ECPS
Educational,
Cultural and
Psychological Studies
JOURNAL of

28
December 2023

Studi e Contributi di Ricerca

Studies and Research Contributions

Note di Ricerca
Research Notes

Hate Speech Recognition: The Role of Empathy and Awareness of Social Media Influence

Francesco M. Melchiori[1] - Sara Martucci[1]
Calogero Lo Destro[1] - Guido Benvenuto[2]

[1] *Università degli Studi Niccolò Cusano - Department of Psychology (Roma, Italy)*

[2] *Sapienza Università di Roma - Department of Psychology of Development and Socialization Processes (Italy)*

DOI: https://doi.org/10.7358/ecps-2023-028-melc

francesco.melchiori@unicusano.it
sara.26febbraio@gmail.com
calogero.lodestro@unicusano.it
guido.benvenuto@uniroma1.it

RICONOSCIMENTO DELL'HATE SPEECH: IL RUOLO DELL'EMPATIA E DELLA CONSAPEVOLEZZA DELL'INFLUENZA DEI SOCIAL MEDIA

Abstract

Hate speech occurs within democratic societies that embrace freedom of expression and is made tangible in the social network context. It is characterized by a specific form of discrimination based on the use of verbal expressions or other media content and, usually, directed at minority groups. Although there is a lack of consensus about a unique and shared definition of hate speech, its social and personal consequences are particularly relevant for the whole society. For these reasons, it seems of crucial importance to identify hate speech recognition antecedents. The present study aimed at analyzing the relationship between hate speech recognition and specific psychological constructs, namely, empathy and awareness of social media influence. More in details, we hypothesized the association between empathy and hate speech recognition was mediated by awareness of social media influence. Data obtained from 146 participants revealed that empathy positively predicted hate speech recognition, and such relationship was mediated by awareness. Implication of such findings are discussed.

Keywords: Awareness of consequences; Discrimination; Empathy; Hate speech; Social media influence.

1. Introduction

Episodes of verbal violence (Infante & Wigley, 2009) are frequently observed and they are significantly increasing in our society. The United Nations organization recognized them on par with sexual and physical violence, and they are considered a de facto crime by the Istanbul Convention (2011; Rossi & Capalbi, 2022).

Since the nature of free speech right may collide with others' dignity respect, an important issue is related to the necessity to balance the guarantee of free speech with the respect for human rights (Touraine, 2009). In the last decades, in fact, especially with the emergence and affirmation of social network sites (SNSs), communicative language is becoming more and more rancorous, cold and detached (Gambäck & Sikdar, 2017; Castaño-Pulgarín *et al.*, 2021), and SNSs seems to create a fertile ground for hate speech to flourish.

1.1. *Hate speech definition and recognition*

The construct of hate speech is complex and multifaceted, encompassing a range of behaviors and expressions that can be motivated by prejudice and discrimination against certain groups of people. Nevertheless, within the scientific community there is no consensus on a shared definition of the term (Mastromattei *et al.*, 2022), and an intense legal and academic debate is still open. One of the main difficulty lies in the formulation of an exhaustive statement of all the hate components, avoiding the risk of colliding with some of the basic principles of democracy, including human dignity and freedom of expression (Hornsby, 2003). Additionally, it seems complicated to incorporate and consider in hate speech definition the large number of minority groups discriminated, which may significantly differ in their characteristics and specificities.

In general, as partially anticipated, hate speech may be considered as a form of communication that expresses hostility or prejudice towards a particular group of people based on their race, ethnicity, religion, sexual orientation, gender identity, or other characteristics (Tontodimamma *et al.*, 2021). In recent years, hate speech has become a widespread and highly controversial issue (Pesce *et al.*, 2020), with serious consequences for individuals and society as a whole. Indeed, hate speech can take many forms (Pollicino, 2017), including verbal or written slurs and insults, fake news, threatening or violent language, and symbols or images used to promote hatred or discrimination. In terms of its dissemination, hate speech can be spread through various channels, including social media, traditional media and face-to-face interactions. Although hate speech is often associated with explicit expressions of prejudice (Allport, 1973), it can also be more subtle and covert. This can make it difficult to identify and address, as it is not always obvious to outsiders that the language or behavior in question is intended to be hateful or discriminatory (Fortuna & Nunes, 2018). It has also been highlighted that the intensity of different types of discriminatory comments can vary and are often mitigated and communicated in more subtle ways. Evidence suggests that direct appeals to violence are more likely to be recognized and reported, whereas less direct and more veiled verbal offences are less likely to be recognized and reported (Wilhelm *et al.*, 2020). One of the main challenges is the general belief that digital platforms facilitate the legitimization of any kind of verbal expression, and there is a common perception that there are no filters or rules in place to control messages or behavior. Despite such concerns, individuals are able to recognize hate speech, especially in order to report it to the relevant authorities. In this regard, it has been shown (Sanguinetti *et al.*, 2019) that individuals are able to distinguish discriminatory posts from neutral ones and to categorize the type of discriminatory comment within three most common categories: stereotyping, aggression and offensiveness. In this sense, a recent work has attempted to identify the emotions involved in the phenomenon of hate speech (Martins *et al.*, 2019). In particular, it has highlighted that the most critical emotions for identifying hate speech are anger, disgust, fear, and sadness, while surprise can be interpreted as a neutral emotion in hate speech recognition. Furthermore, the same research reported a high accuracy in identifying hate speech, indicating that emotionality seems to be crucial for its recognition.

1.2. *Hate speech psychological outcomes*

From a psychological perspective, hate speech can have many negative effects on the well-being of individuals and communities. In particular, it can lead to feelings of shame, anxiety and fear in those targeted by the speech, and feelings of anger and aggression in those who engage in or support it (Saha *et al.*, 2019).

When individuals choose to belong to a particular subject group, they usually do so as a function of several variables, such as shared values and goals, and individual similarities. This selection of distinctive elements is achieved through the activation of cognitive processes that are thought to be responsible for structuring how the other, the self, and the world function. Such categorization is recognized and referred to as social categorization (De Caroli, 2016) and is manifested quite innately in the individual through the simplification and ordering of the surrounding reality. In this regard, hate speech can contribute to the creation and maintenance of social hierarchies and power imbalances, leading to discrimination and violence against marginalized groups. Moreover, anyone who belongs to a social group that is the object of dislike or hatred knows and fears that by publicly expressing his or her identity, he or she risks being exposed to ridicule or insult. In such cases, the awareness of one's social identity acts as a strong disincentive to be who one is in public (Riva, 2019).

1.3. *Social media influence and awareness*

It seems important to underline that today's digital platforms characteristics facilitate the creation and dissemination of hate speech. Since they allow for rapid, effective, permanent and inexpensive dissemination of thought, they have built an open road for the publication of any kind of message, without the presence of any structure (formal or informal) capable of exercising a mediating or controlling function. Indeed, social media permits the message to be extended to a wide audience, but they also allow the acquisition and maintenance over time of that message, which remains tracked and therefore retrievable, possibly leading to a continuous harm to the victims.

For sake of completeness, it should be remarked that technology can not only contribute to hate speech spread, but it can also be useful to fight and reduce it. In this vein, IT research has focused on creating database of hate speech to recognize it through algorithms that assimilate comments on the web to those in the database, with the aim of being able to promptly eliminate such comments (e.g., Calderón *et al.*, 2021). With such tools, social networking sites are able to autonomously identify the use of inappropriate words or content in the area of hate speech. However, it seems very difficult to create a complete and constantly updated database, and for these reasons part of this work is delegated directly to social media users.

In this regard, one of the main issues lies in individual lack of knowledge about social media functioning and about their impact on society.

Indeed, although social networks are daily utilized by numerous users, many of them still do not recognize all the peculiarities of digital platforms. For instance, the research by Proferes (2017) showed that a large amount of Twitter users was unaware of the fact that posted comments could be accessed by the entire population and not just those who followed them. In terms of social media use consequences, a recent study demonstrated that the pattern of revenge thinking and the attitude of ignoring the negative effects of hate speech are powerful predictors of the future online perpetrators pattern of hate speech (Rad, 2020). Thus, according to these results if an individual is pervaded by a thinking pattern built on revenge and if ignorance regarding the negative effects of one's actions online is high, then it is more likely that individual will engage in a form of hate speech. Thus, awareness of social media use and impact on others seems to play an important role in hate speech dissemination and recognition.

1.4. *The role of empathy*

Empathy plays a prominent role within the field of research on hate speech. Among the others, scholars have focused their attention on how empathy varies and differs according to gender. Specifically, it has been reported there are gender differences on empathy itself, and such differences may affect individual perceptions of discriminatory comments. With this regard, it has been found that women scored higher than men on perceived harm of hate incitement and reported freedom of speech as less important (Cowan & Khatchadourian, 2003). In this work, perceived harm of hate speech incitement was positively associated with empathy, whereas free speech was negatively related with it.

Furthermore, it can be hypothesized that such an association between empathy and hate speech may also be influenced by perceived emotional distance, which is far increased when the context becomes virtual. In fact, a recent study showed that empathy significantly explained indirect relationships between witnessing online racism against minorities content and both individual and institutional advocacy behaviours (Keum, 2021). However, such link was significant only for empathy associated with witnessing online content about systemic racism in society (e.g., «Been informed about unfairness in financial gains for racial/ethnic minorities») and not in regard with observing racial discrimination toward minorities in online interactions (e.g., «I have seen other racial/minority users receive racist insults regarding their online profile»). One of the possible explanations, which the study itself also reports, is that online interactions, compared with real-world ones, due to the absence of relational and verbal cues contributes to lower participants' empathy.

Similarly, the study by Bilewicz and Soral (2020) found that exposure to hate speech can have serious effects at the emotional level. In particular, it was shown that frequent exposure to hate speech can lead to empathy being replaced by intergroup contempt as the dominant response to others. Thus, exposure to hate speech can become both an antecedent and a consequence of the use of derogatory language. Therefore, this research has confirmed that an increased presence of derogatory language can create a sense of normativity of this communication. This highlighted how mechanisms that could potentially be effective in preventing the spread of hate speech, such as empathy and prevailing norms, can also be undermined by hate speech itself. Indeed, it has been suggested that even among people who habitually use non-discriminatory language and who would easily support opposition to hate speech, when such language is prevalent in the environment and when political or religious leaders endorse hate speech, the sense of norms may change and hate speech may cease to be a social taboo. This process of desensitization may result in hate speech reducing people's ability to effectively recognize the offensive nature of such language.

In conclusion, despite its negative consequences, research on hate speech is still in its infancy and there is much we do not yet understand about its causes and effects. In this sense, it seems fundamental to recognize and address hate speech as a harmful and dangerous phenomenon in our society. Clearly, many intersections (Sugarman *et al.*, 2018) and dynamics of hate speech are still unknown and will certainly need to be clarified over time, not least because research in this area is still relatively scarce.

1.5. *Hypotheses of the study*

Based on the evidence reported, it was planned to test whether the recognition of discriminatory comments was influenced by the subjects' degree of empathy and by their awareness of social media impact, especially in relation to the effects of what one posts or shares online which may affect other users.

Based on these assumptions, we proposed the following hypotheses:

- H1: Empathy and recognition of discriminatory comments should be positively associated.
- H2: Awareness levels should influence the number of comments recognized as discriminatory.
- H3: The relationship between empathy and recognition should be mediated by awareness of social media influence.

2. Method

2.1. *Participants*

One hundred and forty-six participants (117 females) took part in this research on a voluntary basis, and they did not receive any compensation in exchange for their participation. Participants' mean age was 22.25 (SD = 4.80). Most participants (74.7%) had a high school degree, 17.8% had a bachelor's degree, and 7.5% had a master's degree. Participants were recruited through a snowball sampling procedure. The choice of this recruitment procedure is justified for its ability to tap into networks and communities that might otherwise remain unnoticed (Faugier & Sargeant, 1997) or when it is anticipated that individuals may be reluctant to be identified due to scabrous topics, for instance in our case where racism or stereotyping was involved. Participants were informed about the research and consented to the use of their anonymized data. The study complied with the Declaration of Helsinki.

2.2. *Procedure and materials*

The survey, firstly, consisted of several questions about socio-demographic data. This was followed by a second section concerning empathy measure and self-awareness of social media influence. Lastly, hate speech recognition was assessed.

Empathic Experience Scale. Participants completed the Italian version of the Empathic Experience Scale (Innamorati *et al.*, 2019), a 30-item self-report measure designed to tap individual differences in empathy. Specifically, respondents rated the extent to which they agree with self-descriptive statements (e.g., «Often, I'm able to understand how people feel even before they tell me»). Ratings were made on a 5-point Likert scale ranging from 1 (*Not at all true*) to 5 (*Completely true*). We computed a composite score (α = .92) by summing responses to each item.

Self-awareness. Self-awareness of social media influence was measures via a single item (i.e., «I believe that what I post on my social networks influences other users' thinking») designed ad hoc to assess participants' belief about the influence they can exert through social media use.

Hate speech recognition. A pool of 22 hate speech statements (related to gender, race and homosexual rights) was tested in a pilot study with a smaller sample (73 first year university students). Of these statements, 10 (i.e., 7 hate speech statements and 3 control statements) have been selected and used in the present work. Specifically, participants had to correctly report the degree to which they believed such statements incited hatred, using a scale from 0 (*Not at all*) to 2 (*Completely*). A composite hate speech recognition score was computed by summing responses to each of the 7 hate speech statements (α = .93).

3. Results

Descriptive statistics and correlations between variables are presented in *Table 1*. As can be seen, there is a positive and significant correlation between empathy and social media influence self-awareness and between empathy and hate speech recognition. Similarly, social media influence self-awareness was positively and significantly correlated with hate speech recognition.

Table 1. – Descriptive and correlations between variables.

	M (SD)	1	2	3
1. Empathy	98.25 (15.93)	(.92)		
2. Self-awareness	2.95 (1.25)	.241**	(–)	
3. Hate speech	9.51 (4.88)	.271**	.247**	(.93)

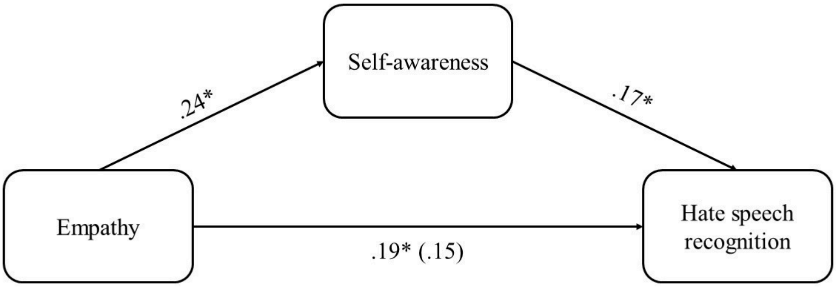
Note: **p ≤ .01; *p ≤ .05; in bracket (Cronbach's alpha); N = 146.

To test the hypothesis that social media influence self-awareness mediated the relationship between empathy and hate speech recognition, we used the PROCESS macro Model 4 (Hayes, 2013), which utilizes the bootstrapping method to extrapolate estimates of direct and indirect effects. Following the recommendation of Aiken and West (1991), predictor variables were centered. Gender (dummy coded as Male = 0 and Female = 1), age, and education were entered as control variables. A summary of the results of these analyses is reported in *Table 2*.

Table 2. – Summary of the regression models.

	Self-awareness			Hate speech		
	B	*SE*	*p*	*B*	*SE*	*p*
Gender	-.07	.27	.798	2.45	.97	**.013**
Age	-.02	.02	.341	.19	.09	**.041**
Education	.24	.19	.220	2.01	.69	**.004**
Empathy	.02	.01	**.006**	.05	.02	.058
Self-awareness				.66	.30	**.031**
R^2		.07			.21	

As can be seen, the results showed that empathy had a positive and significant effect (B = .02, SE = .01, p = .006) on social media influence self-awareness, which in turn had a significant and positive effect on hate speech recognition (B = .66, SE = .30, p = .031). Regarding the covariates considered in the model, it is noteworthy that gender positively predicted hate speech, with females reporting higher levels of hate speech recognition. Similarly, age and educational attainment had a positive effect on hate speech recognition levels, i.e. older and more educated individuals reported higher levels. The association between empathy and hate speech recognition became non-significant after controlling for social media influence self-awareness, indicating that such relationship is fully mediated (*Fig. 1*). Accordingly, the indirect effect of empathy on hate recognition through for social media influence self-awareness was significant (B = .01, BootSE = .01; bootstrapping CI = [.00, .03]).



Note: *p < .05. Standardized regression coefficients (b) are reported. In brackets, the regression coefficient of the predictor when the mediator was included in the model.

Figure 1. – Mediation model.

4. Discussion

The purpose of the investigation was to gain a better understanding of the psychological mechanisms underlying hate speech. The study hypothesized that individuals who were not aware of the phenomenon and its associated consequences were less likely to recognize hate language in selected comments, and may potentially be perceived as complicit, through indifference or sharing of such posts, or in extreme cases, even as the originators of the hate speech. Additionally, the study also took into account the degree of empathy of the subjects, as this construct had been found to be a crucial variable for understanding the phenomenon in literature.

The results of the investigation confirmed the existence of a positive and significant correlation between empathy and self-awareness of the impact of social media, and between empathy and recognition of hate speech (H1). Furthermore, the results also confirmed that self-awareness of social media influence is crucial for hate speech recognition. As stated in the second research hypothesis, the overall model was statistically significant in the empirical referents of hate speech, and the results obtained allowed the study to conclude that hate speech, particularly its recognition, was strongly influenced by the empathy present in the subjects and their awareness of the phenomenon and its consequences (H2). These findings paved the way for new proposals for intervention targeting these specific variables.

Furthermore, the investigations conducted indicated that self-awareness of social media influence played a mediating role in the relationship between empathy and hate speech recognition. Empathy was found to have a positive and significant effect on social media influence self-awareness, which in turn had a significant and positive effect on hate speech recognition. Therefore, to confirm the hypothesis (H3) that supported empathy as a predictor of recognition of discriminatory comments, the variable of awareness of the phenomenon had to be added, which would have a mediating function. Thus, it can be said that the degree of empathy is highly relevant in predicting an individual's relationship with the hate speech phenomenon, both in terms of awareness of the issue, which was found to be functional in mediating the predicted variables, and in terms of recognition and/or eventual disclosure and production.

One of the main complications identified in the literature on this topic was the lack of research that included psychological treatments. Much of the literature used focused on the creation of databases of discriminatory language and the implementation of filters to counter online hate speech. This highlighted the need for further research, as the present study investigated plausible variables with a small sample and an unvalidated instrument. On this matter, the awareness was measured by a single-item therefore usual criticism may arise about lower (or uncertain) reliability and the lack of the capacity for finer-grained assessment. Nevertheless, it is interesting to note that most research published on single-item measures shows that they are not automatically inferior to multi-item measures (Ahmad *et al.*, 2014; Ang & Eisend, 2018) and on the contrary often possess high utility and efficiency (Allen *et al.*, 2022).

Despite the difficulties encountered, the survey carried out helped to broaden the understanding of the phenomenon of hate speech by providing statistically significant observations for the sample size. The sample size and characteristics in this study did not allow for generalized inferences about the entire population, so it would be valuable to investigate hate speech in different Italian environments, analyzing the unique characteristics of these contexts. Additionally, due to the lack of a clear and universally accepted definition of hate speech, it would be beneficial for future research to establish valid indicators of hate speech, even if empathy and awareness continued to play an important role in its recognition. The main limitation of this study was the absence of well-defined indicators to measure hate speech in online posts, which in this case were selected through the Delphi method. Nevertheless, many participants were able to identify hate speech effectively, showing that certain topics were recognized as more discriminatory than others, particularly on an emotional level. Another potential limitation pertains to the reduced external validity of the study deriving from the implementation of a not random sampling method, as the snowball samples have a significant tendency to include those individuals who have numerous connections with, or are linked to, a large group of other people (Berg, 2006). As a result, unbiased estimation is not possible, and inference of findings to the general population or other settings is not immediately feasible. In conclusion, research in this field could be expanded by including other personality traits beyond empathy, or by identifying and comparing the individuals who produced discriminatory comments to those who recognized hate speech, in order to analyze the differences between these two groups and facilitate efforts to counter the phenomenon. In the future, a measurement tool for hate speech will be extremely important for research, prevention and intervention, policy and legal frameworks, social and cultural context, and personal and social well-being. In fact, it would improve the accuracy and consistency of research on hate speech and its effects, aid in the identification and tracking of hate speech, define and enforce laws and regulations, account for cultural and social context, and assess the impact of hate speech on individuals and communities.

References

Ahmad, F., Jhajj, A. K., Stewart, D. E., Burghardt, M., & Bierman, A. S. (2014). Single item measures of self-rated mental health: A scoping review. *BMC Health Services Research*, *14*(1), 1-11. http://www.biomedcentral.com/1472-6963/14/398

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment*, *38*, 1-5. doi: https://doi.org/10.1027/1015-5759/a000699.

Allport, G. W. (1973). *La natura del pregiudizio*. Firenze: La Nuova Italia.

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates thesingle-item measure approach. *Journal of Advertising Research*, *58*(2), 218-227. doi: https://doi.org/10.2501/JAR-2017-001.

Bagnato, K. (2020). Online hate speech. Responsabilità pedagogico-educative. *Annali online della Didattica e della Formazione Docente*, *12*(20), 195-211.

Berg, S. (2006). Snowball sampling – I. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (pp. 7817-7821). Hoboken, NJ: John Wiley & Sons. doi: https://doi.org/10.1002/0471667196.ess2478.pub2.

Bianchi, C. (2021). *Hate speech. Il lato oscuro del linguaggio*. Bari: Laterza.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(S1), 3-33. doi: https://doi.org/10.1111/pops.12670.

Calderón, F. H., Balani, N., Taylor, J., Peignon, M., Huang, Y.-H., & Chen, Y.-S. (2021). Linguistic patterns for code word resilient hate speech identification. *Sensors*, *21*(23), 7859. doi: https://doi.org/10.3390/s21237859.

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

Cowan, G., & Khatchadourian, D. (2003). Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly*, *27*(4), 300-308. doi: https://doi.org/10.1111/1471-6402.00110.

De Caroli, M. E. (2016). *Categorizzazione sociale e costruzione del pregiudizio. Riflessioni e ricerche sulla formazione degli attegiamenti di «genere» ed «etnia»*. Milano: FrancoAngeli.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). doi: https://doi.org/10.1609/icwsm.v12i1.15038.

Faugier, J., & Sargeant, M. (1997). Sampling hard to reach populations. *Journal of Advanced Nursing*, *26*(4), 790-797. doi: https://doi.org/10.1046/j.1365-2648.1997.00371.x.

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1-30. doi: https://doi.org/10.1145/3232676.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. *Proceedings of the First Workshop on Abusive Language Online*, 85-90. doi: https://doi.org/10.18653/v1/W17-3013.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.

Hornsby, J. (2003). Free speech and hate speech: Language and rights. In R. Egidi, M. Dell'Utri, & M. De Caro (a cura di), *Normatività Fatti Valori* (pp. 297-310). Macerata: Quodlibet.

Infante, D. A., & Wigley, C. J. (1986). Verbal aggressiveness: An interpersonal model and measure. *Communication Monographs*, *53*(1), 61-69. doi: https://doi.org/10.1080/03637758609376126.

Innamorati, M., Ebisch, S. J. H., Gallese, V., & Saggino, A. (2019). A bidimensional measure of empathy: Empathic Experience Scale. *PLoS One*, *14*(4), e0216164. doi: https://doi.org/10.1371/journal.pone.0216164.

Keum, B. T. (2021). Does witnessing racism online promote individual and institutional anti-racism advocacy among white individuals? The role of white empathy, white guilt, and white fear of other races. *Cyberpsychology, Behaviour, and Social Networking*, *24*(11), 756-761. doi: https://doi.org/10.1089/cyber.2020.0629.

Lopes, B., & Yu, H. (2017). Who do you troll and Why: An investigation into the relationship between the Dark Triad Personalities and online trolling behaviours towards popular and less popular Facebook profiles. *Computers in Human Behaviour*, *77*, 69-76. doi: https://doi.org/10.1016/j.chb.2017.08.036.

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)* (pp. 61-66). New York: IEEE. doi: https://doi.org/10.1109/BRACIS.2018.00019.

Mastromattei, M., Ranaldi, L., Fallucchi, F., & Zanzotto, F. M. (2022). Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, *8*, e859. doi: https://doi.org/10.7717/peerj-cs.859.

Pesce, F., Loi, D., Ferrari, E., & Paladino, E. (2020). *Il contributo del progetto CO.N.T.R.O. all'analisi del fenomeno dell'odio online e alla definizione di possibili soluzioni utili a contrastarlo*. Istituto per la Ricerca Sociale. [17/01/2023].
https://www.unar.it/portale/documents/20125/50788/RAPPORTO-FINALE-CONTRO_DEFINITIVO.pdf/41d5e8a2-8a00-a389-647b-1fd79f4c65f0?t=1619775596814

Pollicino, O. (2017). La prospettiva costituzionale sulla libertà di espressione nell'era di Internet. In G. Pitruzzella, O. Pollicino, & S. Quintarelli, *Parole e potere. Libertà d'espressione, hate speech e fake news*. Milano: Egea.

Proferes, N. (2017). Information flow solipsism in an exploratory study of beliefs about Twitter. *Social Media + Society*, *3*(1), 205630511769849. doi: https://doi.org/10.1177/2056305117698493.

Rad, D. (2020). Literature review for hate speech perpetuation with regards to empowerment theories: Freire's theory of empowerment or the pedagogy of the oppressed. *Journal Plus Education*, *26*(1), 403-412. doi: https://doi.org/10.24250/JPE/1/2020/DVR.

Riva, N. (2019). Il principio del danno e le espressioni d'avversione o d'odio. *Biblioteca della libertà*, *54*(224), 19. doi: https://doi.org/10.23827/BDL_2019_1_4.

Rosen, L. D., Whaling, K., Rab, S., Carrier, L. M., & Cheever, N. A. (2013). Is Facebook creating «iDisorders»? The link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety. *Computers in Human Behaviour*, *29*(3), 1243-1254. doi: https://doi.org/10.1016/j.chb.2012.11.012.

Rossi, E., & Capalbi, A. (2022). La rappresentazione mediale della violenza verbale, emotiva e psicologica nella comunicazione intima. Analisi delle matrici culturali e delle interazioni in alcuni film. *AG About Gender – Rivista internazionale di studi di genere*, *11*(21), 258-294. doi: https://doi.org/10.15167/2279-5057/AG2022.11.21.1338.

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 255-264). New York: Association for Computing Machinery. doi: https://doi.org/10.1145/3292522.3326032.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In *LREC 2018: Eleventh International Conference on Language Resources and Evaluation* (pp. 1-8). Paris: ELRA.

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, *36*(3), 357-370. doi: https://doi.org/10.1177/0165551510365390.

Sugarman, D. B., Nation, M., Yuan, N. P., Kuperminc, G. P., Hassoun Ayoub, L., & Hamby, S. (2018). Hate and violence: Addressing discrimination based on race, ethnicity, religion, sexual orientation, and gender identity. *Psychology of Violence*, *8*(6), 649-656. doi: https://doi.org/10.1037/vio0000222.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157-179. doi: https://doi.org/10.1007/s11192-020-03737-6.

Touraine, A. (2009). *Libertà, uguaglianza, diversità. Si può vivere insieme?* Milano: il Saggiatore.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research*, *47*(6), 921-944. doi: https://doi.org/10.1177/0093650219855330.

Riassunto

Il discorso d'odio si verifica all'interno delle società democratiche che abbracciano la libertà di espressione ed è reso tangibile nel contesto dei social network. Si caratterizza per una specifica forma di discriminazione basata sull'uso di espressioni verbali o altri contenuti multimediali e, di solito, diretta verso gruppi minoritari. Nonostante la mancanza di un consenso su una definizione unica e condivisa di discorso d'odio, le sue conseguenze sociali e personali sono particolarmente rilevanti per l'intera società. Per questi motivi, sembra di fondamentale importanza identificare gli antecedenti del riconoscimento del discorso d'odio. Il presente studio aveva lo scopo di analizzare la relazione tra il riconoscimento del discorso d'odio e specifici costrutti psicologici, vale a dire, l'empatia e la consapevolezza dell'influenza dei social media. Più nel dettaglio, abbiamo ipotizzato che l'associazione tra empatia e riconoscimento del discorso d'odio fosse mediata dalla consapevolezza dell'influenza dei social media. I dati ottenuti da 146 partecipanti hanno rivelato che l'empatia prediceva positivamente il riconoscimento del discorso d'odio, e tale relazione era mediata dalla consapevolezza. Le implicazioni di tali risultati sono discusse.

Parole chiave: Consapevolezza delle conseguenze; Discorso d'odio; Discriminazione; Empatia; Influenza dei social media.

How to cite this paper: Melchiori, F. M., Martucci, S., Lo Destro, C., & Benvenuto, G. (2023). Hate speech recognition: The role of empathy and awareness of social media influence [Riconoscimento dell'hate speech: il ruolo dell'empatia e della consapevolezza dell'influenza dei social media]. *Journal of Educational, Cultural and Psychological Studies (ECPS)*, *28*, 101-115. https://doi.org/10.7358/ecps-2023-028-melc