International Journal of Communication 19(2025), 974–997 1932–8036/20250005

I'll Share It When I Believe It! An Experimental Study on the Effects of Hateful and False Content on Credibility and Sharing Intention

DAVID BLANCO-HERRERO[1]
University of Amsterdam, Netherlands

DAMIAN TRILLING
Free University of Amsterdam, Netherlands

CARLOS ARCILA-CALDERÓN
University of Salamanca, Spain

The spread of false and hateful messages poses one of the greatest challenges in the current communication ecosystem. Although both phenomena have been widely analyzed, few studies have examined their combined effects. Using an experimental study with 404 Spanish citizens, we aim to understand the mechanism through which the credibility of false and hateful messages affects the intention to share them. We observed that the presence of falsehood and/or hatred seems to reduce the credibility of a message, which in turn has an indirect decreasing effect on the intention to share it, potentially informing future media literacy strategies to combat the spread of such messages by reducing citizens' propensity to believe them. Interestingly, a higher education level does not seem to increase resilience against these toxic discourses.

Keywords: anti-immigration discourses, credibility, disinformation, hate speech, shareworthiness

The transmission of information has become an increasingly complex process. Models that essentially assume one arrow between one sender node and one receiver node must be replaced with larger networks in which multiple users act as nodes and redistribute content (Carlson, 2016). This perspective connects with classical research on the ability of opinion leaders to influence the attitudes of citizens (Katz,

David Blanco-Herrero: d.blancoherrero@uva.nl
Damian Trilling: d.c.trilling@vu.nl
Carlos Arcila-Calderón: carcila@usal.es
Date submitted: 2023-10-12

[1] This study was supported by an FPU Grant awarded to the first author (FPU19/01455) and funded by the Ministry of Universities of Spain, and the project "Desarrollo y evaluación de un prototipo de detección automática de noticias falsas online (FakeDetector)" (Ref: PC-TCUE21-23_003), funded by the Fundación General de la Universidad de Salamanca.

Copyright © 2025 (David Blanco-Herrero, Damian Trilling, and Carlos Arcila-Calderón). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at http://ijoc.org.

1957; Lazarsfeld, Berelson, & Gaudet, 1948). This process is especially relevant in the fragmented environment of online media, in which the intermediation capacity of citizens has grown notably, so that opinion leaders can operate taking information created by the mass media or by other users and sharing it with their followers on social networks (Choi, 2015).

In this context, the spread of toxic discourses, such as disinformation or hate speech, poses a hard-to-control threat. Using the two-step flow of communication theory, we focus not on the origin or features of these messages, as is often the case, but on their reception and subsequent dissemination. This novel theoretical approach helps to understand these processes as a way to help counter the effects of disinformation and hate speech.

Although some experimental studies have tried to understand the factors influencing credibility or sharing intentions of false or hateful messages, mostly in the Anglo-Saxon context (e.g., Guess, Nagler, & Tucker, 2019; Mathew, Dutt, Goyal, & Mukherjee, 2019; Petit, Li, Millet, Ali, & Sun, 2021), it is unclear whether these results hold in other settings. Also, the mechanism through which believing and sharing are connected is not fully understood yet. Moreover, we explore this interaction when both types of toxic discourse are present. Thus, our object of study are false messages that promote hate speech, which includes two interrelated problematics: disinformation and hate speech.

The distinction between disinformation—intentionally false and harmful information—and misinformation—false information without a harm intention—(Wardle & Derakhshan, 2017) is beyond the scope of our study, given that we cannot address the intentionality of the message in our research. However, although the practical implications of combining dis- or misinformation and hate speech would be similar, the concerns about disinformation are bigger given the particular dimension and dangers of its intentionality.

Regarding hate speech, the European Commission against Racism and Intolerance (ECRI, 2016) defines this discourse as "the advocacy, promotion or incitement . . . of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group" (p. 3). These groups can be based on different features, including religion, sexual orientation, gender, ethnicity, or national origin.

Both issues are among the most relevant problems in current societies, and they have been extensively studied, but less is known about the additional challenges posed by their intersection (Schwarz & Holnburger, 2019). With this, our main goal is to comprehend the mechanism connecting the credibility and sharing intention of false and hateful content online. Thus, the objectives of the study are the following:

O1: To check whether the presence of hate and falsehood in a message influences its credibility and the intention to share it.

O2: To identify the role that credibility of a message plays in the intention to share it.

O3: To study whether education level influences the credibility and sharing intention of false and hateful content.

This study focuses on the Spanish setting, given the great concern about disinformation in this country compared with other nations (European Commission, 2023). Similarly, hate speech has been gaining presence in the country, partly associated with the ascension of anti-immigration and nationalistic discourses since the appearance of the far-right party Vox (Ferreira, 2019).

Given the plurality of hateful discourses and the different narratives they present, our study will focus exclusively on anti-immigration hate speech. The reasons for these choices are: (1) the need to use one consistent topic to avoid interferences between the variables; (2) the relevance of racism and xenophobia as main reasons behind hate crimes in Spain and most Western countries in the last years (Organization for Security and Co-operation in Europe [OSCE], n.d.); (3) the abundance of anti-immigration discourses that combine hate speech and disinformation (e.g., Schwarz & Holnburger, 2019); and (4) the observed capacity of this type of hate speech to lead to cases of real violence (Müller & Schwarz, 2020). Although these choices may limit generalizability, they open a path toward replication with other forms of hatred.

Theoretical Background and Related Research

In their meta-research on false news about COVID-19, García-Borrego and Casero-Ripollés (2022) identified studies that referred to belief, susceptibility, discernment capacity, detection, exposure, endorsement, recognition, or perception. Of these, belief has received the most attention in preceding studies because of its impact for (mis)information processing (Freiling, Krause, Scheufele, & Brossard, 2023), which is why our research will also focus on the credibility of a message.

Past works have used different approaches and focused on different aspects to address the credibility of false information. This great interest is explained by the need for disinformation to be believed to have an impact. Similarly, believing real information is essential for it to reach and inform societies. In general, most of these studies have found that false messages are less credible than true ones (Petit et al., 2021), although multiple individual and contextual factors determine this relationship (Allcott & Gentzkow, 2017; Altay, Nielsen, & Fletcher, 2024).

When studying hate speech, credibility has been very seldomly considered. However, even if hateful messages do not need to include factual inaccuracies to be considered as such, they are defined by a harmful intention and a particular discursive strategy that could affect the perceived credibility of the message. In fact, the presence of negative emotions or aggressiveness tends to decrease the credibility of a discourse (Abuín-Vences, Cuesta-Cambra, Niño-González, & Bengochea-González, 2022). At the same time, several studies have also identified the use of false information to promote hate speech (Evolvi, 2018; Schäfer & Schadauer, 2019).

That makes it necessary to comprehend whether the effects of hate speech on perceived credibility can be explained by its combined presence with falsehood or by other features inherent to hateful narratives, such as aggressiveness or negativity. Thus, the following research question is posed:

RQ1: Do the presence of falsehood, the presence of hate, and the joint presence of both affect the credibility of a piece of information?

Different studies have paid attention to the intentions to share information online. For instance, Berger and Milkman (2012) concluded that content that evokes strong feelings is shared more. This might explain why the spreadability of false messages appears to be higher (Vosoughi, Roy, & Aral, 2018). However, other studies show that sharing false information online is less prevalent and the responsibility of a few (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). Similarly, most hate speech seems to be shared by a few actors (Ribeiro, Calais, Santos, Almeida, & Meira, 2018). Previous work has also found that it can spread further and faster (Mathew et al., 2019) and has found the existence of a negativity bias, so that content with negative tones or those that provoke reactions of sadness and anger are shared more often (De León & Trilling, 2021).

Beyond these observations, less research has focused on the sharing intention of messages in which both problematics interact. Thus, the particular challenge posed by the use of false information to spread hate speech (Schwarz & Holnburger, 2019) demands an analysis that goes beyond the understanding of how the isolated presence of falsehood and hatred affect the sharing intentions of a message. Therefore, we pose the following research question:

RQ2: Do the presence of falsehood, the presence of hate, and the joint presence of both affect the intention to share a piece of information?

So far, we are presenting both credibility and intention to share as dependent variables to identify the effects of exposure to toxic discourses on these variables. Previous research is not conclusive about these expected effects, especially about sharing intention, and also the mechanism through which a citizen believes and shares a false and/or hateful message is unclear. To understand this mechanism, key for the spread of false and hateful discourses, we could consider credibility as an intermediate step, important in itself, but that cannot contribute to the spread of disinformation alone. Although some isolated studies have shown that people might be inclined to share information being aware of its falsity to obtain a political gain (Chadwick & Vaccari, 2019), it is generally assumed that believing something acts as a mediator of the intention to share it.

Various works have paid attention to this relationship, and although there is no conclusive evidence that believing something is an essential prerequisite for sharing it, general agreement indicates that credibility determines the intention to share. Petit et al. (2021) found that fake news about vaccination generated lower levels of perceived credibility, which subsequently decreased news-sharing intentions. Montero-Liberona and Halpern (2019) studied the credibility and the sharing process of false health information, noting that greater credibility of the content influences the possibility of sharing false news. Similarly, Bauer and Clemm von Hohenberg (2020) found that credibility is positively correlated with sharing. More related to our approach, Ali, Li, Zain-ul-abdin, and Zaffar (2022) identified perceived credibility as a mediator of the relationship between news veracity and cognitive elaboration, which in turn mediated the relationship between credibility and intention to share news. Although results were not conclusive, Stefanone, Vollmer, and Covert (2019) also explored the mediator role of perceived credibility between different content and individual factors and sharing behavior. This is similar to the approach used by Kumar, Shankar, Behl, Arya, and Gupta (2023),

who observed how perceived believability significantly mediated the relationship between psychological, source, and information features with fake news-sharing intention, thus offering a relevant antecedent for our study.

In total, there is strong evidence about the connection between credibility and the intention to share a message. However, there is no certainty about the mechanism of this interaction. Consequently, as a way to complement the observations about the direct effects in the first two research questions, the following hypothesis is posed:

H1: The credibility of information has a significant mediator role between exposure to false and/or hateful information and the intention to share said information.

These possible effects, however, do not occur in a general way or in all people equally. Previous research has indicated that various factors and personal features act as moderators when it comes to believing or sharing information. Before developing these elements in the following sections, Figure 1 shows the theoretical model that will be contrasted in RQ1, RQ2, and H1:

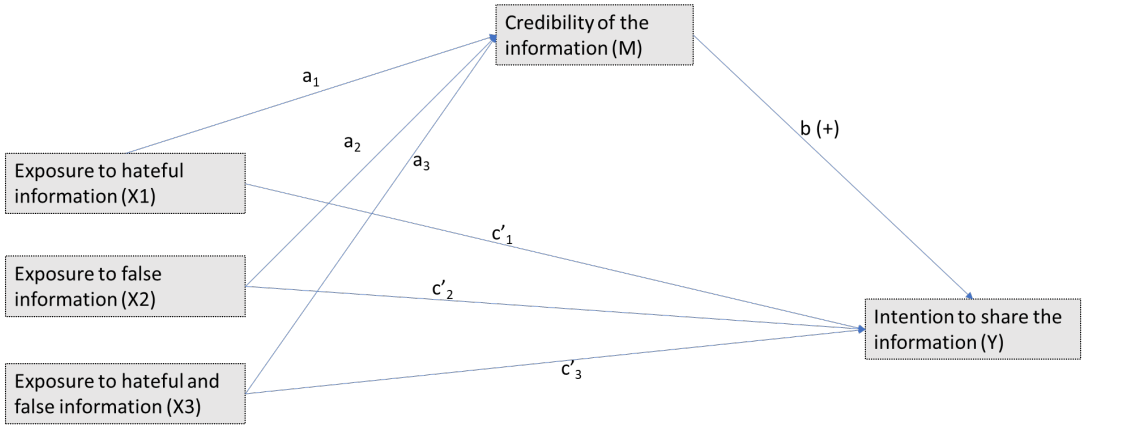

Figure 1. Mediation model. Source: The authors.[2]

The Influence of Education Level on Credibility and Sharing Intentions

Although the individual features that influence both credibility and intention to share are numerous (García-Borrego & Casero-Ripollés, 2022), our focus will be placed on one of the most determining variables, educational level. Parallel tests were run about ideology, another factor that has produced extensive literature; the results of this complementary study can be found as Supplementary Material in https://osf.io/z3c7d/?view_only=41c4be71225444008b05d59e4a945d68.

[2] In brackets, the expected direction of the effect is only hypothesized for the mediation effect (b).

Allcott and Gentzkow (2017) highlight the positive relationship between having obtained a higher education level and holding accurate beliefs about the news in the United States, a trend that has been confirmed in other settings (e.g., Altay et al., 2024; Humprecht, Esser, Van Aelst, Staender, & Morosoli, 2023; Rodríguez-Pérez & Canel, 2023). Other studies that have observed that less educated people are more likely to believe false information are those by García-Borrego and Casero-Ripollés (2022), Melki and colleagues (2021), or Schaewitz, Kluck, Klösters, and Krämer (2020).

There are fewer empirical approaches paying attention to the credibility of online hate messages. Thus, research on what personal features may be associated with a greater perception of credibility in hate messages could benefit from experimental studies. At the same time, the observations made about the credibility of false information often focus on specific case studies, such as the context of the COVID-19 pandemic or the presidential elections in the United States, so it is necessary to continue delving into the potential effects and mechanisms of the variables we are dealing with. Therefore, the following research question is posed:

RQ3: How does educational level moderate the effect of exposure to false and/or hateful information on the credibility of said information?

Among the multiple features studied to comprehend who spreads false information, ideology has been one of the most common (e.g., Guess et al., 2019), but also age (e.g., Guess, Aslett, Bonneau, Nagler, & Tucker, 2021), gender (Laato, Islam, Islam, & Whelan, 2020), or education level have received attention. About this last one, Buchanan (2020) observed that less educated people were more likely to share false information. About the propensity to share hate speech, various works have addressed the characteristics of the groups of users that disseminate these kinds of messages, although generally from the perspective of computational studies (Ribeiro et al., 2018), so it is difficult to establish with precision the influence of an individual's characteristics on their intention to share hateful content.

As was the case with credibility, observations on the effects of educational level on the intention to share false content are far from conclusive. And even fewer are the studies that have tested the effects of this variable on the intention to share hate messages. That is why the last research question is posed (shown together with RQ3 in Figure 2):

RQ4: How does educational level moderate the effect of exposure to false and/or hateful information on the intention to share such information?

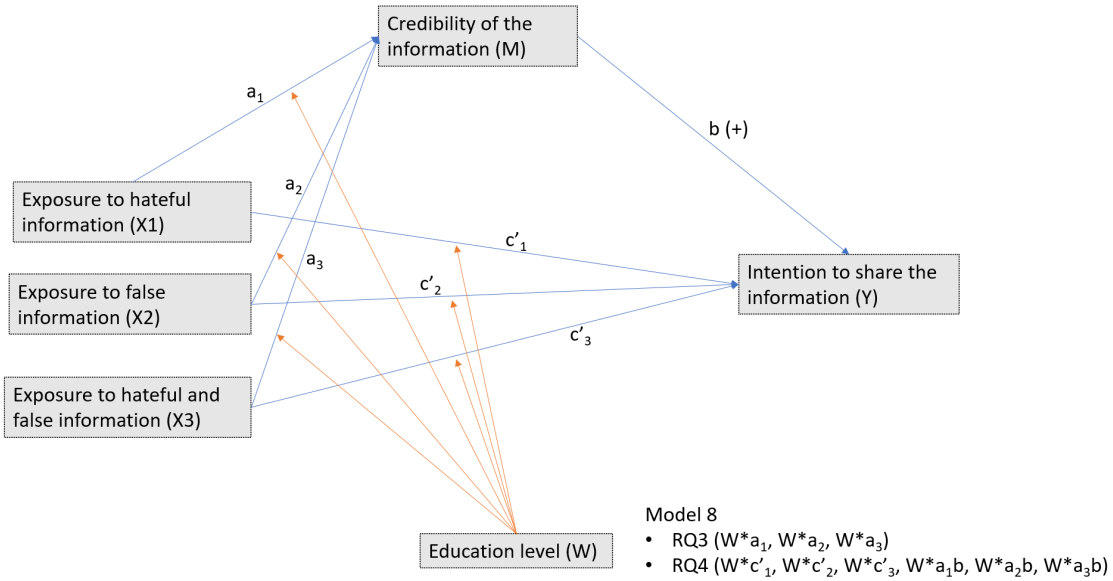

Figure 2. Moderated mediation model. Source: The authors.

Methodology

Sample and Procedure

This study used a 2 (presence or absence of falsehood) × 2 (presence or absence of hate) factorial experimental design. Participants were randomly assigned based on the four experimental conditions, so that a quarter of the sample was exposed to false hateful messages, a quarter to false messages without hate, a quarter to true hateful messages, and a quarter to true messages without hate (which was considered the reference category). To increase the external validity of the study, three messages were used in each condition, following the model of Van Duyn and Collier (2018). The order of the messages in each condition was randomized to avoid internal interaction effects.

The experiment was conducted online using a panel provided by Netquest with a sample stratified by sex, age group, and region. The questionnaire included quality checks, such as a question to ensure that the participant was paying attention, excluding those who did not pass them. A valid sample of 404 adult Spanish citizens was reached, 50.2% of them women and 49.8% men. Participants ranged in age from 18 to 90 years (M = 48.24; SD = 16.739). As can be seen in Table 1, 39.4% had obtained secondary education, and 16.8% had obtained a university bachelor's degree. On a scale between 1 and 10, with 1 being left and 10 being right, the average ideology was 4.83 (SD = 2.332).

Table 1. Sample Characteristics (n = 404).

Variable	Average or Frequency
Age	*M* = 48.24; *SD* = 16,739 (range: 18–90)
Sex	Men, 49.8%; Women, 50.2%
Educational level	Without studies, 2.0%; First grade studies, 1.5%; Second-degree studies of the first cycle, 10.9%; Second-cycle second-grade studies, 39.4%; First cycle third-grade studies, 14.9%; Second-cycle university studies, 16.8%; Third-degree university studies (Master's), 11.9%; Third-degree university studies (Doctorate), 2.7%
Ideology	*M* = 4.83; *SD* = 2,332 (range: 1–10)

Source: The authors.

The questionnaire had two parts: the first, with sociodemographic questions and pretest measures, and in the second, respondents were exposed to three stimuli of the same experimental category, responding after each of them to the posttest measurements. The fieldwork was carried out between April 5 and April 18, 2023.

Sample and Procedure

The stimuli showed brief informative messages, with the approximate length of a tweet or a news headline, that could be understood without further information. The length was always between 21 and 38 words (in Spanish). The information referred to three real cases, each one corresponding to one of the three main frameworks with which immigrants are associated in hateful discourses (Amores, 2022): social burden, crime, and terrorism. The stimuli can be found in the Supplementary Material at https://osf.io/z3c7d/?view_only=41c4be71225444008b05d59e4a945d68.

We followed Allcott and Gentzkow (2017) for the design of false news—understanding as such "intentionally and verifiably false news articles or messages" (p. 213)—collecting them from the database of one of the main fact-checking agencies in Spain, Maldita.es. Thus, the falsehood of the contents could be guaranteed. In the case of true content, we used information from legacy media, whose statements were previously reviewed to confirm their truthfulness. The presence of hate in the messages was validated through the online tool developed by Vrysis and colleagues (2021). Using these strategies, we could guarantee the veracity or falsity and the presence or absence of hate, so it was not deemed necessary to carry out a pretest, although the stimuli were reviewed by three experts.

All the stimuli were presented in the same tweet format, without information about users, preventing interferences from the trust associated with the person who shares or publishes content (Sterrett et al., 2019). The stimuli were designed with the utmost fidelity to the original messages, although on certain occasions they had to be adapted to guarantee that they complied with the appropriate length and style. Whenever the original content included profanity or grammatical errors, they were maintained, assuming that this type of expressions and errors are typical of messages on social networks and pseudomedia that share disinformation and hate. Nonetheless, we tried to use messages in which the language was not extreme or illegible, which could affect their perceptions.

Measures

The main measures used in the questionnaire were the following:

Credibility: It follows the measure developed by Meyer (1988) and used more recently, for instance, by Sterrett and colleagues (2019). It employs a scale between 1 (not at all) and 7 (a lot) to measure five features about the message: It is reasonable, it is biased (this item has a negative value and was inverted to generate the construct), it tells the complete story, it is accurate, and can be trusted (α = 0.820; *M* = 2.92; *SD* = 2.867). This variable is studied as a DV for RQ1 and RQ3, and as a MedV for H1 and RQ4.

Intention to share: The measure used two items. The first (Pennycook & Rand, 2020) asks whether the respondent would consider the possibility of sharing the content between 1 (in no case) and 7 (in all cases). The second (Bobkowski, 2015) asks what the probability is that the respondent shares the information, between 1 (very unlikely) and 7 (very likely). Both were combined, achieving a more robust measure (α = 0.950; *M* = 1.91; *SD* = 1.382).

General credibility: It was used as a control variable, since people who usually consider that the information presented in social networks is credible will be more predisposed to believe the messages of the experiment. Meyer's (1988) measure was used, posing the questions about information in social networks in general and not about specific content (α = 0.791; *M* = 2.97; *SD* = 1.078).

Propensity to share information on social networks: It was also proposed as a control variable, since it is assumed that people who never share as a rule will be less likely to do so with the type of information shown in this experiment. The measure comes from an investigation by the Pew Research Center (Kohut, Doherty, Dimock, & Keeter, 2010) and was used by Weeks and Holbert (2013). It was asked how often the respondent shares information on social networks, ranging between 1 (never) and 4 (regularly; *M* = 2.15; *SD* = 0.841).

In addition to the above, the questionnaire included questions on information consumption and attitudes toward immigration that were not part of this research.

Analysis

The data analysis was carried out mainly through two strategies. To answer RQ1 and RQ2, one-way analysis of variance (ANOVA) tests were performed. The effect size is reported through the η^2_p value, and, as in the rest of the study, the level of significance was set at $p < 0.05$. Second, to contrast H1 and answer the four RQs, the PROCESS macro (v. 4.3.1) by Hayes (2022) was used with a 95% CI. ANOVA tests were also performed to check the validity of the randomization process: the four conditions did not show significant differences in terms of educational level [$F(3, 400) = 0.290$, $p = 0.780$]. All the tests were carried out with SPSS (v. 28).

Results

Effects of the Presence of Falsehood, Hatred, and Both on Credibility

One-way ANOVA tests were carried out to evaluate the existence of differences in the credibility of a content depending on the different conditions. Significant differences were observed [$F(3, 221.465) = 8.900$, $p < 0.001$, $\eta^2_p = 0.059$]. True unhateful messages are considered most credible ($M = 3.35$; $SD = 1.080$), followed by false hateful content ($M = 2.97$; $SD = 1.302$). Less credible are false unhateful messages ($M = 2.78$; $SD = 1.010$), and the least were the true hateful messages ($M = 2.58$; $SD = 1.120$). These values can be seen more clearly in Figure 3, whereas post hoc tests can be found as Supplementary Material.

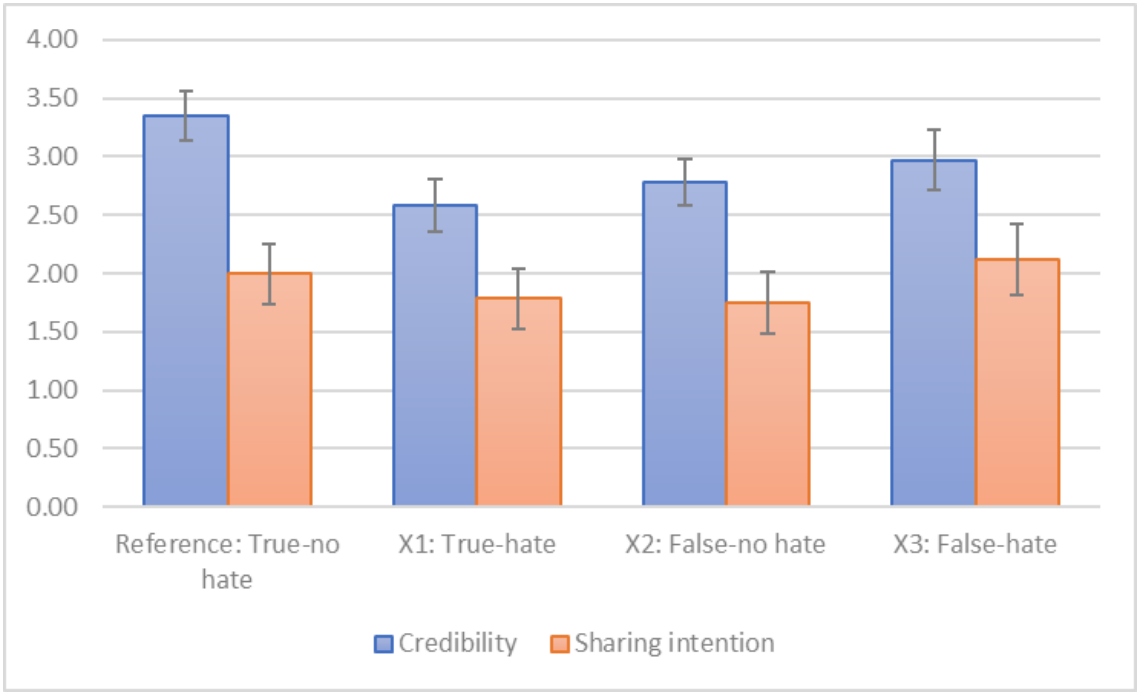

Figure 3. Mean credibility and sharing intention of the four experimental conditions. Source: The authors.

To address RQ1, model 4 was applied using the PROCESS macro for SPSS with 10,000 bootstrapping samples. The dependent variable (presence of falsehood and/or hate) was established as a multicategorical variable, so that the new variables X1 (presence of hate), X2 (presence of falsehood), and X3 (presence of hate and falsehood) are contrasted, with a reference category: true unhateful content. General credibility on social networks, general intention to share content on social networks, ideology, educational level, age, and gender were controlled as covariates. The presence of hate has a direct and negative effect on the credibility of a message [$B = -0.7988$, $SE = 0.1419$, $p < 0.001$], and although with lower coefficients, the same happens with the presence of falsehood [$B = -0.4462$, $SE = 0.1416$, $p < 0.01$]

and the joint presence of hate and falsehood [B = −0.4563, SE = 0.1421, p < 0.01]. These values are seen between the conditions (X) and the mediator variable (M) in Figure 4.

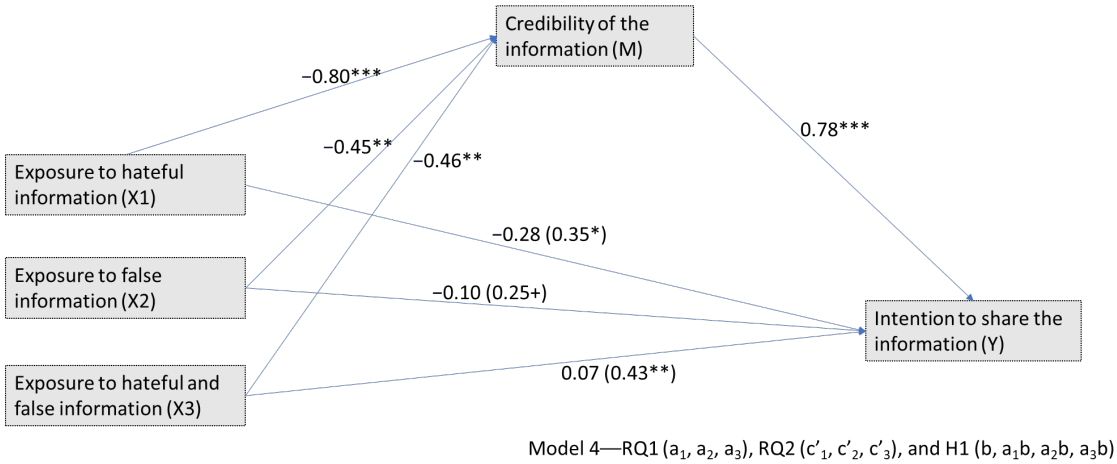

Figure 4. Effects of exposure to hateful and/or false content on credibility and sharing intention.
Source: The authors.[3]

Effects of the Presence of Falsehood, Hatred, and Both on Sharing Intention

One-way ANOVA tests did not find significant differences in the intention to share the four experimental conditions [F(3, 221.766) = 1.575, p = 0.196, η^2_p = 0.012]. Figure 3 shows the mean values, and the post hoc tests appear in the Supplementary Material.

However, the model 4 in the PROCESS macro did show significant direct effects once the covariates were considered. The presence of falsehood (X2) does not have significant effects on the intention to share (B = 0.2529, SE = 0.1371, p = 0.066), but the presence of hate has a direct and significantly positive effect on the sharing intention (B = 0.3492, SE = 0.1419, p < 0.05). Similarly, the joint presence of hate and falsehood has a positive, direct effect on the intention to share a message (B = 0.4285, SE = 0.1376, p < 0.01). These coefficients are found in parentheses between the independent (X) and the dependent variables (Y) in Figure 4.

[3] +, p < 0.1; *, p < 0.05; **, p < 0.01; ***, p < 0.001. The coefficients that appear between the conditions (X) and the independent variable (Y) represent the total effects (c_1, c_2, c_3), whereas the direct ones (c'_1, c'_2, c'_3) appear in brackets.

Mediating Role of Credibility of an Information on the Intention to Share It

PROCESS model 4 also contrasted the mediating role of credibility between exposure to falsehood, hate, and both together, with the intention to share a message hypothesized in H1. As shown in Figure 4, the direct effect of credibility on the intention to share is positive and large ($B = 0.7843$, $SE = 0.0482$, $p < 0.001$). When we evaluated the relative indirect effects of the different conditions on the intention to share mediated by credibility, we found significant effects in all three cases. In other words, although the presence of hate gave rise to a positive direct effect, increasing the intention to share content, when the mediating mechanism of credibility is considered, intention to share decreases. This happens when the message includes hate (*Effect* = −0.6265, $SE = 0.1254$, 95% *CI* [−0.8815, −0.3935]), when it includes falsehood (*Effect* = −0.3500, $SE = 0.1159$, 95% *CI* [−0.5844, −0.1287]), and when it combines both (*Effect* = −0.3579, $SE = 0.167$, 95% *CI* [−0.5908, −0.1317]).

Moderating Effects of Educational Level Between Exposure to the Experimental Conditions and Credibility

The moderation analyses presented in this section and in the next were carried out using model 8 of the PROCESS macro. The moderation was studied on the indirect effects (as it is common in this type of works) but also on the direct ones, given the exploratory approach of our article.

When evaluating the moderating role of educational level on the effect of the three experimental conditions on credibility, no statistically significant interaction effects were found for the presence of hate ($B = −0.0299$, $SE = 0.0981$, $p = 0.7610$), the presence of falsehood ($B = −0.0895$, $SE = 0.1009$, $p = 0.3757$), or the joint presence of hate and falsehood ($B = −0.1128$, $SE = 0.0947$, $p = 0.2339$). Figure 5 illustrates how, although these are not significant interaction effects, the negative effect on credibility of the conditions in which falsehood is present (X2 and X3) seems to be more intense, so that people with higher educational levels believe these messages less (in general, credibility is always lower as the educational level increases). Moreover, the analysis of the conditional effects showed no significant results, so it is assumed that the effects of the dependent variable on credibility occur in a similar way among people with different educational levels.

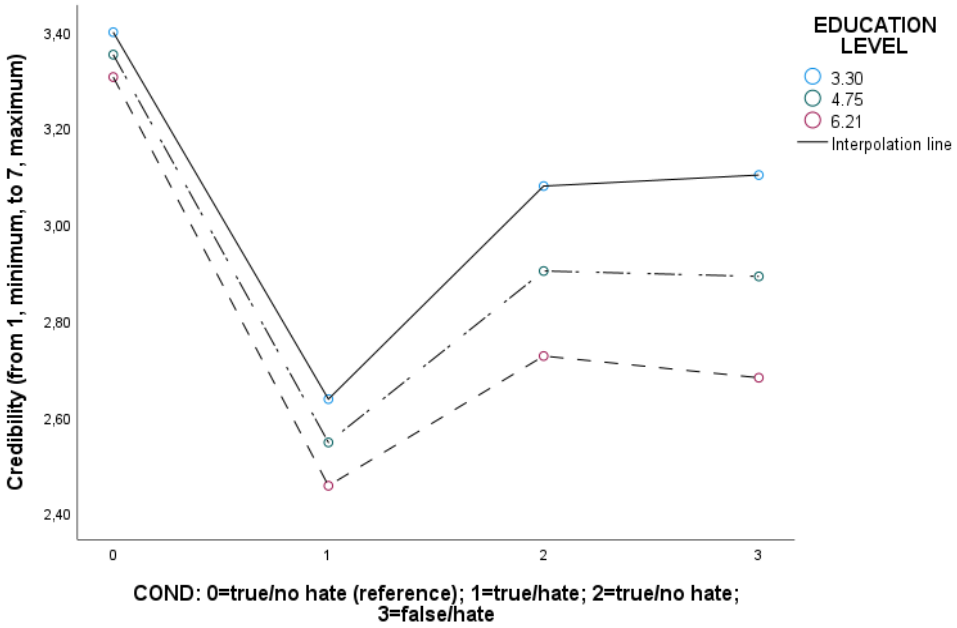

Figure 5. Credibility of the messages of each experimental condition in different educational levels. Source: The authors.[4]

Moderating Effects of Educational Level Between Exposure to Experimental Conditions and Sharing Intention

Also using model 8, we now turn to RQ4, which focused on the moderating role that educational level had on the effect of the presence of hate and/or falsehood on the intention to share information. No statistically significant interaction effects were observed between the presence of hate (B = 0.1096, SE = 0.0937, p = 0.2430), falsehood (B = 0.0403, SE = 0.0965, p = 0.6768), and hate and falsehood jointly (B = 0.1210, SE = 0.0906, p = 0.1826) with educational level on the intention to share. Figure 6 shows that, although the differences are not statistically significant, people with lower educational levels indicate a lower intention to share content without hate (reference condition and X2), whereas the situation is reversed for content including hate (X1 and X3), which tends to be more commonly shared by people with a higher educational level.

Moreover, educational levels moderate the direct effect of the different experimental conditions on the intention to share. Thus, the positive direct effect of the presence of hate (X1) and of the joint presence of hate and falsehood (X3) on the intention to share occurs only in people with a higher educational level. In the case of false and non-hate content (X2), there are no significant effects at any educational level. Table 2 details these values.

[4] The three points of the education level have been obtained by adding and subtracting the standard deviation from the mean, so that the three points correspond to M − SD, M, and M + SD.

On the other hand, the index of moderate mediation is not significant in any of the three conditions of the study. That is, the conditional relative indirect effects of exposure to true and hateful messages (X1) on the intention to share mediated by credibility are negative compared with the reference condition (true and non-hateful), and this is valid regardless of the educational level attained. This process is repeated for false messages without hate (X2) and for false messages with hate (X3), but in these conditions the reduction on the sharing intention mediated by credibility is not significant among people with lower educational levels, as seen in Table 3.

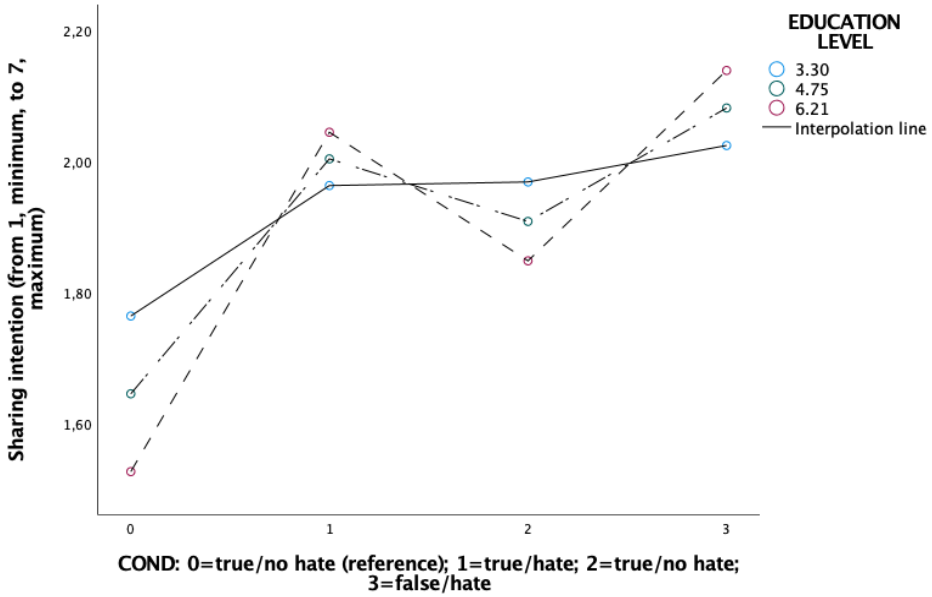

Figure 6. Intention to share the messages of each experimental condition in different educational levels. Source: The authors.

Table 2. Conditional Direct Effects of Exposure to Hateful and False Content on the Intention to Share at Different Educational Levels.

Education Level	Condition	Effects	*HE*	*p*
3.2991	X1	0.1989	0.1929	0.3032
4.7525	X1	0.3582*	0.1414	0.0117
6.2058	X1	0.5175**	0.1997	0.0099
3.2991	X2	0.2044	0.1942	0.2932
4.7525	X2	0.2629+	0.1373	0.0563
6.2058	X2	0.3214	0.1984	0.1060
3.2991	X3	0.2600	0.1871	0.1653
4.7525	X3	0.4358**	0.1379	0.0017
6.2058	X3	0.6116**	0.1942	0.0018

Source: The authors.

Note. +, $p < 0.1$; *, $p < 0.05$; **, $p < 0.01$.

Table 3. Conditional Indirect Effects of Exposure to Hateful and False Content on the Intention to Share Through Credibility at Different Educational Levels.

Condition	Education Level	Effects	Boots *HE*	Boot 95% *IQ*
X1 (IMM = −0.0235, 95% *CI* [−0.1822, 0.1263])	3.2991	−0.6004*	0.1639	[−0.9314, −0.2862]
	4.7525	−0.6345*	0.1264	[−0.8907, −0.4000]
	6.2058	−0.6687*	0.1772	[−1.0354, −0.3369]
X2 (IMM = −0.0704, 95% *CI* [−0.2466, 0.0819])	3.2991	−0.2517	0.1741	[−0.5897, 0.0966]
	4.7525	−0.3541*	0.1168	[−0.5882, −0.1312]
	6.2058	−0.4564*	0.1623	[−0.7928, −0.1553]
X3 (IMM = −0.0888, 95% *CI* [−0.2574, 0.0652])	3.2991	−0.2338	0.1736	[−0.5687, 0.1140]
	4.7525	−0.3628*	0.1174	[−0.5960, −0.1355]
	6.2058	−0.4919*	0.1595	[−0.8205, −0.1961]

Source: The authors.

Note. *, statistically significant indirect effects.

Discussion and Conclusion

In this study, we set out to comprehend the mechanism connecting the presence of falsehood and hatred on an online message with its credibility and sharing intention. First, in addressing RQ1, we found that false content is more likely to be perceived as such. This is consistent with the lower credibility of fake news observed in past research (Petit et al., 2021). When content is hateful, the same thing happens, even to a greater extent, so the presence of hate has a direct negative effect on credibility, making it decrease. This could show that the reduction of perceived credibility of hateful messages might not only be explained by their factual inaccuracy but also by the use of negativity or aggressiveness (Abuín-Vences et al., 2022). Finally, the joint presence of both conditions also undermines the credibility of a message, although the interaction of both phenomena does not lead to a stronger reduction of credibility; on the contrary, the reduction is smaller than when falsehood or hatred appear separately. An exploratory explanation for this is the coherence of the messages including both hatred and falsehood; that is, when a hate narrative does not need to stick to the truth, it can be designed to be more credible. Future studies might be able to shed more light on this issue by treating falsehood not as an independent variable but as a moderator of the credibility of hateful messages.

Answering RQ2, we observed that sharing intention increases when the content includes hate (whether in isolation or together with falsehood). Hence, the direct effect of the presence of hate or hate and falsehood on the sharing intention is positive, something that also happens with the presence of falsehood, although only tendentially. This could be because of what was observed by De León and Trilling (2021), who pointed out that a negative tone led to greater sharing intentions. Additionally, messages that seek to promote hate, especially if they are not limited by reality and introduce false information, use discursive strategies that infuriate or frighten people (Schmid, Kümpel, & Rieger, 2024), something that could also increase the intention to share them. As Maheshwari (2016) argues, when a consumer considers that a message is worrisome or important, even if they are not absolutely certain about its veracity, they could share it in case it was true. In our example, if immigrants commit acts as serious as some false and

hate narratives defend, it is understandable that people want to let their contacts know out of fear or anger. This connects with the cultivation theory and with what was indicated by Berger and Milkman (2012), who pointed out that content that evokes strong feelings is shared more.

This relationship reported on the second research question between the exposure to false and/or hateful messages and the intention to share them is significantly mediated by the credibility of the message, thus confirming H1. This observation helps us understand the mediating role of credibility, confirming previous research addressing the importance of believing a message to share it on social networks (Ali et al., 2022; Bauer & Clemm von Hohenberg, 2020; Kumar et al., 2023).

This discovery opens an important line of work to deal with the spread of false information and hate speech, not acting directly on the sharing action, but on its belief, since reducing this could reduce the spread of harmful content. Since less believed content is shared less, achieving a reduction in the belief of harmful content, for instance, using literacy strategies can help fight toxic discourses. This would not only make citizens less credulous in the face of this content but would indirectly help solve its spread thanks to the observed mediating effect. Accordingly, beyond measures with potential ethical and legislative limits (e.g., the removal or censoring of content), better education to detect falsehood and hatred would help limit its dissemination.

But all this needs to be taken cautiously, and this observation should be considered an exploratory complement in search for practical applications to counter toxic discourses. More studies are needed, as our original observation in RQ2 found that the presence of hate and the joint presence of hate and falsehood (and, to a lesser extent, the presence of falsehood) have a direct positive effect on the intention to share content. In contrast to this, when considering the mediating variable, it seems that the three conditions, compared with the reference condition, decrease credibility, which in turn makes their indirect effect on the sharing intention negative.

As a potential interpretation, this could mean that, although the positive direct effect of the presence of hate/falsehood on the intention to share could be driven by the emotional or provocative nature of the content, the negative indirect effect is explained by the reduced perception of credibility of these messages. In total, it could be claimed that hateful and/or false messages have negative direct effects on credibility, but among those who believe in the message, there is a strong direct effect on the sharing intention.

Although this would be hinting at a competitive mediation model, in which direct and indirect effects are both significant and point in different directions, further analysis is needed. It is also noteworthy that we consider credibility as a mediator, but it is not the only one that may be considered. For instance, Ali and colleagues (2022) also studied cognitive elaboration as a mediator predicting sharing intention.

When interpreting this result, it should also be considered that the values of both credibility (M = 2.92; SD = 1.163) and sharing intention (M = 1.91; SD = 1.382) are rather low, even in true non-hateful content. The general perceived credibility of information on social networks (M = 2.97; SD = 1.078) and the propensity to share it (M = 2.15; SD = 0.841) have been used as control variables, so this would not affect

what was studied in the model, but it is important to point it out. This could be because of the priming effect of asking about the credibility of social media, something that raises the levels of mistrust (Van Duyn & Collier, 2018), but also to the fact that many people do not generally share information on social networks (Altay, Hacquin, & Mercier, 2022).

Finally, evaluating the moderation role of educational level, whose effect was considered in RQ3 and RQ4, some partly unexpected observations were found. It is true that the indirect effects of exposure to harmful content on the intention to share it are greater among more educated people—that is, the reduction in the intention to share a hateful and/or false message through the reduction in the credibility occurs to a greater extent in more educated people—but a higher educational level does not have significant effects on credibility. Moreover, it seems that the presence of hate has a greater direct, positive effect on the sharing intention of a message among more educated people.

This positive effect of the presence of hate (isolated or in connection with falsehood) on the sharing intention among more educated people should be fully explored in future work. One possible explanation is their greater capacity for cognitive processing: If, as Haidt (2001) argues, moral reasoning is a post hoc construction generated after reaching a judgment, once false or hateful content is believed, more educated people will develop better arguments to support it, and, with those arguments, they might feel more confident sharing it.

Moreover, there were no significant interactions between the level of completed studies and the different experimental conditions, which implies that people with more years of education do not necessarily better identify falsehood or hate in a message. One reason for this may be that the pandemic tended to equalize these abilities by increasing the skills of people with a lower educational level (Casero-Ripollés, 2020). Another possible explanation is that formal education does not include media literacy, something that has been demanded in numerous studies (e.g., Sábada & Salaverría, 2023). In this line, Jones-Jang, Mortensen, and Liu (2019) pointed out that information literacy improved people's abilities to identify false news, but this did not happen with other types of literacies. Furthermore, Melki and colleagues (2021) found that having a university education decreased the credibility of false information about COVID-19, but only media literacy activities showed critical posting patterns that slowed its spread.

Because of their lower empowerment and ease of identification, minorities are often targeted by misinformation (Grambo, 2019), which makes it important to stop the spread and impact of rejection discourses, especially when supported by false information. To do that, our study has a clear practical implication in the fight against false and hateful narratives: the need for education to reduce credibility and also sharing intention. But formal education might not be enough, and more specific strategies related to media and technology literacy could be more effective—even more if they are tailored to the features of the target audience, for instance, based on ideological features, as it is also an important moderator (see Supplementary Material in https://osf.io/z3c7d/?view_only=41c4be71225444008b05d59e4a945d68). But also increasing empathy toward the victims of these discourses is important, as improving existing attitudes toward minorities might prevent people from believing and sharing false and hateful discourses against them.

Limitations and Future Lines of Work

As the scant literature on the credibility of hate messages demonstrated, credibility is more often considered in the context of disinformation. Indeed, at least at first glance, studying the credibility of a false message seems more relevant than studying that of a hate message. However, the dissemination of hate is carried out largely through information that is totally or partially false (Schwarz & Holnburger, 2019), manipulated, or taken of context, so there is an obvious relationship with credibility, which also requires an honesty component. In addition, although information may be true, the dissemination of hate is carried out surreptitiously, with a hidden intention in the message. Hence, we argue that reducing the credibility of a message is not only relevant when seeking to curb disinformation but also in the spread of hate.

Methodologically, it should be noted that for the analysis of the mediator (credibility) and dependent (sharing intention) variables, the mean of said variables measured after each of the three stimuli was used. It is assumed that, since the three stimuli belong to the same condition and are rotated, it is not necessary to resort to a multilevel model with random interception. It is true that this approximation, as also happens with grouped standard errors, could have a greater power of analysis; however, it could be problematic with the moderate mediation model that is proposed.

In line with Trilling, Tolochko, and Burscher (2017), we must also consider that asking about the probability of sharing a stimulus is a hypothetical assumption and does not replicate a real situation, since the absence of links or sources gives rise to an artificial situation. An alternative approach for future works, also not fully solving this issue but partly addressing it, is to do conjoint experiments in which people chose which of two alternatives they would *rather* share (e.g., Trilling & Knudsen, 2023). Although no strategy can replicate real-life sharing, with the forced choice variable, we can tap into their relative preferences. This could be complemented with approaches based on content dissemination patterns, for example, through social network analysis. In addition, our work has combined two indicators to obtain a more robust measure than previous studies, which have generally used a single item to measure sharing intention.

Another possible line of future research is to study, with longitudinal approaches, the effects of cultivation, checking whether the effects accelerate over time in those people exposed to false and/or hateful content. Thus, in addition to measuring the immediate effects, more lasting effects could be evaluated, something that is closer to the reality of consumption on social networks since people are not exposed to content in only an occasional and isolated way but rather on a recurring basis. Finally, we should note that our observations are valid for one type of hatred, that against immigrants; future studies would need to investigate whether these observations stand for other forms of rejection.

References

Abuín-Vences, N., Cuesta-Cambra, U., Niño-González, J. I., & Bengochea-González, C. (2022). Análisis del discurso de odio en función de la ideología: Efectos emocionales y cognitivos [Hate speech analysis as a function of ideology: Emotional and cognitive effects]. *Comunicar, 30*(71), 37–48. doi:10.3916/C71-2022-03

Ali, K., Li, C., Zain-ul-abdin, K., & Zaffar, M. A. (2022). Fake news on Facebook: Examining the impact of heuristic cues on perceived credibility and sharing intention. *Internet Research, 32*(1), 379–397. doi:10.1108/INTR-10-2019-0442

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236. doi:10.1257/jep.31.2.211

Altay, S., Hacquin, A. S., & Mercier, H. (2022). Why do so few people share fake news? It hurts their reputation. *New Media & Society, 24*(6), 1303–1324. doi:10.1177/1461444820969893

Altay, S., Nielsen, R. K., & Fletcher, R. (2024). News can help! The impact of news media and digital platforms on awareness of and belief in misinformation. *The International Journal of Press/Politics, 29*(2), 459–484. doi:10.1177/19401612221148981

Amores, J. J. (2022). *De la representación al odio. Desarrollo de nuevas estrategias para entender los discursos de odio hacia los migrantes y refugiados: detectando, midiendo y analizando los encuadres visuales, las actitudes y los mensajes de rechazo a la migración* [From representation to hate. Developing new strategies to understand hate speech towards migrants and refugees: Detecting, measuring and analysing visual frames, attitudes and messages against migration] (Doctoral dissertation). University of Salamanca, Salamanca, Spain. Retrieved from https://gredos.usal.es/handle/10366/150702

Bauer, P. C., & Clemm von Hohenberg, B. (2020). Believing and sharing information by fake sources: An experiment. *Political Communication, 38*(6), 647–671. doi:10.1080/10584609.2020.1840462

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192–205. doi:10.1509/jmr.10.0353

Bobkowski, P. S. (2015). Sharing the news: Effects of informational utility and opinion leadership on online news sharing. *Journalism & Mass Communication Quarterly, 92*(2), 320–345. doi:10.1177/1077699015573194

Buchanan, T. (2020). Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *PLoS One, 15*(10), 1–33. doi:10.1371/journal.pone.0239666

Carlson, M. (2016). Embedded links, embedded meanings. Social media commentary and news sharing as mundane media criticism. *Journalism Studies, 17*(7), 915–924. doi:10.1080/1461670X.2016.1169210

Casero-Ripollés, A. (2020). Impact of COVID-19 on the media system. Communicative and democratic consequences of news consumption during the outbreak. *El Profesional de la Información, 29*(2), 1–11. doi:10.3145/epi.2020.mar.23

Chadwick, A., & Vaccari, C. (2019). *News sharing on UK social media: Misinformation, disinformation, and correction*. Loughborough, UK: Online Civic Culture Centre, Loughborough University.

Choi, S. (2015). The two-step flow of communication in Twitter-based public forums. *Social Science Computer Review, 33*(6), 696–711. doi:10.1177/0894439314556599

De León, E., & Trilling, D. (2021). A sadness bias in political news sharing? The role of discrete emotions in the engagement and dissemination of political news on Facebook. *Social Media + Society, 7*(4), 1–12. doi:10.1177/20563051211059710

European Commission. (2023). *Standard Eurobarometer 98 – Winter 2022–2023: Public opinion in the European Union, Annexes.* Brussels, Belgium*:* European Commission, Directorate-General for Communication. doi:10.2775/460956

European Commission Against Racism and Intolerance. (2016). *ECRI general policy recommendation N°15 on combating hate speech*. Retrieved from https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

Evolvi, G. (2018). Hate in a tweet: Exploring Internet-based Islamophobic discourses. *Religions, 9*(10), 307. doi:10.3390/rel9100307

Ferreira, C. (2019). Vox como representante de la derecha radical en España: un estudio sobre su ideología [Vox as representative of the radical right in Spain: A study of its ideology]. *Revista Española de Ciencia Política, 51*, 73–98. doi:10.21308/recp.51.03

Freiling, I., Krause, N. M., Scheufele, D. A., & Brossard, D. (2023). Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. *New Media & Society, 25*(1), 141–162. doi:10.1177/14614448211011451

García-Borrego, M., & Casero-Ripollés, A. (2022). ¿Qué nos hace vulnerables frente las noticias falsas sobre la COVID-19? Una revisión crítica de los factores que condicionan la susceptibilidad a la Desinformación [What makes us vulnerable to COVID-19 fake news? A critical review of the factors conditioning susceptibility to misinformation]. *Estudios sobre el Mensaje Periodístico, 28*(4), 789–801. doi:10.5209/esmp.82881

Grambo, K. (2019). Fake news and racial, ethnic, and religious minorities: A precarious quest for truth. *Journal of Constitutional Law, 21*(5), 1299–1345. Retrieved from https://scholarship.law.upenn.edu/jcl/vol21/iss5/4

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374–378. doi:10.1126/science.aau2706

Guess, A., Aslett, K., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Cracking open the news feed: Exploring what U.S. Facebook users see and share with large-scale platform data. *Journal of Quantitative Description: Digital Media, 1*(2021), 1–48. doi:10.51685/jqd.2021.006

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), 1–8. doi:10.1126/sciadv.aau4586

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. doi:10.1037//0033-295X.108.4.814

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY: Guilford Press.

Humprecht, E., Esser, F., Van Aelst, P., Staender, A., & Morosoli, S. (2023). The sharing of disinformation in cross-national comparison: Analyzing patterns of resilience. *Information, Communication & Society, 26*(7), 1342–1362. doi:10.1080/1369118X.2021.2006744

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information Literacy helps, but other literacies don't. *American Behavioral Scientist, 65*(2), 371–388. doi:10.1177/0002764219869406

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly, 21*(1), 61–78. doi:10.1086/266687

Kohut, A., Doherty, C., Dimock, M., & Keeter, S. (2010). *Americans spending more time following the news*. Pew Research Center. Retrieved from https://core.ac.uk/download/pdf/30682252.pdf

Kumar, A., Shankar, A., Behl, A., Arya, V., & Gupta, N. (2023). Should I share it? Factors influencing fake news-sharing behaviour: A behavioural reasoning theory perspective. *Technological Forecasting and Social Change, 193*(2023), 1–14. doi:10.1016/j.techfore.2023.122647

Laato, S., Islam, A. N., Islam, M. N., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems, 29*(3), 288–305. doi:10.1080/0960085X.2020.1770632

Lazarsfeld, P., Berelson, B., & Gaudet, H. (1948). *The people's choice*. New York, NY: Columbia University Press.

Maheshwari, S. (2016, November 20). How fake news goes viral: A case study. *The New York Times*. Retrieved from https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In *WebSci '19: Proceedings of the 10th ACM conference on web science* (pp. 173–182). New York, NY: ACM. doi:10.1145/3292522.3326034

Melki, J., Tamim, H., Hadid, D., Makki, M., El Amine, J., & Hitti, E. (2021). Mitigating infodemics: The relationship between news exposure and trust and belief in COVID-19 fake news and social media spreading. *PLoS One*, *16*(6), 1–13. doi:10.1371/journal.pone.0252830

Meyer, P. (1988). Defining and measuring credibility of newspapers: Developing an index. *Journalism Quarterly, 65*(3), 567–574. doi:10.1177/107769908806500301

Montero-Liberona, C., & Halpern, D. (2019). Factores que influyen en compartir noticias falsas de salud online [Factors that influence sharing online fake news on health]. *El Profesional de la Información, 28*(3), 1–9. doi:10.3145/epi.2019.may.17

Müller, K., & Schwarz, C. (2020). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association, 19*(4), 2131–2167. doi:10.1093/jeea/jvaa045

Organization for Security and Co-operation in Europe. (n.d.). *OSCE-ODHIR. Hate crime reporting*. Retrieved from https://hatecrime.osce.org

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*(2), 185–200. doi:10.1111/jopy.12476

Petit, J., Li, C., Millet, B., Ali, K., & Sun, R. (2021). Can we stop the spread of false information on vaccination? How online comments on vaccination news affect readers' credibility assessments and sharing behaviors. *Science Communication, 43*(4), 407–434. doi:10.1177/10755470211009887

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., & Meira Jr., W. (2018). Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media* (pp. 676–679). Washington, DC: AAAI. doi:10.1609/icwsm.v12i1.15057

Rodríguez-Pérez, C., & Canel, M. J. (2023). Exploring European citizens' resilience to misinformation: Media legitimacy and media trust as predictive variables. *Media and Communication, 11*(2), 30–41. doi:10.17645/mac.v11i2.6317

Sábada, C., & Salaverría, R. (2023). Combatir la desinformación con alfabetización mediática: análisis de las tendencias en la Unión Europea [Combatting disinformation with media literacy: Analysis of trends in the European Union] *Revista Latina de Comunicación Social, 81*, 17–33. doi:10.4185/RLCS-2023-1552

Schaewitz, L., Kluck, J. P., Klösters, L., & Krämer, N. C. (2020). When is disinformation (in)credible? Experimental findings on message characteristics and individual differences. *Mass Communication & Society, 23*(4), 484–509. doi:10.1080/15205436.2020.1716983

Schäfer, C., & Schadauer, A. (2019). Online fake news, hateful posts against refugees, and a surge in xenophobia and hate crimes in Austria. In G. Dell'Orto & I. Wetzstein (Eds.), *Refugee news, refugee politics: Journalism, public opinion and policymaking in Europe* (pp. 109–116). Oxfordshire, UK: Routledge.

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2024). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society, 26*(5), 2614–2632. doi:10.1177/14614448221091185

Schwarz, K., & Holnburger, J. (2019). Disinformation: What role does disinformation play for hate speech and extremism on the internet and what measures have social media companies taken to combat it? In J. Baldauf, J. Ebner, & J. Guhl (Eds.), *Hate speech and radicalisation online. The OCCI research report* (pp. 35–43). London, UK: ISD.

Stefanone, M. A., Vollmer, M., & Covert, J. M. (2019, July). In news we trust? Examining credibility and sharing behaviors of fake news. In *Proceedings of the 10th international conference on social media and society* (pp. 136–147). New York, NY: ACM. doi:10.1145/3328529.3328554

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., . . . Loker, K. (2019). Who shared it?: Deciding what news to trust on social media. *Digital Journalism, 7*(6), 783–801. doi:10.1080/21670811.2019.1623702

Trilling, D., & Knudsen, E. (2023). Drivers of news sharing: How context, content, and user features shape sharing decisions on Facebook. *Digital Journalism* [Latest Articles], 1–22. doi:10.1080/21670811.2023.2255224

Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & Mass Communication Quarterly, 94*(1), 38–60. doi:10.1177/1077699016654682

Van Duyn, E., & Collier, J. (2018). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society, 22*(1), 29–48. doi:10.1080/15205436.2018.1511807

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. doi:10.1126/science.aap9559

Vrysis, L., Vryzas, N., Kotsakis, R., Saridou, T., Matsiola, M., Veglis, A., . . . Dimoulas, C. (2021). A web interface for analyzing hate speech. *Future Internet, 13*(3), 1–18. doi:10.3390/fi13030080

Wardle, C., & Derakhshan, H. (2017). *Information disorder. Toward an interdisciplinary framework for research and policymaking*. Strasbourg, France: Council of Europe.

Weeks, B. E., & Holbert, R. L. (2013). Predicting dissemination of news content in social media: A focus on reception, friending, and partisanship. *Journalism & Mass Communication Quarterly, 90*(2), 212–232. doi:10.1177/1077699013482906