social sciences

MDPI

Article

Hate Speech on Social Media: Unpacking How Toxic Language Fuels Anti-Immigrant Hostility

Juan-José Igartua [1,*] and Carlos A. Ballesteros-Herencia [2]

[1] Department of Sociology and Communication, University of Salamanca, Campus Miguel de Unamuno, Paseo Francisco Tomás y Valiente, s/n, 37007 Salamanca, Spain
[2] Faculty of Arts and Humanities, University of Valladolid, Plaza del Campus, s/n, 47011 Valladolid, Spain; carlosantonio.ballesteros@uva.es
* Correspondence: jigartua@usal.es

Abstract

This study investigates the influence of toxic language in hate speech targeting immigrants, particularly through narrative formats like first-person X (Twitter) threads. Hate speech, defined as promotion of hatred based on personal or group characteristics, increasingly escalates on social media, impacting public attitudes and behaviors. While previous research has primarily focused on measuring the scope of hate speech through content analysis and computational methods, there has been limited attention to its effects on audiences. This study presents the results of an online experiment (N = 339) with a 2 × 2 between-subjects design that manipulates the presence of toxic language and message popularity. Results indicate that hate messages lacking toxic language promote greater identity fusion with the author of the message, which in turn increases the intention to share the message, reinforces negative attitudes toward immigrants, and increases support for harsh policies against irregular immigration. Moreover, non-toxic hate messages significantly enhance narrative transportation exclusively for individuals with conservative political views, thereby further increasing their intention to share the message. These findings highlight that subtler forms of hate speech can create strong audience connections with hostile perspectives, emphasizing the need for anti-hate campaigns to address both overt and subtle hate content.

Keywords: online hate speech; toxic language; narrative transportation; identity fusion with the author of the message; indirect effects

Academic Editor: Andreu Casero-Ripollés

Received: 14 November 2025
Revised: 22 January 2026
Accepted: 30 January 2026
Published: 3 February 2026

1. Introduction

The rise of hostile discourse toward immigrants on social media has become an urgent social issue in Europe and beyond (Klein 2024). According to Eurobarometer (European Commission 2024), Russia's war of aggression against Ukraine is currently viewed as the most important issue facing the European Union (31%), closely followed by immigration (28%). Furthermore, nearly one-third of Europeans (31%) regard immigration from outside the EU more as a problem than as an opportunity. Media play a pivotal role in shaping public perceptions of immigration, with 56% of citizens reporting that they obtain information about immigration and integration through traditional media, while 15% rely on social media (European Commission 2022). Meanwhile, digital platforms have become fertile ground for hate speech (Paasch-Colberg et al. 2021), which amplifies negative stereotypes about immigrants and legitimizes exclusionary views (Ahmed et al. 2024). As

hate speech becomes more pervasive online, it is crucial to understand the mechanisms through which it operates and the effects it has on individuals' attitudes and behaviors. This study aims to contribute to that understanding by examining how exposure to toxic discourse shapes public responses to immigration.

1.1. Hate Speech on Social Media: Prevalence, Effects, and Gaps in Research

Hate speech is broadly defined as communication that incites, promotes, or instigates hatred, humiliation, or contempt toward individuals or groups, typically based on characteristics such as race, ethnicity, age, disability, gender, or sexual orientation (Arcila-Calderón et al. 2024; Castaño-Pulgarín et al. 2021; Fino 2020; Hietanen and Eddebo 2023; Villegas-Lirola and Rodriguez-Martinez 2025). Social media have increasingly become platforms for the propagation of hate speech against minority groups, such as immigrants (Saridou et al. 2023; Müller and Schwarz 2021). Within online platforms, this speech frequently targets immigrants by portraying them as economic burdens, cultural threats, or security risks (Essalhi-Rakrak and Pinedo-González 2023; Lilleker and Pérez-Escolar 2023). Hate speech is not only a manifestation of prejudice but also serves as a mechanism for political persuasion, mobilization, fostering polarization and entrenching discriminatory ideologies (Abuín-Vences et al. 2022; Carlson 2020; Ikeanyibe et al. 2018).

To date, much of the empirical research on hate speech has focused on its detection and prevalence through content analysis and computational methods (e.g., Arcila-Calderón et al. 2021; Ayo et al. 2020; Eschmann et al. 2025; Lingiardi et al. 2019; Matamoros-Fernández 2017). However, fewer studies have examined the psychological effects of exposure to hate speech on audiences (e.g., Chen and Dang 2023; Hsueh et al. 2015; Obermaier et al. 2023; Pluta et al. 2023; Rösner et al. 2016; Schäfer et al. 2024; Soral et al. 2018; Wachs et al. 2021; Weber et al. 2020). This line of research, though nascent, reveals the detrimental impact of hate speech on attitudes, emotional responses, and behavioral intentions. For example, exposure to prejudiced online comments can elicit similar prejudiced responses in readers (Hsueh et al. 2015), while hateful content may attenuate empathic neural responses (Pluta et al. 2023) and reduce prosocial behavior (Weber et al. 2020). Furthermore, studies have shown that frequent exposure to hate speech can decrease an individual's sensitivity to its harmfulness, subsequently increasing prejudice towards targeted groups (Soral et al. 2018). Experimental evidence also indicates that exposure to intolerant political speech on social media triggers stronger negative emotions and increases political distrust, particularly among young people (Saumer et al. 2024).

Yet the influence of specific message features within hate speech remains underexplored. In particular, the role of *linguistic toxicity*—marked by dehumanizing, aggressive, or obscene language—is poorly understood in terms of its capacity to enhance or suppress persuasive effects. Previous work suggests that less overtly hostile forms of hate speech may be especially insidious, as they foster identification and engagement without triggering critical resistance (Weber et al. 2020; Obermaier et al. 2023). Another understudied feature is the effects of the presence of *popularity cues*—such as the number of likes, shares, or comments—that often accompany social media posts (Dvir-Gvirsman 2019; Sung and Lee 2015; Woods 2023). These cues act as social endorsements, signaling the perceived approval or agreement with the message by a broader audience (Sung and Lee 2015). The effect of these cues can significantly shape how individuals process and respond to hate speech (Sundar 2008). This study addresses this gap by analyzing the effects of toxic language and the perceived popularity of hate messages, while examining the mediating role of two key mechanisms—identity fusion with the author of the message and narrative transportation. The dependent variables include attitudes toward immigration, support for

harsh policies against irregular immigration, and the intention to share the message, which contributes to the virality and normalization of hate content.

1.2. Theoretical Framework: The THREAD Model

Many online hate messages targeting immigrants emerge in response to news events and often take the form of narrative threads. These narratives, especially first-person testimonials, are prevalent on platforms like Twitter (now X) and enable authors to articulate their perspectives on immigration. To conceptualize how individuals respond to hate speech following sociopolitical or media-disruptive events, we propose the Toxic Hate Responses Emerging After Disruptive Events (THREAD) Model (see Figure 1). This model integrates theories of social identity (Reicher et al. 2005; Tajfel and Turner 1979), narrative persuasion (Green and Brock 2000), and toxic communication (Kim et al. 2021) to explain how specific message elements influence audience processing and alignment with hate-based narratives.

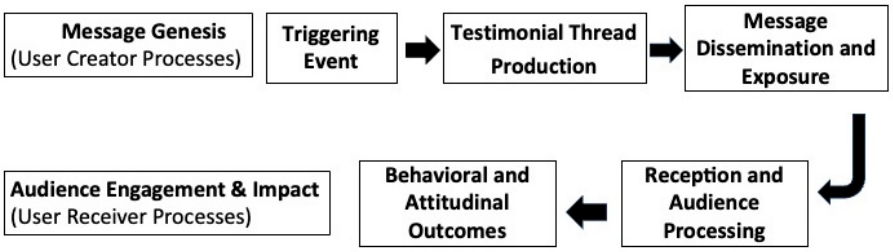

Figure 1. THREAD Model: Toxic Hate Responses Emerging After Disruptive events.

At its core, the THREAD Model posits that hate messages should be understood not merely as expressions of prejudice or aggression, but as narrative artifacts that harness the structural and psychological affordances of storytelling to achieve persuasive aims. These messages often adopt a coherent plot, feature emotionally charged protagonists (typically the message authors themselves), and follow a familiar personal testimonial format that mimics everyday narrative communication. This narrative framing increases engagement by providing a clear storyline, emotional resonance, and a sense of authenticity, which can promote a psychological merging between the audience and the author—a state of identity fusion (Swann et al. 2009). In parallel, these same narrative properties facilitate narrative transportation: the experiential absorption into the story (Green and Brock 2000). Together, identity fusion and narrative transportation operate as dual psychological mechanisms that can make narratively framed hate messages particularly impactful. This perspective aligns with Walther's (2024) assertion that the posting of hate content is frequently motivated by the desire for social validation and approval. In online environments, where reactions such as likes, retweets, and comments serve as indicators of collective endorsement, narratively constructed hate messages may be especially effective at eliciting these forms of social approval. The narrative structure, by fostering both immersion and perceived intimacy, increases the likelihood that others will engage with, amplify, and endorse the content, thereby reinforcing its perceived legitimacy and contributing to the broader diffusion and normalization of hate-based narratives.

In this context, disruptive events—often highlighted in the news, such as crimes attributed to immigrants or sudden surges in irregular immigration—can serve as powerful catalysts that trigger the circulation of hate narratives on social media (Arcila-Calderón et al. 2024). On platforms like Twitter (X), users frequently respond to such events through first-person testimonial threads that emulate personal storytelling. These messages typically embed emotionally charged interpretations and express hostility toward the implicated group, often under the guise of authenticity and concern. A key dimension in which these hate

narratives differ is their level of linguistic toxicity. Some employ overtly dehumanizing or obscene expressions, while others adopt a more restrained tone that masks prejudice behind seemingly civil discourse. Toxic language, as defined by Kim et al. (2021), encompasses the use of derogatory, dehumanizing, and hostile expressions specifically aimed at stigmatizing, discrediting, or inciting hatred against individuals or groups based on their identity. By embedding toxic language within these narratives, authors amplify negative stereotypes and promote discrimination, degrading the targeted group and reinforcing exclusionary and xenophobic attitudes among audiences (Matamoros-Fernández 2017; Müller and Schwarz 2021). As previously noted, the THREAD Model posits that hate messages, structured as narratives, activate psychological processes that mediate their persuasive impact. Specifically, the model highlights two key psychological mechanisms (identity fusion with the author of the message and narrative transportation) that shape how exposure to hate speech influences users' attitudes, beliefs, emotions, and behavioral intentions.

The concept of *identity fusion* describes a visceral sense of oneness between an individual and another person or group, in which the boundaries between personal and social identities become blurred (Swann et al. 2009, 2012). This psychological state entails a profound feeling of alignment that can motivate intense commitment and pro-group behavior (Swann et al. 2012). Building on this framework, the present study adapts the notion of identity fusion to the context of online hate speech, conceptualizing *identity fusion with the author of the message* as the degree to which readers experience a symbolic merging of their self with that of the message's author. In this fused state, readers may come to feel the author's emotions and grievances as if they were their own. They are therefore more likely to adopt the author's perspective, echo their outrage, or amplify their message—as if wearing VR goggles that immerse them in the author's personal experience. This sense of unity facilitates the internalization of the message's viewpoint and can reduce critical scrutiny.

The construct of identity fusion with the author of the message is theoretically distinct from related concepts. Unlike identification—a more symbolic, role-based alignment that typically operates at the cognitive or affective level (Cohen 2001)—fusion involves a deep relational bond in which the self and the other are perceived as functionally equivalent (Swann et al. 2012). It also differs from perceived homophily, which refers to perceived similarity or shared characteristics (McCroskey et al. 1975) but not to the merging of personal and external identities. Prior research shows that such fusion processes can occur not only with groups but also at the interpersonal level (Joo and Park 2017; Chinchilla et al. 2022; Vázquez et al. 2017). Moreover, self-transcendent emotions such as awe have been found to promote fusion by expanding one's sense of connectedness with others (Song et al. 2025). Consistent with this literature, the present study measured identity fusion with the author of the message using the Inclusion of Other in the Self (IOS) scale (Aron et al. 1992; Gächter et al. 2015), a parsimonious and validated pictorial indicator of perceived self–other overlap suitable for dyadic targets. Finally, recent evidence suggests that fusion can, under certain conditions, foster intergroup trust and cooperation (Klein et al. 2025).

Narrative transportation, by contrast, refers to a psychological process of absorption in a story, in which individuals become cognitively and emotionally immersed in the unfolding narrative world (Appel et al. 2015; Green and Brock 2000). Much like immersing oneself in a gripping novel, this absorption makes the reader more receptive to the message's persuasive power, increasing the likelihood that they will internalize the thread's themes and respond in alignment with its emotional and ideological cues (Van Laer et al. 2014). Whereas identity fusion with the author captures a relational alignment with the communicator, narrative transportation describes engagement with the narrative itself. We propose that both mechanisms operate in parallel, contributing to persuasion through

distinct yet complementary pathways: transportation facilitates immersion and openness to the narrative flow, while fusion strengthens alignment with the message source, creating a sense of shared perspective. In the context of online hate speech, these parallel mediating processes are expected to jointly shape attitudes toward immigration, support for harsh policies against irregular immigration, and the intention to share the message—thereby contributing to the dissemination and normalization of hate content.

The THREAD Model also incorporates individual-level and message-level moderators that can influence the psychological processes through which hate narratives exert their effects. At the individual level, political ideology plays a critical role. Political conservatism has been consistently linked to stronger anti-immigrant attitudes (e.g., Davidov et al. 2020), suggesting that individuals with more conservative views may be more susceptible to negative messaging about immigrants. However, the extent to which reception processes are moderated by the recipient's political ideology remains unclear. Toxic language may be perceived as excessively extreme, potentially diminishing both identity fusion with the message source and narrative transportation. Conservatives, who generally value order, tradition, and social cohesion (Jost et al. 2004), might be particularly receptive to anti-immigrant narratives when framed in a non-toxic tone (one that reinforces their existing beliefs without provoking moral discomfort). In contrast, when such narratives are expressed using overtly toxic language, they may elicit psychological reactance (Brehm 1966), thereby weakening both identity fusion and transportation. This raises a research question regarding the moderating effect of political ideology on the impact of toxic language in hate messages on identity fusion with the author and narrative transportation:

RQ1: To what extent does the recipient's political ideology moderate the effect of toxic language in hate messages on identity fusion with the author and narrative transportation?

At the message level, the THREAD Model underscores the relevance of metrics associated with social media messages, as they can exert persuasive effects on user (Dong and Li 2025). Metrics such as likes, shares, and comments function as *popularity cues* (Dvir-Gvirsman 2019; Woods 2023), signaling the extent to which a message has resonated with others. Popularity cues are "metric information about users' behavior" (Haim et al. 2018, p. 188) that function as social signals for individuals navigating social networks. In our model, these cues are considered peripheral aspects of the message (Petty and Cacioppo 1986), as they do not pertain to its substantive content but rather to its perceived social impact. By acting as indicators of collective endorsement, popularity cues can influence how individuals engage with and respond to hate messages, reinforcing their perceived legitimacy and persuasive power (Sundar 2008). Narratives presented in a personal, testimonial style and accompanied by strong popularity cues may be particularly impactful, especially when their tone and content resonate with the audience's ideological predispositions. This study examines whether the presence of popularity cues moderates the effect of varying levels of toxic language in the message on identity fusion with the author and narrative transportation. Furthermore, it explores whether these effects are themselves moderated by political ideology.

RQ2: Does the presence of popularity cues (e.g., number of likes, shares, and comments) moderate the effect of the level of toxic language in the message on identity fusion with the author and narrative transportation?

RQ3: Does the recipient's political ideology further moderate the interaction between level of toxic language and popularity cues in shaping identity fusion with the author and narrative transportation?

Finally, as previously noted, the THREAD Model anticipates both attitudinal and behavioral outcomes, such as the intention to share the message, attitudes toward immigrants, and support for harsh immigration policies. Building on prior

research on the persuasive power of testimonial narratives across various contexts (e.g., Cohen et al. 2023; Dahlstrom and Rosenthal 2018; Igartua and Cachón-Ramón 2023; Igartua and Guerrero-Martín 2022; Rosenthal and Dahlstrom 2019), we examine a full moderated mediation model. Specifically, we test whether narrative transportation and identity fusion with the author of the message function as parallel mediators in the relationship between toxic language and the aforementioned outcome variables. Furthermore, we assess whether political ideology moderates the impact of the level of toxic language in the message on both mediators, thereby influencing their indirect effects on sharing intentions, attitudes, and policy support.

RQ4: Does political ideology moderate the indirect effects of toxic language on the intention to share the message, attitudes toward immigrants, and support for harsh policies against irregular immigration, through parallel mediation by narrative transportation and identity fusion with the author?

2. Materials and Methods

The present study empirically tests predictions derived from the THREAD Model through an experimental design that manipulates both the level of toxic language and the popularity of hate messages targeting immigrants. By employing first-person narrative formats, the study replicates the naturalistic conditions under which social media users typically encounter such content. It focuses on the indirect and conditional effects of such messages on willingness to disseminate the message, attitudes toward immigrants, and support for harsh immigration policies (considering the moderating role of participants' political ideology). In doing so, the study contributes to a deeper understanding of how both subtle and overt forms of online hate influence public opinion on immigration issues.

2.1. Participants

The current study was conducted through an online experiment with a sample of 339 participants. The study was conducted in Spain using Spanish-speaking participants recruited through convenience sampling among volunteers who responded to online invitations disseminated via social media and institutional mailing lists. All participants provided informed consent prior to participation and were debriefed at the end of the experiment. The procedure was fully compliant with institutional ethical standards and Spanish regulations for non-invasive, survey-based research with adults. The initial sample consisted of 429 individuals. Quality control procedures were applied based on message reading time (ranging from 267 to 343 words). The estimated average reading time was calculated using the tool https://legible.es (accessed on 24 February 2023). Only those who spent between 80 and 300 s reading the message were included in the final sample, ensuring adequate engagement without rushing or excessive time, which could indicate disengagement. Additionally, participants who completed the questionnaire in less than 360 s or more than 2400 s were excluded to identify responses that might be rushed or distracted, potentially affecting data reliability (Huang et al. 2012; Meade and Craig 2012).

The final sample consisted of 50.7% women, 47.5% men, 0.6% non-binary individuals, and 1.2% who preferred not to disclose their gender. Participants' ages ranged from 18 to 64 years (M = 26.98, SD = 11.57). Regarding employment status, 63.7% were students and 33.3% were employed. In terms of educational attainment, 36.8% had completed a university degree. Political ideology was assessed on a scale from 0 (left) to 10 (right), with a mean score of 4.15 (SD = 2.50).

Given the absence of directly comparable prior studies to estimate a realistic effect size, we conducted a sensitivity analysis using G*Power Version 3.1.9.4 (Faul et al. 2007) to determine the smallest effect size detectable with the available sample. The analysis was

based on a 2 × 2 between-subjects factorial ANOVA design (see below), with an alpha level of 0.05, a statistical power of 0.80, numerator degrees of freedom set to 1, and a total sample size of 339 participants. The results indicated that the design was sufficiently powered to detect an effect size between small and medium, specifically, an effect size of $f = 0.15$, which corresponds to Cohen's d = 0.30 and partial $\eta^2 = 0.022$. This sensitivity level applies to both main effects and the interaction term in the model.

2.2. Design and Procedure

We used a 2 × 2 between-subjects design, manipulating two key factors: the level of toxic language (low versus high) in the hate speech messages and the number of interactions (low versus high) associated with those posts. First-person testimonial messages were presented in the form of Twitter (X) threads, simulating user-generated posts reacting to a news story about the arrival of immigrants in Spain and highlighting the perceived threat this posed to Spanish society. The triggering event used for all conditions was a real news article published on 24 June 2022, on the online newspaper https://www.vozpopuli.com/, reporting on the arrival of immigrants in Spain. The fieldwork was conducted between 28 February and 13 April 2023. Given the time elapsed between the publication of the article and the data collection period, it is unlikely that participants recalled the news story in detail. Moreover, in Spain, similar news stories about immigration appear regularly throughout the year, which helps maintain the ecological validity of the stimulus without risking recognition bias.

Each thread began with a comment on this news story and continued with a sequence of follow-up tweets that elaborated on the user's perspective. To enhance the external validity of the study, the messages referenced two types of threats: unfair competition in the job market or economic threat versus violence at the borders of Melilla or threats to security.

The online questionnaire comprised three sections: pre-test measures, experimental manipulation, and post-test measures. The first section gathered data on sociodemographic variables, political ideology, and participants' perceptions of the most pressing issues in Spain (included as a filler measure). Next, participants were exposed to the experimental manipulation by reading a social media post presented as a testimonial message. After reading the narrative, the post-test measures were administered, which included questions to evaluate the effectiveness of the experimental manipulation and the outcome variables. All materials related to the research, including the pilot and main study instruments, datasets, syntax files, testimonial messages, and Electronic Supplementary Material (ESM), are available through the Open Science Framework (OSF): https://osf.io/2a4xy/ (accessed on 22 January 2026). For illustrative purposes, anonymized examples of the social media messages analyzed in this study can be consulted in this online repository.

2.3. Pilot Study

The pilot study ($N = 41$) involved participants with the following characteristics: 58.5% female ($n = 24$) and 41.5% male ($n = 17$), with ages ranging from 20 to 50 years ($M = 22.02$, $SD = 5.21$). Most participants (92.7%) were students, while 7.3% were employed. Political ideology was assessed on a scale from 0 (left) to 10 (right), with a mean score of 4.00 ($SD = 2.59$). The majority of participants (90.2%) were born in Spain. The purpose of this pilot study was to assess participants' perceptions of certain expressions commonly used in tweets on Twitter (X) that refer to individuals and social groups. Specifically, the study aimed to evaluate toxic language expressions frequently used in discussions about sensitive social issues and demographic groups.

Participants were asked to evaluate a series of 24 expressions, each presented individually on the screen. These expressions were evaluated on four dimensions: aggressiveness, insulting tone, offensiveness, and humiliating content. Participants rated each expression on a scale from 0 = not at all, to 100 = extremely. A composite measure of expression toxicity was then created based on these four dimensions, with Cronbach's alpha coefficients exceeding 0.93 in all cases. Some of the expressions evaluated included phrases such as "Sucios moros" [Dirty Moors] (M = 88.77, SD = 20.02), "Montón de scoria" [A pile of scum] (M = 70.83, SD = 32.59), and "Marabunta de inmigrantes" [Swarm of immigrants] (M = 61.80, SD = 30.69). Expressions with low scores on the measure of expression toxicity were "Moros" [Moors] (M = 29.96, SD = 31.40), "Multitud de inmigrantes" [Crowd of immigrants] (M = 23.04, SD = 27.09), and "Inmigrantes" [Immigrants] (M = 21.28, SD = 27.38).

Although some expressions used in the pilot study may seem highly offensive and might typically be removed by social media moderators, prior research has shown that such language appears frequently in real-world hate speech against immigrants (Arcila-Calderón et al. 2020). Studies have documented the use of dehumanizing metaphors, ethnic slurs, and aggressive terms like "swarm" or "invaders" in online content, particularly on platforms like Twitter (Essalhi-Rakrak and Pinedo-González 2023). Including these expressions help ensure ecological validity, as they reflect the actual language used in hateful discourse targeting immigrants.

In addition to evaluating toxic language, the pilot study also assessed the interaction levels typically generated by tweets. Participants were asked to indicate the number of interactions (comments, retweets, and "likes") that, in their opinion, would make a tweet important or relevant. The median values for these interactions were as follows: 100 comments, 300 retweets, and 856 "likes." Furthermore, participants were asked about their own behavior when viewing tweets. When asked whether they typically pay attention to the number of comments a tweet has received, 9.8% said they never do, 29.3% said they do so occasionally, 26.8% sometimes, 26.8% often, and 7.3% very often. Regarding retweets, 12.2% said they never pay attention to them, 29.3% do so occasionally, 17.1% sometimes, 26.8% often, and 14.6% very often. Finally, when asked about likes, 7.3% reported never paying attention to them, 22.0% occasionally, 22.0% sometimes, 31.7% often, and 17.1% very often.

The insights gained from this pilot study were then used to craft the experimental stimuli for the main study. This process involved selecting toxic language expressions for the hate speech messages (low vs. high) and determining the levels of user interactions (low vs. high) based on the number of likes, shares, and comments associated with those posts.

2.4. Measures

Fusion with the author of the message. This variable was measured using the Inclusion of the Other in the Self (IOS) Scale (Gächter et al. 2015). Participants were asked the following question: To what extent did you feel "connected" or "fused" with the author of the message you just read—that is, the person who wrote the Twitter thread? Please select the image that best represents that feeling. They responded by selecting one of seven images representing increasing levels of overlap between two circles, depicting a 7-point visual analog scale. The responses were coded from 1 = low fusion to 7 = high fusion (M = 2.35, SD = 1.74).

Narrative transportation. This construct was assessed using five of the six items from the Transportation Scale–Short Form developed by Appel et al. (2015), as the messages featured only one character (the author of the message). Sample items include "I felt highly mentally involved while reading the message" and "While reading the message, I formed a very vivid and clear image of the author of the message". Responses were measured on

a 7-point Likert scale ranging from 1 = strongly disagree to 7 = strongly agree ($\alpha = 0.70$, $M = 3.93$, $SD = 1.22$).

Intention to propagate or share the message. This variable was assessed using a six-item scale (e.g., "I would be willing to share this message with others"), adapted from Barbour et al. (2016) and Igartua et al. (2017). Responses were measured on a 7-point Likert scale ranging from 1 = strongly disagree to 7 = strongly agree ($\alpha = 0.83$, $M = 2.79$, $SD = 1.28$).

Attitudes toward immigrants. This was measured using a feeling thermometer (Wojcieszak et al. 2020), which assessed participants' attitudes towards a variety of social groups, including executives and financiers, immigrants, politicians, judges, law enforcement officers (police), refugees, clergy (priests or ministers), military personnel, and journalists. The scale ranged from 0 (very cold feelings) to 100 (very warm feelings). A composite index was created to reflect participants' attitudes toward both immigrants and refugees, allowing for a more comprehensive evaluation of attitudes toward these groups as distinct, yet related categories ($r[337] = 0.80$, $p < 0.001$; $M = 60.68$, $SD = 23.24$). For brevity, this variable is referred to as attitudes toward immigrants throughout the text, tables, and figures.

Support for harsh policies against irregular immigration. This was measured using a composite scale of four items (e.g., "The government should restrict the entry of additional immigrants into Spain"), adapted from Bilewicz and Soral (2020) and Saleem et al. (2017). Participants rated their agreement on a 7-point scale ranging from 1 = strongly disagree to 7 = strongly agree ($\alpha = 0.85$, $M = 3.44$, $SD = 1.55$).

Manipulation check. The effectiveness of the two experimental manipulations was assessed using a scale composed of six items (e.g., "The message used very aggressive language", "While reading the message, I noticed that it had many likes", from 1 = strongly disagree to 7 = strongly agree). In addition, the emotional impact of the message was measured by asking participants to indicate the extent to which they experienced specific emotions (e.g., anger, disgust, sadness, rage, hostility) while reading the message (from 1 = not at all to 7 = very much). The emotion scale was developed based on previous work by Saleem et al. (2016) and Fernández et al. (2013).

3. Results

3.1. Preliminary Analysis

The effectiveness of the two experimental manipulations (the presence of toxic language in the hate speech messages and the number of interactions associated with those posts) was assessed using a scale composed of six items, three of which were related to the manipulation of the presence of toxic language in the message, while the other three were linked to the manipulation of the number of interactions associated with those posts. Due to space limitations, the full results of the mean difference analyses for the manipulation checks are reported in the ESM on OSF (Tables S1–S3).

It was found that participants randomly assigned to the toxic language hate speech condition (compared to those exposed to the non-toxic language version) were more likely to agree with the statements "The person who wrote the message is very likely to have a negative view of immigration" ($t[337] = -4.55$, $p < 0.001$), "The message used very aggressive language" ($t[337] = -11.08$, $p < 0.001$), and "The tone of the message I read can be considered very offensive or insulting" ($t[337] = -9.65$, $p < 0.001$). On the other hand, individuals exposed to a message that indicated a high number of interactions (compared to those exposed to a message with a low number of interactions) showed a greater agreement with the statements "While reading the message, I noticed that it had many likes" ($t[337] = -5.20$, $p < 0.001$), "The message I read had been shared many times" ($t[337] = -6.44$,

$p < 0.001$), and "The message I read had been retweeted many times" ($t[337] = -8.16$, $p < 0.001$). These results indicate that the manipulation of the two independent variables was effective.

Additionally, participants reported experiencing higher levels of negative emotions (such as anger, disgust, sadness, rage, irritation, discomfort, and hostility) when reading hate messages that included toxic language expressions. These findings suggest that the presence of toxic language in the hate speech condition significantly intensified the emotional response of participants compared to those exposed to the non-toxic language version. This reinforces the effectiveness of the experimental manipulation, highlighting the crucial role that toxic language plays in amplifying the emotional impact of hate speech.

3.2. Testing the Research Questions: Moderation and Moderated Mediation Results

The analyses addressing the research questions were conducted using the PROCESS macro for SPSS version 30 (Model 1; Hayes 2022). To test the first research question (RQ1), a moderation analysis was performed in which the experimental condition—representing the level of toxic language in the hate message—was entered as a dummy-coded independent variable (0 = low, 1 = high), and political ideology served as the moderating variable. Political ideology was coded on an 11-point scale (0 = left, 10 = right). The type of threat depicted in the message (0 = economic, 1 = security) and the number of interactions associated with the post (0 = low, 1 = high) were included as covariates to control for message-level variance. Conditional effects of the independent variable at specific levels of political ideology are displayed in Figures 2 and 3.

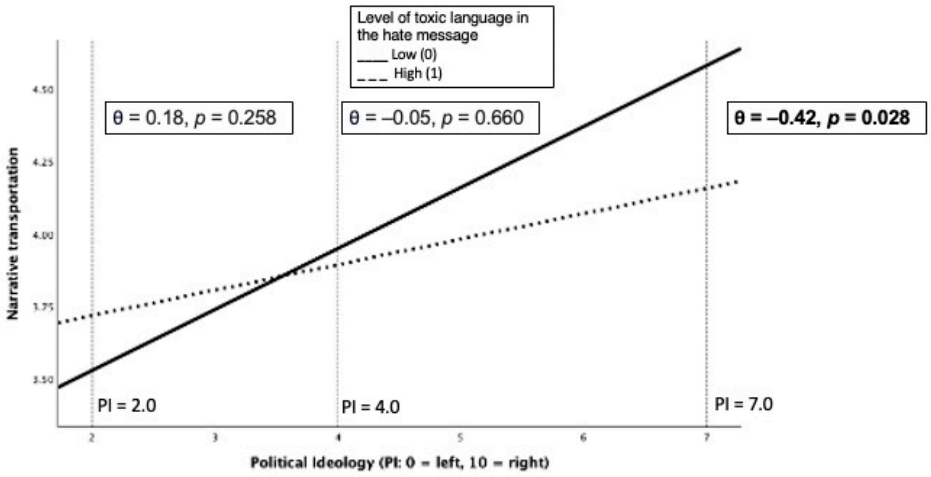

Figure 2. Conditional effects of political ideology on narrative transportation as a function of toxic language. **Note:** The experimental condition (level of toxic language in the hate message) was set up as a dummy variable (0 = low, 1 = high). Narrative transportation (1 = low, 7 = high). Political ideology was assessed on an 11-point scale (0 = left, 10 = right). The type of threat (0 = economic threat, 1 = security threat) and the number of interactions associated with those posts (0 = low, 1 = high) were included as covariates. θ indicates the conditional effect of toxic language on narrative transportation at different levels of political ideology. Conditional effects were estimated for three reference points of the moderator, corresponding to the 18th (2), 50th (4), and 85th (7) percentiles of the political ideology distribution. Conditional effects of the independent variable at these specific values were computed and plotted using PROCESS. The values displayed in the figure therefore represent conditional effects rather than raw regression coefficients. The negative θ value indicates that, at the conservative level of political ideology (value = 7), the low-toxic condition yielded higher narrative transportation than the high-toxic condition.

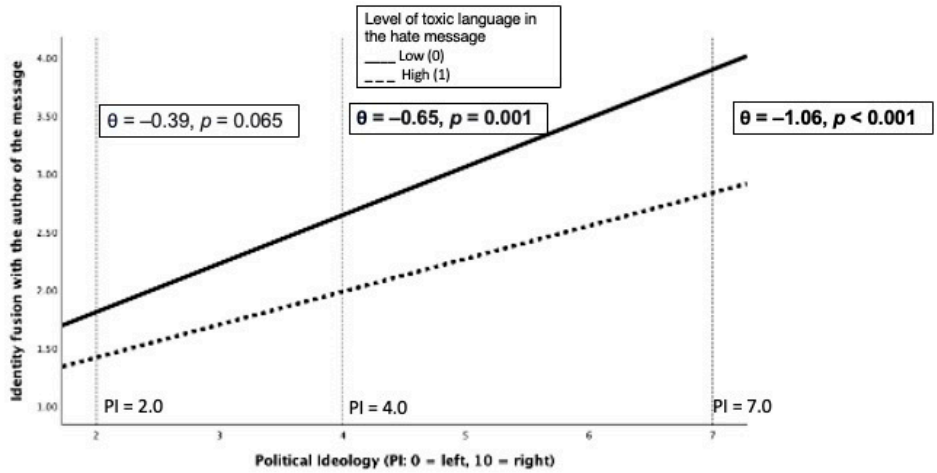

Figure 3. Conditional effects of political ideology on identity fusion with the author of the message as a function of toxic language. **Note:** The experimental condition (level of toxic language in the hate message) was set up as a dummy variable (0 = low, 1 = high). Identity fusion with the author of the message (1 = low, 7 = high). Political ideology was assessed on an 11-point scale (0 = left, 10 = right). The type of threat (0 = economic threat, 1 = security threat) and the number of interactions associated with those posts (0 = low, 1 = high) were included as covariates. θ indicates the conditional effect of toxic language on identity fusion with the author of the message at different levels of political ideology. Conditional effects were estimated for three reference points of the moderator, corresponding to the 18th (2), 50th (4), and 85th (7) percentiles of the political ideology distribution. Conditional effects of the independent variable at these specific values were computed and plotted using PROCESS. The values displayed in the figure therefore represent conditional effects rather than raw regression coefficients. The negative θ value indicates that, at the conservative level of political ideology (value = 7), the low-toxic condition yielded higher identity fusion with the author of the message than the high-toxic condition.

A statistically significant interaction effect was observed between the level of toxic language and political ideology on narrative transportation ($B_{\text{interaction}} = -0.11$, $SE = 0.05$, $p = 0.019$; Table 1). Specifically, it was observed that the absence of toxic language enhanced narrative transportation with the message, but exclusively among individuals with a more conservative political ideology (Figure 2). This suggests that individuals with more conservative ideologies felt a stronger connection to the story when the tone was non-hostile.

Additionally, the interaction between the level of toxic language and political ideology on identity fusion with the author of the message was found to be statistically significant ($B = -0.13$, $SE = 0.06$, $p = 0.038$; Table 2). The conditional effects analysis revealed that the absence of toxic language in the message increased identity fusion, but only among individuals with moderately to strongly conservative political views; this effect was not observed among those with more left-leaning positions (Figure 3).

However, with respect to RQ2, the number of interactions associated with the posts did not moderate the effect the level of toxic language in the message on narrative transportation ($B_{\text{interaction}} = 0.24$, $SE = 0.26$, $p = 0.365$) nor identity fusion with the author ($B_{\text{interaction}} = 0.48$, $SE = 0.37$, $p = 0.192$). Furthermore, no three-way interaction (RQ3) was found among the level of toxic language in the hate speech message, the number of interactions associated with those posts, and political ideology (fusion with the author: narrative transportation: $B_{\text{three-way interaction}} = -0.13$, $SE = 0.10$, $p = 0.173$; $B_{\text{three-way interaction}} = -0.00$, $SE = 0.12$, $p = 0.967$).

Table 1. Political ideology as a moderator in the relationship between the level of toxic language and narrative transportation. Moderation analysis conducted using PROCESS (Model 1).

	Outcome Variable: Narrative Transportation			
Model summary: R^2 = 0.12, p < 0.001	***B***	***SE***	***t***	***p***
Constant	2.92	0.19	14.93	<0.001
Toxic language	0.41	0.24	1.71	0.087
Political ideology	0.20	0.03	5.63	<0.001
Toxic language × Political ideology	−0.11	0.05	−2.34	0.019
Type of threat	0.31	0.12	2.48	0.013
Number of interactions	0.06	0.12	0.52	0.601
Conditional effects of toxic language at different levels of political ideology	***Effect***	***SE***	***t***	***p***
Liberal (progressive)	0.18	0.16	1.09	0.276
Moderate	−0.05	0.12	−0.44	0.657
Conservative	−0.41	0.19	−2.14	0.032

Note: The experimental condition (level of toxic language in the hate message) was set up as a dummy variable (0 = low, 1 = high). Narrative transportation (1 = low, 7 = high). Political ideology was assessed on an 11-point scale (from 0 = left to 10 = right). The type of threat (0 = economic threat, 1 = security threat) and the number of interactions associated with those posts (0 = low, 1 = high) were included as covariates. Effect (θ) = conditional effect of toxic language on narrative transportation at different levels of political ideology. The classification into three groups of political ideology to calculate conditional effects was based on the calculation of the 18th (2), 50th (4), and 85th (7) percentiles.

Table 2. Political ideology as a moderator in the relationship between the level of toxic language and identity fusion with the author of the message. Moderation analysis conducted using PROCESS (Model 1).

	Outcome Variable: Identity Fusion			
Model summary: R^2 = 0.29, p < 0.001	***B***	***SE***	***t***	***p***
Constant	0.84	0.24	3.38	<0.001
Toxic language	−0.12	0.31	−0.41	0.679
Political ideology	0.41	0.04	8.83	<0.001
Toxic language × Political ideology	−0.13	0.06	−2.07	0.038
Type of threat	0.15	0.16	0.97	0.331
Number of interactions	0.09	0.16	0.59	0.551
Conditional effects of toxic language at different levels of political ideology	***Effect***	***SE***	***t***	***p***
Liberal (progressive)	−0.39	0.21	−1.87	0.061
Moderate	−0.66	0.16	−4.12	<0.001
Conservative	−1.06	0.24	−4.35	<0.001

Note: The experimental condition (level of toxic language in the hate message) was set up as a dummy variable (0 = low, 1 = high). Identity fusion with the author of the message (1 = low, 7 = high). Political ideology was assessed on an 11-point scale (from 0 = left to 10 = right). The type of threat (0 = economic threat, 1 = security threat) and the number of interactions associated with those posts (0 = low, 1 = high) were included as covariates. Effect (θ) = conditional effect of toxic language on identity fusion with the author of the message at different levels of political ideology. The classification into three groups of political ideology to calculate conditional effects was based on the calculation of the 18th (2), 50th (4), and 85th (7) percentiles.

Finally, to address the fourth research question (RQ4), which involved testing a moderated mediation model, three separate analyses were conducted using the PROCESS macro for SPSS (Model 7; Hayes 2022), one for each dependent variable (see Table 3). As in the

previous moderation analysis, the independent variable (the level of toxic language in the hate speech message) was dummy-coded (0 = low, 1 = high). Political ideology was included as a continuous moderating variable, as it was assessed on an 11-point scale ranging from 0 (left) to 10 (right). The number of interactions associated with the post (0 = low, 1 = high) and the type of threat conveyed in the message (0 = economic threat, 1 = security threat) were also dummy-coded and entered as covariates in all models.

Table 3. Conditional indirect effects of the level of toxic language on outcome variables through narrative transportation and identity fusion with the author of the message. Moderated mediation model constructed using PROCESS (Model 7).

(a) Narrative transportation as mediating variable				
Toxic language → Narrative transportation → Message sharing intention				
Political ideology	**Values**	**Indirect effect (*SE*)**	**Boot *LLCI***	**Boot *ULCI***
Liberal	2.00	0.05 (0.05)	−0.045	0.159
Moderate	4.00	−0.01 (0.03)	−0.093	0.056
Conservative	7.00	−0.12 (0.06)	−0.255	−0.005
IMM = −0.03 (0.01) [95% CI: −0.071, −0.003]				
Toxic language → Narrative transportation → Attitudes toward immigrants				
Political ideology	**Values**	**Indirect effect (*SE*)**	**Boot *LLCI***	**Boot *ULCI***
Liberal	2.00	−0.01 (0.28)	−0.696	0.554
Moderate	4.00	0.00 (0.16)	−0.355	0.339
Conservative	7.00	0.03 (0.53)	−1.070	1.197
IMM = 0.01 (0.15) [95% CI: −0.292, 0.341]				
Toxic language → Narrative transportation → Support for harsh policies				
Political ideology	**Values**	**Indirect effect (*SE*)**	**Boot *LLCI***	**Boot *ULCI***
Liberal	2.00	−0.00 (0.01)	−0.032	0.030
Moderate	4.00	0.00 (0.00)	−0.018	0.016
Conservative	7.00	0.00 (0.02)	−0.056	0.054
IMM = 0.00 (0.00) [95% CI: −0.015, 0.015]				
(b) Identity fusion as mediating variable				
Toxic language → Identity fusion → Message sharing intention				
Political ideology	**Values**	**Indirect effect (*SE*)**	**Boot *LLCI***	**Boot *ULCI***
Liberal	2.00	−0.14 (0.06)	−0.284	−0.012
Moderate	4.00	−0.23 (0.06)	−0.372	−0.119
Conservative	7.00	−0.38 (0.12)	−0.631	−0.156
IMM = −0.04 (0.02) [95% CI: −0.103, 0.004]				
Toxic language → Identity fusion → Attitudes toward immigrants				
Political ideology	**Values**	**Indirect effect (*SE*)**	**Boot *LLCI***	**Boot *ULCI***
Liberal	2.00	1.78 (0.94)	0.148	3.777
Moderate	4.00	2.99 (0.94)	1.328	5.023
Conservative	7.00	4.80 (1.64)	1.860	8.329
IMM = 0.60 (0.34) [95% CI: −0.063, 1.319]				
Toxic language → Identity fusion → Support for harsh policies				
Political ideology	**MR**	**Indirect effect (*SE*)**	**Boot *LLCI***	**Boot *ULCI***
Liberal	2.00	−0.23 (0.11)	−0.470	−0.021
Moderate	4.00	−0.39 (0.10)	−0.602	−0.208
Conservative	7.00	−0.63 (0.18)	−1.007	−0.280
IMM = −0.08 (0.04) [95% CI: −0.166, 0.008]				

Note: The experimental condition (level of toxic language in the hate message) was set up as a dummy variable (0 = low, 1 = high). Narrative transportation (1 = low, 7 = high). Identity fusion with the author of the message (1 = low, 7 = high). Political ideology was assessed on an 11-point scale (from 0 = left to 10 = right). The type of threat (0 = economic threat, 1 = security threat) and the number of interactions associated with those posts (0 = low, 1 = high) were included as covariates. Indirect effect = indirect effect of toxic language on outcomes variables through narrative transportation and identity fusion with the author of the message. The classification into three groups of political ideology to calculate conditional effects was based on the calculation of the 18th (2), 50th (4), and 85th (7) percentiles. *IMM* = Index of Moderated Mediation.

The moderated mediation analyses were conducted using 10,000 bootstrapped samples to generate 95% confidence intervals based on the percentile method. According to this procedure, a (conditional) indirect effect is considered statistically significant if the confidence interval does not include zero. Additionally, to assess whether the indirect effect of toxic language on the dependent variables was significantly moderated by political ideology, the Index of Moderated Mediation (*IMM*) was examined. *IMM* quantifies the extent to which the indirect effect of an independent variable on a dependent variable (via a mediator) varies as a function of a moderator (Igartua and Hayes 2021). In other words, it assesses whether the mediation process is conditional on another variable. A statistically significant index (indicated by a 95% bootstrap confidence interval that does not contain zero) indicates the presence of a moderated mediation effect.

Regarding narrative transportation, the index of moderated mediation was statistically significant for the variable intention to share the message (*IMM* = −0.03, *SE* = 0.01, 95% *CI* [−0.071, −0.003]), but it was not significant for attitudes toward immigrants (*IMM* = 0.01, *SE* = 0.15, 95% *CI* [−0.292, 0.341]) or for support for harsh policies against irregular immigration (*IMM* = 0.00, *SE* = 0.00, 95% *CI* [−0.015, 0.015]). Hate messages without toxic language increased narrative transportation exclusively among individuals with a more conservative political ideology. Additionally, higher narrative transportation was associated with a greater intention to share the message (*B* = 0.29, *SE* = 0.04, $p < 0.001$). Moreover, narrative transportation was not associated with attitudes toward immigrants (*B* = −0.09, *SE* = 1.13, $p = 0.934$) nor with support for harsh policies against irregular immigration (*B* = −0.00, *SE* = 0.06, $p = 0.975$).

Regarding identity fusion with the author, the index of moderated mediation was not statistically significant for any of the three dependent variables. This indicates that hate messages lacking toxic language fostered identity fusion with the message's author—regardless of political ideology—which in turn was associated with a greater intention to share the message (*B* = 0.29, *SE* = 0.37, $p = 0.192$), more negative attitudes toward immigrants (*B* = −4.49, *SE* = 0.80, $p < 0.001$), and stronger support for harsh policies against irregular immigration (*B* = 0.59, *SE* = 0.04, $p < 0.001$) (Figure 4).

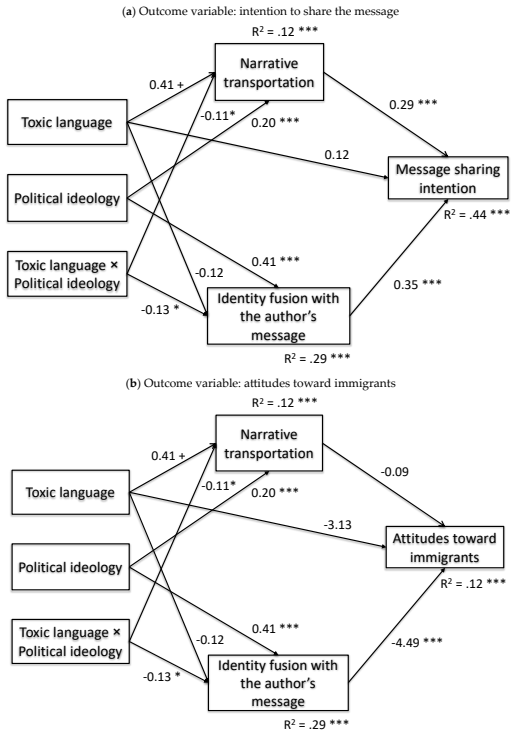

Figure 4. *Cont.*

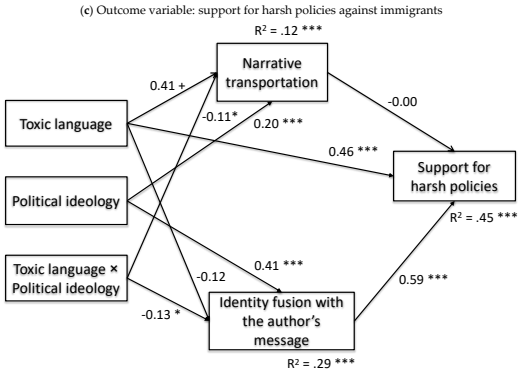

Figure 4. Conditional indirect effects of toxic language on intention to share the message (**a**), attitudes toward immigrants (**b**), and support for harsh policies against immigrants (**c**). ***Note***: The experimental condition (level of toxic language in the hate message) was set up as a dummy variable (0 = low, 1 = high). Narrative transportation (1 = low, 7 = high). Identity fusion with the author of the message (1 = low, 7 = high). Political ideology was assessed on an 11-point scale (0 = left, 10 = right). The type of threat (0 = economic threat, 1 = security threat) and the number of interactions associated with those posts (0 = low, 1 = high) were included as covariates. Effects of the covariates are not displayed to enhance clarity of the main results. Unstandardized regression coefficients are reported. + $p < 0.10$, * $p < 0.05$, *** $p < 0.001$.

4. Discussion

This study aimed to explore the persuasive mechanisms underlying hate speech on social media by testing the THREAD Model. Grounded in social identity theory, narrative persuasion, and toxic communication frameworks, the model was evaluated through a 2 × 2 between-subjects online experiment that manipulated both the level of toxic language and the popularity of testimonial hate messages targeting immigrants. The results challenge the prevailing assumption that the impact of hate speech stems solely from its use of overtly hostile language (e.g., Papcunová et al. 2023; Paasch-Colberg et al. 2021), revealing instead that subtler, less toxic expressions can foster a deeper psychological connection between audiences and the author of the message, particularly through the mechanism of identity fusion.

The results provide strong empirical support for the THREAD model's core premise that identity fusion with the author of the message is a central driver of persuasive outcomes in the context of online hate speech. Specifically, identity fusion was found to mediate the relationship between exposure to non-toxic hate messages and three key outcome variables: intention to share the message, negative attitudes toward immigrants, and support for strict immigration policies. This effect was robust and not moderated by political ideology, suggesting that the absence of toxic language enhances audience alignment with hostile viewpoints broadly, regardless of ideological stance.

Narrative transportation, in contrast, played a more limited role. While it did not emerge as a significant mediator of general persuasive outcomes, it was found to mediate the effect of non-toxic hate language on message sharing intentions, but only among participants with more conservative political ideologies. This conditional pathway reveals that individuals with more conservative political orientations (who are generally more critical of immigration; Davidov et al. 2020) are particularly prone to becoming immersed in the narrative when exposed to hate messages framed in a non-toxic tone. These findings align with Social Judgment Theory (Dal Cin et al. 2004; Perloff 2017), which posits that people evaluate persuasive messages based on their pre-existing attitudes. For conservative individuals, non-toxic hate narratives may fall within their latitude of acceptance, thereby increasing receptivity and reducing resistance.

Contrary to expectations, neither the number of interactions (popularity cues) nor their interaction with toxic language moderated the effects of message exposure on identity fusion or narrative transportation. This suggests that, at least in the context of testimonial hate speech, the linguistic and ideological congruence of the message outweighs the influence of perceived social endorsement, a finding that diverges from prior work emphasizing the role of popularity cues in social media (Sundar 2008; Dong and Li 2025; Dvir-Gvirsman 2019).

These findings validate the THREAD model as a theoretically grounded framework for analyzing the psychological impact of narrative hate messages, particularly those responding to sociopolitical disruptions. By focusing on identity fusion and narrative transportation as distinct but related mediators, the model advances our understanding of how individuals process and internalize online hate content. The model further incorporates message-level and individual-level moderators, such as political ideology and linguistic tone, that shape the persuasive dynamics of hate narratives. Notably, the THREAD model's emphasis on narrative structure and first-person testimonial formats aligns with current trends in social media discourse, making it highly applicable for future research on digital radicalization and stigmatization.

Moreover, the model's potential extends beyond the specific context of anti-immigrant hate speech. Future studies could apply the THREAD model to analyze hate narratives targeting other stigmatized groups (e.g., ethnic minorities, LGBTQ+ individuals, women, or religious communities), examining whether similar psychological mechanisms operate across different forms of online hostility.

Despite its contributions, the present study has some limitations. First, the distinction between "toxic language" and "hate speech" remains conceptually ambiguous in the literature, with some authors using these terms interchangeably (Kavaz et al. 2021; Salminen et al. 2020; Taulé et al. 2021). Although our operationalization of toxicity was empirically validated through a pilot study, future research should further disentangle these constructs and explore how different subtypes of toxicity influence persuasion.

Second, while political ideology emerged as a significant moderator of narrative transportation, other relevant individual difference variables were not included. Future research should examine the potential moderating effects of modern racism, social dominance orientation (SDO), or right-wing authoritarianism (RWA), which have all been linked to receptivity to exclusionary narratives (Cohrs and Asbrock 2009; Golec de Zavala et al. 2017; Igartua et al. 2019; Van Hiel and Mervielde 2005). Additionally, other potential mediators (such as moral disengagement, or empathy) should be explored to refine the explanatory power of the THREAD model. Lastly, the study focused on three main outcomes (message sharing intentions, attitudes toward immigrants, and support for strict immigration policies). Future work should consider alternative or complementary dependent variables, such as dehumanization of immigrants, intergroup anxiety, or threat perception.

From a practical standpoint, the findings have important implications for anti-hate speech interventions. Campaigns aiming to counteract online hate must address not only overtly toxic content but also more subtle, insidious forms of hate speech that use moderate tones to convey extreme ideas. These messages may be especially persuasive, as they avoid triggering audience resistance while enhancing identity fusion. Moreover, interventions should be tailored to individuals' political ideologies: conservative users may require distinct strategies, including narrative deconstruction tools that highlight manipulative rhetorical devices and identity fusion tactics.

Media literacy programs must be expanded to teach audiences how to recognize the emotional and psychological manipulation embedded in testimonial hate narratives. Rather than focusing solely on offensive vocabulary, these programs should emphasize how Rather than focusing solely on offensive vocabulary, these programs should emphasize how

even "civil" language can perpetuate harmful stereotypes and promote alignment with exclusionary ideologies. Funded media education initiatives could incorporate training modules on identity fusion detection and resistance strategies, with particular emphasis on how personal storytelling can distort perception under emotionally charged contexts. This is particularly important given that narrative messages often operate "under the radar," subtly shaping attitudes and beliefs while bypassing critical defenses such as counterarguing or psychological reactance (Bilandzic and Busselle 2013).

5. Conclusions

This study contributes novel empirical evidence to the growing body of work on digital hate speech by demonstrating that hate messages framed in a non-toxic tone can be even more persuasive than overtly hostile ones. It identifies identity fusion with the author of the message as the central mechanism underlying the influence of testimonial hate narratives, while highlighting narrative transportation as a conditional pathway for conservative individuals. The THREAD model provides a comprehensive framework for understanding the psychological impact of online hate, capturing how narratives (particularly those responding to disruptive sociopolitical events) operate to shape public opinion and legitimize hostility toward stigmatized groups.

By foregrounding the persuasive power of subtle hate speech, the study breaks new ground in media psychology and communication research. Its innovative integration of narrative persuasion, identity processes, and toxic discourse represents a significant theoretical and empirical advancement. As hate speech continues to evolve in digital spaces, the tools to analyze and counteract its influence must also become more nuanced, just as this study proposes.

Supplementary Materials: The supporting information can be downloaded at: https://osf.io/2a4xy/ (accessed on 22 January 2026).

Author Contributions: Conceptualization, J.-J.I. and C.A.B.-H.; methodology, J.-J.I.; formal analysis, J.-J.I.; investigation, C.A.B.-H.; data curation, C.A.B.-H.; writing—original draft, J.-J.I.; writing—review & editing, J.-J.I. and C.A.B.-H.; visualization, C.A.B.-H.; supervision, J.-J.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in Spain in accordance with institutional ethical standards, the 1964 Helsinki declaration and its later amendments, and national regulations for non-invasive, survey-based research with adults. In the institutions where the research was carried out, approval from an Institutional Review Board (IRB) or Research Ethics Committee (REC) was not required for this type of study.

Informed Consent Statement: All participants provided informed consent prior to participation, and procedures complied with applicable data-protection regulations (e.g., GDPR).

Data Availability Statement: All de-identified data, stimulus materials, and analysis scripts are available at the Open Science Framework (OSF): https://osf.io/2a4xy/.

Acknowledgments: During the preparation of this work, the authors used ChatGPT (OpenAI, GPT-5, 2025) solely to improve the fluency of the English text and the consistency of technical terminology. No AI tools were used for study design, data collection, statistical analyses, figure generation, or interpretation of results. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Abuín-Vences, Natalia, Ubaldo Cuesta-Cambra, José Niño-González, and Cristina Bengochea-González. 2022. Hate speech analysis as a function of ideology: Emotional and cognitive effects. *Comunicar* 30: 37–48. [CrossRef]

Ahmed, Saifuddin, Kokil Jaidka, Vivian H. H. Chen, Meng Cai, An Chen, Christopher S. Emes, Victor Yu, and Arul Chib. 2024. Social media and anti-immigrant prejudice: A multi-method analysis of the role of social media use, threat perceptions, and cognitive ability. *Frontiers in Psychology* 15: 1280366. [CrossRef]

Appel, Markus, Timo Gnambs, Tobias Richter, and Melanie C. Green. 2015. The transportation scale–short form (TS–SF). *Media Psychology* 18: 243–66. [CrossRef]

Arcila-Calderón, Carlos, Daniel Blanco-Herrero, and María B. Valdez-Apolo. 2020. Rejection and hate speech in Twitter: Content analysis of tweets about migrants and refugees in Spanish. *Revista Española de Investigaciones Sociológicas* 172: 21–40. [CrossRef]

Arcila-Calderón, Carlos, Daniel Blanco-Herrero, María Frías-Vázquez, and Francisco Seoane-Pérez. 2021. Refugees welcome? Online hate speech and sentiments in Twitter in Spain during the reception of the boat Aquarius. *Sustainability* 13: 2728. [CrossRef]

Arcila-Calderón, Carlos, Pedro Sánchez-Holgado, Javier Gómez, Mariana Barbosa, Hong Qi, Ana Matilla, Patricia Amado, Alejandro Guzmán, David López-Matías, and Teresa Fernández-Villazala. 2024. From online hate speech to offline hate crime: The role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanities and Social Sciences Communications* 11: 1369. [CrossRef]

Aron, Arthur, Elaine N. Aron, and David Smollan. 1992. Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology* 63: 596–612. [CrossRef]

Ayo, Femi E., Oludayo Folorunso, Femi T. Ibharalu, and Isaac A. Osinuga. 2020. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review* 38: 100311. [CrossRef]

Barbour, Jennifer B., Meeta J. Doshi, and Laura H. Hernández. 2016. Telling global public health stories: Narrative message design for issues management. *Communication Research* 43: 810–43. [CrossRef]

Bilandzic, Helena, and Rick Busselle. 2013. Narrative persuasion. In *The SAGE Handbook of Persuasion: Developments in Theory and Practice*, 2nd ed. Edited by James P. Dillard and Lijiang Shen. Thousand Oaks: Sage, pp. 200–19.

Bilewicz, Michał, and Wojciech Soral. 2020. Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology* 41: 3–33. [CrossRef]

Brehm, Jack W. 1966. *A Theory of Psychological Reactance*. New York: Academic Press.

Carlson, Christopher R. 2020. Hate speech as a structural phenomenon. *First Amendment Studies* 54: 217–24. [CrossRef]

Castaño-Pulgarín, Santiago A., Natalia Suárez-Betancur, Laura M. Tilano-Vega, and Héctor M. Herrera-López. 2021. Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior* 58: 101608. [CrossRef]

Chen, Gewi, and Jianning Dang. 2023. Crowds' malice behind the screen: The normative influences of online dehumanization on discrimination against foreigners. *Group Processes and Intergroup Relations* 26: 1548–64. [CrossRef]

Chinchilla, Javier, Antonio Vázquez, and Ángel Gómez. 2022. Strongly fused individuals feel viscerally responsible to self-sacrifice. *British Journal of Social Psychology* 61: 1067–85. [CrossRef]

Cohen, Jonathan. 2001. Defining identification: A theoretical look at the identification of audiences with media characters. *Mass Communication and Society* 4: 245–64. [CrossRef]

Cohen, Jonathan, Efrat Atad, and Tal Mevorach. 2023. Does it matter who tells the story? An experimental test of the effects of narrative perspective on credibility, identification, and persuasion. *Communication Research Reports* 40: 101–10. [CrossRef]

Cohrs, J. Christopher, and Felix Asbrock. 2009. Right-wing authoritarianism, social dominance orientation and prejudice against threatening and competitive ethnic groups. *European Journal of Social Psychology* 39: 270–89. [CrossRef]

Dahlstrom, Michael F., and Sonya Rosenthal. 2018. Third-person perception of science narratives: The case of climate change denial. *Science Communication* 40: 340–65. [CrossRef]

Dal Cin, Sonya, Mark P. Zanna, and Geoffrey T. Fong. 2004. Narrative persuasion and overcoming resistance. In *Resistance and Persuasion*. Edited by Eric S. Knowles and Jay A. Linn. New York: Psychology Press, pp. 175–92.

Davidov, Eldad, Daniel Seddig, Asher Gorodzeisky, Rebeca Raijman, Peter Schmidt, and Moshe Semyonov. 2020. Direct and indirect predictors of opposition to immigration in Europe: Individual values, cultural values, and symbolic threat. *Journal of Ethnic and Migration Studies* 46: 553–73. [CrossRef]

Dong, Yuxin, and Wei Li. 2025. Higher numbers = more important? Social media metrics and their agenda-cueing effects in anti-secondhand smoke persuasion. *Media Psychology* 28: 337–61. [CrossRef]

Dvir-Gvirsman, Shira. 2019. I like what I see: Studying the influence of popularity cues on attention allocation and news selection. *Information, Communication and Society* 22: 286–305. [CrossRef]

Eschmann, Randolph, Lian Guo, Jacob Groshek, Patrick Copeland, and Andrew Rochefort. 2025. Uncivil for Civil Rights: A machine learning and qualitative analysis of incivility in the X-based conversation about Black Lives Matter. *Computers in Human Behavior* 166: 108543. [CrossRef]

Essalhi-Rakrak, Abdelaziz, and Raquel Pinedo-González. 2023. #EspañaInvadida. Disinformation and hate speech towards refugees on Twitter: A challenge for critical thinking. *Profesional de la Información* 32: E320310. [CrossRef]

European Commission. 2022. *Integration of Immigrants in the European Union*. Special Eurobarometer 519. Available online: https://ec.europa.eu/migrant-integration/library-document/special-eurobarometer-integration-immigrants-european-union_en (accessed on 4 April 2025).

European Commission. 2024. *Public Opinion in the European Union*. Standard Eurobarometer 102. Available online: https://europa.eu/eurobarometer/surveys/detail/3215 (accessed on 4 April 2025).

Faul, Franz, Edgar Erdfelder, Axel G. Lang, and Albert Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175–91. [CrossRef]

Fernández, Isabel, Juan Igartua, Francisco Moral, Enrique Palacios, Teresa Acosta, and David Muñoz. 2013. Language use depending on news frame and immigrant origin. *International Journal of Psychology* 48: 772–84. [CrossRef]

Fino, Ana. 2020. Defining hate speech: A seemingly elusive task. *Journal of international Criminal Justice* 18: 31–57. [CrossRef]

Gächter, Simon, Chris Starmer, and Fabio Tufano. 2015. Measuring the closeness of relationships: A comprehensive evaluation of the 'Inclusion of the Other in the Self' Scale. *PLoS ONE* 10: e0129478. [CrossRef]

Golec de Zavala, Agnieszka, Rita Guerra, and Catarina Simão. 2017. The relationship between the Brexit vote and individual predictors of prejudice: Collective narcissism, right wing authoritarianism, social dominance orientation. *Frontiers in Psychology* 8: 2023. [CrossRef]

Green, Melanie C., and Timothy C. Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology* 79: 701–21. [CrossRef]

Haim, Mario, Anne S. Kümpel, and Hans-Bernd Brosius. 2018. Popularity cues in online media: A review of conceptualizations, operationalizations, and general effects. *Studies in Communication Media* 7: 186–207. [CrossRef]

Hayes, Andrew F. 2022. *Introduction to Mediation, Moderation, and Conditional Process Analysis. A Regression-Based Approach.*. New York: The Guilford Press.

Hietanen, Mikko, and Johan Eddebo. 2023. Towards a definition of hate speech: With a focus on online contexts. *Journal of Communication Inquiry* 47: 440–58. [CrossRef]

Hsueh, Michelle, Kumar Yogeeswaran, and Sanna Malinen. 2015. "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research* 41: 557–76. [CrossRef]

Huang, Jason L., Patrick G. Curran, Jennifer Keeney, Emily M. Poposki, and Richard P. DeShon. 2012. Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology* 27: 99–114. [CrossRef]

Igartua, Juan José, and Andrew F. Hayes. 2021. Mediation, moderation, and conditional process analysis: Concepts, computations, and some common confusions. *Spanish Journal of Psychology* 24: e49. [CrossRef]

Igartua, Juan José, and Diego Cachón-Ramón. 2023. Personal narratives for improving attitudes towards stigmatized immigrants: A parallel-serial mediation model. *Group Processes and Intergroup Relations* 26: 96–119. [CrossRef]

Igartua, Juan José, and Iñigo Guerrero-Martín. 2022. Personal migrant stories as persuasive devices: Effects of audience–character similarity and narrative voice. *Journal of Social and Political Psychology* 10: 21–34. [CrossRef]

Igartua, Juan José, Magdalena Wojcieszak, and Nuri Kim. 2019. How the interplay of imagined contact and first-person narratives improves attitudes toward stigmatized immigrants: A conditional process model. *European Journal of Social Psychology* 49: 385–97. [CrossRef]

Igartua, Juan José, Magdalena Wojcieszak, Diego Cachón-Ramón, and Iñigo Guerrero-Martín. 2017. "If it hooks you, share it on social networks". Joint effects of character similarity and imagined contact on the intention to share a short narrative in favor of immigration. *Revista Latina de Comunicación Social* 72: 1085–106. [CrossRef]

Ikeanyibe, Obiora M., Chukwuemeka C. Ezeibe, Peter O. Mbah, and Chijioke Nwangwu. 2018. Political campaign and democratisation. *Journal of Language and Politics* 17: 92–117. [CrossRef]

Joo, Minjeong, and Seung W. Park. 2017. Effect of identity fusion on decision to make extreme sacrifices in romantic relationships: The moderating role of impulsiveness. *British Journal of Social Psychology* 56: 819–27. [CrossRef]

Jost, John T., Mahzarin R. Banaji, and Brian A. Nosek. 2004. A decade of system justification Theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology* 25: 881–919. [CrossRef]

Kavaz, Elif, Anna Puig, Irene Rodriguez, Marta Taule, and Marc Nofre. 2021. Data visualisation for supporting linguists in the analysis of toxic messages. *Computer Science Research Notes* 3101: 59–70. [CrossRef]

Kim, Jae W., Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication* 71: 922–46. [CrossRef]

Klein, Jae W., Brock Bastian, Emmanuel N. Odjidja, Samhita S. Ayaluri, Christopher M. Kavanagh, Alimudin M. Mala, and Harvey Whitehouse. 2025. Identity fusion can foster intergroup trust and willingness to cooperate. *Communications Psychology* 3: 124. [CrossRef]

Klein, Olivier. 2024. Anti-immigrant rhetoric of populist radical right leaders on social media platforms. *Communications* 49: 400–20. [CrossRef]

Lilleker, Darren, and María Pérez-Escolar. 2023. Demonising migrants in contexts of extremism: Analysis of hate speech in UK and Spain. *Politics and Governance* 11: 2. [CrossRef]

Lingiardi, Vittorio, Nicola Carone, Giulia Semeraro, Chiara Musto, Marco D'Amico, and Silvia Brena. 2019. Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour and Information Technology* 39: 711–21. [CrossRef]

Matamoros-Fernández, Ariadna. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook, and YouTube. *Information, Communication and Society* 20: 930–46. [CrossRef]

McCroskey, James C., Virginia P. Richmond, and James A. Daly. 1975. The development of a measure of perceived homophily in interpersonal communication. *Human Communication Research* 1: 323–32. [CrossRef]

Meade, Adam W., and Scott B. Craig. 2012. Identifying careless responses in survey data. *Psychological Methods* 17: 437–55. [CrossRef]

Müller, Karsten, and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* 19: 2131–67. [CrossRef]

Obermaier, Martin, Ulrike K. Schmid, and Daniel Rieger. 2023. Too civil to care? How online hate speech against different social groups affects bystander intervention. *European Journal of Criminology* 20: 817–33. [CrossRef]

Paasch-Colberg, Sandra, Christoph Strippel, Jörg Trebbe, and Martin Emmer. 2021. From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication* 9: 171–80. [CrossRef]

Papcunová, Jana, Martin Martončik, Dana Fedáková, Marek Kentoš, Mária Bozogáňová, Ivan Srba, Róbert Moro, Martin Pikuliak, Michal Šimko, and Marek Adamkovič. 2023. Hate speech operationalization: A preliminary examination of hate speech indicators and their structure. *Complex and Intelligent Systems* 9: 2827–42. [CrossRef]

Perloff, Richard M. 2017. *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century.*. New York: Routledge.

Petty, Richard E., and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology* 19: 123–205. [CrossRef]

Pluta, Agnieszka, Jakub Mazurek, Jacek Wojciechowski, Tomasz Wolak, Wojciech Soral, and Michał Bilewicz. 2023. Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports* 13: 4127. [CrossRef] [PubMed]

Reicher, Stephen, Nick Hopkins, Mark Levine, and Robert Rath. 2005. Entrepreneurs of hate and entrepreneurs of solidarity: Social identity as a basis for mass communication. *International Review of the Red Cross* 87: 621–37. [CrossRef]

Rosenthal, Sonya, and Michael F. Dahlstrom. 2019. Perceived influence of proenvironmental testimonials. *Environmental Communication* 13: 222–38. [CrossRef]

Rösner, Laura, Sven Winter, and Nicole C. Krämer. 2016. Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior* 58: 461–70. [CrossRef]

Saleem, Muniba, Grace S. Yang, and Srividya Ramasubramanian. 2016. Reliance on direct and mediated contact and public policies supporting outgroup harm. *Journal of Communication* 66: 604–24. [CrossRef]

Saleem, Muniba, Sarah Prot, Craig A. Anderson, and Andrew F. Lemieux. 2017. Exposure to Muslims in media and support for public policies harming Muslims. *Communication Research* 44: 841–69. [CrossRef]

Salminen, Joni, Sezgin Sengün, Jorge Corporan, Soonho Jung, and Bernard J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLoS ONE* 15: e0228723. [CrossRef]

Saridou, Teodora, Javier J. Amores, Martín Oller-Alonso, Andrea Veglis, and Nikolaos Panagiotou. 2023. Negative perceptions and hate speech toward migrants and refugees in Southern Europe through social media. In *Migrants and Refugees in Southern Europe Beyond the News Stories: Photographs, Hate, and Journalists' Perceptions*. Edited by Carlos Arcila and Andrea Veglis. Lanham: Lexington Books, pp. 101–41.

Saumer, Maria, Kristina Maikovska, Andreas Neureiter, Andrea Čepelova, Hendrik Van Scharrel, and Jörg Matthes. 2024. Angry tweets. How uncivil and intolerant elite communication affects political distrust and political participation intentions. *Journal of Information Technology and Politics*. [CrossRef]

Schäfer, Sebastian, Irene Rebasso, Marie M. Boyer, and Anna M. Planitzer. 2024. Can we counteract hate? Effects of online hate speech and counter speech on the perception of social groups. *Communication Research* 51: 553–79. [CrossRef]

Song, Ji Yeon, Jack W. Klein, Young-Jae Cha, Sarah Goldy, Huan Sun, James Tisch, and Brock Bastian. 2025. From vastness to unity: Awe strengthens identity fusion. *Emotion advance online publication*. [CrossRef] [PubMed]

Soral, Wojciech, Michał Bilewicz, and Mirosław Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior* 44: 136–46. [CrossRef]

Sundar, Shyam S. 2008. The MAIN Model: A heuristic approach to understanding technology effects on credibility. In *Digital Media, Youth, and Credibility*. Edited by Miriam J. Metzger and Andrew J. Flanagin. Cambridge, MA: The MIT Press, pp. 73–100.

Sung, Kyung H., and Min J. Lee. 2015. Do online comments influence the public's attitudes toward an organization? Effects of online comments based on individuals' prior attitudes. *The Journal of Psychology* 149: 325–38. [CrossRef]

Swann, William B., Ángel Gómez, David C. Seyle, Juan F. Morales, and Carmen Huici. 2009. Identity fusion: The interplay of personal and social identities in extreme group behavior. *Journal of Personality and Social Psychology* 96: 995–1011. [CrossRef]

Swann, William B., Jr., Jolanda Jetten, Ángel Gómez, Harvey Whitehouse, and Brock Bastian. 2012. When group membership gets personal: A theory of identity fusion. *Psychological Review* 119: 441–56. [CrossRef]

Tajfel, Henri, and John C. Turner. 1979. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*. Edited by William G. Austin and Stephen Worchel. Monterey: Brooks/Cole, pp. 33–47.

Taulé, Marta, Anna Ariza, Marc Nofre, Enrique Amigó, and Paolo Rosso. 2021. Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments in Spanish. *Procesamiento del Lenguaje Natural* 67: 209–21. [CrossRef]

Van Hiel, Alain, and Ivan Mervielde. 2005. Authoritarianism and social dominance orientation: Relationships with various forms of racism. *Journal of Applied Social Psychology* 35: 2323–44. [CrossRef]

Van Laer, Tom, Ko de Ruyter, Luca M. Visconti, and Martin Wetzels. 2014. The Extended Transportation-Imagery Model: A Meta-Analysis of the antecedents and consequences of consumers' narrative transportation. *Journal of Consumer Research* 40: 797–817. [CrossRef]

Vázquez, Antonio, Ángel Gómez, Juan R. Ordoñana, William B. Swann, and Harvey Whitehouse. 2017. Sharing genes fosters identity fusion and altruism. *Self and Identity* 16: 569–85. [CrossRef]

Villegas-Lirola, Francisco, and Pedro Rodriguez-Martinez. 2025. The mediation (emotional self-control) and moderation (fun) of the relationship between receiving and perpetrating hate speech among boys and girls in Almeria (Spain). *Social Sciences* 14: 349. [CrossRef]

Wachs, Sebastian, Alessia Mazzone, Tijana Milosevic, Michelle F. Wright, Carmen Blaya, Manuel Gámez-Guadix, and John O. H. Norman. 2021. Online correlates of cyberhate involvement among young people from ten European countries: An application of the Routine Activity and Problem Behaviour Theory. *Computers in Human Behavior* 123: 106872. [CrossRef]

Walther, Joseph B. 2024. The effects of social approval signals on the production of online hate: A theoretical explication. *Communication Research*. [CrossRef]

Weber, Matthias, Christian Viehmann, Michael Ziegele, and Carsten Schemer. 2020. Online hate does not stay online—How implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in Human Behavior* 104: 106192. [CrossRef]

Wojcieszak, Magdalena, Nuri Kim, and Juan José Igartua. 2020. How to enhance the effects of mediated intergroup contact? Evidence from four countries. *Mass Communication and Society* 23: 71–106. [CrossRef]

Woods, Jessica. 2023. A systematic literature review of predictors of social media popularity. *Journal of Digital Social Research* 5: 62–92. [CrossRef]