Aggression and Violent Behavior 58 (2021) 101608

Contents lists available at ScienceDirect

Aggression and Violent Behavior

journal homepage: www.elsevier.com/locate/aggviobeh

Internet, social media and online hate speech. Systematic review

Sergio Andrés Castaño-Pulgarín [a,*], Natalia Suárez-Betancur [b], Luz Magnolia Tilano Vega [c], Harvey Mauricio Herrera López [d]

[a] Psychology Department, Corporación Universitaria Minuto de Dios-UNIMINUTO, Colombia
[b] Corporación para la Atención Psicosocial CORAPCO, Medellín, Colombia
[c] Psychology Department, Universidad de San Buenaventura, Medellín, Colombia
[d] Psychology Department, Universidad de Nariño, Pasto, Colombia

ARTICLE INFO

Keywords:
Cyberhate
Internet
Online hate speech
Social Networks

ABSTRACT

This systematic review aimed to explore the research papers related to how Internet and social media may, or may not, constitute an opportunity to online hate speech. 67 studies out of 2389 papers found in the searches, were eligible for analysis. We included articles that addressed online hate speech or cyberhate between 2015 and 2019. Meta-analysis could not be conducted due to the broad diversity of studies and measure units. The reviewed studies provided exploratory data about the Internet and social media as a space for online hate speech, types of cyberhate, terrorism as online hate trigger, online hate expressions and most common methods to assess online hate speech. As a general consensus on what is cyberhate, this is conceptualized as the use of violent, aggressive or offensive language, focused on a specific group of people who share a common property, which can be religion, race, gender or sex or political affiliation through the use of Internet and Social Networks, based on a power imbalance, which can be carried out repeatedly, systematically and uncontrollably, through digital media and often motivated by ideologies.

1. Introduction

Cyberspace offers freedom of communication and opinion expressions. However, the current social media is regularly being misused to spread violent messages, comments, and hateful speech. This has been conceptualized as online hate speech, defined as any communication that disparages a person or a group on the basis of characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or political affiliation (Zhang & Luo, 2018).

The urgency of this matter has been increasingly recognized (Gambäck & Sikdar, 2017). In the European Union (EU), 80% of people have encountered hate speech online and 40% have felt attacked or threatened via Social Network Sites [SNS] (Gagliardone, Gal, Alves, & Martinez, 2015).

Among its main consequences we found harm against social groups by creating an environment of prejudice and intolerance, fostering discrimination and hostility, and in severe cases facilitating violent acts (Gagliardone et al., 2015); impoliteness, pejorative terms, vulgarity, or sarcasm (Papacharissi, 2004); incivility, that includes behaviors that threaten democracy, deny people their personal freedoms, or stereotype social groups (Papacharissi, 2004); and off line hate speech expressed as direct aggressions (Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014; Coe, Kenski, & Rains, 2014), against political ideologies, religious groups or ethnic minorities. For example, racial- and ethnic-centered rumors can lead to ethnic violence and offended individuals might be threatened because of their group identities (Bhavnani, Findley, & Kuklinski, 2009).

A concept that can explain online hate speech is social deviance. This term encompasses all behaviors, from minor norm-violating to law-breaking acts against others, and considers online hate as an act of deviant communication as it violates shared cultural standards, rules, or norms of social interaction in social group contexts (Henry, 2009).

Among the norm violating behaviors we can identify: defamation (Coe et al., 2014), call for violence (Hanzelka & Schmidt, 2017), agitation by provoking statements debating political or social issues displaying discriminatory views (Bhavnani et al., 2009), rumors and conspiracy (Sunstein & Vermeule, 2009).

These issues make the investigation of online hate speech an important area of research. In fact, there are many theoretical gaps on the explanation of this behavior and there are not enough empirical data

* Corresponding author.
E-mail address: scastanopul@uniminuto.edu.co (S.A. Castaño-Pulgarín).

https://doi.org/10.1016/j.avb.2021.101608
Received 14 July 2020; Received in revised form 26 January 2021; Accepted 23 March 2021
Available online 6 April 2021

to understand this phenomenon and its relation with the use of Internet. Based on this need, the main contribution of this systematic review is the understanding of how the Internet and SNS use is related to online hate speech, based on different studies published in indexed databases and gray literature.

2. Methods

This systematic review is based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA] (Moher, Liberati, Tetzlaff, & Altman, 2009). We included studies that tackled any relationship between Internet, SNS and Online Hate Speech involvement.

A search strategy was designed combining the following keywords with Boolean operators as AND, OR, NOT: "online hate, cyber hate, hate material, Internet and online hatred, online victimization, cyber racism, online hate speech and social media hate". The search strategy is not complete due the need for brevity, however, it may be requested to the authors. The data bases were Scopus, PUBMED, PsycArticles and Science Direct. The search was made during August and September 2019. Gray literature, sought in the references of the selected papers, was also considered to mitigate the publication bias.

2.1. Inclusion and exclusion criteria

The following inclusion criteria were employed to identify eligible studies: 1) any paper that addresses the relationship between Internet use or social networks and online hate speech or cyberhate, except annotations, conferences, narrative or systematic reviews, letters and comments; 2) published between January of 2015 and September 2019 given the fast changes in the last five years in the development of Web 2.0 applications that are remarkable even set against the pace of change since the advent of the Internet; 3) focused on any specific type of online hate speech 4) written in English or Spanish, not discriminating by geographical area; 5) there were no restrictions for cross-sectional or longitudinal designs. Articles that assessed offline hate speech or that addressed related terms as cybercrime, cyberterrorism, populism on cyberspace, political propaganda, web nationalism, cyberbullying or sexting were excluded.

2.2. Procedure

Two independent researchers reviewed the papers for compliance with the inclusion criteria. Papers were firstly reviewed through the title and summary to determine whether they met these parameters. Those that accomplished these first screening, were downloaded and systematized through the following variables: authors, country where the study was conducted, characteristics of the sample (people, messages, tweets, comments, etc.), methods for measuring online hate speech (empirical studies, content analysis, etc.), type of online hate (political-ideological, gendered-sexual, racial-ethnic or religious) and most relevant results. This information was gathered in an excel file and then analyzed through Atlas Ti version 8 program for codification, categorization and theorization to identify patterns so that the big quantity of heterogeneous information was more manageable.

The information was analyzed through the methodological orientations of the Grounded Theory (Strauss & Corbin, 1997), with an inductive analysis, as no prior categories (or themes) were pre-settled before the analysis in order to achieve a summarized explanation of the phenomenon by using emerging categories and creating conceptual models (Kaján, 2017), where theory is inductively derived from data, providing evidence for the conclusions using systematic methods of data collection and analysis that inform each other.

The final product of this review was a narrative (qualitative) analysis of findings (Rodgers et al., 2009). A statistical meta-analysis of data was not possible due to the nature of the research designs whose methods ranged from quantitative approaches with big data, studies carried out with automatic word detection software in news comments or social networks to qualitative case studies on specific manifestations of hatred related to certain terrorism events. This made the methods and data very heterogenous. In this variety, it was not possible to gather quantitative information such as: gender, sample size, size effect, odds ratio, means, standard deviation, socio economic status, geographical locality, etc., that are essential for a statistical analysis.

2.3. Synthesis and analysis method

Although most emerging topics are related, we categorized the information into four main topics or categories that emerged from the data by similarity of themes. This approach has been used in different works (Best, Manktelow, & Taylor, 2014) and has been useful in cases where data is highly heterogeneous.

In the first category: *types of online hate speech*, we included 33 papers related to religious hate speech, online racism, gendered online hate and political hate speech. These types of cyberhate were mostly endorsed by ideologies that underlie them. Some of these ideologies are racism, Islamophobia, Alt-Right and White Nationalism, that are narrowly related between themselves to give rise to manifestations of hatred that overlap each other. For example, Islamophobia may have both religious and political motivations, not only because Islamism is a religion, but also because the immigrant status of Muslims in some European countries, promote public and political discussions about the fact that the immigration of Muslims poses a risk to the national security of a country that has been the victim of terrorist attacks in the past.

The second category was *Terrorism as an online hate trigger* with five cases. In this category we grouped different cases of terrorist attacks and their relationship with the social media hate reactions. The third category *Online Hate Expressions* shows the different ways hate is expressed on line and is composed by four papers. Finally, the *Most Used Methods to Assess Online Hate*, conveys how research has identified cyberhate. This last category is composed by 25 cases.

3. Results

2389 articles were downloaded, in addition to 40 investigations extracted from gray literature. Of these, 1482 among duplicates and those dealing with hate content, general hostility, or hostility narratives unrelated to online activity were discarded, with a total of 947 remaining who underwent a review of titles and abstracts. From this review, 67 met the inclusion criteria for analysis. Fig. 1 shows the flow chart of the systematic review process.

3.1. Types of online hate

3.1.1. Online religious hate speech

This type of hate speech is defined as the use of inflammatory and sectarian language to promote hatred and violence against people on the basis of religious affiliation through the cyberspace (Albadi, Kurdi, & Mishra, 2018; Răileanu, 2016).

According to the results, the most attacked religion in world is Islam and it seems to be motivated by an Islamophobia sentiment, favored by the cultural processes of globalization and digital media circulation (Horsti, 2017). Anti-Islam frames are expressed along a wide spectrum of discursive strategies, where people justify opposition to Islam based on Muslims actions of terrorism in different countries (Froio, 2018). One example of this was the Czech initiative in 2016, against Islam characterized by hateful online comments. The targets of these comments were immigrants and refugees, Muslims in general, governments, political elites or people who were in favor of them (Hanzelka & Schmidt, 2017). Far from being harmless, these hateful comments have largely aggravated by offline anti-Islam discourses and Islamophobia, involving narratives that frame Muslims as violent and unable to adapt to Western values (Evolvi, 2019).

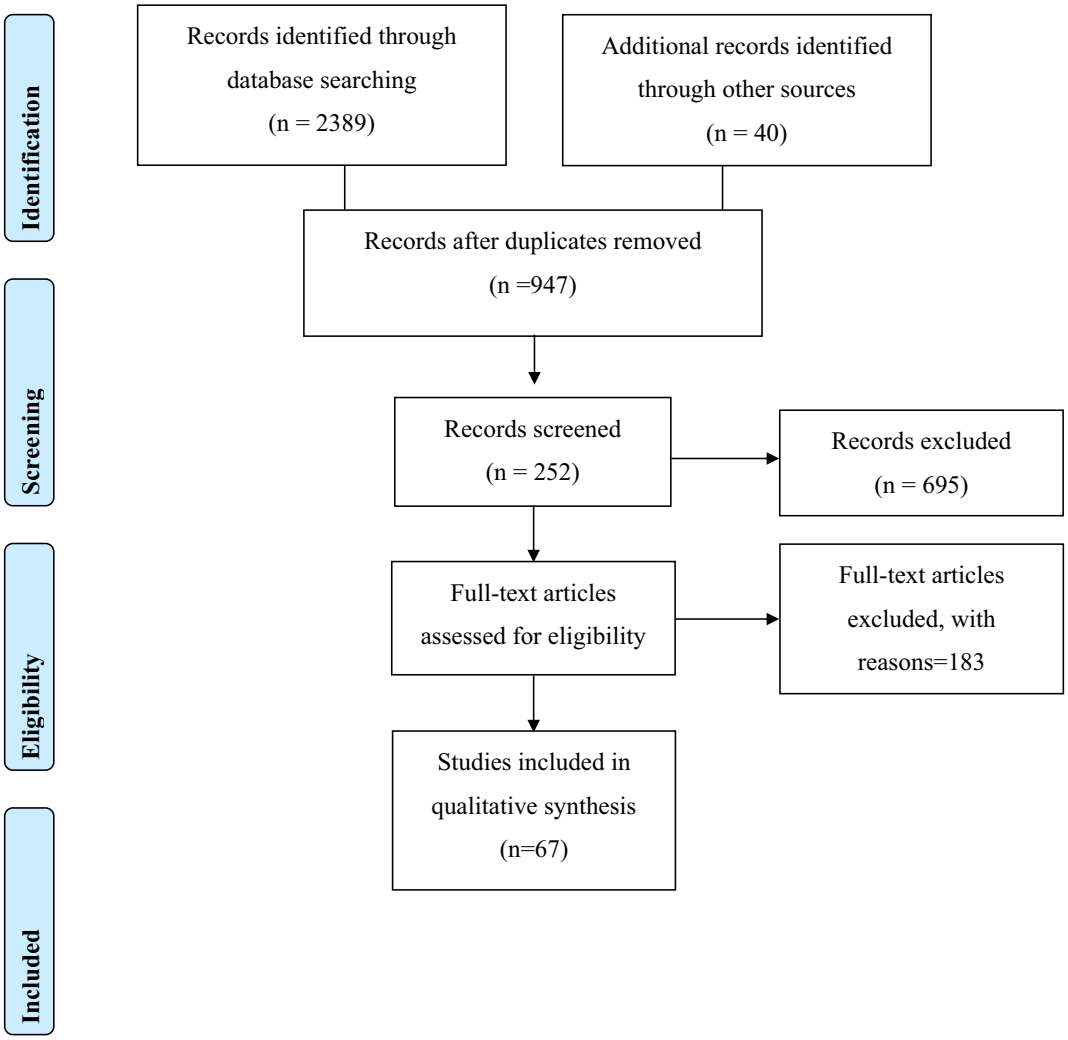

Fig. 1. PRISMA flow diagram.[1]
[1]Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:https://doi.org/10.1371/journal.pmed1000097.

3.1.2. Online racism

Racism seems to be amplified within social media environments. The anonymity and greater accessibility of the Internet, gave platform, identity protection for expressions and online racist attitudes. The analysis of 51,991 public comments posted to 119 news stories in Facebook, about race, racism or ethnicity on the Canadian Broadcasting Corporation News Facebook page, showed the dissemination of hate against indigenous and black people (Chaudhry & Gruzd, 2020), perpetuating dominant discourses on white identities (Ben-David & Matamoros-Fernández, 2016). This kind of representations reflects a racially segregated online traffic pattern.

On the other hand, racist and antiracist groups tend to use words reflecting perceived injustice and moral foundations focused on religion to spread online hate messages. Racist groups focus more on purity, respect for authority, and religion, and less on fairness than anti-racist groups. Racist groups also used less cognitively complex language than non-activist groups (Faulkner & Bliuc, 2018). The KuKuxKlan (KKK), for instance, uses the Internet to rationalize its beliefs by denying its implications or to contextualize it in higher order responsibilities to legitimize deviant behaviors (Cohen, Holt, Chermak, & Freilich, 2018).

3.1.3. Political online hate

Different democratic and political mechanisms may lead to amplify animosity and intolerance against others through SNS. The referendum, for example, rather than generate considerable democratic dialogue around policy alternatives and rationales, are often focused on changing the others arguments around a subset of frequently not related issues to what is being considered into popular vote (Chen, 2019; Siegel et al., 2018).

In Great Britain, Tweets in the aftermath of the 2016 British referendum on European Union membership, also known as “Brexit”, were followed by a surge of Islamophobic episodes. These anti-Islamic sentiments were related to religion, ethnicity, politics, and gender, promoting symbolic violence rather than engaging in constructive conflict (Evolvi, 2019).

Other case was the unexpected NO to the peace agreement between the Colombian government and the Colombian Revolutionary Armed Forces - Fuerzas Armadas Revolucionarias de Colombia (FARC), rebel group that carried out war actions against the State for more than 50 years. The result of the process was influenced by a rhetorical based on false assumptions through SNS, about the content of the agreement, opponents’ political actions, religion, ethnic groups, and justice (Pulido, 2019). On the other hand, the announcement to the candidacy for the presidency of Colombia, of Rodrigo Londoño, alias Timochenko, former head of this rebel group, led to expressions of hatred in the Colombian digital environment but also in his public appearances, forcing the candidate to quit the presidential race (Tabares Higuita, 2018).

Presidential campaigns can also elicit online hate. In the United

States, the Donald Trump presidential race could have risen a particular appeal to individuals drawn to hateful ideologies. This has been observed in anti-Muslims hashtags disseminated mostly by clusters of self-defined conservative actors based in the US (Sainudiin, Yogeeswaran, Nash, & Sahioun, 2019), with expressions related to a racialized, anti-immigration and a white nationalist narrative (Constantinou, 2018; Poole, Giraud, & de Quincey, 2019).

The connection between these conservative and anti-immigrant narratives may be observed in the social networks' endorsement messages to the Trump's executive order in January 2017, to ban seven country immigrants from entering the U.S. This support was accompanied by online comments focused on security, demeaning Muslims, and exclusion (Bresnahan, Chen, & Fedewa, 2018).

But this is not the only way that politics may involve hate. Some politicians convicted of online and offline hate-speech against certain groups as Muslims in Europe, consider hate-speech as an action of trivial mishaps or an act of virtue (Ben-David & Matamoros-Fernández, 2016).

In conclusion, some political discursive patterns may involve the proliferation of racist, ethnic, religious and/or gender stereotypes (Meza, Vincze, & MOGOŞ, 2018).

3.1.4. Gendered online hate

While most cases of online hate speech targets individuals on the basis of ethnicity and nationality, incitements to hatred on the basis of gender and sexual orientation are increasing, as digital media may exacerbate existing patterns of gendered violence and introduce new modes of abuse (Dragiewicz et al., 2018).

The 'Italian Hate Map' project, analyzed 2,659,879 Tweets where women were the most insulted group, having received 71,006 hateful Tweets (60.4%), followed by gay and lesbian persons (12,140 tweets, 10.3%) (KhosraviNik & Esposito, 2018).

In other countries, 73.4% women who blog about politics or identify as feminist have suffered negative experiences online. Most of these negative experiences involved not only abusive comments but also stalking, trolls, rape threats, death threats, unpleasant offline encounters, intimidation, shaming, and discrediting, extreme hostility in the form of digital sexism in discussion rooms, comment sections, gaming communities, and on social media platforms (Sobieraj, 2018).

3.2. Terrorism as an online hate trigger

In this second category, it was found that terrorism events are frequently related to observable public social media reactions (Miro-Llinares & Rodriguez-Sala, 2016; Williams, Burnap, & Sloan, 2017). For instance, #StopIslam hashtag was used to spread racialized hate speech and disinformation directed towards Islam and Muslims, trended on Twitter after the March 2016 terrorist attacks in Brussels (Poole et al., 2019; Urniaz, 2016). In United Kingdom, 200,880 hate tweets against Muslims, were identified following the June 2017 London Bridge terrorist attack (Miró-Llinares, Moneva, & Esteve, 2018), and in France, on the aftermath of the 2015 terrorist attacks (Froio, 2018).

Other actions like the Rotherham scandal in the United Kingdom (UK) were sexual abuse of 1400 children from 1997 to 2013 was made by men of Pakistani origin; the beheading of journalists, James Foley, Steven Sotloff and the humanitarian worker David Haines and Alan Henning by the jihadists group operating in Syria and Iraq (known as ISIS); the Trojan Horse scandal over allegations of an Islamist conspiracy in several schools in Birmingham, England, and the Woolwich attacks in which a British army soldier was killed by two black British men of Nigerian origin who had converted to Islam in 2013, led to the rise of anti-Muslim hate on SNS like Facebook. Overall, these expressions found Muslims being demonized online through negative attitudes, discrimination, stereotypes, physical threats and online harassment which all had the potential to incite violence or prejudicial actions because it disparages and intimidates a protected individual or group. Overall, these expressions found Muslims being demonized online through negative attitudes, discrimination, stereotypes, physical threats and online harassment which all had the potential to incite violence or prejudicial actions because it disparages and intimidates a protected individual or group.

3.3. Online hate expressions

This third category shows how online hate is expressed through SNS. In the case of racism, it was found the use of vicarious observation, racist humor, negative racial stereotyping, racist online media, and racist online hate groups. The online hate against women tends to use shaming (Sundén & Paasonen, 2018).

In politics motivated online hate, social media users tend to assume the role of analysts and judges to confront other political perspectives directly through the use of negative lexis and rhetorical figures to express their negative stance and exert power and dominance over others (Trajkova & Neshkovska, 2018).

Stereotypes, speculation, comparison, degrading comments, slander/defame, sedition, sarcasm, threaten, challenge, criticism, name-calling, and sexual harassments are also used in religious and ethnic online hate (Lingam & Aripin, 2017), where flaming, trolling, hostility, obscenity, high incidence of insults, aggressive lexis, suspicion, demasculinization, and dehumanization can inflict harm to individuals or organizations (Ruzaite, 2018).

3.4. Most used methods to assess online hate

One of the most widely used methods was grounded theory, this method gave researchers the ease of analyzing online discussion threads (Nanney, 2017), and identifying emerging issues to directly develop analytical frameworks (Jane, 2016; Kim, 2017; Pfafman, Carpenter, & Tang, 2015).

The discourse analysis and thematic analysis were also the most used for the qualitative analysis of cyber hate discourses (Bresnahan et al., 2018). Some investigations carried out a discursive analysis based on the perceptions of the sarcastic discursive practice (Malmqvist, 2015), in addition to the dissemination, aesthetics and text as they shape and respond to the discursive signals (Horsti, 2017; Lunstrum, 2016; Nikunen, 2018; Topinka, 2018). The thematic analysis was used to identify specific themes and fields of crime studies (Mondal, Silva, Correa, & Benevenuto, 2018; Ryan, 2018), racism detection, racial micro aggression, inter-ethnic hate incidents, female and male victims, sexual harassment, hate against women and others (Chetty & Alathur, 2019; Faulkner & Bliuc, 2018; Gillett, 2018; Jokanovic, 2018; Tynes, Lozada, Smith, & Stewar, 2018).

Furthermore, the implementation of computational methods or software such as Apache Spark, Python packages, including Tweepy, UCINET and NodeXL among others; through the search for keywords related to hate speech (Bevensee & Ross, 2018; Dias, Welikala, & Dias, 2018; Sainudiin et al., 2019); intend to analyze and create algorithms, by combining with quantitative analysis methods to obtain more reliable scales and reduce misinterpretation of the results obtained (Poole et al., 2019).

In the study of cyber hate trends related to mixed analysis methods emerged, where most of them combined a quantitative analysis method in order to give a modeling of topics and filters using two specialized dictionaries that contained multiple forms of the terms used to refer to the objectives of hate speech (Merrill & Åkerlund, 2018), with qualitative analysis methods, and in-depth interviews, surveys, participatory observations, thematic analysis and grounded theory (Chen, 2019; Eckert, 2018), where they managed to give an index of high reliability to their investigations (Miro-Llinares & Rodriguez-Sala, 2016).

In sum, most of the methods used to assess online hate are based on qualitative approaches: discourse analysis, grounded theory, etc. It is necessary to develop empirical methods to assess cyberhate, especially among adolescents as the most frequent users of the Internet and the

SNS.

4. Discussion

Online hate speech lacks unique, discriminative features and therefore is hard to identify and define (Zhang & Luo, 2018). Among these difficulties are subtleties in language, differing definitions on what constitutes hate speech, and limitations of data availability for identification and prevention (MacAvaney et al., 2019).

In general, on line hate speech refers to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender, their ethnic group or race, their believes and religion or their political preferences (Watanabe, Bouazizi, & Ohtsuki, 2018).

While encouraging freedom of expression, SNS also imply freedom to hate to the extent that individuals exercise their right to voice their opinions while actively silencing others. The identification of the potential targets of hateful or antagonistic speech is key to distinguishing the online hate from arguments that represent political viewpoints protected by freedom of expression rights (Meza et al., 2018).

Online hate is not a harmless matter. Online extremist narratives have been linked to abhorrent real-world events, including hate crimes, mass shootings such as the 2019 attack in Christchurch, stabbings and bombings; recruitment of extremists, including entrapment and sex-trafficking of girls; threats against public figures, including the 2019 verbal attack against an anti-Brexit politician, and renewed anti-western hate in the 2019 post-ISIS landscape associated with support for Osama Bin Laden's son and Al Qaeda. Social media platforms seem to be losing the battle against online hate (Johnson et al., 2019).

Among the psychological explanations, one only study in this review was found to mention how supporters of racism use moral disengagement strategies that allow them to avoid self- and social sanctions for supporting racist activity (Faulkner & Bliuc, 2016). However, the scientific literature in general, allows recognizing several psychological theories to explain the phenomenon of cyberhate: a) Erik Erikson's stages of psychological development: industry vs. inferiority (approximately from 5 to 12 years old), where self-esteem and social skills gain value. On the other hand, the stage of identity vs. role confusion (from 12 to 18 years old), in which adolescents acquires identity in different domains such as the choice of a career, sexual orientation, ethnic racial identity, political affiliations, beliefs and values (social identity); b) Bandura's social cognitive theory, which assumes that human beings learn by imitating social models, which have spread exponentially through social networks and in general in digital media, c) Theory of attribution bias, part from the theory of social information processing, which recognizes that people adopt a cognitive scheme (which becomes a model) to process new information and that can lead to accept as valid information with hostile or discriminatory content; d) the theory of social identity: which explains how social groups improve their cohesion or status, reaffirming their uniqueness by degrading other social groups; and recently, e) the social ecological theory, which gives value to the cultural and interaction context of the person with consecutive levels of development that impact their identity and particularly their decision-making regarding how to interact with others (Bauman, Perry, & Wachs, 2020).

5. Conclusion

There seems to be a current consensus on the definition of cyberhate. The Anti-Defamation League (2016) defined cyberhate as any representation of ideas that promote hatred, discrimination or violence against any individual or group of people, based on aspects such as race, color, ethnic origin, nationality or ethnicity, and religion through digital media. Likewise, Hawdon, Oksanen, and Räsänen (2017) and Wachs and Wright (2019), consider cyberhate as a behavior of denigration, threat, exclusion and provocation that may incite violence through digital media.

Similarly, Watanabe et al. (2018) define cyberhate as the use of violent, aggressive or offensive language, focused on a specific group of people who share a common property, which can be religion, race, gender or sex or political affiliation through the use of Internet and Social Networks, based on a power imbalance, carried out systematically and uncontrollably, through digital media and often motivated by ideologies to which individuals and groups adhere, deriving in behaviors that can be considered as acts of deviant communication as they may violate shared cultural standards, rules or norms of social interaction in group contexts.

As it has been stated in previous lines, cyberhate in general, seems to be amplified by the use of the Internet and social networks, resulting in the proliferation of stereotypes and worse damage. The relevance of this lies in its real effects: hate crimes, offline aggressions, discrimination, racist attitudes, democratic consequences, exacerbation of gendered violence, among others, which affect coexistence and mental health of victims, bystanders or perpetrators.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors do not have any conflict of interests to disclose.

References

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. https://doi.org/10.1109/asonam.2018.8508247

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect": Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication, 19*, 373–387. https://doi.org/10.1111/jcc4.12009

Anti-Defamation League. (2016). *Responding to Cyberhate: Progress and trends*. New York: ADL.

Bauman, S., Perry, V. M., & Wachs, S. (2020). The rising threat of cyberhate for young people around the globe. *Child and Adolescent Online Risk Exposure*, 149–175.

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication, 10*, 1167–1193.

Best, P., Manktelow, R., & Taylor, B. (2014). Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review, 41*, 27–36.

Bevensee, E., & Ross, A. (2018). *The alt-right and global information warfare by Emmi Bevensee and Alexander Ross* (pp. 4393–4440). https://doi.org/10.1109/BigData.2018.8622270

Bhavnani, R., Findley, M. G., & Kuklinski, J. H. (2009). Rumor dynamics in ethnic violence. *Journal of Politics, 71*, 876–892. https://doi.org/10.1017/S002238160909077X

Bresnahan, M., Chen, Y., & Fedewa, K. (2018). Extinguishing Lady Liberty's torch? Online public responses to the U.S. executive order to ban immigrants from 7 countries. *Journal of Intercultural Communication Research, 47*(6), 564–580. https://doi.org/10.1080/17475759.2018.1520737

Chaudhry, I., & Gruzd, A. (2020). Expressing and challenging racist discourse on Facebook: How social media weaken the "spiral of silence" theory. *Policy & Internet, 12*(1), 88–108.

Chen, P. J. (2019). Civic discourse on Facebook during the Australian same-sex marriage postal plebiscite. *Australian Journal of Social Issues, 54*(3), 285–304. https://doi.org/10.1002/ajs4.74

Chetty, N., & Alathur, S. (2019). Racism and social media: A study in Indian context. *International Journal of Web Based Communities, 15*(1), 44–61. https://doi.org/10.1504/IJWBC.2019.098692

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*, 658–679. https://doi.org/10.1111/jcom.12104

Cohen, S. J., Holt, T. J., Chermak, S. M., & Freilich, J. D. (2018). Invisible empire of hate: Gender differences in the Ku Klux Klan's online justifications for violence. *Violence and Gender, 5*(4), 209–225. https://doi.org/10.1089/vio.2017.0072

Constantinou, M. (2018). Resisting Europe, setting Greece free: Facebook political discussions over the Greek referendum of the 5th July 2015. *Lodz Papers in Pragmatics, 14*(2), 273–307.

Dias, D. S., Welikala, M. D., & Dias, N. G. J. (2018). Identifying racist social media comments in Sinhala language using text analytics models with machine learning. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*. https://doi.org/10.1109/icter.2018.8615492

Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies, 18*(4), 609–625. https://doi.org/10.1080/14680777.2018.1447341

Eckert, S. (2018). Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States. *New Media & Society, 20*(4), 1282–1302. https://doi.org/10.1177/1461444816688457

Evolvi, G. (2019). # Islamexit: Inter-group antagonism on Twitter. *Information, Communication & Society, 22*(3), 386–401.

Faulkner, N., & Bliuc, A. M. (2016). "It"s okay to be racist': Moral disengagement in online discussions of racist incidents in Australia. *Ethnic and Racial Studies, 39*(14), 2545–2563. https://doi.org/10.1080/01419870.2016.1171370

Faulkner, N., & Bliuc, A. M. (2018). Breaking down the language of online racism: A comparison of the psychological dimensions of communication in racist, anti-racist, and non-activist groups. *Analyses of Social Issues and Public Policy, 18*(1), 307–322. https://doi.org/10.1111/asap.12159

Froio, C. (2018). Race, religion, or culture? Framing Islam between racism and neo-racism in the online network of the French far right. *Perspectives on Politics, 16*(3), 696–709. https://doi.org/10.1017/S1537592718001573

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. In *Series on internet freedom, pages 1–73*. Unesco: Publishing.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85–90).

Gillett, R. (2018). Intimate intrusions online: Studying the normalisation of abuse in dating apps. *Women's Studies International Forum, 69*, 212–219.

Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology, 11*, 143–160. https://doi.org/10.5281/zenodo.495778

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior, 38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Henry, S. (2009). *Social deviance*. Cambridge, UK: Polity Press.

Horsti, K. (2017). Digital Islamophobia: The Swedish woman as a figure of pure and dangerous whiteness. *New Media & Society, 19*(9), 1440–1457. https://doi.org/10.1177/1461444816642169

Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum, 30*(3), 284–297.

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., … Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature, 573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7

Jokanovic, J. (2018). Hate crime victims in Serbia: A case study of context and social perceptions. *International Journal for Crime, Justice and Social Democracy, 2*(7), 21–37. https://doi.org/10.5204/ijcjsd.v7i2.518

Kaján, E. (2017). Hate online: Anti-immigration rhetoric in Darknet. *Nordia Geographical Publications, 46*(3), 3–22.

KhosraviNik, M., & Esposito, E. (2018). Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics, 14*(1), 45–68. https://doi.org/10.1515/lpp-2018-0003

Kim, J. (2017). #iamafeminist as the "mother tag". *Feminist identification and activism against misogyny on Twitter in South Korea, Feminist Media Studies., 17*(5), 804–820. https://doi.org/10.1080/14680777.2017.1283343

Lingam, R. A., & Aripin, N. (2017). Comments on fire! Classifying flaming comments on YouTube videos in Malaysia. *Jurnal Komunikasi: Malaysian Journal of Communication, 33*(4). https://doi.org/10.17576/JKMJC-2017-3304-07

Lunstrum, E. (2016). Feed them to the lions: Conservation violence goes online. Geoforum. 79, 134–143. doi:DOI: https://doi.org/10.1016/j.geoforum.2016.04.009.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS One, 14*(8), Article e0221152. https://doi.org/10.1371/journal.pone.0221152

Malmqvist, K. (2015). Satire, racist humor and the power of (un)laughter: On the restrained nature of Swedish online racist discourse targeting EU-migrants begging for money. *Discourse & Society.*. https://doi.org/10.1177/0957926515611792

Merrill, S., & Åkerlund, M. (2018). Standing up for Sweden? The racist discourses, architectures and affordances of an anti-immigration Facebook group. *Journal of Computer-Mediated Communication, 23*(6), 332–353. https://doi.org/10.1093/jcmc/zmy018

Meza, R., Vincze, H. O., & MOGOȘ, A. (2018). Targets of online hate speech in context. A comparative digital social science analysis of comments on public Facebook pages from Romania and Hungary. *East European Journal of Society and Politics, 4*(4), 26–50. https://doi.org/10.17356/ieejsp.v4i4.503

Miro-Llinares, F., & Rodriguez-Sala, J. J. (2016). Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design & Nature and Ecodynamics, 11*(3), 406–415. https://doi.org/10.1186/s40163-018-0089-1

Miró-Llinares, F., Moneva, A., & Esteve, M. (2018). Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. *Crime Science, 7*(1), 1–12. https://doi.org/10.1186/s40163-018-0089-1

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). *Characterizing usage of explicit hate expressions in social media. 24(2)* pp. 110–130). https://doi.org/10.1080/13614568.2018.1489001

Nanney, M. (2017). *"I"'m part of the community, too': Women's college alumnae responses to transgender admittance policies. 24* pp. 133–154). ResearchGate. https://doi.org/10.1108/S1529-212620170000024009

Nikunen, K. (2018). From irony to solidarity: Affective practice and social media activism. *Studies of Transition States and Societies, 10*.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*, 259–283. https://doi.org/10.1177/1461444804041444

Pfafman, T. M., Carpenter, C. J., & Tang, Y. (2015). The politics of racism: Constructions of African immigrants in China on ChinaSMACK. *Communication, Culture & Critique, 8*(4), 540–556. https://doi.org/10.1111/cccr.12098

Poole, E. A., Giraud, E., & de Quincey, E. (2019). Contesting# StopIslam: The dynamics of a counter-narrative against right-wing populism. Open Library of Humanities, 5 (1). doi:10.16995/olh.406.

Pulido, A. (2019). Violence, voting & peace: Explaining public support for the peace referendum in Colombia. *Electoral Studies, 61*, 102067. https://doi.org/10.1016/j.electstud.2019.102067

Răileanu, R. (2016). Religion-based user generated content in online newspapers covering the collective nightclub fire. *Romanian Journal of Communication And Public Relations, 18*(2), 55–65. https://doi.org/10.21018/rjcpr.2016.2.209

Rodgers, M., Sowden, A., Petticrew, M., Arai, L., Roberts, H., Britten, N., et al. (2009). Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: Effectiveness of interventions to promote smoke alarm ownership and function. *Evaluation., 15*(1), 49–73.

Ruzaite, J. (2018). In search of hate speech in Lithuanian public discourse: A corpus-assisted analysis of online comments. *Lodz Papers in Pragmatics, 14*(1), 93–116. https://doi.org/10.1515/lpp-2018-0005

Ryan, D. (2018). European remedial coherence in the regulation of non-consensual disclosures of sexual images. *Computer law & security review, 34*(5), 1053–1076. https://doi.org/10.1016/j.clsr.2018.05.016

Sainudiin, R., Yogeeswaran, K., Nash, K., & Sahioun, R. (2019). Characterizing the Twitter network of prominent politicians and SPLC-defined hate groups in the 2016 US presidential election. *Social Network Analysis and Mining, 9*(1), 34. https://doi.org/10.1007/s13278-019-0567-9

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., & Tucker, J. A. (2018). *Measuring the prevalence of online hate speech, with an application to the 2016 US election*.

Sobieraj, S. (2018). Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society, 21*(11), 1700–1714. https://doi.org/10.1080/1369118X.2017.1348535

Strauss, A., & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Sundén, J., & Paasonen, S. (2018). Shameless hags and tolerance whores: Feminist resistance and the affective circuits of online hate. *Feminist Media Studies, 18*(4), 643–656. https://doi.org/10.1080/14680777.2018.1447427

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy, 17*, 202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

Tabares Higuita, L. X. (2018). Análisis del discurso violento y de odio en dos grupos de Facebook contra la candidatura de Rodrigo Londoño "Timochenko" a la presidencia de Colombia. *Revista científica en el ámbito de la Comunicación Aplicada, 8*(3), 157–183 (ISSN-e 2174-1859).

Topinka, R. J. (2018). Politically incorrect participatory media: Racist nationalism on r/ImGoingToHellForThis. *New Media & Society, 20*(5), 2050–2069. https://doi.org/10.1177/1461444817712516

Trajkova, Z., & Neshkovska, S. (2018). Online hate propaganda during election period: The case of Macedonia. *Lodz Papers in Pragmatics, 14*(2), 309–334. https://doi.org/10.1515/lpp-2018-0015

Tynes, B. M., Lozada, F. T., Smith, N. A., & Stewar, A. M. (2018). *From racial microaggressions to hate crimes: A model of online racism based on the lived experiences of adolescents of color* (pp. 194–212). Microaggression Theory: Influence and Implications.

Urniaz, P. (2016). Expanding the nationalist echo-chamber into the mainstream: Swedish anti-immigration activity on twitter, 2010-2013. *First Monday.*. https://doi.org/10.5210/fm.v0i0.5426

Wachs, S., & Wright, M. F. (2019). The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. *Cyberpsychology, Behavior and Social Networking, 22*, 300–306. https://doi.org/10.1089/cyber.2018.0551

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access, 6*, 13825–13835.

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology, 51*(6), 1149–1168.

Zhang, Z., & Luo, L. (2018). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. In *Semantic Web* (pp. 1–21). https://doi.org/10.3233/sw-180338