Extended Abstract

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

Juan-José IGARTUA
University of Salamanca, Spain

Carlos BALLESTEROS
University of Valladolid, Spain

Paper presented to the Intergroup Communication Interest Group of the International Communication Association 75th Annual Conference ("Decoding Hostility: Discrimination, Hate Speech, and Outgroup Perceptions in Media and Organizations," Standard Paper Session). Denver (USA), 12-16 June 2025.

Abstract

This study investigates the influence of toxic language in hate speech targeting immigrants, particularly through narrative formats like first-person Twitter threads. Hate speech, defined as incitement or 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 x 2 between-subjects design that manipulates the presence of toxic language and message popularity. Key dependent variables include identity fusion with the author of the message, narrative transportation, intention to share the message, attitudes toward immigrants, and support for harsh immigration policies. 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: Hate Speech, toxic language, testimonial narratives, online engagement, narrative transportation, identity fusion with the author of the message, message sharing intention, attitudes towards immigrants, support for harsh policy against irregular immigration.

Hate speech is defined as any speech that incites, promotes, or instigates hatred, humiliation, or contempt against individuals or groups. It includes harassment, defamation, negative stereotypes, stigmatization, or threats based on characteristics such as race, ethnicity, age, disability, gender, sexual orientation, and more (Arcila et al., 2024; Castaño-Pulgarín et al., 2021;

Fino, 2020; Hietanen & Eddebo, 2023). Social media platforms have increasingly become spaces for the spread of hate speech against vulnerable and stigmatized groups, such as immigrants (Saridou et al., 2023; Müller & Schwarz, 2021). While prior research has predominantly focused on measuring the scope of the problem through content analysis and computational methods (Arcila et al, 2022; Ayo et al., 2020; Lingiardi et al., 2019; Matamoros-Fernández, 2017), there is limited research examining the effects of hate messages on recipients: how these messages influence the attitudes and behaviors of individuals who read them. Specifically, little is known about how the *toxic language* employed in hate speech influences message processing, and how individual differences moderate these effects.

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 while incorporating toxic language. Toxic language refers to the use of derogatory, dehumanizing, and hostile expressions aimed at stigmatizing, discrediting, or inciting hatred against individuals or groups based on their identity (Kim et al., 2021)[1]. 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 & Schwarz, 2021).

Our study investigates how the language used in hate speech against immigrants influences recipients depending on their political ideology. Previous research has shown that political conservatism is associated with stronger anti-immigration attitudes (Davidov et al., 2020), suggesting that individuals with more conservative views may be more susceptible to negative messaging about immigrants. This work bridges research on hate speech and narrative persuasion, particularly in understanding the mechanisms through which first-person testimonial narratives shape audience attitudes and the intention to propagate hate messages (e.g., Igartua & Cachón-Ramón, 2023; Guerrero-Martín & Igartua, 2021). We specifically examine the mediating roles of narrative transportation and identity fusion with the author (Swann et al., 2009). *Narrative transportation* refers to the psychological process in which individuals become absorbed in a story, leading to a change in their attitudes, beliefs, or behaviors (Appel et al., 2015; Green & Brock, 2000). *Identity fusion with the author of the message* is a concept that describes a psychological state in which individuals feel a strong sense of connection with the author of a message (Swann et al., 2012). This fusion can lead to increased empathy and identification with the author's perspective, making recipients more likely to accept the message and be influenced by it.

Toxic language may be perceived as extreme, potentially reducing both fusion with the author and narrative transportation. However, the extent to which these processes are moderated by the political ideology of the recipient remains unclear. 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).

In addition, the popularity of a message on social media is often indicated by *popularity cues*, such as the number of likes, shares, and comments

that the post accumulates (Dvir-Gvirsman 2019; Woods, 2023). These cues act as social endorsements, signaling the perceived approval or agreement with the message by a broader audience (Sung & Lee, 2015). The effect of these cues can significantly shape how individuals process and respond to hate speech (Sundar, 2008). Our study examines whether the presence of popularity cues moderates the effect of toxic language on identity fusion with the author of the message and on narrative transportation (RQ2). Additionally, we examine whether these effects are moderated by political ideology (RQ3).

Furthermore, we explore a full moderated mediation model by examining whether narrative transportation and identity fusion with the author of the message serve as parallel mediators in the relationship between toxic language and key outcomes, including the intention to share the message, attitudes toward immigrants, and support for stringent measures against irregular immigration. We investigate how political ideology moderates the effects of toxic language on narrative transportation and identity fusion, thereby influencing the indirect effects of these mediators on the dependent variables (RQ4).

Method[2]

The current study was conducted through an online experiment with a sample of 339 participants[3]. We used a 2 x 2 between-subjects design, manipulating two key factors: the presence of toxic language in the hate speech messages and the number of interactions (low vs. high) associated with those posts[4]. Testimonial messages were created in the form of threads, mimicking Twitter posts in which a user commented on a news story about the arrival of immigrants in Spain, highlighting the perceived threat this posed to Spanish society[5].

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. Next, participants were exposed to the experimental manipulation by reading a social media post presented as a testimonial message (Appendix 1). After reading the narrative, the post-test measures were administered, which included questions to evaluate the effectiveness of the experimental manipulation[6] and the outcome variables[7].

Results

The analyses to address the research questions posed in our study were conducted using the PROCESS macro for SPSS (Hayes, 2022). To answer the first research question, a moderation analysis (Model 1) was performed. It was observed that the absence of toxic language in the hate message increased identity fusion with the author, particularly among individuals with a more conservative political ideology (Table 1).

Additionally, it was found that the absence of toxic language enhanced narrative transportation with the message, but exclusively among individuals with a more conservative political ideology (Table 2). However, the number of interactions associated with those posts did not moderate the effect of toxic language in the hate speech message on identity fusion with the author ($B_{\text{interaction}} = 0.48$, $SE = 0.37$, $p = .192$) or narrative transportation ($B_{\text{interaction}} = 0.24$, $SE = 0.26$, $p = .365$). Furthermore, no three-way interaction was found among the presence of toxic

language in the hate speech message, the number of interactions associated with those posts, and political ideology (fusion with the author: $B_{\text{three-way interaction}}$ = -0.00, SE = 0.12, p = .967; narrative transportation: $B_{\text{three-way interaction}}$ = -0.13, SE = 0.10, p = .173).

Table 1

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

	Outcome variable: Identity fusion			
Model summary: R^2 = .29, p < .001	*B*	*SE*	*t*	*p*
Constant	0.84	0.24	3.38	<.001
Toxic language	−0.12	0.31	−0.41	.679
Political ideology	0.41	0.04	8.83	<.001
Toxic language × Political ideology	−0.13	0.06	−2.07	.038
Type of threat	0.15	0.16	0.97	.331
Number of interactions	0.09	0.16	0.59	.551
Conditional effects of toxic language at different levels of political ideology	*Effect*	*SE*	*t*	*p*
Liberal (progressive)	−0.39	0.21	−1.87	.061
Moderate	−0.66	0.16	−4.12	<.001
Conservative	−1.06	0.24	−4.35	<.001

Note.- The experimental condition (presence of toxic language in the hate message) was set up as a dummy variable (0 = without toxic language, 1 = with toxic language). 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.

Table 2

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

	Outcome variable: Narrative transportation			
Model summary: R^2 = .12, p < .001	*B*	*SE*	*t*	*p*
Constant	2.92	0.19	14.93	<.001
Toxic language	0.41	0.24	1.71	.087
Political ideology	0.20	0.03	5.63	<.001
Toxic language × Political ideology	−0.11	0.05	−2.34	.019
Type of threat	0.31	0.12	2.48	.013
Number of interactions	0.06	0.12	0.52	.601
Conditional effects of toxic language at different levels of political ideology	*Effect*	*SE*	*t*	*p*
Liberal (progressive)	0.18	0.16	1.09	.276
Moderate	−0.05	0.12	−0.44	.657
Conservative	−0.41	0.19	−2.14	.032

Note.- The experimental condition (presence of toxic language in the hate message) was set up as a dummy variable (0 = without toxic language, 1 = with toxic language). 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.

Finally, to test the full moderated mediation model (RQ4), three separate analyses were conducted using the PROCESS macro (Model 7), one for each dependent variable (Table 3). 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. 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, more negative attitudes toward immigrants, and stronger support for harsh policies against irregular immigration (Figure 1).

Table 3

Conditional indirect effects of toxic language on outcome variables through narrative transportation and identity fusion with the author of the message. Moderated mediation model 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 (presence of toxic language in the hate message) was set up as a dummy variable (0 = without toxic language, 1 = with toxic language). 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. 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 indirect effects was based on the calculation of the 18th (2), 50th (4), and 85th (7) percentiles. IMM = Index of Moderated Mediation.

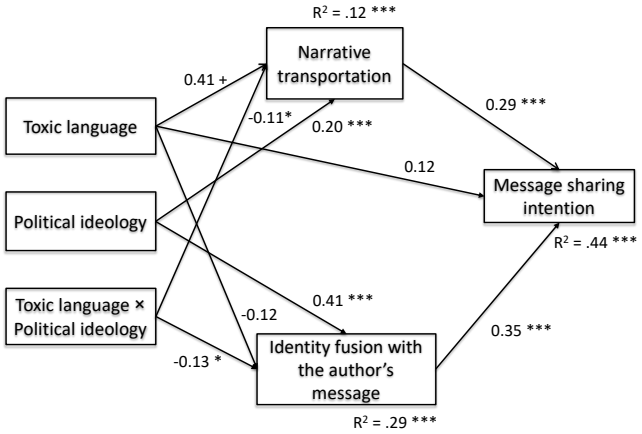
Conclusions and Discussion

Our findings reveal that hate messages without toxic language foster stronger identity fusion with the author than those using aggressive language. This identity fusion, rather than narrative transportation, emerges as the primary mechanism driving higher message sharing intentions, negative attitudes toward immigrants, and support for strict immigration policies. While narrative transportation did not serve as a general mediator, it did mediate the relationship between hate messages without toxic language and sharing intentions specifically among individuals with conservative ideologies. For these individuals, exposure to non-toxic hate messages increased transportation, which in turn heightened their intention to share the message. This pattern suggests that conservative individuals, often more critical of immigration (Davidov et al., 2020), are more receptive to hate messages with a moderate tone but extreme content, as these enhance both transportation and identity fusion. This result aligns with Social Judgment Theory (Dal Cin et al., 2004; Perloff, 2017), which emphasizes that people evaluate issues based on their pre-existing attitudes toward the topic. These findings challenge the assumption that toxic language alone radicalizes attitudes, revealing that more nuanced forms of hate speech can create closer alignment with hostile perspectives.

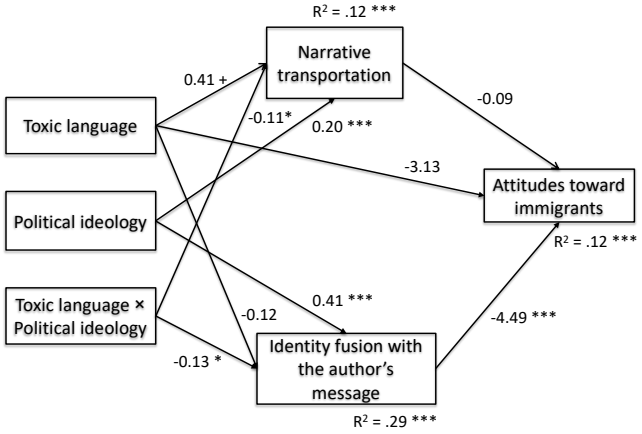
Figure 1

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)

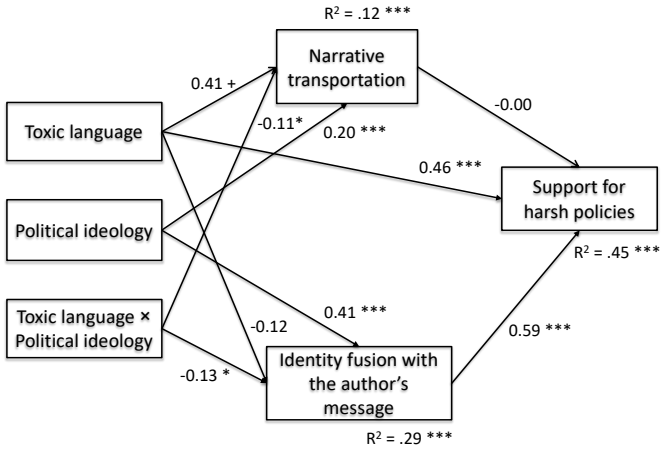
(a) Outcome variable: *intention to share the message*



(b) Outcome variable: *attitudes toward immigrants*



(c) Outcome variable: *support for harsh policies against immigrants*



Note.- The experimental condition (presence of toxic language in the hate message) was set up as a dummy variable (0 = without toxic language, 1 = with toxic language). 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.

Footnotes

[1] In the context of hate speech against immigrants, such language often includes slurs, threats, and inflammatory rhetoric that portray immigrants as a danger, a threat, or a burden to society (Arcila et al., 2024).

[2] All materials related to the online experiments (pilot and main studies), such as measures, datasets, syntax files, and testimonial messages used as experimental stimuli are available via the Open Science Framework (OSF): https://osf.io/2a4xy/?view_only=da390d31c0d24635b01500f2f5153d54

[3] The participant sample comprised 50.7% women, 47.5% men, 0.6% non-binary individuals, and 1.2% who preferred not to disclose their gender. Ages of participants ranged from 18 to 64 years (M = 26.98, SD = 11.57). In terms of employment status, 63.7% were students, while 33.3% were employed. Regarding educational attainment, 36.8% held a university degree. Political ideology was assessed on a scale from 0 (left) to 10 (right), with an average score of 4.15 (SD = 2.50).

[4] A pilot study (N = 41) was conducted to assess perceptions of certain expressions commonly used in tweets on Twitter that refer to individuals and social groups. This pilot study also included questions aimed at evaluating the interaction levels that tweets typically generate among users, reflected in the number of likes, comments, and retweets. Participants were asked to provide their opinions on what constitutes an important or relevant tweet based on these interactions. The insights gained from this pilot study were then used to craft messages that incorporated both mildly toxic and heavily toxic language, along with varying levels of interactions traditionally measured by the number of likes, shares, and comments associated with hate messages.

[5] 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*.

[6] 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. 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[285.38] = -11.08, p < .001), "The message used very aggressive language" (t[312.69] = -4.52, p < .001), and "The tone of the message I read can be considered very offensive or insulting" (t[287.43] = -9.64, p < .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 < .001), "The message I read had been shared many times" (t[333.44] = -6.47, p < .001), and "The message I read had been retweeted many times" (t[328.56] = -8.20, p < .001). These results indicate that the manipulation of the two independent variables was effective.

[7] Measures: fusion with the author of the message, assessed through the Inclusion of the Other in the Self (IOS) Scale (Gächter et al., 2015; M = 2.35, SD = 1.74), narrative transportation (Appel et al., 2015; α = .70, M = 3.93, SD = 1.22), intention to propagate or share the message (Igartua et al., 2017; M = 2.79, SD = 1.28, α = .83), attitudes toward immigrants and towards refugees measured using a feeling thermometer (Wojcieszak et al., 2020; r[337] = .80, p < .001; M = 60.68,

SD = 23.24), and support for harsh policies against irregular immigration (Bilewicz & Soral, 2020; Saleem et al., 2017; α = .85, *M* = 3.44, *SD* = 1.55).

References

Arcila, C., Amores, J. J., Sánchez-Holgado, P., Vrysis, L., Vryzas, N., & Oller-Alonso, M. (2022). How to detect online hate towards migrants and refugees? Developing and evaluating a classifier of racist and xenophobic hate speech using shallow and deep learning. *Sustainability*, *14*, 13094. https://doi.org/10.3390/su142013094

Arcila, C., Sánchez-Holgado, P., Gómez, J., Barbosa, M., Qi, H., Matilla, A., Amado, P., Guzmán, A., López-Matías, D. & Fernández-Villazala, T. (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*(1), 1–14. https://doi.org/10.1057/s41599-024-03899-1

Appel, M., Gnambs, T., Richter, T., & Green, M. C. (2015). The transportation scale–short form (TS–SF). *Media Psychology*, *18*(2), 243–266. https://doi.org/10.1080/15213269.2014.987400

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (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. https://doi.org/10.1016/j.cosrev.2020.100311

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*, 3–33. https://doi.org/10.1111/pops.12670

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Tilano-Vega, L. M., & Herrera-López, H. M. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

Dal Cin, S., Zanna, M. P., & Fong, G. T. (2004). Narrative persuasion and overcoming resistance. In E. S. Knowles & J. A. Linn (Eds.), *Resistance and persuasion* (pp. 175–192). Psychology Press.

Davidov, E., Seddig, D., Gorodzeisky, A., Raijman, R., Schmidt, P., & Semyonov, M. (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*(3), 553–573. https://doi.org/10.1080/1369183X.2018.1550152

Dvir-Gvirsman, S. (2019). I like what I see: Studying the influence of popularity cues on attention allocation and news selection. *Information, Communication & Society*, *22*(2), 286–305. https://doi.org/10.1080/1369118X.2017.1379550

Fino, A. (2020). Defining hate speech: a seemingly elusive task. *Journal of international Criminal Justice*, *18*(1), 31–57. https://doi.org/10.1093/jicj/mqaa023

Gächter S., Starmer C., & Tufano, F. (2015). Measuring the closeness of

relationships: a comprehensive evaluation of the 'Inclusion of the Other in the Self' Scale. *PLoS ONE*, *10*(6), e0129478. https://doi.org/10.1371/journal.pone.0129478

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Guerrero-Martin, I., & Igartua, J. J. (2021). Reduction of prejudice toward unaccompanied foreign minors through audiovisual narratives. Effects of the similarity and of the narrative voice. *Profesional de la Información*, *30*(2), e300203. https://doi.org/10.3145/epi.2021.mar.03

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis. A regression-based approach.* The Guilford Press.

Hietanen, M., & Eddebo, J. (2023). Towards a definition of hate speech: with a focus on online contexts. *Journal of Communication Inquiry*, *47*(4), 440–458. https://doi.org/10.1177/01968599221124309

Igartua, J. J., & Cachón-Ramón, D. (2023). Personal narratives for improving attitudes towards stigmatized immigrants: a parallel-serial mediation model. *Group Processes & Intergroup Relations*, *26*(1), 96–119. https://doi.org/10.1177/13684302211052511

Igartua, J. J., Wojcieszak, M., Cachón-Ramón, D., & Guerrero-Martín, I. (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*, 1085–1106. https://doi.org/10.4185/RLCS-2017-1209

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2019). Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, *39*(7), 711–721. https://doi.org/10.1080/0144929X.2019.1607903

Matamoros-Fernández, A. (2017). Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook, and YouTube. *Information, Communication & Society*, 20(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167. https://doi.org/10.1093/jeea/jvaa045

Perloff, R. M. (2017). *The dynamics of persuasion: communication and attitudes in the 21st century*. Routledge.

Saleem, M., Prot, S., Anderson, C. A., & Lemieux, A. F. (2017). Exposure to Muslims in media and support for public policies harming Muslims. *Communication Research*, *44*(6), 841–869. https://doi.org/10.1177/0093650215619214

Saridou, T., Amores, J. J., Oller- Alonso, M., Veglis, A., & Panagiotou, N.

(2023). Negative perceptions and hate speech toward migrants and refugees in Southern Europe through social media. In C. Arcila & A. Veglis (Eds.). *Migrants and refugees in Southern Europe beyond the news stories: photographs, hate, and journalists' perceptions* (pp. 101–141). Lexington Books.

Sundar, S. S. (2008). The MAIN Model: a heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). The MIT Press.

Swann, W. B., Gómez, A., Seyle, D. C., Morales, J. F., & Huici, C. (2009). Identity fusion: the interplay of personal and social identities in extreme group behavior. *Journal of Personality and Social Psychology*, *96*(5), 995–1011. https://doi.org/10.1037/a0013668

Wojcieszak, M., Kim, N., & Igartua, J. J. (2020). How to enhance the effects of mediated intergroup contact? Evidence from four countries. *Mass Communication and Society*, *23*(1), 71–106. https://doi.org/10.1080/15205436.2019.1630444

Woods, J. (2023). A systematic literature review of predictors of social media popularity. *Journal of Digital Social Research*, *5*(4), 62–92. https://doi.org/10.33621/jdsr.v5i4.181

Appendix

Examples of hate speech messages used in the experiment

Economic threat + No toxic language + Low number of interactions	Economic threat + Toxic language + High number of interactions