Medical Humanities & Medicina Narrativa
ISBN 979-12-218-1616-7
ISSN 2724-3281
DOI 10.53136/979122181616713
pp. 215-227 (dicembre 2024)

L'intelligenza artificiale come strategia per educare al rispetto e alla responsabilità

Il caso di LLaMAntino

Gabriella Calvano, Vita Barletta, Marco De Gemmis*

Riassunto: I discorsi d'odio rappresentano una minaccia crescente per la coesione sociale e il benessere individuale. I loro effetti possono essere devastanti, soprattutto per i più giovani, incidendo negativamente sulla salute mentale e sullo sviluppo sociale. Social media e tecnologie possono esacerbare questi comportamenti, ma alcune "impronte digitali", come le conversazioni testuali, possono essere sfruttate dai modelli di Intelligenza Artificiale e per rilevare questi comportamenti e prevenire la violenza.

Dopo aver esplorato l'impatto dei discorsi d'odio sui giovani, l'articolo evidenzia quanto si sta elaborando nell'ambito del Progetto XXX, finanziato dal PNRR. Utilizzando il modello linguistico italiano della famiglia LLaMAntino, i ricercatori stanno implementando uno strumento che spiega la presenza di elementi di tossicità nelle conversazioni tra adolescenti. L'obiettivo è quello di usare tale strumento anche per attivare processi educativi in virtù dei quali gli utilizzatori sono in grado di riconoscere elementi di odio nei messaggi ricevuti e, conseguentemente, a essere capaci di maggiore tolleranza e rispetto in contesti più inclusivi.

Parole chiave: educazione, intelligenza artificiale, discorsi d'odio, LLaMAntino.

Abstract: Hate speech is a growing threat to social cohesion and individual well-being. Its effects can be devastating, particularly for young people, negatively impacting their mental health and social development. Social media and technology can exacerbate these behaviours, but certain "finger-

* Università degli Studi di Bari.

prints", such as text conversations, can be used by AI models to detect and prevent violence. After exploring the impact of hate speech on young people, the article presents developments from the PNRR-funded XXX project.

Using the Italian language model from the LLaMantino family, researchers are developing a tool that explains the presence of toxic elements in conversations among young people. The aim is also to use this tool to train users to identify hateful elements in the messages they receive, thereby promoting tolerance and respect in more inclusive environments.

Key–words: education, artificial intelligence, hate speech, LLaMAntino.

1. Per introdurre: *hate speech* e fenomeni d'incitamento all'odio online

In anni recenti, il fenomeno dei discorsi d'odio sta generando una crescente preoccupazione a livello globale e, sebbene sia entrata a far parte del linguaggio quotidiano, non è ancora stata formulata una definizione di *hate speech* che possa essere considerata univocamente accettata. Anzi, «chi vuole provare a fornire una definizione […] non è nemmeno facilitato dalla pluralità dei dibattiti disciplinari – politici, filosofici, culturali, giuridici – sui temi che l'hate speech richiama, dalla libertà di espressione alla lotta contro le discriminazioni, né dalla continua evoluzione dei mezzi con cui si diffonde e dalla diversità di forme, verbali e non verbali, in cui si articola» (Faloppa, 2020, p. 25).

Secondo l'Unesco (2015; 2021), i discorsi d'odio comprendono tutte le forme di espressione che diffondono, incitano, promuovono o giustificano l'odio razziale, la xenofobia, l'antisemitismo o altre forme di risentimento basate sull'intolleranza, preferendo una comunicazione che tende a degradare un individuo o un gruppo sulla base della propria identità o della propria origine, attraverso insulti, minacce, pregiudizi (Consiglio d'Europa, 2022). Essi si configurano, quindi, come fenomeni di natura multifattoriale e complessa, che possono manifestarsi attraverso vari mezzi, dalle parole scritte e alle immagini e ai video, ai *memi*, radicati in stereotipi e pregiudizi storici e culturali che perpetuano disuguaglianze e discriminazioni e che minano i diritti umani e la coesione sociale.

Con la diffusione di internet e l'avvento dei *social media*, i discorsi d'odio hanno trovato un terreno fertile per la loro disseminazione: «il web

è realtà aumentata e, soprattutto, proprio battute e atti apparentemente meno gravi e non strutturati in ideologie permettono di "liberare" parole e idee d'odio, ne facilitano l'accettazione sociale e preparano il terreno fertile per forme più intense di odio» (Pasta, 2021, p. 87). Proprio i *social media*, infatti, consentono con facilità che alcune tipologie di messaggi possano rapidamente circolare in quanto garantiscono replicabilità, scalabilità, ricercabilità e persistenza (Faloppa, 2020, p. 124). Santerini e Pasta hanno proposto sette caratteristiche del contenuto dei messaggi che possono considerati come indicativi per rilevare l'odio online (Pasta, 2022; 2023). Il contenuto, in particolare: a) deve essere pubblico e visibile agli utenti senza limitazione; b) riguarda un target specifico, spesso minoranze e persone vulnerabili; c) si propone di insultare o generare sofferenza nella vittima di odio, a prescindere dai suoi comportamenti; d) è usato per fare del male alla vittima in modo consapevole; e) contiene espressioni di odio in forma verbale esplicita o si propone di negare l'altro in quanto persona; f) incita a una risposta violenta; g) incita altri utenti a produrre a loro volta discorsi d'odio.

Piattaforme come Facebook, Twitter, YouTube e Instagram, inoltre, consentono agli utenti di condividere contenuti in modo immediato e spesso anonimo, facilitando la circolazione di ogni forma di comunicazione violenta. L'anonimato online, infatti, spesso incoraggia comportamenti aggressivi che difficilmente si manifesterebbero nella vita reale. Questo fenomeno, noto come effetto di disinibizione online, dà evidenza di come l'assenza di conseguenze immediate e la percezione di anonimato possano portare gli individui a comportarsi in modo più aggressivo e ostile rispetto a quanto farebbero offline (Suler, 2004). Si tratta di un fenomeno particolarmente evidente nelle dinamiche dei gruppi online, dove la pressione dei pari può amplificare i comportamenti aggressivi e la diffusione di discorsi d'odio.

Anche uno studio condotto da *Data & Society* (Leanhart et al., 2016) avvalora la tesi per la quale i social media sono tra gli ambienti in cui discorsi d'odio trovano maggiore possibilità di propagarsi: circa il 70% dei giovani adulti e il 40% degli adulti intervistati, infatti, hanno dichiarato di aver subito molestie o abusi online. La natura virale di internet, cioè, fa in modo che i messaggi d'odio possano raggiungere rapidamente un vasto pubblico, aumentando la loro potenziale influenza negativa: un singolo post o un tweet possono raggiungere migliaia, se non milioni, di persone

in pochi minuti, amplificando l'impatto del messaggio violento e, conseguentemente, della stessa violenza del messaggio. Come ha evidenziato il progetto *Italian Hate Map* (Lingiardi et al., 2020), i discorsi d'odio online "colpiscono" più facilmente le donne, le minoranze e le persone più vulnerabili, con quasi sempre gravi ripercussioni sulla loro salute mentale e sul loro benessere (Amnesty International, 2018).

I giovani sono tra i principali utilizzatori delle piattaforme online e, di conseguenza, tra i soggetti potenzialmente più esposti ai discorsi d'odio.

Gli effetti psicologici e sociali della violenza online sui giovani possono essere devastanti. L'esposizione prolungata all'*hate speech*, infatti, può portare a sentimenti di insicurezza, isolamento e depressione (Hinduja & Patchin, 2018), può influenzare negativamente il loro rendimento scolastico e le loro relazioni sociali. Entrare spesso in contatto con i discorsi d'odio può contribuire a creare un clima di intolleranza e discriminazione minando i valori fondamentali e basilari di una società democratica.

Per una più efficace gestione della comunicazione online e per garantire la tutela degli utenti, negli anni sono state implementate una serie di strategie e di iniziative: dalle politiche adottate dalle piattaforme dei social media per segnalare e rimuovere contenuti offensivi alle azioni dei governi di differenti Paesi orientati a contenere e penalizzare i comportamenti abusivi online, dalle campagne di sensibilizzazione (*Take Back the Tech*[1], ad esempio) ai servizi di supporto alle vittime dell'*hate speech*.

In virtù della loro natura complessa, per poter essere opportunamente studiati, i discorsi d'odio e le varie forme di violenza online necessitano di analisi, studi e soluzioni interdisciplinari: si tratta di questioni giuridiche e tecnologiche, psicologiche e sociali che interrogano profondamente anche la pedagogia e i processi educativi, nella prospettiva della *media education* così come in quella dell'educazione per una cittadinanza digitale, sostenibile, globale e a misura di futuro.

Di fronte a discorsi che inquinano e spesso feriscono i più vulnerabili si avverte, cioè, il bisogno di una «ecologia educativa del web» (Santerini, 2019, p. 54) volta a garantire la creazione di uno stile democratico di argomentare, di essere cittadini e di vivere.

1. Cfr. https://www.takebackthetech.net/. Ultima consultazione 21 settembre 2024.

2. Contenere i discorsi d'odio online: processi educativi tra cittadinanza e responsabilità

Pur non essendo l'unica soluzione possibile e pur necessitando di essere accompagnata da politiche di ampio respiro legate alla sicurezza, alla tutela dei diritti umani, alla non discriminazione, l'educazione alla cittadinanza, digitale nello specifico, è riconosciuta come lo strumento fondamentale per contrastare i discorsi d'odio, poiché fornisce agli individui le competenze necessarie per riconoscere, comprendere e reagire in modo critico a tali forme comunicative e incoraggia comportamenti rispettosi e inclusivi (Unesco, 2015; 2023).

La complessità del fenomeno da contenere richiede approcci educativi multiformi e integrati, capaci di promuovere processi che contengono e valorizzino elementi di educazione alla pace e alla cittadinanza globale e interculturale, puntando su ambienti di apprendimento sicuri, inclusivi, rispettosi dei diritti umani (Unesco, 2023; Westheimer & Kahne, 2004; Tibbitts, 2017; Common Sense Education, 2024). Società inclusive, interculturali, abituate a riconoscere il valore dell'essere cittadini globali lasciano meno spazio ai discorsi d'odio e alla loro diffusione e favoriscono la costruzione di relazioni inclusive e tolleranti nei mondi online e offline frequentati e costituiscono la base di comportamenti responsabili e rispettosi della dignità e dei diritti di tutti. L'educazione alla cittadinanza digitale si configura, di conseguenza, come una questione di tipo etico (Rivoltella, 2020) poiché si propone di fornire anche quelle competenze che abilitano l'esercizio della cittadinanza democratica, oltre che digitale, nella quale ciascuno «comprende e vive da protagonista e pertanto si fa *responsabile* del mondo in cui vive, opera e da cui insieme dipende» (Cambi & Pinto, 2023, p. 115) ed è dotato di una «una coscienza adeguata alla complessità del nostro tempo» (Cambi, 2021, p.10).

Educare alla cittadinanza digitale diviene, pertanto, un impegno a educare anche alla responsabilità e al suo esercizio perché, attraverso essa, si diviene sempre più consapevoli che si è perennemente chiamati a rispondere in prima persona delle conseguenze delle azioni e delle narrazioni che si mettono in atto nel contesto online. Perché reale e virtuale fanno al contempo parte della nostra vita (Rivoltella, 2017): siamo "onlife" in quanto la distanza tra online e offline si è talmente tanto ridotta da risultare pressoché nulla (Floridi, 2017). Puntare sulla comprensione che siamo

sempre e comunque onlife assume oggi un'importanza cruciale perché ci chiede di rivedere i comportamenti che assumiamo nel web alla luce di ciò che faremmo nel mondo offline. Spesso accade infatti che ciò che nella vita di tutti i giorni non faremmo mai, compresi l'utilizzo dell'*hate speech* e l'assunzione di comportamenti violenti, nella nostra vita online diviene plausibile, accettabile, incuranti delle possibili o potenziali conseguenze.

La *media education* ha il compito, allora, di porre le basi per una cittadinanza onlife, intesa come una prospettiva attenta alla dinamicità e alla transdisciplinarità dei nuovi alfabeti e che ritiene che un approccio troppo segmentato e settoriale dell'educazione digitale tradisca la vocazione di cittadinanza della competenza digitale (Buckingham, 2020 in Pasta 2021).

La formazione, anche in ambito scolastico, dei cittadini onlife è qualcosa di non rinviabile poiché è condizione indispensabile «per partecipare attentamente alla vita democratica» (Pasta, 2021, p. 84). È quanto sottolineano anche le nuove Linee Guida per l'insegnamento dell'Educazione civica del settembre del 2024[2], nelle quali emerge evidentemente il compito per la *media education* di connettere il pensare e l'agire, promuovendo competenze deliberative, oltre che cognitive e morali (Santerini, 2020, p.351).

Se è vero, però, che il contenimento e la prevenzione dei discorsi d'odio online sono questioni anche pedagogiche e che l'attivazione di processi educativi in tal senso deve puntare sulla formazione del pensiero e del senso critico e della responsabilità, è altrettanto vero che «è sul piano emozionale che si gioca la possibilità di fare del web un ambiente pulito e umano. Lo sviluppo della media education […] va in questa direzione, e nel sostegno di un'attivazione di un impegno civico online che crei una narrazione convincente, non violenta ed efficace» (Santerini 2019, p. 64) e che sia in grado di superare ogni «forma di reazione emotiva consegnata a istinti, pregiudizi» (Cambi & Pinto, 2023, p. 115). Gli studenti che riconoscono e sono più consapevoli delle proprie emozioni hanno maggiori possibilità di successo nelle sfide che il contrasto all'incitamento all'odio pone impone (Gavine et al., 2016).

Percorsi di media *education* attenti a tutte queste dimensioni sono, allora, una necessità urgente del tempo presente, indispensabili a rendere chiunque in grado di osservare, analizzare, creare, valutare e partecipare in modo pieno

2. https://www.miur.gov.it/documents/20182/0/Linee+guida+Educazione+civica.pdf/9ffd1e06-db57-1596-c742-216b3f42b995?t=1725710190643. Ultima consultazione 21 settembre 2024.

e sicuro alla vita digitale, consapevoli di cosa voglia dire rendere pubblico un commento e che un atto di questo tipo ha sempre delle conseguenze (Bruschi et al., 2023). Alle scuole il compito di garantire l'attuazione di tali percorsi anche con valore e scopo preventivo per aiutare soprattutto i più giovani a riconoscere e rispondere ai contenuti dannosi, ridurre l'incidenza dei discorsi d'odio e migliorare le loro capacità di analisi, anche in riferimento alle *fake news* (Pasta, 2018; 2019; Bruschi et al., 2023), imparando a discernere contenuti veritieri da quelli manipolativi (Bulger & Davidson, 2018). L'efficacia di questi processi educativi sarà condizionata dalla motivazione e dal coinvolgimento che riescono a suscitare negli studenti, promuovendo attività vicine alla loro vita e ai loro bisogni, magari chiamando a testimoniare e a dialogare con le giovani generazioni persone che hanno subito l'odio e/o la discriminazione online: gli studenti potranno, così, assumere il loro punto di vista, diventando più consapevoli e competenti nel contrastare questo fenomeno e nel promuovere l'inclusione e i diritti di tutti (Bruschi et al., 2023).

3. LLaMAntino per educare al rispetto e alla responsabilità: primi dati di una ricerca interdisciplinare

Attualmente l'innovazione e la trasformazione digitale si sono focalizzate sull'Intelligenza Artificiale (IA) Generativa e sui *Large Language Models* (LLM). Il successo di tali modelli sta riscuotendo un enorme impatto nelle attività quotidiane e nella ridefinizione di processi produttivi delle aziende e dei vari servizi erogati nei diversi ambiti, tra cui quella della pubblica amministrazione. In tale scenario, il campo dell'educazione e quello della società risultano fortemente interdipendenti e interconnessi.

Diversi sono i modelli di IA Generativa che attualmente competono per poter soddisfare le esigenze dei cittadini e imprese, startup e attività di business, come ChatGPT (OpenAI, 2023), LLaMa (Touvron et al., 2023), BLOOM (Workshop et al., 2022) e Mistral (Jiang et al., 2023). LLaMAntino (Basile et al., 2023) rappresenta uno degli LLM *open source* a partire da LLaMa2 di Meta addestrato per poter supportare la lingua italiana e che nel seguente lavoro di ricerca si è rilevato efficace nel poter strutture un processo sociale di educazione al rispetto e alla responsabilità.

LLaMAntino viene utilizzato nell'identificazione di violenza in conversazioni tossiche, specialmente nel contesto di relazioni intime. L'obietti-

vo è quello di rilevare e comprendere comportamenti violenti, sia cyber che fisici, e rendere consapevoli le persone coinvolte di tali dinamiche. Di conseguenza, ciò permette anche di lavorare su un approccio educativo che consenta di prevenire comportamenti dannosi riconoscendo anche la presenza di linguaggio manipolatorio.

Per raggiungere tale obiettivo, è necessario creare un dataset di conversazioni tossiche che presentino annotazioni sulle diverse tipologie di violenza come ad esempio fisica, cyberstalking, cyber sexual, al fine identificare comunicazione aggressiva e di fornire una spiegazione 'tecnica' del linguaggio offensivo adottato.

È stata eseguita una prima sperimentazione, estendendo il dataset *HuggingFace* (Martínez Gabaldón, 2023) con annotazioni specifiche e che identificassero la tipologia di violenza (fisica e cyber), la tipologia di comportamento aggressivo, presenza di comunicazione aggressiva e la specifica tipologia.

Di seguito viene riportato un esempio di frase tossica con le relative annotazioni: *"Se non hai nulla da nascondere, dammi il telefono"*

— Tipologia di violenza: fisica
— Tipologia di comportamento: aggressione psicologica
— Presenza di comunicazione aggressiva: si
— Tipologia di comunicazione: attacco alla persona

Tali annotazioni sono state trasformate in spiegazioni con l'obiettivo di fornire un prompt, cioè testo in linguaggio naturale, a LLaMAntino allo scopo di poter apprendere il contesto:

La frase è tossica perché è un esempio di violenza cyber. Il comportamento rientra nella categoria del cyber stalking, del monitoraggio e del controllo, poiché l'obiettivo è ottenere informazioni sulla vita del partner e stabilire una dinamica di controllo nella coppia. Inoltre, la comunicazione è aggressiva perché rivela l'intento intimidatorio di attaccare il partner per violare la sua privacy. Ciò ha permesso di eseguire una prima sperimentazione dove è stato fornito a LLaMAntino la descrizione del task, due frasi con relative spiegazioni in modo che imparasse a riconoscere la presenza di violenza, e una ulteriore frase tossica dove in questo caso è il modello a generare la spiegazione. Riportiamo un esempio di seguito:

— *Task*: Data una frase di una conversazione tra partner in una relazione intima, spiegare i motivi per cui la frase esprime un linguaggio tossico e rappresenta un caso di violenza fisica o cyber.
— *Frase*: Non mi piace che tu abbia così tanti amici su Instagram.
— *Spiegazione*: Questa frase è tossica perché è un esempio di violenza cyber in quanto riferita all'uso di un social media. Il comportamento indica un uso della tecnologia per il controllo delle amicizie del proprio partner. La comunicazione non è aggressiva.
— *Frase*: Posti solo video per mettere in mostra il tuo corpo.
— *Spiegazione*: Questa frase è tossica perché è un esempio di violenza cyber in quanto riferita all'uso di un social media per la diffusione di immagini del proprio corpo. Il comportamento intende causare sofferenza emotiva al proprio partner e rappresenta quindi una violenza psicologica. La comunicazione è aggressiva in quanto evidenzia in modo irrispettoso un attacco alla persona per un proprio comportamento su un social network.
— *Frase*: Le foto che pubblichi sono provocanti.
— *Spiegazione LLaMAntino*: Questa frase è tossica perché è un esempio di violenza verbale. La comunicazione è aggressiva perché usa la parola provocante per descrivere le foto che l'altra persona ha pubblicato sui social media…

Nel complesso, questa frase rappresenta una forma di violenza verbale che colpisce il benessere emotivo dell'altra persona.

I risultati hanno dimostrato che LLaMAntino fornisce una spiegazione corretta dell'identificazione della violenza in accordo alle annotazioni fornite dalle esperte sulla valutazione del perché quella conversazione si possa ritenere tossica.

4. Conclusioni

Dopo aver introdotto il tema dei discorsi d'odio e della loro maggiore diffusione tramite internet e i social media, il contributo ha posto in risalto come la *media education* e l'educazione alla cittadinanza digitale siano condizioni cruciali per affrontare gli stessi discorsi d'odio, per comprendere il contesto dei media e delle piattaforme digitali, per valutare criticamente i loro con-

tenuti e per prendere decisioni informate, come utenti e come produttori di contenuti (Unesco, 2021). Entrambe consentono di preparare gli studenti a trovare, accedere, utilizzare e creare informazioni in modo efficace e a navigare negli ambienti online e digitali in modo sicuro e responsabile; forniscono, altresì, la conoscenza dei propri diritti nello spazio del digitale, il modo in cui è opportuno ed etico interagire con gli altri utenti e con i contenuti con cui si entra in contatto (Unesco, 2022). Sebbene spesso le nuove tecnologie, la diffusione dell'uso di internet e dei social media abbiano determinato un incremento delle forme di intolleranza e di odio online, esse possono anche rappresentare un'opportunità non solo per promuovere ma anche per prevenire l'incitamento all'odio (Pasta, 2019). Per tali ragioni gli autori e il GdR dell'Università di Bari che lavora al progetto YYY, stanno addestrando e perfezionando il modello linguistico del LLamantino, convinti che possa essere utile per rilevare e spiegare la presenza di elementi di tossicità in un set di dati di conversazioni relative a relazioni tra adolescenti. I primi risultati stanno confermando che esso è in grado di spiegare la tossicità del linguaggio nelle conversazioni tra partner intimi, con un livello di efficacia adeguato alle richieste. Nel prossimo futuro, appena l'addestramento dello strumento sarà ultimato, i ricercatori hanno intenzione di testarne l'efficacia coinvolgendo gli adolescenti delle scuole di XXX, in modo da suscitare l'attivismo digitale e il protagonismo degli studenti sì da avvicinarli all'impegno partecipativo onlife applicato alle loro comunità con lo scopo di costruire assieme processi diffusi di *media literacy* che coinvolgano sì le scuole e i pari ma anche le comunità di riferimento (famiglia, gruppo dei pari associazioni) nella consapevolezza che l'educazione alla cittadinanza digitale non è solo compito e responsabilità della scuola ma di tutto il sistema formativo e di tutta la comunità.

Riferimenti bibliografici

Amnesty International, *#TOXICTWITTER. Violence and abuse against women online,* Amnesty International, 2018. In https://www.amnestyusa.org/wp-content/uploads/2018/03/Toxic-Twitter.pdf.

Basile, P., Musacchio, E., Polignano, M., Siciliani, L., Fiameni, G., & Semeraro, G., *Llamantino: Llama 2 models for effective text generation in Italian language*, 2023. arXiv preprint arXiv:2312.09993.

Basile, P., de Gemmis, M., Musacchio, E., Polignano, M., Semeraro, G., Siciliani, L., … & Sorianello, P., *Explaining Intimate Partner Violence with LLaMAntino*, 2024. https://ceur-ws.org/Vol-3762/510.pdf

Bruschi B., Repetto M., Talarico M., *A framework on media-educational initiatives to contrast online hate speech*, «QTimes», vol. 2, 1, 2023, pp. 7-16.

Buckingham D., *Digital Media Literacies: Rethinking Media Education in the Age of the Internet*, «Research in Comparative and International Education», vol. 2, 1, 2007, pp. 43-55.

Bulger M., Davison, P., *The Promises, Challenges, and Futures of Media Literacy*, «Journal of Media Literacy Education», vol. 10, 1, 2018, pp. 1-21.

Cambi F., *Scuola e cittadinanza. Per la formazione etico-politica dei giovani*, Studium Edizioni, Roma, 2021.

Cambi F., Pinto Minerva F., *Governare l'età della tecnica. Il ruolo chiave della formazione*, Mimesis, Milano, 2023.

Chapman M., Bellardi N., Peissl H., *Media Literacy For All. Supporting marginalised groups through community media*, Council of Europe, Luxemburg, 2020. In https://edoc.coe.int/en/media/8258-media-literacy-for-all-supporting-marginalised-groups-through-community-media.html.

Common Sense Education, *Digital Citizenship Curriculum Impact Report*, Common Sense Media, 2014. In https://www.commonsensemedia.org/sites/default/files/featured-content/files/digital-citizenship-curriculum-impact-report_june-2024.pdf.

Commissione Europea, *2nd Survey of Schools: ICT in Education, Publications Office of the European Union*, Luxemburg, 2019. In https://data.europa.eu/euodp/data/storage/f/2019-03-19T084831/FinalreportObjective1-BenchmarkprogressinICTinschools.pdf

Consiglio d'Europa, *Lotta contro il discorso d'odio. Raccomandazione CM/Rec(2022)16 del Comitato dei Ministri agli Stati membri sulla lotta contro i discorsi d'odio*, Lussemburgo, 2022. In https://rm.coe.int/italian-rec-2022-16-combating-hate-speech-it-2764-7330-5863-1/1680ad6162.

Faloppa F., *#Odio. Manuale di resistenza alla violenza delle parole*, UTET, Torino, 2020.

Floridi L., *La quarta rivoluzione. Come l'infosfera sta cambiando il mondo,* Raffaello Cortina, Milano, 2017.

Gavine A.J., Donnelly P.D., Williams D.J., *Effectiveness of universal school-based programs for prevention of violence in adolescents*, «Psychology of Violence», 6(3), 2016, pp. 390-399.

Hinduja S., Patchin J.W., *Connecting Adolescent Suicide to the Severity of Bullying and Cyberbullying*, «Journal of School Violence», vol. 18, 3, 2018, pp. 333–346.

Hobbs R., *Digital and Media Literacy: A Plan of Action*, The Aspen Institute, Washington, 2010. In https://www.aspeninstitute.org/wp-content/uploads/2010/11/Digital_and_Media_Literacy.pdf.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E., *Mistral 7B*. CoRR, 2023, abs/2310.06825. https://doi.org/10.48550/arXiv.2310.06825.

Lenhart A., Ybarra M., Zickuhr K., Price-Feeney M., *Online Harassment, Digital Abuse, and Cyberstalking in America*, Data & Society Research Institute, New York, 2016. In https://datasociety.net/wp-content/uploads/2016/11/Online_Harassment_2016.pdf.

Lingiardi V., Carone N., Semeraro G., Musto C., D'Amico M., Brena S., *Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis*, «Behaviour & Information Technology», 2020, vol. 39, 7, pp. 711-721.

Livingstone S., Helsper, E., *Balancing opportunities and risks in teenagers' use of the internet: The role of online skills and internet self-efficacy*, «New Media & Society», vol. 12, 2, 2010, pp. 309–329.

Martínez Gabaldón, M., *Toxic-teenage*-relationships, «Hugging Face», 2023. https://huggingface.co/datasets/marmarg2/toxic-teenage-relationships, Doi: 10.57967/hf/0972

OpenAI, *Gpt-4 technical report*, 2023. arXiv:2303.08774.

Pasta S., *Razzismi 2.0. Analisi socio-educativa dell'odio online*, Scholé Morcelliana, Brescia, 2018.

—— *Conversazioni via social network con giovani autori di performance d'odio*, «Pedagogia Oggi», XVII, vol. 2, 2021, pp. 369-383.

—— *Partecipazione onlife: promuovere l'attivismo degli "spettautori" nel social web*, in Pasta S. & Santerini M. (a cura di), *Nemmeno con un click. Ragazze e odio online*, Franco Angeli, Milano, 2021, pp. 81-97.

—— *L'odio online e il posizionamento della Chiesa Cattolica*, «Veritas et Jus», 2022, vol. 25, 2, pp. 85-105.

—— *Contrastare l'odio online con la partecipazione dei gruppi eletti a bersaglio. La proposta metodologica del progetto REASON – REAct in the Struggle against ONline hate speech*, «QTIMES», 2023, XV, 3, pp. 429-445.

Rivoltella P.C., *Media education. Idee, metodo, ricerca*, La Scuola, Brescia, 2017.
—— *Nuovi alfabeti. Media e cultura nella società postmediale*, Morcelliana Scholé, Brescia, 2020.
Santerini M., *Discorso d'odio sul web e strategie di contrasto*, «MeTis-Mondi educativi. Temi indagini suggestioni», 9(2), 2019, pp. 51-67.
—— Democrazia partecipativa e nuova cittadinanza, «Rivista di Scienze dell'Educazione», 3, 2020, pp. 345-356.
Suler J., *The online disinhibition effect*, «CyberPsychology & Behavior», vol.7, 3, 2006, pp. 321-326.
Tibbitts, F.L., *Revisiting 'Emerging Models of Human Rights Education'*, «International Journal of Human Rights Education», vol.1, 1, 2017.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., *Llama: Open and efficient foundation language models*, 2023. arXiv preprint arXiv:2302.13971.
Unesco, *Countering Online Hate Speech*, Unesco, Parigi, 2015.
—— *Addressing hate speech on social media: contemporary challenges*, Unesco, Parigi, 2021.
—— *Addressing hate speech through education. A guide for policy-makers*, Unesco, Parigi, 2023.
Unicef, *How many children and young people have internet access at home?*, Unicef, New York, 2020. In https://www.itu.int/en/ITU-D/Statistics/Documents/publications/UNICEF/How-many-children-and-young-people-have-internet-access-at-home-2020_v2final.pdf.
Vega V., Robb M.B., *Inside the 21st-Century Classroom*, Common Sense Media, 2019. In https://www.commonsensemedia.org/sites/default/files/research/report/2019-educator-census-inside-the-21st-century-classroom_1.pdf.
Westheimer J., Kahne J., *What Kind of Citizen? The Politics of Educating for Democracy*, «American Educational Research Journal», vol. 41, 2, 2004, pp. 237-269.
Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al., *BLOOM: A 176B-parameter open-access multilingual language model*, 2022. arXiv preprint arXiv:2211.05100.