Form@re
Open Journal per la formazione in rete

dossier Covid

Antisemitism and Covid-19 on Twitter. The search for hatred online between automatisms and qualitative evaluation

Antisemitismo e Covid-19 in Twitter. La ricerca dell'odio online tra automatismi e valutazione qualitativa

Stefano Pasta[a], Milena Santerini[b], Erica Forzinetti[c], Marco L. Della Vedova[d,1]

[a] *Università Cattolica del Sacro Cuore di Milano, stefano.pasta@unicatt.it*
[b] *Università Cattolica del Sacro Cuore di Milano, milena.santerini@unicatt.it*
[c] *Università Cattolica del Sacro Cuore di Milano, erica.forzinetti@unicatt.it*
[d] *Università Cattolica del Sacro Cuore di Milano, marco.dellavedova@unicatt.it*

Abstract

The article presents a case study on Antisemitic hate speech in Twitter in the period September 2019 - May 2020, with a particular focus on the months of the Covid-19 emergency. The corpus, consisting of 160.646 tweets selected by keywords, was investigated in terms of the amount of hate for each month, rhetoric and forms of Antisemitism. The analysis is carried out through social network analysis (SNA) techniques, with the goal of understanding whether it is possible to automate the process of identifying Antisemitic hatred. 26.11% of tweets contain hatred, that prejudice is the most common rhetoric (44%) and association with financial power the prevailing form (74%). The sample was also compared with another research methodology that only detects the presence of hate words. It emerges that, in addition to an in-depth knowledge of the phenomenon, it is necessary to integrate the automatic classification phase with the manual contribution.

Keywords: social network analysis; hate speech; antisemitism; hate words; hate.

Sintesi

L'articolo presenta un caso studio sul discorso d'odio antisemita in Twitter nel periodo settembre 2019 - maggio 2020, con un particolare affondo sui mesi dell'emergenza Covid-19. Il corpus, composto da 160.646 tweet selezionati per parole chiave, è stato indagato in termini di quantità di odio per mese, retoriche utilizzate e forme di antisemitismo. L'analisi è svolta attraverso le tecniche di *social network analysis* (SNA), con l'obiettivo di capire se sia possibile automatizzare il processo di individuazione dell'odio antisemita. Il 26.11% dei tweet contiene odio, che il pregiudizio è la retorica più presente (44%) e l'associazione al potere finanziario la forma prevalente (74%). Il campione è stato altresì confrontato con un'altra metodologia di ricerca che rileva la sola presenza di *hate words*. Emerge che, oltre una conoscenza approfondita del fenomeno, occorre integrare la fase di classificazione automatica con l'apporto manuale.

Parole Chiave: social network analysis; hate speech; antisemitismo; parole di odio; odio.

[1] La presente ricerca si inserisce nei lavori dell'Osservatorio sull'odio online Mediavox dell'Università Cattolica del Sacro Cuore di Milano, promosso dal Centro di Ricerca sulle Relazioni Interculturali e composto da un'équipe di ricercatori multidisciplinare (www.mediavox.network).

1. Introduzione

Nel linguaggio, nei discorsi e nelle narrazioni pubbliche dell'infosfera e della società *onlife* (Floridi, 2017), ossia segnata dalla continuità di rimandi tra online e offline, si sono innescati concatenamenti discorsivi in grado di costruire nuovi paradigmi di odio (Pasta, 2018; Santerini, 2019a, 2021). Invettive, insulti, offese, retoriche di avversione accompagnano le relazioni e i conflitti umani lungo la storia, ma è in corso oggi una riflessione singolare su questo fenomeno, percepito sempre più come un tema vitale per la tenuta della democrazia e della vita sociale.

La diffusione del Web 2.0 ha aperto nuove dimensioni alla parola d'odio, ne ha trasformato la struttura, la sintassi ma prima ancora il significato e le motivazioni. La maggior parte di questa comunicazione, che dilaga in forma liquida, destrutturata e banalizzata, avviene all'insegna delle emozioni, che orientano e dirigono la nostra mente in modo intelligente ma anche rapido e istintivo (Damasio, 1995; Nussbaum, 2004; Pasta, 2019; Wallace, 2007).

Da un punto di vista giuridico il *discorso d'odio* (*hate speech*) non ha chiara definizione (Faloppa, 2020; Ziccardi, 2016; 2019). Per il Consiglio d'Europa (Racc. n. 97/20), si intendono "tutte le forme di espressione che diffondono, incitano, promuovono o giustificano l'odio razziale, la xenofobia, l'antisemitismo o altre forme di odio basate sull'intolleranza, tra cui: 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". D'altro canto, si potrebbe parlare di odio come sentimento, postura prolungata e sistematica di avversione verso un'altra persona o gruppo, o come di una dinamica intrinseca psicologica che può o meno indurre a comportamenti violenti (Santerini, 2019b).

Proprio per la sua multidimensionalità, il fenomeno necessita di un approccio multidisciplinare. Dal campo giuridico la riflessione è arrivata nelle discipline umanistiche (sociologiche, pedagogiche, antropologiche, filosofiche, linguistiche, semiotiche), convergendo nell'ambito interdisciplinare degli *Hate Studies*, che riunisce studiosi, ricercatori, politici, esperti di comunicazione, attivisti dei diritti umani, responsabili di ONG.

Da un punto di vista pedagogico, l'educazione alla cittadinanza è interrogata dalla narrazione ostile, dalla visione binaria del mondo, diviso in noi/loro, amico/nemico, dentro/fuori, dall'affermarsi di processi di elezione a gruppo bersaglio diffusi e banali, non solo in parole ma anche in immagini (Pasta, 2020). L'hate speech è considerato un grave pericolo per la coesione sociale, tanto più che il confine sottile con la libertà d'espressione e l'esigenza di preservare questo diritto fondamentale forniscono, a volte, il pretesto di lasciarlo diffondere senza limiti[2]. Dalla didattica della Shoah (Santerini, 2005) traiamo l'insegnamento sulla necessità di riflessione educativa sui meccanismi di elezione a bersaglio, sintetizzato nello strumento educativo della *Piramide dell'odio* dell'Anti-

[2] L'European Commission against Racism and Intolerance (ECRI) del Consiglio d'Europa ha raccomandato nel 2015 di porre limiti alla pur fondamentale libertà d'espressione e di opinione quando viola la dignità degli altri (https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01).

Defamation League[3] (https://www.adl.org), che interpreta l'esito estremo dell'odio come soglie che si passano e tabù che si abbattono.

Come si evince dall'analisi di tipo qualitativo-testuale e qualitativo-motivazionale e dalle successive conversazioni con giovani autori di performances d'odio svolte da Stefano Pasta (2018), la riflessione educativa ha l'esigenza di studiare più a fondo il fenomeno per prevenire e contrastare le varie forme espresse in forma di linguaggi e di immagini online.

Le ricerche avviate finora in Italia hanno permesso di analizzare l'hate speech online in diversi modi: in alcuni casi lo si è studiato secondo l'angolatura della mappatura geografica (Vox-Osservatorio italiano sui diritti, 2019)[4], in altri si è tentata una quantificazione di massima utilizzando, per individuare l'odio online, solo termini dichiaratamente offensivi, come nella ricerca di DataMediaHub e KPI6 (2020)[5], i cui risultati indicherebbero che il fenomeno sia più circoscritto e limitato. Le analisi di tipo linguistico che si sono sviluppate sulla base di questo assunto hanno compiuto vari passi avanti nell'individuazione delle componenti lessicali dell'hate speech (Femia, 2019; Ferrini & Paris, 2019). Anche dal punto di vista metodologico sono stati compiuti vari affondi, sperimentando diverse modalità per il trattamento automatico del discorso (sentiment analysis, text mining, etc.) e sulla raccolta ed annotazione dei dati per cogliere le sfaccettature delle espressioni di odio (Sanguinetti, Poletto, Bosco, Patti, & Stranisci, 2018).

Proprio per le implicazioni educative del fenomeno, da queste prime analisi appare necessario affiancare alla *detection* dell'odio online un più approfondito studio delle sue caratteristiche, creando una sinergia tra le linee di ricerca delle scienze informatiche e quelle umanistiche. Si tratta infatti non solo di decidere *cosa sia odio*, quanto di analizzare le numerose forme in cui si esprime, di poter meglio individuare soggetti, bersagli e modalità di espressione, al fine dell'elaborazione di strategie di prevenzione e contrasto più efficaci.

2. Il caso studio

In questo articolo si presenta un caso studio di analisi dell'hate speech, che coniuga l'approccio socio-educativo e il trattamento informatico automatico, a partire dalla metodologia messa a punto dall'Osservatorio Mediavox dell'Università Cattolica del Sacro Cuore e in corso di applicazione anche per altri gruppi bersaglio.

In questo caso specifico si affronta il problema della classificazione del discorso di odio su Twitter, focalizzando l'attenzione sull'antisemitismo in Italia nel periodo del Covid-19. La prima domanda alla quale si vuole rispondere è se sia aumentato l'odio contro gli ebrei durante l'emergenza sanitaria, attraverso analisi temporali su classificazioni campionarie

[3] Online è possibile scaricarne alcune versioni in italiano realizzate dallo Shoah Foundation Institute per differenti target di età.

[4] La IV edizione della Mappa dell'Intolleranza è del 2019; il progetto è stato messo a punto con il contributo di quattro Atenei (Dipartimento di Diritto pubblico, italiano e sovranazionale – Università degli Studi di Milano; Dipartimento di Psicologia dinamica e clinica – Università Sapienza di Roma; Dipartimento di Informatica – Università Aldo Moro di Bari; centro Itstime – Università Cattolica del Sacro Cuore di Milano) (http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-4/).

[5] Si veda il "Report su Hate Speech in Italia su Twitter" (http://www.datamediahub.it/2020/06/22/rapporto-sullhate-speech-in-italia/#axzz6cYynj6H0).

effettuate manualmente da esperti del settore. La seconda domanda di ricerca riguarda invece quali retoriche e forme di odio emergono nell'accostamento tra ebrei e coronavirus, paragonando le retoriche e le forme di antisemitismo riscontrate su Twitter prima dell'emergenza sanitaria e durante la pandemia.

Le retoriche tra cui scegliere sono state: insulti, derisione/ironia, esclusione/separazione, pregiudizio, disumanizzazione, umiliazione/disprezzo, paura, concorrenza, incitamento/violenza. Queste emergono da un'analisi psico-sociale e storico-letteraria sulle forme linguistiche dell'ostilità verso l'altro effettuata da ricercatori esperti del discorso antisemita (Milena Santerini e Stefano Pasta)[6]. Inoltre, fanno riferimento alla costruzione sistematica del disprezzo organizzata dalla propaganda nazista per la distruzione degli ebrei d'Europa durante la Seconda guerra mondiale.

Le forme sono state mutuate dalla definizione di antisemitismo (WDA – Working Definition Antisemitism) accolta dal Governo italiano nel 2020 sulla base di quanto stabilito dall'International Holocaust Remembrance Alliance (IHRA), l'organizzazione intergovernativa fondata nel 1998 che unisce gli Stati e gli esperti per rafforzare, promuovere e divulgare l'educazione sulla Shoah, la ricerca e il ricordo in tutto il mondo e il sostegno degli impegni della Dichiarazione del Forum internazionale di Stoccolma del 2000[7]. Tra le forme antisemite vi sono pertanto: odio contro gli ebrei in quanto tali (incitare, sostenere o giustificare i danni o l'uccisione degli ebrei in nome di un'ideologia radicale o di estremismo religioso), antigiudaismo (accusare gli ebrei di deicidio, di compiere omicidi rituali, di idolatria, di attaccare il cristianesimo), neonazismo e neofascismo e negazionismo della Shoah (considerare gli ebrei come una *razza* inferiore, demonizzandoli e disumanizzandoli, esaltare i simboli del nazismo e del fascismo e denigrarne le vittime), antisionismo/odio verso lo stato di Israele (diffamare e incitare a distruzione, negare a Israele il diritto all'autodeterminazione), potere ebraico sulla finanza (accusare gli ebrei come singoli o collettività di avere il controllo della finanza mondiale, dei media, delle banche, dell'economia, del governo o di altre istituzioni).

3. La metodologia

La metodologia di analisi utilizzata rientra nelle tecniche di *social network analysis* (SNA). I dati sono stati raccolti utilizzando la libreria open-source Python GetOldTweets3 (https://pypi.org/project/GetOldTweets3), con la quale è possibile ottenere i tweet tramite la ricerca per query, per username, per località, linguaggio o date, senza limitazioni temporali. Ogni tweet scaricato è caratterizzato da diverse informazioni oltre al testo del tweet, quali nome utente dell'autore, menzioni, hashtag, numero di retweet e di preferiti, la data, l'ora e la geolocalizzazione.

[6] Milena Santerini è la coordinatrice nazionale per la lotta contro l'antisemitismo scelta del Governo italiano e vicepresidente della Fondazione Memoriale della Shoah (Santerini, 2005); Stefano Pasta è esperto di odio online (Pasta, 2018) e coordinatore della "Ricerca-azione sui discorsi d'odio online di natura antireligiosa" (Pasta, 2020), svolta dal Centro di Ricerca sulle Relazioni Interculturali dell'Università Cattolica del Sacro Cuore insieme al Centro di Documentazione Ebraica Contemporanea (CDEC) e dai Giovani Musulmani d'Italia (GMI).

[7] L'IHRA ha 34 paesi membri, un paese di collegamento e sette paesi osservatori (https://www.holocaustremembrance.com).

Per rispondere alla prima domanda di ricerca "È aumentato l'odio verso gli ebrei durante l'emergenza coronavirus?" sono stati analizzati 160.646 tweet in lingua italiana pubblicati tra il 1° settembre 2019 e il 31 maggio 2020. La stringa di ricerca utilizzata per scaricare i tweet al centro dell'indagine è stata "ebrei OR Soros OR Israele OR sionista OR sionismo", termini individuati a partire dal WDA-IHRA (IHRA, 2016)[8], dai rapporti internazionali dell'Osservatorio sull'antisemitismo del Centro di Documentazione Ebraica Contemporanea (CDEC), del Kantor Center for the Study of Contemporary European Jewry, oltre sulla base della letteratura sull'antisemitismo (Santerini, 2019b; Taguieff, 2016; Wieviorka, 2019) e di una precedente ricerca svolta dall'Osservatorio Mediavox (2020)[9]; in questo modo sono stati quindi scaricati tutti i tweet contenenti almeno una delle parole presenti nella query. Successivamente, seguendo la tecnica del campionamento casuale semplice senza ripetizione, tra questi tweet è stato selezionato un campione composto da 100 tweet per ogni mese, ottenendo così un dataset campionario di 900 post totali (James, Witten, Hastie, & Tibshirani, 2017). Quest'ultimo è stato classificato manualmente da esperti del settore (ruolo di *annotatori*). I tre annotatori, avendo a disposizione il nome dell'autore del post, la data di pubblicazione e il testo del tweet con il rispettivo numero di like e retweet, hanno stabilito se il tweet contenesse odio oppure no. Nel caso in cui si trattasse di un contenuto d'odio, hanno assegnato la retorica e la forma antisemita corrispondente.

Per rispondere alla seconda domanda di ricerca "Quali retoriche e forme di odio ci sono nell'associazione ebrei e coronavirus?", il dataset analizzato è composto da 4.048 tweet pubblicati tra il 1° marzo 2020 e il 31 maggio 2020, in lingua italiana. La stringa di ricerca utilizzata per scaricare i tweet è stata "(covid OR coronavirus OR epidemia OR pandemia) AND (ebrei OR Soros OR Israele OR sionista)"; in questo modo sono stati scaricati tutti i tweet contenenti almeno una delle parole presenti nel primo gruppo e contemporaneamente almeno una delle parole presenti nel secondo gruppo della query. Anche in questo caso è stato selezionato un campione casuale di 900 tweet totali che è stato classificato manualmente da esperti del settore con le stesse modalità adottate per la precedente domanda di ricerca.

4. I risultati

Inizialmente è stata fatta una analisi temporale dei due corpus di tweet scaricati.

La Figura 1 mostra il numero di tweet scaricati mese per mese tra settembre 2019 e maggio 2020. Come si può vedere, mentre gli altri mesi hanno un andamento più o meno costante, a novembre e gennaio si registra il maggior numero di tweet (22.861 e 24.471 rispettivamente), fatto che coincide con il picco di antisemitismo solitamente registrato a gennaio in coincidenza con la Giornata della Memoria e, per il mese di novembre 2019, con le polemiche contro la senatrice Liliana Segre, sopravvissuta ad Auschwitz e testimone della Shoah molto conosciuta in Italia, relativamente all'approvazione della Commissione

[8] https://www.holocaustremembrance.com/working-definition-antisemitism.

[9] Si fa riferimento al corpus di discorsi antisemiti selezionati dal Centro di Documentazione Ebraica Contemporanea (CDEC) nell'ambito della "Ricerca-azione sui discorsi d'odio online di natura antireligiosa", svolta nel 2019-20 dal Centro di Ricerca sulle Relazioni Interculturali dell'Università Cattolica su incarico dell'Ufficio Nazionale Antidiscriminazioni Razziali (UNAR) della Presidenza del Consiglio dei Ministri.

straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza (Mozione n. 1/00136)[10] (Osservatorio antisemitismo della Fondazione CDEC, 2020).

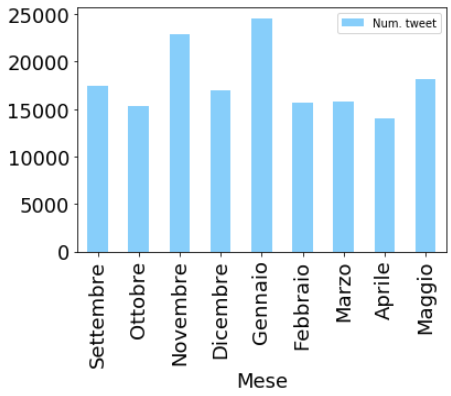

Figura 1. Numero di tweet relativi agli ebrei associati a parole chiave dell'antisemitismo, pubblicati tra il 1° settembre 2019 e il 31 maggio 2020.

La Figura 2, invece, mostra il numero di tweet scaricati tra marzo e maggio 2020 inerenti all'associazione ebrei-coronavirus. Come si può vedere, vi è un andamento decrescente nei mesi presi in considerazione (2.055 tweet nel mese di marzo, 1.274 nel mese di aprile e 719 nel mese di maggio).

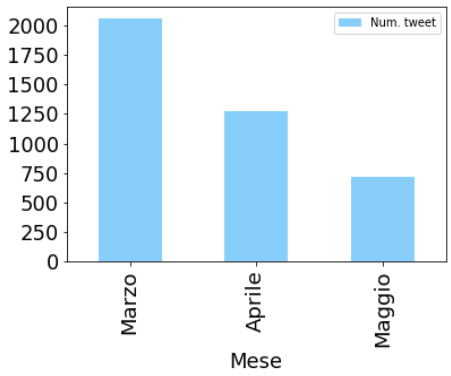

Figura 2. Numero di tweet relativi all'associazione ebrei-coronavirus pubblicati tra il 1° marzo e il 31 maggio 2020.

Successivamente è stata fatta una analisi più approfondita per entrambi i dataset inerente alle classificazioni effettuate dagli annotatori.

[10] Il 30 ottobre 2019, l'Assemblea del Senato ha approvato la mozione per l'istituzione della Commissione con 151 voti favorevoli, nessun contrario, ma 98 astensioni (http://www.senato.it/leg/18/BGT/Schede/Commissioni/0-00143.htm). Per la prima volta dal dopoguerra la memoria della Shoah, implicita nella proposta di Liliana Segre, è stata messa in discussione non tanto nella sostanza, quanto a causa dell'attribuzione di un significato politico alla memoria. In un certo senso, l'Olocausto è stato considerato come una memoria di parte, secondo un processo già iniziato da anni ad esempio con l'istituzione della Giornata in memoria delle vittime delle foibe, creata come una sorta di *contromemoria*.

Il primo step è stato quello di definire se il testo del tweet contenesse odio o meno. I risultati ottenuti sono rappresentati in Figura 3 e Figura 4 e si riferiscono ai dataset campionari contenenti 900 tweet ciascuno. Dalla Figura 3, riferita a "ebrei-antisemitismo", emerge un andamento non lineare del contenuto di odio; nella Figura 4, riferita a "ebrei-coronavirus", si può invece vedere un andamento crescente della percentuale di odio nei tre mesi presi in considerazione.

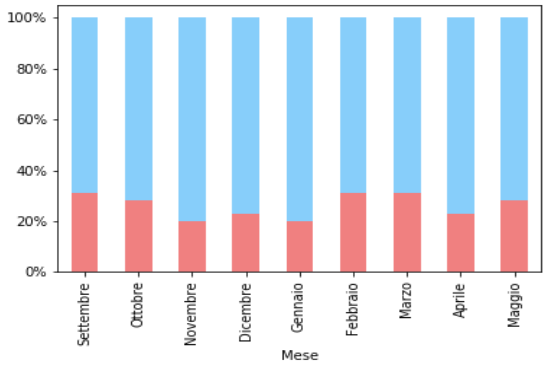

Figura 3. Percentuale di tweet contenenti odio (in rosso) nel dataset campionario relativo a ebrei-antisemitismo.

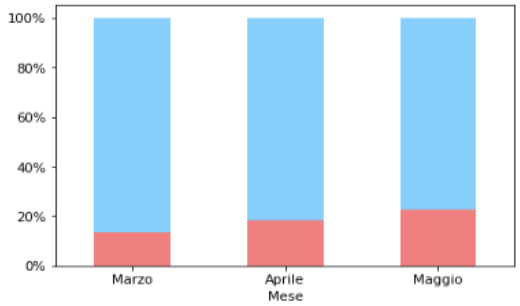

Figura 4. Percentuale di tweet contenenti odio (in rosso) nel dataset campionario relativi all'associazione ebrei-coronavirus.

Una volta stabilito che il tweet contenesse odio, il secondo step dell'analisi è stato quello di classificare le retoriche e le forme antisemite proprie del testo. I risultati si riferiscono quindi ai 235 tweet contenenti odio (26.11%) per il primo dataset e ai 147 tweet (16.3%) per il secondo.

La Figura 5 e la Figura 6 mostrano, attraverso un grafico a barre, le percentuali dei tweet con odio divisi per retorica, la prima inerente al dataset campionario riferito a ebrei-antisemitismo, la seconda inerente al dataset campionario riguardante ebrei-coronavirus. Come si può vedere, la retorica "pregiudizio" è predominante in entrambi i dataset e la retorica "insulti" caratterizza il 21% dei tweet d'odio del primo dataset. Le altre retoriche sono la minoranza.

La Figura 7 e la Figura 8 mostrano, attraverso un grafico a barre, le percentuali dei tweet con odio divisi per forma antisemita, la prima inerente al dataset campionario riferito a ebrei-antisemitismo, la seconda inerente a ebrei-coronavirus. Si rileva che la forma "potere finanziario" è predominante in entrambi i dataset e la forma "antisionismo" caratterizza il 19% dei tweet contenenti odio del secondo dataset. Le altre forme antisemite sono la minoranza.

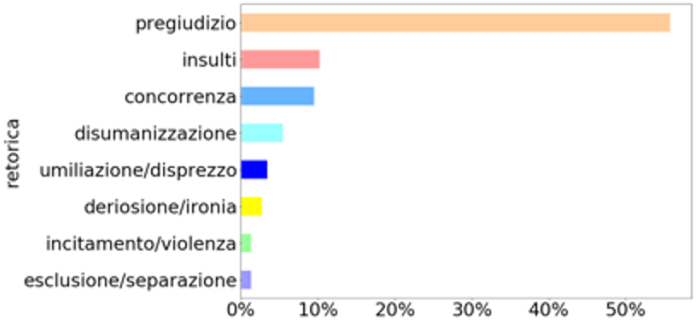

Figura 5. Retoriche dei tweet contenenti odio nel dataset campionario relativo a ebrei-antisemitismo.

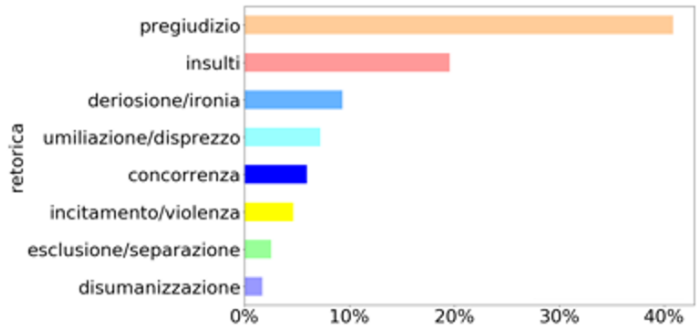

Figura 6. Retoriche dei tweet contenenti odio nel dataset campionario relativo a ebrei-coronavirus.

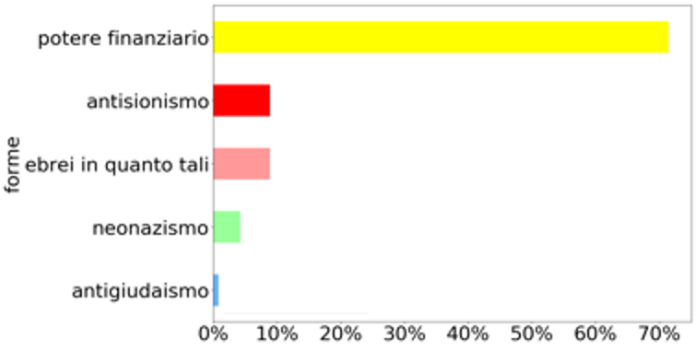

Figura 7. Forme di antisemitismo dei tweet contenenti odio nel dataset campionario relativo a ebrei-antisemitismo.

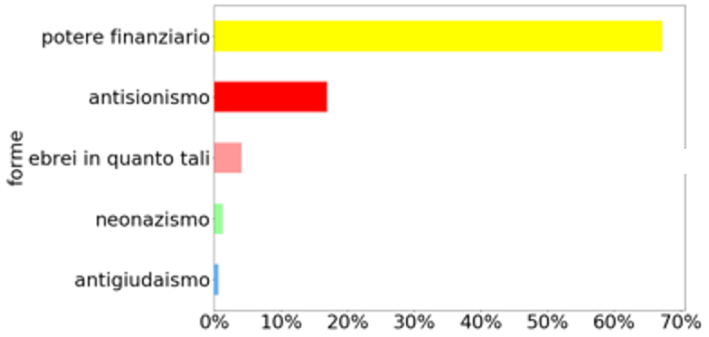

Figura 8. Forme di antisemitismo dei tweet contenenti odio nel dataset campionario relativo all'associazione ebrei-coronavirus.

Per fare un confronto più dettagliato delle retoriche e delle forme antisemite che sono state utilizzate nei due dataset, è stata fatta un'ulteriore analisi con la tecnica dei radarchart, un metodo grafico di visualizzazione dei dati multivariati (Adams, 2014).

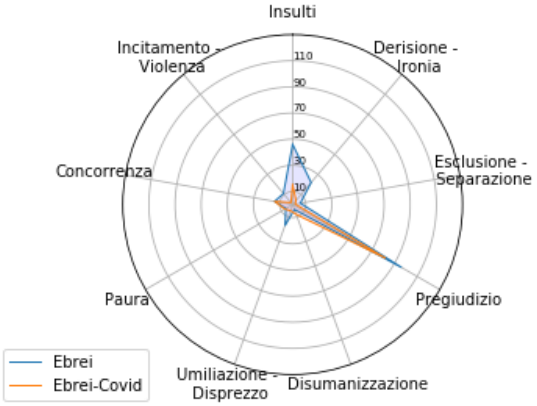

Figura 9. Radarchart delle retoriche di entrambi i dataset campionari analizzati.

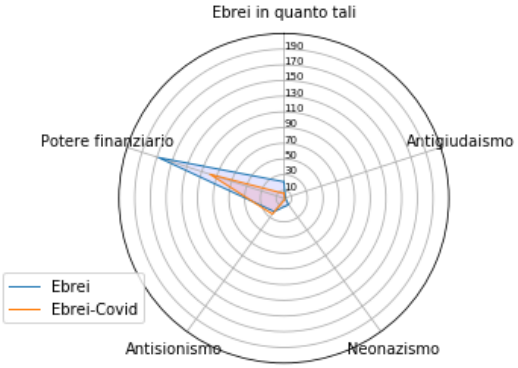

Figura 10. Radarchart delle forme antisemite di entrambi i dataset campionari analizzati.

La Figura 9 e la Figura 10 mostrano rispettivamente il radarchart delle retoriche e quello delle forme antisemite. La Figura 9 mostra come nel dataset relativo all'associazione ebrei-coronavirus siano diminuiti i contenuti con "insulti", "derisione-ironia" e "umiliazione-disprezzo", mentre sono aumentati i contenuti con "concorrenza" e "disumanizzazione". La Figura 10, invece, mostra come nel dataset ebrei-coronavirus siano diminuiti i contenuti di odio verso gli "ebrei in quanto tali" e le forme di "neonazismo" e come al contrario siano aumentati i contenuti con forme di "antisionismo". La percentuale di forme antisemite relative al "potere finanziario" è rimasta invariata.

L'ultimo step di questa analisi è stato quello di trovare un sistema in grado di annotare automaticamente i messaggi pubblicati, fornendo una prima classificazione di odio/non-odio ed eventualmente la tipologia di retorica e forma antisemita corrispondente. Tutti i testi classificati dagli annotatori sono stati valutati anche da un algoritmo in grado di stabilire se il tweet contenesse odio, prendendo in considerazione solamente le radici delle parole. Dopo aver applicato una serie di procedure tipiche del *Natural Language Processing* (NLP) – rimozione dei caratteri superflui (slash, punteggiatura, html, etc.), conversione del testo in minuscolo e rimozione degli *stopwords* –, l'algoritmo classifica il testo *pulito* in base alle radici delle parole che contiene (Bird, Klein, & Loper, 2009). Per la classificazione è stato creato il dizionario di parole negative proposto nella ricerca

effettuata da DataMediaHub e KPI6 (2020). I risultati ottenuti sono mostrati nelle matrici di confusione in Figura 11 e in Figura 12 (James, Witten, Hastie, & Tibshirani, 2017). Come si può vedere, l'algoritmo che tiene in considerazione solo le radici delle parole non è ben performante nell'individuazione dei contenuti di odio: nel primo dataset, dei 235 tweet (221+14) che sono stati classificati come contenuto di odio dagli annotatori, l'algoritmo ne individua correttamente solo 14. Nel secondo dataset, analogamente, dei 147 tweet (137+10) che sono stati classificati come contenuto di odio dagli annotatori, l'algoritmo ne individua correttamente solo 10.

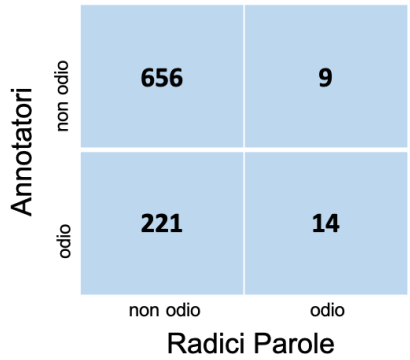

Figura 11. Confusion matrix delle classificazioni fatte manualmente dagli annotatori e quelle ottenute considerando le radici del dataset relativo a ebrei-antisemitismo.

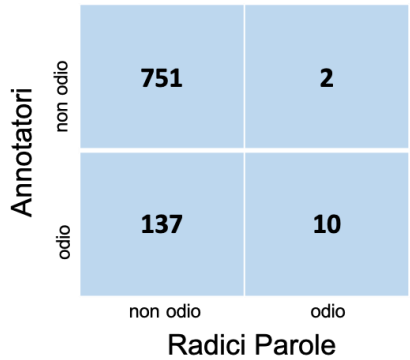

Figura 12. Confusion matrix delle classificazioni fatte manualmente dagli annotatori e quelle ottenute considerando le radici del dataset relativo a ebrei-coronavirus.

Nella Figura 13 sono elencati alcuni esempi di tweet e le relative classificazioni ottenute con i due metodi[11].

Tweet	Annotatori	Radici Parole
#PapaFrancesco inaugura una grande statua in bronzo dedicata ai #migranti. #29settembre 140 persone di diversi luoghi e epoche storiche: ci sono gli indigeni e gli ebrei perseguitati dalla Germania nazista, i siriani che scappano dalla guerra e gli africani che fuggono la fame.	No odio	No odio
@nakamuraelias il nome deriva dall'aramaico eliyáhút o eliyyah e significa "il vero dio è yahvé/yahvè è il mio signore". fu il primo profeta di israele.	No odio	No odio

[11] I tweet sono riportati in forma originale, comprensivi di errori ortografici e sintattici.

Secondo te sa cosa sia la bossi - fini?!? Come minimo pensa che i migranti li abbia inventati Soros.	No odio	No odio
Le armi facili fanno strage di palestinesi in Israele | il manifesto #ilmanifesto	No odio	No odio
Sarà anche uomo di #soros, ma soprattutto è uomo di merda.	Odio	Odio
Siiiiiiiiii comanda e dietro di lui SOROS massoneria e banche!!! che Dio lo possa punire ha venduto L'Italia e gli italiani lui comanda MATTARELLA e il governo	Odio	Odio
Ti regalo una tanica di vaselina da spartirti con gli amici tuoi ebrei che per quanto mi riguarda non sono migliori della Germania nazista di hitler sempre di LURIDI criminali si tratta	Odio	Odio
Un altro maiale come soros.. Maledetti bastardi.	Odio	Odio
Stramaledetti tutti i comunisti che li hanno portati nel nostro paese. In primis Renzi, poi Gentiloni con la benedizione di Soros. Migranti, Viminale: rivolta al Cpr di Torino, ferito un poliziotto	Odio	No odio
Ao non ie lo dite a quel turbocazzaro di Fusaro che ve paga Soros, senno' inizia un pippoto inconcludente che non la finisce piu'	Odio	No odio
Cara Chiesa sei la prostituta vestita di bianco di cui si parla nell'apocalisse di Giovanni. I tuoi clienti sono Merkel, Macron, Soros, Rochfeller e famiglie annesse e connesse. Papa nero gesuita. Ricorda che il diavolo (che avete inventato voi) fa le pentole, ma noi i coperchi.	Odio	No odio
Fatemi capire, Hitler si era affidato agli astri e ha sbagliato, sterminando 6 milioni di ebrei. E se ci avesse azzeccato invece?	Odio	No odio
Solo nel 2019 ci sono stati più di 250 episodi antisemiti. Canali Youtube che denunciano complotti giudaico-massonici e documenti falsi in cui si accusano gli #ebrei di pedofilia. L'odio non è un problema di destra o sinistra, ma un allarme per tutti	No odio	Odio
E allora quando i nazisti dicevano "sporchi ebrei" intendevano solo dire "quelli sudici tra gli ebrei" vero? Non prendiamoci per il culo, dai. La malafede è peggio dell'ignoranza grammaticale...	No odio	Odio
"Froci", "ebrei", "negri", "la guerra razziale è iniziata": i nazisti nelle chat lette dall'Espresso https://ift.tt/2UT9P7c	No odio	Odio
Quanto manca in Italia per vedere le stesse scene che in Israele e USA sono all'ordine del giorno, ovvero l'infamia di arrestare bambini? Abbiamo visto elicotteri rincorrere liberi cittadini come fossero criminali manca poco per vederli piombare sopra bambini che giocano a palla.	No odio	Odio

Figura 13. Esempi di tweet e classificazione degli esperti annotatori e classificazione automatica con radici parole.

5. Conclusioni

In quest'ultima parte dell'articolo si presenteranno due considerazioni tratte dall'analisi dei risultati dello studio, una relativa ai contenuti antisemiti online in lingua italiana diffusi nel periodo settembre 2019-maggio 2020 e una di carattere metodologico rispetto alle prospettive di ricerca automatica e algoritmica per individuare l'odio online.

5.1. Un antisemitismo opportunistico

Le classificazioni fatte manualmente dagli annotatori sono state essenziali per poter rispondere alle due domande di ricerca poste inizialmente dal caso studio. Per rispondere alla prima questione sull'eventuale aumento di odio contro gli ebrei durante l'emergenza sanitaria, si consideri la Figura 3: nel dataset contenente i post antisemiti senza un riferimento esplicito al Covid-19, nel periodo marzo-maggio 2020 non emerge un aumento di odio. Infatti, la percentuale di odio registrata a marzo (31%) è uguale a quella riscontrata nei mesi di febbraio e settembre (31%).

La risposta alla seconda domanda di ricerca, "Quali retoriche e forme di odio ci sono nell'associazione ebrei e coronavirus?" si può invece trovare nella Figura 9 e Figura 10. Per quanto riguarda le forme antisemite, nei tweet relativi all'associazione ebrei e Coronavirus, emerge un aumento del pregiudizio (+18%), della concorrenza (+5%) e della disumanizzazione (+4%), al contrario vi è una diminuzione degli insulti (-10%), della derisione/ironia (-7%), della violenza (-3%) e umiliazione/disprezzo (-4%). Tra le forme antisemite sono prevalenti l'evocazione del potere ebraico sulla finanza (74%) e l'antisionismo/odio verso lo stato di Israele (19%), mentre risultano minoritari i tweet legati alle forme di odio antisemita tradizionali come verso gli ebrei in quanto tali (5%), neonazismo e neofascismo e negazionismo della Shoah (2%), antigiudaismo (1%).

A livello mondiale, come nota il rapporto "Coronavirus and the plague of antisemitism" dell'inglese Community Security Trust di Londra (2020), è circolata l'idea che: a) il virus è una fake, una cospirazione ebraica; b) il virus è reale ma è sempre frutto di un complotto; c) il virus ha vita autonoma ma è diffuso dagli untori ebrei; d) il virus dovrebbe essere diretto contro gli ebrei. Un Rapporto del Kantor Center della Tel Aviv University (2020) scrive: "L'antisemitismo generato dal coronavirus è intenso e feroce è propagato soprattutto dalla destra estremisti, cristiani ultraconservatori e islamisti, attraverso i propri media in varie lingue" (p. 2). Come un parassita, l'antisemitismo riemerge in occasione di una crisi e sparge odio in un organismo indebolito. Le teorie complottiste confermano in modo ossessivo che dietro ogni minaccia ci devono essere per forza *loro*. È interessante però notare che, come sempre, il discorso antisemita è profondamente incoerente e contraddittorio, e allo stesso tempo capace di adattarsi, con la ripetizione del suo schema narrativo, ai cambiamenti storici. Il legame tra la pandemia e il mondo ebraico afferma tutto e il contrario di tutto, in modo illogico: gli ebrei sono vittime e persecutori, tramano nell'ombra inventando minacce oppure sono loro a crearle, sarebbero ricchi dominatori, capitalisti, complottano nell'ombra, ma anche miseri, parassiti, inutili. L'odio ancestrale si adatta ad ogni circostanza in qualsiasi periodo storico, approfittando di un pubblico angosciato e arrabbiato, indebolito dalle crisi (Santerini, 2005). Da parte nazista furono accusati di essere allo stesso tempo bolscevichi e liberali, conservatori e materialisti e quindi "rappresentavano nemici diversi e spesso opposti" (Confino, 2017, p. 43); in questo modo diventa davvero pericoloso un nemico che somma in sé simultaneamente tutti i diversi nemici agli occhi di persone differenti.

Quanto al caso italiano analizzato, l'ostilità online durante il periodo della pandemia non ha registrato picchi significativi, forse perché parallelamente è stata re-diretta più verso i cinesi, accusati di aver propagato il virus. Un'osservazione simile è stata rilevata anche in Australia dove l'ostilità si è canalizzata soprattutto verso i cittadini asiatici (Chew, 2020).

Rispetto alle due tematiche prevalenti, l'odio verso il potere finanziario ebraico di tipo cospiratorio si canalizza verso George Soros[12], citato come il grande vecchio della politica

[12] George Soros è un finanziere e filantropo ebreo ungherese naturalizzato statunitense, presidente del Soros Fund, dell'Open Society Foundations e fondatore e consigliere del Quantum Group.

mondiale secondo il più classico degli stereotipi (sostituisce Rothschild). La fisionomia di Soros oggi invade il Web, comunicando l'equazione tra cosmopolitismo pro-immigrati antinazionalista e l'ebreo di sempre, evocando il pregiudizio dell'ebreo senza patria, *sradicato* e quindi infido e traditore. Nella versione del XXI secolo, il capitalista globale è favorevole agli immigrati e minaccia all'identità nazionale dietro la copertura umanitaria.

Quanto al secondo tema, la demonizzazione di Israele, a differenza che in altri paesi, prevale non la versione strettamente cospirativista (ebrei che producono o diffondono il virus), quanto la versione che vede Israele, o in generale gli ebrei, che sfruttano intenzionalmente la pandemia per motivi politici o economici. Questo almeno in parte coincide con i risultati di "Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England" (Freeman et al., 2020), secondo cui, a maggio 2020, il 20% degli inglesi crede che il virus sia stato creato dagli ebrei per sfruttamento economico (il 45% crede che sia un'arma biologica creata dalla Cina per distruggere l'Occidente).

Pertanto, se *opportunistica* può essere chiamata un'infezione causata da agenti patogeni (batteri, virus) in organismi caratterizzati da un sistema immunitario compromesso, in questo senso l'antisemitismo online registrato in Italia nei mesi della pandemia può essere definito, appunto, di tipo opportunistico, in quanto riemerge col suo carattere del tutto irrazionale e incoerente non tanto per alludere a una cospirazione ebraica che volutamente diffonde il virus, quanto per denigrare il mondo ebraico attribuendogli il suo sfruttamento economico o politico.

5.2. La necessità di contestualizzare le parole

La seconda considerazione è invece di carattere metodologico. In questo lavoro è stato annotato manualmente un corpus di tweet in lingua italiana per riuscire a studiare le caratteristiche dell'hate speech verso un target specifico. L'analisi non si è focalizzata solamente sull'individuare la presenza o meno di odio, ma si è voluto studiarne la tipologia analizzando le retoriche e le forme corrispondenti.

Tuttavia, si è detto della forte discordanza tra il processo di annotazione manuale, detto anche di etichettatura, che richiede la collaborazione con esperti di dominio e quella prodotta dall'algoritmo creato con il dizionario di parole negative di un'altra ricerca (DataMediaHub & KPI6, 2020). Questa alta percentuale di errore prodotta dall'algoritmo conferma come la sola analisi semantica non sia sufficiente per riuscire ad automatizzare correttamente un processo così complesso. È necessario possedere una conoscenza della realtà e quindi del contesto in cui si trova, fatto verificabile dagli annotatori.

Da un lato, come fa notare Jason Stanley (1994), una proprietà degli *slurs* (le *parole per ferire*) è quella di mantenere il loro significato dispregiativo indipendentemente dalla forma dell'enunciato in cui compaiono (Wodak & Meyer, 2011). Eppure, d'altro canto, come ha dimostrato Tullio De Mauro in "Le parole per ferire" (2016)[13], sono i termini stessi di uso

---

Sostenitore di varie ONG e progetti a favore della democrazia e dei diritti umani, diviene nel discorso antisemita il simbolo del potere ebraico, con un'idea complottista sempre presente ma acuita dalle visioni che demonizzano la globalizzazione. Soros viene considerato di volta in volta l'organizzatore occulto di complotti finanziari, ma anche di un piano di sostituzione della popolazione europea con gli immigrati.

[13] Si veda la Relazione finale della Commissione Jo Cox sull'intolleranza, la xenofobia, il razzismo e i fenomeni di odio della Camera dei deputati (2017).

normale che possono essere usati in modo ostile; Federico Faloppa (2019; 2020) nota come online sia possibile ricorrere ad altri mezzi, come grafismi, font, metafore, stereotipi e pregiudizi, che acquistano particolare rilevanza quando possono poggiare sul meccanismo dell'*othering*, ovvero su un insieme di dinamiche, processi, strutture, anche linguistiche, che raggruppano dialetticamente i soggetti in un *noi* e *loro* in gruppi presentati come omogenei e alternativi gli uni agli altri (Powell & Menendian, 2018), e strategie di *perspectivation* di polarizzazione noi vs. loro (Graumann & Kallmeyer, 2002). Allo stesso modo, Stefano Pasta (2018) mostra come l'odio online si annidi spesso in *pedagogie popolari implicite*, filosofie educative legate a una visione interpretativa del mondo (l'altro come nemico, la gerarchia dei gruppi, etc.).

Una delle caratteristiche che rendono l'hate speech così sfuggente è la difficoltà di coglierne il significato quando un discorso è privo di contesto. La logica algoritmica, come ha dimostrato Jerome Bruner (1968; 1988), si occupa di informazioni già codificate, il cui significato è stabilito in anticipo. La logica computazionalista si interessa di stimoli e risposte, non del senso da attribuire alle cose, elabora informazioni, mentre chi fa cultura ed educazione interpreta e produce significato, operazione carica di ambiguità e soprattutto sensibile al contesto.

Pertanto, se si vuole cercare sul Web l'hate speech non basteranno gli algoritmi dato che discorsi e narrazioni dell'odio sono celati, mascherati sotto un lessico ordinario. Il linguaggio d'odio cerca dunque di *mimetizzarsi* e non si esprime necessariamente con un lessico basato su veri e propri insulti, ma può invece servirsi di repertori e registri diversi.

Per questa ragione sono quindi parziali gli studi che individuano i veri e propri insulti, o che rilevano la sola presenza di *hate words*, e diventa quindi contestabile la tesi di una marginalità del discorso d'odio in Twitter, proposta dalle ricerche che usano questa metodologia (DataMediaHub & KPI6, 2020). Al contrario, proseguendo nell'obiettivo di automatizzare un processo così complesso come la ricerca di odio online, per verificarne la possibilità, occorre sperimentare la metodologia qui descritta per l'antisemitismo ad un set di dati molto più ampio e di differenti gruppi bersaglio. È dunque opportuno sperimentare ricerche che integrino le due fasi di classificazione umana e automatica, applichino approcci interdisciplinari e partano da una conoscenza approfondita delle manifestazioni dei fenomeni al centro dell'indagine.

Riferimenti bibliografici

ADL. *Anti-Defamation League*. https://www.adl.org (ver.15.12.2020).

Adams, C. (2014). *Learning Python Data Visualization.* Birmingham: Packt Publishing.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* Sebastopol, CA: O'Reilly Media.

Bruner, J. S. (1968). *Studi sullo sviluppo cognitivo.* Roma: Armando (Original work published 1966).

Bruner, J. S. (1988). *La mente a più dimensioni.* Roma-Bari: Laterza (Original work published 1986).

Camera dei deputati (2017). *Commissione "Jo Cox" sull'intolleranza, la xenofobia, il razzismo e i fenomeni. Relazione finale.* Roma https://www.camera.it/application/xmanager/projects/leg17/attachments/uploadfil

e_commissione_intolleranza/files/000/000/001/RELAZIONE_FINALE.pdf (ver.15.12.2020).

Centro di Ricerca sulle Relazioni Interculturali (2020). *Report Mediavox 2019-2020. Ricerca-azione sui discorsi d'odio online di natura antireligiosa.* Milano: Università Cattolica del Sacro Cuore. www.mediavox.network (ver.15.12.2020).

Chew, E. W. A. (2020). *COVID-19 Racism Incident Report. Reporting Racism Against Asians in Australia Arising due to the COVID-19 Coronavirus Pandemic*. Penrith: Asian Australian Alliance Pty Ltd. http://diversityarts.org.au/app/uploads/COVID19-racism-incident-report-Preliminary-Official.pdf (ver.15.12.2020).

Community Security Trust (2020). *Coronavirus and the plague of antisemitism*. Londra: CST. https://cst.org.uk/data/file/d/9/Coronavirus%20and%20the%20plague%20of%20antisemitism.1586276450.pdf (ver.15.12.2020).

Confino, A. (2017). *Un mondo senza ebrei. L'immaginario nazista dalla persecuzione al genocidio.* Milano: Mondadori (Original work published 2014).

Damasio, A. (1995). *L'errore di Cartesio*. *Emozione, ragione e cervello umano*. Milano: Adelphi (Original work published 1994).

DataMediaHub, & KPI6 (2020). *Report su Hate Speech in Italia su Twitter*. http://www.datamediahub.it/2020/06/22/rapporto-sullhate-speech-in-italia/#axzz6cYynj6H0 (ver.15.12.2020).

De Mauro, T. (27 settembre 2016). Le parole per ferire. *Internazionale*. https://www.internazionale.it/opinione/tullio-de-mauro/2016/09/27/razzismo-parole-ferire (ver.15.12.2020).

ECRI. European Commission against Racism and Intolerance (2015). *ECRI general policy, Recommendation no. 15 on combating hate speech*. Strasbourg: Council of Europe. https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 (ver. 15.12.2020).

Faloppa, F. (2019). *Brevi lezioni sul linguaggio.* Torino: Bollati Boringhieri.

Faloppa, F. (2020). *#ODIO. Manuale di resistenza alla violenza delle parole*. Milano: Utet.

Femia, D. (2019). Discorso dell'odio e risorse per il trattamento automatico delle lingue. Metodi, ipotesi, proposte. In R. Petrilli (Ed.), *Hate speech. L'odio nel discorso pubblico. Politica, media, società* (pp. 147-164). Roma: Round Robin.

Ferrini, C., & Paris, O. (2019). *I discorsi dell'odio. Razzismo e retoriche xenofobe sui social network*. Roma: Carocci.

Floridi, L. (2017). *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo.* Milano: Raffaello Cortina (Original work published 2014).

James, G., Witten, D., Hastie, T., & Tibshirani, T. (2017) *An Introduction to Statistical Learning with Applications in R.* New York, NY: Springer.

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., Jenner, L., Teale, A. L., Carr, L., Mulhall, S., Bold, E., & Lambe, S. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England.

Psychological Medicine, 1–13. https://doi.org/10.1017/S0033291720001890 (ver.15.12.2020).

Graumann, C. F., & Kallmeyer, W. (Eds.). (2002). *Perspective and Perspectivation in Discourse*. Amsterdam: John Benjamins.

IHRA. *International Holocaust Remembrance Alliance*. https://www.holocaustremembrance.com (ver.15.12.2020).

IHRA. International Holocaust Remembrance Alliance (2016). *Working Definition of Antisemitism.* https://www.holocaustremembrance.com/working-definition-antisemitism (ver.15.12.2020).

Kantor Center (2020). *The COVID-19 pandemic has unleashed a unique worldwide wave of antisemitism*. Tel Aviv: Tel Aviv University. https://osservatorioantisemi-c02.kxcdn.com/wp-content/uploads/2020/06/covid_june24.pdf (ver.15.12.2020).

Mozione 1-0013 del Senato della Repubblica, 30 ottobre 2019. *Commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza*. http://www.senato.it/leg/18/BGT/Schede/Commissioni/0-00143.htm (ver.15.12.2020).

Nussbaum, M. (2004). *L'intelligenza delle emozioni*. Bologna: il Mulino (Original work published 2001).

Osservatorio antisemitismo della Fondazione CDEC (2020). *Antisemitismo in Italia 2019*. Milano: Centro di Documentazione Ebraica Contemporanea. https://www.osservatorioantisemitismo.it/approfondimenti/relazione-annuale-sullantisemitismo-in-italia-2019-a-cura-dellosservatorio-antisemitismo-della-fondazione-cdec/ (ver.15.12.2020).

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

Pasta, S. (2019). Conversazioni via social network con giovani autori di performances d'odio. Social network conversations with young online authors of hate speech. *Pedagogia Oggi*, *XVIII*(2), 369 –383.

Pasta, S. (2020). *(S)parlare nel Web. Razzismo online ed educazione alla cittadinanza*. Milano: Ismu.

Python GetOldTweets3. https://pypi.org/project/GetOldTweets3 (ver.15.12.2020).

Powell, J., & Menendian, S. (2018). The Problem of Othering: Towards Inclusiveness and Belonging. *Othering and Belonging Expanding the circle of human concern.* https://www.otheringandbelonging.org/the-problem-of-othering/ (ver.15.12.2020).

Raccomandazione 97/20 del Consiglio d'Europa, 30 ottobre 1997. *Hate Speech*. https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680505d5b (ver.15.12.2020).

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, Japan.

Santerini, S. (2005). *Antisemitismo senza memoria. Insegnare la Shoah nelle società multiculturali.* Roma: Carocci.

Santerini, S. (Ed.). (2019a). *Il nemico innocente. L'incitamento all'odio nell'Europa contemporanea*. Milano: Guerini e Associati.

Santerini, S. (2019b). Discorso d'odio sul web e strategie di contrasto. *Metis, 9(2)*, 51–67.

Santerini, S. (2021). La mente ostile. Forme dell'odio contemporaneo. Milano: Raffaello Cortina.

Stanley, J. (1994). *How Propaganda Works.* Princeton: Princeton University Press.

Taguieff, P. A. (2016). *L'antisemitismo*. Milano: Raffaello Cortina (Original work published 2015).

Vox-Osservatorio italiano sui diritti (2019). *Quarta Mappa sull'Intolleranza in Italia*. http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-4/ (ver.15.12.2020).

Wallace, P. (2007). *La psicologia di Internet*. Milano: Raffaello Cortina (Original work published 2005).

Wieviorka, M. (2019). La haine raciste et antisémite: infra-politique, méta-politique et… politique. In M. Santerini (Ed.), *Il nemico innocente. L'incitamento all'odio nell'Europa contemporanea* (pp. 65-74). Guerini e Associati: Milano.

Wodak, R., & Meyer, M. (Eds.). (2011). *Methods of Critical Discourse Analysis.* London: Sage.

Ziccardi, A. (2016). *L'odio online. Violenza verbale e ossessioni in rete*. Milano: Raffaello Cortina.

Ziccardi, A. (2019). *Tecnologie per il potere. Come usare i social network in politica.* Milano: Raffaello Cortina.