ISSN: 2038-3282

Pubblicato il: ottobre 2024

Social media and Artificial Intelligence in education: Promoting responsible use of technology to counter Hate Speech

Social media e Intelligenza Artificiale nell'educazione: Promuovere l'uso responsabile delle tecnologie per contrastare l'*Hate Speech*

di
Marilena di Padova
marilena.dipadova@unifg.it
Università di Foggia

Abstract:
Social and new digital media are a fundamental part of students' daily lives, shaping social interactions and perceptions of reality. Inappropriate use of these tools can lead to phenomena such as exposure to inappropriate content, hate speech or flame wars. Young protagonists are often left alone in the face of complex dynamics that are difficult to manage. However, it is essential to integrate the education of the new generations in the responsible and ethical use of technologies starting from school education. Through an analysis of the literature, the article aims to try to definea framework for teachers to address the challenges of digital platforms and promote a culture of respect and responsibility online, while exploiting the potential of Artificial Intelligence (AI) to detect and prevent hate speech

Keywords: social media, Artificial Intelligence, AI, hate speech, education.

Abstract:
I Social media e i nuovi media digitali rappresentano una parte fondamentale della vita quotidiana degli studenti, caratterizzando le interazioni sociali e la percezione della realtà. Un uso improprio di questi strumenti può causare fenomeni come l'esposizione a contenuti inadeguati, l'*hate speech* o le *flame wars*. I giovani protagonisti spesso sono lasciati soli di fronte a dinamiche complesse di difficile gestione. Risulta, pertanto, essenziale integrare l'educazione delle nuove generazioni ad un uso responsabile ed etico delle tecnologie a partire dalla formazione scolastica. Mediante un'analisidella letteratura, l'articolo intende provare a definire un framework destinato agli insegnanti per affrontare le sfide delle piattaforme digitali e promuovere una cultura del rispetto e della responsabilità online, sfruttando al contempo le potenzialità dell'Intelligenza Artificiale (IA) per rilevare e prevenire i discorsi d'odio.

Parole chiave: social media, Intelligenza Artificiale, IA, hate speech, educazione.

Introduction

The advent of social media and new digital media has profoundly transformed the way young people interact with their world. Students find themselves living in an environment characterized by the constant presence of connected devices and social platforms, which affect not only interpersonal relationships but also reality perception and identity development (Guinta & John, 2018). An interpsychic conditioning and at the same time intrapsychic in a surrounding that finds in digital technologies, an essential component of daily life, which by stripping itself of simple tool, is redefining the concept of sociality and the way information is accessed (Michikyan & Suárez-Orozco, 2016).

However, while social media offers many opportunities for learning and connecting, it can also expose young people to significant risks (Purba et al., 2023). Like any misuse of tools, this can lead to negative consequences, such as exposure to inappropriate content, the spread of hate speech and participation in flame wars (Douglas et al., 2023). These phenomena, if not properly managed, can have a profound impact on the psychological and social well-being of students (Senekal et al., 2023).

In this context, it becomes essential to promote new forms of literacy aimed at the younger generations for an understanding first, and then a conscious and responsible use, of digital technologies. Schools, as educational institutions, must take an active role in digital education, providing students with the tools they need to navigate safely and productively in digital environments (Purwanto, Fahmi, & Cahyono, 2023).

This article aims to make a qualitative analysis of the literature to identify the challenges posed by social media and new digital media, and suggest, through the proposal of a framework for teachers, Effective educational strategies that can be adopted by teachers to promote a culture of respect and responsibility online.

1. The risks of digital interaction among young people

The use of social media among young people is constantly growing, and with it also increase the associated risks. Recent studies (Yi & Zubiaga, 2023) have shown that a significant proportion of students have experienced or participated in cyberbullying, a phenomenon that can have serious

repercussions on their mental health and psychological well-being. Cyberbullying, unlike traditional bullying, takes place in a virtual context that makes it difficult for victims to escape attacks, as they can happen at any time and reach a wide audience (Zubair, Zubair & Ahmed, 2023).

Furthermore, the anonymous and de-contextualised nature of online interactions can facilitate the escalation of conflicts and the spread of hate speech. Digital platforms are often designed to maximise user engagement, which can encourage polarizing and controversial behaviour. In this environment, characterized by a high level of disinhibition, it can lead young people to express themselves more aggressively and/or to participate in hostile discussions without considering the consequences of their actions (Gracia-Calandínn & Suárez-Montoya, 2023).

In this sense, flame wars and hate speech not only compromise the quality of online interactions, but can also have a significant and lasting impact on the social and emotional development of young people. Continuous exposure to negative and conflictual content can contribute to a leveling towards an inverse empathy, supported by the normalization of verbal violence and desensitization towards the feelings of others. Furthermore, active participation in these dynamics can reinforce intolerant and discriminatory attitudes, compromising the socialization process and the acquisition of ethical values (Pluta et al., 2023). A part of this study will be devoted to identifying clusters through which to direct studies and then constructing research questions to define the framework useful for teachers' professionalism.

2. The role of digital education in schools

Integrating digital education in schools is crucial to help young people develop the skills needed to manage their online presence effectively and securely. Schools can no longer simply provide traditional education and remain disconnected from the reality that graviting around students. They must necessarily broaden their approach to include issues related to digital citizenship, the ethics of social media use and understanding the dynamics of power and influence that characterize the digital world (Timotheou et al., 2023).

One central aspect of digital education is certainly the teaching of critical thinking, a tool that allows students to analyze and evaluate information more effectively, reducing the risk of falling victim to disinformation or engaging in hate speech. Critical thinking not only helps young people distinguish between reliable and unreliable sources, but also promotes greater awareness of their online actions and their possible consequences (López-González, 2023).

At the same time, it is essential to promote a culture of respect online, which encourages young people to interact with others in an empathetic and responsible way. Within this process, teachers can play a crucial role in providing positive role models and creating learning environments where mutual respect and collaboration are valued. Through specific teaching activities and methodologies, such as group discussions, collaborative projects and online situation simulations, students can develop social and digital skills that help them navigate safely and responsibly in digital environments (Prasetiyo et al., 2023). Teachers experience an increasingly complex working reality also due to the presence of different educational needs within the classrooms. The analysis of the phenomena experienced at school is useful in order to be able to build models and tools that can support everyday school work. Through this study, an attempt will be made to provide answers by means of a useful tool for school operators.

3. Artificial Intelligence as a tool for prevention and education

Artificial intelligence (AI) is the new frontier in addressing the challenges of social media. Considering the potential of advanced algorithms through intentional and conscious use, they could be used for example to automatically detect inappropriate content and problematic behaviour, as hate speech, acting in a timely manner. For example, some social platforms have already implemented AI-based systems that analyse texts to identify potential hate speech or offensive content, warning users or automatically removing such content (Bilen, 2023).

However, the use of AI in this context also raises ethical issues as there is a risk of excessive surveillance and invasive control of students' online activities, that is why it is crucial that the implementation of these technologies be guided by sound ethical principles, balancing the need for protection with respect for privacy and freedom of expression. This is why teachers need to be trained to understand the potential and limitations of AI, and to use it as a support rather than control tool (Jamal, 2023).

It is therefore essential that educational institutions recognise the importance of educating young people to use digital technologies in a responsible and conscious way, providing them with the tools they need to navigate online environments safely and ethically. The integration of artificial intelligence as an educational support represents a promising frontier, but requires careful consideration of ethical and pedagogical implications (Wang et al., 2023).

An effective tool can be the development of educational frameworks that combine the effectiveness of traditional teaching with the potential offered by new technologies, promoting a digital culture marked by respect and responsibility. For these reasons, it was decided to consider these potentials supported by the scientific literature to define a role for AI in the context of the framework to be defined below.

4. Methodology

To construct the framework, we started with a qualitative analysis of the literature through the "Scopus" and "Google Scholar" databases with the keywords "social media" AND "education" AND "hate speech" OR "flame wars" OR "cyberbullying". The number of articles found was very high on both search engines. It was decided to refine the search by selecting the first 100 articles from the year 2024 for each search engine, which had content in line with the topic under investigation. The abstracts were read in order to cluster macro areas on which to continue with a more detailed investigation. From the selected articles, we identified four macro areas that are useful for defining the model: education, school policies, communities, monitoring. This was followed by the use of Elicit, an AI-powered research assistant. We then asked Elicti four questions representing our research questions. Table of questions asked, and answers received, from which the framework was designed (Figure n.1).

Question	Elicit's answer
Can teacher and student training create a school environment that promotes respect, inclusion and mutual understanding?	*Research suggests that teacher and student training can indeed create a more inclusive and respectful school environment. Creating a warm classroom atmosphere through*

	thoughtful design, procedures, and strategies can foster cooperation, acceptance, and mutual respect among students (Bucholz & Sheffler, 2009). Educators play a crucial role in promoting inclusive classrooms by understanding social dynamics and addressing various forms of diversity (Hymel & Katz, 2019). However, teachers often perceive a need for additional training to meet the challenges of achieving success for all students in inclusive settings (González-Gil et al., 2015). Implementing universal design for learning approaches can support both academic and social inclusion, teaching students to relate to others with acceptance and respect (Hymel & Katz, 2019). Overall, these studies emphasize the importance of teacher training and classroom strategies in creating an environment that promotes respect, inclusion, and mutual understanding.
Can school policies counter hate speech?	*Research suggests that school policies can play a role in countering hate speech. A positive classroom climate and social skills development can encourage students to actively counter hate speech (Wachs et al., 2023). Arts education has potential for creating counter-narratives to challenge hate speech, though more research is needed to evaluate its effectiveness (Jääskeläinen, 2019). Universities face challenges in balancing free speech with providing a welcoming environment for all students, with many maintaining contradictory policies that both protect and proscribe hate speech (Sarabyn, 2010). To address this, Sarabyn (2010) proposes applying contract law to ensure universities deliver on their promised policies. Wachs et al. (2023) found that classroom climate and social skills (perspective-taking, prosocial behavior, and assertiveness) positively affect students' likelihood to counter hate speech. These findings suggest that schools can implement policies and programs that foster a positive environment and develop students' social skills to effectively counter hate speech.*
How important is the involvement of the school community and society in the fight against hate speech?	*Research highlights the crucial role of schools and communities in combating hate speech. Teachers are encouraged to intervene using*

	various strategies, with teaching-oriented approaches being most common (Bilz et al., 2024). Arts education and artistic expression offer potential for creating counter-narratives to challenge hate speech (Jääskeläinen, 2019). In Poland, a large-scale training program for school social workers was implemented to address hate speech prevention, involving human rights activists and media (Toczyski et al., 2022). Social norms significantly influence adolescents' engagement in hate speech, with anti-hate speech injunctive norms weakening the relationship between witnessing and perpetrating hate speech, while deviant peer pressure strengthens it (Wachs et al., 2021). These findings emphasize the importance of involving the school community and society in hate speech prevention efforts, suggesting that interventions should focus on strengthening positive social norms, enhancing teachers' self-efficacy, and utilizing creative approaches to counter hate speech effectively.
How important is continuous monitoring and evaluation to counter hate speech?	*Continuous monitoring and evaluation are crucial for effectively countering hate speech. While there is limited evidence on the effectiveness of interventions due to difficulties in attribution and long-term impact assessment (Strachan, 2014), recent advancements in evaluation methods show promise. Large Language Models have been proposed as multi-aspect, reference-free evaluators for counter narratives, demonstrating strong alignment with human judgment (Jones et al., 2024). Counter narratives are seen as an effective intervention strategy, with arts and arts education offering potential for creating diverse counter narratives to challenge simplified generalizations (Jääskeläinen, 2019). However, the positive impact of counter hate speech is not guaranteed, highlighting the need for ongoing research and evaluation (Alsagheer et al., 2022). Despite challenges, continuous monitoring and evaluation remain essential for developing and refining strategies to combat hate speech effectively.*

Figure n.1 "Questions and answers with Elicit"

Considering that the answers confirmed the research questions, we proceeded to identify the literature supporting the response of Elicit, filtering as a type of studies: review, meta-analysis, systematic review and asking to show results based on quotes.
The selected results were in Figure n.2:

Can teacher and student training create a school environment that promotes respect, inclusion and mutual understanding?	Lautenbach, F., & Heyder, A. (2019). Changing attitudes to inclusion in preservice teacher education: a systematic review. *Educational Research*, *61*(2), 231-253. (65 citations) Dickens-Smith, M. (1995). The effect of inclusion training on teacher attitude towards inclusion. (52 citations) Kurniawati, F., De Boer, A. A., Minnaert, A. E. M. G., & Mangunsong, F. (2014). Characteristics of primary teacher training programmes on inclusion: A literature focus. *Educational Research*, *56*(3), 310-326. (45 citations) Forlin, C., Kawai, N., & Higuchi, S. (2015). Educational reform in Japan towards inclusion: Are we training teachers for success?. International Journal of Inclusive Education, 19(3), 314-331. (44 citations)
Can school policies counter hate speech?	Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet*, *13*(4), 502-521. (20 citations) Gould, J. B. (2001). The precedent that wasn't: College hate speech codes and the two faces of legal compliance. Law & Society Review, 35(2), 345-392. (19 citations) Neiger, J. A., Palmer, C., Penney, S., & Gehring, D. D. (1998). Addressing hate speech and hate behaviors in codes of conduct: A model for public institutions. NASPA Journal, 35(3), 193-206. (5 citations)
How important is the involvement of the school community and society in the fight against hate speech?	Sanders, M. G. (2003). Community involvement in schools: From concept to practice. Education and urban society, 35(2), 161-180. (103 citations)

	Gagliardone, I. (2014). Mapping and analysing hate speech online. Available at SSRN 2601792. (37 citations) Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. Trauma, violence, & abuse, 24(4), 2598-2615. (30 citations)
How important is continuous monitoring and evaluation to counter hate speech?	Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, 55, 477-523. (344 citations) Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2020). HateCheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*. (201 citations) Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, 126232. (132 citations) Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. Sage Open, 10(4), 2158244020973022. (103 citations)

Figure n. 2 “Collection of evidence on Elicit”

5. The C.A.R.E. framework

From the analysis of the selected articles through research on Elicit, it was decided to identify a keyword representative of the macro area of research in response to the question posed. The initials of the four words identified were used to construct a framework identifier. The framework proposed here, named C.A.R.E, offers a structured and integrated approach to address hate speech and promote a conscious use of technologies. Through awareness, action, networking and continuous evolution, schools can become safe and inclusive places, helping to form responsible and respectful citizens.

The first component of the C.A.R.E. framework is “Consciousness", a crucial element to prevent and effectively counter hate speech in schools, articulated in two main sub-dimensions: training for teachers and education of students, both are essential to creating a school environment that promotes

respect, inclusion and mutual understanding (Ballaschk et al., 2022). The first is the first fundamental step to counter hate speech. Teachers are the main mediators between school and students, and their ability to recognise and address hate speech is essential to prevent such phenomena. Training must be thorough and continuous, allowing teachers to gain a clear and up-to-date understanding of the dynamics of hate speech, its consequences and best practices for countering it (Bilz et al., 2024). The second should include educational sessions that cover various aspects of hate speech, including its origins, forms it manifests itself in and its psychological and social consequences on students. It is crucial that teachers understand the link between hate speech and other forms of discrimination and bullying, as these phenomena are often interconnected. In addition, training should include elements of educational psychology, which help teachers to identify early signs of discomfort among students and to take appropriate action (Košir, 2017).

The second component of the C.A.R.E. framework is "Action", understood as the need to develop and implement effective school policies and proactive interventions to counter hate speech. This section is divided into two main sub-dimensions: the definition and implementation of clear school policies against hate speech, and the promotion of activities and projects that encourage positive values such as inclusion and empathy (Battista & Uva, 2024). Schools must establish clear and enforceable rules that explicitly outline what constitutes hate speech and the consequences for those who violate these rules. A well-structured school policy not only acts as a deterrent, but also provides operational guidance to manage and resolve hate speech incidents when they occur. An inclusive process, involving not only administrative staff but also teachers, students and parents. This participatory approach ensures that policies are understood and accepted by the entire school community, increasing the likelihood of their effective implementation (Bilz et al., 2024). It is important that policies include a clear and comprehensive definition of hate speech, covering all forms of expression that incite hatred, violence or discrimination based on race, religion, gender, sexual orientation, Disability or other personal characteristics, specifying the consequences for anyone who participates in or promotes hate speech, ensuring that those consequences are proportionate and applied consistently (Flick, 2020).

The third component of the C.A.R.E. framework is the "Relationship", which focuses on the importance of engaging the school community and wider society in the fight against hate speech. This section of the framework recognizes that schools cannot tackle hate speech alone, but must work in synergy with families, associations, local institutions and the community (De Leo & Emanuele, 2023). We can consider two main sub-dimensions here: community involvement and the organisation of awareness events. The former is essential to create a supportive environment that can effectively counter hate speech. School is only one part of young people’s lives, and the influence of family, peers, local institutions and associations may be just as significant, if not more so, it is therefore crucial to build a support network that connects the school with the surrounding community, creating synergies that strengthen the values of respect, inclusion and responsibility. Families play a crucial role in shaping the values and behaviour of young people, so they must work closely with the school. Research has shown that when families are actively involved in the upbringing of children, students tend to show a greater sensitivity towards social issues and a reduced propensity for aggressive or discriminatory behaviour (Bernal et al., 2011). Schools, for their part, can organize workshops and information meetings for parents, offering tools and resources to recognize the signs of hate speech

and to talk with their children about online dynamics in a constructive way: Encouraging open communication between parents and school can facilitate early intervention in cases of hate speech by providing coordinated support to the student involved.

The last component of the C.A.R.E. framework is the "Evolution", which legitimizes the importance of continuous monitoring, evaluation and adaptation of policies and activities related to the fight against hate speech. This section recognises that social and technological dynamics are constantly changing and that a flexible and continuous improvement approach is needed to maintain the effectiveness of hate speech interventions (Montero, Laforgue-Bullido & Abril-Hervás, 2022). The evolution is divided into two main sub-dimensions: monitoring and feedback, and adaptation and improvement. Continuous monitoring of policies and activities is essential to ensure that anti-hate speech initiatives are effective and respond to the needs of the school community. This process must be supported by a feedback system involving all members of the school community, including students, teachers, parents and other stakeholders, thus allowing data to be collected on the effectiveness of policies and interventions adopted, Identifying areas of strength and weakness (Ganca & Kyobe, 2022). The whole school community must therefore be actively involved in the process of adaptation and improvement, where students, teachers, parents and other stakeholders must be informed and involved at every stage of the process, From data collection to the formulation of new strategies. This involvement not only improves the quality of decisions made, but also strengthens the sense of belonging and shared responsibility in the fight against hate speech (Pukallus & Arthur, 2024). A continuous evolution-based approach ensures that schools remain resilient to new challenges and that their policies and activities against hate speech are always at the forefront. This not only reduces the incidence of hate speech, but also creates a safer and more inclusive school environment where every student feels valued and respected.

In this process, AI can act as a support to the teacher in the different phases of the framework. It can be used for:

- quickly identify hate speech in online forums or educational platforms used by students, allowing for timely intervention
- conduct training sessions on detecting hate speech online
- generate educational content with the help of chatbots
- predict classroom activity on hate speech, using AI chatbots to explain the phenomenon in a simple way
- analyse and classify content.

Conclusions

The C.A.R.E., proposed in this article, in its four dimensions- Consciousness, Action, Relationship and Evolution - presents itself as a holistic and multidimensional response to the complex challenges posed by hate speech in the educational context. In an age where the digital presence of young people is increasingly pervasive and social media profoundly influences the formation of identities and social relationships, it becomes imperative to adopt educational strategies that go beyond the mere repression of the phenomenon and focus on its prevention and education to a culture of respect and inclusion.

It is not only a set of operational guidelines but wants to be a structured reference for educational interventions that integrate awareness, action, collaboration and adaptation to effectively address hate speech in schools. Its effectiveness will depend on the ability to test and implement it consistently within schools and the willingness of the school community to embrace a cultural change that focuses on respect, inclusion and shared responsibility.

References:

Alsagheer, D., Mansourifar, H., & Shi, W. (2022). *Counter hate speech in social media: A survey*. arXiv preprint arXiv:2203.03584.

Ballaschk, C., Schulze-Reichelt, F., Wachs, S., Krause, N., Wettstein, A., Kansok-Dusche, J., ... & Schubarth, W. (2022). Ist das (schon) Hatespeech?–Eine qualitative Untersuchung zum Verständnis von Hatespeech unter pädagogischem Schulpersonal. *Zeitschrift für Bildungsforschung*, *12*(3), 579-596.

Battista, D., & Uva, G. (2024). Navigating the virtual realm of hate: Analysis of policies combating online hate speech in the Italian-European context. *Law, Technology and Humans*, *6*(1), 48-58.

Bernal, A., Urpí, C., Rivas, S., & Repáraz, R. (2011). Social values and authority in education: Collaboration between school and families. *International Journal about Parents in Education*, *5*(2).

Bilen, A. (2023). A Review: Detection of Discrimination and Hate Speech Shared on Social Media Platforms Using Artificial Intelligence Methods. *Algorithmic Discrimination and Ethical Perspective of Artificial Intelligence*, 171-181.

Bilz, L., Fischer, S. M., Kansok-Dusche, J., Wachs, S., & Wettstein, A. (2024). Teachers' intervention strategies for handling hate-speech incidents in schools. Social Psychology of Education, 1-24.

Bucholz Ed D, J. L., & Sheffler, J. L. (2009). Creating a warm and inclusive classroom environment: Planning for all children to feel welcome. *Electronic Journal for Inclusive Education*, 2(4), 4.

De Leo, A., & Emanuele, R. (2023). Define and Tackle Hate Speech: The Experience of Social Workers in Italy. *Informing Science*, (26), 115-134.

Dickens-Smith, M. (1995). The effect of inclusion training on teacher attitude towards inclusion.

Douglas, K. D., Smith, K. K., Stewart, M. W., Walker, J., Mena, L., & Zhang, L. (2023). Exploring parents' intentions to monitor and mediate adolescent social media use and implications for school nurses. *The Journal of School Nursing*, 39(3), 248-261.

Flick, C. (2020). Good practices to prevent and counter the spread of illegal hate speech online. *Language, Gender and Hate Speech*, *1*, 182.

Forlin, C., Kawai, N., & Higuchi, S. (2015). Educational reform in Japan towards inclusion: Are we training teachers for success?. *International Journal of Inclusive Education*, 19(3), 314-331.

Gagliardone, I. (2014). *Mapping and analysing hate speech online*. Available at SSRN 2601792.

Ganca, S. S., & Kyobe, M. (2022, November). The Effectiveness of School Anti-cyberbullying Policies and Their Compliance with South African Laws: A Conceptual Framework. In *International Development Informatics Association Conference* (pp. 234-248). Cham: Springer Nature Switzerland.

González Gil, F., Martín Pastor, E., Flores Robaina, N. E., Jenaro Río, C., Poy, R., & Gómez Vela, M. (2013). *Inclusión y convivencia escolar: análisis de la formación del profesorado*.

Gould, J. B. (2001). The precedent that wasn't: College hate speech codes and the two faces of legal compliance. *Law & Society Review*, 35(2), 345-392.

Gracia-Calandín, J., & Suárez-Montoya, L. (2023). The eradication of hate speech on social media: a systematic review. *Journal of Information, Communication and Ethics in Society*, 21(4), 406-421.
Guinta, M. R., & John, R. M. (2018). Social media and adolescent health. Pediatric Nursing, 44(4).
Hymel, S., & Katz, J. (2019). Designing classrooms for diversity: Fostering social inclusion. *Educational Psychologist*, 54(4), 331-339.
Jääskeläinen, T. (2020). Countering hate speech through arts and arts education: Addressing intersections and policy implications. *Policy Futures in Education*, 18(3), 344-357.
Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232.
Jamal, A. (2023). The role of Artificial Intelligence (AI) in teacher education: Opportunities & challenges. *International Journal of Research and Analytical Reviews*, 10(1), 140-146.
Jones, J., Mo, L., Fosler-Lussier, E., & Sun, H. (2024). *A Multi-Aspect Framework for Counter Narrative Evaluation using Large Language Models*. arXiv preprint arXiv:2402.11676.
Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, 24(4), 2598-2615.
Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet*, *13*(4), 502-521.
Košir, K., Ambrožič, M., Repnik, R., & Slavinec, M. (2017). *Educational Psychology for Teachers: Selected Topics*.
Kurniawati, F., De Boer, A. A., Minnaert, A. E. M. G., & Mangunsong, F. (2014). Characteristics of primary teacher training programmes on inclusion: A literature focus. *Educational Research*, *56*(3), 310-326.
Lautenbach, F., & Heyder, A. (2019). Changing attitudes to inclusion in preservice teacher education: a systematic review. *Educational Research*, *61*(2), 231-253.
López-González, H., Sosa, L., Sánchez, L., & Faure-Carvallo, A. (2023). Media and Information Literacy and Critical Thinking: A Systematic Review. *Revista Latina de Comunicación Social*, (81), 399-422.
Michikyan, M., & Suárez-Orozco, C. (2016). Adolescent media and social media use: Implications for development. *Journal of Adolescent Research*, 31(4), 411-414.
Montero, A. I., Laforgue-Bullido, N., & Abril-Hervás, D. (2022). Hate speech: a systematic review of scientific production and educational considerations. *Revista Fuentes*, *24*(2), 222-233.
Neiger, J. A., Palmer, C., Penney, S., & Gehring, D. D. (1998). Addressing hate speech and hate behaviors in codes of conduct: A model for public institutions. *NASPA Journal*, 35(3), 193-206.
Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *Sage Open*, 10(4), 2158244020973022.
Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55, 477-523.
Pluta, A., Mazurek, J., Wojciechowski, J., Wolak, T., Soral, W., & Bilewicz, M. (2023). Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports*, 13(1), 4127.

Prasetiyo, W. H., Sumardjoko, B., Muhibbin, A., Naidu, N. B. M., & Muthali'in, A. (2023). Promoting digital citizenship among student-teachers: The role of project-based learning in improving appropriate online behaviors. *Participatory Educational Research*, 10(1), 389-407.
Pukallus, S., & Arthur, C. (2024). Combating Hate Speech on Social Media: Applying Targeted Regulation, Developing Civil-Communicative Skills and Utilising Local Evidence-Based Anti-Hate Speech Interventions. *Journalism and Media*, *5*(2), 467-484.
Purba, A. K., Thomson, R. M., Henery, P. M., Pearce, A., Henderson, M., & Katikireddi, S. V. (2023). *Social media use and health risk behaviours in young people: systematic review and meta-analysis*. bmj, 383.
Purwanto, A., Fahmi, K., & Cahyono, Y. (2023). The benefits of using social media in the learning process of students in the digital literacy era and the education 4.0 era. *Journal of Information Systems and Management (JISMA)*, 2(2), 1-7.
Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2020). HateCheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.
Sanders, M. G. (2003). Community involvement in schools: From concept to practice. *Education and urban society*, 35(2), 161-180.
Sarabyn, K. (2010). Free speech at private universities. *JL & Educ*., 39, 145.
Senekal, J. S., Ruth Groenewald, G., Wolfaardt, L., Jansen, C., & Williams, K. (2023). Social media and adolescent psychosocial development: a systematic review. *South African Journal of Psychology*, 53(2), 157-171.
Strachan, A. L. (2014). Interventions to counter hate speech. *GSDRC Applied Research Services*, 23(11).
Timotheou, S., Miliou, O., Dimitriadis, Y., Sobrino, S. V., Giannoutsou, N., Cachia, R., ... & Ioannou, A. (2023). Impacts of digital technologies on education and factors influencing schools' digital capacity and transformation: A literature review. *Education and information technologies*, 28(6), 6695-6726.
Toczyski, P., Grudzień, M., & Sopyło, M. (2022). *School Social Workers in Human Rights Education Against Hate Speech in Poland*.
Wachs, S., Valido, A., Espelage, D. L., Castellanos, M., Wettstein, A., & Bilz, L. (2023). The relation of classroom climate to adolescents' countering hate speech via social skills: A positive youth development perspective. *Journal of Adolescence*, 95(6), 1127-1139.
Wachs, S., Wettstein, A., Bilz, L., Krause, N., Ballaschk, C., Kansok-Dusche, J., & Wright, M. F. (2022). Playing by the rules? An investigation of the relationship between social norms and adolescents' hate speech perpetration in schools. *Journal of interpersonal violence*, 37(21-22), NP21143-NP21164.
Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., & Pange, J. (2023). Exploring the potential impact of artificial intelligence (AI) on international students in higher education: Generative AI, chatbots, analytics, and international student success. *Applied Sciences*, 13(11), 6716.
Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, 36, 100250.
Zubair, M., Zubair, S., & Ahmed, M. (2023). Cyberbullying instilled in social media. In *Cybersecurity for Smart Cities: Practices and Challenges* (pp. 17-29). Cham: Springer International Publishing.