Unpacking the effects of user anonymity and user popularity on the intensity and diffusion of hate speech on Twitter (X) in Afghanistan

The spread of hate speech on social media, along with its psychological and social harms, potentially even hate crimes, has raised concerns among citizens and policymakers. In response, scholars have explored strategies to reduce hate speech's virality and thus its harms. Using a corpus of 3,210 comments in Persian and Pashtu posted by Twitter users in Afghanistan, we examined how users' anonymity and popularity affect the intensity and diffusion of hate speech. In a series of binary logistic and multiple regression analyses, anonymity showed positive relationships with hate speech's intensity and diffusion on Twitter, whereas user popularity was negatively associated with these factors. A social network analysis also revealed that anonymous accounts were the core nodes in the hate speech cluster and suggested a peer-to-peer (i.e., anonymous user to anonymous user) pattern of interaction. By contrast, non-anonymous users tended to avoid interaction with their anonymous counterparts.

Keywords: *Afghanistan, anonymity, hate speech diffusion, hate speech intensity, Twitter, user popularity*

Acknowledgments

This article has been supported by the National Social Science Foundation of China (Grant No. 23BXW029).

Author information

Qurban Hussain Pamirzad, Xi'an Jiaotong University,
https://orcid.org/0009-0000-0478-1842
Qiang Chen, Xi'an Jiaotong University,
https://orcid.org/0000-0001-8123-077X

How to cite this article:

Pamirzad, Qurban Hussain and Qiang Chen. "Unpacking the effects of user anonymity and user popularity on the intensity and diffusion of hate speech on Twitter (X) in Afghanistan". *Információs Társadalom* XXV, no. 3 (2025): 61–84.
https://dx.doi.org/10.22503/inftars.XXV.2025.3.4

1. Introduction

As the exponential rise of online hate speech promoting animosity and violence gains traction as a global phenomenon (Kilvington 2021; Lingam and Aripin 2017; Williams et al. 2020), its psychological and social negative impacts increasingly attract widespread attention from researchers and policymakers alike (Bilewicz and Soral 2020; Castaño-Pulgarín et al. 2021). Studies have shown that the spread of hate speech in social media feeds and comment sections is facilitated by the technical features and affordances of these platforms —that is, "social media affordances" (Ben-David and Fernández 2016)—with anonymity on social media being one of the most debated and oft-cited factors (Brown 2018; Castaño-Pulgarín et al. 2021; Gorenc 2022; Jaidka et al. 2022). Research has also noted that anonymity contributes to the spread of weaponised information (i.e., mal-, dis-, and misinformation), which itself tends to instigate hate speech (Brown 2018; Gorenc 2022; Nascimento, Cavalcanti and Da Costa-Abreu 2023). Conversely, other findings suggest that anonymity does not directly motivate the spread of hate speech but instead contributes to freedom of speech, deliberative democracy, and other positive outcomes (Ellison et al. 2016; Jaidka et al. 2022; von Essen and Jansson 2018). Beyond that, yet another strand of research has indicated that social media affordances are not solely responsible for inciting and spreading hate speech, for the user's status in the network and their malicious intent, combined with ill-structured language, may also weaponise these tools for the diffusion of hate speech (Ben-David and Fernández 2016; Schmid, Kümpel and Rieger 2024).

Previous studies examining online hate speech have predominantly concentrated on the detection of hate speech (Fortuna and Nunes 2018; Kocoń et al. 2021; Williams et al. 2020), while other aspects have remained underexamined (Chakraborty and Masud 2022), including platform affordances and user elements that influence the intensity and diffusion of hate speech on social media. In our study, we aimed to extend this line of research by investigating how anonymity (i.e., a platform affordance) and user popularity (i.e., a user factor) affect hate speech's intensity and diffusion in the comments section of Twitter. The interactive nature of Twitter's comments section enables users to respond to one another, which often leads to heated discussions on controversial topics that may result in incivility and hate speech (Lingam and Aripin 2017; Zannettou et al. 2020). Despite the abundance of insightful research on hate speech, the factors leading to its intensity and diffusion in comments have received scant attention. Moreover, because research has primarily focused on hate speech in English, its dynamics in other languages have remained largely unclear (Fortuna et al. 2019).

Against this backdrop, we conducted a quantitative content analysis on a dataset of 3,210 tweets in Persian and Pashtu from users in Afghanistan in order to examine how the abovementioned factors affect the intensity and diffusion of hate speech on social media. We also employed social network analysis on the hate speech cluster data to answer two questions:

1. How do anonymous and popular accounts that engage in hate speech rank within hate speech clusters on Twitter?

2. How do levels of user anonymity and popularity influence patterns of interaction in hate speech clusters on Twitter?

The data for our study were collected from Twitter users in Afghanistan, an Asian country that has been embroiled in decades of war and conflict. Afghanistan's social milieu, with profound division along ethnic–lingual and religious lines, has fostered a hostile, even toxic online atmosphere that is an apt case for studying hate speech (Pamirzad 2025). Moreover, aligned with past findings encouraging cross-language and cross-cultural exploration of hate speech (Fortuna et al. 2019), our results provide unique insights into an underexamined topic in Afghanistan as well as in the Persian and Pashtu languages, and thus stand to enrich the literature.

In what follows, we review the relevant literature, articulate the study's hypotheses, and describe the methods employed. After that, we present and discuss the results and provide our conclusions, along with their implications for theory and practice.

2. Literature review and hypotheses

2.1. Hate speech and its intensity

Despite being a buzzword, *hate speech* still lacks a universally agreed-upon definition (Gorenc 2022; Guo and Johnson 2020; Schäfer, Sülflow and Reiners 2021; Ștefăniță and Buf 2021; Vári 2018). As a concept, *hate speech* has been defined from various perspectives. Some scholars have defined it as negative content fraught with swearing, insults, verbal abuse, and hateful derogatory words (Kilvington 2021; Lingam and Aripin 2017) that encompasses all forms of expression that propagate, encourage, support, or legitimize religious hatred, xenophobia, racial hatred, aggressive nationalism, and ethnocentrism, as well as hostility and discrimination targeting minorities, migrants, and other social groups (Parvaresh 2023; Schäfer, Sülflow and Reiners 2021). Other definitions of *hate speech* are associated with its forms, which consist of an array of verbal, nonverbal, symbolic, explicit, and implicit communicative actions involving the use of inappropriate language to attack others (Nascimento, Cavalcanti and Da Costa-Abreu 2023; Parvaresh 2023; Schmid, Kümpel and Rieger 2024; Ștefăniță and Buf 2021).

Hate speech has also been defined from a normative perspective as a form of social deviance—that is, an activity that violates social norms—that runs counter to standard cultural behaviours and interactional norms (Castaño-Pulgarín et al. 2021). Furthermore, social networking sites (SNSs) have their own definitions of *hate speech* that they use as a basis for moderating and filtering out content. Twitter and Facebook, for example, state that any tweet or post that directly attacks or advocates the use of violence against individuals based on their race, ethnicity, national origin, gender, age, disability, or serious illness is considered to be hate speech (Ben-David and Fernández 2016; Mathew et al. 2019).

As a multidimensional concept, hate speech has been classified in different ways. Based on its targeting of social groups, it has been classified into four categories:

political, racial, religious, and gender-based hate speech (Castaño-Pulgarín et al. 2021; Guo and Johnson 2020; Schäfer, Sülflow and Reiners 2021). Posting racist comments, racist humour, and racial stereotypes constitutes racial hate speech, while misogynistic comments containing sexist language represent gender-based hate speech (Saresma, Sanna and Varis 2020). Using derogatory terms and hostile rhetoric, as well as demonising and belittling political opponents, are considered to be forms of political hate speech (Trajkova and Neshkovska 2018). By contrast, posting profane comments, slander or defamation, sarcasm, antisemitism, and Islamophobia can represent religious hate speech (Lingam and Aripin 2017; Ștefăniță and Buf 2021).

Along other lines, scholars interested in detecting hate speech have employed a binary classification—hate speech versus non-hate speech or offensive versus non-offensive content— while considering whether such speech targets a specific group or groups (Fortuna and Nunes 2018; Zampieri et al. 2019). Concerning hate speech's intensity, however, studies have argued that hate speech should be examined beyond that binary classification, for it can range from less offensive and subtly devised to blatantly insulting and violent language (Ruzaite 2018). Bahador (2020) has thus classified hate speech along a spectrum from its lowest (i.e., disagreement) to its highest forms, with the latter being threatening an individual with death or a group with massacre and genocide. This range demonstrates various degrees of the intensity of hate speech, or "hate speech intensity," from mild to highly violent (Fortuna and Nunes 2018; Kocoń et al. 2021; Parvaresh 2023), with the implication that hate speech is not a one-size-fits-all phenomenon but exists on a continuum of hate. In that vein, scholars have compared Gordon W. Allport's (1954) spectrum of racism to the continuum of hate speech (Sachdeva et al. 2022), beginning with prejudiced verbal language as racism's weakest manifestation to actual extermination as its strongest. Consequently, hate speech, beyond its adverse individual-level psychological impacts, including fear, depression, unhappiness, anxiety, desensitisation, and post-traumatic stress (Bilewicz and Soral 2020), can also lead to social avoidance, discrimination, physical attacks, and intended extermination (Bilewicz and Soral 2020; Sachdeva et al. 2022). On social media, an individual may block or unfriend someone to avoid exposing their hate speech (i.e., social avoidance), which can reinforce the discrimination and hostility between social groups (Chakraborty and Masud 2022; Lingam and Aripin 2017). Likewise, high-intensity hate speech that incites and promotes violence and physical harm can spill over into real-life settings by fueling hate crimes such as the attack on a synagogue in Pittsburgh, PA, and the shooting in a mosque in Christchurch, New Zealand (Maarouf, Pröllochs and Feuerriegel 2024; Mathew et al. 2019; Pamirzad 2025).

2.2. Impact of user anonymity on user popularity

The anonymity of users on social media, or "user anonymity," involves using these tools without sharing identifiable information (Backes et al. 2016; Curlew 2019;

Gulyás 2017). Users may choose anonymity based on different reasons; some prefer to remain anonymous or semi-anonymous in order to keep the size of their network manageable and only known to people in real life, whereas others may choose to be anonymous even among their friends and relatives (Ellison et al. 2016; Ma, Hancock and Naaman 2016). Therefore, unlike in the real world, social media anonymity affords users flexibility in selecting their identities. Such a customizable identity may help users to reduce mobbing on their online networks; however, it can also facilitate the spread of hate speech or counter-normative actions by some users (Castaño-Pulgarín et al. 2021). Moreover, users may choose anonymity as a strategy of online activism to reduce threats and perceived risks in repressive political environments (Ellison et al. 2016; Jardine 2018). Considering the pro- and antisocial potential of anonymity, various social media platforms have adopted different measures. For instance, Facebook, addressing the negative aspect of anonymity, has adopted a real-name authentication policy to increase the quality of content and accountability and decrease misconduct such as spamming, bullying, hacking, and spreading hate speech (Peddinti, Ross and Cappos 2017). Conversely, Twitter accentuates positive aspects of anonymity as contributing to freedom of speech and thus allows users to choose their preferred level of identifiability (Backes et al. 2016).

Although the positive and negative aspects of user anonymity have been explored (Brown 2018; Ellison et al. 2016; Jaidka et al. 2022; Kilvington 2021; Ma, Hancock and Naaman 2016; Zannettou et al. 2020), its impacts on the popularity of users on social media, or "user popularity," have not received sufficient attention. *User popularity* refers to the size of a user's network and their number of followers, which enhances their centrality in the network (Garcia et al. 2017; Vedadi and Greer 2021). Research has shown that user popularity on social media is linked to personalisation, authenticity, trust, and perceived realness (Rutledge 2021; Yuan and Lou 2020). Popular users, also known as opinion leaders, use personalisation to enhance and elevate their standing within the network. By actively engaging with their followers, they create an authentic, relatable online presence. By contrast, anonymity is rooted in uncertainty and disingenuousness. Anonymous users who withhold identifiable information cast themselves as enigmatic figures with unknown personalities (Alexopoulou and Pavli 2021).

Even so, some anonymous Twitter accounts defy this norm of identifiability by becoming popular nevertheless. The @YourAnonNews account, for instance, boasts more than 7.5 million followers on Twitter, possibly due to their statuses and the content that they publish, which aligns with the highly polarised global landscape. Events such as Israel–Palestine and Russia–Ukraine conflicts have profoundly divided people worldwide (Milmo 2022), and the polarising posts of anonymous accounts resonate with the polarised public and thus explain their popularity. On a micro level, however, we maintain the conventional argument that identifiability is the primary source of user popularity (Yuan and Lou 2020). Thus, we first hypothesised that:

H1: User anonymity is negatively associated with user popularity on Twitter.

2.3. User anonymity's impact on the intensity and diffusion of hate speech

User anonymity refers to the avoidance of disclosing personal or socially identifiable information on social media (Backes et al. 2016; Gulyás 2017). It is a continuum from identifiability to anonymity (Eklund et al. 2022)—for instance, from complete anonymity on Yik Yak, Whisper, and Secret, where no traceable information of users exists, to partial anonymity and pseudonymity on conventional platforms such as Twitter (Curlew 2019; Ellison et al. 2016; Peddinti, Ross and Cappos 2017). User anonymity can be further classified into personal identity and social identity anonymity (Jaidka et al. 2022). Whereas *personal identity anonymity* refers to the absence of identifiable information about individuals (e.g., name, email address, and phone number), *social identity anonymity* refers to the absence of users' identifiable information about their social, political, and ideological connections. People can be personally anonymous but socially identifiable by exposing signs of affiliation to a social or political group on their accounts. Though they may use pseudonyms, their profile pictures, posted content, hashtags, and bios can reveal their social identities (Jaidka et al. 2022).

Studies have revealed social media anonymity's positive and negative aspects and its use for pro- or antisocial purposes (Ellison et al. 2016). As for positives, it enables people to discuss topics that they might otherwise avoid by protecting their privacy and thus facilitates freedom of speech in suppressive political environments. Such freedoms include criticising an official or flagging flaws and corruption in the system, for anonymity strengthens the user's feeling of perceived safety (Brown 2018). Studies have also indicated that anonymity benefits women, likely by decreasing their identifiability and making them less prone to harassment on social media (Ma, Hancock and Naaman 2016).

Concerning the spread of hate speech, or "hate speech diffusion," studies have additionally shown that anonymity's effect varies based on its type. Research has revealed that having an anonymous personal identity while maintaining a non-anonymous social identity increases the quality of political discussion by fostering rationality and civil discourse (Jaidka et al. 2022). However, most studies have added that anonymity is also associated with offensive and aggressive behaviours; it incites violence, promotes discrimination against individuals and social, political, and gender groups, and motivates extremism, bigotry, and propaganda (Brown 2018; Castaño-Pulgarín et al. 2021; Gorenc 2022; Zannettou et al. 2020). This stream of research has also suggested that anonymity provides users with a sense of safety by making them feel less obliged and accountable to observe conventional behavioural norms and boundaries, which raises their likelihood of disseminating hate speech. Similarly, anonymous users can become deindividuated and disinhibited and turn more violent and aggressive, which amplifies the intensity of hate speech (Brown 2018; Ellison et al. 2016; Kilvington 2021; Zannettou et al. 2020). Consequently, we also hypothesised that:

H2: User anonymity is positively associated with hate speech diffusion on Twitter.
H2a: User anonymity is positively associated with hate speech intensity on Twitter.

2.4. User popularity's impact on hate speech intensity and diffusion

On social media, *user popularity* refers to the size of a user's network, implying their reach and influence based on the number of followers and their centrality within a network (Balaban et al. 2020). Popular users enjoy a high degree of prominence, and their involvement in hate speech diffusion can significantly impact the overall network (Riquelme and González-Cantergiani 2016). Research has shown that promoters of hate speech on Twitter have large numbers of followers, followees, group memberships, and like counts, which indicate the involvement of popular accounts in hate speech diffusion. However, it has remained unknown whether the association between user popularity and hate speech diffusion is statistically significant (Perera et al. 2023). Furthermore, user popularity's influence on hate speech intensity has yet to be investigated.

According to previous studies, users gain popularity through sincere relationship-building with their followers based on respect, mutual trust, and personal affection (Men et al. 2018; Yuan and Lou 2020). Studies have also shown that fame on social media is riskier than offline, because the association between online popular users and their followers hinges on a sense of personal closeness intertwined with the followers' emotions (Rutledge 2021). Thus, any mistakes the popular users commit can sway followers and swiftly diminish their popularity (Rutledge 2021). Thus, popular users may be less likely to participate in hate speech diffusion in order to avoid losing users' trust, respect, and affection. Moreover, to avoid being targeted with reciprocal hostility due to posting hate comments, which could damage their fame (Ellison et al. 2016; Kilvington 2021), popular users are unlikely to spread hate speech at an intense level. Thus, we additionally hypothesised that:

H3: User popularity is negatively associated with hate speech diffusion on Twitter.
H3a: User popularity is negatively associated with hate speech intensity on Twitter.

Figure 1. summarises our research model in relation to our hypotheses.

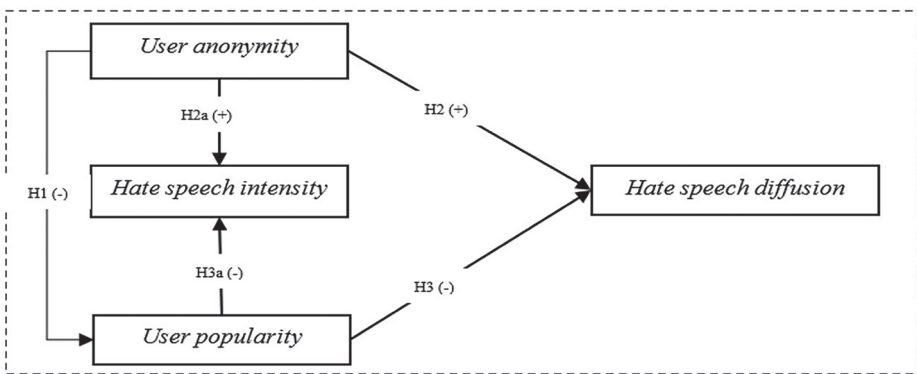

Figure 1. Research model with hypotheses

3. Method

3.1. Sampling and data collection

In recent years, hate speech has increased markedly online, and Twitter has become a widely studied platform regarding the phenomenon (Matamoros-Fernández and Farkas 2021). In Afghanistan, following the Taliban's takeover in 2021, unprecedented social and political restrictions have resulted in censorship and self-censorship, and many social media users have opted to create fake accounts on social media, particularly Twitter. Since then, hate speech has dramatically increased among Twitter users in Afghanistan (Pamirzad 2025), and some mainstream media analysts and popular users have even been involved. Using the keyword inquiry approach, we chose certain contentious viral events susceptible to inciting hatred and searched for the most often recurring terms related to political, social, ethnic, and religious hate speech in Afghanistan's sociopolitical context. The cases include an online campaign named "منم افغان نیستم" ('I am not Afghan'), Afshar "افشار" (i.e., a massacre in 1993 that is remembered every February), and a poem recitation in March 2024 that caused heated discussions and hatred (Pamirzad 2025). Both keyword searching and tracking polarised events that have incited hate speech and hostility online have been used in past studies to extract data.

To narrow our sample, we adopted a criterion that allowed only posts that received more than 20 comments related to keywords and case inquiry to be included. The criterion was adopted based on the idea that comments are a quantitative measure of the virality and profundity of discussions on social media, whereas posts with fewer comments lack such features (Konovalova et al. 2023; Pamirzad 2025), which may not contribute to the depth of knowledge. Consequently, 62 posts that met that condition were selected, and their comments were extracted in June 2024, with comments spanning the period from April 2019 to March 2024. After discarding the duplicates, the final sample in our manual content analysis contained 3,210 comments. Twitter Replies Exporter, a browser extension, was used to extract the data, and SPSS version 27 was used for quantitative analysis. Moreover, a network file was designed using comment sources as nodes and replies received as edges, and Gephi 0.10 software was used for network analysis.

3.2. Operationalisation of variables

For user anonymity, we adopted the approach proposed by Esteve, Moneva and Miró-Llinares (2019) and Peddinti, Ross and Cappos (2017), which categorises users' metadata into four levels of anonymity: anonymous, partly anonymous, identifiable (i.e., non-anonymous), and unclassifiable. Further refining that classification, we followed Jaidka et al. (2021) by differentiating personal from social anonymity. According to Jaidka et al., if an individual is personally anonymous but has a link, a brief bio, or a symbolic feature that signals their social membership, they can be

regarded as partly anonymous. For that reason, we classified anonymity into three categories—non-anonymous, semi-anonymous, and anonymous—and coded the variable using users' metadata in the dataset (see Table 1).

Type of anonymity and codes	Definition	Sources
Non-anonymous (0)	Users with conventional names (i.e., names for humans, particularly in Persian and Pashto), congruency between their name and Twitter ID, clear social identities, or jobs in their bios	Esteve et al. (2019) and Peddinti et al. (2017)
Semi-anonymous (1)	Users with ambiguous personal information (e.g., first and last names) but socially identifiable details (e.g., ethnicity, political party, location, or employer)	Jaidka et al. (2021) and Peddinti et al. (2017)
Anonymous (2)	Users without traceable information on their profiles (e.g., first and last names) or with unconventional names (e.g., names of objects or unknown characters that are not consistent with human names, particularly in Persian and Pashtu), and users without clues in their bios	Esteve et al. (2019) and Peddinti et al. (2017)

Table 1. Coding manual for user anonymity

User popularity refers to the extent of a user's in-degrees and centrality in a network (Balaban et al. 2020), measured by the number of followers. We used the number of users' followers to measure popularity and log-transformed it to achieve a normal distribution (M = 4.78, SD = 2.03; Zhang et al. 2023). Moreover, using the visual binning function in SPSS, we transformed the index into four clusters, with cutoff points based on ±1 *SD* in relation to the mean to measure different levels of popularity. Below the mean, the lowest cluster was labelled *unpopular* (n = 491), and the second-lowest, somewhat unpopular (n = 1,162); above the mean, the first cluster was labelled somewhat popular (n = 1,045), and the cluster above it, *highly popular* (n = 512).

Hate speech intensity refers to the strength of the tone, meaning, and expression of hatred, as well as the targeted group's perception of such meaning. As shown in Table 2, we adopted, integrated, and modified the hate speech intensity classifications of Bahador (2020), Fortuna, Soler-Company and Wanner (2020), and operationalised the variable.

Levels and codes of hate speech intensity	Definition
No hate speech (0)	Comments that do not contain any hate speech
Mild hate speech (1)	Comments that contain offensive, derogatory terms and slurs but do not advocate prejudice, violence, or harm
Moderate hate speech (2)	Comments that contain discriminatory words targeting individuals based on their immutable characteristics (e.g., nationality, religion, ethnicity, gender, age, and sexual orientation) and express dislike or loss of empathy
Strong hate speech (3)	Comments that use harmful stereotypical expressions containing prejudice, demonisation, dehumanisation, and belittlement toward a specific group
Severe hate speech (4)	Violent, abusive comments about a specific individual or group, justifying violence, explicit threats, and/or the incitement of violence against them
Extreme hate speech (5)	Comments containing blatant, abusive, and/or insulting language that promote and glorify violence against a specific group, including threats of death or genocide

Table 2. Coding manual for hate speech intensity based on definitions in Bahador (2020) and Fortuna et al. (2020)

Hate speech diffusion refers to the spread of hate speech in the network (Tontodimamma et al. 2021). In our study, we were particularly interested in identifying agents of hate speech diffusion. Following the approaches of Fortuna and Nunes (2018) and Zampieri et al. (2019), we recoded the data coded for hate speech intensity into a binary variable, such that comments containing hate speech were assigned a value of 1 and those without hate speech were assigned a value of 0.

3.3. Intercoder reliability

After we designed the codebook and trained an undergraduate assistant, 5% of the data (n = 160) was independently coded by the first author and the assistant to ensure intercoder reliability. The Fleiss interrater test was conducted to ensure reliability in user anonymity and hate speech intensity; it is appropriate for multiple coders or variables with more than two categories (Fleiss, Nee and Landis 1979). The overall agreement on user anonymity was .80 with the categories *anonymous* (.88), *semi-anonymity* (.75), and *non-anonymity* (.76). Similarly, the overall agreement

for hate speech intensity was .78 with the categories *no hate speech* (.93), *mild hate speech* (.70), *moderate hate speech* (.72), *strong hate speech* (.71), *severe hate speech* (.70), and *extreme hate speech* (.91). The overall results fall between the acceptable range of moderate to substantial agreement (Fleiss, Nee, and Landis 1979).

4. Results

4.1. Descriptive statistics

In our sample, 52.1% of users were anonymous, 8.8% were semi-anonymous, and 39.1% were non-anonymous. Whereas 50.1% of the sample represented no hate speech, 14.0% represented severe hate speech, 11.3% represented extreme hate speech, 6.7% represented strong hate speech, 11.8% represented moderate hate speech, and 6.1% represented mild hate speech. The majority of highly popular semi-anonymous accounts were involved in spreading extreme hate speech, whereas most highly popular non-anonymous users were engaged in mild and moderate hate speech. Table 3 summarises the correlations between independent and dependent variables included in the data analysis.

Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. Hate speech intensity								
2. Hate speech diffusion	.861**							
3. Anonymous	.052**	.062**						
4. Semi-anonymous	.010	−.003	−.323**					
5. Non-anonymous	−.059**	−.062**	−.836**	−.248**				
6. Unpopular	.039*	.042*	−.054**	−.017	.065**			
7. Somewhat unpopular	.024	.014	.025	−.027	−.010	−.306**		
8. Somewhat popular	−.004	−.002	.054**	−.003	−.053**	−.304**	−.522**	
9. Highly popular	−.065**	−.056**	−.049**	.055**	.019	−.186**	−.319**	−.318**

Table 3. Correlations between dependent and independent variables

4.2. Hypothesis testing

A linear regression was conducted to test H1, which proposed that anonymity negatively influences user popularity. The results, $f(2) = 779.47$ ($p < .001$) and adjusted $R2 = .33$, did not support H1 by showing that anonymity anonymous ($b = 0.14$, $p < .05$) and anonymity semi-anonymous ($b = 0.28$, $p < .01$) compared with non-anonymous users positively and significantly affected user popularity when the user popularity index logged ($M = 4.78$, $SD = 2.03$) was used as a scale variable. This contradictory finding may stem from a lack of accountability associated with anonymity that allows users to post hateful content without fear of personal repercussions (Postmes and Spears 1998). That allowance promotes engagement through conflict and forms echo chambers and leads to their popularity (ElSherief et al. 2018).

A binary logistic regression analysis was conducted to test hypotheses H2 and H3, which proposed that user anonymity positively and popularity negatively influence hate speech diffusion. User anonymity (i.e., non-anonymous, semi-anonymous, and anonymous) and user popularity (i.e., highly popular, somewhat popular, somewhat unpopular, and unpopular) as independent variables, and hate speech diffusion with binary categories (i.e., hate speech vs. no hate speech) as the dependent variable were entered into the model. Although the model showed low variance in the dependent variable (Cox and Snell's $R2 = .09$ and Nagelkerke's $R2 = .011$), the Hosmer and Lemeshow test, $\chi^2 (7) = .689$ ($p = .998$), indicated that the model adequately captured the relationship between the independent and dependent variables.

Variable	b (SE)	Wald	Exp (b)	95% CI
Intercept	0.096 (.09)	.940	1.100	
User anonymity (ref. non-anonymous)				
Anonymous	0.282*** (.07)	14.008	1.326	[1.144, 1.537]
Semi-anonymous	0.161 (.13)	1.478	1.175	[.906, 1.524]
User popularity (ref. unpopular)				
Highly popular	−0.471*** (.13)	13.609	0.625	[.486, .802]
Somewhat popular	−0.303** (.11)	7.494	0.739	[.595, .918]
Somewhat unpopular	−0.244* (.11)	5.023	0.784	[.633, .970]

Note. All entries are unstandardized coefficients with standard errors (*SE*) in parentheses.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 4. Logistic regression predicting the effects of user anonymity and popularity on hate speech diffusion

The results in Table 4, which partly support H2, indicate that user anonymity positively influenced hate speech diffusion. For every unit of increase in anonymous users compared with non-anonymous users as the reference category, the odds ratio of hate speech diffusion increased by 32.6% (Exp (.282) ≈ 1.326, $p < .001$); however, the semi-anonymous users were non-significant predictors of hate speech diffusion. Similarly, the findings supporting H3 showed that for every unit of increase in highly popular users, somewhat popular users, and somewhat unpopular users compared with unpopular users, the odds ratio of hate speech diffusion decreased by 37.5% (Exp (−.471) ≈.625, $p < .001$), 26.1% (Exp (−.303) ≈.739, $p < .01$), and 21.6% (Exp (−.244) ≈.784, $p < .05$), respectively.

Furthermore, a multiple linear regression (MLR) analysis was conducted to examine the association of user anonymity (H2a) and user popularity (H3a) with hate speech intensity.

Variable	**b (SE)**	**t**	**95% CI**
Intercept	1.741 (.092)	18.895	[1.560, 1.921]
User anonymity (Ref. non-anonymous)			
Anonymous	0.235** (.07)	3.333	[.097, .373]
Semi-anonymous	0.226† (.12)	1.820	[−.018, .470]
User popularity (Ref. unpopular)			
Highly popular	−0.545*** (.12)	−4.584	[−.778, −.312]
Somewhat popular	−0.302** (.10)	−2.923	[−.504, −.099]
Somewhat unpopular	−0.211* (.10)	−2.081	[−.410, −.012]
R^2	**.010**		
***f* test**	**(5, 3204) = 6.720*** **		
N	**3,210**		

Note. All entries are unstandardized coefficients with standard errors (*SE*) in parentheses.
CI = confidence interval. $^{\dagger}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 5. MLR predicting the effects of user anonymity and user popularity on hate speech intensity

As shown in Table 5, user anonymity, $_{\text{anonymous}}$ ($b = 0.235$, $p < .01$), compared with non-anonymous as the reference category, significantly and positively predicted hate speech intensity. Similarly, user anonymity, specifically $_{\text{semi-anonymous}}$ (B = 0.226, $p < .1$), compared with non-anonymous as the reference category, positively influenced hate speech intensity, albeit marginally. Thus, H2a was supported. Moreover, user popularity, $_{\text{highly popular}}$ (B = −.545, $p < .001$) compared with the unpopular as the reference category, significantly and negatively influenced hate speech intensity.

Similarly, user popularity $_{\text{somewhat popular}}$ (B = −.302, $p < .01$) and $_{\text{somewhat unpopular}}$ (B = −.211, $p < .05$), compared with the reference category, significantly and negatively influenced hate speech intensity. Therefore, H3a was also supported.

4.3. Social network analysis

A social network analysis was conducted to answer the research questions about the rank of anonymous and popular accounts within the hate speech cluster and how user anonymity and popularity influence patterns of interaction on the social network. The social network was designed by assigning the users' comments as nodes and the resulting interactions as edges.

ID	Anonymity	Popularity	Hate speech intensity	Closeness centrality	Betweenness centrality
18	Non-anonymous	Somewhat popular	Strong	0.42	8,278.31
153	Anonymous	Highly popular	Moderate	0.37	4,614.60
126	Anonymous	Unpopular	Moderate	0.38	3,941.54
129	Anonymous	Highly popular	Moderate	0.34	3,815.92
16	Anonymous	Highly popular	Moderate	0.33	3,319.39
34	Non-anonymous	Highly popular	Moderate	0.38	3,230.55
23	Anonymous	Somewhat popular	Strong	0.36	2,894.82
109	Non-anonymous	Somewhat unpopular	Strong	0.32	2,039.80
32	Non-anonymous	Somewhat popular	Strong	0.28	1,817.11
106	Anonymous	Highly popular	Moderate	0.30	1,736.87

Table 6. Description of network metrics

Table 6 shows the network's top 10 nodes, including anonymous and non-anonymous users, based on their high betweenness centrality. After filtering out the no-hate-speech category of data, six of the top 10 influential nodes in the network were anonymous users. This outcome highlights the centrality of anonymous accounts in the hate speech network (Bloch, Jackson and Tebaldi 2023; Tabassum et al. 2018) and their intermediary role in sustaining and fueling relevant hateful discussions. Concerning hate speech intensity, the top 10 nodes involved in hate speech showed varying levels, from moderate to strong hate speech.

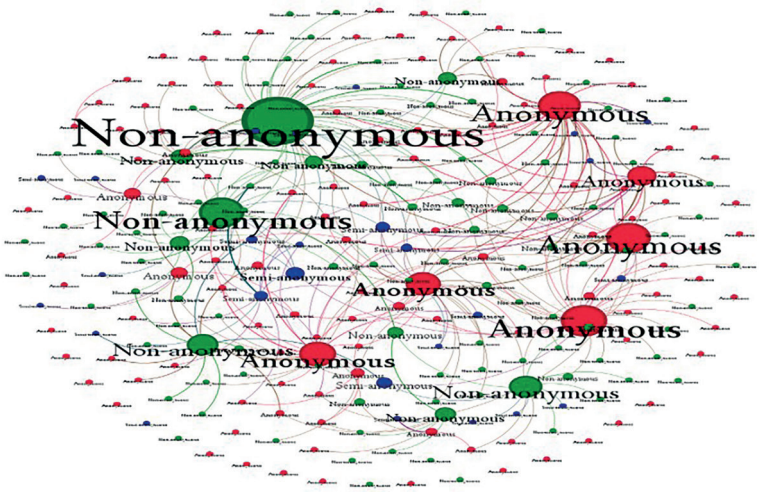

Figure 2. The nodes are coloured by categories of anonymity; red nodes indicate anonymous, green non-anonymous, and blue semi-anonymous users in the network. Edges share the node colour if both endpoints match, or use a mixed colour when categories differ.

Regarding popularity, most anonymous accounts were highly popular, whereas most non-anonymous ones were somewhat popular. Figure 2 also displays a dominant peer-to-peer interaction between anonymous users, as well as interaction avoidance between anonymous and non-anonymous accounts. Meanwhile, semi-anonymous accounts interacted more with non-anonymous users, whereas a cluster of anonymous users also interacted with non-anonymous accounts.

5. Discussion and conclusion

Using a corpus of 3,210 tweets in Persian and Pashtu, we examined how user anonymity and user popularity affect the intensity and diffusion of hate speech among Twitter users in Afghanistan. Our findings suggest that an increase in anonymous users compared with non-anonymous users is associated with a corresponding rise in the diffusion of hateful comments. This finding aligns with the results of past research, which has shown that anonymity affordance on social media engenders a sense of safety that reduces the user's adherence to conventional behavioural norms and their accountability for spreading hate speech (Fortuna and Nunes 2018; Kocoń et al. 2021; Parvaresh 2023; von Essen and Jansson 2018). This finding was further substantiated by the centrality of anonymous users as core nodes and grand connectors in the hate clusters identified in social network analysis (Bloch, Jackson

and Tebaldi 2023; Tabassum et al. 2018). However, our nuanced classification extending beyond the anonymous versus non-anonymous dichotomy revealed that semi-anonymous users, who occupy a rank between anonymity and identifiability, were non-significant predictors of hate speech diffusion. It suggests that certain levels of identity customisation on social media may not inherently lead to adverse outcomes (Jaidka et al. 2022).

The findings also revealed that anonymous users posted more intense hate comments than their non-anonymous counterparts. This outcome can be explained by deindividuation theory, which posits that anonymity prompts the erosion of internal constraints, individual identity, and behavioural accountability (Postmes and Spears 1998). Consequently, individuals become less concerned about guilt, shame, or fear when engaging in aggressive behaviour (Vilanova et al. 2017). According to this theory, anonymous users exhibit less concern about the negative effects of spreading violent and aggressive comments on others and feel less responsible and accountable for their actions (Zapata et al. 2024). Furthermore, beyond the perception of physical safety, anonymity provides a psychological shield that enables individuals inadvertently caught up in hate speech to respond aggressively and simultaneously maintain their social standing. When an individual's personal or social identity is targeted, they may use anonymity to retaliate and vent frustration while concealing their identity to avoid being perceived as impolite and thus safeguard their personality.

When it comes to hate speech, anonymity also fosters a dual psychological protective mechanism. First, anonymous individuals may feel safer and less responsible when engaging in hate speech due to their concealed identity. Second, if they become the target of hate speech themselves, the loss of identity (i.e., deindividuation) shields them from victimisation and reinforces their aggressive behaviour. This explanation gains further relevance considering the peer-to-peer pattern of interaction between anonymous users. Our social network analysis revealed that anonymous users are more likely to interact with one another, which can be attributed to the dual psychological shields that protect them and make them feel less accountable for their behaviour and less aware of reciprocal hatred when targeted (Postmes and Spears 1998). By contrast, non-anonymous users may avoid interaction with anonymous counterparts to prevent becoming the target of aggressive behaviours.

We also investigated user popularity, meaning a user's centrality and influence in a network (Garcia et al. 2017; Vedadi and Greer 2021), regarding its role in hate speech dynamics. Our findings revealed that user popularity was negatively associated with both the intensity and diffusion of hate speech, thereby indicating that the number of hate comments decreased as the number of popular users rose. Similarly, the finding suggests that hate speech intensity dropped significantly as the user's popularity increased. This highlights the potential of popular users in combating the so-called infodemic of hate speech (Masud et al. 2021). Popular users have numerous followers and massive networks on social media, which are considered to be valuable assets in terms of social capital and monetisation (Men et al. 2018; Yuan and Lou 2020). Their positive potential in combating hate speech is promising and

can be leveraged to fight the infodemic. Contrary to previous studies examining high followers, followees, and likes among hate users at the descriptive level (Perera et al. 2023), we found significant evidence that user popularity was inversely associated with hate speech diffusion and intensity. According to previous studies, online popularity is a risk-vulnerable property that can quickly vanish if followers' trust is damaged (Rutledge 2021); hence, our findings can be elucidated based on popular users' perception of risk avoidance. Posting hate speech and targeting others with intensely hateful language by popular users can damage their followers' sentiments and may shrink their audience—that is, the source of their fame and monetisation. Therefore, popular users may avoid engaging in hate speech in order to minimise the risk of becoming the target of hate speech or losing followers (ElSherief et al. 2018). However, some anonymous accounts also become popular, probably because they post hate speech or inflammatory comments. This phenomenon occurs in polarised online echo chambers, where anonymous accounts spearhead hate campaigns, attract like-minded individuals, and thereby increase their centrality (ElSherief et al. 2018).

To conclude, our findings confirm that anonymity is associated with the intensity and diffusion of hate speech. This result is consistent with published findings, which suggest that anonymity promotes users' deindividuation and disinhibition, thereby making them more aggressive and less attentive to the negative impact of their behaviours on others. These findings have practical implications for social media networks. Although studies have shown that discussions on Twitter have been more uncivil than on Facebook (Oz, Pei and Gina 2018), further cross-platform comparative analysis is required to reveal whether the level of incivility on Twitter is associated with its anonymity affordance. If so, then SNSs, particularly Twitter, should adopt a stricter stance against anonymous hate promoters. By contrast, our findings also suggest that user popularity negatively relates to the intensity and diffusion of hate speech, possibly because popular users and opinion leaders on Twitter, primarily politicians, journalists, analysts, and experts, predominantly prefer to be known by their real-life identities. Spreading hate speech, however, contradicts their personae and professions and concurrently damages their reputation. Nonetheless, in other instances, popular accounts, whether anonymous or non-anonymous, become the forerunners of potentially polarising hate-filled discussions.

5.1. Implications and limitations

We have introduced a bifactor model that enriches the literature addressing hate speech on social media by exploring the effects of anonymity affordance and user popularity on the intensity and diffusion of hate speech. Moreover, in response to calls for multilevel anonymity and hate speech intensity (Eklund et al. 2022; Zampieri et al. 2019), we proposed an exploratory taxonomy that warrants further exploration in future research. From a practical perspective, our findings can assist policymakers in formulating legal frameworks and policies regarding anonymity on social media to balance its pro- and antisocial functions and curb the widespread,

harmful virality of hate speech online. Furthermore, these insights can guide SNS companies in adopting filtering policies based on hate speech intensity with varying degrees of tolerance, thereby contributing to a healthy online ecosystem while preserving freedom of speech and relevant criticism (Schäfer, Sülflow and Reiners 2021).

As for our study's limitations, the data were collected from users of Twitter in Afghanistan, which has unique sociocultural features and a hostile, toxic political atmosphere. Therefore, our findings may not be generalizable to other societies and linguistic contexts (Farrand 2023), and further exploration is required to enhance the generalizability of our results. Furthermore, to measure user anonymity, we relied on self-reported profile information, the verification of which is inherently difficult and necessitates innovative techniques in future research (Peddinti, Ross and Cappos 2017). Finally, based on our dataset, we examined the association between dependent and independent variables, rather than causality; further experimental research is required to establish causal relationships.

References

Alexopoulou, Sofia, and Antonia Pavli. "Beneath This Mask There Is More Than Flesh, Beneath This Mask There Is an Idea: Anonymous as the (Super)heroes of the Internet?" *International Journal for the Semiotics of Law – Revue Internationale de Sémiotique Juridique* 34, no. 1 (2021): 237–64. https://doi.org/10.1007/s11196-019-09615-6.

Asimovic, Nejla, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker. "Testing the Effects of Facebook Usage in an Ethnically Polarized Setting." *Proceedings of the National Academy of Sciences* 118, no. 25 (2021): e2022819118. https://doi.org/10.1073/pnas.2022819118

Backes, Michael, Pascal Berrang, Oana Goga, Krishna P. Gummadi, and Praveen Manoharan. "On Profile Linkability Despite Anonymity in Social Media Systems." In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, 25–35. Vienna, Austria: Association for Computing Machinery, 2016. https://doi.org/10.1145/2994620.2994629

Bahador, Babak. "Classifying and Identifying the Intensity of Hate Speech." *Items* (digital forum). Social Science Research Council, November 17, 2020. Accessed October 12, 2025. https://items.ssrc.org/disinformation-democracy-and-conflictprevention/classifying-and-identifying-the-intensity-of-hate-speech

Balaban, Delia, Ioana Iancu, Maria Mustățea, Anișoara Pavelea, and Lorina Culic. "What Determines Young People to Follow Influencers? The Role of Perceived Information Quality and Trustworthiness on Users' Following Intentions." *Romanian Journal of Communication and Public Relations* 22, no. 3 (2020): 5–19. https://doi.org/10.21018/rjcpr.2020.3.306

Ben-David, Anat, and Ariadna Matamoros Fernández. "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain." *International Journal of Communication* 10 (2016): 27–49. https://ijoc.org/index.php/ijoc/article/view/3697/1585

Bilewicz, Michał, and Wiktor Soral. "Hate Speech Epidemic: The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization." *Political Psychology* 41 (2020): 3–33. https://doi.org/10.1111/pops.12670

Bloch, Francis, Matthew O. Jackson, and Pietro Tebaldi. "Centrality Measures in Networks." *Social Choice and Welfare* 61, no. 2 (2023): 413–53. https://doi.org/10.1007/s00355-023-01456-4.

Brown, Alexander. "What Is So Special About Online (as Compared to Offline) Hate Speech?" *Ethnicities* 18, no. 3 (2018): 297–326. https://doi.org/10.1177/1468796817709846

Castaño-Pulgarín, Sergio Andrés, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. "Internet, Social Media and Online Hate Speech: Systematic Review." *Aggression and Violent Behavior* 58 (2021): 101608. https://doi.org/10.1016/j.avb.2021.101608

Chakraborty, Tanmoy, and Sarah Masud. "Nipping in the Bud: Detection, Diffusion and Mitigation of Hate Speech on Social Media." *ACM SIGWEB Newsletter* (Winter 2022): 1–9. https://doi.org/10.1145/3522598.3522601

Curlew, Abigail E. "Undisciplined Performativity: A Sociological Approach to Anonymity." *Social Media + Society* 5, no. 1 (2019). https://doi.org/10.1177/2056305119829843

Eklund, Lina, Emma von Essen, Fatima Jonsson, and Magnus Johansson. "Beyond a Dichotomous Understanding of Online Anonymity: Bridging the Macro and Micro Level." *Sociological Research Online* 27, no. 2 (2022): 486–503. https://doi.org/10.1177/13607804211019760

Ellison, Nicole B., Lindsay Blackwell, Cliff Lampe, and Penny Trieu. "The Question Exists, but You Don't Exist with It: Strategic Anonymity in the Social Lives of Adolescents." *Social Media + Society* 2, no. 4 (2016). https://doi.org/10.1177/2056305116670673

ElSherief, Mai, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. "Peer to Peer Hate: Hate Speech Instigators and Their Targets." In *Proceedings of the International AAAI Conference on Web and Social Media* 12, no. 1 (2018): 52–61. https://doi.org/10.1609/icwsm.v12i1.15038

Esteve, Zoraida, Asier Moneva, and Fernando Miró-Llinares. "Can Metadata Be Used to Measure the Anonymity of Twitter Users? Results of a Confirmatory Factor Analysis." *International E-Journal of Criminal Sciences* 13 (2019): 4. https://ojs.ehu.eus/index.php/inecs/article/view/21157/19010

Farrand, Benjamin. "'Is This a Hate Speech?' The Difficulty in Combating Radicalisation in Coded Communications on Social Media Platforms." *European Journal on Criminal Policy and Research* 29, no. 3 (2023): 477–493. https://doi.org/10.1007/s10610-023-09543-z

Fleiss, Joseph L., John C. Nee, and J. Richard Landis. “Large Sample Variance of Kappa in the Case of Different Sets of Raters.” *Psychological Bulletin* 86, no. 5 (1979): 974–77. https://doi.org/10.1037/0033-2909.86.5.974

Fortuna, Paula, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. “A Hierarchically-Labeled Portuguese Hate Speech Dataset.” In *Proceedings of the Third Workshop on Abusive Language Online*, 94–104. Florence: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/W19-3510

Fortuna, Paula, and Sérgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text.” *ACM Computing Surveys* 51, no. 4 (2018): 1–30. https://doi.org/10.1145/3232676

Fortuna, Paula, Juan Soler-Company, and Leo Wanner. “Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets.” In *Proceedings of the 12th Conference on Language Resources and Evaluation*, 6786–94. Marseille: European Language Resources Association, 2020. https://aclanthology.org/2020.lrec-1.835

Garcia, David, Pavlin Mavrodiev, Daniele Casati, and Frank Schweitzer. “Understanding Popularity, Reputation, and Social Influence in the Twitter Society.” *Policy & Internet* 9, no. 3 (2017): 343–64. https://doi.org/10.1002/poi3.151

Gorenc, Nina. “Hate Speech or Free Speech: An Ethical Dilemma?” *International Review of Sociology* 32, no. 3 (2022): 413–25. https://doi.org/10.1080/03906701.2022.2133406

Gulyás, Gábor György. “Gépi tanulási módszerek alkalmazása deanonimizálásra.” *Information Society/Információs Társadalom* 17, no. 1 (2017): 72–86. https://doi.org/10.22503/inftars.XVII.2017.1.5

Guo, Lei, and Brett G. Johnson. “Third-Person Effect and Hate Speech Censorship on Facebook.” *Social Media + Society* 6, no. 2 (2020). https://doi.org/10.1177/2056305120923003

Jaidka, Kokil, Alvin Zhou, Yphtach Lelkes, Jana Egelhofer, and Sophie Lecheler. “Beyond Anonymity: Network Affordances, Under Deindividuation, Improve Social Media Discussion Quality.” *Journal of Computer-Mediated Communication* 27, no. 1 (2022): zmab019. https://doi.org/10.1093/jcmc/zmab019

Jardine, Eric. “Tor, What Is It Good For? Political Repression and the Use of Online Anonymity-Granting Technologies.” *New Media & Society* 20, no. 2 (2018): 435–52. https://doi.org/10.1177/1461444816639976

Kilvington, Daniel. “The Virtual Stages of Hate: Using Goffman’s Work to Conceptualise the Motivations for Online Hate.” *Media, Culture & Society* 43, no. 2 (2021): 256–72. https://doi.org/10.1177/0163443720972318

Kocoń, Jan, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. “Offensive, Aggressive, and Hate Speech Analysis: From Data-Centric to Human-Centered Approach.” *Information Processing & Management* 58, no. 5 (2021): 102643. https://doi.org/10.1016/j.ipm.2021.102643

Konovalova, Ekaterina, Guillaume Le Mens, and Nicolas Schöll. "Social Media Feedback and Extreme Opinion Expression." *PLOS One* 18, no. 11 (2023): e0293805. https://doi.org/10.1371/journal.pone.0293805

Lingam, Revathy-Amadera, and Norizah Aripin. "Comments on Fire! Classifying Flaming Comments on YouTube Videos in Malaysia." *Jurnal Komunikasi: Malaysian Journal of Communication* 33, no. 4 (2017): 104–118. https://doi.org/10.17576/JKMJC-2017-330407

Ma, Xiao, Jeff Hancock, and Mor Naaman. "Anonymity, Intimacy and Self-Disclosure in Social Media." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3857–69. New York: Association for Computing Machinery, 2016. https://doi.org/10.1145/2858036.2858414

Maarouf, Abdurahman, Nicolas Pröllochs, and Stefan Feuerriegel. "The Virality of Hate Speech on Social Media." *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW1 (2024): 1–22. https://doi.org/10.1145/3641025

Masud, Sarah, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. "Hate Is the New Infodemic: A Topic-Aware Modeling of Hate Speech Diffusion on Twitter." In *Proceedings of the 2021 IEEE 37th International Conference on Data Engineering*, 504–15. New York: IEEE, 2021. https://doi.org/10.1109/ICDE51399.2021.00050

Matamoros-Fernández, Ariadna, and Johan Farkas. "Racism, Hate Speech, and Social Media: A Systematic Review and Critique." *Television & New Media* 22, no. 2 (2021): 205–24. https://doi.org/10.1177/1527476420982230

Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. "Spread of Hate Speech in Online Social Media." In *Proceedings of the 10th ACM Conference on Web Science*, 173–82. Boston: Association for Computing Machinery, 2019. https://doi.org/10.1145/3292522.3326034

Men, Linjuan Rita, Wan-Hsiu Sunny Tsai, Zifei Fay Chen, and Yi Grace Ji. "Social Presence and Digital Dialogic Communication: Engagement Lessons from Top Social CEOs." *Journal of Public Relations Research* 30, no. 3 (2018): 83–99. https://doi.org/10.1080/1062726X.2018.1498341

Milmo, Dan. "Anonymous: The Hacker Collectives That Declare War Against Russia." *The Guardian*, February 27, 2022. https://www.theguardian.com/world/2022/feb/27/anonymous-the-hacker-collective-that-has-declared-cyberwar-on-russia

Nascimento, Francimaria RS, George DC Cavalcanti, and Márjory Da Costa-Abreu. "Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis." *SAGE Open* 13, no. 2 (2023). https://doi.org/10.1177/21582440231181311

Oz, Mustafa, Pei Zheng, and Gina Masullo Chen. "Twitter Versus Facebook: Comparing Incivility, Impoliteness, and Deliberative Attributes." *New Media & Society* 20, no. 9 (2018): 3400–19. https://doi.org/10.1177/1461444817749516

Pamirzad, Qurban Hussain. "A Qualitative Exploration of Hate Speech Manifestation in Afghanistan's Twitter-Sphere." *Jurnal Komunikasi: Malaysian Journal of Communication 41*, no. 3 (2025): 1–20. https://doi.org/10.17576/JKMJC-2025-4103-01

Parvaresh, Vahid. "Covertly Communicated Hate Speech: A Corpus-Assisted Pragmatic Study." *Journal of Pragmatics* 205 (2023): 63–77. https://doi.org/10.1016/j.pragma.2022.12.009

Peddinti, Saikat T., Keith W. Ross, and Justin Cappos. "User Anonymity on Twitter." *IEEE Security & Privacy* 15, no. 3 (2017): 84–87. https://doi.org/10.1109/MSP.2017.74

Perera, Suresha, Nadeera Meedin, Maneesha Caldera, Indika Perera, and Supunmali Ahangama. "A Comparative Study of the Characteristics of Hate Speech Propagators and Their Behaviours over Twitter Social Media Platform." *Heliyon* 9, no. 8 (2023): e19097. https://doi.org/10.1016/j.heliyon.2023.e19097

Postmes, Tom, and Russell Spears. "Deindividuation and Antinormative Behavior: A Meta-Analysis." *Psychological Bulletin* 123, no. 3 (1998): 238–59. https://doi.org/10.1037/0033-2909.123.3.238

Riquelme, Fabián, and Pablo González-Cantergiani. "Measuring User Influence on Twitter: A Survey." *Information Processing & Management* 52, no. 5 (2016): 949–75. https://doi.org/10.1016/j.ipm.2016.04.003.

Rutledge, Pamela. "The Fragility of Social Media Fame." *Psychology Today,* March 30, 2021. https://www.psychologytoday.com/us/blog/positively-media/202103/the-fragility-social-media-fame

Ruzaite, Jurate. "In Search of Hate Speech in Lithuanian Public Discourse: A Corpus-Assisted Analysis of Online Comments." *Lodz Papers in Pragmatics* 14, no. 1 (2018): 93–116. https://doi.org/10.1515/lpp-2018-0005

Sachdeva, Pratik, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. "The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism." In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, 83–94. Marseille: European Language Resources Association, June 2022. https://aclanthology.org/2022.nlperspectives-1.11

Saresma, Tuija, Sanna Karkulehto, and Petri Varis. "Gendered Violence Online: Hate Speech as an Intersection of Misogyny and Racism." In *Violence, Gender and Affect: Interpersonal, Institutional and Ideological Practices*, edited by Marita Husso, Sanna Karkulehto, Tuija Saresma, Aarno Laitila, Jari Eilola, and Heli Siltala, 221–243. Cham: Springer Nature, 2020. https://doi.org/10.1007/978-3-030-56930-3_11

Schäfer, Svenja, Michael Sülflow, and Liane Reiners. "Hate Speech as an Indicator for the State of the Society: Effects of Hateful User Comments on Perceived Social Dynamics." *Journal of Media Psychology* 34, no. 1 (2022): 3–15. https://doi.org/10.1027/18641105/a000294

Schmid, Ursula Kristin, Anna Sophie Kümpel, and Diana Rieger. "How Social Media Users Perceive Different Forms of Online Hate Speech: A Qualitative Multi-Method Study." *New Media & Society* 26, no. 5 (2024): 2614–32. https://doi.org/10.1177/14614448221091185

Ștefăniță, Oana, and Diana-Maria Buf. "Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research." *Romanian Journal of Communication and Public Relations* 23, no. 1 (2021): 47–55.

https://doi.org/10.21018/rjcpr.2021.1.322

Tabassum, Shazia, Fabiola SF Pereira, Sofia Fernandes, and João Gama. "Social Network Analysis: An Overview." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, no. 5 (2018): e1256. https://doi.org/10.1002/widm.1256

Tontodimamma, Alice, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. "Thirty Years of Research into Hate Speech: Topics of Interest and Their Evolution." *Scientometrics* 126 (2021): 157–79. https://doi.org/10.1007/s11192-020-03737-6

Trajkova, Zorica, and Silvana Neshkovska. "Online Hate Propaganda During Election Period: The Case of Macedonia." *Lodz Papers in Pragmatics* 14, no. 2 (2018): 309–34. https://doi.org/10.1515/lpp-2018-0015

Vári, László. "Szabadság határokkal, avagy európai útmutató a szólásszabadság jogszerű gyakorlásához." *Információs Társadalom: Társadalomtudományi Folyóirat* 18, no. 3–4 (2018): 25–45. https://doi.org/10.22503/inftars.XVIII.2018.3-4.2

Vedadi, Ali, and Timothy H. Greer. "The Relationship Between Intention to Use, Popularity Information about a Technology, and Trust in Predecessors and Vendors." *Information Resources Management Journal* 34, no. 1 (2021): 43–65. http://doi.org/10.4018/IRMJ.2021010103

Vilanova, Felipe, Francielle Machado Beria, Ângelo Brandelli Costa, and Silvia Helena Koller. "Deindividuation: From Le Bon to the social identity model of deindividuation effects." *Cogent Psychology* 4, no. 1 (2017): 1308104. https://doi.org/10.1080/23311908.2017.1308104

von Essen, Emma, and Joakim Jansson. "Haters Gonna Hate? Anonymity, Misogyny and Hate Against Foreigners in Online Discussions on Political Topics." In J. Jansson, *We Are (Not) Anonymous: Essays on Anonymity, Discrimination and Online Hate* (PhD diss., Stockholm University, 2018).

Williams, Matthew L., Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *The British Journal of Criminology* 60, no. 1 (2020): 93–117. https://doi.org/10.1093/bjc/azz049

Yuan, Shupei, and Chen Lou. "How Social Media Influencers Foster Relationships with Followers: The Roles of Source Credibility and Fairness in Parasocial Relationship and Product Interest." *Journal of Interactive Advertising* 20, no. 2 (2020): 133–47. https://doi.org/10.1080/15252019.2020.1769514

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. "Predicting the Type and Target of Offensive Posts in Social Media." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1415–20. Minneapolis: Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1144

Zannettou, Savvas, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. "Measuring and Characterizing Hate Speech on News Websites." In *Proceedings of the*

12th ACM Conference on Web Science, 125–134. Southampton: Association for Computing Machinery, 2020.
https://doi.org/10.1145/3394231.3397902

Zapata, Jimena, Justin Sulik, Clemens von Wulffen, and Ophelia Deroy. "Bystanders' Collective Responses Set the Norm Against Hate Speech." *Humanities and Social Sciences Communications* 11, no. 1 (2024): 335.
https://doi.org/10.1057/s41599-024-02761-8

Zhang, Min, Dongxin Zhang, Yin Zhang, Kristin Yeager, and Taylor N. Fields. "An Exploratory Study of Twitter Metrics for Measuring User Influence." *Journal of Informetrics* 17, no. 4 (2023): 101454.
https://doi.org/10.1016/j.joi.2023.101454