

SPECIAL ISSUE COVID-19

L'epidemia dell'odio online: l'hate speech ai tempi del Covid-19

Alice Tontodimamma, Francesca Michetti

Il virus dell'odio: tra passato e presente

Nelle ultime settimane, oltre ai sintomi del Covid-19, a qualcuno sarà capitato di manifestare i sintomi di un altro virus, altrettanto insidioso, quello dell'**odio**.

Numerosi gli episodi di discriminazione, sul web e fuori dal web, che hanno affollato le pagine di **cronaca** ad un ritmo anche più veloce delle notizie inerenti al coronavirus.

«Le parole d'odio sono sempre esistite» (Ghenò, 2019), da ben prima dell'avvento del web e dei social network.

Ce lo ricorda Manzoni, sullo sfondo dell'epidemia di peste del 1630 già imperversava l'odio, assumendo i tratti della ben nota **caccia all'untore**. Del resto, la storia ci insegna che il bisogno di dare un volto e un nome alle nostre paure ci induce, spesso, a cercare un

capro espiatorio a cui imputare l'origine di tutti i mali. E così oggi, come nella Milano della peste manzoniana, in quella che – ricordiamo – è nata come la **prima pandemia social**, riecheggiano vivide antiche forme di intolleranza, odio e discriminazione nella veste nuova e sopraffina dei cosiddetti **«hate speech»**.

Hate speech: che cos'è?

Nuove forme di un fenomeno antico del quale, tutt'oggi, non esiste una definizione univoca (Brown, 2017) e che resta al centro di un intenso dibattito giuridico e accademico a livello internazionale.

La difficoltà di rinvenire una definizione realmente omnicomprensiva di tutte le possibili fattispecie di odio risiede nel rischio di collisione con alcuni dei **principi base dei sistemi democratici**, tra cui l'**uguaglianza**, la **dignità umana** e la **libertà di espressione**; ma deriva, altresì, dalla natura soggettiva, e quindi

facilmente manipolabile, del termine (Isasi e Juanatey, 2016).

L'espressione *hate speech* significa, letteralmente, *discorso d'odio*, talvolta tradotta come *incitamento all'odio*. La sostanza non cambia: si tratta di una definizione-cappello che indica qualsiasi espressione, indipendentemente dalle forme assunte (scritte o orali, verbali o non verbali, esplicite o implicite), contenente insulti, offese, dichiarazioni di intolleranza nei confronti di un singolo o di un gruppo specifico (Ghenò, 2019).

La categoria dell'*Hate Speech* fu elaborata negli anni '70 dalla giurisprudenza statunitense; a livello europeo, soprattutto negli ultimi anni, si è cercato di delimitarne i confini. Stando alla Raccomandazione del Consiglio d'Europa del 1997, essa comprende *«tutte le forme di espressione che diffondono, incitano, promuovono o giustificano l'odio razziale, la xenofobia, l'antisemitismo o altre forme di minaccia basate sull'intolleranza, inclusa l'intolleranza espressa dal nazionalismo aggressivo e dall'etnocentrismo, la discriminazione e l'ostilità contro le minoranze, i migranti e le persone di origine immigrata»*.

Con il tempo la sensibilità sul tema è cresciuta fino a comprendere, tra i potenziali bersagli delle espressioni di odio, anche le minoranze religiose, le donne, le persone LGBT, i disabili e gli anziani. Secondo un Report dell'UNESCO del 2015 intitolato

Countering Online Hate Speech, l'odio online assume delle caratteristiche precise:

1. la **permanenza** nel tempo;
2. il **carattere itinerante** (*itinerancy*) del messaggio che, anche dopo la sua rimozione, potrebbe apparire sulla stessa piattaforma o altrove, con una diversa denominazione e/o tramite un diverso utente;
3. l'**anonimato** – e la conseguente sensazione di impunità – tramite l'uso di pseudonimi o nomi fittizi che incoraggiano l'espressione dell'odio in rete;
4. infine, la **transnazionalità** dei contenuti ovvero l'assenza di confini e la possibilità di diffusione capillare dei messaggi, che alimenta il fenomeno e complica l'individuazione degli strumenti legali per arginarlo.

La comunicazione mediata e la capacità di propagazione virale della rete, poi, ne amplificano gli effetti; senza trascurare le dinamiche della cosiddetta **“polarizzazione di gruppo”**. Come riscontrato da Sunstein (2003), nei dibattiti in rete, gli utenti si confrontano con persone appartenenti al proprio gruppo di opinione – spesso senza sentire alcuna controparte – finendo così per dirigersi ideologicamente verso *«un punto estremo nella direzione in cui i membri del gruppo erano originariamente orientati»*. Da questo punto di vista, i social network, nel favorire lo scambio di opinioni,

amplificano notevolmente la diffusione di questo genere di espressioni, proprio come è accaduto durante la pandemia.

Il che ha portato – già da tempo – a proporre una serie di iniziative internazionali ed europee, volte a definire il problema e a sviluppare contromisure efficaci. Volgendo lo sguardo al panorama italiano, il 31 ottobre 2019 il Senato ha approvato a larga maggioranza la mozione per la costituzione di una **Commissione straordinaria per il contrasto ai fenomeni dell'intolleranza, del razzismo, dell'antisemitismo e dell'istigazione all'odio e alla violenza**. Inoltre, sono stati messi in atto numerosi progetti per contrastare questo fenomeno (“Contro l’Odio”, “Parole O_Stili”, “Odiare ti costa”, “Il barometro dell’odio”).

I bersagli dell’odio durante il Covid-19

Nel mirino degli *haters* nelle scorse settimane, al primo posto, la popolazione cinese e, in generale, gli individui di origine asiatica. Un rapporto di L1ght, un’azienda specializzata nella misurazione della tossicità online, ha riportato – nella fase iniziale della pandemia – un aumento del 900% dei discorsi di odio verso la Cina su Twitter. Del resto, #Chinaliedpeopledied è stato il *leitmotiv* che ha accompagnato i dibattiti social negli ultimi due mesi.

A seguire, i meridionali che hanno dato luogo al celebre esodo da Milano, sul famoso “treno della vergogna”. Tra i casi confermati anche i migranti, risultati misteriosamente immuni al contagio e, pertanto, probabili fautori di un’ennesima “invasione clandestina infetta”; senza tralasciare i leader politici – da Trump a Boris Jhonson – protagonisti indiscussi della scena mediatica.

E non solo. Oltreoceano il Covid-19 ha (ri)acceso lo scontro generazionale: un tweet del 12 marzo, divenuto poi virale, ha definito il virus un “*boomer remover*” (“*rottamatore di boomer*”), con ciò alludendo all’eliminazione fisica di una generazione, quella *over 60*, considerata tutto sommato ‘sacrificabile’.

E, infine, i temibili *runners* che l’opinione pubblica, affacciata alla finestra, non ha tardato ad individuare, fotografare e a consegnare quotidianamente alla gogna mediatica in qualità di novelli untori dei giorni nostri.

La ricerca sull’ hate speech

Alla luce della notevole importanza ed attualità di questo fenomeno, si è riscontrato un crescente interesse per il tema in ambito scientifico ed accademico. Diverse discipline, tra cui le scienze sociali, la psicologia, la statistica e l’informatica, sono impegnate nella ricerca sul discorso dell’odio.

Il rilevamento degli *hate speech* è un compito complesso che presenta molte sfide.

Sul piano metodologico, il primo passo nell'azione di contrasto al fenomeno consiste nella delimitazione del campo di indagine, attraverso l'utilizzo di **tecniche automatiche** che consentono di identificare i discorsi d'odio. Ma poiché – come dicevamo – non esiste una definizione universale di *hate speech*, concordare se si tratti o meno di una espressione d'odio diventa molto difficile. Oltre all'ambiguità del termine, vanno considerati anche i limiti umani e tecnologici nella mappatura dell'odio in rete. Innanzitutto, i programmi utilizzati per l'analisi dei post pubblicati sui social network “imparano” da una banca dati messa a punto da esseri umani. E, se errare è umano per gli individui, lo stesso dicasi per le macchine (Ruffo, 2020). Il basso tasso di accordo (33%) riscontrato nella classificazione dei discorsi d'odio ad opera di esseri umani fa presagire quanto questo sia un compito ancora più complesso per le macchine.

Quanto alla fase di ricerca vera e propria, generalmente si utilizzano datasets contenenti messaggi scaricati da social networks (Twitter è la fonte più comunemente usata); l'approccio prevalente si fonda sulla creazione di modelli di *machine learning* per la classificazione degli *hate speech*; la lingua più utilizzata è l'inglese.

Una delle principali difficoltà riscontrate risiede nel fatto che, spesso, gli autori non utilizzano datasets pubblici e non pubblicano i nuovi dati raccolti, il che rende molto difficile la comparazione dei risultati e, quindi, delle conclusioni. Inoltre, l'annotazione di un set di dati richiede specifiche competenze su ciascuna cultura e sulle relative caratteristiche giuridiche e sociali; a ciò si aggiunga che l'**evoluzione del linguaggio** e dei fenomeni sociali rende particolarmente difficile tenere costantemente traccia di tutte le espressioni di odio rivolti verso i gruppi di minoranza (Fortuna and Nunes, 2018).

Un'altra sfida è legata alla **comprensione del linguaggio naturale**, ambiguo e spesso informale, contrariamente a quello delle macchine. I post pubblicati sui social network, in particolare su Twitter, sono brevi, contengono spesso errori di ortografia, slang, abbreviazioni, *hashtag* ed *emoji*. C'è poi da considerare l'impiego di metafore, sarcasmo o ironia, che può distorcere il senso generale; infine, da non trascurare il contesto di ciascun messaggio.

A tal proposito, sottolineiamo l'importanza di sviluppare **risorse lessicali** per la rilevazione dell'odio online e, in generale, del linguaggio abusivo. Nella fattispecie, alcuni ricercatori, basandosi sull'assunto generale che i messaggi d'odio contengono specifiche parole negative

(come parolacce e insulti), utilizzano la presenza di tali parole come *feature* per rilevare gli *hate speech* (Schmidt and Wiegand, 2017).

In Italia, ad esempio, il gruppo di ricerca del Dipartimento di Informatica dell'Università di Torino ha sviluppato un **lessico computazionale multilingue**, tradotto semi-automaticamente in più di 50 lingue, composto da parole utilizzate per ferire, chiamato **Hurtlex** (Bassignana et al., 2018). Come punto di partenza utilizza il lessico italiano "*Le Parole per Ferire*" del linguista Tullio De Mauro, che contiene circa 1000 parole ed è organizzato in 17 categorie che mirano a catturare il contesto di ogni parola (De Mauro, 2016).

Senza dubbio lo sviluppo e il miglioramento di tali risorse lessicali è uno strumento prezioso nell'azione di contrasto all'odio online. Ma per sradicare l'*hate speech*, oltre ad identificare le parole 'incriminate',

bisogna anche offrire una narrazione qualitativamente diversa – una **contronarrazione** – a cui aderire.

Per concludere

Nella dimensione online, «*a rappresentare ciò che siamo ci sono solo le nostre parole*» (Ghenò, 2018). Va da sé quanto queste, in un tale contesto, siano ancora più importanti.

E se è vero che il virus in questi giorni sembra rallentare, lo stesso non può dirsi per l'odio, nella sua dimensione sia online sia offline.

In proposito, risuonano quanto mai attuali le toccanti parole di Andrea Camilleri "*stiamo perdendo la **misura**, il **peso**, il **valore** della **Parola** [...]. Bisogna pesare ogni parola che si dice e soprattutto far cessare questo **vento dell'odio**, che è veramente atroce, lo si sente, palpabile, attorno a noi*".

Bibliografia

1. Gheno V. (2019), "Potere alle parole", <https://volerelaluna.it/cultura/2019/09/20/potere-alle-parole-intervista-a-vera-gheno/?print=pdf>
2. Brown A. (2017). What is hate speech? Part 1: the Myth of Hate, *Law and Philosophy*, 36, pp. 419-468.
3. Isasi, A. C., & García Juanatey, A. (2017). Hate Speech in Social Media: A state-of-the-art review.
4. Gheno V., (2019). Cos'è l'hate speech e come lo si combatte, <https://semplicecome.it/tendenze-hate-speech/>
5. Recommendation No. R (97) 20, of the Council of Europe Committee of Ministers on "Hate Speech," 30 October 1997, available at <https://rm.coe.int/1680505d5b>
6. UNESCO, (2015), Countering Online Hate Speech <https://en.unesco.org/news/unesco-launches-countering-online-hate-speech-publication>
7. Sunstein C.R. (2003), Republic.com. Cittadini informati o consumatori di informazioni?, Bologna, Il Mulino.
8. Ruffo G. (2020) https://lavalibera.libera.it/it-schede-22-odio_online
9. Fortuna P, Nunes S (2018) A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)* 51(4).
10. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics, pp 1–10.
11. Bassignana E, Basile V, Patti V, (2018), Hurltlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018* (Vol. 2253, pp. 1-6). CEUR-WS.
12. De Mauro T. (2016). Le parole per ferire. *Internazionale*. Compiled for the "Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena, of the Italian Chamber of Deputies, which issued a Final Report in 2017. <https://www.internazionale.it/opinione/tullio-de-mauro/2016/09/27/razzismo-parole-ferire>
13. Gheno V., (2018). Tutti i modi dell'hate speech sui social media: quando la lingua separa e ferisce, *Agenda Digitale*, <https://www.agendadigitale.eu/cultura-digitale/quando-la-lingua-separa-e-ferisce44241/>