for evaluating which types of content warrant scrutiny, how content should be audited, and what interventions are most effective [52]. In this sense, understanding serves as a key dimension for enhancing the integrity of social media platforms by informing auditing and intervention strategies.

Understanding harmful content is often concerned with *who* is sharing hate speech, misinformation, or propaganda. In this vein, research has examined the differences between harmful and non-harmful users. For instance, hateful users on X (formerly Twitter) tend to be more active and exhibit stronger connections within the social network [36]. Similarly, users who engage with conspiracy theories often display higher posting frequency and maintain larger friendship networks, suggesting greater influence and reach [119]. However, the dissemination of harmful content is not always intentional. The sharing of false information is frequently habitual and shaped by the platform mechanics of social media [173]. These findings suggest that harmful users can exert a disproportionate impact on social media discourse [119], making them important targets for both audits and intervention strategies. At the same time, these results underscore the potential of platform-level changes to reduce harmful behavior by addressing structural incentives [173].

Understanding is further concerned with studying why and how content is propagating through online environments. To this end, research has studied the diffusion dynamics of harmful content [33, 37, 38, 46, 118, 174, 175]. This body of work typically models resharing cascades to examine how content propagates, providing insights into the speed, breadth, and depth of diffusion, essentially, its virality [37, 38, 174, 176]. Overall, harmful content tends to exhibit greater virality than non-harmful content [37, 174]. For example, false information has been found to spread six times faster and reach a given number of users twenty times earlier than true information [37]. Homogeneous friendship networks, or so-called "echo chambers," may further amplify this effect by reinforcing like-mindedness and limiting exposure to corrective content [38, 46]. Similarly, content characteristics, such as negativity, drive online news consumption [177], which may facilitate the spread of misinformation.

User characteristics also influence the virality of harmful content [174]. On X (formerly Twitter), hateful posts from verified users are more likely to spread than those from non-verified accounts [174]. However, some scholars argue that the greater virality of false information may primarily stem from differences in cascade size, rather than content properties [176]. Despite these nuances, the diffusion patterns of harmful content have important implications for platform governance. For instance, interventions could prioritize users whose content is more likely to spread broadly, or aim to reduce the "infectiousness" of harmful posts through structural interventions [174, 176]. In this way, understanding diffusion dynamics supports the design of targeted interventions to enhance the integrity of social media.

Understanding the online effects of harmful content is essential for assessing how it shapes user behavior, social dynamics, and platform-level discourse. Studies show that exposure to harmful content, such as hate speech or misinformation, triggers immediate psychological responses, including stress, anxiety, and perceived discrimination, especially among vulnerable populations [11–13]. Over time, this exposure can contribute to behavioral changes, such as reduced participation in online discussions, increased self-censorship, or a decline in empathy and civility [13, 72]. These effects are further exacerbated by platform mechanics [68, 69]. For example, research has found that algorithmic curation tends to amplify emotionally charged and uncivil content [68, 69]. This amplification may facilitate the creation of echo chambers, which intensify polarization and accelerate the spread of misinformation and hate speech [38, 46]. Additionally, users who frequently disseminate harmful content tend to be more active and well-connected, enhancing their capacity to shape online discourse [36, 119]. These findings highlight that harmful content does not exist in isolation but systematically alters platform dynamics and social norms. Understanding these dynamics is critical to developing platform-level interventions that go beyond content removal to address the structural incentives that sustain harmful behavior [173].

Research also aims to understand how the consequences of harmful content extend beyond digital spaces and affect the offline world [14, 16, 19, 77, 126]. For example, studies link misinformation and conspiracy theories to adverse public health outcomes [14, 16, 77]. During the COVID-19 pandemic, for example, such content was associated with vaccine hesitancy and non-compliance with protective health measures, ultimately contributing to preventable illness and mortality [16, 126]. Similarly, hate speech and online radicalization have been implicated in real-world acts of violence, such as the Christchurch mosque shooting in 2019 and the storming of the U.S. Capitol in 2021 [19, 81, 82]. Beyond isolated incidents, harmful content has been shown to erode trust in democratic institutions, polarize political attitudes, and weaken collective decision-making [10, 17, 25]. These effects are often disproportionately borne by marginalized communities, who face greater exposure to discrimination, misinformation, and targeted attacks [11, 102]. Collectively, these studies provide a better understanding of the societal threats from harmful content and its capacity to influence the offline world. This emphasizes the importance of holistic, evidence-based interventions that account for both the digital and societal dimensions of online harm.

Understanding how harmful content spreads and shapes both online and offline environments is essential for safeguarding the integrity of social media. Research has shown that harmful content tends to be more viral [37, 176], exerts disproportionate influence on user behavior and platform dynamics [11–13], and contributes to serious offline consequences, ranging from public health risks to the erosion of democratic values [10, 16, 19, 25]. These findings highlight that effective interventions to counter harmful content must be grounded not only in detection, but also in a nuanced understanding of how harmful content functions within social media [52].

This dissertation contributes to this literature by presenting two case studies: the first examines how emotional language influences the virality of online rumors on X (formerly Twitter), and the second compares how emotional cues affect the diffusion of true versus false rumors. Together, these studies outline key mechanisms behind the spread of harmful content and provide insights for developing more targeted and scalable interventions. The case studies are in Section 8 and Section 9, respectively.

4 Auditing social media platforms

Social media platforms exert substantial influence over the content that users encounter online [46, 69, 73, 178]. Therefore, it is essential not only to detect and understand harmful content hosted on these platforms but also to audit how this content is delivered. In the early stages of platform development, content was typically displayed in reverse chronological order, showing posts from users' followers [69]. However, content delivery has since evolved, with algorithms now governing the ranking of content in users' feeds [40, 69]. Despite this shift, it remains unclear how harmful content is managed or even amplified by these algorithms [68, 69]. Since the content presented to users can significantly shape both online behavior and real-world actions [3–6], auditing social media platforms—particularly the algorithms that govern content delivery—is crucial. Ensuring that these algorithms align with democratic principles such as transparency, fairness, and accountability is vital to maintaining the integrity of social media platforms.

Algorithmic content delivery has increasingly come under scrutiny from researchers, policymakers, and the public in recent years [44, 45, 69]. On an individual level, privacy concerns are eminent, as content delivery often relies on private or even sensitive user data [179–181]. On a societal level, fears persist that algorithmic content delivery may create filter bubbles and echo chambers, reinforcing stereotypes, polarizing societies, and facilitating the spread of harmful content [38, 46, 69, 70, 182]. For example, Facebook has been blamed for promoting divisive content harmful to democratic processes [94, 183]. Furthermore, algorithms may introduce bias by favoring certain content [40] or disproportionately targeting specific audiences [41–43, 53–55, 184, 185]. This is particularly problematic as it limits content providers' ability to reach diverse audiences and restricts users' access to a balanced range of perspectives. However, the algorithms responsible for content delivery are often opaque, proprietary, and beyond societal control [41, 42, 53, 54, 69]. These concerns underscore the urgent need for greater transparency, fairness, and accountability in the delivery of content by social media algorithms.

Concerns about algorithmic bias and the integrity of democratic processes have led to calls for greater monitoring of algorithmic content delivery on social media [40, 44–46, 69, 186, 187]. However, granular data that would enable independent audits, such as view counts, reactions, or engagement metrics, is often inaccessible or unreliable [44, 45]. This lack of transparency complicates efforts to hold platforms accountable and may even facilitate misconduct [44, 70, 94]. This is especially concerning given evidence suggesting that social media content can significantly influence offline behaviors such as voting patterns [188–191] and, in extreme cases, contribute to real-world violence [19, 81, 87]. As such, monitoring and auditing content delivery on social media platforms has become essential for ensuring the integrity of democratic processes and maintaining public trust.

In response to public and regulatory pressure, social media platforms have begun improving transparency around content delivery. Legislative initiatives such as the Digital Services Act in the European Union [32] or the Platform Accountability and Transparency Act in the United States [192] have mandated greater transparency from major platforms

[32, 56, 192]. Consequently, platforms like Meta (i.e., Facebook and Instagram) [193], YouTube [194], TikTok [195], and X (formerly Twitter) [196] have launched research platforms that provide access to internal data, facilitating external audits [56, 191]. Researchers have utilized these tools to examine various aspects of social media, including advertising [43, 53], political communication [191, 197–202], and harmful content [36, 37, 91].

There have also been collaborative efforts between social media companies and researchers to audit platforms and enhance platform integrity [39, 40, 68, 69, 203, 204]. For example, Meta has partnered with researchers to analyze how social media and algorithmic content delivery on Facebook and Instagram influenced attitudes and behaviors before the 2020 U.S. presidential elections [39, 68, 69, 203, 205]. Similarly, teams at X (formerly Twitter) have collaborated with researchers to audit how algorithms amplify political content on their platform [40]. Other efforts have focused on evaluating the impact of automated content moderation on adherence to platform guidelines [204]. While these collaborations have provided valuable insights into the inner workings of social media platforms, they have also faced criticism regarding the independence of researchers, particularly when they are funded by the companies or lack full control over the research design and implementation [206].

To complement these efforts, community-driven auditing initiatives have emerged, often relying on data donations and volunteer participation from social media users [207–209]. These initiatives have been used to monitor political advertising [207, 208], detect online risks for adolescents [210], and assess the spread of misinformation or propaganda [94, 112]. While these analyses offer an independent view of social media platforms, they face significant challenges in monitoring content or users at scale and often lack access to critical data points such as engagement metrics and reach. Furthermore, the community-driven nature of these initiatives introduces biases toward the community donating the data, which can affect the representativeness of the findings [211].

Auditing social media platforms typically pursues three primary goals. First, researchers are often interested in *who* views the content delivered by social media algorithms [41–43, 53–55, 184]. This involves studying biases in content delivery, such as whether certain groups, based on gender, age, or socio-cultural factors, are disproportionately exposed to specific types of content. For example, research on the delivery of paid advertising has revealed that minority groups are particularly likely to be shown problematic ads [55]. In contrast, content providers may also abuse such features to discriminate against certain user groups [179, 184]. Second, scholars focus on *what* content is amplified by social media algorithms. Studies have shown that algorithms tend to amplify certain types of content, such as uncivil discourse or politically biased (e.g., right- or left-leaning) content [39, 40, 46, 68, 203]. Finally, auditing addresses *how* algorithmic content delivery shapes user attitudes and behaviors. Research has demonstrated that algorithmic delivery influences not only the time users spend online but also their political engagement and news knowledge [39, 69]. Overall, auditing aims to address a broad set of goals that are critical to enhancing the integrity of social media platforms.

Methodologically, auditing social media platforms involves both experimental and observational approaches. Experimental approaches, for example, often alter users' experience on a platform by intervening in content delivery, such as modifying feed algorithms [39, 68, 69]. Other experiments induce content, such as paid advertisements, to evaluate how content is delivered [41, 212]. However, ethical considerations may limit the applicability of experiments, particularly when interventions withhold information or favor certain audiences [191]. Moreover, experimental studies are typically conducted over a limited time period, which restricts their ability to assess long-term effects or audit platforms continuously [39, 68, 69]. Additionally, experiments often focus on specific audiences or contexts, which may not generalize to the platform as a whole [69, 191, 213].

In contrast, observational studies can audit platforms over extended periods and at scale, reducing the risk of sampling or selection bias. These studies typically rely on data collected from platform APIs or publicly available content [42–46, 53–55, 184, 185, 197, 199, 202]. Observational methods are particularly useful for large-scale audits, as they allow researchers to study real-world interactions at the population level without interfering with the platform's operations. However, a key limitation of observational studies is that they rely on strict assumptions to establish causal relationships [191]. These assumptions are often difficult to justify, making it challenging to draw causal conclusions [191].

Overall, both types of studies are essential for auditing social media platforms. Experimental studies facilitate the testing of hypotheses about causal mechanisms, while observational studies provide valuable insights into real-world behavior and large-scale trends. Together, they offer a complementary approach for understanding the complex dynamics of algorithmic content delivery and ensuring that platforms are held accountable for their societal impact.

Auditing social media platforms is essential for enhancing their integrity and accountability [44, 45, 56]. Concerns regarding transparency, algorithmic bias, and unfair content delivery are prevalent and often conflict with democratic values such as fairness, inclusivity, and open discourse [44, 45, 56, 69, 70]. Addressing these issues requires a comprehensive approach involving regulatory frameworks, collaborations between social media companies and researchers, and independent, community-driven initiatives. Methodologically, auditing demands diverse empirical strategies, including both experimental and observational studies, to reliably detect, understand, and mitigate threats to the integrity of social media.

In this dissertation, we illustrate a social media audit through a case study examining political advertising on Meta's platforms, Facebook and Instagram, during the 2021 German federal election. Specifically, we evaluate the use of targeting features for political advertising and audit discrepancies between targeted and actual audiences due to the algorithmic delivery of political ads. This case study highlights the importance of auditing as a critical tool for safeguarding democratic processes against potential misuse of targeted political ads. The complete case study is in Section 10.

5 Interventions to counter harmful content

Content moderation and effective interventions are essential for enhancing the integrity of social media platforms. While detection, understanding, and auditing are critical for identifying and analyzing threats posed by harmful content, they do not directly offer strategies for mitigating such content. Content moderation refers to the broader, ongoing process of curating user-generated content on a platform to ensure compliance with community standards and societal norms [58, 59]. In contrast, interventions are specific, targeted measures designed to address particular issues [10]. For instance, platforms may deploy interventions such as accuracy prompts or automated content removal to counter the spread of misinformation [10, 51]. Each of these constitutes a distinct intervention, but together they form an integral part of a platform's overall content moderation strategy. Developing effective content moderation practices, therefore, relies on designing, implementing, and evaluating individual interventions that collectively safeguard platforms from harmful content.

Interventions to address harmful content typically follow two main strategies: (1) top-down strategies, where enforcement is carried out by the platform itself, and (2) bottom-up strategies, which encourage users to take an active role in countering harmful content. Importantly, users generally agree that content moderation is necessary and support measures to counter harmful content [103], providing legitimacy to interventions and moderation efforts.

Top-down interventions are directly enforced by platforms and are guided by predefined policies and community standards. Examples include the downranking of harmful posts to reduce their visibility, the removal of content, or the suspension of users who violate community standards [26]. These interventions are appealing because they ensure consistency in enforcement and can be scaled to manage the vast volume of content generated on social media platforms [58]. Typically, platforms define a set of rules that are operationalized through automated systems or human moderators to address harmful content [57, 59]. However, top-down measures may also be perceived as intrusive or censorship, particularly when enforcement lacks transparency or when moderation errors occur [58]. For example, satirical or ironic content may be misclassified as harmful [214, 215], leading to unjustified content removal or account penalties. Such errors can undermine user trust and spark criticism over bias, censorship, or opaque decision-making processes [57].

In contrast to top-down strategies, bottom-up strategies, often referred to as community-based interventions, encourage users to actively participate in content moderation [57]. For example, users can flag posts for violating community standards, report inappropriate behavior, or respond to harmful content by offering corrective information or counterspeech [27, 49, 57, 59]. These mechanisms grant considerable agency to users and are typically perceived as less invasive, as they promote community self-regulation and support freedom of expression [57]. Despite their advantages, bottom-up strategies also face significant challenges. They often lack consistency and scalability, making them less effective in addressing the vast volume of content generated on large platforms [57, 58].

Moreover, such approaches may inadvertently reinforce existing biases within online communities [57]. For instance, majority groups may disproportionately flag content from minority users, potentially silencing marginalized voices. These limitations highlight the need for carefully designed mechanisms that balance user empowerment with safeguards against abuse and bias.

Content moderation is typically implemented using two main approaches: (1) manual moderation carried out by human moderators [216], and (2) automated moderation through algorithms [58]. Manual moderation offers flexibility and the ability to tailor decisions to the specific context of abusive content or behavior [58]. However, it requires substantial human labor, which limits scalability on large social media platforms [57]. Moreover, exposure to harmful content may adversely affect the well-being of human moderators [58, 216]. In addition, manual decisions can be subject to individual biases, which may compromise the consistency and fairness of moderation outcomes [217, 218].

In contrast, automated moderation relies on algorithms, typically machine learning models, to identify and manage harmful content [57, 204]. Automated systems are highly scalable and can ensure more consistent application of moderation policies [57]. Yet, they often struggle to capture the nuanced context of online communication. For instance, detecting irony, sarcasm, or coded language remains a persistent challenge for automated classifiers [214, 215]. Consequently, such systems face an inherent trade-off between precision and overreach, where overly strict models risk removing benign posts, while lenient models may fail to address harmful content [57, 219]. Moreover, automated systems may also exhibit biases that result in disproportionate moderation for specific types of content or user groups [218].

To leverage the strengths of both manual and automated approaches, many social media platforms have adopted hybrid moderation systems [58]. In such systems, automated algorithms are used for the initial detection of potentially harmful content, while human moderators review edge cases or high-impact decisions [57]. This combination allows platforms to scale moderation efforts while retaining the contextual judgment and flexibility that human reviewers provide [57]. For example, machine learning models may flag posts for containing hate speech [204], but final decisions about content removal or account suspension may be delegated to human moderators [57]. Hybrid systems can thus improve both the efficiency and the fairness of content moderation. However, they also introduce coordination challenges, such as aligning algorithmic thresholds with human standards [204] and ensuring accountability in decision-making chains [57]. As moderation becomes increasingly complex, hybrid systems are likely to remain central to the evolving architecture of content moderation on social media [58].

Policymakers have introduced a range of frameworks to address harmful content on social media platforms [32, 192]. Early regulatory approaches often relied on voluntary commitments, wherein platforms collaborated with governments and civil society to self-regulate harmful content [52]. However, growing public pressure and continued evidence of platform inaction [69, 220–222] have led to the development of more robust legal mandates with clear enforcement mechanisms [223]. For instance, the Digital Services Act (DSA) in the European Union requires large platforms to actively mitigate harmful content and increase transparency around their moderation practices [32, 224]. While the DSA sets standards on *what* content needs to be monitored, it also grants platforms the flexibility to design and implement tailored content moderation and intervention strategies that align with their specific operational and community contexts [223].

Social media platforms play a pivotal role in implementing interventions against harmful content [52, 57, 59]. For example, Meta has adopted a "Remove, Reduce, Inform" policy aimed at mitigating content that violates its community standards [26]. This strategy includes downranking content from users or groups that frequently disseminate misinformation and collaborating with external experts to develop new interventions [26]. Similarly, X (formerly Twitter) has removed accounts associated with conspiracy theories or misinformation [73, 124] and introduced the "Community Notes" feature, which allows users to collaboratively add context to potentially misleading posts [27]. The latter has been deemed effective for reducing the engagement and diffusion of false content [175]. In addition, platforms have deployed automated systems to identify and suppress harmful content at scale [204], while also enabling users to report or flag abusive behavior [57]. Despite these efforts, platforms have faced consistent criticism from both non-governmental organizations [221, 222] and policymakers [32, 60] for failing to moderate harmful content adequately. In this context, academic research can play a crucial role by developing, testing, and refining evidence-based content moderation policies and interventions to support platform efforts.

A large body of research has examined the design and efficacy of content moderation policies on social media platforms [111, 204, 218, 225–228]. Scholars have explored how platform norms and governance structures evolve over time [225, 226], as well as how different moderation strategies, i.e., manual [227, 229, 230], automated [204], and hybrid systems [228], shape user behavior and content quality. Overall, these studies suggest that content moderation can effectively reduce the visibility and spread of harmful content. However, this literature also highlights implementation challenges, including algorithmic biases [218], moderator well-being [58, 216], and the difficulty of establishing universally accepted norms across diverse user communities [57, 227, 229, 230]. These findings underscore the need for nuanced, adaptable interventions that balance efficacy, fairness, and transparency.

In response to the diverse types of harmful content, researchers have developed and evaluated a range of targeted interventions. For instance, to combat misinformation, common strategies include accuracy prompts that nudge users to consider the veracity of information before sharing [51, 104–108], media literacy tips that educate users on identifying false or misleading claims [50, 107–109], and debunking or rebuttal techniques that provide corrective information [76, 139]. Similarly, interventions against conspiracy theories challenge beliefs in conversations with chatbots [138]. Interventions against online hate speech have focused on counterspeech, responses intended to challenge or neutralize hateful content [21, 47–49]. These studies have analyzed both the source of counterspeech (e.g., whether the intervention is delivered by in-group or out-group members) [21, 48] and the conversational strategy of the response (e.g., empathy vs. humor) [49]. Recently, interventions are frequently relying on large language models to be flexible and scalable [138, 139, 231, 232]. This illustrates the importance of tailoring interventions to the social dynamics between users and the context of harm.

Methodologically, the evaluation of interventions to counter harmful content predominantly relies on experimental research designs. This results from two major limitations of observational studies in this context: First, internal platform data necessary to track intervention effects is often inaccessible to external researchers [228]. Second, causal inference in observational settings is challenging [111, 191, 230], as it is typically unclear which users or content were exposed to interventions [204]. In contrast, experimental designs offer the advantage of yielding robust causal evidence regarding the effectiveness of interventions.

To assess the efficacy of interventions, researchers have employed both laboratory [51, 76, 104–106, 108] and field experiments [21, 47–49, 51, 107]. Laboratory experiments offer high internal validity and allow for precise control over experimental conditions. They also pose fewer ethical concerns, as participants provide informed consent in a controlled setting. However, their external validity may be limited, as behaviors in lab settings do not always generalize to real-world social media environments [51]. For example, the social desirability bias suggests that individuals may report prosocial attitudes, such as disapproval of hate speech, under lab conditions due to perceived social pressure, but fail to act accordingly in actual online settings [233]. Field experiments, by contrast, offer stronger external validity by testing interventions in actual platform contexts [51]. However, they face technical and ethical challenges: access to platform infrastructure and data may be restricted, and participants are often unaware they are part of a study, raising concerns about informed consent and user safety [234, 235]. Moreover, both lab and field experiments frequently rely on non-representative samples, which may limit the generalizability of findings [191]. Despite these limitations, experiments remain the gold standard for understanding causal mechanisms and are critical for assessing the real-world impact of interventions aimed at mitigating harmful content on social media.

Content moderation encompasses a broad set of efforts aimed at enhancing the integrity of social media platforms [57, 59]. In contrast to detection, understanding, and auditing, which focus on identifying and analyzing harmful content, interventions shift the perspective toward *how* harmful content can be effectively mitigated. Achieving this requires a combination of clear platform policies and active community engagement, enabling moderation systems that can both scale to the volume of online content and account for diverse community standards [57]. As such, collaboration between policymakers, platform owners, and researchers is essential [52]. Policymakers must define the regulatory framework and normative boundaries for interventions, which platform owners can then operationalize through technological and procedural measures. This process should be informed by empirical research that helps strike a balance between effective mitigation and the risk of overreach, while also delivering evidence-based interventions that can be implemented at scale [52].

In line with this objective, this dissertation discusses whether contextualized counterspeech generated by large language models can effectively mitigate online hate speech. We evaluate this in a large-scale field experiment on X (formerly Twitter) and discuss how different conversational strategies, such as empathy and warning-of-consequences, may influence the effectiveness of counterspeech. The complete case study is provided in Section 11.

6 Discussion

Social media has become an integral part of our online ecosystem, shaping both individual communication and public discourse [3–6]. However, the widespread dissemination of harmful content, such as hate speech, misinformation, and conspiracy theories, poses significant challenges to platform integrity and constitutes a serious threat to society [10, 17–25]. Empirical evidence shows that harmful content undermines individual well-being [11–16, 72], exacerbates societal polarization [21, 23, 24, 78, 79], threatens democratic principles [10], and, in extreme cases, incites real-world violence [21–25]. These developments underscore the urgent need for effective and scalable approaches to mitigate the proliferation of harmful content and to enhance the integrity of social media.

In this dissertation, we propose a computational approach to enhance the integrity of social media platforms. By integrating state-of-the-art methods from computer science with theoretical and empirical insights from the social sciences, we develop a holistic framework to address the challenges posed by harmful content. Along five case studies, this dissertation demonstrates how to (1) detect and understand harmful content, (2) audit social media platforms, and (3) design and evaluate effective interventions to mitigate online harm.

Specifically, we show how (1) machine learning can be used to detect QAnon conspiracy theorists among regular users (Section 7) and study the influence of emotions for the spread of online rumors to gain a better understanding of the diffusion dynamics of harmful content (see Section 8 and Section 9). (2) We demonstrate how computational methods can be applied to audit the algorithmic delivery of political advertisements and uncover systematic discrepancies in the delivery of ads by different political parties (see Section 10). (3) We design and evaluate a scalable intervention based on large language models to mitigate online hate speech, where our findings warrant caution when implementing AI-based solutions for platform moderation (see Section 11).

The interdisciplinary approach adopted in this dissertation underscores the necessity of integrating computational and social science perspectives to effectively address the spread of harmful content on social media. Computational methods provide the analytical foundation to detect harmful content and examine diffusion patterns at scale. For instance, in the first case study on QAnon conspiracy theorists, machine learning techniques are used to classify conspiratorial accounts (see Section 7), while the second and third case study on rumor diffusion rely on emotion detection and cascade modeling to understand how harmful content spreads (see Section 8 and Section 9). Similarly, scalable interventions such as AI-generated counterspeech in the fifth case study depend on advanced natural language processing to produce contextualized responses to hateful posts (see Section 11).

At the same time, social science contributes essential theoretical and normative insights. It informs which types of content warrant scrutiny, generates hypotheses about their societal impact, and explains how users may respond to moderation and intervention. For example, the design of the counterspeech builds on behavioral theories related to empathy and warning-of-consequences (see Section 11), while the audit of political ad delivery draws on democratic theory to assess fairness and accountability (see Section 10). Moreover, social science perspectives draw attention to ethical risks, including the potential for backfiring effects, reinforcement of stereotypes, and bias introduced through training data. Ultimately, this dissertation demonstrates how interdisciplinary research enables the development of empirically grounded, ethically responsible tools to enhance the integrity of social media.

This dissertation employs a range of methodological approaches, including exploratory analyses (e.g., quantifying QAnon conspiracy theorists), explanatory studies (e.g., identifying drivers of rumor diffusion), and experimental designs (e.g., testing counterspeech interventions against online hate). This reflects the breadth of empirical strategies necessary to effectively address harmful content, while also highlighting the inherent challenges of social media analysis. As outlined previously, exploratory and explanatory approaches often rely on observational data that offer a comprehensive perspective on specific phenomena [191]. For instance, our audit of political advertising during the 2021 German federal election draws on a complete dataset of ads published on Facebook and Instagram throughout the campaign period (see Section 10). However, identifying causal effects in observational settings is challenging [111, 191, 230], which warrants caution in interpreting the findings.

In contrast, experimental approaches offer stronger causal evidence for the mechanisms under investigation. Our counterspeech experiment, for example, provides a robust indicator of how AI-generated interventions may backfire on social media platforms (see Section 11). Nevertheless, experimental studies are inherently limited in scope [191]. Even large-scale experiments involving millions of users [188] may not generalize to the global scale of social media platforms, which host billions of users across diverse sociocultural contexts [1, 52]. Additionally, the long-term effectiveness of interventions remains difficult to assess when outcomes can only be observed over a restricted time horizon [39, 68, 69]. Overall, the combination of exploratory, explanatory, and experimental methodologies is, therefore, essential for generating robust, policy-relevant evidence to inform effective countermeasures against harmful content.

This dissertation presents a series of studies with important implications for both platform owners and policymakers. Across all case studies, we demonstrate that harmful content is widespread on social media and has detrimental societal consequences [10, 17–25, 119, 197, 236–238]. We further propose and empirically evaluate interventions to mitigate these harms, situating them within a broader computational framework aimed at enhancing platform integrity. This research relies on access to data controlled by private companies such as Meta, as well as on the regulatory environments shaped by policymakers. As such, the collaboration of both platforms and policymakers with academia is critical for enabling impactful and ethically responsible research to mitigate threats from harmful content [52].

Platforms play an essential role in enhancing the integrity of social media, as they set the rules, boundaries, and infrastructure that enable research and intervention to enhance the integrity of social media [57, 59]. While many platforms have implemented measures to mitigate harmful content [26, 27], they have also faced persistent criticism for inadequate moderation policies [44, 69, 87, 179, 221, 222]. In response to public pressure [69, 87, 179, 220–222], some platforms have introduced transparency tools [196, 239, 240] and begun collaborating with independent researchers [39, 68, 69, 205]. These efforts have enabled important research on platform design, content moderation, and political communication [39, 68, 198, 204], and have facilitated novel studies that were previously infeasible due to data restrictions [54, 191, 198–202, 241, 242]. Our own work auditing the delivery of political ads during the 2021 German federal election contributes to this line of work enabled through transparency measures (see Section 4). However, platforms should adopt a more proactive stance in fostering transparency and academic collaboration. Systematic engagement with the research community could help improve platform design, strengthen content moderation systems, and ultimately enhance the user experience and the overall integrity of social media ecosystems.

In recent years, policymakers have taken an increasingly active role in regulating social media platforms [32, 60, 65, 192]. New regulations, such as the Digital Services Act (DSA) in the European Union [32], mandate that platform owners take concrete measures to mitigate harmful content or face significant fines [32]. Importantly, these regulations also facilitate academic research by requiring platforms to provide data access [56, 191], thereby enabling the study of societal threats emerging from digital ecosystems. Such regulatory developments have made research like our experiment on the effectiveness of counterspeech possible (see Section 5). However, the current regulatory frameworks often lack specificity, leaving loopholes that allow platforms to meet formal obligations without enabling meaningful oversight [223]. To address this, policymakers should actively engage with researchers to close these gaps and foster research that informs evidence-based interventions. Conversely, researchers must collectively advocate for stronger regulatory mandates by emphasizing both the societal harms of harmful content and the empirical needs required to study and address them. This collaborative effort becomes even more urgent in light of recent trends, including announcements by some platform owners to reduce or abandon content moderation efforts [243, 244] or political shifts calling for fewer interventions against harmful content [103, 245], often under false pretense of free speech [245]. In this context, sustained cooperation between policymakers and the research community is critical to ensuring that platform regulation remains robust, informed, and democratically accountable.

This dissertation offers a comprehensive framework to detect and understand, audit, and intervene against harmful content on social media. In doing so, it provides a broad foundation to guide future research. One important direction concerns the distinction between illegal and legally permissible yet socially harmful content (e.g., misinformation about climate change). While illegal content requires interventions based on established legal standards, socially harmful content demands more nuanced approaches that respect freedom of expression while mitigating potential harms [103]. Another promising direction involves the development of new data collection methods. For instance, data donation platforms represent a novel yet underexplored mechanism for obtaining user-level insights while preserving privacy and ethical standards [207–209]. Similarly, advances in psychological science could inspire innovative experimental designs such as the use of mock social media platforms to study how users interact with and spread harmful content in controlled environments [232, 246–248]. This could inform specific designs of social media platforms that foster pro-democratic values [249]. Moreover, novel methodological approaches such as causal machine learning [250–259] or sensitivity analysis [260–263] may be a promising avenue for inferring causal effects from observational data in the context of social media analysis. Finally, future research should expand its scope to include understudied or emerging platforms. TikTok, for example, has transformed into an important platform of today's online ecosystem [264] that requires different computational tools to be analyzed [59]. Likewise, messenger platforms such as WhatsApp and Telegram are increasingly adopting social media-like features (e.g., broadcast channels), which raise new challenges for transparency, moderation, and accountability [122, 265–268].

Overall, this dissertation contributes to a growing body of research that combines computational methods with insights from the social sciences to better understand and address harmful content on social media. By integrating detection and understanding, auditing, as well as interventions within a unified framework, we offer practical tools and perspectives for researchers studying the societal impact of digital platforms. Moreover, this work demonstrates how evidence-based approaches can help reduce online harms and foster healthier public discourse. It also emphasizes the importance of collaboration among platform owners, policymakers, and researchers to enhance the integrity of social media and promote equity, inclusivity, and democratic values within digital ecosystems.

References

[1] Ani Petrosyan. “Number of internet and social media users worldwide as of February 2025”. In: *Statista* (2025). URL: `https://www.statista.com/statistics/617136/digital-population-worldwide/`.

[2] Stacy Jo Dixon. “Daily time spent on social networking by internet users worldwide from 2012 to 2024”. In: *Statista* (2025). URL: `https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/`.

[3] John T. Jost et al. “How social media facilitates political protest: Information, motivation, and social networks”. In: *Political Psychology* 39.S1 (2018), pp. 85–118.

[4] David Garcia et al. “Analyzing gender inequality through large-scale Facebook advertising data”. In: *PNAS* 115.27 (2018), pp. 6958–6963.

[5] Matthew H. Goldberg et al. “Shifting Republican views on climate change through targeted advertising”. In: *Nature Climate Change* 11 (2021), pp. 573–577.

[6] Emily Breza et al. “Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: A cluster randomized controlled trial”. In: *Nature Medicine* 27 (2021), pp. 1622–1628.

[7] Clay Shirky. “The political power of social media: Technology, the public sphere, and political change”. In: *Foreign Affairs* 90.1 (2011), pp. 28–41.

[8] Ruben Enikolopov, Alexey Makarin, and Maria Petrova. “Social media and protest participation: Evidence from Russia”. In: *Econometrica* 88.4 (2020), pp. 1479–1514.

[9] Robert M. Bond et al. “A 61-million-person experiment in social influence and political mobilization”. In: *Nature* 489 (2012), pp. 295–298.

[10] Anastasia Kozyreva et al. “Toolbox of individual-level interventions against online misinformation”. In: *Nature Human Behaviour* 8 (2024), pp. 1044–1052.

[11] Brendesha M. Tynes et al. “Online racial discrimination and psychological adjustment among adolescents”. In: *Journal of Adolescent Health* 43.6 (2008), pp. 565–569.

[12] Koustuv Saha, Eshwar Chandrasekharan, and Munmun de Choudhury. “Prevalence and psychological effects of hateful speech in online college communities”. In: *WebSci*. 2019.

[13] Cinoo Lee et al. “People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise”. In: *PNAS* 121.38, e2322764121 (2024).

[14] Israel Júnior Borges do Nascimento et al. “Infodemics and health misinformation: A systematic review of reviews”. In: *Bulletin of the World Health Organization* 100.9 (2022), pp. 544–561.

[15] Arne Dreißigacker et al. “Online hate speech victimization: Consequences for victims’ feelings of insecurity”. In: *Crime Science* 13, 4 (2024).

[16] Jon Roozenbeek et al. "Susceptibility to misinformation about COVID-19 around the world". In: *Royal Society Open Science* 7.10, 201199 (2020).

[17] Katherine Levine Einstein and David M. Glick. "Do I think BLS data are BS? The consequences of conspiracy theories". In: *Political Behavior* 37.3 (2015), pp. 679–701.

[18] Karen M. Douglas, Robbie M. Sutton, and Aleksandra Cichocka. "The psychology of conspiracy theories". In: *Current Directions in Psychological Science* 26.6 (2017), pp. 538–542.

[19] Elise M. Roberts-Ingleson and Wesley S. McCann. "The link between misinformation and radicalisation: Current knowledge and areas for future inquiry". In: *Perspectives on Terrorism* 17.1 (2023), pp. 36–49.

[20] Alexandra A. Siegel. "Online hate speech". In: *Social Media and Democracy: The State of the Field and Prospects for Reform*. Ed. by Nathaniel Persily and Joshua A. Tucker. Cambridge, UK: Cambridge University Press, 2020, pp. 56–88.

[21] Alexandra A. Siegel and Vivienne Badaan. "#No2Sectarianism: Experimental approaches to reducing sectarian hate speech online". In: *American Political Science Review* 114.3 (2020), pp. 837–855.

[22] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. "New threats to society from free-speech social media platforms". In: *Communications of the ACM* 66.10 (2023), pp. 37–40.

[23] Anni Sternisko, Aleksandra Cichocka, and Jay J. van Bavel. "The dark side of social movements: social identity, non-conformity, and the lure of conspiracy theories". In: *Current Opinion in Psychology* 35 (2020), pp. 1–6.

[24] Karen M. Douglas et al. "Understanding conspiracy theories". In: *Political Psychology* 40.S1 (2019), pp. 3–35.

[25] Stephan Lewandowsky et al. "Misinformation, disinformation, and violent conflict: From Iraq and the "War on Terror" to future threats to peace". In: *The American Psychologist* 68.7 (2013), pp. 487–501.

[26] Guy Rosen. *Remove, reduce, inform: New steps to manage problematic content*. 2019. URL: `https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/`.

[27] Nicolas Pröllochs. "Community-based fact-checking on Twitter's birdwatch". In: *ICWSM*. 2022.

[28] HateAid. *HateAid*. 2025. URL: `https://hateaid.org/en/`.

[29] Center for Countering Digital Hate. *Center for Countering Digital Hate*. 2025. URL: `https://counterhate.com/`.

[30] United Nations. *The Rabat Plan of Action*. 2012. URL: `https://www.ohchr.org/en/freedom-of-expression`.

[31] United Nations. *Countering and addressing online hate speech: A Guide for policy makers and practitioners*. 2023. URL: https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf.

[32] European Commission. *The Digital Services Act: Ensuring a safe and accountable online environment*. 2022. URL: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.

[33] Christof Naumzik and Stefan Feuerriegel. "Detecting false rumors from retweet dynamics on social media". In: *WWW*. 2022.

[34] Dominique Geissler et al. "Russian propaganda on social media during the 2022 invasion of Ukraine". In: *EPJ Data Science* 12.1 (2023).

[35] Marco Minici et al. "IOHunter: Graph foundation model to uncover online information operations". In: *AAAI* (2025).

[36] Manoel Horta Ribeiro et al. "Characterizing and detecting hateful users on Twitter". In: *ICWSM*. 2018.

[37] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151.

[38] Michela Del Vicario et al. "The spreading of misinformation online". In: *PNAS* 113.3 (2016), pp. 554–559.

[39] Andrew M. Guess et al. "Reshares on social media amplify political news but do not detectably affect beliefs or opinions". In: *Science* 381.6656 (2023), pp. 404–408.

[40] Ferenc Huszár et al. "Algorithmic amplification of politics on Twitter". In: *PNAS* 119.1, e2025334119 (2022).

[41] Anja Lambrecht and Catherine Tucker. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads". In: *Management Science* 65.7 (2019), pp. 2966–2981.

[42] Muhammad Ali et al. "Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes". In: *CSCW*. 2019.

[43] Basileal Imana, Aleksandra Korolova, and John Heidemann. "Auditing for racial discrimination in the delivery of education ads". In: *FAccT*. 2024.

[44] Laura Edelson, Tobias Lauinger, and Damon McCoy. "A security analysis of the Facebook Ad Library". In: *Symposium on Security and Privacy*. 2020.

[45] Victor Le Pochat et al. "An audit of Facebook's political ad policy enforcement". In: *USENIX Security Symposium*. 2022.

[46] Matteo Cinelli et al. "The echo chamber effect on social media". In: *PNAS* 118.9, e2023301118 (2021).

[47] Kevin Munger. "Tweetment effects on the tweeted: Experimentally reducing racist harassment". In: *Political Behavior* 39.3 (2017), pp. 629–649.

[48] Kevin Munger. “Don’t @ me: Experimentally reducing partisan incivility on Twitter”. In: *Journal of Experimental Political Science* 8.2 (2021), pp. 102–116.

[49] Dominik Hangartner et al. “Empathy-based counterspeech can reduce racist hate speech in a social media field experiment”. In: *PNAS* 118.50, e2116310118 (2021).

[50] Andrew M. Guess et al. “A digital media literacy intervention increases discernment between mainstream and false news in the United States and India”. In: *PNAS* 117.27 (2020), pp. 15536–15545.

[51] Gordon Pennycook et al. “Shifting attention to accuracy can reduce misinformation online”. In: *Nature* 592 (2021), pp. 590–595.

[52] Robert Gorwa. “The platform governance triangle: Conceptualising the informal regulation of online content”. In: *Internet Policy Review* 8.2 (2019).

[53] Basileal Imana, Aleksandra Korolova, and John Heidemann. “Auditing for discrimination in algorithms delivering job ads”. In: *WWW*. 2021.

[54] Muhammad Ali et al. “Ad delivery algorithms: The hidden arbiters of political messaging”. In: *WSDM*. 2021.

[55] Muhammad Ali et al. “Problematic advertising and its disparate exposure on Facebook”. In: *USENIX Security Symposium*. 2023.

[56] Basileal Imana, Aleksandra Korolova, and John Heidemann. “Having your privacy cake and eating it too: Platform-supported auditing of social media algorithms for public interest”. In: *CSCW*. 2023.

[57] Tarleton Gillespie. *Custodians of the internet: Plattforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT, US: Yale University Press, 2018.

[58] Tarleton Gillespie. “Content moderation, AI, and the question of scale”. In: *Big Data & Society* 7.2 (2020).

[59] Sarah T. Roberts. *Behind the screen: Content moderation in the shadows of social media*. New Haven, CT, US: Yale University Press, 2019.

[60] United Nations. *Countering disinformation*. 2022. URL: `https://www.un.org/en/countering-disinformation`.

[61] European Commission. *Identifying conspiracy theories*. 2020. URL: `https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation/identifying-conspiracy-theories_en`.

[62] Sinan Aral and Dean Eckles. “Protecting elections from social media manipulation”. In: *Science* 365.6456 (2019), pp. 858–861.

[63] Christopher A. Bail. “Social-media reform is flying blind”. In: *Nature* 603 (2022), p. 766.

[64] Ullrich Ecker et al. “Misinformation poses a bigger threat to democracy than you might think”. In: *Nature* 630 (2024), pp. 29–32.

[65] United Nations. *United Nations strategy and plan of action on hate speech*. 2019. URL: `https://www.un.org/en/hate-speech`.

[66] Ullrich K. H. Ecker et al. "The psychological drivers of misinformation belief and its resistance to correction". In: *Nature Reviews Psychology* 1 (2022), pp. 13–29.

[67] OECD. *Disinformation and Russia's war of aggression against Ukraine: Threats and governance responses*. 2022. URL: `https://www.oecd.org/en/publications/disinformation-and-russia-s-war-of-aggression-against-ukraine_37186bde-en.html`.

[68] Brendan Nyhan et al. "Like-minded sources on Facebook are prevalent but not polarizing". In: *Nature* 620 (2023), pp. 137–144.

[69] Andrew M. Guess et al. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" In: *Science* 381.6656 (2023), pp. 398–404.

[70] Wajeeha Ahmad et al. "Companies inadvertently fund online misinformation despite consumer backlash". In: *Nature* 630 (2024), pp. 123–131.

[71] United Nations. *Engage - how to deal with hate speech?* 2024. URL: `https://www.un.org/en/hate-speech/take-action/engage`.

[72] Agnieszka Pluta et al. "Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain". In: *Scientific Reports* 13, 4127 (2023).

[73] Stefan D. McCabe et al. "Post-January 6th deplatforming reduced the reach of misinformation on Twitter". In: *Nature* 630 (2024), pp. 132–140.

[74] Eli J. Finkel et al. "Political sectarianism in America". In: *Science* 370.6516 (2020), pp. 533–536.

[75] Jevin D. West and Carl T. Bergstrom. "Misinformation in and about science". In: *PNAS* 118.15, e1912444117 (2021).

[76] Philipp Schmid and Cornelia Betsch. "Effective strategies for rebutting science denialism in public discussions". In: *Nature Human Behaviour* 3 (2019), pp. 931–939.

[77] Annalise Baines, Muhammad Ittefaq, and Mauryne Abwao. "#Scamdemic, #Plandemic, or #Scaredemic: What Parler social media platform tells us about COVID-19 vaccine". In: *Vaccines* 9.421 (2021).

[78] Jan-Willem van Prooijen, André P. M. Krouwel, and Thomas V. Pollet. "Political extremism predicts belief in conspiracy theories". In: *Social Psychological and Personality Science* 6.5 (2015), pp. 570–578.

[79] Emma F. Thomas et al. "Do conspiracy beliefs fuel support for reactionary social movements? Effects of misbeliefs on actions to oppose lockdown and to "stop the steal"". In: *British Journal of Social Psychology* 63.3 (2024), pp. 1297–1317.

[80] Christopher A. Bail et al. "Exposure to opposing views on social media can increase political polarization". In: *PNAS* 115.37 (2018), pp. 9216–9221.

[81] Adam Taylor. "New Zealand suspect allegedly claimed "brief contact" with Norwegian mass murderer Anders Breivik". In: *The Washington Post* (2019). URL: https://www.washingtonpost.com/world/2019/03/15/new-zealand-suspect-allegedly-claimed-brief-contact-with-norwegian-mass-murderer-anders-breivik/.

[82] Darren W. Davis and David C. Wilson. ""Stop the steal": Racial resentment, affective partisanship, and investigating the January 6th insurrection". In: *The ANNALS of the American Academy of Political and Social Science* 708.1 (2023), pp. 83–101.

[83] European Parliament. *Fighting the rise in antisemitism and anti-Muslim hatred.* 2024. URL: https://www.europarl.europa.eu/news/en/agenda/briefing/2024-02-05/8/fighting-the-rise-in-antisemitism-and-anti-muslim-hatred.

[84] Binny Mathew et al. "Hate begets hate: A temporal study of hate speech". In: *CSCW*. 2020.

[85] Sheera Frenkel and Steven Lee Myers. "Antisemitic and anti-muslim hate speech surges across the internet". In: *New York Times* (2023). URL: https://www.nytimes.com/2023/11/15/technology/hate-speech-israel-gaza-internet.html.

[86] Kevin Roose. "On Gab, an extremist-friendly site, Pittsburgh shooting suspect aired his hatred in full". In: *New York Times* (2018). URL: https://www.nytimes.com/2018/10/28/us/gab-robert-bowers-pittsburgh-synagogue-shootings.html.

[87] Amnesty International. *Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya.* 2022. URL: https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.

[88] Abdurahman Maarouf et al. "HQP: A human-annotated dataset for detecting online propaganda". In: *ACL Findings*. 2024.

[89] Dominique Geissler, Abdurahman Maarouf, and Stefan Feuerriegel. "Analyzing user characteristics of hate speech spreaders on social media". In: *WWW 2025*.

[90] Chikashi Nobata et al. "Abusive language detection in online user content". In: *WWW*. 2016.

[91] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. "Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media". In: *CSCW*. 2019.

[92] Binny Mathew et al. "Thou shalt not hate: Countering online hate speech". In: *ICWSM*. 2019.

[93] Bruce Lannes Smith. "Propaganda". In: *Encyclopedia Britannica* (2022).

[94] Gregory Eady et al. "Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior". In: *Nature Communications* 14, 62 (2023).

[95] Aaron M. McCright et al. “Examining the effectiveness of climate change frames in the face of a climate change denial counter-frame”. In: *Topics in Cognitive Science* 8.1 (2016), pp. 76–97.

[96] Stephan Lewandowsky, Ullrich K. H. Ecker, and John Cook. “Beyond misinformation: Understanding and coping with the "post-truth" era”. In: *Journal of Applied Research in Memory and Cognition* 6.4 (2017), pp. 353–369.

[97] Sander van der Linden et al. “Inoculating the public against misinformation about climate change”. In: *Global Challenges* 1, 1600008 (2017).

[98] Lisa K. Fazio et al. “Knowledge does not protect against illusory truth”. In: *Journal of Experimental Psychology: General* 144.5 (2015), pp. 993–1002.

[99] Gordon Pennycook and David G. Rand. “The psychology of fake news”. In: *Trends in Cognitive Sciences* 25.5 (2021), pp. 388–402.

[100] Justin Farrell, Kathryn McConnell, and Robert Brulle. “Evidence-based strategies to combat scientific misinformation”. In: *Nature Climate Change* 9 (2019), pp. 191–195.

[101] Kathie M. d’I. Treen, Hywel T. P. Williams, and Saffron J. O’Neill. “Online misinformation about climate change”. In: *Wiley Interdisciplinary Reviews: Climate Change* 11.5, e665 (2020).

[102] Mohamed Mostagir and James Siderius. “Social inequality and the spread of misinformation”. In: *Management Science* 69.2 (2023), pp. 904–921.

[103] Anastasia Kozyreva et al. “Resolving content moderation dilemmas between free speech and harmful misinformation”. In: *PNAS* 120.7, e2210666120 (2023).

[104] Gordon Pennycook et al. “Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention”. In: *Psychological Science* 31.7 (2020), pp. 770–780.

[105] Jon Roozenbeek, Alexandra L. J. Freeman, and Sander van der Linden. “How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020)”. In: *Psychological Science* 32.7 (2021), pp. 1169–1178.

[106] Gordon Pennycook and David G. Rand. “Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation”. In: *Nature Communications* 13, 2333 (2022).

[107] Molly Offer-Westort, Leah R. Rosenzweig, and Susan Athey. “Battling the coronavirus "infodemic" among social media users in Kenya and Nigeria”. In: *Nature Human Behaviour* 8 (2024), pp. 823–834.

[108] Antonio A. Arechar et al. “Understanding and combatting misinformation across 16 countries on six continents”. In: *Nature Human Behaviour* 7 (2023), pp. 1502–1513.

[109] Emma Hoes et al. “Prominent misinformation interventions reduce misperceptions but increase scepticism”. In: *Nature Human Behaviour* 8 (2024), pp. 1545–1553.

[110] Ethan Porter and Thomas J. Wood. “The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom”. In: *PNAS* 118.37, e2104235118 (2021).

[111] Manoel Horta Ribeiro et al. “Deplatforming norm-violating influencers on social media reduces overall online attention toward them”. In: *CSCW*. 2025.

[112] Andrew Guess, Jonathan Nagler, and Joshua Tucker. “Less than you think: Prevalence and predictors of fake news dissemination on Facebook”. In: *Science Advances* 5.1, eaau4586 (2019).

[113] Andrew M. Guess, Brendan Nyhan, and Jason Reifler. “Exposure to untrustworthy websites in the 2016 US election”. In: *Nature Human Behaviour* 4 (2020), pp. 472–480.

[114] Francesco Pierri et al. “Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine”. In: *WebSci*. 2023.

[115] Valeria Pantè et al. “Beyond interaction patterns: Assessing claims of coordinated inter-state information operations on Twitter/X”. In: *WWW 2025*.

[116] Ryan C. Moore, Ross Dahlke, and Jeffrey T. Hancock. “Exposure to untrustworthy websites in the 2020 US election”. In: *Nature Human Behaviour* 7 (2023), pp. 1096–1105.

[117] Adrien Friggeri et al. “Rumor cascades”. In: *ICWSM*. 2014.

[118] Dominique Geissler and Stefan Feuerriegel. “Analyzing the strategy of propaganda using inverse reinforcement learning: Evidence from the 2022 Russian invasion of Ukraine”. In: *CSCW*. 2024.

[119] Dominik Bär, N. Pröllochs, and Stefan Feuerriegel. “Finding Qs: Profiling QAnon supporters on Parler”. In: *ICWSM*. 2023.

[120] Max Aliapoulios et al. “The gospel according to Q: Understanding the QAnon conspiracy from the perspective of canonical information”. In: *ICWSM*. 2022.

[121] Kristen Engel et al. “Characterizing Reddit participation of users who engage in the QAnon conspiracy theories”. In: *CSCW*. 2022.

[122] Mohamad Hoseini et al. “On the globalization of the QAnon conspiracy theory through Telegram”. In: *WebSci*. 2023.

[123] Daniel Röchert et al. “Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube”. In: *Information Systems* 103, 101866 (2022).

[124] Karishma Sharma, Emilio Ferrara, and Yan Liu. “Characterizing online engagement with disinformation and conspiracies in the 2020 U.S. Presidential Election”. In: *ICWSM*. 2022.

[125] Jing Zeng and Mike S. Schäfer. “Conceptualizing "Dark Platforms". Covid-19-related conspiracy theories on 8kun and Gab”. In: *Digital Journalism* 9.9 (2021), pp. 1321–1343.

[126] Daniel Romer and Kathleen Hall Jamieson. "Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S". In: *Social Science & Medicine* 263, 113356 (2020).

[127] Reid McIlroy-Young and Ashton Anderson. "From "Welcome new Gabbers" to the Pittsburgh synagogue shooting: The evolution of Gab". In: *ICWSM*. 2019.

[128] Daniel Jolley and Karen M. Douglas. "The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint". In: *British Journal of Psychology* 105.1 (2014), pp. 35–56.

[129] Deen Freelon, Alice Marwick, and Daniel Kreiss. "False equivalencies: Online activism from left to right". In: *Science* 369.6508 (2020), pp. 1197–1201.

[130] Ethan Zuckerman. "QAnon and the emergence of the unreal". In: *Journal of Design and Science* 6 (2019).

[131] Antonis Papasavva et al. ""Is it a Qoincidence?": An exploratory study of QAnon on Voat". In: *WWW*. 2021.

[132] Amarnath Amarasingam and Marc-André Argentino. "The QAnon conspiracy theory: A security threat in the making?" In: *CTC Sentinel* 13.7 (2020), pp. 37–44.

[133] Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. "No calm in the storm: Investigating QAnon website relationships". In: *ICWSM*. 2022.

[134] J. Hunter Priniski, Mason McClay, and Keith J. Holyoak. "Rise of QAnon: A mental model of good and evil stews in an echochamber". In: *CogSci*. 2021.

[135] Brian J. Holoyda. "QAnon: A modern conspiracy theory and the assessment of its believers". In: *The Journal of the American Academy of Psychiatry and the Law* 50.1 (2022).

[136] Daniel Taninecz Miller. "Characterizing QAnon: Analysis of YouTube comments presents new conclusions about a popular conservative conspiracy". In: *First Monday* 26.2 (2021).

[137] Savvas Zannettou et al. "What is Gab? A bastion of free speech or an alt-right echo chamber?" In: *WWW*. 2018.

[138] Thomas H. Costello, Gordon Pennycook, and David Gertler Rand. "Durably reducing conspiracy beliefs through dialogues with AI". In: *Science* 385.6714, eadq1814 (2024).

[139] Mitchell Linegar et al. "Prebunking elections rumors: Artificial intelligence assisted interventions increase confidence in American elections". In: *arXiv* 2410.19202 (2024).

[140] Kai Shu et al. "Fake news detection on social media". In: *KDD*. 2017.

[141] Jing Ma et al. "Detecting rumors from microblogs with recurrent neural networks". In: *IJCAI*. 2016.

[142] Kai Shu, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection". In: *WSDM*. 2019.

[143] Natali Ruchansky, Sungyong Seo, and Yan Liu. “CSI: A hybrid deep model for fake news detection”. In: *CIKM*. 2017.

[144] Prashanth Vijayaraghavan and Soroush Vosoughi. “TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations”. In: *NACCL*. 2022.

[145] Yang Liu and Yi-Fang Wu. “Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks”. In: *AAAI*. 2018.

[146] Francesco Barbieri et al. “TweetEval: Unified benchmark and comparative evaluation for tweet classification”. In: *EMNLP Findings*. 2020.

[147] Sai Saketh Aluru et al. “Deep learning models for multilingual hate speech detection”. In: *ECML-PKDD*. 2020.

[148] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach”. In: *Multimedia Tools and Applications* 80.8 (2021), pp. 11765–11788.

[149] Arnav Arora et al. “Detecting harmful content on online platforms: What platforms need vs. where research efforts go”. In: *ACM Computing Surveys* 56.3 (2024), pp. 1–17.

[150] Francesco Marra et al. “Detection of GAN-generated fake images over social networks”. In: *Conference on Multimedia Information Processing and Retrieval*. 2018.

[151] Greeshma Arya et al. “Multimodal hate speech detection in memes using contrastive language-image pre-training”. In: *IEEE Access* 12 (2024), pp. 22359–22375.

[152] Stefan Feuerriegel et al. “Using natural language processing to analyse text data in behavioural science”. In: *Nature Reviews Psychology* 4 (2025), pp. 96–111.

[153] Onur Varol et al. “Online human-bot interactions: Detection, estimation, and characterization”. In: *ICWSM*. 2017.

[154] Kai-Cheng Yang et al. “Scalable and generalizable social bot detection through data selection”. In: *ICWSM*. 2020.

[155] Sneha Kudugunta and Emilio Ferrara. “Deep neural networks for bot detection”. In: *Information Sciences* 467 (2018), pp. 312–322.

[156] Mohsen Sayyadiharikandeh et al. “Detection of novel social bots by ensembles of specialized classifiers”. In: *CIKM*. 2020.

[157] Despoina Chatzakou et al. “Mean birds: Detecting aggression and bullying on Twitter”. In: *WebSci*. 2017.

[158] Chengcheng Shao et al. “The spread of low-credibility content by social bots”. In: *Nature Communications* 9, 4787 (2018).

[159] Dominik Bär et al. “Analyzing social media activities at Bellingcat”. In: *WebSci*. 2023.

[160] Giuseppe Russo et al. “Understanding online migration decisions following the banning of radical communities”. In: *WebSci*. 2023.

[161] Wajiha Shahid et al. "Are you a cyborg, bot or human? A survey on detecting fake news spreaders". In: *IEEE Access* 10 (2022), pp. 27069–27083.

[162] Valerio Basile et al. "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter". In: *International Workshop on Semantic Evaluation*. 2019.

[163] Arianna D'Ulizia et al. "Fake news detection: A survey of evaluation datasets". In: *PeerJ. Computer science* 7, e518 (2021).

[164] William Yang Wang. ""Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection". In: *arXiv* 1705.00648 (2017).

[165] Binny Mathew et al. "HateXplain: A benchmark dataset for explainable hate speech detection". In: *AAAI*. 2021.

[166] Jing Qian et al. "A benchmark dataset for learning to intervene in online hate speech". In: *EMNLP and IJCNLP*. 2019.

[167] Manuel Tonneau et al. "NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data". In: *ACL*. 2024.

[168] "SemEval-2020 task 11: Detection of propaganda techniques in news articles". In: *Workshop on Semantic Evaluation*. 2020.

[169] Giovanna Menardi and Nicola Torelli. "Training and assessing classification rules with imbalanced data". In: *Data Mining and Knowledge Discovery* 28.1 (2014), pp. 92–122.

[170] Sayan Ghosh et al. "Detecting cross-geographic biases in toxicity modeling on social media". In: *Workshop on noisy user-generated text*. 2021.

[171] Maximilian Wich et al. "Investigating annotator bias in abusive language datasets". In: *International Conference on Recent Advances in Natural Language Processing*. 2021.

[172] Fabio Poletto et al. "Resources and benchmark corpora for hate speech detection: A systematic review". In: *Language Resources and Evaluation* 55.2 (2021), pp. 477–523.

[173] Gizem Ceylan, Ian A. Anderson, and Wendy Wood. "Sharing of misinformation is habitual, not just lazy or biased". In: *PNAS* 120.4, e2216614120 (2023).

[174] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. "The virality of hate speech on social media". In: *CSCW*. 2024.

[175] Isaac Slaughter et al. "Community notes moderate engagement with and diffusion of false information online". In: *arXiv* 2502.13322 (2025).

[176] Jonas L. Juul and Johan Ugander. "Comparing information diffusion mechanisms by matching on cascade size". In: *PNAS* 118.46 (2021), e2100786118.

[177] Claire Robertson et al. "Negativity drives online news consumption". In: (2022).

[178] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239 (2015), pp. 1130–1132.

[179] Aleksandra Korolova. “Privacy violations using microtargeted ads: A case study”. In: *International Conference on Data Mining Workshops*. 2011.

[180] Brooke Auxier. “54% of Americans say social media companies shouldn’t allow any political ads”. In: *Pew Research Center* (2020). URL: https://www.pewresearch.org/short-reads/2020/09/24/54-of-americans-say-social-media-companies-shouldnt-allow-any-political-ads/.

[181] José González Cabañas et al. “Does Facebook use sensitive data for advertising purposes?” In: *Communications of the ACM* 64.1 (2021), pp. 62–69.

[182] Jon Green et al. “Curation bubbles”. In: *American Political Science Review* (2025), pp. 1–19.

[183] Siva Vaidhyanathan. “Facebook wins, democracy loses”. In: *The New York Times* (2017). URL: https://www.nytimes.com/2017/09/08/opinion/facebook-wins-democracy-loses.html.

[184] Till Speicher et al. “Potential for discrimination in online targeted advertising”. In: *FAccT*. 2018.

[185] Aruna Sankaranarayanan, Erik Hemberg, and Una-May O’Reilly. “The Facebook algorithm’s active role in climate advertisement delivery”. In: *arXiv* 2308.03191 (2023).

[186] Jim Isaak and Mina J. Hanna. “User data privacy: Facebook, Cambridge Analytica, and privacy protection”. In: *Computer* 51.8 (2018), pp. 56–59.

[187] Katharine Dommett. “The rise of online political advertising”. In: *Political Insight* 10.4 (2019), pp. 12–15.

[188] Minali Aggarwal et al. “A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout”. In: *Nature Human Behaviour* 7 (2023), pp. 332–341.

[189] Anselm Hager. “Do online ads influence vote choice?” In: *Political Communication* 36.3 (2019), pp. 376–393.

[190] Alexander Coppock, Donald P. Green, and Ethan Porter. “Does digital advertising affect vote choice? Evidence from a randomized field experiment”. In: *Research & Politics* 9.1, 20531680221076901 (2022).

[191] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. “The role of social media ads for election outcomes: Evidence from the 2021 German election”. In: *PNAS Nexus* 4.3, pgaf073 (2025).

[192] United States Senate. *The Platform Accountability and Transparency Act*. 2023. URL: https://www.congress.gov/bill/118th-congress/senate-bill/1876/text.

[193] Meta. *Meta Content Library and Meta Content Library API*. 2025. URL: https://transparency.meta.com/researchtools/meta-content-library/.

[194] YouTube. *YouTube researcher program*. 2025. URL: https://research.youtube/.

[195] TikTok. *Research tools*. 2025. URL: `https://developers.tiktok.com/products/research-api/`.

[196] X. *Academic research*. 2025. URL: `https://developer.x.com/en/use-cases/do-research/academic-research`.

[197] Dominik Bär et al. "Systematic discrepancies in the delivery of political ads on Facebook and Instagram". In: *PNAS Nexus* 3.7, pgae247 (2024).

[198] Erika Franklin Fowler et al. "Political advertising online and offline". In: *American Political Science Review* 115.1 (2021), pp. 130–149.

[199] Arthur Capozzi et al. "Facebook ads: Politics of migration in Italy". In: *International Conference on Social Informatics*. 2020.

[200] Arthur Capozzi et al. "Clandestino or rifugiato? Anti-immigration Facebook ad targeting in Italy". In: *CHI*. 2021.

[201] Arthur Capozzi et al. "The thin ideology of populist advertising on Facebook during the 2019 EU elections". In: *WWW*. 2023.

[202] Laurenz Aisenpreis, Gustav Gyrst, and Vedran Sekara. "How do US congress members advertise climate change: An analysis of ads run on Meta's platforms". In: *ICWSM*. 2023.

[203] Sandra González-Bailón et al. "Asymmetric ideological segregation in exposure to political news on Facebook". In: *Science* 381.6656 (2023), pp. 392–398.

[204] Manoel Horta Ribeiro, Justin Cheng, and Robert West. "Automated content moderation increases adherence to community guidelines". In: *WWW*. 2023.

[205] Hunt Allcott et al. "The effects of Facebook and Instagram on the 2020 election: A deactivation experiment". In: *PNAS* 121.21, e2321584121 (2024).

[206] Kai Kupferschmidt. "Studies find little impact of social media on polarization". In: *Science* 381.6656 (2023), pp. 367–368.

[207] Márcio Silva et al. "Facebook ads monitor: An independent auditing system for political ads on Facebook". In: *WWW*. 2020.

[208] J. Nathan Matias, Austin Hounsel, and Nick Feamster. "Software-supported audits of decision-making systems: Testing Google and Facebook's political advertising policies". In: *CSCW*. 2022.

[209] Kiran Garimella and Simon Chauchard. "WhatsApp Explorer: A data donation tool to facilitate research on WhatsApp". In: *Mobile Media & Communication* Forthcoming (2025).

[210] Afsaneh Razi et al. "Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection". In: *CHI*. 2022.

[211] Jakob Ohme et al. "Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking". In: *Communication Methods and Measures* 18.2 (2024), pp. 124–141.

[212] Fabio Votta et al. "The cost of reach: Testing the role of ad delivery algorithms in online political campaigns". In: *Political Communication* 42.3 (2024), pp. 476–508.

[213] John Sides, Lynn Vavreck, and Christopher Warshaw. "The effect of television advertising in United States elections". In: *American Political Science Review* 116.2 (2022), pp. 702–718.

[214] Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. "Automatic sarcasm detection". In: *ACM Computing Surveys* 50.5 (2018), pp. 1–22.

[215] Micol Spitale, Fabio Catania, and Francesca Panzeri. "Understanding non-verbal irony markers: Machine learning insights versus human judgment". In: *International Conference on Multimodel Interaction*. 2024.

[216] Miriah Steiger et al. "The psychological well-being of content moderators". In: *CHI*. 2021.

[217] Nicholas Diakopoulos and Mor Naaman. "Towards quality discourse in online news comments". In: *CSCW*. 2011.

[218] Shan Jiang, Ronald E. Robertson, and Christo Wilson. "Reasoning about political bias in content moderation". In: *ICWSM*. 2020.

[219] Jialun Aaron Jiang et al. "A trade-off-centered framework of content moderation". In: *ACM Transactions on Computer-Human Interaction* 30.1 (2023), pp. 1–34.

[220] Erika Franklin Fowler, Michael M. Franz, and Travis N. Ridout. "Online political advertising in the United States". In: *Social Media and Democracy: The State of the Field and Prospects for Reform*. Ed. by Nathaniel Persily and Joshua A. Tucker. Cambridge, UK: Cambridge University Press, 2020, pp. 111–138.

[221] Center for Countering Digital Hate. *X content moderation failure: How Twitter/X continues to host posts we reported for extreme hate speech*. 2023. URL: `https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/`.

[222] Center for Countering Digital Hate. *Twitter fails to act on 99% of Twitter Blue accounts tweeting hate*. 2023. URL: `https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/`.

[223] Julian Jarusch. *The digital services act is in effect - now what?* 2024. URL: `https://www.stiftung-nv.de/sites/default/files/snv-digital-services-act-now-what.pdf`.

[224] European Commission. *DSA Transparency Database*. 2025. URL: `https://transparency.dsa.ec.europa.eu/`.

[225] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. "Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia". In: *CHI*. 2008.

[226] Eshwar Chandrasekharan et al. "The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales". In: *CSCW*. 2018.

[227] Sarah A. Gilbert. “"I run the world’s largest historical outreach project and it’s on a cesspool of a website." Moderating a public scholarship site on Reddit: A case study of r/AskHistorians”. In: *CSCW*. 2020.

[228] Shagun Jhaver et al. “Human-machine collaboration for content regulation: The case of Reddit automoderator”. In: *TOCHI*. 2019.

[229] Joseph Seering, Robert Kraut, and Laura Dabbish. “Shaping pro and anti-social behavior on Twitch through moderation and example-setting”. In: *CSCW*. 2017.

[230] Kumar Bhargav Srinivasan et al. “Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView Community”. In: *CSCW*. 2019.

[231] Lorenzo Cima et al. “Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation”. In: *WWW 2025*.

[232] Lisa P. Argyle et al. “Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale”. In: *PNAS* 120.41, e2311627120 (2023).

[233] Pamela Grimm. “Social desirability bias”. In: *Wiley International Encyclopedia of Marketing*. Ed. by Jagdish Sheth and Naresh Malhotra. Hoboken, NJ, US: Wiley, 2010.

[234] Mohsen Mosleh, Gordon Pennycook, and David G. Rand. “Field experiments on social media”. In: *Current Directions in Psychological Science* 31.1 (2022), pp. 69–75.

[235] Retraction Watch. *Experiment using AI-generated posts on Reddit draws fire for ethics concerns*. 2025. URL: https://retractionwatch.com/2025/04/28/experiment-using-ai-generated-posts-on-reddit-draws-fire-for-ethics-concerns/.

[236] Dominik Bär, Abdurahman Maarouf, and Stefan Feuerriegel. “Generative AI may backfire for counterspeech”. In: *arXiv* 2411.14986 (2024).

[237] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. “Emotions in online rumor diffusion”. In: *EPJ Data Science* 10, 51 (2021).

[238] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. “Emotions explain differences in the diffusion of true vs. false social media rumors”. In: *Scientific Reports* 11.1, 22721 (2021).

[239] Meta. *Ad Library*. 2022. URL: https://www.facebook.com/ads/library/.

[240] Meta. *Ad Targeting Dataset*. 2023. URL: https://developers.facebook.com/docs/fort-ads-targeting-dataset.

[241] Francesco Pierri. “Political advertisement on Facebook and Instagram in the run up to 2022 Italian general election”. In: *WebSci*. 2023.

[242] Piotr Sapiezynski et al. “On the use of proxies in political ad targeting”. In: *CSCW*. 2024.

[243] Clare Duffy. “Meta is getting rid of fact checkers. Zuckerberg acknowledged more harmful content will appear on the platforms now”. In: *CNN* (2025). URL: https://edition.cnn.com/2025/01/07/tech/meta-censorship-moderation.

[244] Adrian Kopps. "Two years after the takeover: Four key policy changes of X under Musk". In: *Alexander von Humboldt Institute for Internet and Society* (2024). URL: https://doi.org/10.5281/zenodo.14040407.

[245] Yona T. R. Golding. "How politics broke content moderation". In: *Columbia Journalism Review* (2024). URL: https://www.cjr.org/covering_the_election/how-politics-broke-content-moderation-hamilton-68-elon-musk.php.

[246] Aidan Combs et al. "Reducing political polarization in the United States with a mobile chat platform". In: *Nature Human Behaviour* 7 (2023), pp. 1454–1461.

[247] Petter Törnberg et al. "Simulating social media using large language models to evaluate alternative news feed algorithms". In: *arXiv* 2310.05984 (2023).

[248] Yue Jia and Sandy Schumann. "Tackling hate speech online: The effect of counter-speech on subsequent bystander reactions". In: *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 19.1, 4 (2025).

[249] Chenyan Jia et al. "Embedding democratic values into social media AIs via societal objective functions". In: *CSCW*. 2024.

[250] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests". In: *The Annals of Statistics* 47.2 (2019).

[251] Stefan Wager and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242.

[252] Konstantin Hess and Stefan Feuerriegel. "Stabilized neural prediction of potential outcomes in continuous time". In: *ICLR*. 2025.

[253] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. "Causal transformer for estimating counterfactual outcomes". In: *ICML*. 2022.

[254] Dennis Frauen and Stefan Feuerriegel. "Estimating individual treatment effects under unobserved confounding using binary instruments". In: *ICLR*. 2023.

[255] Jonas Schweisthal et al. "Learning representations of instruments for partial identification of treatment effects". In: *ICML*. 2025.

[256] Yuxin Wang et al. "Constructing confidence intervals for average treatment effects from multiple datasets". In: *ICLR*. 2025.

[257] Maresa Schröder, Valentyn Melnychuk, and Stefan Feuerriegel. "Differentially private learners for heterogeneous treatment effects". In: *ICLR*. 2025.

[258] Yuchen Ma et al. "DiffPO: A causal diffusion model for learning distributions of potential outcomes". In: *NeurIPS*. 2024.

[259] Konstantin Hess et al. "Bayesian neural controlled differential equations for treatment effect estimation". In: *ICLR*. 2024.

[260] Carlos Cinelli and Chad Hazlett. "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1 (2020), pp. 39–67.

[261] Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. “Sharp bounds for generalized causal sensitivity analysis”. In: *NeurIPS*. 2023.

[262] Dennis Frauen et al. “A neural framework for generalized causal sensitivity analysis”. In: *ICLR*. 2024.

[263] Maresa Schröder, Dennis Frauen, and Stefan Feuerriegel. “Causal fairness under unobserved confounding: A neural sensitivity framework”. In: *ICLR*. 2024.

[264] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. “Dancing to the partisan beat: A first analysis of political communication on TikTok”. In: *WebSci*. 2020.

[265] Simon Chauchard and Kiran Garimella. “What circulates on partisan WhatsApp in India? Insights from an unusual dataset”. In: *Journal of Quantitative Description: Digital Media* 2 (2022), pp. 1–42.

[266] Tiago Ventura et al. “WhatsApp increases exposure to false rumors but has limited effects on beliefs and polarization: Evidence from a multimedia-constrained deactivation”. In: *SSRN* 4457400 (2023).

[267] Rajeshwari Majumdar. “Reducing prejudice and support for religious nationalism through conversations on WhatsApp”. In: *SSRN* 4569815 (2023).

[268] Punyajoy Saha et al. “"Short is the Road that Leads from Fear to Hate": Fear speech in Indian WhatsApp groups”. In: *WWW*. 2021.

Part I: Detecting and Understanding harmful content

7 Finding Qs: Profiling QAnon supporters on Parler

Title: Finding Qs: Profiling QAnon supporters on Parler

Abstract: The social media platform "Parler" has emerged into a prominent fringe community where a significant part of the user base are self-reported supporters of QAnon, a far-right conspiracy theory alleging that a cabal of elites controls global politics. QAnon is considered to have had an influential role in the public discourse during the 2020 U.S. presidential election. However, little is known about QAnon supporters on Parler and what sets them aside from other users. Building up on social identity theory, we aim to profile the characteristics of QAnon supporters on Parler. We analyze a large-scale dataset with more than 600,000 profiles of English-speaking users on Parler. Based on users' profiles, posts, and comments, we then extract a comprehensive set of user features, linguistic features, network features, and content features. This allows us to perform user profiling and understand to what extent these features discriminate between QAnon and non-QAnon supporters on Parler. Our analysis is three-fold: (1) We quantify the number of QAnon supporters on Parler, finding that 34,913 users (5.5 % of all users) openly report supporting the conspiracy. (2) We examine differences between QAnon vs. non-QAnon supporters. We find that QAnon supporters differ statistically significantly from non-QAnon supporters across multiple dimensions. For example, they have, on average, a larger number of followers, followees, and posts, and thus have a large impact on the Parler network. (3) We use machine learning to identify which user characteristics discriminate QAnon from non-QAnon supporters. We find that user features, linguistic features, network features, and content features, can – to a large extent – discriminate QAnon vs. non-QAnon supporters on Parler. In particular, we find that user features are highly discriminatory, followed by content features and linguistic features.

Author contributions: Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel contributed to conceptualization. Dominik Bär performed the data analysis. Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel contributed to results interpretation and manuscript writing.

Reference: Bär, D., Pröllochs, N., & Feuerriegel, S. 2023. *Finding Qs: Profiling QAnon supporters on Parler.* In: **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)** 17 (1), 34-46, DOI: `https://doi.org/10.1609/icwsm.v17i1.22124` (Full paper).

Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Finding Qs: Profiling QAnon Supporters on Parler

Dominik Bär[1,2], Nicolas Pröllochs[3], Stefan Feuerriegel[1,2]
[1]LMU Munich, Munich, Germany
[2]Munich Center for Machine Learning, Munich, Germany
[3]JLU Giessen, Giessen, Germany
baer@lmu.de, nicolas.proellochs@wi.jlug.de, feuerriegel@lmu.de

Abstract

The social media platform "Parler" has emerged into a prominent fringe community where a significant part of the user base are self-reported supporters of QAnon, a far-right conspiracy theory alleging that a cabal of elites controls global politics. QAnon is considered to have had an influential role in the public discourse during the 2020 U.S. presidential election. However, little is known about QAnon supporters on Parler and what sets them aside from other users. Building up on social identity theory, we aim to profile the characteristics of QAnon supporters on Parler. We analyze a large-scale dataset with more than 600,000 profiles of English-speaking users on Parler. Based on users' profiles, posts, and comments, we then extract a comprehensive set of user features, linguistic features, network features, and content features. This allows us to perform user profiling and understand to what extent these features discriminate between QAnon and non-QAnon supporters on Parler. Our analysis is three-fold: (1) We quantify the number of QAnon supporters on Parler, finding that 34,913 users (5.5 % of all users) openly report supporting the conspiracy. (2) We examine differences between QAnon vs. non-QAnon supporters. We find that QAnon supporters differ statistically significantly from non-QAnon supporters across multiple dimensions. For example, they have, on average, a larger number of followers, followees, and posts, and thus have a large impact on the Parler network. (3) We use machine learning to identify which user characteristics discriminate QAnon from non-QAnon supporters. We find that user features, linguistic features, network features, and content features, can – to a large extent – discriminate QAnon vs. non-QAnon supporters on Parler. In particular, we find that user features are highly discriminatory, followed by content features and linguistic features.

Introduction

The social media platform "Parler" has emerged into a prominent fringe community, where a significant user base identifies with far-right viewpoints. Parler was founded in August 2018 and promotes open self-expression and free speech (Aliapoulios et al. 2021; Otala et al. 2021). Its user base grew during the 2020 U.S. presidential election and, in January 2021, counted 13.255 M users (Aliapoulios et al. 2021).

In recent years, Parler gained widespread public interest when conservative thought leaders endorsed the platform as an alternative to mainstream social media platforms (Aliapoulios et al. 2021). As such, a large portion of the content on Parler is characterized by conservative viewpoints. For example, many discussions revolve around former U.S. President Donald Trump, the 2020 U.S. presidential election, and religion (Aliapoulios et al. 2021). In addition, Parler users were associated with the storming of the U.S. Capitol on January 6, 2021 (Hitkul et al. 2021). The latter eventually led to the removal of the Parler app from both the Google Play Store and Apple's App Store, as well as Amazon stopping to host the website (Aliapoulios et al. 2022). As a result, Parler went offline on January 11, 2021 but was reinstated on February 15, 2021 after finding a new hosting service (Robertson 2021).

Besides conservative viewpoints, Parler users frequently identify with various conspiracy theories (Aliapoulios et al. 2022; Hitkul et al. 2021). Different from (necessary) healthy skepticism of government and elites, beliefs in false narratives and conspiracies can pose serious threats to democracy and public safety (Geissler et al. 2022; Sternisko, Cichocka, and van Bavel 2020). As a particularly concerning example, Parler users have been associated with QAnon, a far-right conspiracy theory in which supporters allege that a cabal of elites controls global politics, conspires in a pedophile ring, and engages in destroying society (Aliapoulios et al. 2022; Amarasingam and Argentino 2020; Hanley, Kumar, and Durumeric 2022). QAnon combines several conspiracy theories, and, as such, research has characterized it as a "meta narrative" (Zuckerman 2019) or "super-conspiracy theory" (Papasavva et al. 2021) that quickly gained a large base of supporters and expanded globally (Hoseini et al. 2021).

The prominence of QAnon has led to widespread concerns. One reason is that QAnon supporters were frequently associated with violent incidents (Hoseini et al. 2021). For example, several QAnon supporters participated in the storming of the U.S. Capitol on January 6, 2021 (Aliapoulios et al. 2022). As a result, the U.S. Federal Bureau of Investigation (FBI) has labeled QAnon a national security threat (Amarasingam and Argentino 2020). Moreover, major social media networks such as Reddit, Twitter, and Facebook have banned QAnon-related content from their platforms (Papasavva et al. 2021), and, as a consequence, QAnon supporters migrated – to a large extent – to fringe communities such as Parler (Aliapoulios et al. 2022). However, a user-centric view profiling QAnon supporters on Parler is missing and presents our novelty.

Despite the above, the total size of the QAnon community on Parler remains unclear. Yet, understanding the size of the community is crucial to evaluate the security threat that emanates from QAnon on Parler and motivates us to quantify the number of self-reported QAnon supporters on the platform. Furthermore, little is known about characteristics that discriminate QAnon supporters from non-supporters. Parler is characterized by a largely right-leaning community (Otala et al. 2021), which leads to a rather homogeneous user base compared to mainstream platforms such as Twitter or Facebook. As a consequence, QAnon supporters and other users on Parler might – at least to some extent – share a similar social identity and might hardly be distinguishable. Social identity theory argues that the behavior of people depends on – and varies across – group membership (Tajfel and Turner 1986), which can strengthen beliefs in conspiracy theories and drive behavior (Sternisko, Cichocka, and van Bavel 2020). Adherents of conspiracy theories might, for example, want to improve their self-image or the reputation of the group they belong to and find explanations for their beliefs (Sternisko, Cichocka, and van Bavel 2020). In fact, previous research on Facebook and Twitter showed that conspiracy theorists behave differently compared to other users (Bessi et al. 2015; Sharma, Ferrara, and Liu 2022). However, it is unknown if the same holds for QAnon users on Parler. Gaining such knowledge would have important practical implications as it might help surveillance of conspiracy theorists and improve early detection of upcoming security threats.

Research Questions: In this paper, we profile self-reported QAnon supporters on Parler based on a large-scale public snapshot of the Parler network (Aliapoulios et al. 2021). In particular, we seek to answer the following research questions (RQs):

- **RQ1:** *How many users on Parler are self-reported QAnon supporters?*
- **RQ2:** *How do user characteristics differ between self-reported QAnon supporters and other users on Parler?*
- **RQ3:** *Which user characteristics allow machine learning to discriminate self-reported QAnon supporters from other users on Parler?*

Contributions: By addressing the above RQs, we aim at profiling the characteristics of self-reported QAnon supporters on Parler. Our contribution is three-fold: (1) We quantify the number of self-reported QAnon supporters on Parler. (2) We make use of a user's profile, friendships, posts, and comments, based on which we extract a comprehensive set of user features, linguistic features, network features, and content features. We then compare QAnon vs. non-QAnon supporters along these features to identify user characteristics that distinguish both groups. (3) We use machine learning methods to identify which user characteristics can be leveraged to discriminate QAnon supporters from non-QAnon supporters.

Related Work

User Profiling: Social networks attract a large and diverse user base. While users disclose some personal information, other characteristics can be kept private. As a result, research has leveraged data from social media platforms to profile different user groups. For example, one study seeks to understand what distinguishes verified from non-verified users on Twitter (Paul et al. 2019). Other works use public information of users to predict their gender, age, and geographic origin, even if such information is kept private (Burger et al. 2011; Rao et al. 2010). Even other works leverage personal information to learn about the socioeconomic status and income of users (Lampos et al. 2016; Preoţiuc-Pietro et al. 2015; Rao et al. 2010). Previous literature has also studied the characteristics of users engaging in online abuse (Chatzakou et al. 2017b,a; ElSherief et al. 2018; Gorrell et al. 2018; Hua, Naaman, and Ristenpart 2020; Maity et al. 2018; Ribeiro et al. 2018). Here, we add by profiling QAnon supporters.

Fringe Online Communities: Fringe online communities (sometimes also referred to as alt-tech) are characterized by open self-expression and free speech (Aliapoulios et al. 2021; Bär, Pröllochs, and Feuerriegel 2023; Papasavva et al. 2021). Examples of corresponding social media platforms are Gab, 4chan, Voat, BitChute, Gettr, and Parler. Over the last years, they attracted large numbers of new users that were dissatisfied with or banned from mainstream communities such as Twitter or Facebook (i. e., de-platforming) as many of these are increasingly invested in content moderation efforts (Aliapoulios et al. 2021; Otala et al. 2021). As such, fringe communities attract a different user base compared to mainstream communities characterized by extreme viewpoints. In particular, many fringe communities host users from the far-right political spectrum or conspiracy theorists (Aliapoulios et al. 2021, 2022; Papasavva et al. 2021; Zannettou et al. 2018).

There is a growing interest in computational social science to research the behavior of fringe communities at Gab (Ali et al. 2021; Zannettou et al. 2018), 4chan (Hine et al. 2017; Papasavva et al. 2020), Voat (Papasavva et al. 2021), BitChute (Trujillo et al. 2020), Gettr (Paudel et al. 2022), or Parler (Aliapoulios et al. 2021; Baines, Ittefaq, and Abwao 2021; Jakubik et al. 2023; Munn 2021; Otala et al. 2021; Pieroni et al. 2021). In this work, we focus on Parler due to its alleged role in inciting violence, disseminating extreme far-right content (Aliapoulios et al. 2021; Hitkul et al. 2021), and its relevance for hosting conspiracy theories as is the case for QAnon (Aliapoulios et al. 2021).

Parler: Parler operates as a microblogging service similar to Twitter and has grown a significant fringe online community over the last years. The Parler community is characterized by extreme viewpoints and accused of partially coordinating the storming of the U.S. Capitol on January 6, 2021 (Aliapoulios et al. 2021). Parler is subject to heavy political polarization (Munn 2021; Otala et al. 2021). In fact, previous research has shown that many politicians from the Republican Party, as well as their followers, migrated from Twitter to Parler during the 2020 U.S. presidential election and around the storming of the U.S. Capitol (Otala et al. 2021) dissatisfied with increasing content moderation on Twitter. The authors suggest that this migration has contributed to a Parler community that is – to a large extent – right-leaning.

Research examining the social media platform Parler has only recently received traction. One stream of literature is interested in cross-platform comparisons, whereby content from Parler is compared against, e. g., Twitter. Here, findings suggest that Parler encompasses views and emotions that differ from those shared on Twitter, especially with regards to the storming of the U.S. Capitol on January 6, 2021 (Hitkul et al. 2021; Jakubik et al. 2023). For example, many Parler users are supportive of the strongly conservative ideology of the rioters, whereas most users on Twitter condemned the riots.

Another stream of literature has studied *what* content is posted on Parler. On the one hand, content appears to frequently reference former U.S. President Donald Trump (Aliapoulios et al. 2021). On the other hand, content on Parler was found to cover various conspiracy theories (Aliapoulios et al. 2021; Pieroni et al. 2021). In this regard, content on Parler was also found to make frequent references to QAnon (Aliapoulios et al. 2021; Sipka, Hannak, and Urman 2022). However, other than that, there is a paucity of works analyzing the online behavior of QAnon supporters on Parler.

QAnon: QAnon emerged supposedly in 2017 (Papasavva et al. 2021), when an anonymous user named "Q" posted a thread named "Calm before the storm" stating that former U.S. President Donald Trump is leading the fight against a cabal of elites (Aliapoulios et al. 2022). Ever since, cryptic pieces of information have appeared online, which are decrypted by QAnon supporters informing them about the fight against the cabal of elites (Aliapoulios et al. 2022).

A large part of the literature on QAnon focuses on the online diffusion of the conspiracy theory (Aliapoulios et al. 2022; Hanley, Kumar, and Durumeric 2022; Hoseini et al. 2021; Priniski, McClay, and Holyoak 2021). Scholars have analyzed posts attributed to "Q" (so-called "Q drops") from different aggregation sites (Aliapoulios et al. 2022). However, they find only little agreement between the Q drops from different websites, thus implying that Q drops are written by multiple people. In this regard, prior literature has provided a detailed overview of QAnon discussions on, e. g., Voat (Papasavva et al. 2021), or compared content across platforms (Sipka, Hannak, and Urman 2022). Furthermore, there exists evidence that the QAnon conspiracy theory is also discussed on other platforms such as Twitter (Sharma, Ferrara, and Liu 2022), YouTube (Miller 2021), 8kun (Aliapoulios et al. 2022), or Parler (Aliapoulios et al. 2022).

Previous research has also studied QAnon supporters (Engel et al. 2022; Papasavva et al. 2021; Priniski, McClay, and Holyoak 2021; Sharma, Ferrara, and Liu 2022). In the case of Twitter, QAnon supporters engage in active discussions and effectively circumvent content moderation (Sharma, Ferrara, and Liu 2022). Moreover, QAnon supporters tend to disseminate rather than produce content (Priniski, McClay, and Holyoak 2021). Similarly, QAnon supporters actively participated in discussions on Reddit and frequently shared low-quality links (Engel et al. 2022). For the QAnon-related subverse on Voat, it was found that only a few users are responsible for writing the majority of posts, while comments are made by a large base of subscribers (Papasavva et al. 2021). However, none of these studies profile QAnon supporters on Parler, especially not with the objective of understanding which user characteristics discriminate QAnon supporters from non-QAnon supporters.

Research Gap: Little is known about which user characteristics can discriminate QAnon supporters from non-QAnon supporters. In this paper, we close this research gap by profiling the characteristics of QAnon supporters on Parler.

Data

Our analysis is based on a large-scale public snapshot of the Parler social network (Aliapoulios et al. 2021). Specifically, we collect data from 638,865 English-speaking users between the founding of Parler in August 2018 to its shutdown on January 11, 2021. These users posted approximately 158 M posts and 42 M comments. Overall, the size of our dataset is comparatively large-scale, especially when compared with user profiling in other contexts (e. g., Lampos et al. 2016; Matero et al. 2019; Preoţiuc-Pietro et al. 2015).

Our dataset contains information on a user's friendship network (i. e., followers, followees), online activity (i. e., number of posts, number of comments), and popularity (i. e., number of upvotes per post, impressions per post, up/downvotes per comment). Furthermore, for each account, we collect information about the date on which users joined Parler and the user bios. User bios on Parler are short self-descriptions of Parler users. They appear on the front page of a user's profile and are similar to user bios on Twitter. For the purpose of our study, we only focus on users with non-empty bios. Our data also contains the content posted by each user from August 2018 to January 11, 2021.

Methods[1]

Theoretical Motivation

In this work, we aim to quantify the number of self-reported QAnon supporters on Parler (**RQ1**); compare them to non-QAnon supporters across different user characteristics (**RQ2**); and, eventually, examine which user characteristics can discriminate both QAnon vs. non-QAnon supporters using machine learning (**RQ3**).

There are several reasons why we may or may not find differences with regard to **RQ2/RQ3**. On the one hand, the political orientation on Parler is largely right-leaning (Otala et al. 2021). This leads to a generally more homogeneous user base compared to mainstream communities such as Twitter or Facebook (where, in the latter, users are still exposed to diverse ideological content (Bakshy, Messing, and Adamic 2015)). Within Parler's right-leaning environment, users largely share similar views on many matters, such as their support for Donald Trump. Hence, QAnon supporters on Parler might exhibit – at least to some extent –

[1]Code is available via https://github.com/DominikBaer95/Parler_QAnon_UserProfiling.

the social identity of non-supporters. This would imply that QAnon supporters and other users are non-distinguishable; we would thus see no differences with regard to **RQ2** and **RQ3**. On the other hand, social identity theory argues that the behavior of people varies across different groups (Tajfel and Turner 1986). Hence, people are drawn towards conspiracy theories due to specific social identity motivators (Sternisko, Cichocka, and van Bavel 2020). Here, adherents might want to improve their self-image or the reputation of the group they belong to and find explanations for their environment (Sternisko, Cichocka, and van Bavel 2020). Hence, adherents of conspiracy theories might behave differently to align with these motivations. In fact, previous research on Facebook showed that conspiracy theorists behave differently (e. g., share more content related to their views) as compared to other communities (Bessi et al. 2015). In addition, research from Twitter points out that QAnon supporters are particularly active in engaging on the platform (Sharma, Ferrara, and Liu 2022). Overall, this would imply that there are indeed differences with regard to **RQ2** and **RQ3**. Resolving this opposition motivates our work to profile user characteristics of QAnon supporters on Parler.

Identifying QAnon Supporters

To quantify the size of the QAnon community on Parler, we first classify users into whether they self-report as QAnon supporters. For this, we follow the approach in Sharma, Ferrara, and Liu (2022) and apply a keyword list to a user's bio in order to classify users. The theoretical reasoning behind this approach is that the user bios are supposed to reflect the personal and social identity of a user (Li et al. 2020; Pathak, Madani, and Joseph 2021; Rogers and Jones 2021). As such, user bios are predictive of one's identity, including gender (Burger et al. 2011; Pathak, Madani, and Joseph 2021), personal interests (Ding and Jiang 2014), and political orientation (Pathak, Madani, and Joseph 2021; Rogers and Jones 2021), while we here adapt the approach to supporters of conspiracy theories.

We adopt an extensive list of QAnon-specific keywords from prior literature (Sharma, Ferrara, and Liu 2022) to identify QAnon supporters. These capture different aspects of the underlying conspiracy theory, including the final fight against the cabal (e. g., "thestorm", "thegreatawakening"), the alleged child sex trafficking ring run by the cabal (e. g., "saveourchildren", "adrenochrome"), or related conspiracy theories (e. g., "pizzagate", "obamagate"). Hence, a user is counted as a *QAnon supporter* if one of the keywords appears on a user's bio. Otherwise, we refer to the user as *non-QAnon supporter*. Example user bios of QAnon vs. non-QAnon supporters are shown in Tbl. 1.

Validation Study: We validate the reliability of the above keyword approach as follows. Specifically, we performed a user study on the online survey platform Prolific (www.prolific.co) with $n = 7$ participants in order to assess whether our keyword approach leads to similar results as human assessment. Prior to our user study, all participants were trained in our task by providing them with background information on the QAnon conspiracy theory (and what sets it apart from conservative ideology). Subsequently, the participants were shown 100 randomly sampled user bios (50 QAnon and 50 non-QAnon) and then asked to assess to what extent the corresponding identifies with QAnon. For this, we use a Likert scale ranging from -3 to $+3$, where -3 corresponds to a user who does not identify with QAnon at all, while $+3$ refers to a strong identification.

We find a statistically significant interrater agreement in terms of Kendall's W ($W = 0.76$; $p < 0.01$). We then evaluate to what extent the rater assessments and our keyword approach agree. For this, we first map the Likert scale onto a binary label (QAnon yes/no), whereby a user bio is classified as QAnon (non-QAnon) if a rater has labeled the user bio with a Likert rating of $+1$ or higher (zero or below). We find a large overlap between our keyword approach and the human assessment. In fact, we find that only 4 % of QAnon and 2 % of non-QAnon supporters were classified incorrectly by the keyword approach, whereas the majority of class labels are in agreement[2]. Consequently, our keyword approach can reliably discern (non-)QAnon supporters on Parler.

Class label	Bios (examples)
QAnon	*"Truth Seeking Psychic Passionate About Revealing False Narratives Perpetrated by The 1 % #WWG1WGA Here To Help Provide Insight And Awaken Humanity #The Great Awakening"*
	"Lover of God, Family, Country, and First Responders. #TrumpTrain #WWG1WGA"
	"Conservative. Trump supporter. Q army"
Non-QAnon	*"a fair-minded thinker that believes in one constitution and equal justice for all."*
	"Believer, Patriot, Conservative, Father, Friend"
	"I am a patriot, an American. I support our president. I will never kneel before our flag..."

Table 1: Examples of user bios classified as QAnon and non-QAnon supporters. Class labels via keyword matching.

Feature Extraction

We extract a comprehensive set of features that should characterize our user base (see Tbl. 2), namely (1) user features, (2) linguistic features, (3) network features, and (4) content features as follows.

(1) User Features: To characterize users, we rely on information on a user's friendship network. This information was discriminatory in other research inferring verified status

[2]We find that misclassified accounts are related to other conspiracy theories, yet, with no other reference to QAnon. For example, users reference the keyword "plandemic" which might not provide enough evidence for human annotators to classify the profile as QAnon. To quantify how this would extrapolate to the whole sample, we checked for the number of profiles only mentioning the keyword "plandemic" with no other reference to QAnon in their bios: This only accounts for 99 of all accounts in our sample. This thus corroborates the reliability of our approach.

(Paul et al. 2019), the socioeconomic circumstances (Lampos et al. 2016), or discriminating bots and human users (Kudugunta and Ferrara 2018) on Twitter. Consistent with (Kudugunta and Ferrara 2018; Lampos et al. 2016; Paul et al. 2019), we also include information on a user's activity on Parler, namely, the number of posts and comments as well as the account age.

(2) Linguistic Features: We extract linguistic characteristics that capture the style (i. e., the *how*) with which content is written. Previously, such linguistic features were found to be a reliable predictor of group membership in other settings (Hu et al. 2016; Khalid and Srinivasan 2020). Analogous to prior literature (Chatzakou et al. 2017a; Khalid and Srinivasan 2020; Lampos et al. 2016; Paul et al. 2019), we thus control for handles, hashtags, external URLs, long words, and part-of-speech tags (POS).

We also compute the sentiment and the levels of toxicity, identity attacks, insult, profanity, and threat of content on Parler. Here, we follow prior research (Aliapoulios et al. 2022; Naumzik and Feuerriegel 2022; Papasavva et al. 2021; Pröllochs, Bär, and Feuerriegel 2021a,b; Pröllochs and Feuerriegel 2023) and extract sentiment using the NRC dictionary (Mohammad 2021) and use Google's Perspective API (Jigsaw and Google 2022) to extract the other features (i. e., toxicity, identity attacks, etc.). Overall, this is motivated by the fact that sentiment was found to be discriminatory in similar applications and that different levels of toxicity and threat for QAnon users compared to non-QAnon supporters were observed on other platforms (Aliapoulios et al. 2022; Papasavva et al. 2021; Paul et al. 2019)

We further calculate the average stance of a user towards QAnon. This accounts for the fact that sentiment and stance are not necessarily correlated (AlDayel and Magdy 2021). Inspired by prior research (Kawintiranon and Singh 2021), we pre-trained BERT on a large corpus of $\sim$5 M posts from Parler and added the most important stance tokens towards/against QAnon to the original BERT vocabulary. Overall, this should allow our model to capture Parler-specific language more accurately compared to using standard BERT for the downstream task of stance detection (Kawintiranon and Singh 2021). Subsequently, we fine-tuned the new language model on a sample of 1250 stance-labeled posts and computed the average stance of a user towards QAnon (details are in our GitHub).

Lastly, we include features extracted using the LIWC 2015 (Pennebaker et al. 2015) and the Empath library (Fast, Chen, and Bernstein 2016), which provide a multifaceted view on various lexical categories (e. g., negative/positive emotions, aggression) and were used in various applications studying online content (Kratzwald et al. 2018; Maarouf, Pröllochs, and Feuerriegel 2022; Ribeiro et al. 2018; Robertson et al. 2023).

(3) Network Features: We extract network features as a representation of how users are connected on Parler. Following prior research characterizing hateful users on Twitter (Chatzakou et al. 2017a; Ribeiro et al. 2018), we compute betweenness, eigenvector, and in-/out-degree centrality for each user in the repost network.

(4) Content Features: For each post and comment, we extract text embeddings as a representation of the textual content. In doing so, each post and comment is cleaned and tokenized; i. e., we replace contractions and remove hashtags, handles, emojis, and other alphanumeric characters. Next, we extract text embeddings from the standard SBERT model *all-MiniLM-L6-v2* with mean pooling (Reimers and Gurevych 2019). In particular, we map each post and comment on a 383-dimensional vector. SBERT is based on Google's BERT model (Devlin et al. 2019) and is specifically trained to create meaningful embeddings for short text and chosen because it is computationally more efficient and showed significantly better results on common benchmarks (Reimers and Gurevych 2019). Subsequently, we follow prior literature (Yu, Wan, and Zhou 2016) and average all sentence embeddings corresponding to a specific user. This provides a representation of the textual content posted by each user.

Machine Learning Approach

We use machine learning in order to discriminate QAnon from non-QAnon supporters based on the above features. For this, we use extreme gradient boosting (XGBoost) (Chen and Guestrin 2016). The outcome variable y is set to $y = 1$ if a user is classified as a QAnon supporter, and $y = 0$ otherwise. The choice of XGBoost is consistent with research performing user profiling in other settings (e. g., to predict verified status on Twitter (Paul et al. 2019)). In our analysis, we fit a classifier using the combination of all feature groups to evaluate the overall performance. To study the predictive power of the different feature groups individually, we further fit separate classifiers to the respective set of user features, linguistic features, network features, and content features. Finally, we evaluate the performance based on the area under the receiver operating characteristic curve (ROC AUC).

Our original sample is heavily skewed towards non-QAnon supporters (see Tbl. 3). Therefore, we randomly selected a balanced subsample of $n = 62,084$ users. Before training, we split our data into a training set (80 %) and a hold-out set (20 %) for the model evaluation. XGBoost is tuned using 10-fold cross-validation in combination with a grid search (details are in our GitHub).

Results

Size of QAnon Community on Parler (RQ1)

To answer **RQ1**, we classify users on Parler into self-reported QAnon supporters vs. others by matching user bios against an extensive list of keywords that are characteristic of QAnon. This allows us to quantify the size of the QAnon community on Parler. Reassuringly, we remind that we refer to all other users as "non-QAnon supporters." The results are shown in Tbl. 3. Out of all English-speaking users in our sample, we find that a large number of Parler users (34,913) are self-reported QAnon supporters. This amounts to 5.5 % of all users, thus providing evidence that Parler hosts a comparatively large QAnon community.

In addition, we provide summary statistics on the number of posts and comments associated with QAnon supporters

Feature Group	Dimension	Description
User features	*Followers*	Number of followers per user
	Followees	Number of followees per user
	Posts	Number of posts per user
	Comments	Number of comments per user
	Impressions (post)	Average number of impressions per post
	Upvotes (post)	Average number of upvotes per post
	Upvotes (comment)	Average number of upvotes per comment
	Downvotes (comment)	Average number of downvotes per comment
	Account age	Time passed since a user joined Parler
Linguistic Features	*Handles*	Frequency of handles per post/comment
	Hashtags	Frequency of hashtags per post/comment
	URLs	Frequency of URLs per post/comment
	Long words	Frequency of long words ($\geq$ 6 letters) per post/comment
	POS	Frequency of tokens per post/comment identified as part of speech after POS tagging[1]
	Stance	Average stance of user towards QAnon
	Sentiment	Sentiment, where sentiment scores are weighted over all posts and comments by length.
	Toxicity	Average toxicity by user
	Severe toxicity	Average severe toxicity by user
	Identity attack	Average level of identity attacks by user
	Insult	Average level of insults by user
	Profanity	Average level of profanities by user
	Threat	Average level of threat by user
	LIWC features	Average LIWC scores by user
	Empath features	Average Empath scores by user
Network Features	*Betweenness*	Betweenness centrality of user
	Eigen	Eigenvector centrality of user
	In-degree	In-degree centrality of user
	Out-degree	Out-degree centrality of user
Content Features	*Embeddings*	Text embeddings by user

[1] Part-of-speech tags comprise nouns, pronouns, adjectives, verbs, adpositions, and determiners, thus counting words in "natural" language.

Table 2: List of features extracted per user.

on Parler (see Tbl. 3). We find that QAnon supporters on Parler shared approximately 21.55 M posts and 3.65 M comments. This accounts for 14 %, and 9 %, respectively, of the overall posts and comments shared by users in our sample. Hence, QAnon supporters share proportionately more posts and comments than non-QAnon supporters on Parler. In the following, we will further study how user characteristics of QAnon and non-QAnon supporters on Parler differ.

Class Label	#Users [%]	#Posts [%]	#Comments [%]
QAnon	34,913 [5.5 %]	21,532,766 [14 %]	3,619,857 [9 %]
Non-QAnon	603,952 [94.5 %]	136,316,477 [86 %]	38,701,991 [91 %]
Overall	638,865 [100 %]	157,849,243 [100 %]	42,321,848 [100 %]

Table 3: Descriptives summarizing the QAnon community on Parler.

Comparison of User Characteristics between QAnon vs. non-QAnon Supporters (RQ2)

To answer **RQ2**, we now compare QAnon vs. non-QAnon supporters on Parler across different user characteristics. For this purpose, we analyze descriptive statistics with respect to the extracted user features, linguistic features, network features, and content features.

How active are QAnon supporters on Parler?

QAnon supporters are a bigger threat if they are especially active. Abusive users tend to be more active on social media (Ribeiro et al. 2018), which might also apply to QAnon users. Hence, we compare the number of posts and comments shared by QAnon vs. non-QAnon supporters on Parler. Specifically, we compute the complementary cumulative distribution function (CCDF) for both variables and test for statistically significant differences in the distributions using a Kolmogorov-Smirnov (KS) test (Smirnov 1939). We find that QAnon supporters share significantly more posts and comments compared to non-QAnon supporters ($p < 0.01$) (see Fig. 1a,b). In particular, the average number of posts (comments) is 621.67 (104.51) for QAnon supporters vs. 227.04 (64.46) for non-supporters on Parler. Hence, QAnon users have a large impact on the content discussed on Parler.

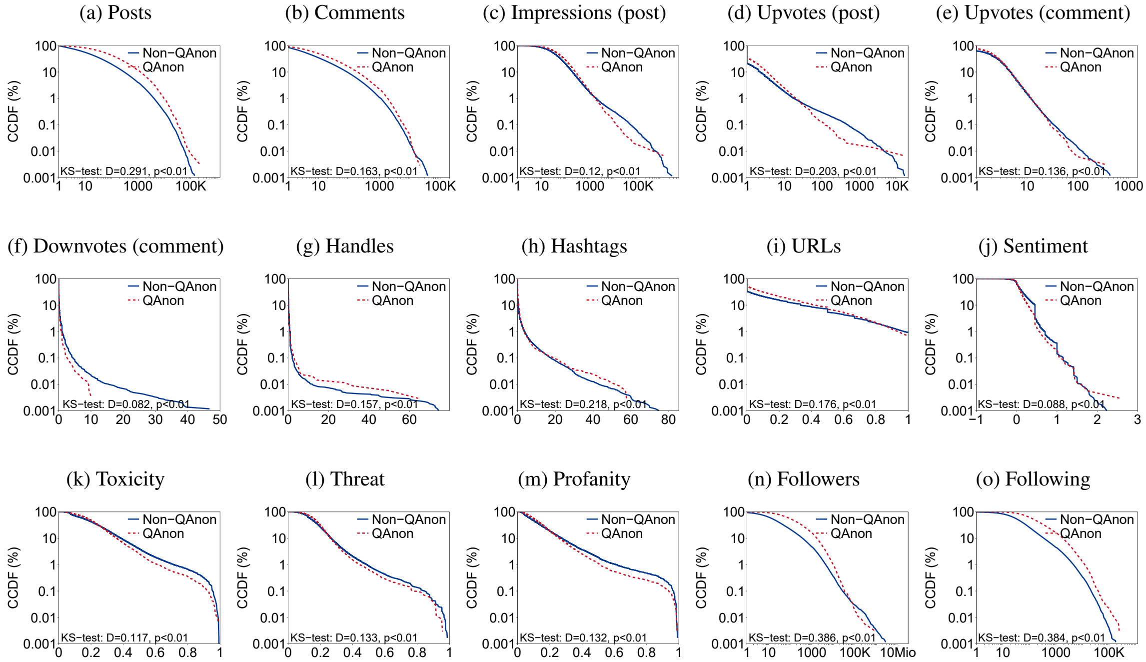

Figure 1: Complementary cumulative distribution functions (CCDFs) for selected features.

When did QAnon supporters join Parler?

The timing of when QAnon supporters migrated to Parler is important, as it informs which user base – Qanon supporters or non-supporters – drives the growth of the platform. Hence, we compare the account age of QAnon and non-QAnon supporters. Overall, the majority of users joined only in a later stage of the platform (≈ from June 2020 onward), which coincides, for example, with the 2020 U.S. presidential election and Twitter's increased efforts to manage content (Aliapoulios et al. 2021). However, on average, QAnon supporters joined Parler earlier than non-QAnon supporters. The average account age for QAnon supporters is 266.25 days (interquartile range [IQR]: 193.85 to 248.84), whereas, for non-QAnon supporters, it is 203.60 days (interquartile range [IQR]: 84.14 to 221.68) and thus around 2 months earlier. Overall, this suggests that QAnon supporters were early adopters of so-called "free speech" platforms such as Parler (e. g., to openly discuss the conspiracy).

How popular is content from QAnon supporters on Parler?

We now examine the virality of QAnon-related content. In the past, QAnon-related content has gone viral on mainstream social media (Sternisko, Cichocka, and van Bavel 2020). As such, we expect content by QAnon supporters to be more popular on Parler compared to content shared by non-supporters. To check this, we compare the popularity of posts and comments shared by QAnon vs. non-QAnon supporters on Parler. In doing so, we compare the number of upvotes per post, upvotes per comment, downvotes per comment, and impressions per user. We observe mixed patterns regarding the popularity of the posted content on Parler. Posts from both groups are, on average, almost equally likely to be upvoted (QAnon: 3.00 vs. non-QAnon: 3.62) (see Fig. 1d). Similarly, comments by QAnon supporters receive on average a comparable number of upvotes (QAnon: 2.53 vs. non-QAnon: 2.27) (see Fig. 1e). Nevertheless, a KS-test for each of the variables suggests statistically significant distributional differences for all variables ($p < 0.01$). In contrast, posts by non-QAnon supporters on average receive substantially less impressions (see Fig. 1c). In particular, non-QAnon supporters on average receive 406.21 impressions, while, for QAnon supporters, the average number of impressions is 249.09. Even though there is more QAnon-related content on Parler, the posts are less viral and far-reaching (as opposed to non-QAnon-related content).

How do QAnon share content on Parler?

QAnon supporters have developed specific hashtags (Sharma, Ferrara, and Liu 2022) and collectively investigate the cabal and decipher Q drops (Aliapoulios et al. 2022). This might relate to an extensive use of hashtags, handles, and URLs. Thus, we now compare *"how"* content by QAnon and non-QAnon supporters differs on Parler. We find that QAnon supporters use more handles than non-QAnon supporters (see Fig. 1g). Evidently, QAnon supporters also use more hashtags and share URLs more frequently than non-QAnon supporters (see Fig. 1h and Fig. 1i, respectively). The distributional difference is statistically significant for all variables ($p < 0.01$). Overall, this suggests that QAnon supporters make strategic use of handles, hashtags, and URLs. A possible explanation is that this might help to actively discuss the conspiracy with other supporters, reach certain audiences, and share external content to sources of conspiratorial materials.

How does sentiment, toxicity, threat, and profanity vary between QAnon and non-QAnon supporters?

On the one hand, the QAnon narrative (and many theories around cabals, sex trafficking, etc.) lets one expect high levels of toxicity, threat, and profanity. On the other hand, observations from other alt-right social media platforms such as Gab, 4chan, and 8kun suggest lower levels of toxicity, threat, and profanity for posts authored by QAnon supporters compared to other content on those platforms (Papasavva et al. 2021; Aliapoulios et al. 2022). Hence, we study differences in sentiment, toxicity, threat, and profanity. For all variables, we find distributional differences at statistically significant levels (see Fig. 1j–1m). QAnon supporters appear to use more positive sentiment compared to non-QAnon supporters. Furthermore, we find that posts and comments by QAnon supporters express lower levels of toxicity, threat, and profanity. Overall, this resembles the general narrative of QAnon pointing to a "brighter" future after the cabal is defeated (Zuckerman 2019).

How are (non-)QAnon supporters connected on Parler?

We compare the network structure of QAnon vs. non-QAnon supporters on Parler by analyzing (i) the friendship network and (ii) the repost network:

(i) Friendship Network: Abusive users on social media tend to maintain larger friendship networks (Ribeiro et al. 2018). QAnon users might also build a large friendship network to reach other supporters or convince non-supporters. Hence, we compare the number of followers and followees of QAnon and non-QAnon supporters. QAnon supporters generally have more followers than non-QAnon supporters (see Fig. 1n). The corresponding means amount to 799.20 (for QAnon supporters) vs. 416.25 (for non-QAnon supporters). In a similar vein, QAnon supporters are following more users than non-QAnon supporters (see Fig. 1o). On average, QAnon supporters follow 856.54 users vs 272.44 users for non-QAnon supporters. For both variables, the distributional differences are statistically significant as suggested by a KS-test ($p < 0.01$). This implies that QAnon supporters connect to a large number of users on Parler and thus attain an influential role in the network.

QAnon supporters, on average, joined Parler earlier compared to non-supporters. The additional time on the platform might have allowed QAnon supporters to grow a larger friendship network. Thus, we check how the size of the friendship network depends on the account age and compute the median number of followers and followees of QAnon supporters and non-QAnon supporters for 30-day intervals. The results are shown in Fig. 2. We find that the median number of followers is higher for QAnon supporters compared to non-QAnon supporters for each interval (see Fig. 2a). Similarly, QAnon supporters also follow more accounts compared to non-QAnon supporters regardless of their account age (see Fig. 2b). Overall, this shows how users grow on Paler and suggests that QAnon supporters quickly gain followers and followees on the platform.

(a) Followers (b) Followees

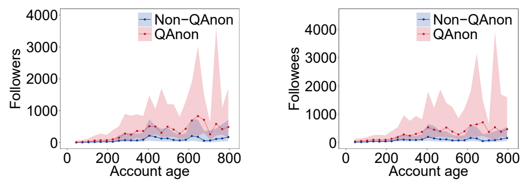

Figure 2: Median number of (a) followers and (b) followees of QAnon and non-QAnon supporters by account age. Shading represents the interquartile range.

(ii) Repost Network: Content by QAnon supporters tends to receive substantially fewer impressions compared to non-supporters (see above). This implies that the repost network of QAnon and non-QAnon supporters on Parler differs. Fig. 3 illustrates the repost network for highly connected users on Parler between November 2020 and January 2021. We find that QAnon supporters (colored in red) take a central position in the Parler network. Furthermore, they are closely connected to other QAnon accounts but also maintain links to spreaders of fake news (e. g., thegatewaypundit.com), far-right activists (e. g., Tommy Robinson), and conservative thought leaders (e. g., Sean Hannity).

We also compare the in- and out-degree centrality of users in both groups. Here, in-degree centrality measures the number of times a user's post is reposted/commented by another user and vice versa. We find that QAnon supporters are more central compared to non-QAnon supporters. In particular, posts by QAnon supporters are reposted/commented by 29.78 other users on average while posts by non-QAnon supporters are only reposted/commented on by 16.66 other users. Similarly, QAnon supporters repost/comment on average 24.37 posts by other users compared to 17.01 for non-QAnon supporters. Overall, this underlines the strong impact of QAnon supporters on Parler network.

What content do (non-)QAnon supporters share on Parler?

We expect that QAnon supporters discuss QAnon-related topics in addition to other topics on Parler. To check this, we now compare the content (i. e., *"what"* users write) shared by QAnon and non-QAnon supporters on Parler. In particular, we compare the most frequently used words by QAnon and non-QAnon supporters, which should point to how topics of interest vary across both groups. Fig. 4 reports the 10 most frequent words appearing in content composed by QAnon vs. non-QAnon supporters.

There are several similarities between QAnon and non-QAnon supporters. We find that both groups mention former U.S. President Donald Trump and "god". This may be expected: while many non-QAnon supporters do not self-disclose interest in the QAnon conspiracy theory, many of them are still conservatives (thus engaging in frequent

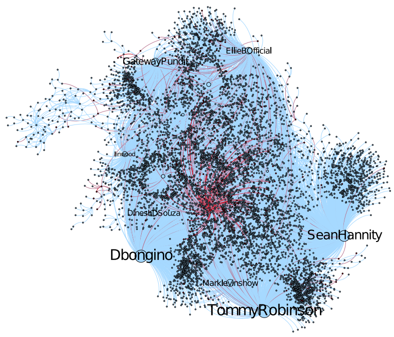

Figure 3: Social network plot of Parler showing user interactions between November 2020 and January 2021. QAnon supporters are colored in red. The size of the nodes varies by the overall number of interactions of a user (i. e., reposts and comments). For readability, only users who interacted at least 100 times are shown.

discussions around politics or religion) (Aliapoulios et al. 2021). Along these lines, we find that QAnon supporters frequently use words such as "America" and "patriots" which are often associated with a strong political polarization towards the right-wing.

However, there are also differences. We find that QAnon supporters frequently share terms such as "WWG1WGA" (as a short form for "Where we go one we go all") that are inherently QAnon-specific. Evidently, QAnon supporters frequently engage in discussing conspiratorial content.

Overall, the analysis shows that QAnon supporters discuss many conservative topics similar to those of regular users but, beyond that, also refer to words that are unique to the QAnon conspiracy theory.

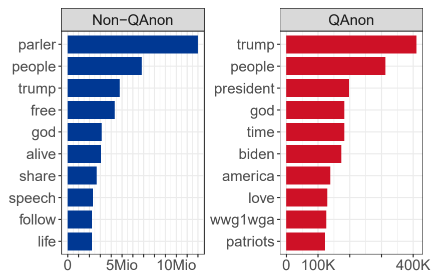

Figure 4: Top 10 most common words in Parler content (i. e., posts and comments) as a breakdown by QAnon and non-QAnon supporters.

Discriminating QAnon and non-QAnon Supporters Using Machine Learning (RQ3)

To answer **RQ3**, we examine which of the different feature groups allows to discriminate QAnon and non-QAnon supporters using machine learning. We thus fit four XGBoost classifiers, each using one of the four different feature groups. In addition, we examine the predictive power of a classifier trained on a combination of all four feature groups.

Tbl. 4 reports the prediction performance of the different classifiers. Mann–Whitney U-tests (Mason and Graham 2002) are used to confirm that the performance is above that of a random guess. We find that all feature sets have predictive power based on which QAnon vs. non-QAnon supporters can be discerned. For each feature group, the respective ROC AUC is above 0.50. Further, the improvement is statistically significant ($p < 0.01$). Consistent with this, we find that the performance of a classifier trained based on a combination of all features achieves a ROC AUC of 0.76. Again, this is statistically significant ($p < 0.01$).

Beyond that, we observe several notable patterns. First, we find that user features have the highest discriminatory power wrt. distinguishing (non-)QAnon supporters (ROC AUC = 0.74). This is followed by content features and linguistic features, where the performance of the classifier amounts to a ROC AUC of 0.69 and 0.67, respectively. In contrast, network features have the lowest discriminatory power (ROC AUC = 0.63).

We now check the feature importance for our model using all features. We find that user features (e. g., followers, account age) have overall large feature importance scores. In addition, several linguistic and content features (e. g., stance, SBERT 336) are also important. This corroborates our previous findings showing that QAnon supporters differ significantly from non-QAnon supporters along these dimensions (e. g., build larger friendship networks, use more hashtags).

Input	**ROC AUC**	**Sensitivity**	**Specificity**	**F1**
User features	0.74^{***}	0.77	0.71	0.75
Linguistic features	0.67^{***}	0.62	0.72	0.66
Network features	0.63^{***}	0.62	0.65	0.63
Content features	0.69^{***}	0.67	0.70	0.68
All features	0.76^{***}	0.77	0.75	0.77

p-values are obtained using the Mann–Whitney U-test (Mason and Graham 2002): $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Performance of classifying users into QAnon (=1) vs. non-QAnon (=0) supporters based on different features.

Robustness Checks: We check the robustness of our machine learning approach by conducting a series of additional analyses (details are in our GitHub): First, we check the predictive performance of each feature group with alternative classifiers. Overall, we find: XGBoost performs best. Various other classifiers reach a similar performance, and the ordering of the different feature groups remains the same. Second, we test the predictive performance with all possible combinations (i. e., powerset) of our four feature groups. In line with our main results, we find that combinations including user features exhibit higher ROC AUC scores and that no combination can statistically significantly achieve higher predictive performance compared to a model based on the complete set of features. Third, we check if features with low predictive performance influence the overall classification performance and perform feature selection using a LASSO. Subsequently, we train a model on all features with non-zero coefficients. Our results indicate that the exclusion of features with low predictive power does not increase the classification performance (ROC AUC=0.76). Nevertheless, this may yield a reduced model for applications in practice.

Discussion

Relevance: The QAnon conspiracy theory has been deemed as a significant threat to public security by the U.S. Federal Bureau of Investigation (Amarasingam and Argentino 2020). Thus, a better understanding of QAnon supporters is necessary to identify potentially dangerous communities on the platform. To the best of our knowledge, this paper provides the first study profiling QAnon supporters on Parler.

Summary of Findings: Our findings contribute to the existing research on Parler by quantifying the number of QAnon supporters on the platform. Previously, it was frequently observed that the Parler social media platform hosts QAnon (e. g., Aliapoulios et al. 2021; Baines, Ittefaq, and Abwao 2021); however, the actual size of the community has remained unclear. Here, our findings show that there is indeed a large community of self-reported QAnon supporters on Parler. Specifically, we find that around 5.5 % of the users in our sample are self-reported QAnon supporters.

Our results further show that QAnon supporters differ from non-QAnon supporters across multiple dimensions. This is in line with social identity theory which predicts that the behavior of people differs according to their group membership (Tajfel and Turner 1986). For example, we find that QAnon supporters have a large impact on the platform as they are more active and maintain larger friendship and repost networks compared to non-QAnon supporters. As such, QAnon supporters appear to behave similarly to other abusive users on social media (Ribeiro et al. 2018). Thereby, we establish a better understanding of the behavior of QAnon supporters in online communities.

A potential reason for the different behavior of QAnon supporters might lie in the participatory nature of QAnon (Zuckerman 2019) and the motivations behind conspiracy theory beliefs (Sternisko, Cichocka, and van Bavel 2020). Psychological research shows that conspiracy theorists want to make sense of their environment (Sternisko, Cichocka, and van Bavel 2020). In the case of QAnon, supporters collectively investigate the cabal and decipher Q drops. Such efforts require a high level of outreach and discussion to be successful (Bär et al. 2023) and could thus lead to larger networks and activity on Parler. Along similar lines, the higher number of URLs shared by QAnon supporters might indicate an increased effort to explain their views. Furthermore, conspiracy theorists are often driven by social identity motives and adherents want to feel good about themselves / their group (Sternisko, Cichocka, and van Bavel 2020). The frequent use of specific hashtags of QAnon supporters might indicate the development of a unique social identity, whereas the relatively more positive sentiment and the lower levels of toxicity, threat, and profanity might be related to a positive self-image of QAnon supporters. Of note, the differences for all these features are statistically significant, showing that QAnon supporters differ from the otherwise rather homogeneously conservative user base on Parler.

We further find that machine learning together with a representative set of features (chosen analogous to prior research (Paul et al. 2019; Rao et al. 2010)) can discriminate QAnon vs. non-QAnon supporters with a ROC AUC of up to 0.76. The performance is lower compared to studies profiling verified users (Paul et al. 2019) or detecting bots on Twitter (Kudugunta and Ferrara 2018). This implies that there is still some unexplained variance (beyond the information from user characteristics that is typically used for user profiling in social media). However, the lower ROC AUC is also an indication of similarities among QAnon supporters and other users on Parler, who, for example, may often post somewhat similar content (e. g., posts with similar conservative viewpoints). Out of the different feature groups, we find that user features are especially discriminatory to distinguish QAnon vs. non-QAnon supporters, suggesting that both groups are characterized by different behavior (and not necessarily *"how"* or *"what"* users write). This is in line with other research profiling users on mainstream online communities such as Twitter, where user features consistently have large predictive power (Kudugunta and Ferrara 2018; Paul et al. 2019).

Implications: The QAnon community is expanding globally (Hoseini et al. 2021) and poses a significant threat to public security (Amarasingam and Argentino 2020). As such, the growing number of violent acts by QAnon supporters (National Consortium for the Study of Terrorism and Responses to Terrorism 2021) that peaked in an attack on democracy with the storming of the U.S. Capitol on January 6, 2021 (Hitkul et al. 2021) call for action by research, platform owners, and policymakers. Here, our work provides first insights into the behavioral attributes of QAnon supporters. We demonstrate that machine learning in combination with a comprehensive set of features can help to identify QAnon supporters – even on a social media platform that largely resembles a right-wing echo chamber. From a practical perspective, our features rely on data available for most social media platforms. As such, our machine learning framework is directly applicable to other platforms and may help surveillance and early detection of upcoming threats through a group of conspiracy theorists that have repeatedly been associated with violent incidents.

Moreover, it is concerning that QAnon supporters seem to use Parler to cultivate their social identity (e. g., by using different linguistic styles) while also growing larger friendship networks faster than other user groups. The latter may be expected as QAnon supporters chose the Parler platform for a particular reason (e. g., actively discussing politics based on a certain ideology rather than pure news consumption or curiosity). However, it also renders it likely that false information spreads particularly fast and viral among QAnon supporters due to their larger reach and central role in the network. Also, we see that community mechanisms to control information (i. e., upvotes and downvotes) may not be functioning as desired. Upvotes are distributed fairly similarly for both QAnon supporters and other users on Parler. However, more importantly, comments from QAnon supporters receive *considerably* fewer downvotes. This may exacerbate and even reinforce the spread of false information due to the absence of content moderation (as well as having a segregated "echo chamber" platform with users originating primarily from a single, right-leaning ideology).

Limitations and Future Research: As with other research, ours is not free of limitations that offer opportunities for future research: (1) We identified QAnon supporters based on an extensive series of keywords consistent with earlier research (Sharma, Ferrara, and Liu 2022). Previously, the reliability of such an approach was unclear (e. g., users might express their opposition to QAnon and still get classified as supporters). However, our validation study confirms that the approach is highly accurate. (2) We infer QAnon supporters based on their user bios, while future research could identify QAnon supporters based on posts and comments, though this might also include users arguing against QAnon and may thus provide less reliable labels. (3) Our analysis is based on Parler, which has attracted a large community of QAnon supporters and which makes it particularly relevant for research to understand differences between QAnon vs. non-QAnon supporters. However, QAnon supporters might behave differently on mainstream social media. As a result, taking a cross-platform perspective presents a promising avenue for future research. (4) Our data provides a static snapshot of Parler. Hence, future research should study how the characteristics of QAnon supporters change over time. (5) We compare QAnon vs. non-QAnon supporters. Yet, there may also be further heterogeneity among QAnon supporters, which could be analyzed by future research.

Conclusion

The social media platform Parler has emerged into a prominent fringe community where a significant proportion of the user base are self-reported supporters of QAnon. Yet, little is known about QAnon supporters on Parler. To fill this void, we analyze a large-scale public snapshot of Parler, based on which we profile QAnon supporters. Self-reported QAnon supporters make up a significant portion (5.5 %) of the user base on Parler. These users are significantly different from non-QAnon supporters on Parler. Following social identity theory, the self-reported appraisal of QAnon manifests in different online behavior such as larger friendship networks and greater activity. These differences allow machine learning to discriminate QAnon vs. non-QAnon supporters. Here, user features such as the size of the friendship network or activity are more discriminatory compared to *how* or *what* users write. Our machine learning framework may thus allow for real-time surveillance and early warnings.

Ethics Statement

This research did not involve interventions with human subjects, and, thus, no approval from the Institutional Review Board was required by the author institutions. All analyses are based on publicly available data and we do not make any attempt to track users across different platforms. We neither de-anonymize nor de-identify their accounts. Furthermore, all analyses conform with national laws. To respect privacy, we explicitly do not publish usernames in our paper (except for celebrity profiles) and only report aggregate results.

References

AlDayel, A.; and Magdy, W. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4): 102597.

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; de Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the effect of deplatforming on social networks. In *WebSci*.

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; de Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021. A large open dataset from the Parler social network. In *WWW*.

Aliapoulios, M.; Papasavva, A.; Ballard, C.; de Cristofaro, E.; Stringhini, G.; Zannettou, S.; and Blackburn, J. 2022. The gospel according to Q: Understanding the QAnon conspiracy from the perspective of canonical information. In *ICWSM*.

Amarasingam, A.; and Argentino, M.-A. 2020. The QAnon conspiracy theory: A security threat in the making? *CTC Sentinel*, 13(7): 37–44.

Baines, A.; Ittefaq, M.; and Abwao, M. 2021. #Scamdemic, #Plandemic, or #Scaredemic: What Parler social media platform tells us about COVID-19 vaccine. *Vaccines*, 9(5): 421.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Bär, D.; Calderon, F.; Lawlor, M.; Licklederer, S.; Totzauer, M.; and Feuerriegel, S. 2023. Analyzing social media activities at Bellingcat. In *WebSci*.

Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. New threats to society from free-speech social media platforms. *Communications of the ACM*.

Bessi, A.; Coletto, M.; Davidescu, G. A.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, 10(2): e0118093.

Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating gender on Twitter. In *EMNLP*.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; de Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017a. Mean birds: Detecting aggression and bullying on Twitter. In *WebSci*.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; de Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017b. Measuring #GamerGate: A tale of hate, sexism, and bullying. In *WWW*.

Chen, T.; and Guestrin, C. 2016. XGBoost. In *KDD*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Ding, Y.; and Jiang, J. 2014. Extracting interest tags from Twitter user biographies. In *Asia Information Retrieval Societies Conference*.

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. 2018. Peer to peer hate: Hate speech instigators and their targets. In *ICWSM*.

Engel, K.; Hua, Y.; Zeng, T.; and Naaman, M. 2022. Characterizing Reddit participation of users who engage in the QAnon conspiracy theories. In *CSCW*.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *CHI*.

Geissler, D.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2022. Russian propaganda on social media during the 2022 invasion of Ukraine. *arXiv*, (2211.04154).

Gorrell, G.; Greenwood, M.; Roberts, I.; Maynard, D.; and Bontcheva, K. 2018. Twits, twats and twaddle: Trends in online abuse towards UK politicians. In *ICWSM*.

Hanley, H. W. A.; Kumar, D.; and Durumeric, Z. 2022. No calm in the storm: Investigating QAnon website relationships. In *ICWSM*.

Hine, G.; Onaolapo, J.; de Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*.

Hitkul; Prabhu, A.; Guhathakurta, D.; jain, J.; Subramanian, M.; Reddy, M.; Sehgal, S.; Karandikar, T.; Gulati, A.; Arora, U.; Shah, R. R.; and Kumaraguru, P. 2021. Capitol (pat)riots: A comparative study of Twitter and Parler. *arXiv*, (2101.06914).

Hoseini, M.; Melo, P.; Benevenuto, F.; Feldmann, A.; and Zannettou, S. 2021. On the globalization of the QAnon conspiracy theory through Telegram. *arXiv*, (2105.13020).

Hu, T.; Xiao, H.; Nguyen, T.-v. T.; and Luo, J. 2016. What the language you tweet says about your occupation. In *ICWSM*.

Hua, Y.; Naaman, M.; and Ristenpart, T. 2020. Characterizing Twitter users who engage in adversarial interactions against political candidates. In *CHI*.

Jakubik, J.; Vössing, M.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Online emotions during the storming of the U.S. Capitol: Evidence from the social media network Parler. In *ICWSM*.

Jigsaw; and Google. 2022. Perspective API. www.perspectiveapi.com. Accessed: 2022.

Kawintiranon, K.; and Singh, L. 2021. Knowledge enhanced masked language model for stance detection. In *NAACL*.

Khalid, O.; and Srinivasan, P. 2020. Style matters! Investigating linguistic style in online communities. In *ICWSM*.

Kratzwald, B.; Ilić, S.; Kraus, M.; Feuerriegel, S.; and Prendinger, H. 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115: 24–35.

Kudugunta, S.; and Ferrara, E. 2018. Deep neural networks for bot detection. *Information Sciences*, 467: 312–322.

Lampos, V.; Aletras, N.; Geyti, J. K.; Zou, B.; and Cox, I. J. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *European Conference on Information Retrieval*.

Li, J.; Longinos, G.; Wilson, S.; and Magdy, W. 2020. Emoji and self-identity in Twitter bios. In *Workshop on Natural Language Processing and Computational Social Science*.

Maarouf, A.; Pröllochs, N.; and Feuerriegel, S. 2022. The virality of hate speech on social media. *arXiv*, (2210.13770).

Maity, S. K.; Chakraborty, A.; Goyal, P.; and Mukherjee, A. 2018. Opinion conflicts: An effective route to detect incivility in Twitter. In *CSCW*.

Mason, S. J.; and Graham, N. E. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584): 2145–2166.

Matero, M.; Idnani, A.; Son, Y.; Giorgi, S.; Vu, H.; Zamani, M.; Limbachiya, P.; Guntuku, S. C.; and Schwartz, H. A. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Workshop on Computational Linguistics and Clinical Psychology*.

Miller, D. T. 2021. Characterizing QAnon: Analysis of YouTube comments presents new conclusions about a popular conservative conspiracy. *First Monday*, 26(2).

Mohammad, S. M. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion Measurement (Second Edition)*, 323–379.

Munn, L. 2021. More than a mob: Parler as preparatory media for the U.S. Capitol storming. *First Monday*, 26(3).

National Consortium for the Study of Terrorism and Responses to Terrorism. 2021. QAnon Crime Maps. https://www.start.umd.edu/qanon-crime-maps. Accessed: note = Accessed: 2021.

Naumzik, C.; and Feuerriegel, S. 2022. Detecting false rumors from retweet dynamics on social media. In *WWW*.

Otala, J. M.; Kurtic, G.; Grasso, I.; Liu, Y.; Matthews, J.; and Madraki, G. 2021. Political polarization and platform migration. In *WWW*.

Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and de Cristofaro, E. 2021. "Is it a Qoincidence?": An exploratory study of QAnon on Voat. In *WWW*.

Papasavva, A.; Zannettou, S.; de Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *ICWSM*.

Pathak, A.; Madani, N.; and Joseph, K. 2021. A method to analyze multiple social identities in Twitter bios. In *CSCW*.

Paudel, P.; Blackburn, J.; de Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2022. A longitudinal study of the Gettr social network. In *ICWSM*.

Paul, I.; Khattar, A.; Chopra, S.; Kumaraguru, P.; and Gupta, M. 2019. What sets verified users apart? Insights, analysis and prediction of verified users on Twitter. In *WebSci*.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. In *Texas ScholarWorks*.

Pieroni, E.; Jachim, P.; Jachim, N.; and Sharevski, F. 2021. Parlermonium: A data-driven UX design evaluation of the Parler platform. *arXiv*, (2106.00163).

Preoţiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; and Aletras, N. 2015. Studying user income through language, behaviour and affect in social media. *PLOS ONE*, 10(9): e0138717.

Priniski, J. H.; McClay, M.; and Holyoak, K. J. 2021. Rise of QAnon: A Mental Model of Good and Evil Stews in an Echochamber. In *CogSci*.

Pröllochs, N.; Bär, D.; and Feuerriegel, S. 2021a. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, 11(1): 22721.

Pröllochs, N.; Bär, D.; and Feuerriegel, S. 2021b. Emotions in online rumor diffusion. *EPJ Data Science*, 10(1).

Pröllochs, N.; and Feuerriegel, S. 2023. Mechanisms of true and false rumor sharing in social media: Collective intelligence or herd behavior? In *CSCW*.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in Twitter. In *International Workshop on Search and Mining User-generated Contents*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*.

Ribeiro, M.; Calais, P.; Santos, Y.; Almeida, V.; and Meira Jr., W. 2018. Characterizing and detecting hateful users on Twitter. In *ICWSM*.

Robertson, A. 2021. Parler is back online after a month of downtime: An Amazon suspension shut it down in January. *The Verge*.

Robertson, C.; Pröllochs, N.; Kaoru, S.; Parnamets, P.; van Bavel, J. J.; and Feuerriegel, S. 2023. Negativity drives online news consumption. *Nature Human Behaviour*.

Rogers, N.; and Jones, J. J. 2021. Using Twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time? *Journal of Social Computing*, 2(1): 1–13.

Sharma, K.; Ferrara, E.; and Liu, Y. 2022. Characterizing online engagement with disinformation and conspiracies in the 2020 U.S. Presidential Election. In *ICWSM*.

Sipka, A.; Hannak, A.; and Urman, A. 2022. Comparing the language of QAnon-related content on Parler, Gab, and Twitter. In *WebSci*.

Smirnov, N. V. 1939. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2(2): 3–16.

Sternisko, A.; Cichocka, A.; and van Bavel, J. J. 2020. The dark side of social movements: social identity, nonconformity, and the lure of conspiracy theories. *Current Opinion in Psychology*, 35: 1–6.

Tajfel, H.; and Turner, J. C. 1986. The Social Identity Theory of Intergroup Behavior. In *Psychology of Intergroup Relations*, 7–24.

Trujillo, M.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2020. What is BitChute? Characterizing the “free speech” alternative to YouTube. In *ACM HT*.

Yu, Y.; Wan, X.; and Zhou, X. 2016. User embedding for scholarly microblog recommendation. In *ACL*.

Zannettou, S.; Bradlyn, B.; de Cristofaro, E.; Kwak, H.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2018. What is Gab? A bastion of free speech or an alt-right echo chamber? In *WWW*.

Zuckerman, E. 2019. QAnon and the emergence of the unreal. *Journal of Design and Science*, 6.

8 Emotions in online rumor diffusion

Title: Emotions in online rumor diffusion

Abstract: Emotions are regarded as a dominant driver of human behavior, and yet their role in online rumor diffusion is largely unexplored. In this study, we empirically study the extent to which emotions explain the diffusion of online rumors. We analyze a large-scale sample of 107,014 online rumors from Twitter, as well as their cascades. For each rumor, the embedded emotions were measured based on eight so-called basic emotions from Plutchik's wheel of emotions (i.e., anticipation–surprise, anger–fear, trust–disgust, joy–sadness). We then estimated using a generalized linear regression model how emotions are associated with the spread of online rumors in terms of (1) cascade size, (2) cascade lifetime, and (3) structural virality. Our results suggest that rumors conveying anticipation, anger, and trust generate more reshares, spread over longer time horizons, and become more viral. In contrast, a smaller size, lifetime, and virality is found for surprise, fear, and disgust. We further study how the presence of 24 dyadic emotional interactions (i.e., feelings composed of two emotions) is associated with diffusion dynamics. Here, we find that rumors cascades with high degrees of aggressiveness are larger in size, longer-lived, and more viral. Altogether, emotions embedded in online rumors are important determinants of the spreading dynamics.

Author contributions: Nicolas Pröllochs and Stefan Feuerriegel designed the study. Nicolas Pröllochs and Dominik Bär analyzed the data. Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel wrote and revised the manuscript.

Reference: Pröllochs, N., **Bär, D.**, & Feuerriegel, S. 2021. *Emotions in online rumor diffusion.* In: **EPJ Data Science** 10, 51, DOI: `https://doi.org/10.1140/epjds/s13688-021-00307-5`.

Supplementary materials: Supplementary materials for this article are in Supplementary Material A.

Pröllochs et al. *EPJ Data Science* (2021) 10:51
https://doi.org/10.1140/epjds/s13688-021-00307-5

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Emotions in online rumor diffusion

Nicolas Pröllochs[1*], Dominik Bär[2] and Stefan Feuerriegel[2,3]

*Correspondence:
nicolas.proellochs@wi.jlug.de
[1]University of Giessen, Licher Str. 62, 35394 Giessen, Germany
Full list of author information is available at the end of the article

Abstract

Emotions are regarded as a dominant driver of human behavior, and yet their role in online rumor diffusion is largely unexplored. In this study, we empirically study the extent to which emotions explain the diffusion of online rumors. We analyze a large-scale sample of 107,014 online rumors from Twitter, as well as their cascades. For each rumor, the embedded emotions were measured based on eight so-called basic emotions from Plutchik's wheel of emotions (i.e., anticipation–surprise, anger–fear, trust–disgust, joy–sadness). We then estimated using a generalized linear regression model how emotions are associated with the spread of online rumors in terms of (1) cascade size, (2) cascade lifetime, and (3) structural virality. Our results suggest that rumors conveying anticipation, anger, and trust generate more reshares, spread over longer time horizons, and become more viral. In contrast, a smaller size, lifetime, and virality is found for surprise, fear, and disgust. We further study how the presence of 24 dyadic emotional interactions (i.e., feelings composed of two emotions) is associated with diffusion dynamics. Here, we find that rumors cascades with high degrees of aggressiveness are larger in size, longer-lived, and more viral. Altogether, emotions embedded in online rumors are important determinants of the spreading dynamics.

Keywords: Online rumors; Information cascade; Online diffusion; Emotions; Regression analysis

1 Introduction

Social media platforms such as Facebook, Sina Weibo, and Twitter allow users to disseminate content through sharing (e.g., called retweeting in the case of Twitter). As a result, content can go viral and reach a large audience despite that fact that it originated from a single broadcast. To this end, understanding the diffusion of online content is relevant for a number of reasons. Marketers are interested in identifying what makes content go viral, so that marketing content can be designed accordingly [1–4]. Humanitarian organizations leverage the potential of online diffusion in social media to collect information for effective responses to natural disasters and to inform the wider public [5–7]. Public stakeholders are confronted with the diffusion of political content and, by understanding the underlying mechanics, can help prevent the spread of rumors [8–11].

Previous research has identified several drivers of online diffusion (see Additional file 1 for an overview). These drivers are primarily located in the different characteristics of senders. For instance, senders with a larger follower base (i.e., with more outgoing ties in the network) also reach, on average, a larger audience [12]. Other characteristics of senders are the number of followees (i.e., how many incoming ties a user has [13–15]) or their past engagement (i.e., the number of posts or reshares [11]). A different stream of research has examined online diffusion around specific topics (e.g., a specific election [9] or a specific disaster [5–7, 16–19]). In this work, we add by studying the role of emotions in the diffusion of online rumors.

Emotions have been established as an important determinant of human behavior in offline behavior [20–22]. Emotions typically arise as a response to environmental stimuli that are of relevance to the needs, goals, or concerns of users and, as a consequence, also guide user behavior in online settings [23]. Emotions influence what type of information users seek, what they process, how they remember it, and ultimately what judgments and decisions they derive from it. Emotions are themselves contagious and can spread among people, both offline (i.e., in person) [24] and online (i.e., via social media) [25–29].

Following the above, an important driver of online behavior are emotions embedded in online content. For instance, it was previously confirmed that emotions influence posting and liking activities [30], users' willingness-to-share [1], and actual sharing behavior [2, 31–33]. As such, embedded emotions explain, to a large extent, the propensity to share posts, as well as user response time. Here, emotional stimuli such as emotion-laden wording trigger cognitive processing [34], which in turn results in the behavioral response of information sharing [35–37]. In particular, emotions embedded in online content also explain the dynamics of online diffusion. For instance, emotions describe different properties of diffusion cascades, such as their size, branching, or lifetime [38–41]. Especially misinformation relies upon emotions in order to attract attention [11, 38, 42–46]. Given the importance of emotions in online behavior, we investigate how emotions are linked to the spread of online rumors.

Hypothesis *Emotions embedded in online rumors are associated with the size, lifetime, and structural virality of the cascade.*

In this study, we empirically analyze to what extent emotions explain the diffusion of online rumors. For this, we infer the emotions embedded in replies to online rumors through the use of affective computing (see Methods). For each rumor, the degree of emotion is rated along so-called basic emotions. Basic emotions refer to a subset of emotions that are universally recognized across cultures and through which other, more complex emotions can be derived. In this work, we adopt Plutchik's wheel of emotions [22], comprising 8 basic emotions (ANTICIPATION, SURPRISE, ANGER, FEAR, TRUST, DISGUST, JOY, SADNESS). Based on these, we infer 24 dyadic emotional interactions, each representing a more complex emotion composed of two basic emotions (e.g., AGGRESSIVENESS as a combination of ANGER and ANTICIPATION). These emotions are then linked to the spread of online rumors using regression analysis. Thereby, we estimate to what extent emotions embedded in online rumors explain: (1) cascade size, that is, how many reshares a rumor generates; (2) cascade lifetime, that is, how long a rumor is active; and (3) structural virality, that is, how effectively it spreads. The latter, structural virality, provides a quantitative metric [47] aggregating the depth-breadth variation in rumor diffusion.

One work [11] contains summary statistics reporting which emotions are present in online rumors but not how emotions affect *sharing*. Hence, any statistical claims measuring the emotion effect (= which emotions drive a faster and wider rumor spreading) are precluded. This presents the added value of our work. We measure how emotions are associated with the diffusion dynamics (e.g., TRUST as an emotion is present in only a small portion of rumors but it has a large influence on virality). Because of this, our work is different in several ways: (i) we focus not only on basic emotions but also dyadic emotions, (ii) we infer the emotion effect on diffusion dynamics, and, because of that, (iii) we use a regression analysis as opposed to summary statistics. Therefore, this work is—to the best of our knowledge—the first comprehensive study assessing the link between emotions and the spread of online rumors.

We analyze a large-scale, representative sample of Twitter rumors and their corresponding cascades [11]. Specifically, our data cover the complete time frame from the launch of Twitter in 2006 until (and including) 2017. Altogether, this results in 2189 rumors associated with 107,014 cascades. The sample comprises approx. 3.7 million reshares that originate from almost 3 million different users. Based on the cascades, various control variables are constructed. Specifically, in our regression analysis, we capture time- and rumor-effects through the use of random effects, based on which we control for the heterogeneity among rumors (see Materials and Methods).

2 Materials and methods

2.1 Dataset

A rumor is defined as a piece of content that is propagated between users but without confirmation of its veracity. This definition is rooted in social psychology literature [43, 48]. For this study, a large-scale dataset comprising of rumor cascades from Twitter [11] was analyzed. The resulting sample comprises ***all*** rumors from Twitter between its founding in the year 2006 until (and including) 2017. Ethics approval was obtained from ETH Zurich (2020-N-44). Overall, our sample includes 2189 rumors with a total of $N = 107{,}014$ cascades (i.e., some rumor contents were shared as part of multiple but different cascades). The rumors had approx. 3.7 million reshares originating from 3 million users (see [11] for details).

2.2 Characteristics of online rumor diffusion

The cascades were then processed as follows in order to generate additional variables. These variables refer to different characteristics of online rumor diffusion and later represent the dependent variables in the regression analysis. For simplicity, we introduce the following notation. We refer to the cascades via $j = 1, \ldots, N$. These belong to $i = 1, \ldots, 2189$ different rumors. Each cascade is a three-tuple $T_j = (r_j, t_{j0}, R_j)$, where r_j is the root post that corresponds to the original broadcast and where t_{j0} is its timestamp and R_j the set of reshares. A reshare k has a parent p_{jk} and a timestamp t_{jk}, i.e., $R_j = \{(p_{jk}, t_{jk})\}_k$.

(1) *Cascade size*: The cascade size counts how many reshares a cascade generated. Formally, it amounts to all reshares plus 1 (for the root), i.e., $|R_j| + 1$.
(2) *Cascade lifetime*: The cascade lifetime is the timespan during which a rumor cascade was active, thus the elapsed time between the root broadcast and the last reshare. It is calculated via $\max_k t_{jk} - t_{j0}$.
(3) *Structural virality*: Structural virality [47] provides an aggregated metric combining the depth and breadth of a cascade. A higher structural virality corresponds to a cascade that is both of great depth *and* where each reshare generated a large relative number of additional reshares (i.e., a high branching factor). As proposed in [47],

structural virality is based on the idea of the Wiener index, i.e.,

$$\nu(T_j) = \frac{1}{|R_j| \times (|R_j| + 1)} \sum_{j_1=0}^{|R_j|} \sum_{j_2=0}^{|R_j|} d_{j_1,j_2}, \tag{1}$$

where d_{j_1,j_2} is the shortest path between nodes j_1 and j_2 in the tree T_j. Intuitively, structural virality reflects the average distance between all reshares in the graph.

2.3 Model variables on heterogeneity between rumor cascades

Model variables x_j, concerning the heterogeneity among rumor cascades, were computed as in earlier research [11, 12, 31, 38]. These later act as controls. In our study, controls are (1) account age; (2) a binary dummy representing whether the account is officially labeled as "verified" (= 1 if yes, i.e., Twitter displays a blue badge next to it); (3) the number of followers (outgoing ties); (4) the number of followees (incoming ties); and (5) user engagement, that is, the average number of posts, reshares, and likes relative to the account age as in [11]. These variables reflect that the senders of rumors vary in their social influence.

Note that all of the above variables were computed at the level of cascades (which is later our unit of analysis). Additional sources of heterogeneity among rumors are captured via rumor-level random effects.

2.4 Computing emotions embedded in online rumors

For all cascades, we measured the emotions embedded in replies to rumor cascades. Here, we distinguish *basic emotions*, *bipolar emotion pairs*, and *dyadic emotional interactions* comprising primary, secondary, tertiary dyads. The computation of the emotions is detailed below (see [22] for further details).

Basic emotions: Basic emotions refer to a subset of emotions that are universally recognized across cultures and through which other, more complex emotions can be derived [20, 21]. In our study, Plutchik's wheel of emotions [22] is adopted as it is a common tool in affective computing [49]. It defines 8 basic emotions (see Fig. 1, petals): ANTICIPATION, SURPRISE, ANGER, FEAR, TRUST, DISGUST, JOY, and SADNESS.

Our computation follows a dictionary-based approach as in [11]. Dictionary-based approaches are widely used when large-scale analyses of emotions are performed with the objective of explanatory modeling and thus reliable interpretations [38, 41]. In our work, the NRC emotion lexicon was used [50], which classifies English words into the 8 basic emotions. For all cascades j, the content of the replies was tokenized and the frequency of dictionary terms per basic emotion was counted, resulting in an 8-dimensional emotion score e_j. Afterwards, the vector was normalized to sum to one across basic emotions (i.e., $e'_j = \frac{1}{\|e_j\|_1} e_j$). We omit rumor cascades that do not contain any emotional words from the NRC emotion lexicon (since, otherwise, the denominator is not defined). As a result, the 8 emotion dimensions in $e'_j \in [0,1]^8$ range from zero to one. Owing to this fact, replies to rumors can embed a combination of multiple emotions (e.g., 40% ANGER and 60% FEAR).

Bipolar emotion pairs: In Plutchik's wheel of emotions, the 8 basic emotions are organized according to 4 pairs of bipolar emotions (i.e., the opposite petals in Fig. 1). The 4 pairs of bipolar emotions are ANTICIPATION–SURPRISE, ANGER–FEAR, TRUST–DISGUST, JOY–SADNESS. In each case, one dimension of the pair is considered to be positive and the other

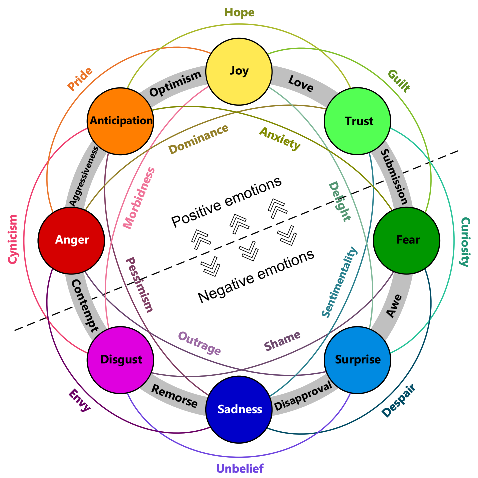

Figure 1 Plutchik's wheel of emotions [22]

negative. We calculate a 4-dimensional score ϕ_j^{pairs} that measures the difference between a specific positive emotion and its complement from the set of negative emotions. For example, ANGER–FEAR refers to the difference between ANGER and FEAR.

Dyadic emotional interactions: Plutchik's wheel of emotions further defines 24 dyadic emotional interactions, which are more complex emotions composed of two basic emotions (see Fig. 1, round lines). The dyadic emotional interactions comprise:

1 *Primary dyads* that are one petal apart from each other (e.g., AGGRESSIVENESS = ANGER + ANTICIPATION). The 8 primary dyadic emotional interactions are OPTIMISM, DISAPPROVAL, LOVE, REMORSE, SUBMISSION, CONTEMPT, AWE, and AGGRESSIVENESS.
2 *Secondary dyads* that are two petals apart from each other (e.g., HOPE = ANTICIPATION + TRUST). The 8 secondary dyadic emotional interactions are HOPE, UNBELIEF, GUILT, ENVY, CURIOSITY, CYNICISM, DESPAIR, and PRIDE.
3 *Tertiary dyads* that are three petals apart from each other (e.g., ANXIETY = ANTICIPATION + FEAR). The 8 tertiary dyadic emotional interactions are ANXIETY, OUTRAGE, DELIGHT, PESSIMISM, SENTIMENTALITY, MORBIDNESS, SHAME, and DOMINANCE.

Similar to the bipolar emotion pairs, the dyadic emotional interactions are arranged such that each has an opposite emotion. For example, LOVE is the opposite of REMORSE. Hence, for each pair, we again compute a score that is the difference between the opposing emotions. This yields $\phi_j^{primary}, \phi_j^{secondary}, \phi_j^{tertiary} \in [0,1]^4$.

2.5 Regression analysis

To analyze the role of emotions in online rumor diffusion, we apply a generalized regression model. Regression models are generally regarded as an explanatory approach with the ability to document statistical relationships and, in particular, estimate effect sizes [51]. Furthermore, regression models are widely used to estimate the marginal effect of content on diffusion characteristics [11, 31, 38, 41]. This allows us to later make inferences that test our research hypothesis statistically.

Let y_j denote a characteristic of the cascade of interest, namely cascade size, cascade lifetime, or structural virality. We then model y_j of the cascade via a two-level generalized hierarchical regression:

$$\text{Level 1:} \quad y_j = \alpha_i + \beta^T \phi_j + \gamma^T x_j + \varepsilon_j, \tag{2}$$

$$\text{Level 2:} \quad \alpha_i = \gamma_0 + \gamma_i, \tag{3}$$

where level 1 refers to the cascade level and level 2 to the rumor level. The other variables are as follows. The coefficient β captures the marginal effect of emotions. This is later our variable of interest as it measures the contribution of emotions to rumor diffusion. The coefficient γ is used to control for other model variables at the rumor cascade level. Both γ_0 and γ_i are assumed to be independent and identically normally distributed with mean zero. Then γ_0 reflects the base diffusion in the sample, while γ_i controls for variation at rumor level. Notably, this turns α_i into a rumor-specific random effect. The error term ε_j is assumed to be independent and identically normally distributed with mean zero.

The use of regression analysis is imperative for the scope of our study. The reasons are as follows. (1) Our objective is different from predictive modeling [51], where the focus is on accurate estimates of the outcome variable. Instead, we are concerned with the model logic as it allows us to interpret the model coefficients. (2) Our objective is also different from analyzing summary statistics as in [11]. Summary statistics deal with comparisons across groups and thereby ignore other sources of heterogeneity in the sample. For instance, the summary statistics on rumor emotions in [11] only report which emotions are common but not how emotions are associated with sharing dynamics. This is especially relevant for our research as we expect that some properties of rumor diffusion are also due to the social influence of the sender. Hence, by combining emotions and further controls in a joint regression model, we can isolate the *marginal* effect of emotions on the diffusion dynamics, which would not be possible with summary statistics.

Later, a regression analysis based on basic emotions is precluded due to multicollinearity (recall that the emotion scores e_j sum to one across basic emotions). Instead, the regression analysis is performed using bipolar emotion pairs ϕ_j^{pairs} and the dyadic emotional interactions $\phi_j^{primary}$, $\phi_j^{secondary}$, $\phi_j^{tertiary}$. For the latter, we fit 12 separate models, i.e., one for each pair among the emotional dyads, due to linear dependencies between the dyads.

In our implementation, the estimator depends on the distribution of y_j as follows:

1 Cascade size is modeled via a negative binomial regression with log-transformation. The reason is that cascade size denotes count data with overdispersion (i.e., variance larger than the mean).

2 Cascade lifetime is first log-transformed and then modeled via a normal distribution. This is consistent with previous research assuming a log-normal distribution for response times [12].

3 Structural virality is modeled via a gamma regression with a log-link. This allows us to account for a skewed distribution of continuous, non-negative variables.

All estimations are conducted based on the R package `lme4`. Before estimation, all model variables are z-standardized. Owing to this, the regression coefficients quantify changes in the dependent variable in standard deviations. This is beneficial as it allows us to compare the estimated coefficients across emotions in a straightforward manner.

3 Results

3.1 Summary statistics

The diffusion dynamics in our data are as follows. Figure 2 compares cascade size, lifetime, and structural virality via complementary cumulative distribution functions (CCDF). On average, a rumor cascade reaches 31.95 users and has a lifetime of 123.18 hours. The mean structural virality is 1.26.

Basic emotions: Fig. 3 plots the CCDFs for each of the eight basic emotions, while Fig. 4 reports the relative proportion of emotional intensity averaged over all rumors. We find that a large proportion of rumors embed DISGUST and SURPRISE, whereas comparatively few rumors embed JOY and SADNESS. Evidently, rumors embed more ANGER (relative share of 12.34%) than FEAR (10.74%), more SURPRISE (16.44%) than ANTICIPATION (14.23%), more DISGUST (23.58%) than TRUST (9.05%), and more JOY (7.39%) than SADNESS (6.23%). Overall,

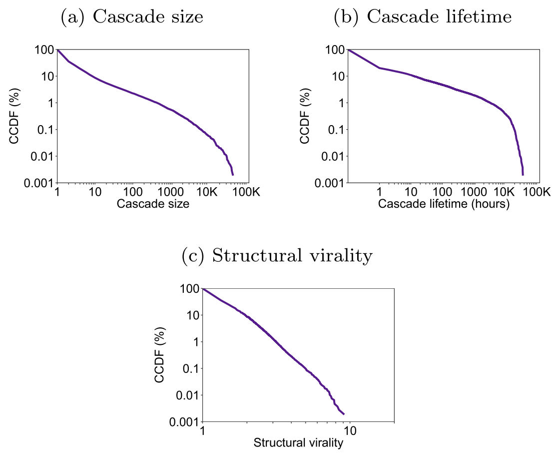

Figure 2 Complementary cumulative distribution functions (CCDFs) for (**a**) cascade size, (**b**) cascade lifetime, and (**c**) structural virality

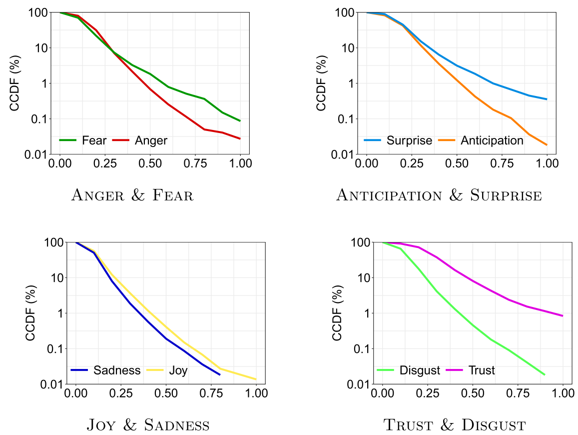

Figure 3 Complementary cumulative distribution functions (CCDFs) for basic emotions

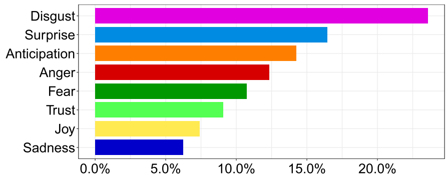

Figure 4 Relative proportion of basic emotions in cascades

(a) Primary dyadic emotional interactions (b) Secondary dyadic emotional interactions

(c) Tertiary dyadic emotional interactions

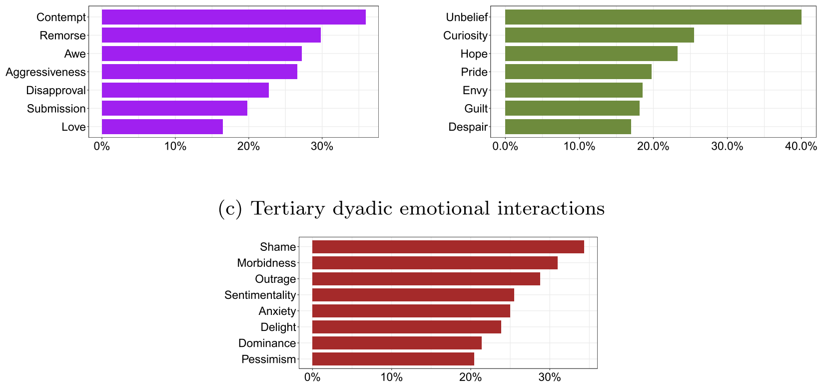

Figure 5 Relative proportion of dyadic emotional interactions in cascades

43.01% of the embedded emotions originate from the group of positive emotions, while 56.98% belong to the group of negative emotions. Hence, rumors comprise more negative than positive emotions.

Dyadic Emotional Interactions: Fig. 5 shows the distribution of the dyadic emotional interactions. For the primary emotion dyads, we find that a large proportion of rumors embed CONTEMPT and REMORSE, whereas fewer rumors embed LOVE and SUBMISSION. For the secondary and tertiary emotion dyads, we find that many rumor cascades embed UNBELIEF and SHAME. In contrast, only a relatively small proportion of rumors embed DESPAIR and PESSIMISM.

Note that the above summary statistics only report the relative frequency of emotions but do not allow one to draw conclusions regarding how users respond to emotions. This is studied in the following regression analyses.

3.2 Regression results from bipolar emotion pairs

In the following, we report results for the bipolar emotion pairs ϕ_j^{pairs}.

We use regression analysis to explain different characteristics of cascades based on the bipolar emotion pairs. The parameter estimates in Fig. 6 show that the 8 basic emotions are important determinants of the spreading dynamics of rumors. Across all dependent variables, we find coefficients that are positive and statistically significant for the ANTICIPATION–SURPRISE, ANGER–FEAR, and TRUST–DISGUST dimensions. Hence, rumors are estimated to diffuse more pronouncedly when embedding positive emotions. For instance,

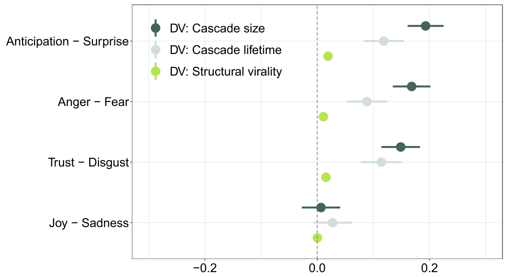

Figure 6 Standardized parameter estimates and 95% confidence intervals

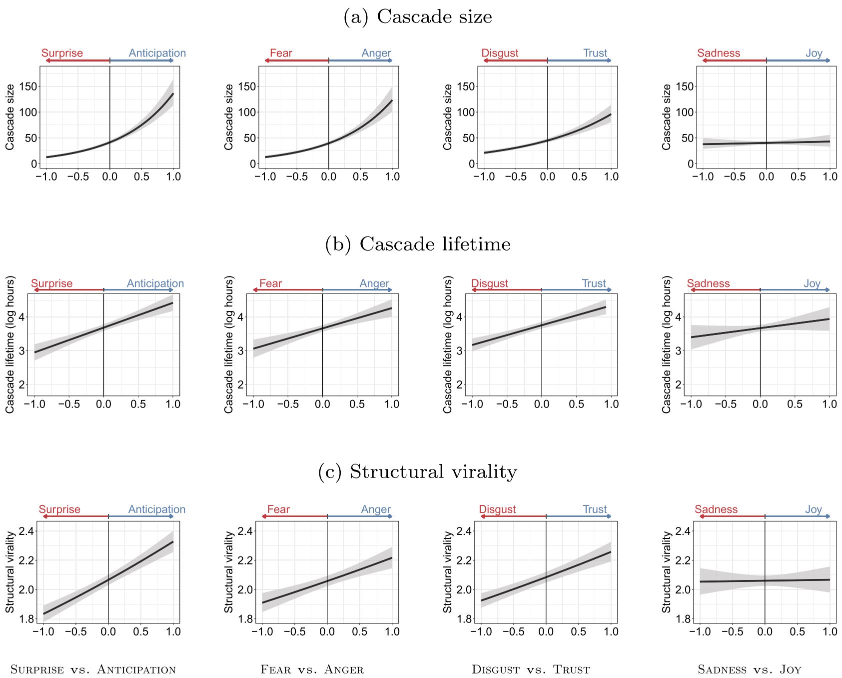

Figure 7 Predicted marginal effect of bipolar emotions on (**a**) cascade size, (**b**) cascade lifetime, and (**c**) structural virality. The 95% confidence intervals are highlighted in gray

the estimated effect sizes for the ANTICIPATION–SURPRISE pair are as follows: the coefficients amount to 0.193 for cascade size (p-value < 0.001), to 0.118 for cascade lifetime (p-value < 0.001), and to 0.019 for structural virality (p-value < 0.001). Hence, a one standard deviation change in this bipolar emotion pair is linked to a 21.29% increase in the cascade size, a 12.52% increase in the cascade lifetime, and a 1.92% increase in structural virality.

The predicted marginal effects for the bipolar emotion pairs are shown in Fig. 7. Rumors embedding ANTICIPATION, ANGER, and TRUST generate more reshares, spread over a longer time horizon, and become more viral. The coefficient for the JOY–SADNESS emotion pair is not significant.

Our regression model controls for heterogeneity in users' social influence. The corresponding estimates are omitted for the sake of brevity (their findings have been discussed elsewhere, e.g., in [31]). In short, rumor cascades initiated from accounts that are verified and younger are linked to a larger, longer, and more viral spread. Similar relationships are observed for users exhibiting a higher engagement level and a greater number of followers. In contrast, a higher number of followees is negatively associated with the size, lifetime, and structural virality of a cascade.

We calculated the pseudo-R^2 for each model, resulting in relatively high values of 0.64 for cascade size, 0.43 for cascade lifetime, and 0.31 for structural virality. Evidently, the model variables explain the variation in the dependent variables to a large extent. Furthermore, a visual inspection of the actual vs. fitted plot and goodness-of-fit tests indicate that the models are well specified. This is also supported when considering the differences between the AIC models for individual models estimated with/without emotion variables. For each dependent variable, the difference is greater than the threshold [52] of 10 (difference in cascade size: 226.16; lifetime: 52.22; structural virality: 121.03), indicating strong support for the corresponding candidate models. Therefore, the inclusion of the emotion variables in the regression model is to be preferred.

3.3 Regression results from dyadic emotional interactions

We now study how the presence of 24 dyadic emotional interactions is associated with the diffusion dynamics of online rumors. For this purpose, we employ the previous regression model, but this time include the emotion variables $\phi_j^{primary}$, $\phi_j^{secondary}$, and $\phi_j^{tertiary}$. Figure 8 shows the predicted marginal effects for the 8 primary, 8 secondary, and 8 tertiary dyadic emotional interactions.

Primary dyadic emotional interactions: Rumor cascades with higher values of AGGRESSIVENESS, LOVE, OPTIMISM are larger in size, longer-lived, and more viral. We observe no statistically significant effect for the SUBMISSION–CONTEMPT pair. Overall, the largest positive association is observed for AGGRESSIVENESS (i.e., the combination of ANTICIPATION and ANGER). An increase of one standard deviation in this dimension is linked to a 19.18% increase in the cascade size, an 8.33% increase in the cascade lifetime, and a 1.69% increase in structural virality.

Secondary dyadic emotional interactions: Rumor cascades with higher values of HOPE vs. UNBELIEF generate more reshares, spread over a longer time horizon, and become more viral. We further find that rumor cascades embedding GUILT, and DESPAIR are negatively associated with the size, lifetime, and structural virality of a cascade. The CURIOSITY–CYNICISM pair is not statistically significant at common statistical significance levels.

Tertiary dyadic emotional interactions: Rumor cascades with higher values of ANXIETY are larger in size, longer-lived, and more viral. We also find a larger size, lifetime, and virality for rumor cascades embedding high levels of DOMINANCE, PESSIMISM, and ANXIETY. We find no statistically significant effect for the SENTIMENTALITY–MORBIDNESS pair.

The control variables tend in a similar direction as in the analysis of the basic emotions. Again, the difference in AIC (comparing the model with and without emotions) is above the common threshold of 10 [52]. Therefore, the models that include emotions are to be preferred.

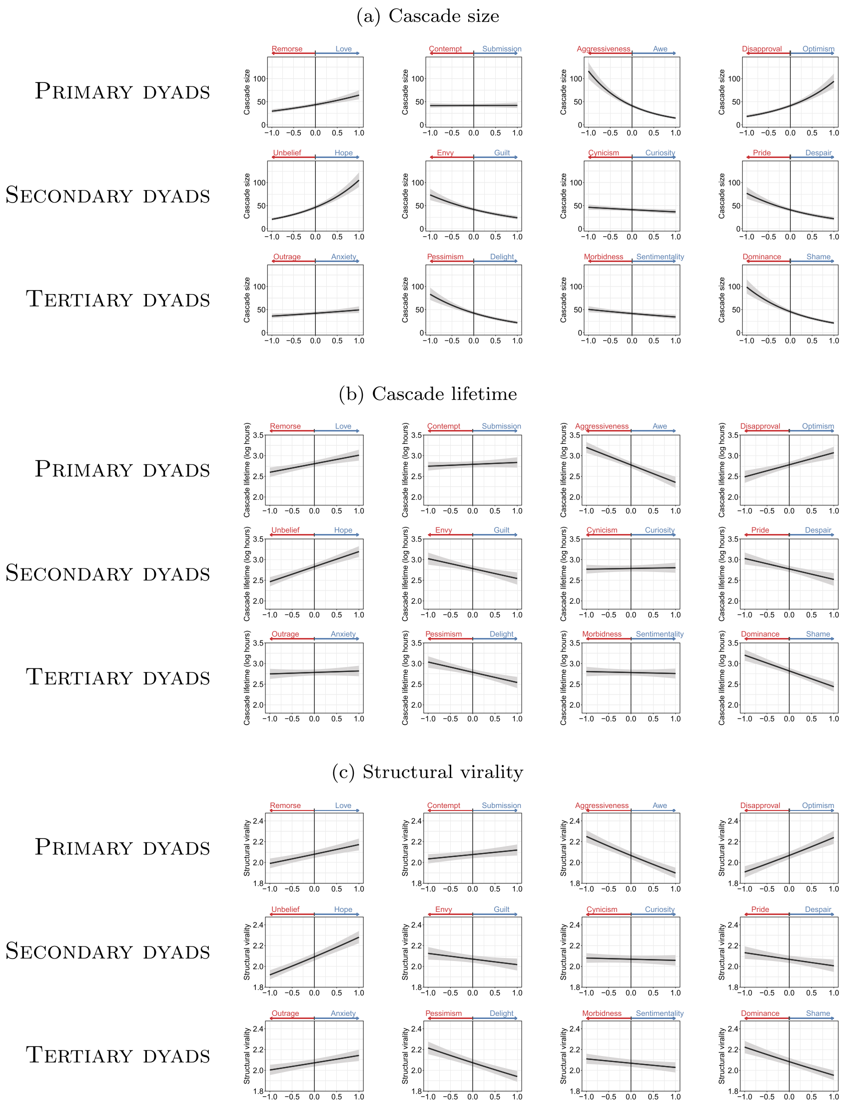

Figure 8 Predicted marginal effect of dyadic emotional interactions on (**a**) cascade size, (**b**) cascade lifetime, and (**c**) structural virality. The 95% confidence intervals are highlighted in gray

3.4 Sensitivity across rumor topics

Our empirical analysis is based on a large-scale dataset with Twitter rumors across varying topics. We now study topic-specific variations. For this purpose, we employ the topic categorization from [11], which classifies Twitter rumors into topics. Here, we focus on the topics Politics, Business, and Science given their high relevance for society. Note that the topic Science is broadly defined and also comprises related topics such as health-related rumors. For each of the three topics, we generate a subset of the data and re-estimate our models. The results are visualized in Fig. 9. We find that emotions explain

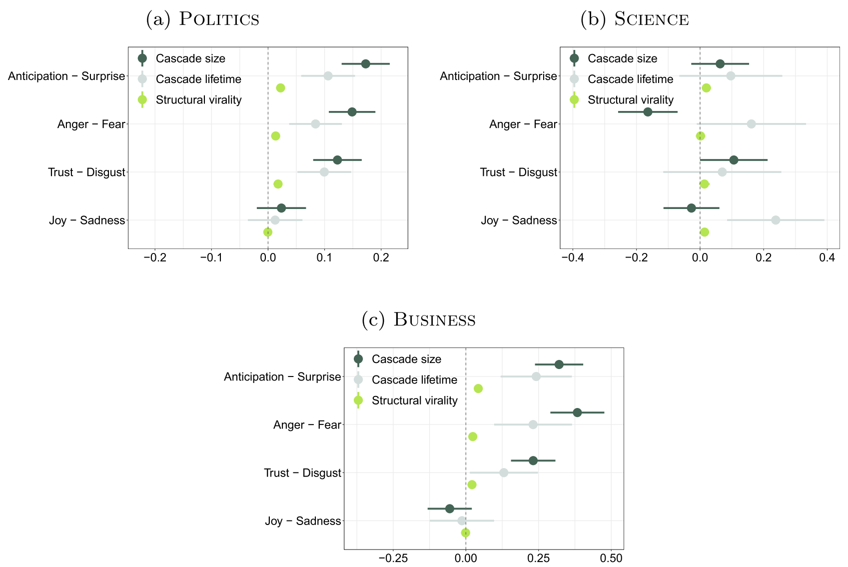

Figure 9 Standardized parameter estimates and 95% confidence intervals for different subsets of rumors filtered by topic

differences in cascade size, cascade lifetime, and structural virality at a statistically significant level for the topics Politics and Business. In contrast, we find mixed results for Science. These results are in line with existing literature. For example, [31] find a pronounced role of political content in social media sharing. The authors argue that political topics are more controversial and thus attract more attention, which itself influences sharing behavior.

3.5 Robustness checks

3.5.1 Model checks

We conducted a series of additional model checks that contribute to the robustness of our findings. First, we followed common practice in regression analysis and checked that variance inflation factors as an indicator of multicollinearity were below five [53]. This check led to the desired outcome. Second, we controlled for year-level time effects (i.e., via clustered standard errors and different study horizons) in addition to rumor-level random effects that are already included in our regression model. We obtained conclusive findings. Third, we controlled for non-linear relationships via quadratic terms. In all cases, our findings were supported.

3.5.2 Validation of emotion scores

Our results rely on the validity of dictionaries to extract emotions from online rumors. To check how perceived emotions in rumors align with the dictionary-based emotions, we conducted a survey using the online survey platform Prolific (https://www.prolific.co/). We asked $n = 7$ participants (English native speakers) to rate the presence of the eight basic emotions on a Likert scale from −3 to 3 (here: −3 indicates no emotion present while 3 refers to a high degree of emotion present) for a set of 100 randomly sampled rumors. As

Table 1 Kendal's *W* coefficient for the interrater agreement between survey participants

Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
0.474***	0.198***	0.427***	0.406***	0.364***	0.408***	0.227***	0.230***

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

shown in Table 1, the participants exhibited a statistically significant interrater agreement according to Kendall's W for each of the 8 basic emotions ($p < 0.01$).

Overall, when aggregating across all 8 basic emotions, the correlation between the dictionary-based emotion scores and human annotations is $\rho = 0.17$ ($p < 0.01$) and thus statistically significant at common significance thresholds. This demonstrates that dictionaries are able to capture emotions in online rumors.

3.5.3 Negation handling

We performed negation scope detection [54, 55] to analyze the robustness to how negations (e.g., "not," "no") are handled by the dictionary approach. For example, phrases like "I am surprised" and "I am not surprised" contain the same number of emotional words but convey different emotions to the reader. We analyzed emotional words that are negated by surrounding negation words as follows: (i) We searched for negations using a predefined list of negation words. Here, we used the list of negations from the R package `sentimentr`. (ii) We recalculated the emotion scores by counting all emotional words in the neighborhood of the negation word as belonging to the opposite emotional dimension (e.g., $Joy = Joy + Sadness_{negated}$). The neighborhood is set to 5 words before and 2 words after the negation. We then compared the emotion scores with negation handling to the values obtained without negation handling. As a result, we found that merely 5.58% of the emotional words in rumors are affected by negations (i.e., lie within negation scopes). Furthermore, the emotion scores with negation handling are highly correlated with the emotion scores without negation handling ($\rho > 0.9$). Altogether, this implies that our analysis and findings are robust to negations.

4 Discussion

In this work, we provided a large-scale study of emotions in online rumor diffusion. For this purpose, 2189 rumors from Twitter with approx. 3.7 million reshares were analyzed with regard to the embedded emotions. Overall, we found that negative emotions are frequently embedded in rumors. Especially frequent are DISGUST (relative share of 23.58%) and SURPRISE (16.44%). (2) The relationship between emotions and the structure of cascade is statistically significant at common significance levels for almost all emotions under study. (3) Rumors embedding ANTICIPATION, ANGER, and, TRUST are estimated to reach a significantly larger number of individuals and diffuse significantly longer and more virally. Interestingly, while negative emotions are more often embedded in rumors, positive emotions are particularly relevant for explaining the diffusion dynamics. (4) A particularly large effect of emotions on the diffusion characteristics is found for AGGRESSIVENESS (which is a derived emotion composed of ANTICIPATION and ANGER). A one standard deviation higher level of AGGRESSIVENESS is predicted to generate 19.18% more reshares, to be active for 8.33% longer, and to spread 1.69% more virally. Overall, our study establishes emotions as important determinants that describe the spread of online rumor.

Our results contribute to the understanding of online rumor diffusion. As shown by our analysis, emotions are important determinants in explaining the structure of rumor cascades, specifically how many users are involved, the active lifespan and, to a lesser extent, structural virality. The findings are consistent across basic emotions and also dyadic emotion interaction (primary, secondary, tertiary). In addition, our results suggest considerable heterogeneity in the role of emotions. Strong effects are found for most basic emotions (ANTICIPATION, SURPRISE, ANGER, FEAR, TRUST, DISGUST), albeit with the exception of JOY and SADNESS. Similar patterns are observed when studying more complex (derived) emotions. Here, the largest estimated effect size is associated with AGGRESSIVENESS. A one standard deviation higher level of AGGRESSIVENESS is predicted to generate 19.18% more reshares, cascade that are 8.33% longer, and a 1.69% increase in structural virality. Thereby, we reveal AGGRESSIVENESS as a dominant driver of rumor diffusion.

Our work also expands upon rumor theory from offline settings. Offline rumors have a higher chance of dissemination when conveying anxiety [56] and, in particular, negative emotions [42, 43]. However, the underlying evidence stems from offline rumors rather than online rumors. Our work adds in two ways: First, we study the role of emotions in the diffusion of *online* rumors. While rumor diffusion in offline settings is more pronounced for negative emotions, we observe the opposite for online rumors, for which positive emotions appear more influential. Second, we not only compare positive vs. negative emotions but perform a granular study across primary, secondary, and tertiary emotional dyadic interactions. This provides rich findings on the *heterogeneity* of emotion effects. As such, we confirm that ANXIETY is an important driver for rumor diffusion not only in offline but also in online settings. However, further emotions are also relevant: a particularly pronounced role is found with regard to AGGRESSIVENESS. To the best of our knowledge, the importance of AGGRESSIVENESS in rumor diffusion was previously overlooked.

In our study, inferences were made based on data from Twitter. Twitter has a wide popularity with more than 300 million active users. In addition, it plays an important part in rumor diffusion due to its influential role in the political discourse [10]. This makes our findings directly relevant to both social media platforms and, in particular, public stakeholders. For the same reason, established procedures were followed when compiling the data [11], as this ensures that findings are drawn from a realistic, large-scale dataset of Twitter rumors. To the best of our knowledge, our work is the first statistical analysis linking emotions to online rumor diffusion.

As with other studies, ours is subject to limitations that provide opportunities for future research. First, this study is based on observational inferences, while we leave the extension to (quasi-)experimental settings, and thus causal inferences, to future work. Nevertheless, our study design ensures that many potential confounding factors can be ruled out. This is because of the temporal order (i.e., the emotion-laden wording *precedes* the actual cascade) and the fact that further sources of variability among rumors are captured through rumor-level random effects. Second, our study employs statistical inferences that provide explanatory insights. This allows us to quantify the *marginal* contribution of emotions to online rumor diffusion. A different objective is to use emotions for predictive modeling, which is discussed elsewhere [57–60].

Our work entails several implications. It emphasizes the necessity of considering emotions when studying rumor diffusion. Emotions are also relevant in practice, particularly for social media platforms. To counter the proliferation of online rumors, social media platforms should seek solutions, based on which emotions can be actively managed. Our study also encourages a granular investigation of emotions for related research questions, whereby not only basic emotions but also derived emotions are considered. Such granular analyses are comparatively more challenging in lab experiments; however, a remedy is offered by computational social science based on which large-scale datasets from online behavior can be mined.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-021-00307-5.

Additional file 1. S1—Background Literature (PDF 90 kB)

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
All data needed to evaluate the conclusions in the paper are publicly available (and the source reported in the paper). Replication code for this study is available via https://github.com/DominikBaer95/Emotions_Rumor_Diffusion.

Declarations

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
NP and SF designed the study. NP and DB analyzed the data. NP, DB, and SF wrote and revised the manuscript. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Author details
[1]University of Giessen, Licher Str. 62, 35394 Giessen, Germany. [2]LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany. [3]ETH Zurich, Weinbergstr. 56/58, 8092 Zurich, Switzerland.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 18 May 2021 Accepted: 20 September 2021 Published online: 18 October 2021

References

1. Berger J (2011) Arousal increases social transmission of information. Psychol Sci 22(7):891–893. https://doi.org/10.1177/0956797611413294
2. Berger J, Milkman KL (2012) What makes online content viral? J Mark Res 49(2):192–205. https://doi.org/10.1509/jmr.10.0353
3. Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web 1(1):5. https://doi.org/10.1145/1232722.1232727
4. Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. Mark Sci 23(4):545–560. https://doi.org/10.1287/mksc.1040.0071
5. de Domenico M, Lima A, Mougel P, Musolesi M (2013) The anatomy of a scientific rumor. Sci Rep 3:2980. https://doi.org/10.1038/srep02980
6. Starbird K, Maddock J, Orand M, Achterman P, Mason RM (2014) Rumors, false flags, and digital vigilantes: misinformation on Twitter after the 2013 Boston marathon bombing. In: iConference
7. Starbird K (2017) Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In: International AAAI conference on web and social media (ICWSM)
8. Aral S, Eckles D (2019) Protecting elections from social media manipulation. Science 365(6456):858–861. https://doi.org/10.1126/science.aaw8243
9. Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239):1130–1132. https://doi.org/10.1126/science.aaa1160
10. Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. Science 363(6425):374–378. https://doi.org/10.1126/science.aau2706
11. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559
12. Zaman T, Fox EB, Bradlow ET (2014) A Bayesian approach for predicting the popularity of tweets. Ann Appl Stat 8(3):1583–1611. https://doi.org/10.1214/14-AOAS741

13. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: International world wide web conference (WWW)
14. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: International world wide web conference (WWW). https://doi.org/10.1145/1772690.1772751.
15. Lerman K, Ghosh R (2010) Information contagion: an empirical study of spread of news on Digg and Twitter social networks. In: International AAAI conference on web and social media (ICWSM)
16. Arif A, Shanahan K, Chou F-J, Dosouto Y, Starbird K, Spiro ES (2016) How information snowballs: exploring the role of exposure in online rumor propagation. In: ACM conference on computer-supported cooperative work & social computing (CSCW). https://doi.org/10.1145/2818048.2819964
17. Spiro ES, Fitzhugh S, Sutton J, Pierski N, Greczek M, Butts CT (2012) Rumoring during extreme events: a case study of deepwater horizon 2010. In: ACM web science conference (WebSci). https://doi.org/10.1145/2380718.2380754
18. Kryvasheyeu Y, Chen H, Obradovich N, Moro E, van Hentenryck P, Fowler J, Cebrian M (2016) Rapid assessment of disaster damage using social media activity. Sci Adv 2(3):1500779
19. Zeng L, Starbird K, Spiro ES (2016) Rumors at the speed of light? Modeling the rate of rumor transmission during crisis. In: Hawaii international conference on system sciences (HICSS). https://doi.org/10.1109/HICSS.2016.248
20. Sauter DA, Eisner F, Ekman P, Scott SK (2010) Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc Natl Acad Sci USA 107(6):2408–2412. https://doi.org/10.1073/pnas.0908239106
21. Ekman P (1992) An argument for basic emotions. Cogn Emot 6(3–4):169–200. https://doi.org/10.1080/02699939208411068
22. Plutchik R (2001) The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
23. Zhang P (2013) The affective response model: a theoretical framework of affective concepts and their relationships in the ICT context. MIS Q 37(1):247–274
24. Barsade SG (2002) The ripple effect: emotional contagion and its influence on group behavior. Adm Sci Q 47(4):644. https://doi.org/10.2307/3094912
25. Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci USA 111(24):8788–8790. https://doi.org/10.1073/pnas.1320040111
26. Goldenberg A, Gross JJ (2020) Digital emotion contagion. Trends Cogn Sci 24(4):316–328. https://doi.org/10.1016/j.tics.2020.01.009
27. Fan R, Varol O, Varamesh A, Barron A, van de Leemput IA, Scheffer M, Bollen J (2019) The minute-scale dynamics of online emotions reveal the effects of affect labeling. Nat Hum Behav 3(1):92–100. https://doi.org/10.1038/s41562-018-0490-5
28. Ferrara E, Yang Z (2015) Measuring emotional contagion in social media. PLoS ONE 10(11):0142390. https://doi.org/10.1371/journal.pone.0142390
29. Alvarez R, Garcia D, Moreno Y, Schweitzer F (2015) Sentiment cascades in the 15M movement. EPJ Data Sci 4(1):407. https://doi.org/10.1140/epjds/s13688-015-0042-4
30. Zollo F, Novak PK, Del Vicario M, Bessi A, Mozetič I, Scala A, Caldarelli G, Quattrociocchi W (2015) Emotional dynamics in the age of misinformation. PLoS ONE 10(9):0138740. https://doi.org/10.1371/journal.pone.0138740
31. Stieglitz S, Dang-Xuan L (2013) Emotions and information diffusion in social media: sentiment of microblogs and sharing behavior. J Manag Inf Syst 29(4):217–248. https://doi.org/10.2753/MIS0742-1222290408
32. Naveed N, Gottron T, Kunegis J, Alhadi AC (2011) Bad news travel fast: a content-based analysis of interestingness on Twitter. In: International web science conference (WebSci). https://doi.org/10.1145/2527031.2527052
33. Kim J, Yoo J (2012) Role of sentiment in message propagation: reply vs. retweet behavior in political communication. In: International conference on social informatics. https://doi.org/10.1109/SocialInformatics.2012.33
34. Kissler J, Herbert C, Peyk P, Junghofer M (2007) Buzzwords: early cortical responses to emotional words during reading. Psychol Sci 18(6):475–480. https://doi.org/10.1111/j.1467-9280.2007.01924.x
35. Luminet O, Bouts P, Delie F, Manstead ASR, Rimé B (2000) Social sharing of emotion following exposure to a negatively valenced situation. Cogn Emot 14(5):661–688. https://doi.org/10.1080/02699930050117666
36. Rimé B (2009) Emotion elicits the social sharing of emotion: theory and empirical review. Emot Rev 1(1):60–85. https://doi.org/10.1177/1754073908097189
37. Peters K, Kashima Y, Clark A (2009) Talking about others: emotionality and the dissemination of social information. Eur J Soc Psychol 39(2):207–222. https://doi.org/10.1002/ejsp.523
38. Chuai Y, Zhao J (2020) Anger makes fake news viral online. arXiv:2004.10399
39. Wu S, Tan C, Kleinberg J, Macy M (2011) Does bad news go away faster? In: International AAAI conference on web and social media (ICWSM)
40. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer. In: International conference on web search and data mining (WSDM). https://doi.org/10.1145/1935826.1935845
41. Brady WJ, Wills JA, Jost JT, Tucker JA, van Bavel JJ (2017) Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci USA 114(28):7313–7318. https://doi.org/10.1073/pnas.1618923114
42. Anthony S (1973) Anxiety and rumor. J Soc Psychol 89(1):91–98. https://doi.org/10.1080/00224545.1973.9922572
43. Knapp RH (1944) A psychology of rumor. Public Opin Q 8(1):22–37
44. Martel C, Pennycook G, Rand DG (2020) Reliance on emotion promotes belief in fake news. Cogn Res Princ Implic 5(1):47. https://doi.org/10.1186/s41235-020-00252-3
45. Weeks BE (2015) Emotions, partisanship, and misperceptions: how anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. J Commun 65(4):699–719. https://doi.org/10.1111/jcom.12164
46. Acerbi A (2019) Cognitive attraction and online misinformation. Palgrave Commun 5(1):15. https://doi.org/10.1057/s41599-019-0224-y
47. Goel S, Anderson A, Hofman J, Watts DJ (2016) The structural virality of online diffusion. Manag Sci 62(1):180–196. https://doi.org/10.1287/mnsc.2015.2158
48. Allport GW, Postman L (1947) The psychology of rumor. Holt, New York
49. Kratzwald B, Ilić S, Kraus M, Feuerriegel S, Prendinger H (2018) Deep learning for affective computing: text-based emotion recognition in decision support. Decis Support Syst 115:24–35. https://doi.org/10.1016/j.dss.2018.09.002

50. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x
51. Breiman L (2001) Statistical modeling: the two cultures. Stat Sci 16(3):199–231
52. Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304
53. Akinwande MO, Dikko HG, Samson A et al (2015) Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. Open J Stat 5(7):754–767
54. Pröllochs N, Feuerriegel S, Neumann D (2019) Learning interpretable negation rules via weak supervision at document level: a reinforcement learning approach. In: Conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT)
55. Pröllochs N, Feuerriegel S, Lutz B, Neumann D (2020) Negation scope detection for sentiment analysis: a reinforcement learning framework for replicating human interpretations. Inf Sci 536:205–221. https://doi.org/10.1016/j.ins.2020.05.022
56. Rosnow RL (1991) Inside rumor: a personal journey. Am Psychol 46(5):484–496
57. Castillo C, Mendoza M, Poblete B (2011) Information credibility on Twitter. In: International world wide web conference (WWW). https://doi.org/10.1145/1963405.1963500
58. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Prominent features of rumor propagation in online social media. In: International conference on data mining (ICDM). https://doi.org/10.1109/ICDM.2013.61
59. Kwon S, Cha M, Jung K (2017) Rumor detection over varying time windows. PLoS ONE 12(1):0168344. https://doi.org/10.1371/journal.pone.0168344
60. Ducci F, Kraus M, Feuerriegel S (2020) Cascade-LSTM: a tree-structured neural classifier for detecting misinformation cascades. In: ACM SIGKDD conference on knowledge discovery and data mining (KDD)

9 Emotions explain differences in the diffusion of true vs. false social media rumors

Title: Emotions explain differences in the diffusion of true vs. false social media rumors

Abstract: False rumors (often termed "fake news") on social media pose a significant threat to modern societies. However, potential reasons for the widespread diffusion of false rumors have been underexplored. In this work, we analyze whether sentiment words, as well as different emotional words, in social media content explain differences in the spread of true vs. false rumors. For this purpose, we collected rumor cascades from Twitter, comprising more than 4.5 million retweets that have been fact-checked for veracity. We then categorized the language in social media content to (1) sentiment (i.e., positive vs. negative) and (2) eight basic emotions (i.e., anger, anticipation, disgust, fear, joy, trust, sadness, and surprise). We find that sentiment and basic emotions explain differences in the structural properties of true vs. false rumor cascades. False rumors (as compared to true rumors) are more likely to go viral if they convey a higher proportion of terms associated with a positive sentiment. Further, false rumors are viral when embedding emotional words classified as trust, anticipation, or anger. All else being equal, false rumors conveying one standard deviation more positive sentiment have a 37.58 % longer lifetime and reach 61.44 % more users. Our findings offer insights into how true vs. false rumors spread and highlight the importance of managing emotions in social media content.

Author contributions: Nicolas Pröllochs and Stefan Feuerriegel designed the study. Nicolas Pröllochs and Dominik Bär analyzed the data. Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel wrote and revised the manuscript.

Reference: Pröllochs, N., **Bär, D.**, & Feuerriegel, S. 2021. *Emotions explain differences in the diffusion of true vs. false social media rumors.* In: **Scientific Reports** 11, 22721, DOI: `https://doi.org/10.1038/s41598-021-01813-2`.

Supplementary materials: Supplementary materials for this article are in Supplementary Material B.

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Emotions explain differences in the diffusion of true vs. false social media rumors

Nicolas Pröllochs[1]✉, Dominik Bär[2] & Stefan Feuerriegel[2,3]

False rumors (often termed "fake news") on social media pose a significant threat to modern societies. However, potential reasons for the widespread diffusion of false rumors have been underexplored. In this work, we analyze whether sentiment words, as well as different emotional words, in social media content explain differences in the spread of true vs. false rumors. For this purpose, we collected *N* = 126,301 rumor cascades from Twitter, comprising more than 4.5 million retweets that have been fact-checked for veracity. We then categorized the language in social media content to (1) sentiment (i.e., positive vs. negative) and (2) eight basic emotions (i. e., anger, anticipation, disgust, fear, joy, trust, sadness, and surprise). We find that sentiment and basic emotions explain differences in the structural properties of true vs. false rumor cascades. False rumors (as compared to true rumors) are more likely to go viral if they convey a higher proportion of terms associated with a positive sentiment. Further, false rumors are viral when embedding emotional words classified as trust, anticipation, or anger. All else being equal, false rumors conveying one standard deviation more positive sentiment have a 37.58% longer lifetime and reach 61.44% more users. Our findings offer insights into how true vs. false rumors spread and highlight the importance of managing emotions in social media content.

A vast number of social media users have been exposed to knowingly false content. This was confirmed to be the case during humanitarian crises[1] and elections[2–5]. For example, in the 2016 U. S. presidential election, each adult was shown, on average, more than one item with false content[6]. On top of that, there were more user interactions with deliberately false content than with reliable information sources[7]. To this end, false content on social media poses a threat to individuals, organizations, and even whole societies[8,9].

Understanding the spread of false content is of wide interest[2,9]. For users, understanding this phenomenon could yield certain signals based on which true and false content can be recognized. For social media platforms, a better understanding could inform the design of early warning systems that automatically detect the spread of false content[10]. Specifically, it would allow one to derive features from the propagation dynamics of false content that could then be fed into machine learning classifiers[11–14]. For policy makers, understanding the spread of false content is necessary for developing mitigation strategies that directly target the viral effects of false content (e. g., educating users to exercise more critical thinking when confronted with emotional content). This is especially critical as repeated exposure to false information has led many users to erroneously believe that it was true[15].

Only a few studies have focused on understanding differences in the spread of true vs. false social media content. True vs. false rumors have been compared across different characteristics of resharing cascades by Refs.[16,17]. They observed larger, wider, and deeper cascades for false rumors. Further, some emotions are more often found in false rumors[18]; however, it does not link emotions to differences in diffusion across true vs. false rumors.

In this work, we hypothesize that differences in the diffusion of true vs. false rumors can be explained by the conveyed sentiment and basic emotions. Our rationale is motivated by prior literature. Emotions are highly influential for human judgment and decision making[19], and strongly affect how humans draw or capture attention[20]. Emotions are highly contagious and thus spread through direct interaction within a social network[21–23]. Emotions have also been found to impact retweeting[24], thus driving diffusion[21,25,26]. To this end, emotional stimuli trigger cognitive processing[27], which in turn results in the behavioral response of information sharing[28–30]. Reliance on emotions further promotes belief in false information[31]. Altogether, this suggests that sentiment and emotions might offer a potential explanation for differences in the spreading dynamics of true vs. false rumors; however, empirical evidence is lacking.

Prior literature has established sentiment, as well as emotions, to be drivers of online diffusion[24,26,32–37]. However, these works suggest that their roles regarding different types of online content vary. For example, the

[1]JLU Giessen, Giessen 35394, Germany. [2]LMU Munich, Munich 80539, Germany. [3]ETH Zurich, Zurich 8092, Switzerland. ✉email: nicolas.proellochs@wi.jlug.de

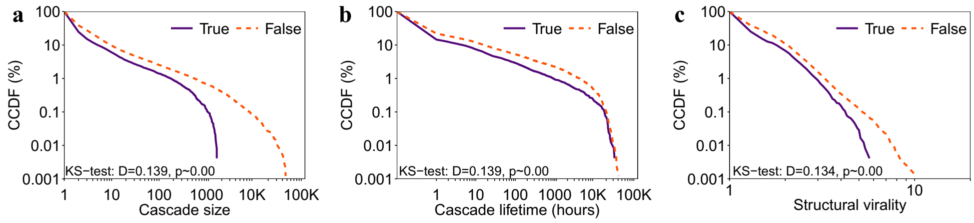

Figure 1. Complementary cumulative distribution functions (CCDFs) for different diffusion properties of social media cascades, namely, cascade size (**a**), cascade lifetime (**b**), and structural virality (**c**).

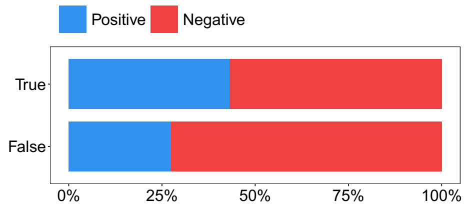

Figure 2. Relative frequency of true and false rumor cascades associated with positive vs. negative language.

spreading of news has been found to be promoted by positive sentiment[26,34], whereas the diffusion of health-related content is driven by negative sentiment[35]. Another work studies how sentiment promotes the diffusion of online rumors[38]. However, the sample used in this study only comprises rumors for a single crisis event, thus motivating us to analyze the role of sentiment and emotions in the spreading of true vs. false rumors.

We perform a large-scale explanatory analysis from observational data and, based on this, quantify to what extent language characterized by sentiment and basic emotions explain cascades of true vs. false rumors (see "Materials and methods"). We focus our analysis on three common structural properties of cascades: (1) size, (2) lifetime, and (3) the so-called "structural virality"[39]. These metrics quantify (1) how many users they reach, (2) how long rumors persist, and (3) how effectively they spread through the social network (i. e., a breadth-depth trade-off[39]).

Using a text mining framework, we extract sentiment and emotions embedded in replies to rumor cascades according to Plutchik's emotion model[40]. Plutchik's emotion model provides a comprehensive categorization across 8 basic emotions (i. e., anger, anticipation, joy, trust, fear, surprise, sadness, and disgust) that are regarded as universally recognized across cultures[41,42]. We compute a sentiment score that measures the overall valence of the text, that is, whether words are categorized more often as positive or negative. We then use hierarchical generalized linear models with one-way interactions in order to capture differences in the effects of sentiment and basic emotions across veracity. Here we control for between-rumor heterogeneity, specifically the social influence of senders (e. g., we correct for the number of followers, etc.).

To address our research questions, we analyze $N = 126{,}301$ rumor cascades from Twitter. Our data provides a large-scale, cross-sectional sample based on a comprehensive set of cascades on Twitter during the time period from the founding of Twitter in 2006 through 2017. In particular, our sample contains all English-language tweets that were subject to fact-checking by one of five different fact-checking organizations (see "Materials and methods"). Overall, this amounts to $\sim$ 4.5 million retweets by $\sim$ 3 million different users.

In summary, we study whether variations in language characterized as (1) positive and negative sentiment and (2) certain emotions (e. g., anger, anticipation, trust) explain differences in the structural properties of true vs. false rumor cascades on social media. For this, we draw upon a large-scale dataset of true and false rumors from Twitter and, on this basis, analyze the effect across a comprehensive, fine-grained set of emotions.

Results

Cascades of true and false rumors exhibit different structural properties. Figure 1 compares the diffusion based on the complementary cumulative distribution functions (CCDF). Overall, we find that false rumors are characterized by cascades of larger size and longer lifetime. For instance, the average cascade lifetime for false rumors is 149.61 h, whereas it is 71.62 h for true rumors. Furthermore, false rumors also entail cascades with higher structural virality.

True and false rumors also convey language of different sentiment and with different emotions. As shown in Fig. 2, the language in false rumors is more often associated with negative sentiment than in true rumors. In addition, Fig. 3 shows that false rumors convey a higher proportion of words classified as disgust, fear, and

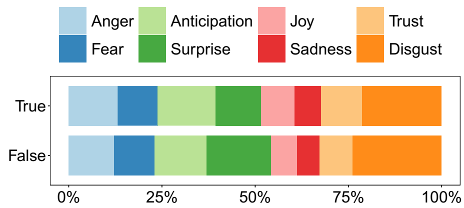

Figure 3. Average emotion score in true and false rumor cascades, following Plutchik's wheel of emotions[40].

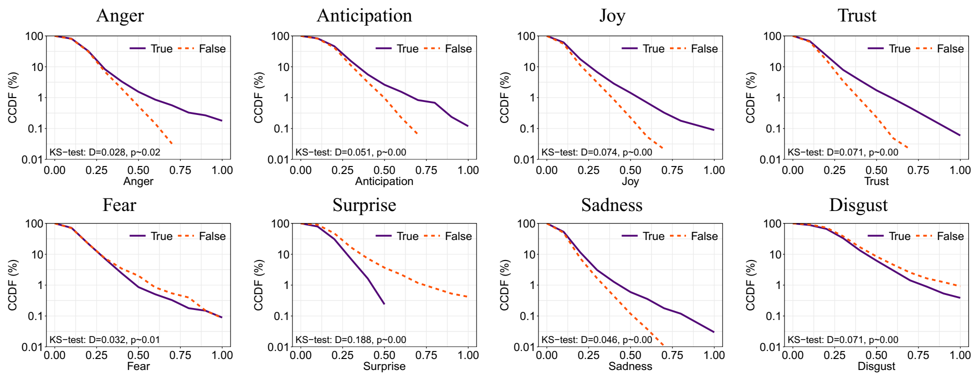

Figure 4. Complementary cumulative distribution functions (CCDFs) for conveyed emotions. Statistical comparisons are based on a Kolmogorov–Smirnov (KS) test.

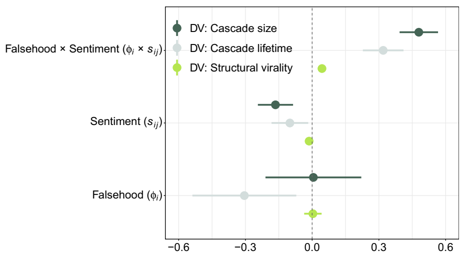

Figure 5. Standardized parameter estimates and 95% confidence intervals.

surprise, while true rumors are more likely to be linked to anger, anticipation, joy, sadness, and trust. In Fig. 4, we plot the CCDFs for each of the eight basic emotions. Evidently, false rumors are more likely to contain words associated with fear, disgust, and surprise, whereas true rumors contain words associated with sadness but also anger, anticipation, joy, and trust. Kolmogorov-Smirnov (KS) tests confirm that these differences are statistically significant.

Analysis of sentiment. We fit explanatory regression models to evaluate how variations in sentiment (i. e., the difference between positive vs. negative word counts) are associated with differences in the structural properties of true vs. false rumor cascades (see "Materials and methods" and Supplementary Table S1). In Fig. 5, the parameter estimates establish a pronounced role of sentiment (s_{ij}) with significantly different estimates for true vs. false rumors. For each dependent variable (DV), we observe negative coefficients for the sentiment variable, meaning that true rumors diffuse more pronouncedly if negative language is present. The positive coefficient

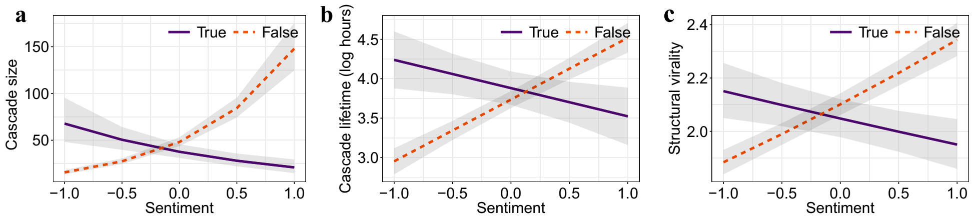

Figure 6. Predicted marginal means of cascade size (**a**), cascade lifetime (**b**), and structural virality (**c**) for different values of the sentiment variable. The 95% confidence intervals are highlighted in gray.

for the interaction term (*Sentiment* × *Falsehood*) suggests the opposite effect for false rumors. Compared to true rumors, a one standard deviation more positive sentiment for false rumors is linked to a 61.44% increase in the cascade size, a 37.58% increase in the cascade lifetime, and a 4.81% increase in structural virality. Notably, the estimated effect sizes are larger for false as compared to true rumors. Hence, positive sentiment appears to promote the diffusion of false rumors (while negative sentiment is estimated to promote the diffusion of true rumors).

Figure 6 shows the predicted marginal mean effect of the sentiment variable on the DVs. For each DV, we find relatively large effect sizes for the sentiment variable that significantly differ between true vs. false rumors. All else being equal, false rumors have cascades that are of larger size, longer duration, and greater virality if the sentiment is positive. Hence, a (positive) sentiment in the language of rumors explains the pronounced diffusion of false rumors.

Our regression model controls for heterogeneity in users' social influence (see Supplementary Table S1). In short, rumor cascades initiated from accounts that are verified and younger are linked to a larger, longer, and more viral spread. Similar relationships are observed for users exhibiting greater numbers of followers and followees. In contrast, a higher engagement level is negatively associated with the size, lifetime, and structural virality of a cascade.

We calculated the pseudo-R^2 for each model, resulting in relatively high values of 0.64 for cascade size, 0.43 for cascade lifetime, and 0.31 for structural virality. Evidently, the model variables explain a large proportion of the DV variations. Furthermore, visual inspection of the actual vs. fitted plot and goodness-of-fit tests indicate that the models are well specified. This is also supported when considering the differences between the AIC models for individual models estimated with/without sentiment variables. For each DV, the difference is greater than 10 (cascade size: 303.43; lifetime: 110.56; structural virality: 170.01), indicating strong support for the corresponding candidate models[43]. Therefore, the inclusion of sentiment variables in the regression model is to be preferred.

Analysis of emotions. Plutchik's emotion model arranges the eight basic emotions into four pairs of bipolar emotions (see "Materials and methods"). We now evaluate how these bipolar emotion pairs are associated with differences in the structural properties of true vs. false rumor cascades (see coefficient estimates in Supplementary Table S2). The reason for using bipolar emotions is the strong linear dependence among the 8 basic emotions. Adding all basic emotions to the same model would make the estimation rank-deficient. As a remedy, we focus on bipolar emotions, which allow for all eight basic emotions to be examined in the same model.

The predicted marginal effects for the bipolar emotion pairs are shown in Fig. 7. Changes in the emotional language dimensions are associated with greater changes in size, lifetime, and structural virality for false rumors vs. true rumors, as evidenced by steeper slopes of the curves. We observe that false rumor cascades containing words associated with anticipation, anger, and trust have a more extensive diffusion than their true counterparts. We find no statistically significant coefficient for language related to joy vs. sadness. In summary, false rumors spread more extensively than true rumors in the presence of emotional language embedding anticipation, anger, and trust, whereas we observe opposite effects, albeit of smaller magnitude, for language connected to surprise, fear, and disgust.

Discussion

Here we analyze to what extent language embedded in online content can explain differences in the spread of true vs. false social media rumors. Specifically, we study two dimensions: (1) sentiment and (2) basic emotions. Our results establish that both are important determinants of the different spread of true vs. false rumors. For sentiment, we find that positive language is associated with a wider, longer, and more viral spread for false rumors. For basic emotions, we find that language characterized as anger, anticipation, and trust is associated with a wider, longer, and more viral spread for false rumors.

Our research is based on the following rationale as to why sentiment (and emotions) should have the ability to influence the spread of true vs. false rumors. Sentiment (and emotions) are highly relevant for diffusion of online content[24,26,34–36,44,45]. For instance, prior research has studied the role of sentiment in the diffusion of online rumors during crisis[38]. Similarly, online rumors are characterized by a distinctive set of emotions[17]. Hence, this motivated our research to examine whether sentiment (and emotions) are determinants for the distinct spread

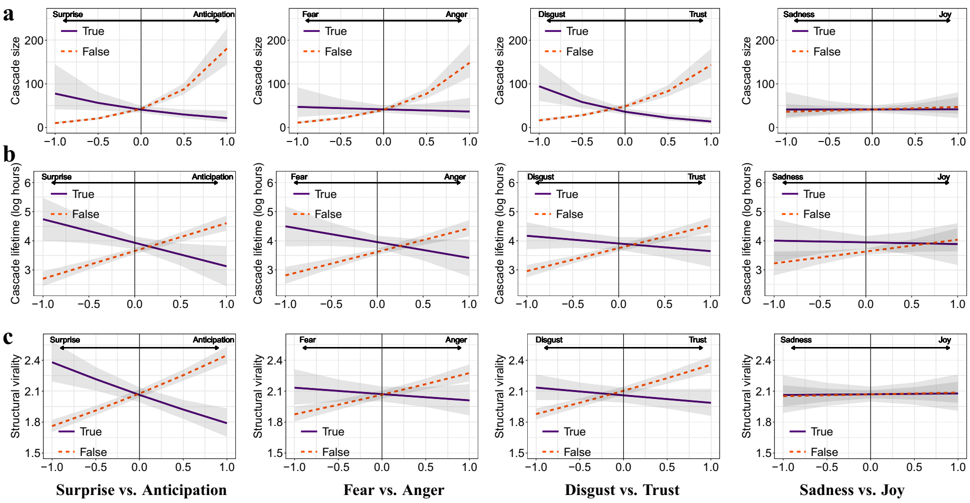

Figure 7. Predicted marginal effect of language classified by bipolar emotions on cascade size (**a**), cascade lifetime (**b**), and structural virality (**c**). The 95% confidence intervals are highlighted in gray.

of true vs. false rumors. Different from previous works, we demonstrate that language in the form of sentiment and emotions can explain the unique structural properties of false rumors.

In our research, we studied the role of different discrete emotions (e. g., anger) in promoting the spread of true vs. false rumors. This choice was made for two reasons. First, discrete emotions are commonly used in affective computing. Specifically, we build upon the NRC emotion lexicon which provides a prominent and comprehensive dictionary for examining discrete emotions[46]. This choice renders our analysis comparable to other research. Second, and more importantly, discrete emotions such as anger have been identified as being relevant for offline rumors[47,48] and online rumors[17,18,37]. Because of this, our analysis also involves discrete emotions. Future research could expand our work and follow a physiological constructionist perceptive as an alternative emotion model (where emotions form a 2×2 dimensional space around valence-arousal).

This study is subject to the typical limitations inherent in observational inferences. First, we report associations and refrain from making causal claims. Other studies[18] argue that estimates should resemble those from causal inferences due to the temporal nature whereby the tweet precedes the cascade formation. Second, our inferences are limited by the accuracy and availability of fact-checking labels. Possible selection biases might arise from the preferences and processes of the used fact-checking websites (e. g., partisan biases). Reassuringly, the fact-checking websites reveal high pairwise agreement[17]. Third, our objective was to compare true vs. false rumors. Future research might further investigate rumors that cannot be clearly attributed to one of the two fact-checking labels. Fourth, our dictionary approach does not allow us to infer the physiological state of users and whether certain emotions are inspired. Instead, our dictionary approach quantifies the use of language in text. Thus, it is possible that even if rumors embed words associated with positive language, they may still elicit negative emotions in readers. More research is necessary to understand the relationship between expression and elicitation of emotions in online rumors, i. e., author vs. receiver effects[49]. Fifth, our study builds upon Plutchik's emotion model and does not account for the the extent of emotionality in rumor cascades, i. e., the extent to which emotional words are present at all. Future research might complement our analysis, by distinguishing the roles of total emotionality and emotional valence in rumor diffusion. Sixth, we follow earlier research and quantify online diffusion by extracting the size, lifetime, and structural virality of cascade. Therefore, our unit of analysis is at the cascade level, which is consistent with earlier research[37,39,50–54]. As such, we expect interesting research opportunities by studying the within-cascade diffusion dynamics.

Policy initiatives around the world require social media platforms to limit the spread of false rumors[9]. To detect them early, our findings emphasize the importance of considering variations in positive and negative words as well as emotional language. In machine learning predictions, sentiment and emotions have been employed in comparatively few works[11–14,55], despite the fact that sentiment and emotions promise benefits in platform-wide settings: they are likely to be more robust against manipulation than other predictors (e. g., content features, for which predictive power is limited if an unseen topic or keyword is encountered). Sentiment and emotions are also available in the early stages of the diffusion, at which point features from the propagation dynamics are scarce (cf. the discussion in[56]). By managing sentiment and emotions in social media content, platforms might develop an effective strategy for reducing the proliferation of false rumors.

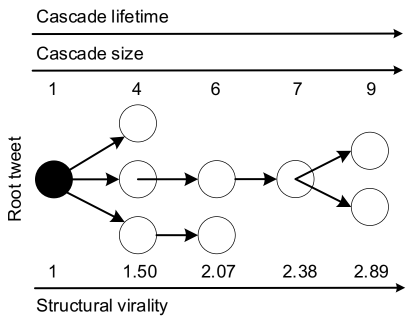

Figure 8. Example tree structure of a rumor cascade.

Materials and methods

Data collection. We analyze a comprehensive dataset with rumor cascades from Twitter[17]. In particular, we examine a sample of English-language cascades on Twitter from its founding in 2006 through 2017. To this end, rumors were matched against established fact-checking websites (see below). Permission to process this dataset for the purpose of our study was granted by Twitter. This ensures a real-world, large-scale sample. Each rumor in our sample involves one or more *rumor cascades*. A rumor has more than one rumor cascade if it exhibits multiple independent retweet chains started by different users but pertaining to the same story/claim. In sum, our data contains $N = 126{,}301$ rumor cascades corresponding to 2448 rumors. The rumors were retweeted more than 4.5 million times by around 3 million different users. The rumors in the dataset cover varying topics (e. g., Politics, Business, Natural Disasters), while the largest proportion of rumors are political rumors[17].

As per terminology, we adopt the definition of *rumors* used in[17]. In this work, rumors refer to content that can be identified as true or false through fact-checking. This definition emerged in the 1940's in social psychology literature[57,58], formalizing it as a proposition involving person-to-person propagation but without necessarily being truthful, such that fact-checking can determine the underlying veracity.

Twitter was selected for this study for the following reasons. First, Twitter represents a social media platform with tremendous popularity[59]. In 2019, it counted ~ 330 million active users[60]. Second, Twitter is extensively used for news consumption. Twitter is consulted for information on political matters by one in ten U.S. adults[61]. Third, Twitter is regarded as highly influential in the public discourse, especially concerning political matters[5], in which deceptive content poses a threat to the functioning of societies.

Our dataset further contains information regarding the retweet path of each rumor cascade, i. e. the temporal propagation dynamics of a rumor cascade on Twitter. Figure 8 shows an exemplary tree structure of a rumor cascade. The root node is the original tweet containing a rumor, whereas the children are retweets of the original tweet and all other nodes are retweets of retweets of the original tweet. We use the retweet path to calculate structural characteristics of each rumor cascade, namely the size (the number of users involved in a cascade), lifetime (the time difference between the root tweet and the terminal tweet), and structural virality.

IRB approval was received from ETH Zurich (2020-N-44). The above data collection results in a large-scale dataset on online rumors.

Fact-checking. Our data sample comprises a comprehensive set of Twitter cascades that were subject to fact-checking based on at least one of six independent organizations: http://factcheck.org, http://hoax-slayer.com, http://politifact.com, http://snopes.com, http://truthorfiction.com, and http://urbanlegends.about.com. Fact-checking returns labels that denote the veracity of the content according to three categories: true, false, or mixed. Fact-checking websites show high pairwise agreement[17], ranging between 95 and 98%. True and false labels are even completely disjunct.

In our data, the frequencies of fact-checking labels at cascade level are: 24,409 (= true) and 82,605 (= false). For 19,287 rumors, no clear assignment to true or positive was possible; these rumors were discarded in our analysis as we aim at comparing true vs. false rumors. Examples of analyzed rumors are given in Table 1.

Calculation of scores for sentiment and emotions. Scores for sentiment and emotions were computed based on affective computing[62]. Here we use (1) sentiment giving the overall valence across positivity and negativity and (2) eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The basic emotions are defined in Plutchik's wheel of emotions[40]; see Fig. 9. Basic emotions are rooted in human evolution and are thus stable across ethnic or cultural differences[41,42]. Furthermore, according to emotion theory, basic emotions represent a small subset of core emotion based on which other more complex emotions are derived. As shown in the Plutchik's wheel of emotions, basic emotions exhibit a bipolar categorization, where each emotion has a corresponding opposite emotion.

The underlying computation of the emotion scores followed the procedure from[17]. For all rumor cascades j of rumor i, the scores were determined based on the NRC emotion lexicon[46] that contains a comprehensive list of

Rumor	Label
"From 2010 to 2014, median household income has actually gone up 7.4%."	FALSE
"Increasing the min. wage to $15 an hour would reduce spending on food stamps, public housing and other programs by over $7.6 billion a year."	FALSE
"Thanks to #ObamaCare, average E.R. wait in California is 5 hours: http://gop.cm/6015YqKd *And "it's only going to get worse.""*	FALSE
"California Gov says yes to poisoning more children with mercury and aluminum in manditory vaccines. This corporate fascist must be stopped."	FALSE
"It's the longest running congressional investigation ever. It's cost taxpayers $4 million. And what's it about?"	FALSE

Table 1. Examples of rumors posted on Twitter. Fact-checking labels from http://politifact.com. Fact-checking labels from the other websites show high pairwise agreement, with true and false labels being completely disjunct[17].

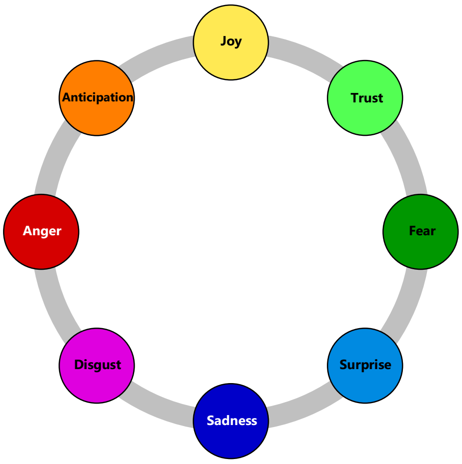

Figure 9. Plutchik's wheel of emotions[40].

141,820 English words and their associations with each of the eight basic emotions. Each reply to the root tweet of a rumor cascade was cleaned, tokenized, and then compared against the NRC emotion lexicon. To illustrate examples of emotional words, Table 2 categorizes a set of online rumors across the eight basic emotions using the NRC emotion lexicon.

Sentiment. We calculate a sentiment score s_{ij} that only measures the extent of positive/negative polarity in replies to rumor cascades. Based on Plutchik's wheel of emotions, we compute the word count of all positive words, denoted by $Positivity_{ij}$, and the word count of all negative words, denoted by $Negativity_{ij}$, respectively. Both scores were normalized so that they add to one, and thus measure the relative extent to which language leans toward a positive or negative polarity. The sentiment score s_{ij} is then defined as the difference between positivity and negativity, i. e., $s_{ij} = Positivity_{ij} - Negativity_{ij}$.

Bipolar emotion pairs. We start by computing the fraction of words in the reply tweets that relate to each of the eight emotions. These were then aggregated and averaged to create a vector of emotion weights that sum to one across the emotions. The eight emotion dimensions in e_{ij} thus range from zero to one, while most rumor cascades exhibit multiple emotions. For instance, emotion scores in replies to rumor cascades can be 70% surprise and 30% fear.

We calculate a 4-dimensional score b_{ij} for the bipolar emotion pairs in Plutchik's wheel of emotions, one for each of the four axes: "anticipation–surprise", "anger–fear", "trust–disgust", and "joy–sadness". Each of the four bipolar emotion pairs thus measures the difference between an emotion (e. g., joy) and its complement at the opposite side of the wheel (e. g., sadness). We use bipolar emotions due to the strong linear dependence among the eight basic emotions. Adding all basic emotions to the same model would make the estimation rank-deficient. Therefore, we focus on bipolar emotions as these allow for all basic emotions to be examined in the same model.

Emotion	Online rumor
Anger	*"Reports That IS Leader Abu Bakr Al-Baghdadi Was Wounded In A Coalition Air* ***Strike*** *Are "Unconfirmed" - Us Defense Sec John Kirby #R4today"*
	"Y'all, I Just Read That ABC Paid Darren Wilson $500k For The Interview. ***Destroying*** *Black Life Remains A Lucrative American Career. #Ferguson"*
Fear	*"Reports Of* ***Shooting*** *At Dammartin En Goele On Route N2 North East Of Paris - French Media Says Car Chase Under Way"*
	"Sydney ***Hostage*** *- Taker - Man Monis, 49 - Originally From Iran - Self-Styled Sheikh -****Accused*** *Of Sexually Assaulting 7 Women Developing.."*
Anticipation	*"Obama's Daughter Is Pregnant LOL Michelle Should've Spent More* ***Time*** *With Her Instead Of Taking Away Our French Fries"*
	"Obama Has Filed Federal Charges Against Zimmerman For Violating Trayvon's Civil Rights. ***God*** *Is* ***Good****"*
Trust	*"Darren Wilson Is A Six Year* ***Veteran*** *Of The #Ferguson* ***Police*** *And Had No Disciplinary Actions Against Him."*
	"Canadian Authorities Have Given Name Of Suspect In Ottawa Attacks To U.S. Feds; Ask For FBI Assistance: Sr U.S. ***Law*** *Enforcement* ***Official****"*
Surprise	*"Breaking: 15.1 Foot Tsunami Reported In Coquimbo. #****Earthquake*** *#Tsunami"*
	"Walmart Donates $10,000 To Support Darren Wilson, But Won't Give Tracy Morgan A Penny For The ***Accident*** *Their Company Caused. #Boycottwalmart"*
Sadness	*"What's So Frustrating Is That Now We Are Talking About A* ***Robbery*** *And Not The* ***Killing*** *Of An Unarmed Kid. #Ferguson"*
	"Conservative Caucas Informed ***Soldier Shot*** *At War Memorial In Ottawa This Morning Has Died. A Sad Development On A Shocking Day."*
Joy	*"Disney Are Making Their First Movie To Feature Two Gay Princes Who Fall In* ***Love****, Amazing."*
	"Paula Deen: "Forget The ***Food*** *Network. I've Already Been Offered A Cooking Show On The New Fox News* ***Food*** *Channel.""*
Disgust	*"#Psa Please Do Not Drink Any Pepsi Soda, A Worker From That Company Has Put Blood* ***Contaminated*** *With Aids Inside The Bottles!!! Please Rt!!"*
	"People ***Blame*** *The* ***Massacre*** *In Orlando On The NRA. Newsflash: The Orlando Shooter Wasn't A NRA Member... But He Was A Registered Democrat."*

Table 2. Examples for rumors posted on Twitter and the emotional words they contain. The emotional words are classified according to the NRC emotion lexicon using eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. Emotional word corresponding to the basic emotion in column 1.

In Plutchik's emotion model, emotion scores sum up to one across the basic emotions. We thus omit 149 rumor cascades that do not contain any emotional words from the NRC emotion lexicon (since, otherwise, the denominator is not defined).

Validation of dictionary approach. Our results rely on the validity of dictionaries to extract sentiment and emotions from online rumors. We thus checked how the perceived sentiment and emotions in rumors align with the lexicon-based sentiment score and emotion scores. For this, we conducted two user studies (see Supplementary Section A), where participants were asked to rate the perceived sentiment, as well as the perceived emotions, in a given rumor. In both studies, the participants exhibited a statistically significant interrater agreement (using Kendall's W). Importantly, we found Spearman's correlation coefficients for the human labels and the dictionary-based scores to be positive and statistically significant; both for sentiment ($r_s = 0.11$, $p < 0.01$) and emotions ($r_s = 0.13$, $p < 0.01$). In sum, the results add to the validity of our lexicon-based approach. The lexicon-based approach should thus capture the perceived sentiment, as well as the perceived emotions, in online rumors.

Variable description. A rumor cascade $j = 1, \ldots, N$ belonging to rumor i is given by a tree structure $T_{ij} = (r_{ij}, t_{ij0}, R_{ij})$ with root tweet r_{ij}, the root node's timestamp t_{ij0}, and a set of retweets $R_{ij} = \{(p_{ijk}, t_{ijk})\}_k$, where each retweet is a 2-tuple comprising a parent p_{ijk} and a timestamp t_{ijk}. The root denotes the original sender of the tweet.

Cascade structure. Based on the tree structure T_{ij}, we compute the following variables y_{ij} characterizing the underlying diffusion dynamics (Fig. 8):

- *Size* The size refers to the overall number of retweets in the cascade, that is, $|R_{ij}| + 1$. Hence, it measures how many users interacted with a tweet.
- *Lifetime* This is the overall timespan during which the tweet travels through the network, defined as $\max\{t_{ijk}\}_k - t_{ij0}$.
- *Structural virality*[39] This metric measures the trade-off between a cascade that stems from a single retweet and a cascade that has a chain structure, thus quantifying how frequently and how extensively a message is retweeted. Formally, it is defined as the average "distance" between all pairs of retweeters[39], i. e., $v(T_{ij}) = \frac{1}{n\,(n-1)} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} d_{ij_1,ij_2}$ for a cascade T_{ij} with n nodes and where d_{ij_1,ij_2} is the shortest path between nodes ij_1 and ij_2 (similar to the Wiener index).

Social influence. Following earlier research[17,24,63], the social influence of the root r_{ij} is quantified by the following covariates x_{ij}:

- *Account age* The age of the root's account (in years).
- *Out-degree* The number of followers, i. e., the number of accounts that follow the user (in 1000s).
- *In-degree* The number of followees, i. e., the number of accounts whom the user follows (in 1000s).
- *User engagement* For the sender, past engagement is measured by the past number of interactions on Twitter (i. e., tweets, shares, replies, and likes) relative to the account age[17]. Formally, it computes to $(T + R + P + L)/A$ given the past volume of tweets T, retweets R, replies P, and likes L divided by the root's account age A (in days).
- *Verified account* A binary dummy indicating whether the account of the root has been officially verified by Twitter ($= 1$; otherwise $= 0$). This is shown by a blue badge that is reserved for users of public interest (e. g., celebrities, politicians).

All of the above variables are computed at the level of cascades as our unit of analysis. Time is not explicitly included but later captured in the random effects (we also performed a separate analysis with time effects as part of our robustness checks).

Research framework. In this work, our objective is to attribute differences in the structural properties of true vs. false rumors to positive and negative language as well as words associated with certain emotions. For this purpose, we link the structural properties to the sentiment and emotions conveyed by the language in the replies to rumor cascades. Specifically, we address the following questions: (1) How are variations in language characterized by positive and negative sentiment associated with differences in the structural properties of true vs. false rumor cascades? (2) How does the presence words conveying certain emotions (e. g., anger, trust) explain differences in the structural properties of true vs. false rumor cascades?

Our research questions aim to explain why false rumors (as compared to true rumors) have a longer lifetime, a larger size, and higher structural virality. As defined before, sentiment is a one-dimensional measure along with positive and negative polarity, while emotions refer to a granular, bipolar assessment of arousal along multiple dimensions. In answering the above research questions, we are interested in the marginal effects (that is, by controlling for other sources of heterogeneity).

Model specification. We specify regression models that explain the cascade structure based on positive and negative language as well as emotional words, while also accounting for further sources of heterogeneity. Recall that the cascade structure (i. e., the lifetime, size, and structural virality) is given by y_{ij}. Furthermore, let ϕ_i denote the veracity of rumor i. Here we define a true rumor as $\phi_i = 0$ and a false rumor as $\phi_i = 1$. Rumors of mixed veracity are included later as part of the robustness checks.

Controls. In order to estimate marginal effects, we include several control variables. The control variables are: the social influence of the root x_{ij} (as cascades are likely to diffuse more extensively from influential users) and the veracity ϕ_i. The latter measures, all else being equal, the relative contribution of veracity to a rumor going viral. In addition, we control for heterogeneity among rumors by using rumor-specific random effects. The latter is important as it accounts for other unobserved factors (e. g., rumor topic, links to external websites, posting date) that may influence the spreading dynamics.

Regression. Based on the above, we yield the following hierarchical generalized linear model for our analysis of language classified by positive and negative sentiment:

$$y_{ij} = \beta_0 + \beta_1^T x_{ij} + \beta_2 \phi_i + \beta_3 s_{ij} + \beta_4 (\phi_i \times s_{ij}) + u_i \tag{1}$$

with intercept β_0, rumor-specific random effects u_i, and coefficients $\beta_1, \ldots, \beta_4$ (out of which β_1 is a vector). Here the dependent variable is given by y_{ij} (i. e., lifetime, size, or structural virality). Depending on the actual choice of the dependent variable, a different distribution is modeled and, hence, a different estimator must be used. This is detailed later. The notation $(\phi_i \times s_{ij})$ refers to a one-way interaction term.

For our analysis of emotional language, a hierarchical generalized linear model is analogously obtained whereby the sentiment variable s_{ij} is replaced by the bipolar emotions pairs $b_{ij} \in \mathbb{R}^4$, i. e.,

$$y_{ij} = \beta_0 + \beta_1^T x_{ij} + \beta_2 \phi_i + \beta_3^T b_{ij} + \beta_4^T (\phi_i \odot b_{ij}) + u_i \tag{2}$$

with parameters $\beta_0, \ldots, \beta_4$ (out of which β_1, β_3, and β_4 are vectors and where $\odot$ is the element-wise multiplication).

Model coefficients. The estimation results for the parameters $\beta_0, \ldots, \beta_4$ characterize the spread of true vs. false rumors as follows:

- β_1 is the intercept. It represents the baseline structure for a cascade with average properties.
- β_2 assesses the overall contribution of veracity to diffusion dynamics (after correcting for different emotions and social influence as in true vs. false rumors). Hence, all else being equal, this parameter quantifies to what extent false rumors last longer, spread more widely, and go more viral as compared to true rumors.

- β_3 measures how tweets with emotional language link to cascade structures. Estimation results for this coefficient have been discussed elsewhere[18,24,26] and, for reasons of brevity, are thus omitted from our results section. We note that the influence directly attributed to emotional language is consistent with previous research.

Of particular interest is the following parameter:

- β_4 estimates the relative differences in how emotional language is received in relation to true vs. false rumors. This is captured by the one-way interaction between the emotion variables and veracity. Hence, a positive β_4 indicates that an increase in the fraction of emotional words of a certain category is associated with larger increases of the dependent variable for false vs. true rumors. As we controlled for other sources of heterogeneity, these estimates are "ceteris paribus," that is, all else being equal, they measure how much larger/smaller is the effect of language classified by emotions on size, lifetime, and structural virality if the rumor is false.

Estimation details. The actual choice of the dependent variable requires a different estimator in order to account for the underlying distribution. Cascade size represents count data and its variance is larger than its mean. We thus adjust for overdispersion and apply negative binomial regression with log-transformation. For lifetime, prior research has suggested that response times are log-normally distributed[63]. Accordingly, we log-transform the lifetimes. Results of the Shapiro-Wilk test for normality applied to the log-transformed variable suggest that the null hypothesis of normal distribution cannot be rejected. This allows us to estimate the model using ordinary least squares (OLS). For structural virality, we use a gamma regression with log-link, which is a common choice for modeling positively skewed, non-negative continuous variables.

Our implementation uses the lme4 package in R 3.6.3. This ensures that random effects are considered. Approaches for winsorizing or censoring the data (or other filtering options) were intentionally disregarded, as we consider all observations to be informative, especially those in the tails. We nevertheless performed a robustness check with winsorizing, yielding consistent outcomes. We z-standardized all variables in order to facilitate interpretability. Accordingly, the regression coefficients measure the relationship with the dependent variable in standard deviations.

Robustness checks. We conducted the following checks to validate the robustness of our results.

Fine-grained emotions. Instead of having four bipolar dimensions, we ran a regression with all eight fine-grained emotions as separate variables (see Supplementary Tables S3–S5). Consistent with our previous findings, we again find that words associated with emotions like anticipation, trust, and anger accelerate the spread of false rumors. However, the estimation is rank-deficient and, hence, our main analysis is instead based on bipolar emotion pairs.

Additional checks. We conducted additional checks to validate the robustness of our results: (1) we ran separate regressions for true vs. false rumors. (2) to ensure robustness across the complete time period of the study, we used clustered standard errors at the annual level and repeated the analysis for different time periods. Furthermore, we included dummy variables for each year in our sample to control for year-level effects. The results show a good agreement of the coefficients of all variables and support the robustness of our results across time periods (see Supplementary Section B). (3) The validity of our estimates was ensured by following common practice for regression modeling. In particular, we determined the variance inflation factor (VIF) to be below the critical threshold of five[64]. (4) We added non-linear regressors (i. e., quadratic terms) for each emotion to our regression models. In all cases, our results are robust consistently support our findings. (5) We analyzed how the diversity of emotion scores is association with the spread of rumors. Here we extended our regression models with a variable that measures the sum of squares over the 8-dimensional vector comprising the different emotion scores. We find that a higher diversity of emotion scores is associated with higher values for cascade size, duration, and structural virality (see Supplementary Section B).

Data availability

Permission from Twitter to analyze the dataset was obtained. All data needed to evaluate the conclusions in the paper are publicly available (and the source reported in the paper). Replication code for this study is available via https://github.com/DominikBaer95/Emotions_true_vs_false_online_rumors.

Received: 23 June 2021; Accepted: 3 November 2021
Published online: 22 November 2021

References

1. Starbird, K., Maddock, J., Orand, M., Achterman, P. & Mason, R. M. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. in *iConference* (2014).
2. Aral, S. & Eckles, D. Protecting elections from social media manipulation. *Science* **365**, 858–861. https://doi.org/10.1126/science.aaw8243 (2019).
3. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132. https://doi.org/10.1126/science.aaa1160 (2015).
4. Bovet, A. & Makse, H. A. Influence of fake news in twitter during the 2016 us presidential election. *Nat. Commun.* **10**, 7. https://doi.org/10.1038/s41467-018-07761-2 (2019).

5. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378. https://doi.org/10.1126/science.aau2706 (2019).
6. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236. https://doi.org/10.1257/jep.31.2.211 (2017).
7. Economist, The. How the world was trolled. *Economist* **425**, 21–24 (2017).
8. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539. https://doi.org/10.1126/sciadv.aay3539 (2020).
9. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096. https://doi.org/10.1126/science.aao2998 (2018).
10. Shao, C. *et al.* The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 4787. https://doi.org/10.1038/s41467-018-06930-7 (2018).
11. Castillo, C., Mendoza, M. & Poblete, B. Information credibility on Twitter. in *International World Wide Web Conference (WWW)*. https://doi.org/10.1145/1963405.1963500 (2011).
12. Kwon, S., Cha, M., Jung, K., Chen, W. & Wang, Y. Prominent features of rumor propagation in online social media. in *International Conference on Data Mining (ICDM)*. https://doi.org/10.1109/ICDM.2013.61 (2013).
13. Kwon, S., Cha, M. & Jung, K. Rumor detection over varying time windows. *PLOS ONE* **12**, e0168344. https://doi.org/10.1371/journal.pone.0168344 (2017).
14. Ducci, F., Kraus, M. & Feuerriegel, S. Cascade-LSTM: A tree-structured neural classifier for detecting misinformation cascades. in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (2020).
15. Pennycook, G., Cannon, T. D. & Rand, D. G. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. General* **147**, 1865–1880. https://doi.org/10.1037/xge0000465 (2018).
16. Friggeri, A., Adamic, L. A., Eckles, D. & Cheng, J. Rumor cascades. in *International AAAI Conference on Web and Social Media (ICWSM)* (2014).
17. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151. https://doi.org/10.1126/science.aap9559 (2018).
18. Chuai, Y. & Zhao, J. Anger makes fake news viral online. arXiv (2020).
19. Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. Emotion and decision making. *Annu. Rev. Psychol.* **66**, 799–823. https://doi.org/10.1146/annurev-psych-010213-115043 (2015).
20. Fox, E., Russo, R., Bowles, R. & Dutton, K. Do threatening stimuli draw or hold visual attention in subclinical anxiety?. *J. Exp. Psychol. General* **130**, 681–700 (2001).
21. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci. (PNAS)* **114**, 7313–7318. https://doi.org/10.1073/pnas.1618923114 (2017).
22. Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. (PNAS)* **111**, 8788–8790. https://doi.org/10.1073/pnas.1320040111 (2014).
23. Goldenberg, A. & Gross, J. J. Digital emotion contagion. *Trends Cognit. Sci.* **24**, 316–328. https://doi.org/10.1016/j.tics.2020.01.009 (2020).
24. Stieglitz, S. & Dang-Xuan, L. Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *J. Manag. Inf. Syst.* **29**, 217–248. https://doi.org/10.2753/MIS0742-1222290408 (2013).
25. Berger, J. Arousal increases social transmission of information. *Psychol. Sci.* **22**, 891–893. https://doi.org/10.1177/0956797611413294 (2011).
26. Berger, J. & Milkman, K. L. What makes online content viral?. *J. Market. Res.* **49**, 192–205. https://doi.org/10.1509/jmr.10.0353 (2012).
27. Kissler, J., Herbert, C., Peyk, P. & Junghofer, M. Buzzwords: Early cortical responses to emotional words during reading. *Psychol. Sci.* **18**, 475–480. https://doi.org/10.1111/j.1467-9280.2007.01924.x (2007).
28. Luminet, O., Bouts, P., Delie, F., Manstead, A. S. R. & Rimé, B. Social sharing of emotion following exposure to a negatively valenced situation. *Cognit. Emot.* **14**, 661–688. https://doi.org/10.1080/02699930050117666 (2000).
29. Rimé, B. Emotion elicits the social sharing of emotion: Theory and empirical review. *Emot. Rev.* **1**, 60–85. https://doi.org/10.1177/1754073908097189 (2009).
30. Peters, K., Kashima, Y. & Clark, A. Talking about others: Emotionality and the dissemination of social information. *Eur. J. Soc. Psychol.* **39**, 207–222. https://doi.org/10.1002/ejsp.523 (2009).
31. Martel, C., Pennycook, G. & Rand, D. G. Reliance on emotion promotes belief in fake news. *Cognit. Res. Principles Implications.* **5**, Article 47. https://doi.org/10.1186/s41235-020-00252-3 (2020).
32. Naveed, N., Gottron, T., Kunegis, J. & Alhadi, A. C. Bad news travel fast: A content-based analysis of interestingness on Twitter. in *International Web Science Conference (WebSci)*. https://doi.org/10.1145/2527031.2527052 (2011).
33. Kim, J. & Yoo, J. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. in *International Conference on Social Informatics*. https://doi.org/10.1109/SocialInformatics.2012.33 (2012).
34. Heimbach, I. & Hinz, O. The impact of content sentiment and emotionality on content virality. *Int. J. Res. Market.* **33**, 695–701. https://doi.org/10.1016/j.ijresmar.2016.02.004 (2016).
35. Meng, J. *et al.* Diffusion size and structural virality: The effects of message and network features on spreading health information on twitter. *Comput. Hum. Behav.* **89**, 111–120. https://doi.org/10.1016/j.chb.2018.07.039 (2018).
36. Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J. Everyone's an influencer. in *International Conference on Web Search and Data Mining (WSDM)*. https://doi.org/10.1145/1935826.1935845 (2011).
37. Pröllochs, N., Bär, D. & Feuerriegel, S. Emotions in online rumor diffusion. *EPJ Data Sci.* **10**, Article 51. https://doi.org/10.1140/epjds/s13688-021-00307-5 (2021).
38. Zeng, L., Starbird, K. & Spiro, E. S. Rumors at the speed of light? Modeling the rate of rumor transmission during crisis. in *Hawaii International Conference on System Sciences (HICSS)*. https://doi.org/10.1109/HICSS.2016.248 (2016).
39. Goel, S., Anderson, A., Hofman, J. & Watts, D. J. The structural virality of online diffusion. *Manag. Sci.* **62**, 180–196. https://doi.org/10.1287/mnsc.2015.2158 (2016).
40. Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **89**, 344–350 (2001).
41. Ekman, P. An argument for basic emotions. *Cognit. Emot.* **6**, 169–200. https://doi.org/10.1080/02699939208411068 (1992).
42. Sauter, D. A., Eisner, F., Ekman, P. & Scott, S. K. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl. Acad. Sci. (PNAS)* **107**, 2408–2412. https://doi.org/10.1073/pnas.0908239106 (2010).
43. Burnham, K. P. & Anderson, D. R. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304 (2004).
44. Tsugawa, S. & Ohsaki, H. On the relation between message sentiment and its virality on social media. *Social Netw. Anal. Mining.*https://doi.org/10.1007/s13278-017-0439-0 *(2017).*
45. Tellis, G. J., MacInnis, D. J., Tirunillai, S. & Zhang, Y. What drives virality (sharing) of online digital content? the critical role of information, emotion, and brand prominence. *J. Market.* **83**, 1–20. https://doi.org/10.1177/0022242919841034 (2019).
46. Mohammad, S. M. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. *Emot. Meas. (Second Edition)*https://doi.org/10.1016/B978-0-12-821124-3.00011-9 *(2021).*
47. Anthony, S. Anxiety and rumor. *J. Social Psychol.* **89**, 91–98. https://doi.org/10.1080/00224545.1973.9922572 (1973).

48. Rosnow, R. L. Inside rumor: A personal journey. *Am. Psychol.* **46**, 484–496 (1991).
49. Kato, Y., Kato, S. & Akahori, K. Effects of emotional cues transmitted in e-mail communication on the emotions experienced by senders and receivers. *Comput. Hum. Behav.* **23**, 1894–1905 (2007).
50. Goel, S., Watts, D. J. & Goldstein, D. G. The structure of online diffusion networks. in *ACM Conference on Electronic Commerce (EC)*. https://doi.org/10.1145/2229012.2229058 (2012).
51. Leskovec, J., Adamic, L. A. & Huberman, B. A. The dynamics of viral marketing. *ACM Trans. Web.* **1**, Article 5. https://doi.org/10.1145/1232722.1232727 (2007).
52. Myers, S. A. & Leskovec, J. The bursty dynamics of the twitter information network. in *International World Wide Web Conference (WWW)*. https://doi.org/10.1145/2566486.2568043 (2014).
53. Taxidou, I. & Fischer, P. M. Online analysis of information diffusion in Twitter. in *International Conference on World Wide Web (WWW) Companion*. https://doi.org/10.1145/2567948.2580050 (2014).
54. Zang, C., Cui, P., Song, C., Faloutsos, C. & Zhu, W. Quantifying structural patterns of information cascades. in *International Conference on World Wide Web (WWW) Companion*. https://doi.org/10.1145/3041021.3054214 (2017).
55. Wu, S., Tan, C., Kleinberg, J. & Macy, M. Does bad news go away faster? in *International AAAI Conference on Web and Social Media (ICWSM)* (2011).
56. Conti, M., Lain, D., Lazzeretti, R., Lovisotto, G. & Quattrociocchi, W. It's always april fools' day! on the difficulty of social network misinformation classification via propagation features. in *IEEE Workshop on Information Forensics and Security (WIFS)*. https://doi.org/10.1109/WIFS.2017.8267653 (2017).
57. Allport, G. W. & Postman, L. *The Psychology of Rumor* (Henry Holt, New York, NJ, 1947).
58. Knapp, R. H. A psychology of rumor. *Public Opin. Quart.* **8**, 22–37 (1944).
59. Scharkow, M., Mangold, F., Stier, S. & Breuer, J. How social network sites and other online intermediaries increase exposure to news. *Proc. Natl. Acad. Sci. (PNAS)* **117**, 2761–2763. https://doi.org/10.1073/pnas.1918279117 (2020).
60. Statista. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (2020).
61. Pew Research Center. News use across social media platforms 2016 (2016).
62. Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S. & Prendinger, H. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decis. Supp. Syst.* **115**, 24–35. https://doi.org/10.1016/j.dss.2018.09.002 (2018).
63. Zaman, T., Fox, E. B. & Bradlow, E. T. A Bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.* **8**, 1583–1611. https://doi.org/10.1214/14-AOAS741 (2014).
64. Akinwande, M. O. *et al.* Variance inflation factor: As a condition for the inclusion of suppressor variable (s) in regression analysis. *Open J. Stat.* **5**, 754–767 (2015).

Author contributions

N.P. and S.F. designed the study. N.P. and D.B. analyzed the data. N.P., D.B., and S.F. wrote and revised the manuscript. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-01813-2 .

Correspondence and requests for materials should be addressed to N.P.

Part II: Auditing social media platforms

10 Systematic discrepancies in the delivery of political ads on Facebook and Instagram

Title: Systematic discrepancies in the delivery of political ads on Facebook and Instagram

Abstract: Political advertising on social media has become a central element in election campaigns. However, granular information about political advertising on social media was previously unavailable, thus raising concerns regarding fairness, accountability, and transparency in the electoral process. In this article, we analyze targeted political advertising on social media via a unique, large-scale dataset of over 80,000 political ads from Meta during the 2021 German federal election, with more than 1.1 billion impressions. For each political ad, our dataset records granular information about targeting strategies, spending, and actual impressions. We then study (i) the prevalence of targeted ads across the political spectrum; (ii) the discrepancies between targeted and actual audiences due to algorithmic ad delivery; and (iii) which targeting strategies on social media attain a wide reach at low cost. We find that targeted ads are prevalent across the entire political spectrum. Moreover, there are considerable discrepancies between targeted and actual audiences, and systematic differences in the reach of political ads (in impressions-per-EUR) among parties, where the algorithm favor ads from populists over others.

Author contributions: Dominik Bär, Francesco Pierri, Gianmarco de Francisci Morales, and Stefan Feuerriegel contributed to conceptualization. Dominik Bär and Francesco Pierri contributed to data analysis. Dominik Bär, Francesco Pierri, Gianmarco de Francisci Morales, and Stefan Feuerriegel contributed to results interpretation and manuscript writing.

Reference: Bär, D., Pierri, F., Morales, G. d. F., & Feuerriegel, S. 2024. *Systematic discrepancies in the delivery of political ads on Facebook and Instagram*. In: **PNAS Nexus** 3 (7), pgae247, DOI: `https://doi.org/10.1093/pnasnexus/pgae247`.

Supplementary materials: Supplementary materials for this article are in Supplementary Material C.

PNAS Nexus, 2024, **00**, pgae247
https://doi.org/10.1093/pnasnexus/pgae247
Advance access publication 18 June 2024
Research Report

Systematic discrepancies in the delivery of political ads on Facebook and Instagram

Dominik Bär[a,b,*,1], Francesco Pierri[c,1], Gianmarco De Francisci Morales[d] and Stefan Feuerriegel[a,b]

[a]LMU Munich, Munich 80539, Germany
[b]Munich Center for Machine Learning, Munich 80539, Germany
[c]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan 20133, Italy
[d]CENTAI, Turin 10138, Italy
*To whom correspondence should be addressed: Email: baer@lmu.de
[1]D.B. and F.P. contributed equally to this work.
Edited By: Marc Meredith

Abstract

Political advertising on social media has become a central element in election campaigns. However, granular information about political advertising on social media was previously unavailable, thus raising concerns regarding fairness, accountability, and transparency in the electoral process. In this article, we analyze targeted political advertising on social media via a unique, large-scale dataset of over 80,000 political ads from Meta during the 2021 German federal election, with more than 1.1 billion impressions. For each political ad, our dataset records granular information about targeting strategies, spending, and actual impressions. We then study (i) the prevalence of targeted ads across the political spectrum; (ii) the discrepancies between targeted and actual audiences due to algorithmic ad delivery; and (iii) which targeting strategies on social media attain a wide reach at low cost. We find that targeted ads are prevalent across the entire political spectrum. Moreover, there are considerable discrepancies between targeted and actual audiences, and systematic differences in the reach of political ads (in impressions-per-EUR) among parties, where the algorithm favor ads from populists over others.

Keywords: targeted advertising, social media, algorithmic discrepancies, politics, election campaigns

Significance Statement

Social media platforms have become important tools for political campaigning worldwide. In our study of over 80,000 political ads from the 2021 German federal election, we reveal extensive use of targeted political advertising across the full political spectrum, significant discrepancies between targeted and actual audiences, and a systematic bias in the algorithmic delivery of ads favoring populist parties. These findings highlight the complex relationship between digital campaigning strategies and democratic processes, and caution about the potential for algorithmic biases to influence election campaigns. Overall, our work contributes to a better understanding of targeted political advertising on social media and informs policymakers about the design of effective regulatory frameworks to promote fairness, accountability, and transparency.

Introduction

With around 4.6 billion users globally (1), social media platforms such as Facebook, Instagram, and Twitter/X have become important tools for political campaigning worldwide (2). For example, in the United States, expenditure on online political advertising rose from USD ~70 million in 2014 to USD ~1.8 billion in 2018 (3). Similarly, politicians in Europe have recognized the importance of social media for their campaigns. For instance, the majority of candidates in the 2021 German federal election believed that social media can influence voters (4).

An essential feature of advertising on social media is *targeting*, which allows advertisers to select specific user groups and deliver tailored political messages to particularly receptive audiences (2, 5–7). For example, campaigns can send tailored ads that align with the interests of distinct voter groups (8), thus ensuring that their content resonates with the unique political perspectives of each audience (9). However, targeting in political advertising is problematic (10–12). First, targeting is concerning if parties cater ads to specific groups (13) or send conflicting messages on political issues to different audiences (14). Second, targeted ads are distributed by proprietary algorithms that are beyond societal scrutiny and that may exhibit biases that influences the audience of specific ads (11, 15–17). Third, privacy concerns are eminent given that political targeting heavily depends on potentially sensitive

Competing Interest: The authors declare no competing interest.
Received: March 18, 2024. **Accepted:** June 11, 2024

information (e.g. ethnic origin, sexual orientation) to identify receptive audiences (18–20). Essentially, the use of targeting raises concerns regarding fairness, accountability, and transparency in electoral processes.

The concerns regarding democratic integrity have spurred calls to monitor political advertising on social media (21, 22). However, granular information about political ads (e.g. impression counts, price per ad) is in the hands of proprietary social media platforms and, so far, has been either unavailable or deemed imprecise (23, 24). This lack of transparency may prevent accountability for misconduct, which is particularly concerning given recent evidence suggesting that political advertising on social media directly affects voter turnout (25) and vote choice (26–28). As such, there is a growing need to monitor targeted political advertising on social media to safeguard democratic integrity.

A combination of public pressure (3) and regulatory efforts (e.g. the Honest Ads Act in the United States (29) and the Digital Services Act in the E.U. (30)) have pushed social media platforms to strengthen their transparency efforts around political advertising. Indeed, Meta has launched the Meta Ad Library, which provides public access to all political and social ads published on Facebook and Instagram, and allows researchers to study political advertising at scale (31–35) (see Supplementary Material S1 for a comprehensive overview of the literature). However, existing analyses had only limited access to political ads, since crucial information about *targeting* was missing. As such, it remains unclear how targeting is used, how targeted ads are distributed, what the different targeting strategies behind political ads are, and how their reach at a given budget varies.

There are good reasons to believe that parties adopt diverse targeting strategies (8, 9, 36). Previous research on the content of online political ads has shown that parties pursue different communication strategies (32, 37). Some parties advertise by particularly focusing on issues related to their base (32, 37) such as, for example, environmentalism in the case of the *Grüne*. In contrast, other parties avoid ads related to specific political issues (9, 37) and tend to publish more "generic" ads, for example, introducing a candidate or calling to vote. Since the latter does not refer to a concrete political issue, such ads may reach audiences across party boundaries, which may vary with respect to age, gender, and interests (32). As such, targeting may limit the political participation of disadvantaged groups due to a party's targeting strategy (differently from ads on broadcast media that can be received by all voter groups). Furthermore, there is reason to expect that the algorithmic delivery of social media ads may introduce further bias (15–17), thus resulting in differences between actual and intended audiences. For example, social media algorithms exhibit a tendency to target fewer women due to differences in advertising costs (17). Such algorithmic bias can lead to discrimination as women may be less frequently exposed to political campaigns and thus harm political participation. Moreover, such bias can harm political competition when, for instance, some parties consistently pay higher prices for political ads, thus leading to fairness issues.

In this article, we analyze targeted political advertising on social media using a large-scale dataset with $N = 81{,}549$ political ad contracts (henceforth simply ads)[a] from Meta during the 2021 German federal election (see Supplementary Material S2 for additional context on the election). Overall, these ads generated more than 1.1 billion impressions with an overall cost of EUR 9.8 million. Our dataset provides a unique view of the targeting strategies that parties use across the entire political spectrum, during an election with more than 60 million eligible voters. In particular, our dataset comprises granular information about each ad, including targeting strategies, spending, and actual impressions, and thus allows us to study targeted political advertising on social media.

Our analysis is three-fold: (i) We assess the prevalence of targeted political ads across the full political spectrum and infer detailed targeting strategies used by political parties for their election campaigns. (ii) We evaluate discrepancies between targeted and actual audiences due to algorithmic bias in the ad delivery and how such discrepancies vary across parties. (iii) We analyze the characteristics of targeted ads with far reach at a given budget during elections and analyze whether parties are discriminated by algorithmic ad delivery in that they pay a higher price per impression.

Results

Targeted political advertising during the 2021 German federal election

We analyze targeted political ads on Meta during the 2021 German federal election (see Supplementary Material S2 for additional context on the election) and compare how parties across the political spectrum use targeting for their campaign purposes. Overall, we analyze $N = 81{,}549$ political ads that generated more than 1.1 billion impressions with an overall cost of EUR 9.8 million. For a breakdown of our dataset in terms of total number of ads, ad spending, and impressions by party, see Supplementary Material S3. The ads in our dataset are designed mostly to mobilize and persuade voters and tend to focus on parties more broadly rather than candidates. Details are in Supplementary Materials S4 and S5.

Targeting has spurred concerns regarding political advertising on social media (12, 38). However, it is unclear to what extent parties use targeting during election campaigns. Throughout the article, we consider an ad to employ "targeting" if it uses any targeting category available to advertisers on Meta in addition to demographics (i.e. gender and age) and location (since the latter must always be specified in the ad creation). For details, see Materials and methods section.

To analyze the prevalence of targeted political ads on Meta during the 2021 German federal election, we start by quantifying the number of targeted ads. In the run-up to the election, 72.3 % of all ads used targeting, which corresponds to 72.6 % of the total ad spending on Meta during the election. This highlights the importance of targeted ads for political campaigns on social media.

Meta allows advertisers to target users based on various targeting categories (see Supplementary Material S8 for an overview). We expect that some targeting categories are more popular than others and thus study how the campaign budget is distributed between categories. Figure 1 shows the top-10 targeting categories in terms of spending across all parties. We find that parties tend to use exclusion rather than inclusion criteria to target users. This result suggests that most parties rely on broad audiences and allow Meta to optimize ad delivery among users. To define inclusion criteria, parties largely rely on so-called interests (e.g. social equality, environmentalism, and international relations), behaviors (e.g. early adopters of new technology, commuters, international travelers), or employers (e.g. business owners, police officers, Ford Deutschland). Parties also frequently define a specific list of users to be targeted ("Custom") or an audience that is similar to a previously defined target group ("Lookalike"). Lastly, parties commonly target users based on their location. For example, 88.72 % of the ads have a precise location targeting (beyond Germany). Overall, these results show that parties employ a wide

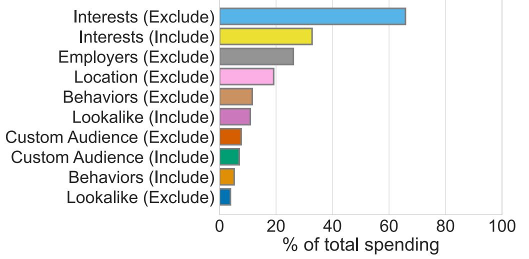

Fig. 1. Top-10 targeting criteria by total spending.

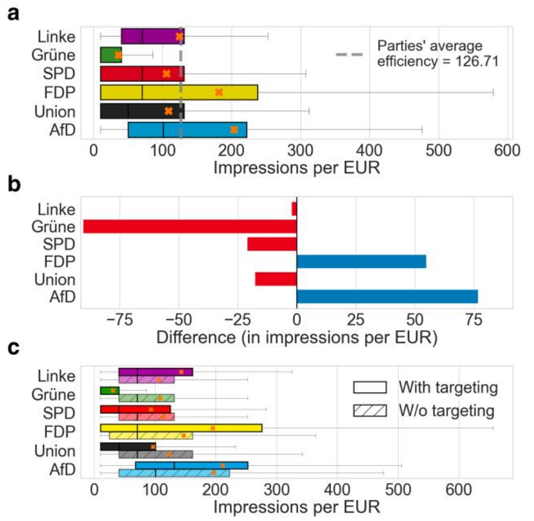

Fig. 2. a) Distributions of impressions-per-EUR across the ads of each party. The cross indicates the mean of the distribution. b) Difference between average impressions-per-EUR of a party and average impressions-per-EUR in the overall sample. c) Distribution of impressions-per-EUR across the ads of each party. The cross indicates the mean of the distribution.

variety of targeting strategies. For a breakdown of the top-10 targeting criteria by party, see Supplementary Material S3.

We now study the reach of political ads at a given budget. To do so, we focus on the number of impressions an ads generates per EUR spent (henceforth simply impressions-per-EUR). Political parties employ different campaign strategies on social media (37), which may also lead to differences in the impressions-per-EUR of political ads. We thus compare impressions-per-EUR across political parties and test for statistically significant differences using a Kruskal–Wallis test (39). The results are in Fig. 2a and b. We find that impressions-per-EUR significantly differs across parties ($P < 0.05$). On average, political ads generate 126.71 impressions-per-EUR. However, the *Grüne* receives, on average, only 36.18 impressions-per-EUR. In contrast, ads published by the *FDP* and the *AfD* are considerably more efficient, reaching, on average, 181.53 and 203.49 impressions per EUR, respectively. In summary, impressions-per-EUR of social media ads varies greatly across the political spectrum.

Given the high prevalence of targeted ads, we also compare impressions-per-EUR for targeted ads and ads without targeting. Our results are mixed (see Fig. 2c). While targeted ads achieve, on average, more impressions-per-EUR compared to ads without targeting for the *Linke*, *FDP*, and *AfD*, the opposite is true for the *Grüne*, *SPD*, and *Union*. Performing multiple pairwise Kruskal–Wallis tests with Benjamini–Hochberg correction based on a family-wise error rate (FWER) of FWER = 0.05, we find that these differences are all statistically significant ($P < 0.05$). Generally, targeting may result in less impressions-per-EUR than ads without targeting in terms of impressions per EUR. This may be due to more narrow and thus more expensive targeting, or to an audience for which targeting is more expensive.

Discrepancies between targeted and actual audiences

Previous research has shown that algorithmic ad delivery can lead to discrepancies between intended and actual audiences on Meta (15, 16). This may be concerning if algorithmic ad delivery would propagate existing biases in voting patterns that restrict the reach of a party among a certain population, and thus hamper fair electoral competition. For example, some parties (e.g. *Grüne*) tend to receive more votes from younger female audiences while others are more popular among older male audiences (e.g. *AfD*). We thus explore potential discrepancies between the demographic distribution of the intended (= targeted) and the actual audience reached by political campaigns due to the algorithmic ad delivery. Meta does not provide detailed information on the actual audience with respect to other targeting criteria, which is why we focus on age and gender. These demographic characteristics are nevertheless important determinants for a plethora of life outcomes and ideological stances (40).

Let us first focus on age (Fig. 3). For each party, we compute the proportion of actual impressions generated by different age buckets, where we weight each ad by the amount of money spent. We refer to this vector as the *actual* audience of an ad. Similarly, we compute the proportion of targeted impressions by different age buckets, where we again weight each ad by the amount of money spent. This vector is the *target* audience of an ad. Figure 3a shows the *discrepancy* (percentage difference) between the *actual* and *target* audience. All parties reach an actual audience that is younger than the one targeted, except for the far-right *AfD* which generally reaches an older audience. We also compute the Wasserstein distance WS (see Supplementary Material for details) to quantify the difference (in years) between the age distribution of *actual* and *target* audiences for all parties. It yields: *Linke* = 4.69, *Grüne* = 6.35, *SPD* = 4.61, *FDP* = 5.18, *Union* = 5.69, and *AfD* = 4.89. Overall, the *Linke* and the *SPD* show the smallest discrepancy between the targeted and actual audience, while the *Grüne* exhibits the largest one. Figure 3b shows the values for *actual* and *target* for the *AfD*, thereby highlighting in red (green) the age buckets where the actual audience is smaller (larger) than the targeted one.

We repeat the procedure above to compute the variables for the *actual* and *target* audience of an ad by gender. For all parties, there is a large *discrepancy* between the targeted vs. actual audience in terms of gender, with ads shown to fewer female users than intended, except for the *Grüne* (see Fig. 3c). This is particularly pronounced for right-wing parties such as the *Union* and the *AfD*. For example, the *AfD* reaches 12.91% more male individuals than originally targeted. Interestingly, the *Grüne* is the only party for which the opposite is true: its ads reach 5 % more female individuals and 8 % fewer male individuals than targeted. To quantify the discrepancies in the gender distribution between *actual* and *target* audience, we again compute the Wasserstein distance for each

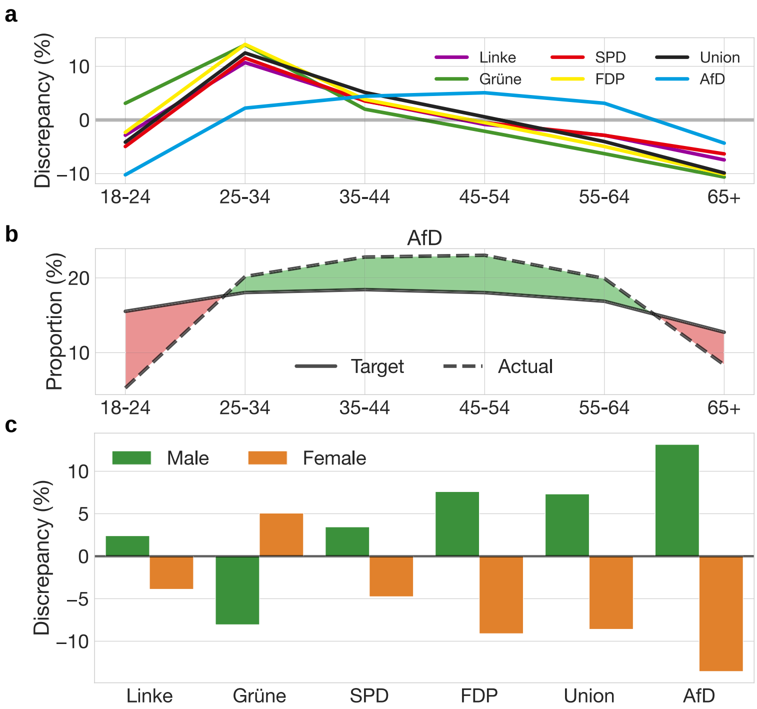

Fig. 3. a) Discrepancy in the age distribution between the *actual* and *target* audience (in %). We find that ads by most parties (except *AfD*) are seen by more users between 25 and 34 than originally intended. b) Comparison of *actual* and *target* audience by age for political ads published by *AfD*. A red color (i.e. proprotion of target audience is larger than actual audience) indicates areas where the difference between the actual and targeted audience is negative (green for positive, i.e. proportion of target audience smaller than actual audience). Younger and older users see the ads less often than originally intended by the party. c) Discrepancy in the gender distribution between the *actual* and *target* audience (in %). We find large differences between male and female audiences for right-wing parties (e.g. *Union*, *AfD*), implying that ads are seen by considerably fewer females than originally intended due to the algorithmic ad delivery.

party: *Linke* = 2.35, *Grüne* = 6.55, *SPD* = 3.27, *FDP* = 7.81, *Union* = 7.95, and *AfD* = 13.35, which corroborate the previous observations.

Systematic differences in impressions-per-EUR across parties

Regression analysis

To evaluate how key aspects of online political advertising are associated with impressions-per-EUR, we perform a regression analysis. For this, we estimate three separate linear regression models that focus on different determinants for impressions-per-EUR, namely, (i) targeting strategies, (ii) demographics, and (iii) ad characteristics (see Materials and methods for methodological details).

1. *Targeting strategy:* Our first regression model assesses how targeting strategies are related to impressions-per-EUR (see Fig. 4a). Ads that use more targeting criteria are linked to lower impressions-per-EUR as indicated by the negative and statistically significant coefficients for "No. Include criteria" ($P < 0.05$) and "No. Exclude criteria" ($P < 0.05$). This link is particularly strong for exclusions, where, all else equal, an additional 18.49 excluded criteria predicts a decrease of 68.37 impressions-per-EUR. Our regression results further show that targeting has a heterogeneous effect on impressions-per-EUR. For example, the usage of the targeting categories "Behaviors (Include)" and "Interests (Exclude)" is associated with considerably more impressions-per-EUR as indicated by a positive and significant coefficient ($P < 0.05$). In contrast, the negative and significant coefficient for "Employers (exclude)" and "Interests (include)" suggest that exclusion criteria for employers or inclusion criteria for interests correspond to lower levels of impressions-per-EUR. Interestingly, excluding custom audiences is linked to higher ad efficiency, while excluding lookalike audiences is negatively associated with ad efficiency as shown by a positive and statistically significant coefficient for "Custom audience (exclude)" ($P < 0.05$) and a negative and statistically significant for "Lookalike (exclude)" ($P < 0.05$). Given that ad delivery heavily relies on Meta's algorithm, more transparency would be crucial to explain these findings.
2. *Demographic segments:* Our second regression model evaluates how different demographics explain impressions-per-EUR. The regression results are in Fig. 4b. We find a positive and statistically significant coefficient for "Female" ($P < 0.05$) and "Male" ($P < 0.05$), suggesting that targeting only female or male audiences rather than all genders is associated with more impressions-per-EUR. All else equal, only targeting users from a single gender predicts, on average, an additional 43.90 ("Female only") and 38.08 ("Male only") impressions-per-EUR. Furthermore, the coefficients for *Age:* 18–24, *Age:* 25–34, and *Age:* 45–54 are positive and statistically significant ($P < 0.05$), while the coefficients for *Age:* 35–44, and *Age:* 65+ are negative

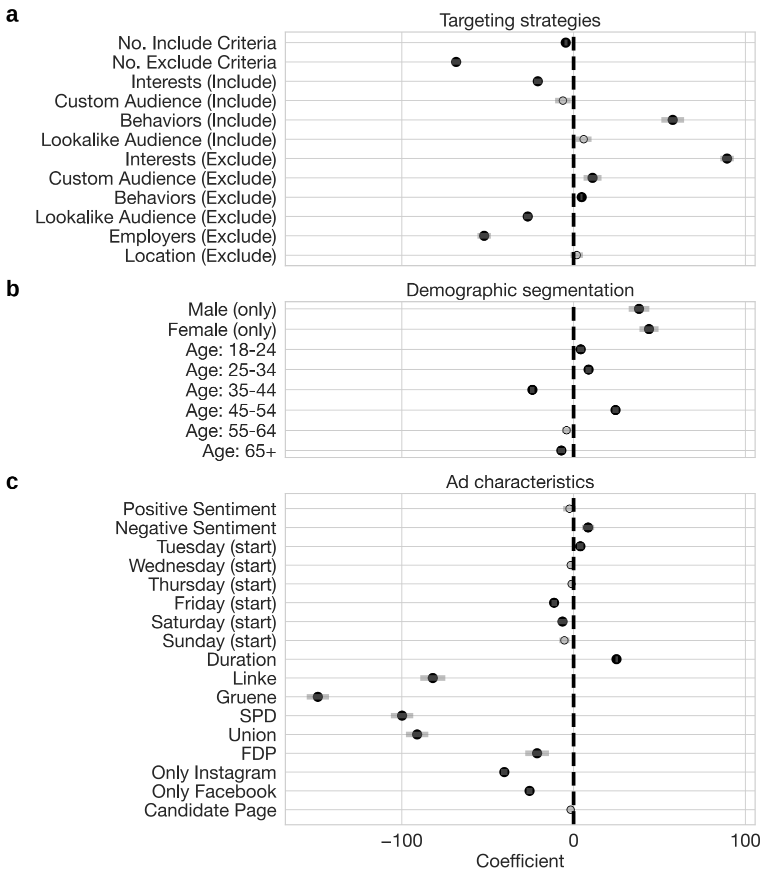

Fig. 4. Coefficient estimates and 95 % confidence intervals for a) targeting strategies, b) demographics, and c) ad characteristics. Statistically significant coefficients ($P < 0.05$) are indicated by black circles ●, all others by gray circles ○. Sentiment, weekday, platform dummy, and party dummy are categorical variables, and the reference categories are “neutral” sentiment, “Monday”, “both platforms”, and “AfD”.

and statistically significant ($P < 0.05$). As such, addressing younger audiences is linked to more impressions-per-EUR except for the age group between 35 and 44.

3. *Ad characteristics:* In our third regression model, we study how ad characteristics are linked to impressions-per-EUR (Fig. 4c). For example, the timing and content of an ad explain its impressions-per-EUR. In particular, ads that are online for a longer period and published earlier in the week tend to receive more impressions-per-EUR as indicated by the positive and statistically significant coefficient for *Duration* ($P < 0.05$) and *Tuesday* ($P < 0.05$) as well as the negative coefficients for the other weekdays. As such, all else equal, for each extra week that an ad remains online our model predicts 24.98 additional impressions-per-EUR. When assessing the link between impressions-per-EUR and whether a candidate has published an ad, we do not find a statistically significant coefficient ($P = 0.31$). However, the party dummy is an important determinant of impressions-per-EUR. In particular, ads published by the *AfD* are linked to more impressions-per-EUR compared to all other parties, which is seen by the negative and statistically significant coefficient for all other party variables ($P < 0.01$). All else equal, ads by the *AfD* reach +148.88, +99.86, +91.09, +81.97, and +21.25 additional impressions-per-EUR compared to the *Grüne*, *SPD*, *Union*, *Linke*, and *FDP*, respectively.

Of note, sentiment, weekday, platform dummy, and party dummy are categorical variables. Hence, they need a reference condition to include them in our regression model. The coefficients of these variables should be interpreted relative to the reference categories. In this analysis, we chose “neutral” sentiment, “Monday”, “both platforms”, and “AfD” as reference categories.

We further conducted two additional analyses, which are motivated by the fact that larger competition may influence ad impressions. Generally, we expect competition among political ads to be small given that political ads represent only a small fraction out of all ads on Meta. Nevertheless, we perform two analyses where we control for the time-to-election-day and the number of competing political ads. We find that publishing ads earlier in the campaign is related to higher levels of impressions-per-EUR. Furthermore, higher competition in terms of more active ads at the publishing day is negatively related to impressions-per-EUR. Of note, all other results remain consistent with our main analysis except for the coefficient of negative sentiment, which is no longer statistically significant. Details are in Supplementary Material S6.

Machine learning approach

We now employ a machine learning approach to predict the reach of political ads for a given budget based on the information provided by Meta. The rationale for this design is two-fold. (i) If the available information in the Meta Ad Library and the Meta Ad Targeting Dataset is sufficiently complete, we should be able to make accurate predictions of impressions-per-EUR given the available features. In other words, a low prediction performance suggests that other unobserved features explain the heterogeneity in the reach of different ads for a given budget (e.g. factors related to the algorithmic delivery of ads), but these features are not captured in the dataset and are thus not available for external stakeholders. Hence, this can provide a glimpse into the transparency provided by the available data. (ii) The machine learning predictions can also be used to examine empirically whether there are systematic differences between the predicted and actual reach of an ad for a given budget, across different political parties. In other words, if a party consistently receives more views for a given budget than others targeting the same audience, this could indicate that the algorithmic delivery of ads is advantaging said party, which would undermine fair competition.

We fit a random forest model (41) by using all variables from our regression analysis related to key determinants of the reach of an ad for a given budget, namely, (1) targeting strategies (e.g. the type and frequency of targeting categories), (2) demographics (age, gender of target group), and (3) ad characteristics (e.g. sentiment of an ad, ad duration, publishing party). In addition, we use the full set of targeting variables and add variables that indicate whether (i) the advertisers supplied a data file to include/exclude a custom audience and (ii) the data supplied by the advertisers to include/exclude a custom/lookalike audience was complete. For a full list of variables, see Materials and methods. To measure the reach of an ad for a given budget, we again rely on the number of impressions generated per EUR spent (i.e. impressions-per-EUR). Details on the implementation are in Materials and methods.

Our model achieves an average root mean squared error RMSE = 123.79 over 10 runs (± a s.d. of 4.21) when predicting impressions-per-EUR on the hold-out set. More importantly, our model can only explain a fraction of the variance in our data ($R^2 = 0.40 \pm$ a s.d. of 0.02 over 10 runs). The low R^2 suggests that the available information about targeting strategies, demographics, and ad characteristics is not sufficient to fully characterize the impressions-per-EUR of the ad campaign.

Next, we compute the mean difference between the predicted and actual impressions-per-EUR across parties. This measure indicates whether specific parties consistently achieve more or fewer impressions-per-EUR while controlling for all other available sources of heterogeneity in targeting strategies, demographics, and ad characteristics. Figure 5 shows that most left-leaning parties (i.e. the *Grüne* and the *SPD*, except for *Linke*) and *Union* consistently achieve fewer impressions-per-EUR than predicted by our model. In contrast, the *FDP* and *AfD*, on average, receive 12.83 and 2.81 additional impressions per EUR, respectively. This result implies a relative advantage of 10.13 and 2.22% in impressions-per-EUR compared to the average ad that achieves 126.71 impressions-per-EUR. Overall, our results suggest that the algorithmic delivery of political ads may advantage specific parties.

We perform a series of checks to ensure the robustness of our results. First, we train an XGBoost model using the same training procedure as outlined for the random forest model. Second, we check the heterogeneity of our results across platforms and re-train our random forest model by using political ads that were solely published on (i) Facebook and (ii) Instagram. Across both checks, our main findings regarding (i) and (ii) remain consistent. Details are in Supplementary Material S10.

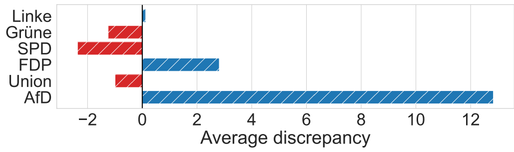

Fig. 5. Average difference between actual vs. predicted impressions-per-EUR based on our machine learning model over 10 runs.

Discussion

Targeted political advertising on social media has raised significant concerns regarding fairness, accountability, and transparency among researchers, policymakers, and society at large. Given the known and significant impact of political advertising on voter turnout and vote choice (25–28), it is crucial to analyze how parties across the political spectrum employ targeting during election campaigns and inform policymakers on the implications of targeted political advertising for democracies.

Our findings contribute to the existing literature on political advertising on social media by providing evidence of the prevalence of targeting across the political spectrum. While social media was found to be an essential communication channel for parties during election campaigns (9, 32, 37), it was previously unclear whether and how parties make use of targeting.

Targeting was used in 72.3% of all ads published on Meta during the 2021 German federal election. In fact, parties rely on a wider range of targeting categories available to advertisers on Meta. For example, a majority of campaign budget is spent for ads that exclude users based on so-called interests (e.g. social equality, environmentalism, and international relations), behaviors (e.g. early adopters of new technology, commuters, international travelers), or employers (e.g. business owners, police officers, Ford Deutschland). We further find significant differences in the reach of ads at a given budget. For example, the far-right *AfD* achieves significantly more impressions-per-EUR compared to other parties.

Our results further show that the algorithms of the platforms drive—to a large extent—*who* views an ad, which can lead to discrepancies between the intended (targeted) and actual audience. Indeed, we find considerable discrepancies between both. For instance, the *Grüne* reaches considerably more female users than intended, while the *Union* and the *AfD* reach a larger male audience than intended. Algorithmic bias in the delivery of political ads may amplify stereotypes associated with vote choices among distinct segments of the electorate. In fact, previous research has shown that algorithmic bias is responsible for systematically delivering ads based on gender stereotypes, but in a job context (17). This aligns with the fact that the *Grüne* traditionally receives more support from female voters compared to both the *Union* and the *AfD*, which more strongly depend on male voters.

The results of our regression analysis provides evidence for significant heterogeneity in the impressions-per-EUR of targeted ads. For example, ads targeting only a single gender achieve significantly more impressions-per-EUR. This may be due to a higher level of personalization but also raises concerns on whether algorithmic bias propagates gender-specific ads particularly well.

Moreover, ads by *Grüne* receive considerably fewer impressions-per-EUR than other parties. Given that their voters have an above-average income targeting them may be more costly than the average. In contrast, ads published by the *AfD* generate the most impressions-per-EUR. This could be explained by the fact that incendiary political issues promoted by populist parties (e.g. anti-immigration), tend to attract high attention on social media (32, 42, 43).

Finally, the results of our machine learning approach show that detailed information on targeting strategies, demographics, and ad characteristics cannot fully explain the reach of an ad for a given budget in terms of impressions-per-EUR. This finding indicates that current transparency measures do not suffice to evaluate how proprietary algorithms deliver political ads. In fact, we find systematic differences between predicted and actual impressions-per-EUR, which is particularly wide for the far-right *AfD*. This gap is concerning as it possibly indicates the presence of algorithmic biases that may favor populist ads.

As with other research, ours is not free from limitations that offer opportunities for future work. We focus on the 2021 German federal election; extending our results to other countries is important for generalizability. Nevertheless, the federal structure in Germany is similar to other countries, and the election is comparatively large, with more than 60 million eligible voters and candidates from parties across the political spectrum (44). Our analysis is based on a unique, large-scale dataset that is proprietary and, as such, offers granular insights. However, more research is needed to assess the accuracy of the Meta Ad Library (23). Furthermore, the effects of algorithmic ad delivery are not yet fully understood and may require further data disclosure by platforms. For instance, Meta only offers specific audience details related to age and gender, without providing comprehensive pricing information. This lack of data precludes a broader analysis of how pricing mechanisms vary and the discrepancies between the targeted and actual audiences. We thus encourage future research to focus on pricing mechanisms for political ads on social media and more granular analyses of actual and targeted audiences. Lastly, while our current focus is on targeting, we also provide insights on the purpose (see Supplementary Material S4) and focus of ads (i.e. candidates vs. parties; see Supplementary Material S5). Here, interdisciplinary approaches combining computational methods and theory from political science (45) could expand our analysis on the communication strategies employed by various parties on social media beyond targeting.

Our results contribute to the political science literature by shedding light on the prevalence of targeting during real-world elections. Targeting is highly prevalent in political campaigns and allows parties to focus on specific voter segments. This result mandates additional efforts by researchers to unravel the impact of targeted political advertising on society.

Our results have further implications for parties and policymakers. For parties, we provide valuable insights into the delivery mechanisms of targeted ads on social media. For policymakers, our research emphasizes the necessity to intensify auditing and regulation regarding political advertising on social media. The discrepancies between actual and targeted audiences that we identified potentially originate from algorithmic bias that favors certain voter segments. This is concerning as such bias may harm political participation and reinforce discrimination against disadvantaged groups.

Policymakers should pay particular attention to addressing these issues in regulatory frameworks and hold platforms accountable for ensuring fairness, accountability, and transparency in political advertising. For example, previous research (46) recommends that platforms make adaptations to the auction mechanisms to subsidize political advertisers or provide quotas in combination with separate auctions for political and commercial advertisers to lower competition for political advertisers. Policymakers could also require randomization of ad delivery among the target population as discussed by the European Parliament (47), thereby preventing discrimination due to algorithmic bias. Most importantly, policymakers should incentivize additional transparency measures. For example, current measures are insufficient to evaluate how the platform's pricing mechanism influences ad delivery, limiting independent assessment by researchers. Mandatory disclosure of information about click rates could help to better understand the effectiveness of social media ads. Overall, there is a pressing need for policymakers to mandate such disclosures, allowing for third-party monitoring and contributing to a more accountable system of political advertising on social media.

Materials and methods

Data

We collect $N = 81{,}549$ political ads from the 2021 German federal election published by candidates and parties (including their youth organizations) on Facebook and Instagram. Specifically, we first obtain all ads with a starting day in the period [2021 July 01, 2021 September 26] from the *Meta Ad Library API* (48). We then filter ads where the sponsor or page name matches any of the six major parties (i.e. *Linke*, *Grüne*, *SPD*, *FDP*, *Union*, and *AfD*) or one of their candidates running for office.[b] Of note, advertisers on Meta may purchase ads with the same creatives (i.e. content, image, etc.) multiple times. Therefore, we use the term "ad" to refer to a single ad contract, including its timing, budget, and targeting settings besides its content.

The Meta Ad Library API provides detailed information about content, page/sponsor name, money spent, start and stop dates, and the number of impressions distributed across gender and age (i.e. in buckets corresponding to 18–24, 25–34, 35–44, 45–54, 55–64, 65+ years). It also indicates whether an ad was published (a) only on Facebook, (b) only on Instagram, or (c) simultaneously on both platforms. For ad spending and the number of impressions, Meta only reports discretized buckets. Following previous research (9, 26, 31–34), we average the maximum and minimum of each bucket to obtain conservative point estimates of the ad spending and the number of impressions per ad. A full list of variables is available in Supplementary Material S7.

We further use the *Meta Ad Targeting Dataset* (49) to access ad targeting information. The *Meta Ad Targeting Dataset* contains targeting information on all social issue, electoral, and political ads published after 2020 August 3, on Facebook or Instagram (49). The dataset can be accessed after approval by Meta via https://developers.facebook.com/docs/ad-targeting-dataset. We accessed the dataset in August 2023 and queried targeting information for all $N = 81{,}549$ political ads retrieved from the *Meta Ad Library API* based on the unique ad ids. In particular, we queried information about the users' demographics (age and gender) targeted by each ad, as well as additional targeting categories such as interests, behaviors, jobs, or locations. Throughout the article, we consider an ad to employ "targeting" if it uses any category provided in the dataset in addition to demographics (i.e. gender and age) and location (since the latter must always be specified in the ad creation). We further choose to ignore the "Include location" category, which is defined by default for every ad (e.g.

"Germany"), but we do consider "Exclude location" (only present for a very small minority of ads = 3.35%). Targeting information for each ad is specified in groups of categories (e.g. "Interests" such as *Politics* and *Environment*), and the ad will include/exclude people that match at least one category in each group. We notice that the majority of the "include" conditions in our data contains 1–2 groups, while all "exclude" conditions contain a single group. For a comprehensive overview of the targeting information provided in the dataset, see Supplementary Material S8. For an example ad with targeting information, see Fig. S4.

Wasserstein distance

Following Capozzi et al. (32), we compute the Wasserstein distance *WS* to quantify discrepancies between the demographic distribution (i.e. age and gender) of the *Target* and *Actual* audience of an ad. The *WS* distance, also known as the earth mover's distance (EMD) or Kantorovich–Rubinstein distance, is a measure of the dissimilarity between two probability distributions. It quantifies the minimum cost required to transform one distribution into another, where the cost is determined by the amount of "mass" that needs to be moved from each point in one distribution to its corresponding point in the other distribution. We compute it using the `scipy` Python library, which is based on the following formula:

$$l_1(u, v) = \int_{-\infty}^{+\infty} |U - V|,$$

where u and v are the two distributions to compare, and U and V the respective cumulative distribution functions.

Regression analysis

We hypothesize that targeting affects the impressions-per-EUR of political ads on social media. We thus use regression analysis to study what factors drive the reach of an ad at a given budget for political advertising on social media. To measure the reach of an ad, we rely on the number of impressions an ad generated per EUR spent (or simply impressions-per-EUR). Let y_i denote impressions-per-EUR for an ad i, and let x_i refer to a vector with different characteristics belonging to that ad. We then estimate the following linear regression model

$$y_i = \alpha + \beta^T x_i, \tag{1}$$

where α represents the model intercept, and β measures the associations between the variables in x_i and y_i. For estimation, we use ordinary least squares regression (OLS) and test whether the coefficients are significantly different from zero using two-sided *t*-tests. To facilitate the interpretability of our results, we further *z*-standardize all numerical variables in *x*. Hence, we interpret our coefficients as follows: For numerical variables, a one standard deviation increase in *x* predicts an increase in the reach of an ad at a given budget by β impressions-per-EUR. For dummy variables, a dummy variable in *x* set to one predicts an increase in the reach of an ad at a given budget by β impressions-per-EUR.

We use three different regression models to analyze how impressions-per-EUR varies across ads. To do so, we use a comprehensive set of variables available to us in the Meta Ad Library or the Meta Ad Targeting Dataset that represent three key determinants of impressions-per-EUR. In particular, we focus on (i) different targeting strategies, (ii) different demographics, and (iii) different ad characteristics. We estimate separate models for each of these dimensions to avoid multicollinearity (50) and facilitate interpretability.

Targeting strategy: To assess the role of targeting strategies as a determinant of impressions-per-EUR, we compute two variables that correspond to the overall number of inclusion and exclusion criteria. The hypothesis is that more granular targeting may incur higher costs. Further, we consider which targeting categories have been used by a party. Here, we include two variables for each targeting category available to advertisers on Meta that indicate whether a certain category was used (=1 if the category was used, and =0 otherwise) to distribute an ad. The first (second) variable indicates whether users should be included (excluded) from the audience based on the corresponding targeting category. Table S3 lists all targeting categories available to advertisers on Meta. Targeting categories are often employed together. To improve interpretaiblity and mitigate multicollinearity concerns, we thus focus on the top-10 targeting categories (by total spending). The top-10 targeting categories (by total spending) are shown in Fig. S1.

Demographics: To study which demographics are linked to higher impressions-per-EUR, we include whether an ad is targeting only users from a single gender (i.e. only female or male users; =1 if yes, and =0 otherwise) as well as the share of targeted users across different age groups (i.e. 18–24, 25–34, 35–44, 45–54, 55–64, 65+).

Ad characteristics: Ad characteristics are likely to influence the reach of political ads at a given budget. Hence, we study how (i) content, (ii) timing, (iii) the platform an ad was published on, and (iv) the publisher of an ad is linked to impressions-per-EUR. (i) *Content*: We thus analyze the sentiment conveyed by an ad and classify the content of each ad as "positive", "neutral", or "negative". We use German Sentiment Bert, a state-of-the-art transformer-based sentiment model for German text that was trained on 5.4 million labeled samples (51). (ii) *Timing*: We analyze whether launching an ad on specific weekdays is beneficial and how ad duration (i.e. the timespan an ad remains online) relates to impressions-per-EUR. (iii) *Platform:* Platform characteristics such as audience, user behavior, and ad competition can influence political campaigning (52). Hence, we study the association between impressions-per-EUR and the platform on which an ad was published, by using dummy variables to encode whether an ad was published on Facebook, Instagram, or both platforms simultaneously. (iv) *Publisher*: In the context of the German dual-vote system, which features both party and candidate votes, political science literature shows the role of both party and candidate behavior in shaping voter perceptions (53). These perceptions may affect the impressions-per-EUR of ads authored by different party and candidate pages. Thus, we encode the different parties using dummy variables. We further encode whether the ad was distributed through a candidate's page (=1 if yes, and =0 otherwise).

Sentiment, weekday, platform dummy, and party dummy are multilevel categorical variables. Hence, we have to choose a reference condition to include them in our regression model and interpret the coefficients relative to the reference categories. In analysis, we choose "neutral" sentiment, "Monday", "both platforms", and "AfD" as reference categories.

Machine learning approach

Objective: We employ a machine learning approach to evaluate whether we can predict the reach of an ad for a given budget as measured by impressions-per-EUR. The aim is two-fold. (i) We study whether a machine learning model can accurately predict impressions-per-EUR of ads based on the information provided in the Meta Ad Library and Meta Ad Targeting Dataset. A high prediction performance implies that the provided information is sufficient to understand the ad delivery mechanisms of the platforms. Vice versa, a low prediction performance suggests the presence of unobserved confounders that explain the observed heterogeneity but which are currently unavailable for external audits. (ii) We analyze the difference between predicted and actual impressions-per-EUR across parties. In a fair environment, we would expect no systematic differences in terms of impressions-per-EUR between different parties. In contrast, if a party consistently receives more views at a given budget than others targeting the same audience, this could indicate that the algorithmic delivery of ads is advantaging said party, which would undermine fair competition.

Features: For our machine learning model, we use all variables from our regression model that represent key determinants of impressions-per-EUR: (i) different targeting strategies, (ii) different demographics, and (iii) different ad characteristics (see above).

In addition, we use the full set of targeting variables (see Table S3 for a full list) and add variables that indicate whether (i) the advertisers supplied a data file to include/exclude a custom audience (=1 if the data file was supplied, and =0 otherwise) and (ii) the data supplied by the advertisers to include/exclude a custom/lookalike audience was complete (=1 if the data were complete, and = 0 otherwise).

Implementation: We (i) study whether machine learning can accurately predict impressions-per-EUR based on the above variables and (ii) analyze the difference between predicted and actual impressions-per-EUR. To do so, we fit a random forest model (41) by using all features from above. We split our data into a training (80%) and a hold-out set (20%) for evaluation, and tune the model via 10-fold cross-validation in combination with a grid search (see below). To control for the number of ads published by each party, we weigh observations inverse proportionally to the ad frequency by party in our training set when fitting the model. We further z-standardize numeric variables.

Hyperparameter tuning: For the training of the random forest model, we use 10-fold cross-validation in combination with a grid search. In particular, we vary (i) the number of trees used for a forest (*N estimators*), (ii) the number of variables to consider at each split (*Max features*), (iii) maximum depth of the tree (*Max depth*), (iv) minimum samples to split a node (*Min node*), (v) minimum samples in a leaf (*Min leaf*), and (vi) whether to bootstrap samples when building trees (*Bootstrap*). Details on the hyperparameter tuning are in Supplementary Material S9.

Evaluation: We evaluate the predictive power of our model based on the average root mean squared error (RMSE) over 10 runs with different seeds. We further rely on the R^2 to determine how well our set of variables is able to explain variance in impressions-per-EUR. We compute the R^2 via

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}, \quad (2)$$

where n is the number of observations, with x_i referring to a vector with different ad characteristics and y_i representing predicted impressions-per-EUR, and $\bar{y}$ and $\bar{x}$ as their respective means. In a fully transparent setting, we would expect the R^2 to be close to 1.

Notes

[a] Advertisers on Meta may purchase ads with the same creatives (i.e. content, image, etc.) multiple times. Therefore, we use "ad" to refer to the specific ad contract, including its timing, budget, and targeting settings besides its creative parts.

[b] For detailed election results, see www.bundeswahlleiterin.de/en/bundestagswahlen/2021/ergebnisse.html.

Acknowledgments

This manuscript was posted on a preprint server: https://arxiv.org/abs/2310.10001.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

F.P. is partially funded by the European Union (NextGenerationEU project PNRR-PE-AI FAIR), the Italian Ministry of Education (PRIN PNRR grant CODE and PRIN grant DEMON). This manuscript reflects only the authors' views and opinions, and funding bodies are not responsible for them.

Author Contributions

D.B., F.P., G.D.F.M., and S.F. contributed to conceptualization. D.B. and F.P. contributed to data analysis. D.B., F.P., G.D.F.M., and S.F. contributed to results interpretation and manuscript writing. D.B., F.P., G.D.F.M., and S.F. approved the manuscript.

Data Availability

All code to replicate our analyses is available via our GitHub repository at https://github.com/DominikBaer95/auditing_targeted_political_advertising. All data used for the analysis is publicly available. Data on political ads on Facebook and Instagram is available via the Meta Ad Library: https://www.facebook.com/ads/library/. Targeting data is available via the Meta Ad Targeting Dataset: https://developers.facebook.com/docs/fort-ads-targeting-dataset/. To ensure reproducibility, we provide ids for all ads in our dataset, which can be used to retrieve the original data through the Meta Ad Library (both through the API and with the web interface by simply using the id as query) via our GitHub repository at https://github.com/DominikBaer95/auditing_targeted_political_advertising. Due to Meta's ToS we cannot share any further information.

References

1 Statista. Statista - empowering people with data [accessed 2023 Oct 05]. https://www.statista.com.

2 Votta F, *et al.* 2024. Who does(n't) target you? Mapping the worldwide usage of online political microtargeting. *J Quant Descr Digit Media*. 4:1–47.

3 Fowler EF, Franz MM, Ridout TN. 2020. Online political advertising in the United States. In: Persily N, Tucker JA, editors. *Social media and democracy: the state of the field, prospects for reform*. Cambridge, (UK): Cambridge University Press. p. 111–138.

4 GLES. 2022. GLES Kandidierendenstudie 2021. https://doi.org/10.4232/1.14100.

5 Breza E, *et al.* 2021. Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: a cluster randomized controlled trial. *Nat Med*. 27(9):1622–1628.

6 Goldberg MH, Gustafson A, Rosenthal SA, Leiserowitz A. 2021. Shifting republican views on climate change through targeted advertising. *Nat Clim Change*. 11(7):573–577.
7 Matz SC, Kosinski M, Nave G, Stillwell DJ. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proc Natl Acad Sci U S A*. 114(48):12714–12719.
8 Ridout TN, Fowler EF, Franz MM. 2021. The influence of goals and timing: how campaigns deploy ads on Facebook. *J Inf Technol Polit*. 18(3):293–309.
9 Fowler EF, Franz MM, Martin GJ, Peskowitz Z, Ridout TN. 2021. Political advertising online and offline. *Am Polit Sci Rev*. 115(1): 130–149.
10 Dommett K. 2019. Data-driven political campaigns in practice: understanding and regulating diverse data-driven campaigns. *Internet Policy Rev*. 8(4). https://policyreview.info/articles/analysis/data-driven-political-campaigns-practice-understanding-and-regulating-diverse-data.
11 Imana B, Korolova A, Heidemann J. 2021. Auditing for discrimination in algorithms delivering job ads. In: *Proceedings of the Web Conference (WWW)*. New York (NY): Association for Computing Machinery.
12 Tappin BM, Wittenberg C, Hewitt LB, Berinsky AJ, Rand DG. 2023. Quantifying the potential persuasive returns to political microtargeting. *Proc Natl Acad Sci U S A*. 120(25):e2216261120.
13 Hersh ED, Schaffner BF. 2013. Targeted campaign appeals and the value of ambiguity. *J Polit*. 75(2):520–534.
14 Votta F, Noroozian A, Dobber T, Helberger N, de Vreese C. 2023. Going micro to go negative? Targeting toxicity using Facebook and Instagram ads. *Comput Commun Res*. 5(1):1–50.
15 Ali M. 2019. Discrimination through optimization: how Facebook's ad delivery can lead to biased outcomes. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. New York (NY): Association for Computing Machinery.
16 Ali M, Sapiezynski P, Korolova A, Mislove A, Rieke A. 2021. Ad delivery algorithms: the hidden arbiters of political messaging. In: *International Conference on Web Search and Data Mining (WSDM)*. New York (NY): Association for Computing Machinery.
17 Lambrecht A, Tucker C. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manage Sci*. 65(7):2966–2981.
18 Auxier B. 2020. 54% of Americans say social media companies shouldn't allow any political ads. Pew Research Center. [accessed 2023 Oct 05]. https://www.pewresearch.org/short-reads/2020/09/24/54-of-americans-say-social-media-companies-shouldnt-allow-any-political-ads/.
19 Cabañas JG, Cuevas Á, Arrate A, Cuevas R. 2021. Does Facebook use sensitive data for advertising purposes? *Commun ACM*. 64(1): 62–69.
20 Korolova A. 2011. Privacy violations using microtargeted ads: a case study. *J Priv Confidentiality*. 3:1.
21 Dommett K. 2019. The rise of online political advertising. *Polit Insight*. 10(4):12–15.
22 Isaak J, Hanna MJ. 2018. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*. 51(8):56–59.
23 Edelson L, Lauinger T, McCoy D. 2020. A security analysis of the Facebook ad library. In: *IEEE Symposium on Security and Privacy*. New York (NY): IEEE.
24 Le Pochat V. 2022. An audit of Facebook's political ad policy enforcement. In: *USENIX Security*. Berkeley (CA): USENIX.
25 Aggarwal M, *et al.* 2023. A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. *Nat Hum Behav*. 7(3):332–341.
26 Bär D, Pröllochs N, Feuerriegel S. 2024. The effect of social media ads on election outcomes. OSF Preprints. https://doi.org/10.31219/osf.io/q8mxj, preprint: not peer reviewed.
27 Coppock A, Green DP, Porter E. 2022. Does digital advertising affect vote choice? Evidence from a randomized field experiment. *Res Polit*. 9(1):20531680221076901.
28 Hager A. 2019. Do online ads influence vote choice? *Polit Commun*. 36(3):376–393.
29 United States House of Representatives. 2021. For the people act [accessed 2023 Oct 05]. https://www.congress.gov/117/bills/hr1/BILLS-117hr1ih.pdf.
30 European Commission. 2022. The digital services act: ensuring a safe and accountable online environment [accessed 2023 Oct 05]. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.
31 Aisenpreis L, Gyrst G, Sekara V. 2023. How do US congress members advertise climate change: an analysis of ads run on Meta's platforms. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. Washington (DC), USA: AAAI Press.
32 Capozzi A, de Francisci Morales G, Mejova Y, Monti C, Panisson A. 2023. The thin ideology of populist advertising on Facebook during the 2019 EU elections. In: *Proceedings of the Web Conference (WWW)*. New York (NY): Association for Computing Machinery.
33 Capozzi A. 2020. Facebook ads: politics of migration in Italy. In: *International Conference on Social Informatics*. Cham, (Germany): Springer.
34 Capozzi A. 2021. Clandestino or rifugiato? Anti-immigration Facebook ad targeting in Italy. In: *Conference on Human Factors in Computing Systems (CHI)*. New York (NY): Association for Computing Machinery.
35 Pierri F. 2023. Political advertisement on Facebook and Instagram in the run up to 2022 Italian general election. In: *ACM Web Science Conference (WebSci)*. New York (NY): Association for Computing Machinery.
36 Brodnax NLM, Sapiezynski P. 2022. From home base to swing states: the evolution of digital advertising strategies during the 2020 US presidential primary. *Polit Res Q*. 75(2):460–478.
37 Kruschinski S, Haßler J, Jost P, Sülflow M. 2022. Posting or advertising? How political parties adapt their messaging strategies to Facebook's organic and paid media affordances. *J Polit Marketing*. Forthcoming. London, (UK): Taylor and Francis.
38 Dommett K, Power S. 2019. The political economy of Facebook advertising: election spending, regulation and targeting online. *Polit Q*. 90(2):257–265.
39 Kruskal WH, Wallis WA. 1952. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 47(260):583–621.
40 Harting H, Daniller A, Keeter S, van Green T. 2023. Voting patterns in the 2022 elections. Pew Research Center. [accessed 2024 Jun 14]. https://www.pewresearch.org/politics/2023/07/12/republican-gains-in-2022-midterms-driven-mostly-by-turnout-advantage/.
41 Breiman L. 1984. *Classification and regression trees*. New York (NY): Routledge.
42 Bene M, *et al.* 2022. Keep them engaged! Investigating the effects of self-centered social media communication style on user engagement in 12 European countries. *Polit Commun*. 39(4):429–453.
43 Klinger U, Koc-Michalska K, Russmann U. 2023. Are campaigns getting uglier, and who is to blame? Negativity, dramatization and populism on Facebook in the 2014 and 2019 EP election campaigns. *Polit Commun*. 40(3):263–282.

44 Bundeswahlleiter. 2022. The federal returning officer [accessed 2023 Oct 05]. https://www.bundeswahlleiter.de/en/bundeswahlleiter.html.

45 Bär D, Pröllochs N, Feuerriegel S. 2023. New threats to society from free-speech social media platforms. *Commun ACM*. 66(10):37–40.

46 Turkel E. 2022. Regulating online political advertising. In: *Proceedings of the Web Conference (WWW)*. New York (NY), USA: Association for Computing Machinery.

47 Council of the European Union. 2023. Proposal for a regulation of the European Parliament and of the Council on the transparency and targeting of political advertising: latest state of play of the interinstitutional negotiations [accessed 2024 Jun 05]. https://www.asktheeu.org/de/request/12988/response/45245/attach/5/st09184.en23.pdf?cookie_passthrough=1.

48 Meta. 2022. Ad library [accessed 2023 Oct 05]. https://www.facebook.com/ads/library/.

49 Meta. 2023. Ad targeting dataset [accessed 2023 Oct 05]. https://developers.facebook.com/docs/fort-ads-targeting-dataset.

50 Pröllochs N, Bär D, Feuerriegel S. 2021. Emotions in online rumor diffusion. *EPJ Data Sci*. 10(1):51.

51 Guhr O, Schumann A-K, Bahrmann F, Böhme HJ. 2020. Training a broad-coverage German sentiment classification model for dialog systems. In: *Language Resources and Evaluation Conference (LREC)*. Paris, France: European Language Resources Association.

52 Bossetta M. 2018. The digital architectures of social media: comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. election. *J Mass Commun Q*. 95(2): 471–496.

53 Gschwend T, Zittel T. 2015. Do constituency candidates matter in German Federal Elections? The personal vote as an interactive process. *Elect Stud*. 39:338–349.

Part III: Interventions to counter harmful content

11 Generative AI may backfire for counterspeech

Title: Generative AI may backfire for counterspeech

Abstract: Online hate speech poses a serious threat to individual well-being and societal cohesion. A promising solution to curb online hate speech is counterspeech. Counterspeech is aimed at encouraging users to reconsider hateful posts by direct replies. However, current methods lack scalability due to the need for human intervention or fail to adapt to the specific context of the post. A potential remedy is the use of generative AI, specifically large language models (LLMs), to write tailored counterspeech messages. In this paper, we analyze whether contextualized counterspeech generated by state-of-the-art LLMs is effective in curbing online hate speech. To do so, we conducted a large-scale, pre-registered field experiment ($N = 2,664$) on the social media platform Twitter/X. Our experiment followed a 2x2 between-subjects design and, additionally, a control condition with no counterspeech. On the one hand, users posting hateful content on Twitter/X were randomly assigned to receive either (a) contextualized counterspeech or (b) non-contextualized counterspeech. Here, the former is generated through LLMs, while the latter relies on predefined, generic messages. On the other hand, we tested two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. We then measured whether users deleted their initial hateful posts and whether their behavior changed after the counterspeech intervention (e.g., whether users adopted a less toxic language). We find that non-contextualized counterspeech employing a warning-of-consequence strategy significantly reduces online hate speech. However, contextualized counterspeech generated by LLMs proves ineffective and may even backfire.

Author contributions: Dominik Bär, Abdurahman Maarouf, and Stefan Feuerriegel contributed to conceptualization. Dominik Bär performed the data analysis. Dominik Bär, Abdurahman Maarouf, and Stefan Feuerriegel contributed to results interpretation and manuscript writing.

Reference: **Bär, D.**, Maarouf, A., & Feuerriegel, S. 2024. *Generative AI may backfire for counterspeech.* In: **arXiv**, 2411.14986, DOI: `https://doi.org/10.48550/arXiv.2411.14986`.

Supplementary materials: Supplementary materials for this article are in Supplementary Material D.

Generative AI may backfire for counterspeech

DOMINIK BÄR*, LMU Munich, Munich Center for Machine Learning, Germany
ABDURAHMAN MAAROUF, LMU Munich, Munich Center for Machine Learning, Germany
STEFAN FEUERRIEGEL, LMU Munich, Munich Center for Machine Learning, Germany

Online hate speech poses a serious threat to individual well-being and societal cohesion. A promising solution to curb online hate speech is counterspeech. Counterspeech is aimed at encouraging users to reconsider hateful posts by direct replies. However, current methods lack scalability due to the need for human intervention or fail to adapt to the specific context of the post. A potential remedy is the use of generative AI, specifically large language models (LLMs), to write tailored counterspeech messages. In this paper, we analyze whether contextualized counterspeech generated by state-of-the-art LLMs is effective in curbing online hate speech. To do so, we conducted a large-scale, pre-registered field experiment ($N = 2,664$) on the social media platform Twitter/X. Our experiment followed a 2x2 between-subjects design and, additionally, a control condition with no counterspeech. On the one hand, users posting hateful content on Twitter/X were randomly assigned to receive either (a) contextualized counterspeech or (b) non-contextualized counterspeech. Here, the former is generated through LLMs, while the latter relies on predefined, generic messages. On the other hand, we tested two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. We then measured whether users deleted their initial hateful posts and whether their behavior changed after the counterspeech intervention (e.g., whether users adopted a less toxic language). We find that non-contextualized counterspeech employing a warning-of-consequence strategy significantly reduces online hate speech. However, contextualized counterspeech generated by LLMs proves ineffective and may even backfire.

Additional Key Words and Phrases: hate speech, content moderation, counterspeech, social media, field experiment

Warning: Content in this paper may be upsetting or offensive. Reader discretion is advised

1 Introduction

Online hate speech poses a serious threat to individual well-being and societal cohesion. Individuals who experience online hate speech frequently suffer from psychological consequences that negatively affect their mental and physical well-being [17, 28, 46, 59]. Additionally, online hate speech is known to foster hostility between societal groups [48, 49] and may even motivate real-world violence such as witnessed in the 2017 Rohingya genocide in Myanmar [5] and the 2019 Christchurch mosque shooting [55]. Reducing online hate speech is thus a pressing issue for society. Here, we evaluate whether generative AI, specifically large language models (LLMs), can help in writing counterspeech and thereby reduce hate speech on social media.

Counterspeech refers to direct responses (typically posted publicly) intended to encourage users to reconsider their hateful posts [32]. A benefit of counterspeech is that it does not infringe on users' freedom of speech since no content is removed. Generally, there are different strategies to counter online hate speech. For example, one counterspeech strategy is to promote **empathy** toward the attacked group or individual (e.g., *"Imagine how it feels for group X to see people be attacked like this ..."*) [23, 54]. Another strategy is commonly referred to as **warning-of-consequences** and reminds offenders of social norms and warns of the consequences of online misbehavior (e.g., *"This is hate speech!*

*To whom correspondence should be addressed. E-mail: baer@lmu.de; Mail: LMU Munich, Geschwister-Scholl-Platz 1, 80539 München

Authors' Contact Information: Dominik Bär, LMU Munich, and Munich Center for Machine Learning, Munich, Germany, baer@lmu.de; Abdurahman Maarouf, LMU Munich, and Munich Center for Machine Learning, Munich, Germany, a.maarouf@lmu.de; Stefan Feuerriegel, LMU Munich, and Munich Center for Machine Learning, Munich, Germany, feuerriegel@lmu.de.

Such posts can damage your personal and professional reputation") [3, 4, 49, 54]. Previous research has demonstrated the effectiveness of counterspeech for reducing online hate speech across multiple field experiments [23, 37, 38, 49]. Later, we build upon the two strategies, which we then automated for contextualized counterspeech through the use of LLMs.

Counterspeech is traditionally implemented via two approaches: (1) manual counterspeech through human counterspeakers [53] or (2) scalable approaches with predefined, yet generic counterspeech messages under an "one-fits-all" paradigm [23, 37, 38, 49]. Manual counterspeech done by human counterspeakers is highly flexible and can be specifically tailored to the context of a hateful post [19]. However, manual counterspeech requires significant human effort and is thus **not** scalable to large social media platforms. Counterspeakers are further exposed to large amounts of online hate speech, which may negatively affect their well-being [53], rendering this approach impractical. In contrast, the "one-fits-all" approach from previous research [23, 37, 38, 49] can be automated and is thus scalable to a wider population of social media users. However, this approach ignores the context of a hateful post, potentially limiting the effectiveness of such a non-contextualized approach [36, 54]. In sum, existing studies have only studied scalable approaches based on a "one-fits-all" paradigm where hateful users received predefined, yet generic counterspeech messages. However, a counterspeech approach that is scalable and contextualized is missing (see Supplementary Material 1 for a comprehensive overview of the literature).

There are good reasons to believe that contextualized counterspeech generated by an LLM is effective in curbing online hate speech. Context generally plays an important role in countering hate speech [36, 54]. For example, educating an aggressor on "why" a post is offensive may be more effective and lead to more lasting behavior change than a generic message (e.g., a user may not be fully aware of why a post is perceived as offensive) [54]. Previous research has demonstrated that LLM-generated messages are generally persuasive across various applications [33] but outside of counterspeech. For example, LLMs can generate messages that successfully mediate between opposing groups [58], decrease conspiracy beliefs [16] and promote civility in online conversations [6, 20]. Thus, it is likely that crafting custom messages through an LLM could also encourage online offenders to reconsider their hateful posts and, therefore, potentially reduce hate speech.

However, there are also reasons why counterspeech generated by LLMs may be ineffective. Generally, whether LLMs are persuasive varies across different use cases [56, 62]. For example, outside of counterspeech, some works ask users to have long discussions with chatbots and then assess whether their beliefs have changed as a result [6, 16, 20, 58]. In contrast, one-time interventions such as counterspeech are minimally invasive and may thus be ineffective. Additionally, studies suggest that contextualized messages are more likely to be identified as LLM-generated, which could lessen their impact compared to non-contextualized messages [21]. In fact, the identity of the source delivering counterspeech is crucial to be effective [37, 49, 52]. Users who recognize that they are interacting with an LLM may resist changing their behavior or feel deceived. There is also evidence that counterspeech may be perceived as intrusive and therefore can backfire and even escalate hostility [32].

Motivated by the above, we study the effectiveness of generative AI in the form of LLMs as a scalable approach for writing contextualized counterspeech. Modern LLMs can generate human-like text tailored to specific contexts [18, 21, 25, 51, 67], which allows to generate counterspeech at scale and further enables to provide counterspeech that is contextualized to a specific topic.

In this paper, we analyze whether contextualized counterspeech generated by LLMs is effective in curbing online hate speech. To do so, we conducted a large-scale, pre-registered field experiment ($N = 2,664$) on the social media platform Twitter/X (see Fig. 1 for an overview). A particular strength of our study is its external validity. In contrast to survey or lab experiments that rely on simulated online environments, we provide real-world evidence from actual social media users who posted hate speech.

Our experiment followed a 2x2 between-subjects design, with an additional control condition without counterspeech. Users posting hateful content on Twitter/X were randomly assigned to receive either (a) contextualized counterspeech or (b) non-contextualized counterspeech. Here, the former is generated through an LLM, while the latter relies on predefined, generic messages. In doing so, we test whether counterspeech is more effective when carefully tailored to the context of the original hate speech post. Additionally, we employed two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. Here, we test which strategies are effective and whether the effectiveness may be positively (or negatively) influenced when contextualized counterspeech messages are crafted through an LLM. Eventually, we measured whether users reconsidered their actions (i.e., whether they deleted their initial hateful post) and whether their behavior changed as a response to the intervention (i.e., whether they posted fewer hateful posts and adopted a less toxic language). Thereby, we contribute new insights into the role of LLMs in promoting online civility. Importantly, we later find that the use of LLMs may even backfire and thus call for caution when LLMs are used to improve online safety.

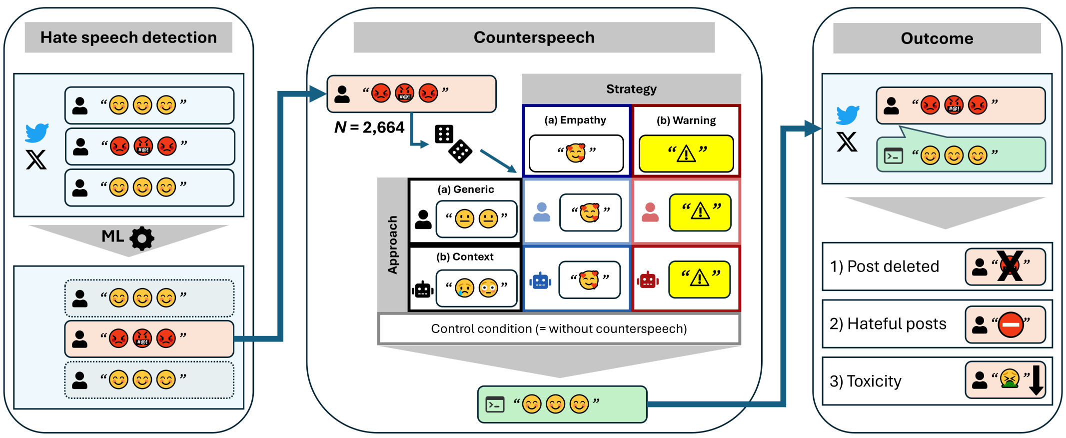

Fig. 1. Overview of our field experiment.

2 Results

To evaluate the effectiveness of contextualized counterspeech generated by an LLM to reduce hateful content on social media, we conducted a large-scale, pre-registered[1] field experiment on Twitter/X (www.X.com). We randomly assigned users that posted hate speech on Twitter/X to one of five treatment conditions (see Table 1). Users receive either (a) contextualized counterspeech or (b) non-contextualized counterspeech. Additionally, we employed two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. Details on the experimental procedure are described in Section 4.

To evaluate the effectiveness of our intervention, we then relied on three outcome variables: (1) We measured whether users reconsidered their actions (i.e., whether they deleted their initial hateful post). (2) We monitor whether they

[1]Pre-registration available at https://osf.io/38saz/?view_only=263687bff9824852b8ed204f257de8d5

Table 1. Experimental conditions for the field experiment. We randomly assigned users to one of five experimental conditions: Users receive either (a) contextualized LLM-generated counterspeech or (b) non-contextualized counterspeech. We further tested two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. Additionally, we used a control condition without intervention (i.e., without counterspeech).

	Empathy	Warning-of-Consequences
Non-contextualized	GENERIC-EMPATHY	GENERIC-WARNING
Contextualized	CONTEXT-EMPATHY	CONTEXT-WARNING
CONTROL CONDITION (=no counterspeech intervention)		

posted fewer hateful posts. (3) We check whether users adopted a less toxic language. For (2) and (3), we monitor users' posts in the two weeks following our intervention. Overall, our outcome variables should capture users' immediate reactions to the intervention as well as any changes in their behavior during the subsequent period.

2.1 Effectiveness of counterspeech

To evaluate the effectiveness of counterspeech in reducing online hate speech, we compare each intervention to the CONTROL CONDITION (=no counterspeech intervention) across our outcome variables. Fig. 2 shows the result, which we discuss in the following:

• (1) *Rate of deleted posts*: Fig. 2a shows the average rate of deleted posts across each condition in our experiment. In the control group, on average, 7.13 % of users deleted their original posts. Across all conditions, except CONTEXT-WARNING, fewer users deleted their hateful posts following counterspeech. On average 3.94 % (−3.19 p.p. compared to control), 3.74 % (−3.39 p.p.), and 5.21 % (−1.92 p.p.) of users deleted their hateful posts for GENERIC-EMPATHY, CONTEXT-EMPATHY, and CONTEXT-WARNING counterspeech, respectively. In contrast, counterspeech based on GENERIC-WARNING resulted in an average of 7.72 % (+0.59 %) of users deleting their hateful posts, indicating that users are encouraged to remove hateful content when receiving non-contextualized warning-of-consequences counterspeech.

• (2) *Number of hateful posts*: The average number of hateful posts shared by each user within two weeks after the intervention for each experimental condition is shown in Fig. 2b. In the control group, users shared an average of 9.07 hateful posts in the two weeks following the intervention. Users shared fewer hateful posts following counterspeech when receiving CONTEXT-EMPATHY or GENERIC-WARNING counterspeech. Specifically, users shared an average of 8.18 (−0.89 compared to control) and 8.04 (−1.03) hateful posts for CONTEXT-EMPATHY, and GENERIC-WARNING counterspeech, respectively. In contrast, GENERIC-EMPATHY and CONTEXT-WARNING counterspeech resulted in an average of 9.20 (+0.13) and 9.16 (+0.09) hateful posts, respectively. These results suggest that hostility increased among users who received either non-contextualized empathy-based counterspeech or contextualized LLM-generated warning-of-consequences counterspeech in the two weeks following the intervention.

• (3) *Relative change in toxicity*: The mean relative change in toxicity of users' posts within 2 weeks after our intervention for each experimental condition is shown in Fig. 2c. On average, toxicity increased by 3.44 % for users that did not receive any counterspeech (i.e., CONTROL CONDITION (=no counterspeech intervention)). Non-contextualized counterspeech led to a reduction in toxicity: Users that received GENERIC-EMPATHY (2.99 %) and GENERIC-WARNING (1.88 %) counterspeech are, on average, less toxic (−0.45 p.p., and −1.56 p.p. compared to the control, respectively). In contrast, LLM-generated counterspeech led to an increase in toxicity for CONTEXT-EMPATHY (9.74 %) and CONTEXT-WARNING (4.54 %) by, on average, +6.30 p.p. and +1.10 p.p. compared to the control. Overall, this suggests that LLM-generated counterspeech increases toxicity regardless of the counterspeech strategy.

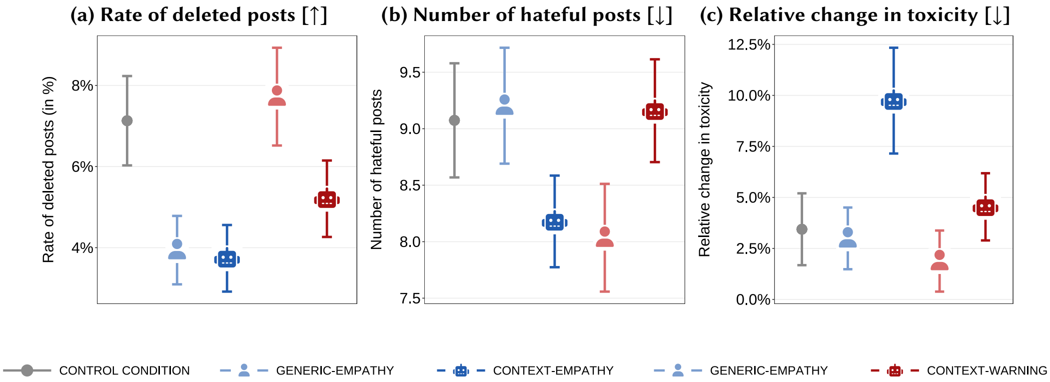

Fig. 2. Average (a) rate of deleted posts, (b) number of hateful posts after the intervention, and (c) relative change in toxicity and standard errors (bars) by experimental condition. [↑] ([↓]) indicates that a [positive] ([negative]) outcome is associated with an [increase] ([decrease]) in the outcome values.

2.2 Regression analysis

To statistically compare the effectiveness of the different interventions, we use a linear regression model. Our regression analysis is split in two: (1) First, we evaluate the effectiveness of the different counterspeech interventions compared to the CONTROL CONDITION (=no counterspeech intervention). (2) Second, we study the effectiveness of LLM-generated counterspeech compared to non-contextualized counterspeech.

As before, we start by estimating the treatment effect of our counterspeech compared to the CONTROL CONDITION (=no counterspeech intervention) across our three outcomes (see Fig. 3 and Supplementary Material 8.1):

• (1) *Rate of deleted posts:* Fig. 3a shows the treatment effects on the rate of users deleting their posts following our intervention vs. the control. In line with our descriptive analysis, empathy-based counterspeech negatively affects the likelihood of users deleting their hateful posts. All else equal, users who received GENERIC-EMPATHY and CONTEXT-EMPATHY counterspeech were, on average, 2.62 percentage points ($p = 0.055$) and 2.89 percentage points ($p = 0.0345$) less likely to delete their posts, respectively. We also observe a positive coefficient for GENERIC-WARNING, yet this effect is not statistically significant ($\theta = -0.83\,\%$; $p = 0.608$).

• (2) *Number of hateful posts:* The effect of counterspeech on the number of hateful posts shared by users after the intervention is shown in Fig. 3b. Compared to the control, we observe a negative coefficient for both non-contextualized and LLM-generated counterspeech across both strategies (i.e., both empathy and warning-of-consequences). This effect is statistically significant for GENERIC-WARNING, where users shared, all else equal, on average, 1.03 fewer hateful posts ($p = 0.022$) after receiving non-contextualized warning-of-consequences counterspeech.

• (3) *Relative change in toxicity:* Fig. 3c presents the estimated effects of counterspeech on the relative change in the toxicity of a user's posts. We do not observe a statistically significant effect of counterspeech on the relative change in toxicity across all experimental conditions compared to the control. However, the negative coefficients for generic-empathy and generic-warning, alongside the positive coefficients for context-empathy and context-warning, suggest a potential adverse effect of LLM-generated counterspeech.

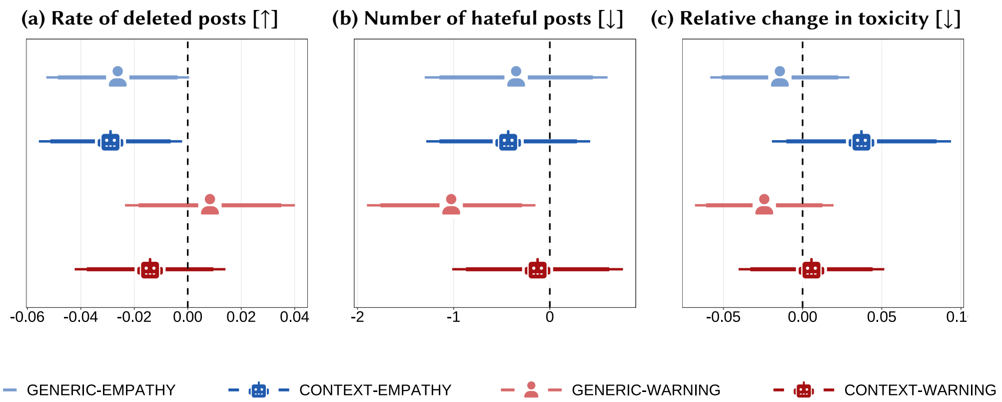

Fig. 3. Treatment effect of an intervention relative to the control condition (=no counterspeech intervention) for (a) *Rate of deleted posts*, (b) *Number of hateful posts*, and (c) *Relative change in toxicity*. Shown are the estimated coefficients from our linear regression model (symbol) as well as 95 % (thin), and 90 % (thick) confidence intervals. [↑] ([↓]) indicates that a [positive] ([negative]) outcome is associated with an [increase] ([decrease]) in the outcome values. Detailed estimation results for all coefficients are in Supplementary Material 8.1.

2.3 Deep-dive: contextualized vs. non-contextualized counterspeech

Our descriptive analysis revealed that contextualized counterspeech generated by LLMs led to worse outcomes compared to non-contextualized counterspeech. We thus use the linear regression model form above to statistically compare contextualized counterspeech vs. non-contextualized counterspeech across our outcome variables. The results are shown in Fig. 4 (see Supplementary Material 8.2 for detailed regression results).

• (1) *Rate of deleted posts:* When comparing non-contextualized to contextualized counterspeech for *rate of deleted posts*, we do not find any significant effects (see Fig. 4a). Nevertheless, the negative coefficients for context-empathy and context-warning indicate that contextualized counterspeech generated by LLMs may be less effective than non-contextualized counterspeech in reducing online hate speech.

• (2) *Number of hateful posts:* Here, we do not find a statistically significant difference for empathy, when comparing non-contextualized and contextualized counterspeech. However, we find a positive and statistically significant coefficient for context-warning ($p = 0.032$). Hence, all else equal, context-warning increased the number of hateful posts shared within the two weeks following the intervention, on average, by 0.84 posts compared to generic-warning. As such, contextualized warning-of-consequences increases online hostility compared to non-contextualized warning-of-consequences.

• (3) *Relative change in toxicity:* The treatment effect of contextualized vs. non-contextualized counterspeech is shown in Fig. 4c. We find a positive and statistically significant coefficient for context-empathy compared to generic-empathy ($p = 0.048$). All else equal, context-empathy led to an increase in toxicity by 2.80 percentage points, on average, compared to generic-empathy. While we also observe a positive coefficient for context-warning, this effect is not statistically significant at common significance thresholds.

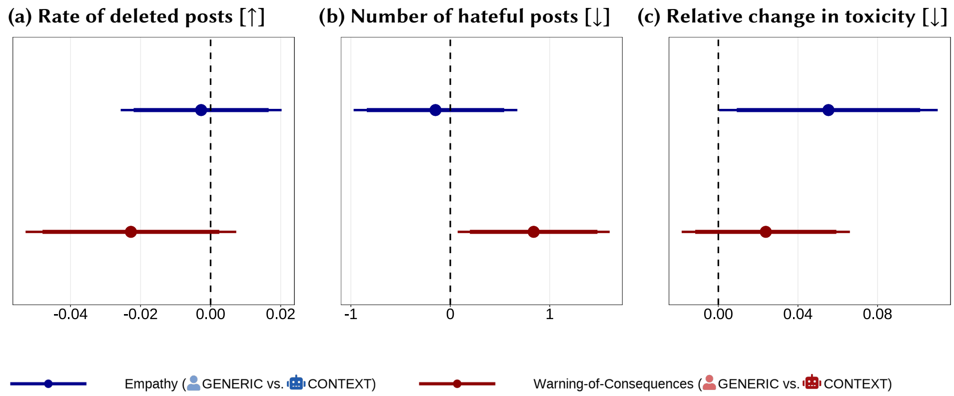

Fig. 4. Treatment effect of contextualized vs. non-contextualized counterspeech for (a) *Rate of deleted posts*, (b) *Number of hateful posts*, and (c) *Relative change in toxicity*. Shown are the estimated coefficients from our linear regression model (dot) measuring the relative effect of generic (Non-contextualized) vs. contextualized (Contextualized) counterspeech for the respective strategy as well as 95 % (thin), and 90 % (thick) confidence intervals. [↑] ([↓]) indicates that a [positive] ([negative]) outcome is associated with an [increase] ([decrease]) in the outcome values. Detailed estimation results for all coefficients are in Supplementary Material 8.2.

2.4 Additional analysis: Counterspeech for Twitter/X Premium users

Hateful users who subscribe to Twitter/X Premium are less likely to have their content removed by the platform, and their posts are algorithmically boosted [14]. Hence, we evaluate whether our intervention is effective for Twitter/X Premium users. To do so, we re-estimate the regression model from our main analysis, adding an interaction term between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise).

We find no significant interaction between our intervention and Twitter/X Premium subscription status for the *rate of deleted posts*. However, Twitter/X Premium users who received CONTEXT-EMPATHY counterspeech shared significantly more hateful posts ($p = 0.012$) than non-subscribers. Additionally, Premium users exhibited higher toxicity levels when receiving GENERIC-EMPATHY ($p = 0.049$) and CONTEXT-EMPATHY ($p = 0.024$) counterspeech. This suggests that empathetic counterspeech, particularly when LLM-generated, may backfire for Premium users. Importantly, all treatment effects remain consistent with our primary analysis across all models and dependent variables, except for CONTEXT-WARNING vs. GENERIC-WARNING and the *number of hateful posts*, which is no longer significant ($p = 0.107$). Detailed regression results are in Supplementary Material 8.3.

3 Discussion

Online hate speech poses a serious threat to societal cohesion and individual well-being [17, 28, 46, 48, 49, 59] and can even incite real-world violence [61]. Hence, curbing online hate speech is a crucial challenge for society. In this paper, we evaluate the effectiveness of contextualized counterspeech generated by an LLM in reducing online hate speech through a large-scale, pre-registered field experiment.

Our field experiment offers only limited evidence that counterspeech can significantly reduce online hate speech. While we find that GENERIC-WARNING leads to a slight but statistically significant reduction in the sharing of hateful posts compared to the control, we observe only weak directional evidence or even adversarial effects for other counterspeech strategies and outcomes. In particular, for both GENERIC-EMPATHY and CONTEXT-EMPATHY, we even see a significantly lower rate of deleted posts (and for Twitter/X Premium users even more hateful posts and increased toxicity), indicating a negative outcome.

Our results contrast with previous research reporting that counterspeech is effective [23, 37, 38, 49]. Given that our study design and non-contextualized messages are inspired by prior work [23], this discrepancy may be attributed to changes in the ecosystem of Twitter/X, which is reported to have become more hostile and toxic [14, 15]. This shift could make it increasingly difficult to persuade users to behave civilly, as they may face fewer repercussions for their actions. Another possible explanation is a lack of statistical power to detect small positive effects. However, given that our sample size ($N = 2,664$) is significantly larger than in previous studies [23, 37, 38, 49], this seems unlikely.

Our results even show that LLM-use may backfire: when comparing contextualized LLM-generated vs. non-contextualized counterspeech, we see that LLM-generated counterspeech is less effective in reducing online hate speech and may even increase hostility. For instance, context-warning led to significantly more hateful posts vs. generic-warning. Similarly, context-empathy resulted in greater toxicity than generic-empathy.

One possible explanation is that users often react negatively when recognizing LLM-generated content intended to convey empathy [41, 45]. Similarly, the identity of the messenger is crucial for counterspeech based on warning-of-consequences, which aims to reinforce social norms [49]. Given that people are more likely to recognize tailored LLM-generated texts [21], users may realize they are interacting with an LLM and thus might resist changing their behavior or feel deceived, which could lead to negative outcomes.

As with other research, ours is not free of limitations that offer opportunities for future work. For instance, our analysis is based on a large-scale, pre-registered field experiment conducted on Twitter/X, a platform often criticized for hosting hate speech and inadequately removing harmful content [14, 15]. While Twitter/X presents a challenging case, the effects of contextualized counterspeech generated by LLMs may differ across platforms, highlighting the need for future research to explore the potential of counterspeech in other online environments. Furthermore, we use LLama-3, a state-of-the-art open-source LLM developed by Meta [34], to generate contextualized counterspeech. This allows for reproducibility and accessibility [50]. Future research may also explore the use of proprietary models (e.g., GPT-4). Nevertheless, we experimented with proprietary models such as GPT-4 by Open AI but did not find qualitative differences in the counterspeech generated by Llama 3.

Our findings contribute to the literature on content moderation, specifically, counterspeech to curb hate speech on social media. Unlike previous studies that employed predefined, generic counterspeech messages [23, 37, 38, 49], we consider the importance of context in countering hate [36, 54]. Our approach uses LLMs to generate counterspeech tailored to individual hateful posts, aiming to promote civil online behavior. In doing so, we contribute to the ongoing debate on *when* LLMs can enhance persuasion [56, 62]. While LLMs have shown promise in mediating opposing groups [58], countering conspiracy theories [16], and fostering civil online conversations [6, 20], it was unclear whether LLMs could encourage more civil behavior through counterspeech. Our findings indicate that LLM-generated counterspeech is ineffective in promoting civil behavior and may even backfire, highlighting the need for further research into the conditions under which LLM-generated messages influence behavior effectively.

For platforms and policymakers, our results offer new insights into the role of LLMs in promoting online civility and highlight the need for caution when deploying LLM-driven societal interventions at scale. While counterspeech is promising in addressing hate speech, our findings suggest that LLM-generated interventions may be ineffective or even backfire. This indicates that relying on LLMs alone to foster behavioral change in online environments may yield limited results without a deeper understanding of the conditions under which LLM interventions are most persuasive. LLM-generated messages may thus need to be part of a broader strategy that includes repeated interventions or human moderation to have a meaningful impact. Additionally, given the potential for LLM-generated content to be perceived as inauthentic or deceiving [21], transparency and careful design are essential to prevent backlash. Therefore, the use of LLMs should be accompanied by thorough testing, ongoing evaluation, and flexibility to adjust strategies based on platform-specific dynamics and user behavior.

4 Materials and methods

To evaluate the effectiveness of contextualized counterspeech generated by an LLM to reduce hateful content on social media, we conducted a large-scale, pre-registered[2] field experiment on Twitter/X (www.X.com). In the following, we describe our (1) interventions, (2) experimental procedure, (3) study population, (4) statistical analysis, and (5) ethical considerations.

4.1 Interventions

Our experiment followed a 2x2 between-subjects design where, in addition, we included a control condition with no counterspeech. Overall, we thus randomly assigned users to one of five experimental conditions. Our experimental conditions are shown in Table 1.

Our messages for non-contextualized counterspeech are inspired by [23]. The messages either promote empathy (GENERIC-EMPATHY) or warning-of-consequences (GENERIC-WARNING). To avoid a strongly repetitive reply pattern that may irritate users, we used five different counterspeech messages for each non-contextualized condition that are qualitatively the same.[3] The exact messages are reported in Table S4.

To generate contextualized counterspeech, we used Llama-3 70B Chat[4], a state-of-the-art open-source LLM developed by Meta [1]. Depending on the condition, we prompt the model to generate counterspeech aimed at promoting empathy (CONTEXT-EMPATHY) or warning-of-consequences (CONTEXT-WARNING).[5] Our prompts followed best practices in prompt engineering [63] and prior research [6, 16, 21]. The exact prompts are in Table S2. Examples of contextualized counterspeech are in Table S3.

4.2 Procedure

Our experimental procedure is as follows (see Fig. 1): We sampled hateful users on Twitter/X by searching for hateful posts using a comprehensive list of keywords (see Table S1 for a full list) via the Twitter/X API v2.[6] We then manually filtered posts by users matching our keywords for hateful content. Note that this step could be automated. We opted for a manual validation for ethical considerations and, in particular, to comply with requirements from our Institutional Review Board (IRB), which allows us to ensure participants' safety (see our extensive discussion of ethical considerations in Sec. 4.5).

Subsequently, we retrieved the user profiles associated with each hateful post and filtered users according to pre-registered exclusion criteria (see Section 4.3 for details). The remaining users were then randomly assigned to one of the five experimental conditions.

[2] Pre-registration available at https://osf.io/38saz/?view_only=263687bff9824852b8ed204f257de8d5
[3] Throughout our paper, we label non-contextualized counterspeech with an emoji showing a "human" ( ) to indicate that these messages were crafted by human experts. However, all non-contextualized counterspeech messages are predefined and not customized to address specific posts.
[4] Model card: https://huggingface.co/meta-llama/Meta-Llama-3-70B
[5] LLMs are known for their ability to generate empathetic content [29]. However, it is unclear whether they can produce convincing warnings. To address this, we conducted an online survey with 500 participants recruited from Prolific www.prolific.com to evaluate perceived differences between human- and LLM-generated warnings. Our results indicate that LLM-generated and human-generated warnings are equally likely to be recognized as such.
[6] https://developer.twitter.com/en/products/twitter-api

Next, we assigned counterspeech to each user and replied to their hateful post. Of note, each user is only treated once. For users receiving non-contextualized counterspeech (i.e., generic-empathy or generic-warning), we randomly selected one of five predefined counterspeech messages based on the assigned strategy (see Table S4). For users receiving contextualized counterspeech (i.e., context-empathy or context-warning), we prompted Llama 3 to generate contextualized counterspeech using the corresponding prompt template (see Table S2).

We administered our intervention via multiple human-controlled accounts. The accounts were designed to appear politically neutral and natural to users on Twitter/X, which was inspired by the design in [23]. Each account was assigned a unisex English name, with no disclosure of gender, ethnicity, nationality, or beliefs. Furthermore, to appear as natural users, we regularly posted neutral posts via our accounts (e.g., *"Just witnessed the most breathtaking sunset!"*) and re-posted content from diverse accounts (e.g., NASA, WWF, ESPN). The accounts were created at least 3 months before the start of the experiment. Screenshots of example profiles are in Fig. S1.

Following our intervention, we monitored users' behavior on Twitter/X for two weeks to assess its effectiveness. Specifically, we analyzed the following three outcome variables (see Table S5 for summary statistics):

(1) ***Rate of deleted posts:*** A dichotomous variable that indicates whether a user deleted their original hateful post (= 1 if the post was deleted, = 0 otherwise) following our intervention. We later operationalize this by computing the average rate of deleted posts per experimental condition.
(2) ***Number of hateful posts:*** The number of hateful posts by a user after the intervention.
(3) ***Relative change in toxicity:*** The relative change in toxicity of a user's posts after the intervention.

Our outcome variables are designed to measure the effect of counterspeech on two distinct psychological processes: (a) whether users reconsidered their action and deleted the original tweet and (b) whether users changed their behavior and posted less hate speech but also adapted their tone and engaged in more civil conversations. We chose this approach since it should reflect users' recent activities on Twitter/X and their immediate reaction to our intervention. Details for each outcome are below:

• **Rate of deleted posts:** To measure whether a user deleted their original post, we queried the respective post via the Twitter/X API two weeks after our intervention. If the post was no longer available, we then used Twitter/X's compliance API endpoint to confirm whether the post had been actually deleted by the user (i.e., we do not count cases where the user has changed their privacy settings or was suspended by the platform).

• **Number of hateful posts:** To measure the number of hateful posts by users after our intervention, we collected up to the most recent 100 posts from the two weeks following the intervention. We then classified each post as hateful or not using Twitter-roBERTa-base-hate, which is trained on ≈58 million Twitter posts and fine-tuned for hate speech detection [10]. Eventually, we counted the number of hateful posts by each user.

• **Relative change in toxicity:** To measure the relative change in the toxicity of a user's posts following the intervention, we also collected up to 100 posts from the two weeks before the intervention. We then computed the average toxicity of posts shared before and after the intervention using Google's Perspective API [30], which is frequently used by previous research to study the toxicity of online content [8, 9, 66] and yields a toxicity score $\in [0, +1]$.

4.3 Study population

We recruited $N = 2,778$ users who posted hateful content on Twitter/X on weekdays between June 7 and July 26, 2024. Users are identified following the procedure described above. Specifically, we searched for hateful posts and retrieved the associated user profiles. Subsequently, each user was randomly assigned to one of the experimental conditions and received the corresponding treatment. Importantly, each user is treated only once.

Following our pre-registration, we excluded users based on the following criteria: (1) We excluded verified accounts (i.e., we excluded organizations with a golden checkmark). (2) We only considered users that posted original posts (i.e., we excluded retweets and replies). (3) We excluded users that had been inactive (i.e., they had shared fewer than 5 posts in the past 7 days) to ensure sufficient data to compare later if users had changed their behavior after our intervention. (4) We excluded users likely to be bot accounts.

As per our pre-registration, we have excluded 99 accounts that altered their privacy settings or were suspended by Twitter/X during the experiment. To check for differential attrition rates between treatment and control conditions, we used a χ^2-test [40]. We found no significant differences in attrition rates between treatment and control groups ($p > 0.1$). Following our debrief, we have excluded 15 users who opted out of the study (see section 4.5 for details). Overall, we have thus excluded 114 accounts and, eventually, have $N = 2,664$ accounts in the subsequent analysis.

4.4 Statistical analysis

To statistically compare the effectiveness of the different interventions, we use a linear regression model. Our unit of analysis is users who posted hate speech. Let y_i denote one of our three outcome variables, namely, (1) the *rate of deleted posts*, (2) the *number of hateful posts*, and (3) the *relative change in toxicity*, for user i. Let t_i denote the intervention received by user i, and let x_i refer to a vector of different characteristics belonging to that user (described later). We then estimate the following linear regression model

$$y_i = \alpha + \theta t_i + \beta^T x_i + \varepsilon_i, \tag{1}$$

where α represents the model intercept, θ measures the effect of the intervention t_i, β captures the effect of all control variables in x_i on y_i, and ε_i is the error term. For estimation, we use ordinary least squares regression (OLS) with robust standard errors to account for heteroskedasticity in the error term due to variations in our intervention and control variables. We test whether the coefficients are significantly different from zero using two-sided t-tests.

Our analysis is split into two: (1) We evaluate the effectiveness of counterspeech compared to the CONTROL CONDITION (=no counterspeech intervention). (2) We study the effectiveness of contextualized counterspeech generated by an LLM compared to non-contextualized counterspeech. Both are as follows:

(1) *Intervention vs. control:* To evaluate the effectiveness of each intervention compared to the control, we separately estimate the linear regression model described above for each type of counterspeech. Specifically, we set $t_i = 1$ if a user received a counterspeech reply and $t_i = 0$ for users assigned to the control condition.

(2) *Contextualized vs. non-contextualized:* We compare the effectiveness of contextualized counterspeech generated by an LLM to non-contextualized counterspeech. To do so, we re-estimate the linear regression model from above but set $t_i = 1$ if a user received contextualized counterspeech and $t_i = 0$ if a user received non-contextualized counterspeech. We perform this comparison separately for each counterspeech strategy. Hence, we estimate one model to compare CONTEXT-EMPATHY vs. GENERIC-EMPATHY and a separate model to compare CONTEXT-WARNING vs. GENERIC-WARNING.

For each regression model described above, we estimate three versions, each with a different outcome variable. Thus, we estimate separate models for (1) the *rate of deleted posts*, (2) the *number of hateful posts*, and (3) the *relative change in toxicity*. The analysis was implemented in R 4.4.1. using the `stats` and `lmtest` packages.

Following our pre-registration, we included a set of pre-treatment covariates to account for variability in the outcome explained by pre-treatment covariates. Specifically, we included a user's account age (in days), follower count, following count, tweet count, like count, and whether the user subscribes to Twitter/X Premium (= 1 if premium, = 0 otherwise) as indicated by a blue checkmark on a user's profile page. We further included the number of hateful posts shared by a user in the two weeks before the intervention and the average toxicity of posts shared by a user in the two weeks before the intervention. To classify whether a post is hateful, we again used the Twitter-roBERTa-base model [10]. Lastly, we included the average toxicity of a user's posts shared within two weeks before the intervention, measured by Google's Perspective API [30]. Of note, we only collected up to 100 posts before our interventions, which should reflect the recent activities of users on Twitter/X. Summary statistics are in Table S5.

Robustness checks: To ensure the robustness of our results, we conducted a series of checks: (1) We re-estimated our analysis using a single model that included separate dummy variables for each intervention, instead of estimating separate models for each combination of treatment and control. (2) We pooled all observations in the treatment conditions to assess the overall effect of counterspeech compared to the control group. (3) We pooled observations based on the counterspeech strategies (i.e., empathy vs. warning-of-consequences) and re-estimated our regression model to evaluate their overall effects compared to the control. (4) To compare the effectiveness of contextualized and non-contextualized counterspeech, we pooled all users assigned to contextualized counterspeech across both strategies and repeated the analysis. All robustness checks led to consistent findings. Details are in Supplementary Material 9.

4.5 Ethics

Ethics approval (EK-MIS-2024-254) for the field experiment was obtained from the Institutional Review Board at the Faculty of Mathematics, Informatics, and Statistics at LMU Munich. This ethics approval complies with regulations for studies involving human participants at LMU Munich. The experimental task, data collection, and analysis closely follow related works involving counterspeech on social media [23, 37, 38]. Our study solely relies on publicly available data and follows common guidelines for ethical research with social media [44]. We only report aggregated and anonymized results to protect users' privacy.

Ethical considerations were of utmost importance for our study. Following previous studies [23, 37, 38] and ethical guidelines on experimental research on social media [35], we designed our interventions to be minimally invasive and socially beneficial. Specifically, our interventions are designed to mitigate hate speech while preserving users' right to free expression.

To minimize ethical risks and protect the well-being of all participants, we have further implemented a detailed experimental protocol that includes comprehensive safety measures. This includes explicit guidelines for continuous human monitoring and specific countermeasures. Our experimental protocol ensures (a) the appropriateness of our interventions and (b) immediate actions to guarantee the safety of all participants. For example, we manually ensured that each counterspeech conveyed an appropriate tone, avoided biases, and was culturally sensitive.

Of note, users provide informed consent to receive public replies when they agree to the terms of service at Twitter/X when signing up for the platform [65]. Users agree that they may receive replies from other users when engaging on Twitter/X (Section 3, Twitter/X Terms of Service [65]). As such, our intervention fully complies with users' informed consent to the use of Twitter/X and aligns with the platform's goal to combat hate speech without infringing users' rights to freedom of expression [65].

Our study was carefully designed to ensure users' privacy. Specifically, our study was compliant with the General Data Protection Regulation (GDPR) of the European Union. We have implemented measures to repeatedly enforce the privacy settings of all participants by frequently calling the Twitter/X Compliance API to check if users have changed their privacy settings and delete private data accordingly. We further abide by the privacy regulations of GDPR and ensure that users can fully opt out of data collection. To do so, we have debriefed users following our experiment. Our debrief included detailed information on the goals, methods, and interventions of our study. We have further informed participants about their privacy rights concerning GDPR and provided contact addresses for questions. We sent our debrief collectively at the end of the experiment and granted users an extended period to request additional information on the study or have their data removed.

References

[1] AI @ Meta, Llama Team. 2024. The Llama 3 herd of models. *arXiv* 2407.21783 (2024).

[2] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2019. Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. In *CSCW*.

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

[4] Amalia Álvarez-Benjumea and Fabian Winter. 2020. The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. *PNAS* 117, 37 (2020), 22800–22804.

[5] Amnesty International. 2022. Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya. https://www.amnesty.org/en/documents/ASA16/5933/2022/en/

[6] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *PNAS* 120, 41 (2023), e2311627120.

[7] Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Workshop on Online Abuse and Harms*.

[8] Michele Avalle, Niccolò Di Marco, Gabriele Etta, Emanuele Sangiorgio, Shayan Alipour, Anita Bonetti, Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli, Matteo Cinelli, and Walter Quattrociocchi. 2024. Persistent interaction patterns across social media platforms and over time. *Nature* 628, 8008 (2024), 582–589.

[9] Dominik Bär, N. Pröllochs, and Stefan Feuerriegel. 2023. Finding Qs: Profiling QAnon supporters on Parler. In *ICWSM*.

[10] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *EMNLP Findings*.

[11] Susan Benesch, Derek Ruths, Kelly P. Dillon, Haji Mohammad Salem, and Lucas Wright. 2016. Considerations for successful counterspeech. *Dangerous Speech Project* (2016). https://www.dangerousspeech.org/libraries/considerations-for-successful-counterspeech

[12] Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Workshop on CounterSpeech for Online Abuse*.

[13] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech Countering. In *EMNLP*.

[14] Center for Countering Digital Hate. 2023. Twitter fails to act on 99% of Twitter Blue accounts tweeting hate. https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/

[15] Center for Countering Digital Hate. 2023. X content moderation failure: How Twitter/X continues to host posts we reported for extreme hate speech. https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/

[16] Thomas H. Costello, Gordon Pennycook, and David Gertler Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385 (2024), eadq1814.

[17] Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. Online hate speech victimization: Consequences for victims' feelings of insecurity. *Crime Science* 13, 1 (2024).

[18] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative AI. *Business & Information Systems Engineering* 66, 1 (2024), 111–126.

[19] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020).

[20] Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-driven mediation strategies for audience depolarisation in online debates. In *CHI*.

[21] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *PNAS* 121, 24 (2024), e2403116121.

[22] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. In *CSCW*.

[23] Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *PNAS* 118, 50 (2021), e2116310118.

[24] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *WWW*.

[25] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *PNAS* 120, 11 (2023), e2208839120.

[26] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation. In *CHI*.

[27] Sander L. Koole, Nils B. Jostmann, and Nicola Baumann. 2023. From intentions to action: An integrative review of action control theory and research. In *Advances In Experimental Social Psychology*. Advances in Experimental Social Psychology, Vol. 68. Elsevier, 291–375.

[28] Cinoo Lee, Kristina Gligorić, Pratyusha Ria Kalluri, Maggie Harrington, Esin Durmus, Kiara L. Sanchez, Nay San, Danny Tse, Xuan Zhao, MarYam G. Hamedani, Hazel Rose Markus, Dan Jurafsky, and Jennifer L. Eberhardt. 2024. People who share encounters with racism are silenced online by

humans and machines, but a guideline-reframing intervention holds promise. *PNAS* 121, 38 (2024), e2322764121.

[29] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models produce responses perceived to be empathic. *arXiv* 2403.18148 (2024).

[30] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of Perspective API: Efficient multilingual character-level transformers. In *KDD*.

[31] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2024. The virality of hate speech on social media. In *CSCW*.

[32] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *ICWSM*.

[33] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14, 1 (2024), 4692.

[34] Meta. 2024. Build the future of AI with Meta Llama 3.

[35] Mohsen Mosleh, Gordon Pennycook, and David G. Rand. 2022. Field experiments on social media. *Current Directions in Psychological Science* 31, 1 (2022), 69–75.

[36] Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *EMNLP Findings*.

[37] Kevin Munger. 2017. Tweetment effects on the Tweeted: Experimentally reducing racist harassment. *Political Behavior* 39, 3 (2017), 629–649.

[38] Kevin Munger. 2021. Don't @ me: Experimentally reducing partisan incivility on Twitter. *Journal of Experimental Political Science* 8, 2 (2021), 102–116.

[39] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*.

[40] Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50, 302 (1900), 157–175.

[41] Anat Perry. 2023. AI will never convey the essence of human empathy. *Nature Human Behaviour* 7, 11 (2023), 1808–1809.

[42] Jakub Podolak, Szymon Łukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bąkowicz, and Piotr Sankowski. 2024. LLM generated responses to mitigate the impact of hate speech. In *EMNLP Findings*.

[43] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *ICWSM*.

[44] Caitlin M. Rivers and Bryan L. Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014), 38.

[45] Matan Rubin, Joanna Z. Li, Federico Zimmerman, Desmond C. Ong, Amit Goldenberg, and Anat Perry. 2024. The value of perceiving a human response: Comparing perceived human versus AI-generated empathy. *OSF Preprints* ng97s (2024).

[46] Koustuv Saha, Eshwar Chandrasekharan, and Munmun de Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *WebSci*. 255–264.

[47] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.

[48] Alexandra A. Siegel. 2020. Online hate speech. In *Social media and democracy*, Nathaniel Persily and Joshua A. Tucker (Eds.). Cambridge University Press, Cambridge, 56–88.

[49] Alexandra A. Siegel and Vivienne Badaan. 2020. #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review* 114, 3 (2020), 837–855.

[50] Arthur Spirling. 2023. Why open-source generative AI models are an ethical way forward for science. *Nature* 616, 7957 (2023), 413.

[51] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis)informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850.

[52] Andreas Spitz, Ahmad Abu-Akel, and Robert West. 2021. Interventions for softening can lead to hardening of opinions: Evidence from a randomized controlled trial. In *WWW*.

[53] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators. In *CHI*.

[54] Derald Wing Sue, Sarah Alsaidi, Michael N. Awad, Elizabeth Glaeser, Cassandra Z. Calle, and Narolyn Mendez. 2019. Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *The American Psychologist* 74, 1 (2019), 128–142.

[55] Adam Taylor. 2019. New Zealand suspect allegedly claimed 'brief contact' with Norwegian mass murderer Anders Breivik. *The Washington Post* (2019). https://www.washingtonpost.com/world/2019/03/15/new-zealand-suspect-allegedly-claimed-brief-contact-with-norwegian-mass-murderer-anders-breivik/

[56] Jacob D. Teeny and Sandra C. Matz. 2024. We need to understand "when" not "if" generative AI can enhance personalized persuasion. *PNAS* 121, 43 (2024), e2418005121.

[57] Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: A comparative study. In *ACL Findings*.

[58] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in

democratic deliberation. *Science* 386, 6719 (2024), eadq2852.

[59] Brendesha M. Tynes, Michael T. Giang, David R. Williams, and Geneene N. Thompson. 2008. Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health* 43, 6 (2008), 565–569.

[60] United Nations. 2019. UN strategy and plan of action on hate speech. https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech

[61] United Nations. 2023. Countering and addressing online hate speech: A Guide for policy makers and practitioners. https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf

[62] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* (2024).

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

[64] Christopher Y. K. Williams, Brenda Y. Miao, Aaron E. Kornblith, and Atul J. Butte. 2024. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications* 15, 1 (2024), 8236.

[65] X. 2024. Terms of Service. https://x.com/en/tos

[66] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions. In *CSCW*.

[67] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *WWW*.

Data availability

To ensure reproducibility, we provide the anonymized data used for the analysis via OSF at https://osf.io/2nhsm/?view_only=5c628dd0e4fa4c309a6a2b0c8dfc69a8.

Code availability

All code to replicate our analyses is available via OSF at https://osf.io/2nhsm/?view_only=5c628dd0e4fa4c309a6a2b0c8dfc69a8.

Author contributions

DB, AM, and SF contributed to conceptualization. DB conducted the experiment. DB performed the data analysis. DB, AM, and SF contributed to results interpretation and manuscript writing. All authors approved the manuscript.

Funding information

Funding by the German Research Foundation (Grant: 543018872) and the Civic Health Project (www.civichealthproject.org) is acknowledged.

Competing interests

The authors declare no competing interests.

Supplementary materials

A Supplementary materials: Emotions in online rumor diffusion

Supplementary Materials

Emotions in Online Rumor Diffusion

Nicolas Pröllochs,*[1] Dominik Bär,[2] Stefan Feuerriegel,[2,3]

[1]University of Giessen, Giessen, 35394, Germany
[2]LMU Munich, Munich, 80539, Germany
[3]ETH Zurich, Zurich, 8092, Switzerland

*E-mail: nicolas.proellochs@wi.jlug.de

S1 Background Literature

Rumor Theory

In social psychology literature, rumors are defined as the spread of a piece of content from person to person without confirmation of its veracity [1, 2]. Hence, rumors refer to content that might be true, false, or a mix thereof. The concept of rumors has been employed in computational social science with an analogous definition; see Friggeri et al. (2014) [3] or Vosoughi et al. (2018) [4].

In offline settings, the diffusion of rumors has been subject to research. Offline rumors are often perceived as important and therefore spread both extensively [2] and quickly [5]. One of the primary reasons for which rumors are propagated between people is that rumors typically fulfill an emotional need [1]. In particular, rumors with a strong polarization serve motivational aspects of human behavior, as they potentially allow people to justify or legitimize their feelings [2]. This raises the question concerning to what extent rumor spreading is driven by emotions.

The role of emotions in rumor diffusion has been studied previously, albeit only for offline rumors. Offline rumors have a higher chance of dissemination if they convey negative emotions [1]. A laboratory study has further revealed similar effects for different emotional categories, such as perceived uncertainty and anxiety [6, 7]. However, the role of emotions in online rumor diffusion has heretofore remained unknown.

Diffusion of Online Content

Diffusion of online content has been the focus of an extensive stream of literature [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Here different objectives are of interest, including predictive modeling [10, 22, 18, 23, 24, 25, 26] and explanatory inferences. Due to the wealth of studies, we provide only a brief summary in the following and, in particular, concentrate on explanatory studies as this is also the scope of the present work.

Several properties have been identified along which online diffusion dynamics can be characterized. *Cascade size* counts the number of involved users and thus reflects the overall audience [8, 15, 16, 19, 20, 27]. The likelihood of observing a cascade is inversely linked to its size, for which reason cascades with a large size are rare [13]. *Cascade lifetime* measures the timespan between the initial broadcast and the last sharing, thereby indicating how long a cascade was active [8, 15, 16, 19, 20, 27]. A longer lifetime is usually associated with posts that generate more attention and are thus regarded as being of particular relevance. To measure virality, an aggregated metric has been developed, called *structural virality* [13]. Structural virality is based on the Wiener index and combines both the depth and breadth of cascades. Later, the aforementioned characteristics represent the dependent variables of our regression analysis.

Drivers of online diffusion have been located primarily in attributes of the sender, that is, her social influence. In this regard, social influence is typically quantified by the number of followers (i. e., outgoing ties) or followees (i. e., incoming ties) [9, 14, 15, 26]. Here users with more ties are assumed to reach a larger audience and thus initiate cascades of larger size. Other determinants are, for instance, past levels of user engagement (e. g., posts, sharings, or likes) [4].

Diffusion dynamics have been studied with a specific focus on online rumors. Some works compare the spread of rumors vs. non-rumors [28, 29, 30, 31] or science vs. conspiracy [32]. Other works report summary statistics from past rumor cascades [3, 4, 33] or make statistical inferences between social influence and sharing probability [4]. However, none of the works provides statistical inferences that explain the structure of rumor cascades in terms of embedded emotions.

Emotions in Online Diffusion

Emotions, defined as the response to environmental stimuli, play an important role in online behavior [34]. Emotions can be mined for the purpose of sensing public mood [35] and, more importantly, are known to guide the way in which users engage in information collection, information processing, judgment, and decision-making. According to psychology, emotions also promote social interactions such as offline information sharing [36, 37, 38]. Analogous to emotional exchanges in offline environments, emotional states of users are highly contagious and are propagated through social media [39, 40, 41]. However, this stream of literature studies the emotions of users but not emotions embedded in social media content, as in this work.

Prior literature offers explanatory evidence concerning how emotions impact the diffusion of online content, yet with clear differences from the present study. First, it is common that emotions are studied through simpler constructs. Examples are arousal [42, 39], sentiment (i. e., positive/negative valence) [43, 44, 8], or discrete emotions [45]. However, a comprehensive analysis involving derived emotions (i. e., dyadic emotional interactions) is missing. Second, existing literature builds upon a variety of dependent variables such as the likelihood of email forwarding [42], response time [43], sharing counts [39], or sharing probability [43]. With the exception of size [8, 45], the dependent variables refer solely to individual-level sharing behavior and not the structure of cascades. Third, the studies comprise specific content such as New York Times articles [42], political content [39, 43], humanitarian crises [44, 45], or fake news [46], but **not** rumors.

One work [4] reports summary statistics on emotions that are embedded in online rumors. However, the results are purely descriptive and therefore do not allow for statistical inferences of how emotions are associated with diffusion dynamics.

Research gap: None of the above references has analyzed the link between emotions and online rumor diffusion through statistical inferences. To fill this void, we perform a large-scale regression analysis based on which we estimate the effect of emotions embedded in rumors on the structure of their cascade.

References

[1] Knapp, R.H.: A psychology of rumor. The Public Opinion Quarterly **8**(1), 22–37 (1944)

[2] Allport, G.W., Postman, L.: The Psychology of Rumor. Henry Holt, New York, NJ (1947)

[3] Friggeri, A., Adamic, L.A., Eckles, D., Cheng, J.: Rumor cascades. In: International AAAI Conference on Web and Social Media (ICWSM) (2014)

[4] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). doi:10.1126/science.aap9559

[5] Shibutani, T.: Improvised News: A Sociological Study of Rumor. Bobbs-Merrill, Indianapolis, IN (1966)

[6] Anthony, S.: Anxiety and rumor. Journal of Social Psychology **89**(1), 91–98 (1973). doi:10.1080/00224545.1973.9922572

[7] Rosnow, R.L.: Inside rumor: A personal journey. American Psychologist **46**(5), 484–496 (1991)

[8] Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer. In: International Conference on Web Search and Data Mining (WSDM) (2011). doi:10.1145/1935826.1935845

[9] Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the Flickr social network. In: International World Wide Web Conference (WWW) (2009)

[10] Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: International World Wide Web Conference (WWW) (2014). doi:10.1145/2566486.2567997

[11] Cheng, J., Kleinberg, J., Leskovec, J., Liben-Nowell, D., Adamic, L., Subbian, K.: Do diffusion protocols govern cascade growth? In: International AAAI Conference on Web and Social Media (ICWSM) (2018)

[12] Goel, S., Watts, D.J., Goldstein, D.G.: The structure of online diffusion networks. In: ACM Conference on Electronic Commerce (EC) (2012). doi:10.1145/2229012.2229058

[13] Goel, S., Anderson, A., Hofman, J., Watts, D.J.: The structural virality of online diffusion. Management Science **62**(1), 180–196 (2016). doi:10.1287/mnsc.2015.2158

[14] Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: International World Wide Web Conference (WWW) (2010). doi:10.1145/1772690.1772751

[15] Lerman, K., Ghosh, R.: Information contagion: An empirical study of spread of news on Digg and Twitter social networks. In: International AAAI Conference on Web and Social Media (ICWSM) (2010)

[16] Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Transactions on the Web **1**(1), 5 (2007). doi:10.1145/1232722.1232727

[17] Macskassy, S.A., Michelson, M.: Why do people retweet? Anti-homophily wins the day! In: International AAAI Conference on Web and Social Media (ICWSM) (2011)

[18] Myers, S.A., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2012). doi:10.1145/2339530.2339540

[19] Myers, S.A., Leskovec, J.: The bursty dynamics of the twitter information network. In: International World Wide Web Conference (WWW) (2014). doi:10.1145/2566486.2568043

[20] Taxidou, I., Fischer, P.M.: Online analysis of information diffusion in Twitter. In: International Conference on World Wide Web (WWW) Companion (2014). doi:10.1145/2567948.2580050

[21] Wu, F., Huberman, B.A.: Novelty and collective attention. Proceedings of the National Academy of Sciences **104**(45), 17599–17601 (2007). doi:10.1073/pnas.0704916104

[22] Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., Kustarev, A.: Prediction of retweet cascade size over time. In: ACM International Conference on Information and Knowledge Management (CIKM) (2012). doi:10.1145/2396761.2398634

[23] Subbian, K., Prakash, B.A., Adamic, L.: Detecting large reshare cascades in social networks. In: International World Wide Web Conference (WWW) (2017). doi:10.1145/3038912.3052718

[24] Weng, L., Menczer, F., Ahn, Y.-Y.: Virality prediction and community structure in social networks. Scientific Reports **3**, 2522 (2013). doi:10.1038/srep02522

[25] Wu, S., Tan, C., Kleinberg, J., Macy, M.: Does bad news go away faster? In: International AAAI Conference on Web and Social Media (ICWSM) (2011)

[26] Zaman, T., Fox, E.B., Bradlow, E.T.: A Bayesian approach for predicting the popularity of tweets. The Annals of Applied Statistics **8**(3), 1583–1611 (2014). doi:10.1214/14-AOAS741

[27] Zang, C., Cui, P., Song, C., Faloutsos, C., Zhu, W.: Quantifying structural patterns of information cascades. In: International Conference on World Wide Web (WWW) Companion (2017). doi:10.1145/3041021.3054214

[28] Bessi, A., Coletto, M., Davidescu, G.A., Scala, A., Caldarelli, G., Quattrociocchi, W.: Science vs conspiracy: Collective narratives in the age of misinformation. PLOS ONE **10**(2), 0118093 (2015). doi:10.1371/journal.pone.0118093

[29] Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: International World Wide Web Conference (WWW) (2011). doi:10.1145/1963405.1963500

[30] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences of the United States of America **113**(3), 554–559 (2016). doi:10.1073/pnas.1517441113

[31] Nekovee, M., Moreno, Y., Bianconi, G., Marsili, M.: Theory of rumor spreading in complex social networks. Physica A: Statistical Mechanics and its Applications **374**(1), 457–470 (2007). doi:10.1016/j.physa.2006.07.017

[32] Zollo, F., Novak, P.K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., Quattrociocchi, W.: Emotional dynamics in the age of misinformation. PLOS ONE **10**(9), 0138740 (2015). doi:10.1371/journal.pone.0138740

[33] Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Wu, J., Havlin, S.: Fake news propagates differently from real news even at early stages of spreading. EPJ Data Science **9**(1), 3035 (2020). doi:10.1140/epjds/s13688-020-00224-z

[34] Zhang, P.: The affective response model: A theoretical framework of affective concepts and their relationships in the ICT context. MIS Quarterly **37**(1), 247–274 (2013)

[35] Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: International AAAI Conference on Web and Social Media (ICWSM) (2011)

[36] Luminet, O., Bouts, P., Delie, F., Manstead, A.S.R., Rimé, B.: Social sharing of emotion following exposure to a negatively valenced situation. Cognition & Emotion **14**(5), 661–688 (2000). doi:10.1080/026999300501176666

[37] Rimé, B.: Emotion elicits the social sharing of emotion: Theory and empirical review. Emotion Review **1**(1), 60–85 (2009). doi:10.1177/1754073908097189

[38] Peters, K., Kashima, Y., Clark, A.: Talking about others: Emotionality and the dissemination of social information. European Journal of Social Psychology **39**(2), 207–222 (2009). doi:10.1002/ejsp.523

[39] Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences (PNAS) **114**(28), 7313–7318 (2017). doi:10.1073/pnas.1618923114

[40] Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences (PNAS) **111**(24), 8788–8790 (2014). doi:10.1073/pnas.1320040111

[41] Goldenberg, A., Gross, J.J.: Digital emotion contagion. Trends in Cognitive Sciences **24**(4), 316–328 (2020). doi:10.1016/j.tics.2020.01.009

[42] Berger, J., Milkman, K.L.: What makes online content viral? Journal of Marketing Research **49**(2), 192–205 (2012). doi:10.1509/jmr.10.0353

[43] Stieglitz, S., Dang-Xuan, L.: Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. Journal of Management Information Systems **29**(4), 217–248 (2013). doi:10.2753/MIS0742-1222290408

[44] Zeng, L., Starbird, K., Spiro, E.S.: Rumors at the speed of light? Modeling the rate of rumor transmission during crisis. In: Hawaii International Conference on System Sciences (HICSS) (2016). doi:10.1109/HICSS.2016.248

[45] Li, L., Wang, Z., Zhang, Q., Wen, H.: Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters. Information Processing & Management **57**(6), 102313 (2020). doi:10.1016/j.ipm.2020.102313

[46] Chuai, Y., Zhao, J.: Anger makes fake news viral online. arXiv (2004.10399) (2020)

B Supplementary materials: Emotions explain differences in the diffusion of true vs. false social media rumors

Supplementary Materials

Emotions Explain Differences in the Diffusion of True vs. False Social Media Rumors

Nicolas Pröllochs,*[1] Dominik Bär,[2] Stefan Feuerriegel,[2,3]

[1]JLU Giessen, Giessen, 35394, Germany
[2]LMU Munich, Munich, 80539, Germany
[3]ETH Zurich, Zurich, 8092, Switzerland

*E-mail: nicolas.proellochs@wi.jlug.de.

Tables S1 – S5

Table S1: Regression results for sentiment. The dependent variables are cascade size (column 1), cascade duration (column 2), and structural virality (column 3). Rumor-specific random effects are included.

	Size	**Duration**	**Virality**
Intercept	3.710***	4.001***	0.742***
	(0.101)	(0.109)	(0.015)
Falsehood	0.006	−0.305*	−0.021
	(0.110)	(0.119)	(0.016)
User Engagement	−0.184***	−0.587***	0.035***
	(0.027)	(0.036)	(0.005)
Verified Account	0.848***	0.754***	0.156***
	(0.056)	(0.063)	(0.009)
Account Age	−0.180***	−0.232***	−0.003
	(0.015)	(0.018)	(0.002)
Followers	0.755***	0.235***	0.002
	(0.041)	(0.016)	(0.002)
Followees	0.227***	0.131***	0.017***
	(0.022)	(0.016)	(0.002)
Sentiment	−0.165***	−0.100*	−0.014**
	(0.040)	(0.042)	(0.005)
Falsehood × Sentiment	0.479***	0.319***	0.047***
	(0.044)	(0.047)	(0.006)
Rumor-Specific Random Effects	Yes	Yes	Yes
AIC	182,185	71,979	32,108

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

2

Table S2: Regression results for bipolar emotion pairs. The dependent variables are cascade size (column 1), cascade duration (column 2), and structural virality (column 3). Rumor-specific random effects are included.

	Size	**Duration**	**Virality**
Intercept	3.725***	4.023***	0.746***
	(0.101)	(0.110)	(0.015)
Falsehood	−0.007	−0.327**	−0.025
	(0.110)	(0.120)	(0.016)
User Engagement	−0.188***	−0.588***	0.035***
	(0.027)	(0.036)	(0.005)
Verified Account	0.840***	0.751***	0.155***
	(0.056)	(0.063)	(0.009)
Account Age	−0.179***	−0.232***	−0.003
	(0.015)	(0.018)	(0.002)
Followers	0.751***	0.235***	0.002
	(0.041)	(0.016)	(0.002)
Followees	0.223***	0.130***	0.017***
	(0.022)	(0.016)	(0.002)
JoySadness	0.000	−0.006	0.000
	(0.033)	(0.036)	(0.004)
TrustDisgust	−0.189***	−0.051	−0.007
	(0.043)	(0.045)	(0.005)
AngerFear	−0.020	−0.080	−0.004
	(0.046)	(0.047)	(0.005)
AnticipationSurprise	−0.105*	−0.131*	−0.023***
	(0.047)	(0.056)	(0.006)
Falsehood × JoySadness	0.013	0.047	0.001
	(0.039)	(0.041)	(0.005)
Falsehood × TrustDisgust	0.402***	0.205***	0.031***
	(0.047)	(0.049)	(0.006)
Falsehood × AngerFear	0.212***	0.200***	0.020***
	(0.050)	(0.052)	(0.006)
Falsehood × AnticipationSurprise	0.339***	0.285***	0.051***
	(0.050)	(0.060)	(0.007)
Rumor-Specific Random Effects	Yes	Yes	Yes
AIC	182,147	72,013	32,072

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

3

Table S3: Regression results with basic emotions. The dependent variable is cascade size. We fit separate regression models, each including one of the 8 basic emotions. This estimation procedure is adopted because basic emotions sum to 1 and are thus subject to a linear dependency. Rumor-specific random effects are included.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	3.610^{***}	3.604^{***}	3.621^{***}	3.637^{***}	3.635^{***}	3.608^{***}	3.618^{***}	3.662^{***}
	(0.100)	(0.004)	(0.100)	(0.100)	(0.101)	(0.100)	(0.100)	(0.092)
Falsehood	0.136	0.149^{***}	0.131	0.101	0.107	0.144	0.126	0.085
	(0.109)	(0.004)	(0.110)	(0.109)	(0.110)	(0.109)	(0.109)	(0.101)
User Engagement	-0.197^{***}	-0.202^{***}	-0.203^{***}	-0.197^{***}	-0.198^{***}	-0.202^{***}	-0.197^{***}	-0.201^{***}
	(0.027)	(0.004)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.031)
Verified Account	0.896^{***}	0.896^{***}	0.896^{***}	0.872^{***}	0.879^{***}	0.894^{***}	0.894^{***}	0.894^{***}
	(0.056)	(0.004)	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)	(0.055)
Account Age	-0.172^{***}	-0.178^{***}	-0.179^{***}	-0.182^{***}	-0.177^{***}	-0.178^{***}	-0.179^{***}	-0.180^{***}
	(0.015)	(0.004)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Followers	0.766^{***}	0.774^{***}	0.770^{***}	0.767^{***}	0.763^{***}	0.774^{***}	0.773^{***}	0.771^{***}
	(0.041)	(0.004)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.014)
Followees	0.235^{***}	0.242^{***}	0.242^{***}	0.239^{***}	0.231^{***}	0.242^{***}	0.242^{***}	0.244^{***}
	(0.022)	(0.004)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.014)
Anger	0.014							
	(0.039)							
Falsehood × Anger	0.181^{***}							
	(0.044)							
Fear		-0.039^{***}						
		(0.004)						
Falsehood × Fear		-0.011^{**}						
		(0.004)						
Anticipation			-0.072^{*}					
			(0.035)					
Falsehood × Anticipation			0.116^{**}					
			(0.041)					
Trust				-0.076^{*}				
				(0.031)				
Falsehood × Trust				0.313^{***}				
				(0.038)				
Surprise					0.083			
					(0.069)			
Falsehood × Surprise					-0.286^{***}			
					(0.071)			
Sadness						-0.026		
						(0.033)		
Falsehood × Sadness						0.141^{***}		
						(0.039)		
Joy							-0.041	
							(0.031)	
Falsehood × Joy							0.165^{***}	
							(0.038)	
Disgust								0.244^{***}
								(0.036)
Falsehood × Disgust								-0.333^{***}
								(0.039)
Rumor-Specific Random Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	182,393	182,484	182,484	182,359	182,337	182,462	182,458	182,449

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S4: Regression results with basic emotions. The dependent variable is cascade lifetime. We fit separate regression models, each including one of the 8 basic emotions. This estimation procedure is adopted because basic emotions sum to 1 and are thus subject to a linear dependency. Rumor-specific random effects are included.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	3.946***	3.946***	3.957***	3.955***	3.982***	3.941***	3.944***	3.952***
	(0.108)	(0.108)	(0.109)	(0.108)	(0.110)	(0.108)	(0.108)	(0.109)
Falsehood	−0.224	−0.221	−0.233*	−0.244*	−0.264*	−0.217	−0.227	−0.232
	(0.118)	(0.118)	(0.119)	(0.118)	(0.119)	(0.118)	(0.118)	(0.119)
User Engagement	−0.590***	−0.591***	−0.591***	−0.590***	−0.589***	−0.590***	−0.588***	−0.589***
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Verified Account	0.779***	0.778***	0.779***	0.771***	0.771***	0.779***	0.777***	0.779***
	(0.063)	(0.063)	(0.063)	(0.063)	(0.063)	(0.063)	(0.063)	(0.063)
Account Age	−0.229***	−0.229***	−0.230***	−0.231***	−0.229***	−0.229***	−0.229***	−0.230***
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Followers	0.236***	0.236***	0.236***	0.236***	0.235***	0.236***	0.236***	0.236***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Followees	0.134***	0.135***	0.135***	0.134***	0.133***	0.135***	0.136***	0.135***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Anger	−0.045							
	(0.039)							
Falsehood × Anger	0.111*							
	(0.044)							
Fear		0.038						
		(0.057)						
Falsehood × Fear		−0.086						
		(0.060)						
Anticipation			−0.051					
			(0.041)					
Falsehood × Anticipation			0.077					
			(0.046)					
Trust				−0.036				
				(0.035)				
Falsehood × Trust				0.202***				
				(0.040)				
Surprise					0.140			
					(0.073)			
Falsehood × Surprise					−0.270***			
					(0.075)			
Sadness						0.019		
						(0.037)		
Falsehood × Sadness						0.033		
						(0.042)		
Joy							0.005	
							(0.036)	
Falsehood × Joy							0.100*	
							(0.041)	
Disgust								0.039
								(0.051)
Falsehood × Disgust								−0.096
								(0.055)
Rumor-Specific Random Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	72,092	72,097	72,101	72,033	72,051	72,097	72,079	72,095

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S5: Regression results with basic emotions. The dependent variable is structural virality. We fit separate regression models, each including one of the 8 basic emotions. This estimation procedure is adopted because basic emotions sum to 1 and are thus subject to a linear dependency. Rumor-specific random effects are included.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	0.716***	0.715***	0.719***	0.716***	0.720***	0.715***	0.715***	0.719***
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Falsehood	0.016	0.017	0.012	0.014	0.010	0.016	0.015	0.012
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
User Engagement	0.034***	0.034***	0.034***	0.034***	0.034***	0.034***	0.035***	0.034***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Verified Account	0.167***	0.166***	0.166***	0.165***	0.164***	0.166***	0.166***	0.166***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Account Age	−0.001	−0.001	−0.001	−0.002	−0.001	−0.001	−0.001	−0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Followers	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Followees	0.016***	0.016***	0.016***	0.016***	0.016***	0.016***	0.017***	0.016***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Anger	−0.005							
	(0.004)							
Falsehood × Anger	0.013**							
	(0.005)							
Fear		−0.007						
		(0.006)						
Falsehood × Fear		0.003						
		(0.006)						
Anticipation			−0.014**					
			(0.004)					
Falsehood × Anticipation			0.020***					
			(0.005)					
Trust				0.000				
				(0.004)				
Falsehood × Trust				0.022***				
				(0.005)				
Surprise					0.017*			
					(0.008)			
Falsehood × Surprise					−0.041***			
					(0.009)			
Sadness						0.003		
						(0.004)		
Falsehood × Sadness						0.015**		
						(0.005)		
Joy							0.002	
							(0.004)	
Falsehood × Joy							0.014**	
							(0.005)	
Disgust								0.013*
								(0.006)
Falsehood × Disgust								−0.022***
								(0.006)
Rumor-Specific Random Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	32,267	32,274	32,263	32,194	32,158	32,222	32,233	32,259

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

6

A User Studies to Validate Dictionary Approach

Our results rely on the validity of dictionaries to extract emotions from online rumors. We thus check how (i) perceived sentiment aligns with the lexicon-based sentiment and (ii) how perceived emotions align with the lexicon-based emotion scores. For this, we conducted two user studies using the online survey platform Prolific (`https://www.prolific.co/`). For both studies, we randomly sampled 100 rumors from Twitter and presented them to $n = 7$ participants (English native speakers).
The participants of the first study were asked to rate the sentiment conveyed in each tweet on a Likert scale from -3 to $+3$ (here: -3 indicates negative sentiment, while $+3$ refers to a positive sentiment). The participants exhibited a statistically significant interrater agreement according to Kendall's W ($p < 0.01$). Furthermore, we compute the correlation between the human labels and the lexicon-based sentiment score. We found Spearman's correlation coefficient to be positive ($r_s = 0.11$) and statistically significant ($p < 0.01$). In sum, the results add to the validity of our lexicon-based approach. The lexicon-based approach should thus largely match the perceived sentiment in online rumors.
In the second study, the participants were instructed to rate the presence of the eight basic emotions on a Likert scale from -3 to $+3$ (here: -3 indicates no emotion present while $+3$ refers to a high degree of emotion present) for each tweet. As shown in the following table, the participants exhibited a statistically significant interrater agreement according to Kendall's W for each of the 8 basic emotions ($p < 0.01$). Overall, the correlation between the dictionary-based emotion scores and human annotations is $r_s = 0.13$ ($p < 0.01$) and thus statistically significant at common significance thresholds. This demonstrates that our dictionary approach is able to capture emotions in online rumors.

Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
0.474***	0.198***	0.427***	0.406***	0.364***	0.408***	0.227***	0.230***

Kendall's W coefficient for the interrater agreement between survey participants.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B Additional Robustness Checks

B.1 Analysis of Annual Effects

Our sample contains a comprehensive set of fact-checked rumors from Twitter during the time period from its founding in 2006 through 2017. We validate the robustness of our results with respect to different time periods by incorporating dummy variables for each year of our observation period. The estimation results with year-level effects are shown in Table S6 and Table S7. The coefficients of all variables are in good agreement and support the robustness of our results across time periods. These findings suggest that the observed effects of language classified by (i) sentiment and (ii) emotions on size, lifetime, and structural virality should generalize well.

Table S6: Regression results for sentiment. The dependent variables are cascade size (column 1), cascade duration (column 2), and structural virality (column 3). Rumor-specific random effects are included. Year-level effects are included.

	Size	Duration	Virality
Intercept	-0.279	0.673	-0.068
	(1.304)	(0.965)	(0.330)
Falsehood	0.032	-0.308^{**}	-0.020
	(0.107)	(0.118)	(0.016)
User Engagement	-0.218^{***}	-0.598^{***}	0.029^{***}
	(0.027)	(0.036)	(0.005)
Verified Account	0.901^{***}	0.812^{***}	0.161^{***}
	(0.056)	(0.063)	(0.009)
Account Age	-0.194^{***}	-0.247^{***}	-0.005^{*}
	(0.015)	(0.018)	(0.002)
Followers	0.715^{***}	0.228^{***}	0.001
	(0.040)	(0.016)	(0.002)
Followees	0.209^{***}	0.128^{***}	0.016^{***}
	(0.022)	(0.016)	(0.002)
Sentiment	-0.159^{***}	-0.094^{*}	-0.014^{**}
	(0.040)	(0.042)	(0.005)
Falsehood × Sentiment	0.459^{***}	0.304^{***}	0.046^{***}
	(0.044)	(0.046)	(0.006)
Rumor-Specific Random Effects	Yes	Yes	Yes
Year-Level Effects	Yes	Yes	Yes
AIC	181,873	71,806	32,003

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table S7: Regression results for bipolar emotion pairs. The dependent variables are cascade size (column 1), cascade duration (column 2), and structural virality (column 3). Rumor-specific random effects are included. Year-level effects are included.

	Size	Duration	Virality
Intercept	-0.249	0.683	-0.079
	(2.163)	(0.965)	(0.309)
Falsehood	0.025	-0.329^{**}	-0.002
	(0.099)	(0.119)	(0.020)
User Engagement	-0.220^{***}	-0.599^{***}	0.030^{***}
	(0.030)	(0.036)	(0.005)
Verified Account	0.897^{***}	0.809^{***}	0.165^{***}
	(0.054)	(0.063)	(0.009)
Account Age	-0.193^{***}	-0.247^{***}	-0.004
	(0.015)	(0.018)	(0.002)
Followers	0.708^{***}	0.228^{***}	0.001
	(0.014)	(0.016)	(0.002)
Followees	0.206^{***}	0.127^{***}	0.015^{***}
	(0.014)	(0.016)	(0.002)
JoySadness	-0.003	-0.005	0.000
	(0.026)	(0.036)	(0.004)
TrustDisgust	-0.177^{***}	-0.047	-0.007
	(0.032)	(0.045)	(0.005)
AngerFear	-0.018	-0.079	-0.005
	(0.032)	(0.047)	(0.005)
AnticipationSurprise	-0.106^{**}	-0.125^{*}	-0.023^{***}
	(0.037)	(0.056)	(0.006)
Falsehood × JoySadness	0.019	0.038	0.000
	(0.030)	(0.041)	(0.005)
Falsehood × TrustDisgust	0.384^{***}	0.195^{***}	0.029^{***}
	(0.035)	(0.049)	(0.006)
Falsehood × AngerFear	0.194^{***}	0.196^{***}	0.019^{**}
	(0.035)	(0.051)	(0.006)
Falsehood × AnticipationSurprise	0.327^{***}	0.275^{***}	0.049^{***}
	(0.040)	(0.059)	(0.006)
Rumor-Specific Random Effects	Yes	Yes	Yes
Year-Level Effects	Yes	Yes	Yes
AIC	181,284	71,839	31,958

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

9

Analysis of Emotional Uniformity

We calculate the sum of squares over the 8-dimensional vector comprising the different emotion scores. This metric provides a measure of emotional uniformity by attributing higher scores to rumors that have lower diversity among emotion classifications. For instance, if a rumor contains words associated with only one emotion (e. g., anger) and no other emotion, then it receives a high score. In contrast, a lower score is assigned if language is associated with all emotions to a similar extent. The regression estimates are shown in Table S8. We find that a higher level of uniformity across emotion scores (i. e., lower diversity) is associated with smaller values for cascade size, duration, and structural virality.

Table S8: Regression results for emotional uniformity. The dependent variables are cascade size (column 1), cascade duration (column 2), and structural virality (column 3). Rumor-specific random effects are included.

	Size	**Duration**	**Virality**
Intercept	3.498***	3.837***	0.719***
	(0.098)	(0.107)	(0.014)
Falsehood	0.162	−0.198	−0.006
	(0.107)	(0.116)	(0.016)
User Engagement	−0.171***	−0.575***	0.036***
	(0.027)	(0.036)	(0.005)
Verified Account	0.793***	0.700***	0.146***
	(0.056)	(0.062)	(0.009)
Account Age	−0.179***	−0.235***	−0.003
	(0.015)	(0.018)	(0.002)
Followers	0.740***	0.234***	0.001
	(0.040)	(0.016)	(0.002)
Followees	0.218***	0.129***	0.016***
	(0.021)	(0.015)	(0.002)
Emotional Uniformity	−0.355***	−0.327***	−0.052***
	(0.012)	(0.018)	(0.002)
Rumor-Specific Random Effects	Yes	Yes	Yes
AIC	181,781	71,767	31,673

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

10

B.2 Analysis of Rumors With Mixed Veracity

In our main analysis, we focused on rumors that can clearly be determined as "true" or "false." However, some rumors were categorized as being of "mixed" veracity by the fact-checking organizations (where the veracity could not be clearly designated as true or false). As an exploratory analysis, we analyze differences in the spreading dynamics of mixed rumors. Due to data limitations, we re-estimate our model for a smaller set of independent variables, i. e., only the control variables. The regression results for cascade size, cascade duration, and cascade virality are shown in Table S9. Compared to true rumors, we find that rumor cascades of mixed veracity exhibit a smaller cascade size and virality but spread over a longer time horizon.

Table S9: Regression results for mixed rumors. The dependent variables are cascade size (column 1), cascade duration (column 2), and structural virality (column 3).

	Size	**Duration**	**Virality**
Intercept	1.662***	1.732***	0.177***
	(0.012)	(0.028)	(0.002)
Falsehood	1.240***	0.285***	0.058***
	(0.013)	(0.031)	(0.003)
Mixed	−0.096***	0.200***	−0.021***
	(0.017)	(0.043)	(0.003)
User Engagement	−0.193***	−0.262***	0.077***
	(0.005)	(0.008)	(0.001)
Verified Account	1.553***	1.753***	0.439***
	(0.036)	(0.049)	(0.007)
Account Age	−0.147***	−0.177***	0.014***
	(0.005)	(0.010)	(0.001)
Followers	2.940***	0.168***	0.030***
	(0.005)	(0.006)	(0.001)
Followees	0.567***	0.195***	0.089***
	(0.005)	(0.007)	(0.001)
Rumor-Specific Random Effects	No	No	No
AIC	761,187	182,629	86,614

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

C Supplementary materials: Systematic discrepancies in the delivery of political ads on Facebook and Instagram

Systematic discrepancies in the delivery of political ads on Facebook and Instagram
Supplemental Materials

DOMINIK BÄR[*†], LMU Munich, Munich Center for Machine Learning, Munich, Germany
FRANCESCO PIERRI[*], Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Milan, Italy
GIANMARCO DE FRANCISCI MORALES, CENTAI, Turin, Italy
STEFAN FEUERRIGEL, LMU Munich, Munich Center for Machine Learning, Munich, Germany

[*]Both authors contributed equally to this research.
[†]To whom correspondence should be addressed. E-mail: baer@lmu.de; Mail: LMU Munich, Geschwister-Scholl-Platz 1, 80539 München

Table of contents

1 RELATED WORK

Social media has led to a major shift in political advertising [21, 22, 33]. Due to the large user base of social media platforms, advertisers are able to run campaigns with wide reach at comparatively low costs [21]. This capability helps campaigns with smaller budgets and may democratize elections by fostering electoral competition [21, 22]. However, the use of social media for political advertising has also introduced new challenges. For example, political ads on social media tend to be more partisan compared to traditional forms like television ads [21]. Additionally, far-right and populist parties, known for promoting anti-democratic narratives, appear to benefit from advertising on social media [10] and even use it to spread misinformation [9]. Given that political advertising was shown to influence voter turnout and vote choice [1, 15, 26], it is crucial to audit political advertising on social media in order to ensure the fairness, accountability, and transparency of electoral processes.

An important benefit of advertising on social media is targeting. Targeting allows advertisers to craft customized messages directed at specific user groups, which makes it particularly valuable for political parties during election campaigns [21]. In fact, targeted advertising has been shown to be effective in various settings outside of elections. Examples include increasing conversion rates for products [29], promoting public health measures to reduce COVID-19 infection rates [6], and shifting views on climate change [24]. Furthermore, previous research has shown that parties strategically tailor social media ads to target specific demographic groups [20]. However, the prevalence of targeted political ads as well as the detailed targeting strategies across the political spectrum remain unclear. To fill this gap, we provide a comprehensive analysis of targeted advertising on social media

Auditing online political advertising follows a long tradition of scrutiny in our democratic process [3, 18, 19]. A major concern is that hyper-personalization could lead to political filter bubbles and echo chambers [3, 13, 23]. Moreover, concerns have been raised that targeted advertising may discriminate against parts of the electorate [34] or use personally sensitive data (e.g., ethnic origin, sexual orientation) to identify receptive audiences [8]. In fact, previous research has shown that the algorithmic delivery of ads discriminates against women [27], which may reinforce existing stereotypes and harm political participation. Overall, this underlines the need for a better understanding of political advertising on social media (as a first step to developing regulatory frameworks).

There have been community-driven efforts to independently monitor political ads online and thus improve the transparency of political advertising [28, 32]. These systems often rely on data donations [32], or volunteers to audit political advertising on social media platforms [28]. As such, they are independent from platforms and offer a transparent view on online political advertising. However, such efforts are unable to monitor political advertising at scale, are limited to information that is publicly available on the platforms (e.g., cannot capture ad spending), and are biased toward the community and thus not representative. To address these issues, we leverage a novel dataset that offers in-depth insights into targeting strategies for political ads.

Following public pressure and regulatory efforts, platforms have started to release internal data that records political ads at scale and provides comprehensive insights into advertising behavior beyond the scope of community-driven systems, including information about actual spending and real-world impressions. For example, the Meta Ad Library provides public access to all political ads published on Facebook and Instagram. Researchers have used these resources to study how politicians advertise on climate change [2] and immigration [11, 12], address Spanish vs. English-speaking audiences during the 2020 U.S. election [14], and analyze political ads by populist and mainstream parties during the 2019 European elections [10] and the 2022 Italian election [30]. However, a significant gap exists in our understanding of the targeting strategies employed in political advertising. To close this gap, our study offers the first analysis of *targeting* for political advertising on social media.

Given the importance of algorithmic ad delivery for targeting, researchers and policymakers are worried that safeguarding the integrity of the democratic process is now in the hands of commercial actors, "who may have differing understandings of fundamental democratic norms" [17]. Beyond researchers and policymakers, a large part of the public is also concerned about targeted political advertising: according to the Pew Research Center, more than half of the adult U.S. population finds that social media platforms should ban political advertising, and more than three-quarters find that targeting for political campaigns is not acceptable [4]. Therefore, auditing is crucial to ensure fairness, accountability, and transparency in electoral processes.

2 THE 2021 GERMAN FEDERAL ELECTION

On September 26, 2021, more than 60 million Germans were called to elect a new parliament (called "Bundestag") [7]. The election marked a turning point in German politics as the previous chancellor, Angela Merkel, did not stand for re-election after 16 years [36].

In Germany, each voter casts two votes: The *first vote* ("Erststimme") selects a preferred candidate as a constituency representative, where a majority vote determines who secures a seat in the parliament. The *second vote* ("Zweitstimme") is for a political party, for which seats in the parliament are allocated based on each party's share of second votes. The dual-vote system in Germany ensures both individual representation and party influence. Furthermore, due to the dual-vote system, election campaigns in Germany have to simultaneously appeal to the local concerns of voters and broader national issues. While candidates may run their own campaigns, overarching party campaigns typically dominate [25].

The German multi-party system has six major political parties that compete for voters across the political spectrum [16][1]: (1) ***Linke*** (*The Left*) is a democratic socialist party located on the political left that promotes progressive social and economic policies. (2) ***Grüne*** (*Alliance 90/The Greens*) is located at the center-left of the political spectrum with a strong emphasis on environmental topics and social equality, because of which the party is particularly popular among young urban voters. (3) ***SPD*** (*The Social Democratic Party of Germany*) is also located at the center-left of the political spectrum and traditionally focuses on social equality. (4) ***FDP*** (*The Free Democratic Party*) is a liberal party located in the center of the political spectrum advocating a liberal and market-oriented agenda that gained large support from young voters. (5) ***Union*** (The *Union*) is the main center-right party in Germany. It is a coalition of the *CDU* (*Christian Democratic Union*), which operates in all federal states except Bavaria, and the *CSU* (*Christian Social Union*, which only operates in Bavaria). The *Union* promotes conservative values and supports a market-oriented economy with a balanced approach to social policies and Christian values. (6) ***AfD*** (*The Alternative for Germany*) was founded in 2013 and is a far-right party with a strong focus on immigration and public security. It is popular among an older and male-dominated electorate. The *AfD* has been a source of contention in German society, with criticisms highlighting its anti-immigrant rhetoric, affiliations with far-right extremism, and tendencies towards historical revisionism. The party's divisive stances and oversimplification of complex issues from a populist perspective have led to a situation where the *AfD* is politically isolated from the other main parties.

[1]There are further candidates running for smaller parties or as independents. However, smaller parties and independents play only a minor role given their limited resources and particularities of the German electoral system in that parties need a second vote share of at least 5 % or win three constituencies to gain seats in the parliament. This, in turn, means that it is intentionally made highly unlikely for smaller parties and independents to enter the parliament. Hence, we focus on political ads by the six main parties—*Linke*, *Grüne*, *SPD*, *FDP*, *Union*, and *AfD*—throughout our paper. These parties have also been part of the parliament in the legislative period before the 2021 election.

3 POLITICAL ADVERTISING ON META DURING THE 2021 GERMAN FEDERAL ELECTION

Our data comprises $N = 81\,549$ ads with an overall cost of EUR 9.8 million for more than 1.1 billion impressions. Generally, larger parties ran a larger number of ads. However, we also see two exceptions: the winning party (*SPD*) was not particularly active (~ 16k ads), while one of the smaller parties (*Grüne*) ran a disproportionately high number of ads (over 39k). In Tbl. S1, we show a breakdown in terms of ads, money, and impressions for all parties.

Table S1. Breakdown of our dataset in terms of the total number of ads, money spent, and impressions generated by each party. Money and impressions data come in brackets; for closed ranges, we consider the average of the endpoints of the range, and, for open-ended ranges, we take the known closed endpoint.

Party	Number of ads	Spending (EUR)	Impressions
Linke	2257	390 950	75 087 865
Grüne	38 604	3 608 103	205 728 143
SPD	12 525	1 526 741	169 188 312
FDP	10 327	1 369 329	319 367 015
Union	15 349	2 337 002	299 763 160
AfD	2487	577 429	87 361 955
Total	81 549	9 809 553	1 156 496 450

Fig. S1 shows the top-10 targeting criteria employed by different parties in terms of spending.

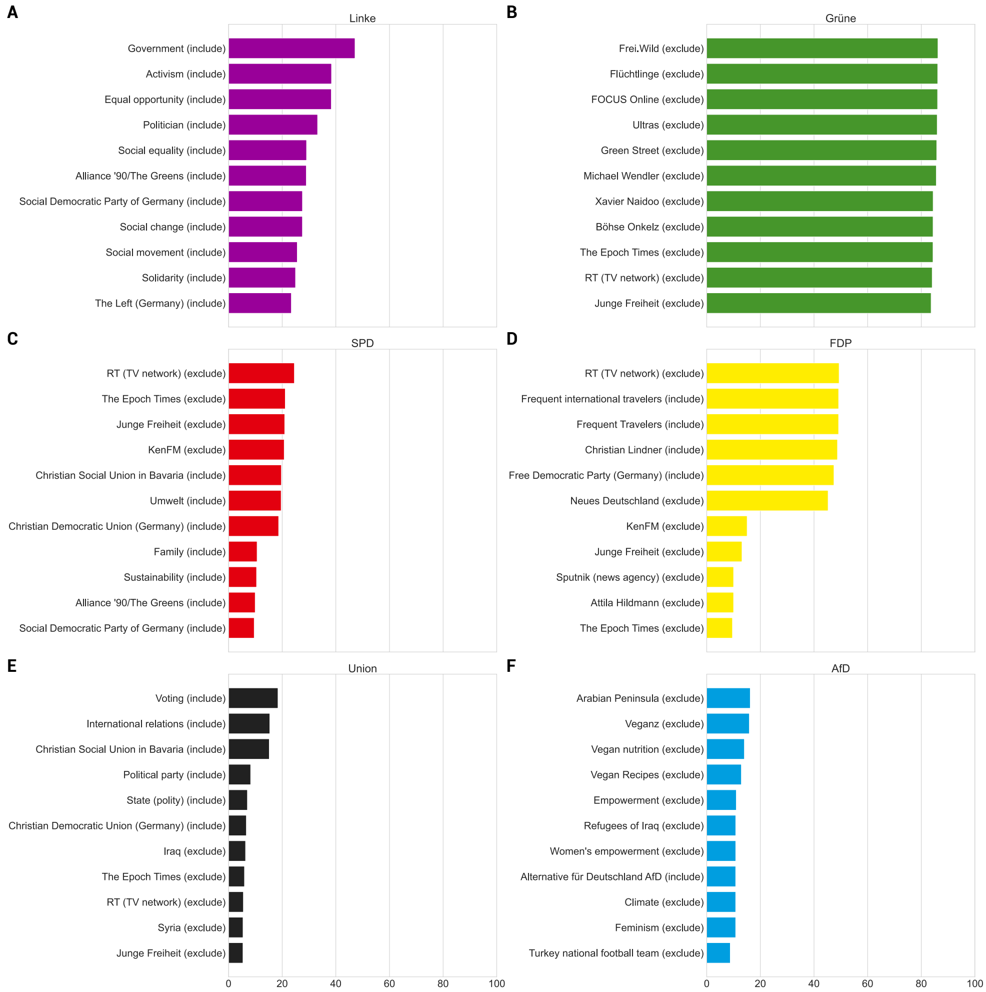

Fig. S1. Top-10 targeting criteria employed by a *Linke*, b *Grüne*, c *SPD*, d *FDP*, e *Union*, and f *AfD* in terms of overall spending (in EUR).

4 GOALS OF POLITICAL ADVERTISING ON SOCIAL MEDIA

Political campaigning follows various purposes [5, 31, 35, 37]. To explore the purpose of political ads on social media, we conduct an additional qualitative analysis. Specifically, we manually analyzed and coded a random sample of $N = 100$ ads for their purpose. Inspired by political science literature [5, 31, 35, 37], we distinguish ads that mobilize, persuade, promote events, or call for donations. Furthermore, we distinguish whether an ad is informational (i.e., purely focusing on explaining a policy) or attacking opponents.

Our analysis shows that a large majority of ads are designed to mobilize (50 %) or persuade (31 %) voters of a political party or candidate. Few ads are informational (9 %) or promote events (8 %). We find only 2 ads attacking another party, while no ad is calling for donations. The latter is not surprising since parties in Germany are mostly funded publicly or via membership fees and would not call for donations during electoral campaigns.

Overall, our analysis suggests that political ads on social media are designed for various purposes but predominantly focus on mobilization and persuasion of voters.

5 POLITICAL ADVERTISING BY PARTIES VS. CANDIDATES

In the German voting system, voters cast two votes: The *first vote* ("Erststimme") selects a preferred candidate as a constituency representative and the *second vote* ("Zweitstimme") is for a political party (see Section 2 for details). As such, political campaigns may focus on parties more broadly but also specific candidates.

To check whether the ads in our sample are focusing on parties more broadly or tend to focus on specific candidates we ran two additional analysis. First, we classify ads by whether they have been published on pages run by candidates. This follows the rationale that ads published by candidates are likely to be tailored to the specific candidate. In contrast, parties are likely to advertise more broadly. We collected all candidate names for the 2021 German federal election from the Federal Returning Officer [7] and matched the candidate names with the page name of an ad. We find that 18135 ads (i.e., 22.24 % of all ads in our sample) are published by candidates and thus likely to specifically focus on a candidate. Second, since it is likely that parties sometimes also specifically advertise for a candidate, we employed a qualitative approach and manually classified a random sample of $N = 100$ ads to check whether ads tend to be focused on specific candidates or parties broadly. We find that 59 % of ads are focused more broadly on a party while 41 % tend to focus on a specific candidate. We further find that all ads published on a candidate page are also focusing on the specific candidate thus corroborating our first analysis.

Overall, we find that ads published by parties tend to be focused more broadly on the party while ads published by candidates tend to focus on the specific candidates.

6 THE ROLE OF COMPETITION ON IMPRESSIONS-PER-EUR

We conducted two additional analyses to study the role of competition. Due to the absence of other variables on competition at Meta, we focus on two variables: the timing-to-election-day and the number of competing political ads.

For the first analysis, we re-run the regression model for ad characteristics from our main analysis but included an additional variable measuring the time between publication and election day. The results are in Fig. S2. We find a positive and statistically significant coefficient for *Time to election (days)* ($p < 0.05$), suggesting that publishing ads earlier in the campaign is related to higher levels of impressions-per-EUR. All else equal, ads published one standard deviation of days earlier receive, on average, 6.26 additional impressions-per-EUR. This may be due to less competition for political audiences at the beginning of the campaign peri

Our first analysis indicates that publishing ads earlier during the campaign is linked to higher levels of impressions-per-EUR. While this may be related to lower competition, it could also be due to other idiosyncrasies pertaining to the specific time an ad was published unrelated to competition. Hence, we also studied how the total number of active political ads (across all parties) on the day an ad was published is linked to impressions-per-EUR as a more appropriate measure of competition. To do so, we again re-run the regression model for ad characteristics from our main analysis but now included an additional variable measuring the number of active ads on the publishing day of an ad. The results are in Fig. S3. We find a negative and statistically significant coefficient for *No. Active Ads* ($p < 0.05$), indicating that higher competition is related to less impressions-per-EUR. All else equal, a one-standard-deviation increase in competing political ads ($\approx$ 4000 ads) is related to a decrease of 10.45 impressions-per-EUR.

Of note, for both analyses, all other results remain consistent with our main analysis except for the coefficient of negative sentiment, which is no longer statistically significant.

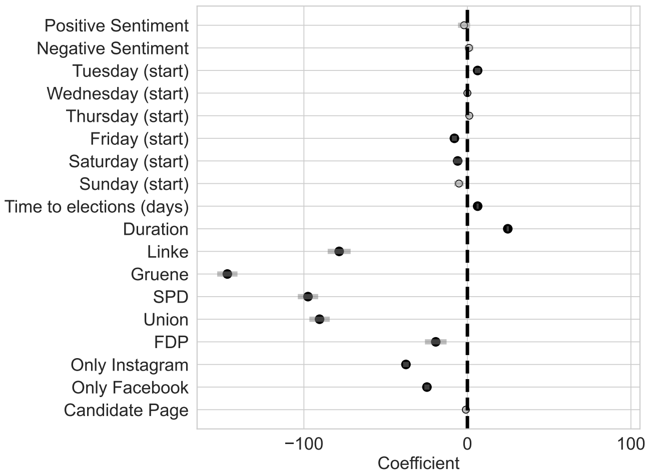

Fig. S2. Coefficient estimates and 95 % confidence intervals for ad characteristics including the time to election in days (*Time to elections (days)*). Statistically significant coefficients ($p < 0.05$) are indicated by black circles ●, all others by gray circles ○.Sentiment, weekday, platform dummy, and party dummy are categorical variables, and the reference categories are "neutral" sentiment, "Monday", "both platforms", and "AfD".

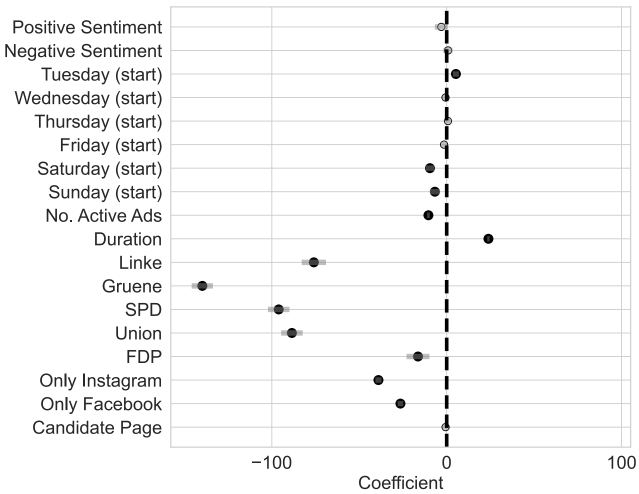

Fig. S3. Coefficient estimates and 95 % confidence intervals for ad characteristics including the the number of active ads at the day an ad was first published (*No. Active Ads*). Statistically significant coefficients ($p < 0.05$) are indicated by black circles ●, all others by gray circles ○. Sentiment, weekday, platform dummy, and party dummy are categorical variables, and the reference categories are "neutral" sentiment, "Monday", "both platforms", and "AfD".

7 META AD LIBRARY

The Meta Ad Library provides the variables as described in Tbl. S2 for "ads about social issues, election or politics that were delivered anywhere in the world during the past seven years". We refer the reader to the API documentation (https://www.facebook.com/ads/library/api/) for further details.

Table S2. List of variables provided in the Meta Ad Library API.

Creation time
Text
Link descriptions, captions and titles
Start and stop time
Snapshot URL
Sponsor name
Regional distribution
Demographic distributions (age and gender)
Estimated audience size
Impressions
Amount spent (EUR)
Platform(s) on which the ad is published

8 META TARGETING CATEGORIES

Meta allows advertisers to include or exclude users based on a set of targeting categories to reach specific user groups on their platforms Facebook and Instagram. The full list of available targeting categories is shown in Tbl. S3. For a more detailed description, we refer readers to https://developers.facebook.com/docs/fort-ads-targeting-dataset/table-schema. Fig. S4 provides an example ad with targeting information as provided by Meta.

Table S3. Targeting categories available to advertisers on Meta. Advertisers can include or exclude users based on the targeting categories.

Include/exclude interest
Include/exclude industry
Include/exclude parents
Include/exclude job title
Include/exclude employer
Include/exclude behavior
Include/exclude field of study
Include/exclude life event
Include/exclude school
Include/exclude education level
Include/exclude relationship status
Include/exclude income
Include/exclude undergrad years
Include/exclude custom audience
Include/exclude lookalike audience
Include/exclude connection
Include/exclude friend connection

Fig. S4. Example of an ad with targeting information as provided by Meta.

9 HYPERPARAMETER TUNING

For the training of the classifiers reported in the main paper, we use 10-fold cross-validation in combination with a grid search. In particular, we vary (1) the number of trees used for a forest (*N estimators*), (2) the number of variables to consider at each split (*Max features*), (3) maximum depth of the tree (*Max depth*), (4) minimum samples to split a node (*Min node*), (5) minimum samples in a leaf (*Min leaf*), and (6) whether to bootstrap samples when building trees (*Bootstrap*). The corresponding tuning range for each parameter is shown in Supplementary Tbl. S4.

Table S4. Tuning range of the grid search for random forest (RF) hyperparameters. n refers to the total number of variables available for training.

Hyperparameter	Tuning Range
N estimators	[100, 200, 300]
Max features	[$\sqrt{n}$, $\log_2 n$]
Max depth	[None, 10, 20, 30]
Min node	[2, 5, 10]
Min leaf	[1, 2, 4]
Bootstrap	[True, False]

10 ROBUSTNESS CHECKS

We performed a series of checks to ensure the robustness of our results. In particular, we performed robustness checks regarding the (1) machine learning model and (2) platform heterogeneity.

First, we checked how our results change when using another machine learning classifier to predict impressions-per-EUR. In particular, we trained an XGBoost model using the same training procedure as outlined for the random forest model. However, the XGBoost model resulted in a lower prediction performance ($RMSE$ =127.75 over 10 runs ± a s.d. of 3.91) when predicting impressions-per-EUR on the hold-out set. Nevertheless, our main findings remain consistent, i.e., (1) our model can only explain a fraction of the variance in our data (R^2 =0.36 ± a s.d. of 0.02 over 10 runs) and (2) the far-right *AfD* consistently achieves more impressions-per-EUR than predicted by our model.

Second, we checked the heterogeneity of our results with respect to the platform an ad was published. Specifically, we re-trained our random forest model but only used (1) ads that were published on Facebook and (2) ads that were published on Instagram. Hence, we trained two additional random forest model focusing on political ads from (1) Facebook or (2) Instagram. The results are shown in Fig. S5. Consistent with the results from our main analysis, we find systematic differences between actual impressions-per-EUR on the platform and predicted by our model. Importantly, we find that the far-right *AfD* consistently achieves more impressions-per-EUR than predicted by our model.

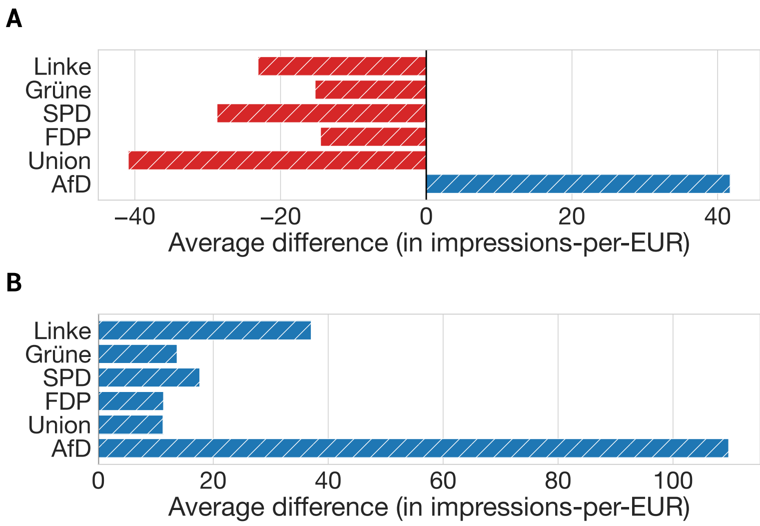

Fig. S5. Average difference between actual vs. predicted impressions-per-EUR based on our machine learning model over 10 runs for political ads published on a, Facebook and b, Instagram.

REFERENCES

[1] Minali Aggarwal, Jennifer Allen, Alexander Coppock, Dan Frankowski, Solomon Messing, Kelly Zhang, James Barnes, Andrew Beasley, Harry Hantman, and Sylvan Zheng. 2023. A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. *Nature Human Behaviour* 7 (2023), 332–341.

[2] Laurenz Aisenpreis, Gustav Gyrst, and Vedran Sekara. 2023. How do US congress members advertise climate change: An analysis of ads run on Meta's platforms. In *ICWSM*.

[3] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2021. Ad delivery algorithms: The hidden arbiters of political messaging. In *WSDM*.

[4] Brooke Auxier. 2020. 54% of Americans say social media companies shouldn't allow any political ads. *Pew Research Center* (2020). https://www.pewresearch.org/short-reads/2020/09/24/54-of-americans-say-social-media-companies-shouldnt-allow-any-political-ads/

[5] Andrew O. Ballard, D. Sunshine Hillygus, and Tobias Konitzer. 2016. Campaigning online: Web display ads in the 2012 presidential campaign. *PS: Political Science & Politics* 49, 03 (2016), 414–419.

[6] Emily Breza, Fatima Cody Stanford, Marcella Alsan, Burak Alsan, Abhijit Banerjee, Arun G. Chandrasekhar, Sarah Eichmeyer, Traci Glushko, Paul Goldsmith-Pinkham, Kelly Holland, Emily Hoppe, Mohit Karnani, Sarah Liegl, Tristan Loisel, Lucy Ogbu-Nwobodo, Benjamin A. Olken, Carlos Torres, Pierre-Luc Vautrey, Erica T. Warner, Susan Wootton, and Esther Duflo. 2021. Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: A cluster randomized controlled trial. *Nature Medicine* 27, 9 (2021), 1622–1628.

[7] Bundeswahlleiter. 2022. The Federal Returning Officer. https://www.bundeswahlleiter.de/en/bundeswahlleiter.html

[8] José González Cabañas, Ángel Cuevas, Aritz Arrate, and Rubén Cuevas. 2021. Does Facebook use sensitive data for advertising purposes? *Communications of the ACM* 64, 1 (2021), 62–69.

[9] Lorena Cano-Orón, Dafne Calvo, Guillermo López García, and Tomás Baviera. 2021. Disinformation in Facebook ads in the 2019 Spanish general election campaigns. *Media and Communication* 9, 1 (2021), 217–228.

[10] Arthur Capozzi, Gianmarco de Francisci Morales, Yelena Mejova, Corrado Monti, and André Panisson. 2023. The thin ideology of populist advertising on Facebook during the 2019 EU elections. In *WWW*.

[11] Arthur Capozzi, Gianmarco de Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. 2020. Facebook ads: Politics of migration in Italy. In *International Conference on Social Informatics*.

[12] Arthur Capozzi, Gianmarco de Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. 2021. Clandestino or rifugiato? Anti-immigration Facebook ad targeting in Italy. In *CHI*.

[13] Matteo Cinelli, Gianmarco de Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *PNAS* 118, 9 (2021).

[14] Bruno Coelho, Tobias Lauinger, Laura Edelson, Ian Goldstein, and Damon McCoy. 2023. Propaganda política pagada: Exploring U.S. political Facebook ad en Español. In *WWW*.

[15] Alexander Coppock, Donald P. Green, and Ethan Porter. 2022. Does digital advertising affect vote choice? Evidence from a randomized field experiment. *Research & Politics* 9, 1 (2022), 20531680221076901.

[16] Deutscher Bundestag. 2023. The Federal Republic of Germany (since 1949). https://www.bundestag.de/en/parliament/history/parliamentarism/frg_parliamentarism/frg_parliamentarism-200324

[17] Katharine Dommett and Sam Power. 2019. The political economy of Facebook advertising: Election spending, regulation and targeting online. *The Political Quarterly* 90, 2 (2019), 257–265.

[18] Laura Edelson, Tobias Lauinger, and Damon McCoy. 2020. A security analysis of the Facebook Ad Library. In *IEEE Symposium on Security and Privacy*.

[19] Laura Edelson, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. An analysis of United States online political advertising transparency. *arXiv* (2019), 1902.04385.

[20] Cornelius Erfort. 2023. Gendered targeting: Do parties tailor their campaign ads to women? *OSF Preprints* (2023).

[21] Erika Franklin Fowler, Michael M. Franz, Gregory J. Martin, Zachary Peskowitz, and Travis N. Ridout. 2021. Political advertising online and offline. *American Political Science Review* 115, 1 (2021), 130–149.

[22] Erika Franklin Fowler, Michael M. Franz, and Travis N. Ridout. 2020. Online political advertising in the United States. In *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press, Cambridge, UK, 111–138.

[23] Kiran Garimella, Gianmarco de Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *WWW*.

[24] Matthew H. Goldberg, Abel Gustafson, Seth A. Rosenthal, and Anthony Leiserowitz. 2021. Shifting Republican views on climate change through targeted advertising. *Nature Climate Change* 11, 7 (2021), 573–577.

[25] Thomas Gschwend and Thomas Zittel. 2015. Do constituency candidates matter in German Federal Elections? The personal vote as an interactive process. *Electoral Studies* 39 (2015), 338–349.

[26] Anselm Hager. 2019. Do online ads influence vote choice? *Political Communication* 36, 3 (2019), 376–393.

[27] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science* 65, 7 (2019), 2966–2981.

[28] J. Nathan Matias, Austin Hounsel, and Nick Feamster. 2022. Software-supported audits of decision-making systems: Testing Google and Facebook's political advertising policies. In *CSCW*.

[29] Sandra C. Matz, Michal Kosinski, Gideon Nave, and David J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *PNAS* 114, 48 (2017), 12714–12719.

[30] Francesco Pierri. 2023. Political advertisement on Facebook and Instagram in the run up to 2022 Italian general election. In *WebSci*.

[31] Travis N. Ridout, Erika Franklin Fowler, and Michael M. Franz. 2021. The Influence of goals and timing: How campaigns deploy ads on Facebook. *Journal of Information Technology & Politics* 18, 3 (2021), 293–309.

[32] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabricio Benevenuto. 2020. Facebook ads monitor: An independent auditing system for political ads on Facebook. In *WWW*.

[33] Vera Sosnovik and Oana Goga. 2021. Understanding the complexity of detecting political ads. In *WWW*.

[34] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Mislove Alan. 2018. Potential for discrimination in online targeted advertising. In *FAccT*.

[35] Jennifer Stromer-Galley, Patrícia Rossini, Jeff Hemsley, Sarah E. Bolden, and Brian McKernan. 2021. Political messaging over time: A comparison of US presidential candidate Facebook posts and tweets in 2016 and 2020. *Social Media + Society* 7, 4 (2021).

[36] The Economist. 2021. After Merkel. https://www.economist.com/special-report/2021/09/20/after-merkel

[37] Feifei Zhang, Jennifer Stromer-Galley, Sikana Tanupabrungsun, Yatish Hegde, Nancy McCracken, and Jeff Hemsley. 2017. Understanding discourse acts: Political campaign messages classification on Facebook and Twitter. In *Social, cultural, and behavioral modeling*. Lecture Notes in Computer Science, Vol. 10354. Springer, New York, NY, USA, 242–247.

D Supplementary materials: Generative AI may backfire for counterspeech

Generative AI may backfire for counterspeech

Supplementary Materials

Dominik Bär, LMU Munich & Munich Center for Machine Learning, Germany
Abdurahman Maarouf, LMU Munich & Munich Center for Machine Learning, Germany
Stefan Feuerriegel, LMU Munich & Munich Center for Machine Learning, Germany

1 Related work

Online hate speech is a significant threat to individual well-being and social cohesion [17, 28, 46, 48, 49, 59]. The United Nations defines hate speech as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." [60] Previous research has, for example, studied the virality of hate speech [31], the characteristics of users sharing hate speech [43] but also how to detect [2, 39, 43] and curb online hate speech [23, 37, 38, 49].

Curbing online hate speech is challenging [19, 26]. Automated measures, such as content removal and account suspensions, are scalable and can effectively reduce online hate speech [19]. However, automated measures that are not properly calibrated may falsely remove content, which may be perceived as an infringement on individuals' freedom of speech [20, 22, 36] and thus even spur more hostility [24, 32]. In contrast, manual moderation, such as the removal of problematic content or accounts by human moderators, can be more precise [19]. However, the high prevalence of online hate speech makes such efforts impractical [26]. Research has also shown that manual moderation can negatively impact the well-being of moderators [53], raising ethical concerns about its use. In this paper, we focus on counterspeech, which is seen as a promising approach to address the rise in online hate speech [11, 23, 61].

Counterspeech refers to direct responses intended to encourage users to reconsider their hateful posts [32]. Since no content is removed, a key advantage of counterspeech is that it does not infringe on users' freedom of speech [32]. Previous research has suggested various strategies for counterspeech, such as empathy or reminding of social norms through warning-of-consequences [23, 32, 36–38, 54]. The effectiveness of counterspeech has been demonstrated in multiple field experiments [23, 37, 38, 49]. For example, counterspeech reminding of social norms reduced religious hate speech [49]. Furthermore, empathy-based counterspeech significantly reduced racist hate on Twitter/X [23, 37]. However, these studies typically follow a "one-fits-all" paradigm where predefined, generic counterspeech messages are sent to all offenders (e.g., *"This post is disrespectful. Please stop posting such hateful content!"*). In other words, this "one-fits-all" approach ignores the context of the underlying hateful post, potentially limiting its persuasiveness.

The emergence of LLMs has greatly improved the quality of automated text generation. LLMs take so-called prompts as inputs and then generate human-like text [18, 21, 25, 51, 67]. LLMs are nowadays applied across various fields. For example, LLMs have been used to write political messages [21], aid mental health support [47], and provide recommendations in the emergency department [64]. Here, we explore the use of LLMs to generate contextualized counterspeech to curb online hate speech. A key strength of LLMs in this setting is that LLMs are scalable to the large volume of hate speech on social media platforms.

Previous research has demonstrated the potential of LLMs to produce counterspeech [7, 12, 13, 42, 57], yet with important limitations. On the one hand, these studies [7, 12, 13, 57] do not evaluate the effectiveness of LLM-generated counterspeech in the field but merely use surveys for evaluation. This is problematic since, even if people report that they are willing to behave civilly, it does not mean they act accordingly. This observation is known as the "intention-behavior gap" and poses a severe limitation when measuring intentions instead of actual behavior [27]. This can lead to inflated reports of civil behavior in surveys that may not translate to real-world social media environments. On the other hand, there is research [42] that analyzes secondary outcomes (e.g., views, likes) and thus fails to measure the effectiveness of counterspeech. Consequently, it remains unclear whether LLM-generated counterspeech can reduce online hate speech on real-world social media platforms.

To the best of our knowledge, there is no evidence on whether LLM-generated counterspeech can effectively change real-world social media users' behavior to reduce online hate speech. To close this gap, we conducted a field experiment on Twitter/X to test whether contextualized counterspeech generated by state-of-the-art LLMs is effective in curbing online hate speech.

2 Human-controlled accounts

We administered our intervention via multiple human-controlled accounts. The accounts were designed to appear politically neutral and natural to users on Twitter/X, which was inspired by the design in [23]. Each account was assigned a unisex English name, with no disclosure of gender, ethnicity, nationality, or beliefs. Furthermore, to appear as natural users, we regularly posted neutral posts via our accounts (e.g., *"Just witnessed the most breathtaking sunset!"*) and re-posted content from diverse accounts (e.g., NASA, WWF, ESPN). The accounts were created at least 3 months before the start of the experiment. Screenshots of example profiles are in Fig. S1.

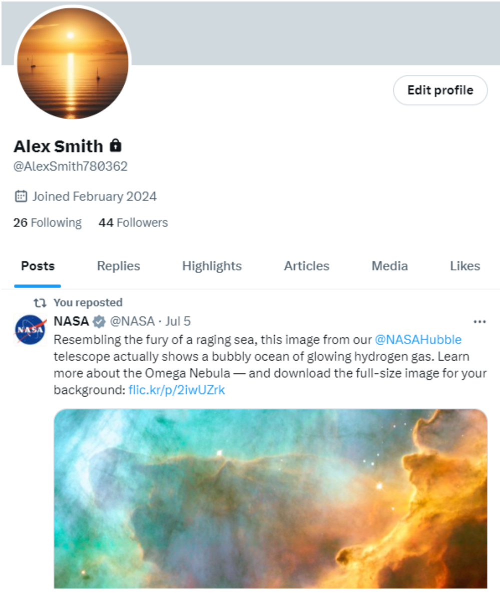

Fig. S1. Example of one of our human-controlled accounts.

3 Keywords to identify hatespeech

Table S1. Keywords to identify hate speech on Twitter/X.

Group	Terms
Religious	jew, zionist, goy, goyim, heeb, hebe, muslim, islam, jihad, nazi, terror
Ethnic/Racial	arab, paki, cameljockey, cameltoe, kanake, palesimian, spaghettibender, raghead, kike, spic, wetback, chink, gook
Cultural	gypsy, redneck, hillbilly, beaner, mite, dink, injun, jigaboo
Other	parasite

4 Prompt templates

Table S2. Prompt templates to generate contextualized LLM-generated counterspeech. We use two different templates for our two strategies, namely, (1) empathy and (2) warning-of-consequences.

Empathy	**System Prompt**: ### Persona ### User Persona: As a social media user committed to countering hate speech, you embody empathy and skill in responding to harmful content. Your role is to foster understanding and promote more respectful and inclusive dialogue. ### Concepts ### - Counter Speech: This involves crafting responses to directly address, mitigate, or refute hateful or harmful speech in a way that discourages further negativity. - Empathy: Empathy, the ability to understand and share others' feelings, helps combat hate speech by encouraging compassion over hostility. This fosters a culture of respect and reduces the spread of harmful language. ### Your Task ### - Contextual Relevance: Your response must directly relate to the specific content of the original hateful post. - Conciseness: Limit your response to 2-3 SHORT sentences, 30-40 words, no subclauses, no hashtags or emojis. - Avoid Confrontational Language: Keep the tone respectful and constructive. ### Approach ### 1) Ask yourself why the post is harmful and what consequences it might have. 2) Ask yourself what is the relevant context of the post. 3) Think about the feelings of those targeted by the hate speech. 4) Craft a counterspeech that: - Explains why the post is harmful following the analysis in 1) and 2). - Demonstrates empathy towards those targeted following the analysis in 3). ### Output ### Ensure the response is in a JSON-compatible format: {“counterspeech”: “Your contextually relevant counter speech here”}
	User Prompt: Create a counterspeech response using empathy for the following post on social media: Post: [POST]
Warning-of-consequences	**System Prompt**: ### Persona ### User Persona: As a social media user committed to countering hate speech, you maintain a neutral and factual tone responding to harmful content. Your role is to foster understanding and promote more respectful and inclusive dialogue. ### Concepts ### - Counter Speech: This involves crafting responses to directly address, mitigate, or refute hateful or harmful speech in a way that supports the victim and discourages further negativity. - Warning of Consequences: Warn the poster about the potential consequences to their personal and professional life from promoting hate speech. This can help deter future harmful behavior and encourage more responsible online conversations. ### Your Task ### - Contextual Relevance: Your response must directly relate to the specific content of the original hateful post. - Conciseness: Limit your response to 2-3 SHORT sentences, 30-40 words, no subclauses, no hashtags or emojis. - Avoid Confrontational Language: Keep the tone respectful and constructive. ### Approach ### 1) Ask yourself why the post is harmful and what consequences it might have. 2) Ask yourself what is the relevant context of the post. 3) Think about the potential personal and professional consequences for the poster of the hate speech. 4) Craft a counterspeech that: - Explains why the post is harmful following the analysis in 1) and 2). - Warns of the consequences the poster might face based on the analysis in 3). ### Output ### Ensure the response is in a JSON-compatible format: {“counterspeech”: “Your contextually relevant counter speech here”}
	User Prompt: Create a counterspeech response warning of consequences for the following post on social media: Post: [POST]

5 Examples of counterspeech

Table S3. Examples for hateful posts from Twitter/X and counterspeech generated using our LLM approach together.

Strategy	Post	Counterspeech
Empathy	*"Why do the muslims in the comm hate Christians so much? They never support us unless it benefits them or some muslim scholar tells them to."*	*"I understand that you may feel frustrated by perceived lack of support, but it's unfair to generalize an entire community. Muslims, like people of any faith, are individuals with diverse views and experiences. Let's focus on building bridges, not stereotypes."*
	"This guy is an absolute braindead ass who just hates irish and suddenly blames 'All irish as jews who send Americans to war' lmfao ðŸ¤£ he's a pawn probably a jew himself deceiving creating hate for others and protection for them as always.. Literally guy is an ass."	*"Let's not generalize and blame entire groups. Irish and Jewish communities have faced discrimination, we should promote understanding and respect."*
	"JOE BIDEN IS A NAZI"	*"Labeling someone a Nazi without evidence is harmful and disrespectful to the millions of people who suffered under the real Nazi regime. It also creates a toxic environment where constructive dialogue is impossible."*
Warning-of-consequences	*"I can't stand this kike's music @Drake"*	*"Using derogatory terms like 'kike' is harmful and offensive, promoting anti-Semitism and hate. This kind of language can lead to real-world discrimination and violence, and can also damage your personal and professional reputation."*
	"Fuck Islam"	*"This kind of hate speech can lead to discrimination and violence against Muslims, and can also damage your personal and professional reputation. It's important to respect all religions and promote inclusivity."*
	"Bill Ackman is a parasite, a taker of the world. His opinion is meaningless"	*"Labeling individuals as 'parasites' can perpetuate harmful stereotypes and discrimination. Such language can have serious consequences, including damage to one's professional reputation and potential legal repercussions."*

6 Non-contextualized counterspeech

Table S4. Non-contextualized counterspeech for our two strategies, namely, (1) empathy and (2) warning-of-consequences. We used five different counterspeech messages for each strategy to avoid a strongly repetitive reply pattern that may irritate users.

Strategy	Counterspeech
Empathy	**Variant 1**: "*Language like this may hurt people. Words have power. Imagine those words aimed at you or someone you love. We must communicate with openness, and respect, and promote empathy in every interaction. Together, we can build a kinder world. Let's lead by example.*"
	Variant 2: "*Divisive rhetoric like this does more harm than good. Remember, there's a human with feelings behind every screen. Use words to uplift, not hurt. Compassion is our most powerful tool for change. Let's commit to kindness and build connections, not divisions.*"
	Variant 3: "*Hate speech like this can cause real harm. We must strive to understand and respect one another, even when we disagree. Let's work together to build a world where everyone feels safe and valued. Our words have the power to heal or to hurt. Let's choose to heal.*"
	Variant 4: "*Words matter. We must strive to understand and respect one another, even when we disagree. Let's work together to build a world where everyone feels safe and valued. Let's spread messages of hope and inclusion, not fear. Together, we can make a difference.*"
	Variant 5: "*It's easy to forget the impact our words can have on others, especially online. Before we speak, let's pause and consider the person on the other side. Empathy and understanding are the foundations of a strong community. Let's choose to be kind.*"
Warning of consequence	**Variant 1**: "*Be mindful of what you post online. Hateful comments can not only hurt others but also come back at you, affecting job prospects, personal relationships, and how the community views you. In a world where digital footprints are permanent, think before you share.*"
	Variant 2: "*Every public post you make is permanent and can impact your future in ways you might not anticipate. Employers, family members, and friends may see what you post online. Consider the long-term impact of your words on your reputation and opportunities.*"
	Variant 3: "*Your words on social media carry weight and consequences. A moment of frustration or anger can translate into a lifetime of regret if it leads to legal issues or social isolation. Think about the broader impact of your posts before making them public.*"
	Variant 4: "*While social media offers some anonymity, remember that hateful posts can lead to serious offline consequences, including legal action or personal backlash. Your online actions reflect on your real-life identity. Pause and consider the consequences of your post.*"
	Variant 5: "*The internet has a long memory, and today's post could easily become tomorrow's regret. What you share today could shape your future in unexpected ways. Protect your future self by taking a moment to reflect on the potential personal consequences of your post.*"

7 Summary statistics

Table S5. Summary statistics of dependent variables and pre-treatment covariates.

Variable	Mean	Median	Min	Max	Std	Skew
Rate of deleted posts	0.055	0.000	0.000	1.000	0.228	3.896
Number of hateful posts (post-intervention)	8.747	5.000	0.000	98.000	10.922	2.746
Relative change in toxicity	0.046	−0.012	−0.964	6.815	0.429	4.418
Account age (in days)	1770.637	954.500	2.000	6392.000	1813.210	0.912
Follower count	7646.771	419.500	0.000	7 157 501.000	143 838.344	46.490
Following count	1715.756	447.500	0.000	270 342.000	8027.240	22.938
Tweet count	33 358.907	10 652.000	12.000	858 800.000	70 174.820	5.461
Like count	36 769.592	10 824.500	0.000	1 310 634.000	76 956.586	6.374
Twitter/X premium (= 1 if premium)	0.109	0.000	0.000	1.000	0.312	2.512
Number of hateful posts (pre-intervention)	7.953	5.000	0.000	89.000	9.787	2.722
Average toxicity (pre-intervention)	0.287	0.269	0.020	0.889	0.123	0.831

8 Regression results

8.1 Effectiveness of counterspeech compared to control condition

Table S6. Estimation results for the treatment effects of different counterspeech interventions compared to the control condition for *Rate of deleted posts*. The first column reports the regression results for GENERIC-EMPATHY counterspeech. The second column shows the results for CONTEXT-EMPATHY. The third column provides the results for GENERIC-WARNING, and the fourth column reports on CONTEXT-WARNING.

	GENERIC-EMPATHY	CONTEXT-EMPATHY	GENERIC-WARNING	CONTEXT-WARNING
Intercept	0.086^{***}	0.074^{***}	0.106^{***}	0.068^{**}
	(0.027)	(0.026)	(0.027)	(0.027)
GENERIC-EMPATHY	-0.026^{*}			
	(0.014)			
CONTEXT-EMPATHY		-0.029^{**}		
		(0.014)		
GENERIC-WARNING			0.008	
			(0.016)	
CONTEXT-WARNING				−0.014
				(0.014)
Account age (in days)	-0.000^{**}	-0.000^{**}	-0.000^{**}	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Followers count	0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Following count	−0.000	−0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Tweet count	-0.000^{**}	−0.000	-0.000^{**}	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Like count	0.000	0.000	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0.024	0.003	−0.034	−0.013
	(0.018)	(0.022)	(0.021)	(0.020)
Average toxicity (pre-intervention)	0.015	0.098	0.001	0.064
	(0.085)	(0.088)	(0.087)	(0.085)
Number of hateful posts (pre-intervention)	−0.000	-0.002^{**}	-0.002^{*}	−0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Adj. R^2	0.006	0.011	0.006	0.000
Obs.(N)	1073	1074	1030	1093

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.01$

Table S7. Estimation results for the treatment effects of different counterspeech interventions compared to the control condition for *Number of hateful posts*. The first column reports the regression results for GENERIC-EMPATHY counterspeech. The second column shows the results for CONTEXT-EMPATHY. The third column provides the results for GENERIC-WARNING, and the fourth column reports on CONTEXT-WARNING.

	GENERIC-EMPATHY	CONTEXT-EMPATHY	GENERIC-WARNING	CONTEXT-WARNING
Intercept	0.208	0.251	0.034	−0.554
	(0.758)	(0.699)	(0.698)	(0.721)
GENERIC-EMPATHY	−0.350			
	(0.485)			
CONTEXT-EMPATHY		−0.433		
		(0.434)		
GENERIC-WARNING			-1.026^{**}	
			(0.447)	
CONTEXT-WARNING				−0.127
				(0.453)
Account age (in days)	−0.000	0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Followers count	−0.000	0.000	−0.000	-0.000^{**}
	(0.000)	(0.000)	(0.000)	(0.000)
Following count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Tweet count	0.000^{***}	0.000^{*}	0.000	0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Like count	−0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−1.125	-1.513^{**}	−1.097	0.300
	(0.696)	(0.668)	(0.730)	(0.693)
Average toxicity (pre-intervention)	7.421^{***}	7.275^{***}	7.312^{***}	9.497^{***}
	(2.750)	(2.377)	(2.494)	(2.470)
Number of hateful posts (pre-intervention)	0.852^{***}	0.837^{***}	0.827^{***}	0.822^{***}
	(0.048)	(0.052)	(0.045)	(0.050)
Adj. R^2	0.575	0.555	0.605	0.588
Obs.(N)	1064	1062	1018	1082

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S8. Estimation results for the treatment effects of different counterspeech interventions compared to the control condition for *Relative change in toxicity*. The first column reports the regression results for GENERIC-EMPATHY counterspeech. The second column shows the results for CONTEXT-EMPATHY. The third column provides the results for GENERIC-WARNING, and the fourth column reports on CONTEXT-WARNING.

	GENERIC-EMPATHY	CONTEXT-EMPATHY	GENERIC-WARNING	CONTEXT-WARNING
Intercept	0.400^{***}	0.501^{***}	0.398^{***}	0.376^{***}
	(0.054)	(0.062)	(0.052)	(0.054)
GENERIC-EMPATHY	−0.014			
	(0.022)			
CONTEXT-EMPATHY		0.037		
		(0.029)		
GENERIC-WARNING			−0.024	
			(0.022)	
CONTEXT-WARNING				0.006
				(0.023)
Account age (in days)	-0.000^{**}	0.000	-0.000^{***}	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Followers count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Following count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Tweet count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Like count	−0.000	0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	-0.059^{*}	-0.149^{***}	−0.045	−0.048
	(0.031)	(0.041)	(0.035)	(0.029)
Average toxicity (pre-intervention)	-1.167^{***}	-1.624^{***}	-1.194^{***}	-1.176^{***}
	(0.138)	(0.197)	(0.131)	(0.145)
Number of hateful posts (pre-intervention)	0.002^{*}	0.003^{**}	0.003^{***}	0.003^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Adj. R^2	0.115	0.129	0.132	0.100
Obs.(N)	1063	1060	1016	1080

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

8.2 Contextualized vs. non-contextualized counterspeech

Table S9. Estimation results for the treatment effects of interventions with contextualized counterspeech generated by an LLM, compared to generic counterspeech, for the *Rate of deleted posts*. The first column presents the regression results comparing GENERIC-EMPATHY with CONTEXT-EMPATHY counterspeech. The second column provides the results comparing GENERIC-WARNING with CONTEXT-WARNING counterspeech.

	Empathy	Warning-of-consequences
Intercept	0.049^{**}	0.099^{***}
	(0.022)	(0.027)
Empathy	−0.003	
	(0.012)	
Warning-of-consequences		−0.023
		(0.015)
Account age (in days)	-0.000^{**}	−0.000
	(0.000)	(0.000)
Followers count	−0.000	−0.000
	(0.000)	(0.000)
Following count	-0.000^{**}	−0.000
	(0.000)	(0.000)
Tweet count	−0.000	−0.000
	(0.000)	(0.000)
Like count	−0.000	−0.000
	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0.000	−0.023
	(0.019)	(0.021)
Average toxicity (pre-intervention)	0.053	0.005
	(0.072)	(0.072)
Number of hateful posts (pre-intervention)	−0.001	−0.001
	(0.001)	(0.001)
Adj. R^2	0.002	0.001
Obs.(N)	1065	1041

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S10. Estimation results for the treatment effects of interventions with contextualized counterspeech generated by an LLM, compared to generic counterspeech, for the *Number of hateful posts*. The first column presents the regression results comparing GENERIC-EMPATHY with CONTEXT-EMPATHY counterspeech. The second column provides the results comparing GENERIC-WARNING with CONTEXT-WARNING counterspeech.

	Empathy	Warning-of-consequences
Intercept	0.305	-1.428^{**}
	(0.629)	(0.636)
Empathy	-0.149	
	(0.420)	
Warning-of-consequences		0.839^{**}
		(0.390)
Account age (in days)	-0.000^{**}	-0.000
	(0.000)	(0.000)
Followers count	0.000	-0.000
	(0.000)	(0.000)
Following count	0.000	-0.000
	(0.000)	(0.000)
Tweet count	0.000^{**}	0.000^{**}
	(0.000)	(0.000)
Like count	0.000	0.000
	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	-0.793	1.268^{**}
	(0.587)	(0.584)
Average toxicity (pre-intervention)	7.743^{***}	9.785^{***}
	(2.699)	(2.479)
Number of hateful posts (pre-intervention)	0.811^{***}	0.780^{***}
	(0.044)	(0.041)
Adj. R^2	0.588	0.661
Obs.(N)	1056	1030

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S11. Estimation results for the treatment effects of interventions with contextualized counterspeech generated by an LLM, compared to generic counterspeech, for the *Relative change in toxicity*. The first column presents the regression results comparing GENERIC-EMPATHY with CONTEXT-EMPATHY counterspeech. The second column provides the results comparing GENERIC-WARNING with CONTEXT-WARNING counterspeech.

	Empathy	Warning-of-consequences
Intercept	0.470^{***}	0.324^{***}
	(0.055)	(0.042)
Empathy	0.055^{**}	
	(0.028)	
Warning-of-consequences		0.024
		(0.022)
Account age (in days)	-0.000	-0.000
	(0.000)	(0.000)
Followers count	-0.000	-0.000
	(0.000)	(0.000)
Following count	-0.000	-0.000
	(0.000)	(0.000)
Tweet count	-0.000	-0.000
	(0.000)	(0.000)
Like count	0.000	-0.000
	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	-0.093^{**}	0.014
	(0.041)	(0.034)
Average toxicity (pre-intervention)	-1.564^{***}	-1.107^{***}
	(0.201)	(0.127)
Number of hateful posts (pre-intervention)	0.002	0.003^{***}
	(0.001)	(0.001)
Adj. R^2	0.127	0.111
Obs.(N)	1055	1028

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

8.3 Counterspeech for Twitter/X Premium users

Hateful users who subscribe to Twitter/X Premium are less likely to have their content removed by the platform, and their posts are algorithmically boosted [14]. Hence, we evaluate whether our intervention is effective for Twitter/X Premium users. To do so, we re-estimate the regression model from our main analysis, adding an interaction term between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise).

We find no significant interaction between our intervention and Twitter/X Premium subscription status for the *rate of deleted posts*. However, Twitter/X Premium users who received CONTEXT-EMPATHY counterspeech shared significantly more hateful posts ($p = 0.012$) than non-subscribers. Additionally, Premium users exhibited higher toxicity levels when receiving GENERIC-EMPATHY ($p = 0.049$) and CONTEXT-EMPATHY ($p = 0.024$) counterspeech. This suggests that empathetic counterspeech, particularly when LLM-generated, may backfire for Premium users. Importantly, all treatment effects remain consistent with our primary analysis across all models and dependent variables, except for CONTEXT-WARNING vs. GENERIC-WARNING and the *number of hateful posts*, which is no longer significant ($p = 0.107$). Detailed regression results are in Table S12 to Table S17.

Table S12. Estimation results for the treatment effects of different counterspeech interventions compared to the control condition for *Rate of deleted posts*. To evaluate whether our intervention is effective for Twitter/X Premium users, we re-estimate the regression model from our main analysis, adding an interaction term (*Treated x Twitter/X premium (*= 1 *if premium)*) between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise). The first column reports the regression results for GENERIC-EMPATHY counterspeech. The second column shows the results for CONTEXT-EMPATHY. The third column provides the results for GENERIC-WARNING, and the fourth column reports on CONTEXT-WARNING. *Treated* corresponds to the intervention indicated by the column.

	GENERIC-EMPATHY	CONTEXT-EMPATHY	GENERIC-WARNING	CONTEXT-WARNING
Intercept	0.086^{***}	0.076^{***}	0.105^{***}	0.069^{**}
	(0.027)	(0.027)	(0.027)	(0.027)
GENERIC-EMPATHY	-0.026^{*}			
	(0.015)			
CONTEXT-EMPATHY		-0.033^{**}		
		(0.015)		
GENERIC-WARNING			0.011	
			(0.017)	
CONTEXT-WARNING				−0.016
				(0.016)
Account age (in days)	-0.000^{**}	-0.000^{**}	-0.000^{**}	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Followers count	0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Following count	−0.000	−0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Tweet count	-0.000^{**}	−0.000	-0.000^{**}	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Like count	0.000	0.000	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0.025	−0.018	−0.021	−0.023
	(0.029)	(0.028)	(0.029)	(0.029)
Treated x Twitter/X premium (= 1 if premium)	0.001	0.041	−0.030	0.016
	(0.034)	(0.041)	(0.039)	(0.039)
Average toxicity (pre-intervention)	0.015	0.097	0.001	0.063
	(0.085)	(0.088)	(0.087)	(0.085)
Number of hateful posts (pre-intervention)	−0.000	-0.002^{**}	-0.002^{*}	−0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Adj. R^2	0.005	0.010	0.005	−0.001
Obs.(N)	1073	1074	1030	1093

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S13. Estimation results for the treatment effects of different counterspeech interventions compared to the control condition for *Number of hateful posts*. To evaluate whether our intervention is effective for Twitter/X Premium users, we re-estimate the regression model from our main analysis, adding an interaction term (*Treated x Twitter/X premium (*= 1 *if premium)*) between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise). The first column reports the regression results for GENERIC-EMPATHY counterspeech. The second column shows the results for CONTEXT-EMPATHY. The third column provides the results for GENERIC-WARNING, and the fourth column reports on CONTEXT-WARNING. *Treated* corresponds to the intervention indicated by the column.

	GENERIC-EMPATHY	CONTEXT-EMPATHY	GENERIC-WARNING	CONTEXT-WARNING
Intercept	0.279	0.284	0.108	−0.328
	(0.778)	(0.712)	(0.711)	(0.740)
GENERIC-EMPATHY	−0.496			
	(0.520)			
CONTEXT-EMPATHY		−0.515		
		(0.468)		
GENERIC-WARNING			-1.202^{**}	
			(0.474)	
CONTEXT-WARNING				−0.536
				(0.480)
Account age (in days)	−0.000	0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Followers count	−0.000	0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Following count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Tweet count	0.000^{***}	0.000^{*}	0.000	0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Like count	−0.000	0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−1.830	-1.896^{*}	-1.923^{*}	−1.643
	(1.157)	(1.104)	(1.123)	(1.138)
Treated x Twitter/X premium (= 1 if premium)	1.336	0.775	1.913	3.432^{**}
	(1.392)	(1.294)	(1.319)	(1.361)
Average toxicity (pre-intervention)	7.363^{***}	7.261^{***}	7.307^{***}	9.340^{***}
	(2.756)	(2.378)	(2.495)	(2.470)
Number of hateful posts (pre-intervention)	0.853^{***}	0.838^{***}	0.827^{***}	0.825^{***}
	(0.048)	(0.052)	(0.045)	(0.049)
Adj. R^2	0.575	0.555	0.605	0.590
Obs.(N)	1064	1062	1018	1082

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S14. Estimation results for the treatment effects of different counterspeech interventions compared to the control condition for *Relative change in toxicity*. To evaluate whether our intervention is effective for Twitter/X Premium users, we re-estimate the regression model from our main analysis, adding an interaction term (*Treated x Twitter/X premium (= 1 if premium)*) between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise). The first column reports the regression results for GENERIC-EMPATHY counterspeech. The second column shows the results for CONTEXT-EMPATHY. The third column provides the results for GENERIC-WARNING, and the fourth column reports on CONTEXT-WARNING. *Treated* corresponds to the intervention indicated by the column.

	GENERIC-EMPATHY	CONTEXT-EMPATHY	GENERIC-WARNING	CONTEXT-WARNING
Intercept	0.404^{***}	0.500^{***}	0.402^{***}	0.383^{***}
	(0.054)	(0.062)	(0.052)	(0.055)
GENERIC-EMPATHY	−0.023			
	(0.024)			
CONTEXT-EMPATHY		0.040		
		(0.031)		
GENERIC-WARNING			−0.036	
			(0.024)	
CONTEXT-WARNING				−0.008
				(0.026)
Account age (in days)	-0.000^{**}	0.000	-0.000^{**}	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Followers count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Following count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Tweet count	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Like count	−0.000	0.000	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	-0.101^{***}	-0.134^{***}	-0.103^{***}	-0.114^{***}
	(0.038)	(0.039)	(0.038)	(0.038)
Treated x Twitter/X premium (= 1 if premium)	0.079	−0.031	0.133^{**}	0.116^{**}
	(0.058)	(0.071)	(0.067)	(0.051)
Average toxicity (pre-intervention)	-1.170^{***}	-1.623^{***}	-1.193^{***}	-1.180^{***}
	(0.138)	(0.197)	(0.131)	(0.145)
Number of hateful posts (pre-intervention)	0.002^{*}	0.003^{**}	0.003^{***}	0.003^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Adj. R^2	0.115	0.129	0.133	0.101
Obs.(N)	1063	1060	1016	1080

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S15. Estimation results for the treatment effects of interventions with contextualized counterspeech generated by an LLM, compared to generic counterspeech, for the *Rate of deleted posts*. To evaluate whether our intervention is effective for Twitter/X Premium users, we re-estimate the regression model from our main analysis, adding an interaction term (*Treated x Twitter/X premium (= 1 if premium)*) between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise). The first column presents the regression results comparing GENERIC-EMPATHY with CONTEXT-EMPATHY counterspeech. The second column provides the results comparing GENERIC-WARNING with CONTEXT-WARNING counterspeech. *Treated* corresponds to the intervention with contextualized counterspeech generated by an LLM with the strategy indicated by the column.

	Empathy	Warning-of-consequences
Intercept	0.051**	0.102***
	(0.023)	(0.027)
Empathy	−0.008	
	(0.012)	
Warning-of-consequences		−0.027
		(0.017)
Account age (in days)	−0.000**	−0.000
	(0.000)	(0.000)
Followers count	−0.000	−0.000
	(0.000)	(0.000)
Following count	−0.000**	−0.000
	(0.000)	(0.000)
Tweet count	−0.000	−0.000
	(0.000)	(0.000)
Like count	−0.000	−0.000
	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0.022	−0.051*
	(0.020)	(0.029)
Treated x Twitter/X premium (= 1 if premium)	0.044	0.044
	(0.037)	(0.038)
Average toxicity (pre-intervention)	0.054	0.003
	(0.072)	(0.072)
Number of hateful posts (pre-intervention)	−0.001	−0.001
	(0.001)	(0.001)
Adj. R^2	0.003	0.001
Obs.(N)	1065	1041

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S16. Estimation results for the treatment effects of interventions with contextualized counterspeech generated by an LLM, compared to generic counterspeech, for the *Number of hateful posts*. To evaluate whether our intervention is effective for Twitter/X Premium users, we re-estimate the regression model from our main analysis, adding an interaction term (*Treated x Twitter/X premium (= 1 if premium)*) between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise). The first column presents the regression results comparing GENERIC-EMPATHY with CONTEXT-EMPATHY counterspeech. The second column provides the results comparing GENERIC-WARNING with CONTEXT-WARNING counterspeech. *Treated* corresponds to the intervention with contextualized counterspeech generated by an LLM with the strategy indicated by the column.

	Empathy	Warning-of-consequences
Intercept	0.275	-1.324^{**}
	(0.631)	(0.639)
Empathy	-0.080	
	(0.458)	
Warning-of-consequences		0.677
		(0.419)
Account age (in days)	-0.000^{**}	-0.000
	(0.000)	(0.000)
Followers count	0.000	-0.000
	(0.000)	(0.000)
Following count	0.000	-0.000
	(0.000)	(0.000)
Tweet count	0.000^{**}	0.000^{**}
	(0.000)	(0.000)
Like count	0.000	0.000
	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	-0.488	0.277
	(0.811)	(0.711)
Treated x Twitter/X premium (= 1 if premium)	-0.640	1.559
	(1.066)	(1.064)
Average toxicity (pre-intervention)	7.726^{***}	9.724^{***}
	(2.702)	(2.478)
Number of hateful posts (pre-intervention)	0.811^{***}	0.781^{***}
	(0.044)	(0.041)
Adj. R^2	0.587	0.661
Obs.(N)	1056	1030

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table S17. Estimation results for the treatment effects of interventions with contextualized counterspeech generated by an LLM, compared to generic counterspeech, for the *Relative change in toxicity*. To evaluate whether our intervention is effective for Twitter/X Premium users, we re-estimate the regression model from our main analysis, adding an interaction term (*Treated x Twitter/X premium (= 1 if premium)*) between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise). The first column presents the regression results comparing GENERIC-EMPATHY with CONTEXT-EMPATHY counterspeech. The second column provides the results comparing GENERIC-WARNING with CONTEXT-WARNING counterspeech. *Treated* corresponds to the intervention with contextualized counterspeech generated by an LLM with the strategy indicated by the column.

	Empathy	Warning-of-consequences
Intercept	0.464***	0.322***
	(0.054)	(0.042)
Empathy	0.069**	
	(0.030)	
Warning-of-consequences		0.027
		(0.023)
Account age (in days)	−0.000	−0.000
	(0.000)	(0.000)
Followers count	−0.000	−0.000
	(0.000)	(0.000)
Following count	−0.000	−0.000
	(0.000)	(0.000)
Tweet count	−0.000	−0.000
	(0.000)	(0.000)
Like count	0.000	−0.000
	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0.033	0.032
	(0.047)	(0.058)
Treated x Twitter/X premium (= 1 if premium)	−0.125	−0.029
	(0.079)	(0.068)
Average toxicity (pre-intervention)	−1.568***	−1.106***
	(0.201)	(0.127)
Number of hateful posts (pre-intervention)	0.002	0.003***
	(0.001)	(0.001)
Adj. R^2	0.128	0.110
Obs.(N)	1055	1028

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

9 Robustness checks

To ensure the robustness of our results, we conducted a series of checks: (1) We re-estimated our analysis using a single model that included separate dummy variables for each intervention, instead of estimating separate models for each combination of treatment and control. (2) We pooled all observations in the treatment conditions to assess the overall effect of counterspeech compared to the control group. (3) We pooled observations based on the counterspeech strategies (i.e., empathy vs. warning-of-consequences) and re-estimated our regression model to evaluate their overall effects compared to the control. (4) To compare the effectiveness of contextualized and non-contextualized counterspeech, we pooled all users assigned to contextualized counterspeech across both strategies and repeated the analysis. All robustness checks led to consistent findings. Detailed regression results for each check are in Supplementary Material 9.1 to Supplementary Material 9.4.

9.1 Alternative model specification

Table S18. Estimation results for our regression model where we re-estimated our analysis using a single model that included separate dummy variables for each intervention, instead of estimating separate models for each combination of treatment and control. The first column presents the regression results for the *Rate of deleted posts*. The second column provides the results for *Number of hateful posts*. The third column shows the regression results for *Relative change in toxicity*.

	Rate of deleted posts	Number of hateful posts	Relative change in toxicity
Intercept	0.083^{***}	0.022	0.413^{***}
	(0.018)	(0.500)	(0.037)
GENERIC-EMPATHY	-0026^{*}	-0319	-0016
	(0.014)	(0.478)	(0.022)
CONTEXT-EMPATHY	-0029^{**}	-0440	0.042
	(0.014)	(0.437)	(0.029)
GENERIC-WARNING	0.009	-1020^{**}	-0022
	(0.016)	(0.442)	(0.022)
CONTEXT-WARNING	-0014	-0121	0.005
	(0.014)	(0.446)	(0.023)
Account age (in days)	-0000^{***}	-0000	-0000
	(0.000)	(0.000)	(0.000)
Followers count	-0000	-0000	-0000
	(0.000)	(0.000)	(0.000)
Following count	-0000	-0000	-0000
	(0.000)	(0.000)	(0.000)
Tweet count	-0000^{*}	0.000^{***}	-0000
	(0.000)	(0.000)	(0.000)
Like count	-0000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	-0014	-0205	-0057^{**}
	(0.012)	(0.383)	(0.023)
Average toxicity (pre-intervention)	0.035	8.559^{***}	-1313^{***}
	(0.050)	(1.613)	(0.102)
Number of hateful posts (pre-intervention)	-0001	0.810^{***}	0.002^{***}
	(0.001)	(0.028)	(0.001)
Adj. R^2	0.009	0.605	0.118
Obs.(N)	2647	2621	2617

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

9.2 Pooled regression (Counterspeech vs. control)

Table S19. Estimation results for the treatment effects where we pooled all observations in the treatment conditions to assess the overall effect of counterspeech compared to the control group across our dependent variables. The first column presents the regression results for the *Rate of deleted posts*. The second column provides the results for *Number of hateful posts*. The third column shows the regression results for *Relative change in toxicity*. *Treated* corresponds to whether a user received counterspeech (= 1 if user received counterspeech; = 0 else).

	Rate of deleted posts	Number of hateful posts	Relative change in toxicity
Intercept	0.082***	0.014	0.414***
	(0.018)	(0.500)	(0.037)
Treated (= 1 if user received counterspeech)	−0015	−0460	0.003
	(0.012)	(0.376)	(0.019)
Account age (in days)	−0000***	−0000	−0000
	(0.000)	(0.000)	(0.000)
Followers count	−0000	−0000	−0000
	(0.000)	(0.000)	(0.000)
Following count	−0000	−0000	−0000
	(0.000)	(0.000)	(0.000)
Tweet count	−0000*	0.000***	−0000
	(0.000)	(0.000)	(0.000)
Like count	−0000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0015	−0159	−0056**
	(0.012)	(0.384)	(0.023)
Average toxicity (pre-intervention)	0.039	8.532***	−1317***
	(0.049)	(1.608)	(0.103)
Number of hateful posts (pre-intervention)	−0001	0.810***	0.002***
	(0.001)	(0.028)	(0.001)
Adj. R^2	0.007	0.605	0.116
Obs.(N)	2647	2621	2617

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

9.3 Pooled regression based on counterspeech strategies

9.3.1 Pooled regression comparing control vs. empathy counterspeech.

Table S20. Estimation results for the treatment effects where we pooled all users who received an empathy counterspeech (i.e., GENERIC-EMPATHY and CONTEXT-EMPATHY) to assess the overall effect empathy counterspeech compared to the control group across our dependent variables. The first column presents the regression results for the *Rate of deleted posts*. The second column provides the results for *Number of hateful posts*. The third column shows the regression results for *Relative change in toxicity*. *Treated (Empathy)* corresponds to whether a user received empathy counterspeech (= 1 if user received empathy counterspeech; = 0 else).

	Rate of deleted posts	Number of hateful posts	Relative change in toxicity
Intercept	0.078^{***}	0.345	0.463^{***}
	(0.022)	(0.607)	(0.049)
Treated (Empathy)	-0027^{**}	−0388	0.011
	(0.012)	(0.408)	(0.022)
Account age (in days)	-0000^{***}	−0000	−0000
	(0.000)	(0.000)	(0.000)
Followers count	−0000	0.000	−0000
	(0.000)	(0.000)	(0.000)
Following count	−0000	−0000	−0000
	(0.000)	(0.000)	(0.000)
Tweet count	-0000^{**}	0.000^{***}	−0000
	(0.000)	(0.000)	(0.000)
Like count	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0008	-1115^{**}	-0102^{***}
	(0.016)	(0.516)	(0.031)
Average toxicity (pre-intervention)	0.055	7.499^{***}	-1450^{***}
	(0.067)	(2.123)	(0.147)
Number of hateful posts (pre-intervention)	−0001	0.835^{***}	0.002^{**}
	(0.001)	(0.039)	(0.001)
Adj. R^2	0.009	0.573	0.121
Obs.(N)	1606	1591	1589

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

9.3.2 Pooled regression comparing control vs. warning-of-consequences counterspeech.

Table S21. Estimation results for the treatment effects where we pooled all users who received a warning-of-consequences counterspeech (i.e., GENERIC-WARNING and CONTEXT-WARNING) to assess the overall effect warning-of-consequences counterspeech compared to the control group across our dependent variables. The first column presents the regression results for the *Rate of deleted posts*. The second column provides the results for *Number of hateful posts*. The third column shows the regression results for *Relative change in toxicity*. *Treated (Warning-of-consequences)* corresponds to whether a user received empathy counterspeech (= 1 if user received warning-of-consequences counterspeech; = 0 else).

	Rate of deleted posts	Number of hateful posts	Relative change in toxicity
Intercept	0.088^{***}	−0360	0.373^{***}
	(0.023)	(0.591)	(0.044)
Treated (Warning-of-consequences)	−0003	−0535	−0008
	(0.013)	(0.403)	(0.020)
Account age (in days)	-0000^{**}	0.000	−0000
	(0.000)	(0.000)	(0.000)
Followers count	−0000	-0000^{**}	−0000
	(0.000)	(0.000)	(0.000)
Following count	−0000	−0000	−0000
	(0.000)	(0.000)	(0.000)
Tweet count	−0000	0.000^{***}	−0000
	(0.000)	(0.000)	(0.000)
Like count	−0000	0.000	−0000
	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0025	0.301	−0024
	(0.017)	(0.549)	(0.026)
Average toxicity (pre-intervention)	0.023	8.955^{***}	-1159^{***}
	(0.067)	(2.017)	(0.110)
Number of hateful posts (pre-intervention)	−0001	0.809^{***}	0.003^{***}
	(0.001)	(0.036)	(0.001)
Adj. R^2	0.003	0.615	0.113
Obs.(N)	1582	1565	1562

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

9.4 Pooled regression (Contextualized vs. non-contextualized counterspeech)

Table S22. Estimation results for the treatment effects where we pooled all users who received a contextualized counterspeech (i.e., CONTEXT-EMPATHY and CONTEXT-WARNING) to assess the overall effect contextualized counterspeech generated by an LLM compared to the generic counterspeech across our dependent variables. The first column presents the regression results for the *Rate of deleted posts*. The second column provides the results for *Number of hateful posts*. The third column shows the regression results for *Relative change in toxicity*. *Contextualized counterspeech* corresponds to whether a user received contextualized counterspeech (= 1 if user received CONTEXT-EMPATHY or CONTEXT-WARNING counterspeech; = 0 if users receive GENERIC-EMPATHY or GENERIC-WARNING counterspeech).

	Rate of deleted posts	Number of hateful posts	Relative change in toxicity
Intercept	0.073***	−0586	0.393***
	(0.017)	(0.448)	(0.034)
Contextualized counterspeech	−0012	0.357	0.042**
	(0.010)	(0.288)	(0.018)
Account age (in days)	−0000**	−0000*	−0000
	(0.000)	(0.000)	(0.000)
Followers count	−0000	−0000	−0000
	(0.000)	(0.000)	(0.000)
Following count	−0000	−0000	−0000
	(0.000)	(0.000)	(0.000)
Tweet count	−0000	0.000***	−0000
	(0.000)	(0.000)	(0.000)
Like count	−0000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Twitter/X premium (= 1 if premium)	−0013	0.242	−0045*
	(0.014)	(0.396)	(0.027)
Average toxicity (pre-intervention)	0.034	8.895***	−1335***
	(0.051)	(1.820)	(0.119)
Number of hateful posts (pre-intervention)	−0001	0.795***	0.002***
	(0.001)	(0.029)	(0.001)
Adj. R^2	0.005	0.623	0.119
Obs.(N)	2106	2086	2083

Standard errors are in parentheses; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 28.05.2025

Dominik Bär

Dominik Bär