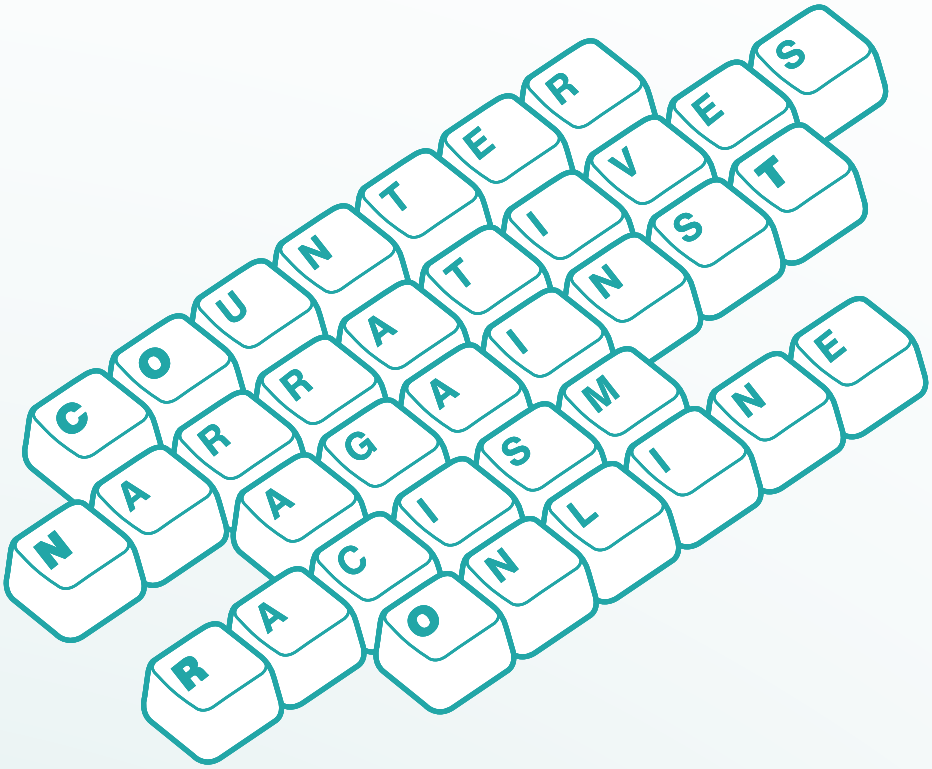
Presidenza del Consiglio dei Ministri
Dipartimento per le Pari Opportunità

unar
Ufficio Nazionale Antidiscriminazioni Razziali
a difesa delle differenze

RIGHT BY YOUR SIDE

CO.N.T.R.O.!



Il contributo del progetto CO.N.T.R.O. all'analisi del fenomeno dell'odio online e alla definizione di possibili soluzioni utili a contrastarlo

ISTITUTO PER LA RICERCA SOCIALE
irs

Co-funded by the European Union

Questo rapporto è stato redatto da un gruppo di lavoro IRS coordinato da Flavia Pesce e composto da Daniela Loi, Elena Ferrari e Emma Paladino che si sono avvalsi del supporto di alcuni collaboratori: Davide Coero Borga, Daria Denti, Pasquale Pavone.

Il contenuto di questa pubblicazione rappresenta solo il punto di vista degli autori ed è esclusiva responsabilità degli stessi. La Commissione Europea non si assume alcuna responsabilità per l'uso che può essere fatto delle informazioni in essa contenute.

Questa pubblicazione è stata finanziata nell'ambito del Programma REC (Rights, Equality and Citizenship Programme) dell'Unione Europea (2014-2020).

Indice

Indice di box, figure e tavole

Premessa

L'Ufficio per la promozione della parità di trattamento e la rimozione delle discriminazioni fondate sulla razza o sull'origine etnica, brevemente denominato UNAR – Ufficio Nazionale Antidiscriminazioni Razziali, è l'ufficio deputato dallo Stato italiano a garantire il diritto alla parità di trattamento di tutte le persone, indipendentemente dalla origine etnica o razziale, dalla loro età, dal loro credo religioso, dal loro orientamento sessuale, dalla loro identità di genere o dal fatto di essere persone con disabilità. L'Ufficio è stato istituito nel 2003 (d.lgs. n. 215/2003) in seguito a una direttiva comunitaria (n. 2000/43/CE), che impone a ciascun Stato Membro di attivare un organismo appositamente dedicato a contrastare le forme di discriminazione.

L'UNAR, da diversi anni, segue con attenzione e preoccupazione la crescita pervasiva del discorso d'odio on-line. I tre compiti fondamentali che la normativa gli affida, dal prevenire e rimuovere la discriminazione, alla tutela delle vittime passando per la comprensione del fenomeno discriminatorio *tout court*, in riferimento all'*hate speech* si intrecciano inesorabilmente in un rapporto di interdipendenza.

Quali sono gli effetti dei discorsi d'odio a carattere discriminatorio sui gruppi target a cui questo odio è diretto? Esiste una correlazione, un nesso di causalità-effetto tra ciò che avviene online e gli eventi, altrettanto violenti, che prendono forma nel mondo reale? Individuare e rimuovere le migliaia di contenuti che ogni minuto vengono postati sui social media può risultare efficace senza comprenderne al contempo motivazioni e modalità, in definitiva senza elaborare una strategia educativa e comunicativa di risposta?

Risulta di immediata evidenza la difficoltà di rispondere alle domande appena accennate. Come non è agevole ricostruire il dibattito scientifico che si è sviluppato attorno ad una questione complessa qual è l'aggressività verbale e la violenza che circola sul Web. Molteplici temi si intrecciano in una discussione che coinvolge studiosi ed esperti di estrazione disciplinare differente. Bisogna quindi procedere con ordine e cautela per evitare di sovrapporre diversi piani di analisi; si è difatti in presenza di un fenomeno in parte conosciuto, in parte inedito. L'odio è moneta corrente nei rapporti sociali, sin dalla notte dei tempi, ben prima dell'avvento di Internet. Non è un caso che nella home page di uno dei maggiori portali di studio sull'argomento[1] campeggi un libro del 1954, *La natura del pregiudizio*, opera fondamentale di Gordon Allport, psicologo sociale americano che, poco meno di un decennio dopo la fine della Seconda guerra mondiale, aveva elaborato una scala per misurare il livello di discriminazione/pregiudizio esistenti nella società del tempo.

Nella piena consapevolezza della complessità del fenomeno risiede la scelta dell'UNAR di estendere e approfondire il proprio livello di analisi (e quindi di intervento). Il progetto CO.N.T.R.O "Counter Narratives AgainsT Racism Online", in partneriato con l'Istituto per la Ricerca Sociale si inserisce in questa strategia, cogliendo alcuni degli aspetti fondamentali del dibattito, con l'obiettivo di aggiornare e sviluppare pratiche e strumenti per prevenire e combattere efficacemente il razzismo, la xenofobia e altre forme di intolleranza diffuse attraverso discorsi di incitamento all'odio (*hate speech*) online.

Dal considerare l'odio online come un fenomeno in evoluzione continua, in linea con la rapidità a cui la trasformazione digitale ci ha abituato in ogni settore, deriva l'esigenza per le istituzioni di affrettare il proprio passo, cogliendo e valorizzando rapidamente quanto fatto fino ad ora, dalla società civile all'accademia. In questo primo e corposo rapporto del Progetto si dà ampiamente conto di questa consapevolezza e di questa accelerazione.

Roberto Bortone
Coordinatore scientifico del progetto

1 Si tratta dell'International Network for Hate Studies (INHS), rete di studiosi accademici, ricercatori e operatori che funge da archivio on line delle pubblicazioni sugli hate studies, promuovendo scambi interdisciplinari e dibattiti scientifici fra gli esperti del tema (*http://www.internationalhatestudies.com*).

Introduzione

Il presente rapporto si inserisce nelle attività promosse dal progetto CO.N.T.R.O. - "*Counter Narratives AgainsT Racism Online*" per aggiornare e sviluppare pratiche e strumenti per prevenire e combattere efficacemente il razzismo, la xenofobia e altre forme di intolleranza diffuse attraverso discorsi di incitamento all'odio (*hate speech*) online.

Promosso e finanziato dalla Commissione europea, CO.N.T.R.O. è un progetto ideato e coordinato da UNAR (Ufficio Nazionale Antidiscriminazioni Razziali), in partenariato con IRS (Istituto per la Ricerca Sociale).

L'*hate speech* online è un fenomeno in forte crescita che sfrutta la rete per diffondere i propri messaggi in maniera veloce e pervasiva. Nel tentativo di fornire risposte sempre più adeguate ed efficaci alle varie forme di discriminazione presenti in rete, sono nate molte iniziative di contrasto, prevenzione ed informazione sui discorsi di odio.

Il progetto CO.N.T.R.O. si inserisce in questo contesto, ponendosi come obiettivo generale quello di contribuire al contrasto del razzismo, della xenofobia e di altre forme di intolleranza diffuse tramite discorsi di incitamento all'odio online. Attraverso una prima fase di studio e ricerca sui discorsi di odio online e sulle migliori strategie in uso per contrastarli, il progetto ha promosso una mirata campagna di comunicazione e sensibilizzazione sul fenomeno e ha messo in atto le condizioni per lo sviluppo dell'Osservatorio Media e Internet dell'Ufficio Nazionale Antidiscriminazioni Razziali (UNAR) coinvolgendo i principali attori istituzionali e non coinvolti dal fenomeno.

Questo rapporto, a partire dalle attività progettuali realizzate[1], intende presentare i principali risultati ottenuti e i possibili sviluppi futuri.

Nello specifico, il primo capitolo delinea l'ambito di analisi - l'odio online - fornendo informazioni sulla sua rilevanza e definizioni utili per la sua comprensione. In particolare, l'evoluzione del dibattito attorno al tema dell'odio online viene affrontata mediante una disamina delle principali definizioni e inquadramenti giuridici nonché del ruolo della società civile e dei social network nel dibattito legislativo e nell'implementazione di politiche di contrasto.

Il secondo e il terzo capitolo presentano i risultati ottenuti dall'attività di ricerca svolta nel corso del progetto con il fine ultimo di predisporre una mappatura ed analisi:

1) dei principali approcci e metodologie di analisi per l'identificazione dei messaggi di odio veicolati dai social network;

2) di alcune significative esperienze e metodologie per la produzione di contro-narrative realizzate a livello nazionale, europeo ed internazionale a cui si è affiancato il racconto della genesi e del processo che ha caratterizzato l'attività di contro-narrativa del progetto con la realizzazione di una serie di video.

Gli ultimi due capitoli, infine, presentano il contributo del progetto per la finalizzazione di una metodologia comune di monitoraggio e analisi dei messaggi di odio online. In particolare il quarto capitolo, a partire da una breve introduzione all'analisi automatica dei testi, effettua una disamina critica dell'attuale metodologia utilizzata dall'Ufficio Nazionale Antidiscriminazioni Razziali (UNAR) al fine di evidenziare i possibili scenari di sviluppo dell'Osservatorio stesso che, nel quinto capitolo, vengono presentati mediante la definzione di alcune linee guida operative per lo sviluppo di un sistema di monitoraggio e contrasto dei messaggi di odio online a regia istituzionale.

1 Quanto svolto è presentato in maniera più dettagliata in singoli rapporti specifici disponibili sul sito dell'UNAR al seguente link: *http://www.unar.it/cosa-facciamo/azioni-positive-e-progetti/progetto-co-n-t-r-o/*

1

L’odio online: un fenomeno in continua crescita

1.1

L'odio online: definizioni e rilevanza nel dibattito italiano ed europeo

Gli ultimi anni sono stati caratterizzati da un'attenzione crescente da parte delle istituzioni verso i comportamenti devianti caratterizzati da violazioni violente delle norme sociali e aventi come obiettivi specifici gruppi sociali come ad esempio i migranti e le minoranze. L'aumento dell'interesse verso questi fenomeni è stato indubbiamente sollecitato da una propagazione estremamente rapida di questo tipo di comportamenti. Contestualmente si è diffusa anche la consapevolezza degli ingenti impatti negativi nei confronti delle vittime e della collettività nel suo complesso.

All'interno della categoria dei comportamenti oppressivi violenti, uno dei fenomeni che registra i più alti tassi di crescita è l'*hate speech* online, riconosciuto come una delle sfide più rilevanti poste dalle piattaforme di social media su Internet (EC, 2018; Brocato, 2016; ENAR, 2016; OSCE-ODIHR, 2010; CoE, 2004) e definito generalmente come: "*parole o simboli diffusi attraverso Internet, che sono dispregiativi e/o intimidatori in base alla razza, alla religione, all'orientamento sessuale e altri ambiti simili*" (McGonagle, 2013). I discorsi di odio online sono finalizzati a danneggiare, molestare, intimidire e umiliare gruppi specifici, promuovendo la violenza e l'insensibilità (Perry e Olsson, 2009). L'odio online si caratterizza, infatti, come azione oppressiva contro le minoranze configurabile come atto-messaggio: in altre parole, colpendo un singolo bersaglio si vuole colpire un intero gruppo sociale (generalmente una minoranza) (Gagliardone et al., 2015; Cohen-Almagor, 2014).

L'importanza crescente dell'*hate speech* online è supportata anche dalle cifre: in un recente speciale Eurobarometro, il **75%** degli intervistati ha dichiarato di aver assistito a discorsi di odio online su piattaforme social (Eurobarometro, 2016); negli Stati Uniti, la quota è pari al **66%** (PEW, 2017). Un recente rapporto dell'ECRI[1] sottolinea l'aumento senza precedenti dei discorsi d'odio a sfondo razziale (ECRI, 2016), confermato anche dallo *Shadow Report* 2016 dell'ENAR[2]. I dati mostrano anche grandi quote di discorsi d'odio online prodotte da persone estranee a qualsiasi gruppo di odio organizzato (Hall, 2013).

I trend sulla crescita del fenomeno hanno determinato in questi ultimi anni anche un incremento delle attività di ricerca sui temi dell'*hate speech* online in diverse discipline: dall'informatica alla criminologia, dall'economia alla psicologia (i.a. Müller e Schwarz, 2018a; Silva et al, 2016; Gagliardone et al., 2015; Hall, 2013). La rilevanza dell'*hate speech* è confermata anche dall'aumento delle iniziative istituzionali di natura legislativa, informativa e di prevenzione (Assimakopoulos et al., 2017).

Pur avendo diverse analogie con l'odio offline (finalità oppressiva, attitudine al rafforzamento dell'identità di gruppo contro fenomeni che sfidano tali identità come etnie, religioni e culture differenti), l'odio online differisce dall'odio offline in molte dimensioni rilevanti:

1. **si manifesta a livello verbale o mediante immagini**, quindi, non comporta né vandalismo né attacchi fisici;
2. le caratteristiche delle **piattaforme social** rappresentano un elemento facilitatore della creazione e della diffusione degli *hate speech* online, in quanto consentono una **rapida, efficace, permanente e poco costosa diffusione dei contenuti online** (Silva et al., 2016), democratizzandone la pubblicazione al punto da aver aperto la strada a qualsiasi tipo di messaggio, senza alcuna struttura - formale o informale - in grado di esercitare un'azione di mediazione (McGonagle, 2013). Sebbene l'*hate speech* online si manifesti solo verbalmente o attraverso le immagini, il suo impatto è potenzialmente molto **duraturo**: attraverso l'*hyperlinking*, i motori di ricerca e i contenuti condivisi dagli utenti, i messaggi di odio rimangono infatti tracciabili e recuperabili, determinando un **perdurare significativo del danno alla vittima** e alla minoranza a cui essa appartiene (McGonagle, 2013). Non solo: la possibilità di acquisire immagini dei contenuti verbali diffusi (i cosiddetti *screenshots*) fa sì che l'effetto dannoso possa mentenersi per sempre.
3. **coltiva stereotipi e pregiudizi** attraverso due canali:
   - un'elevata propensione da parte degli utilizzatori delle piattaforme di social media a collegarsi con utenti che condividono le stesse opinioni (Himelboin et al, 2013), creando "*echo chambers*": un sistema di relazioni denso dove prospettive, credenze, stereotipi e pregiudizi sono amplificati e rafforzati (Sunstein, 2017);
   - subendo il fenomeno dell'auto-veridicità, poiché una parte non trascurabile degli utenti che ricorrono a Internet come principale fonte di informazioni spesso non è dotata delle competenze critiche necessarie per valutare la legittimità delle informazioni che vengono presentate (Perry, 2001);
4. *L'hate speech* online è caratterizzato da una percezione di anonimato, o **de-individuazione** (Burnamp and Williams, 2015), che stimola **comportamenti più aggressivi e radicali** sotto l'apparente illusione di non essere identificati per quanto detto online e di non doverne, di conseguenza, rispondere (Gerstenfeld, 2017; Sunstein, 2017; Hall, 2013). Le conseguenze dell'associazione tra piattaforme social che consentono una comunicazione più immediata con i minori freni sociali in virtù della percezione di anonimato o pseudo-anonimato sono rappresentate dalla crescita esponenziale dei contenuti di odio sui social network (Citron e Norton, 2011). Ciò è reso oltremodo evidente dai dati 2018 sulla rimozione degli *hate speech* che forniscono un quadro del volume dei discorsi di odio prodotto su alcune delle principali piattaforme social: Facebook[3] ha rimosso circa 7,9 milioni di contenuti relativi ai discorsi di odio e YouTube[4] ha cancellato più di 15.000 canali a trimestre per le stesse motivazioni.

Alla luce degli elementi distintivi succitati che stanno emergendo come caratterizzanti gli *hate speech* online si è ulteriormente rafforzata la consapevolezza della portata specifica del loro impatto negativo. Infatti, l'*hate speech* online non solo comporta molti degli stessi effetti dell'odio esercitato offline (traumi psicologici, impatto negativo sulla comunità, etc.), ma favorisce, sfruttando le modalità estremamente veloci e pervasive del web, un'atmosfera in cui la violenza motivata da pregiudizi viene incoraggiata, in modo sottile o esplicito (Gagliardone et al. 2015).

1.2

Evoluzione del dibattito e definizioni giuridiche dell'*hate speech*

Accanto ad una conoscenza sempre più accurata del fenomeno, si sta inoltre sviluppando un ricco dibattito sull'**inquadramento giuridico** dell'*hate speech* online. Infatti, sebbene questo fenomeno sia sempre più analizzato e dibattuto in ambito accademico e istituzionale, si è ancora distanti da una identificazione normativa comune ai diversi contesti nazionali ed internazionali. Basti considerare che in alcuni casi, come la Germania, si è arrivati all'adozione di provvedimenti legislativi che impongono ai principali social network di agire tempestivamente per la rimozione dei contenuti riconducibili all'odio, mentre in altri contesti, come quello statunitense, la **contrapposizione** tra **diritto alla libertà di espressione** e **diritti umani/della persona** ha generato un acceso dibattito tra criminologi, psicologi sociali, sociologi e *policymaker* su quali confini tracciare (e se tracciarli) tra libertà di espressione e tutela della dignità delle persone (Gerstenfeld, 2017). La contrapposizione tra questi diritti è anche il motivo prevalente della difficoltà tuttora esistente di fornire una definizione condivisa a livello internazionale dell'*hate speech* (McGonagle, 2013).

1 ECRI European Commission against Racism and Intolerance.

2 ENAR European Network Against Racism.

3 *https://transparency.Facebook.com/community-standards-enforcement#hate-speech.*

4 *https://transparencyreport.Google.com/Youtube-policy/removals?hl=en.*

La finalità alla base della regolamentazione normativa degli *hate speech* è la salvaguardia dei diritti della persona e la prevenzione dell'insorgenza di rilevanti danni, individuali e collettivi. L'*hate speech* può avere effetti lesivi sui diritti umani e sui valori fondanti di una società, quali quelli sanciti dai principi costituzionali. Allo stesso tempo, può generare danni morali, psicologici e materiali per le vittime e costi sociali ed economici alla comunità. Nel complesso, l'estensione dei danni da prevenire è varia e complessa e il dibattito in corso si focalizza prevalentemente **sull'individuazione di criteri per distinguere tra danni che giustifichino restrizioni e quelli che non le giustificano** e su quali tipologie di restrizioni applicare.

Il lavoro di analisi che consegue questo tipo di discussione si sta concentrando sull'elaborazione di **approcci olistici** che prevedono l'adozione di un quadro normativo per la regolamentazione delle espressioni più eclatanti di *hate speech* e un sistema di *policies* non giuridiche (educative, culturali, informative, economiche) in grado di rimuovere i fattori di rischio legati alla proliferazione del*l'hate speech* (McGonagle, 2013).

1.2.1 Inquadramento giuridico e principali politiche per il contrasto dell'*hate speech* a livello istituzionale

Nel dibattito in corso un ruolo rilevante è stato svolto dalle diverse iniziative istituzionali intraprese negli ultimi anni a livello internazionale e nazionale che si sono concretizzate in primo luogo in un insieme di normative volte a fornire una definizione giuridica di *hate speech* quale presupposto fondamentale per inquadrare il fenomeno su basi legali e consentirne il contrasto. L'attenzione all'inquadramento giuridico si è accompagnata, inoltre, alla definizione di rilevanti documenti di policy che, attraverso specifiche raccomandazioni, hanno contribuito a delineare un sistema di politiche di intervento in materia.

Alcuni degli attori internazionali più attivi in riferimento all'azione normativa di contrasto dell'*hate speech* sono rappresentati dall'Organizzazione delle Nazioni Unite (ONU), dal Consiglio d'Europa e dalla Commissione Europea.

L'**ONU**, attraverso la sua vocazionale azione di contrasto alle violazioni dei diritti umani, rappresenta una delle istituzioni con il più ampio spettro di provvedimenti che contribuiscono alla definizione dell'attuale quadro di diritto internazionale sul tema del contrasto all'*hate speech*.

A questo proposito, particolarmente rilevanti risultano essere alcune disposizioni. In primo luogo, la **Dichiarazione Universale dei Diritti dell'Uomo (UDHR)** del 1948 che sancisce l'esistenza di limiti all'esercizio del diritto fondamentale della libertà di espressione, tra cui l'incitamento alla discriminazione. A questo atto fondamentale sono seguite altre disposizioni rilevanti: la **Convenzione per la Prevenzione e la Repressione del Crimine di Genocidio** del 1948, la **Convenzione Internazionale per l'Eliminazione di ogni Forma di Razzismo (ICERD)** del 1965 e la **Convenzione Internazionale sui Diritti Politici e Civili (ICCPR)** del 1966.

Anche se questi provvedimenti sono stati adottati ben prima della proliferazione dell'*hate speech* online, contengono l'identificazione di fattispecie di espressione verbale riconosciute a livello internazionale come lesive della dignità e dei diritti umani e punibili per legge.

La Convenzione Internazionale sui Diritti Politici e Civili, in particolare, contiene uno dei più rilevanti dispositivi prodotti a livello di istituzioni internazionali sul tema dell'*hate speech*, sancendo, all'interno dell'art. 20, che *qualsiasi forma di incoraggiamento all'odio razziale, religioso, nazionale che costituisca incitamento alla discriminazione, al conflitto e alla violenza debba essere proibita per legge*. Infatti, la Convenzione è lo strumento giuridico a cui si fa più comunemente riferimento nei dibattiti in materia di *hate speech* e loro regolamentazione (Gagliardone et al., 2015). Il fenomeno dell'*hate speech* online sta però evidenziando una serie di problematiche relative all'applicazione dell'art. 20 da parte dei singoli stati, rese evidenti sia dai bassi tassi di reazione e contrasto al fenomeno sia dall'assenza di strategie nazionali contro la proliferazione dei discorsi di incitamento all'odio razziale e religioso sul web.

Per superare i limiti nell'attuazione dell'art. 20 e, in generale, della Convenzione, l'Alto Commissariato delle Nazioni Unite per i diritti umani (OHCHR), ha organizzato una serie di incontri consultivi che hanno portato nel 2012 alla formulazione del **Piano di azione Rabat** sul divieto di "odio nazionale, razziale o religioso che

costituisca un incitamento alla discriminazione, all'ostilità o alla violenza". Il Piano mira a superare alcune evidenti limitazioni all'applicazione dell'art 20 dell'ICCPR da parte degli stati nazionali e l'esistente eterogeneità nelle azioni di contrasto, attraverso una serie di raccomandazioni per l'identificazione dei messaggi di odio che considerino il contesto, l'oratore, l'intento, il contenuto, l'estensione del discorso e il danno potenziale (Gagliardone et al., 2015). Il Piano, che si è sviluppato secondo un impianto *multistakeholder*, ha coinvolto la società civile, il mondo del giornalismo e le organizzazioni per la difesa dei diritti umani. Tuttavia, non ha previsto il coinvolgimento diretto delle piattaforme di *social networking*, che invece rappresentano un attore cruciale per la diffusione e, conseguentemente, il contrasto all'*hate speech* online.

Spostando il focus su scala europea, uno degli atti fondamentali è rappresentato dalla **Convezione Europea per i Diritti Umani emanata dal Consiglio d'Europa**, in cui si sancisce il valore del diritto alla libertà di espressione con la specificazione che tale diritto non possa intendersi come assoluto, ma limitabile attraverso provvedimenti legislativi per soddisfare alcune finalità, tra cui la protezione dei diritti di terzi. Sotto questo profilo, la Convenzione Europea per i Diritti Umani si muove in analogia con la Dichiarazione Universale dei Diritti dell'Uomo succitata.

Accanto a questo atto fondamentale, il Consiglio d'Europa ha realizzato altri provvedimenti rilevanti finalizzati a contrastare il proliferare dell'*hate speech*, tra cui la **Convenzione Quadro per la Protezione delle Minoranze Nazionali (1995)** che individua, tra le altre cose, la centralità del ruolo dei media nella promozione della tolleranza e del rispetto delle diversità. Il Comitato dei Ministri ha, nella fattispecie, adottato nel 1997 le Raccomandazioni **R (97) 20** sui discorsi d'odio e **R (97) 21** sui media e sulla cultura della tolleranza, in cui si prevede che gli Stati realizzino strategie di contrasto esaustive, in grado di agire sulla prevenzione e sul contrasto e che l'industria dei media e delle comunicazioni realizzi prodotti in grado di promuovere la cultura del rispetto e della tolleranza.

Alle attività di indirizzo e normativa, il Consiglio d'Europa ha inoltre affiancato attività di policy, istituendo nel 1993 la **Commissione europea contro il razzismo e l'intolleranza (ECRI)** che si occupa operativamente di monitorare e approfondire le tendenze e caratteristiche dell'*hate speech*, favorendo il coinvolgimento attivo della società civile. L'ECRI, in particolare, ritiene che un approccio efficace per contrastare l'incitamento all'odio online, debba prevedere l'autoregolamentazione da parte di istituzioni pubbliche e private, media e industria di Internet, attraverso l'adozione di codici di condotta e relative sanzioni, così come basarsi sull'istruzione e la contro-narrativa[5].

Un importante provvedimento del Consiglio d'Europa sul tema dell'*hate speech* è costituito dal **Protocollo addizionale alla Convenzione sulla Criminalità Informatica** entrato in vigore nel 2006, relativo alla penalizzazione degli atti di natura razzista e xenofoba commessi attraverso i sistemi informatici. In dettaglio, il Protocollo inserisce un'estensione dell'applicabilità dei dispositivi sulla criminalità informatica ai reati legati alla propaganda a sfondo razzistico e xenofobo, consentendo ai Paesi firmatari di poter ricorrere agli strumenti della cooperazione internazionale stabiliti nella Convenzione per il contrasto di tali reati. Focalizzandosi sulla criminalizzazione, prevede una serie di misure penali che possono essere adottate da ciascuno Stato relativamente a: diffusione di materiale razzista e xenofobo attraverso sistemi informatici; minaccia razzista e xenofoba e insulto, negazione, minimizzazione, approvazione o giustificazione del genocidio o dei crimini contro l'umanità, nonché in materia di favoreggiamento o complicità.

A livello comunitario, è la **Decisione Quadro sulla Lotta contro il Razzismo e la Xenofobia attraverso l'Azione Penale** (2008) a rappresentare il tassello fondamentale per l'identificazione di un quadro europeo comune per il contrasto all'*hate speech*. Tale decisione impegna infatti gli Stati membri dell'Unione europea a punire "l'istigazione pubblica alla violenza o all'odio nei confronti di un gruppo di persone, o di un suo membro, definito in riferimento alla razza, al colore, alla religione, all'ascendenza o all'origine nazionale o etnica», nonché «l'apologia, la negazione o la minimizzazione grossolana dei crimini di genocidio, dei crimini contro l'umanità e dei crimini di guerra», quando

5 In particolare, l'ECRI ha adottato due raccomandazioni: la Raccomandazione di politica generale n.6 (2000) che intende contrastare la diffusione di materiale razzista tramite Internet, chiedendo ai governi di adottare le misure necessarie, a livello nazionale e internazionale; la Raccomandazione di politica generale n. 15 sulla lotta all'incitamento all'odio (adottata l'8 Dicembre 2015), secondo la quale" l'incitamento all'odio" si basa sul presupposto ingiustificato che una persona o un gruppo di persone sia superiore agli altri, incita ad atti di violenza o discriminazione, minando il rispetto per le minoranze e danneggiando la coesione sociale. In questa raccomandazione, l'ECRI chiede di aumentare la consapevolezza delle conseguenze dell'incitamento all'odio e di ritirare il sostegno finanziario e di altro tipo ai partiti politici che fanno ricorso attivamente all'incitamento all'odio, criminalizzandone le manifestazioni più estreme. Sottolinea, tuttavia, che le misure contro l'incitamento all'odio debbano essere fondate, proporzionate, non discriminatorie e non utilizzate in modo improprio per limitare la libertà di espressione.

però tali comportamenti siano posti in essere in modo atto a istigare alla violenza o all'odio (...)" [6]. La decisione quadro impone agli Stati membri di perseguire penalmente l'istigazione pubblica alla violenza o all'odio per motivi di colore della pelle, religione, ascendenza, razza o origine etnica, anche quando essa avvenga online.

La decisione quadro è integrata dalla **Direttiva europea 2012/29 sui diritti delle vittime di reato** che mira, tra l'altro, a garantire giustizia, protezione e sostegno alle vittime di reati basati sull'odio e sull'incitamento all'odio[7].

Molto rilevante per il contrasto dell'hate speach il **Codice di Condotta per Contrastare l'Illecito Incitamento all'Odio Online** (EC, 2016)[8] che, su iniziativa della Commissione Europea, impegna le piattaforme di social media a realizzare protocolli e azioni per limitare la pubblicazione di contenuti di incitamento all'odio.

Più recentemente, anche la **Direttiva 2018/1808 sui servizi di media audiovisivi**[9] stabilisce norme che tutelano gli utenti dei servizi di media audiovisivi e delle piattaforme per la condivisione di video dall'istigazione alla violenza o all'odio, nonché da comunicazioni commerciali audiovisive discriminatorie. Impone inoltre a tali piattaforme di adottare misure adeguate per proteggere gli utenti dai contenuti razzisti e xenofobi.

Infine, la recente **Comunicazione della Commissione Europea COM (2020) 565 *"Un'Unione dell'uguaglianza: il piano d'azione dell'UE contro il razzismo 2020-2025***", definisce una serie di misure volte ad intensificare gli interventi e a ad aiutare le persone appartenenti a minoranze razziali o etniche a far sentire la loro voce. "I recenti episodi di tensione razziale fanno temere, infatti, che le tutele giuridiche contro la discriminazione razziale o etnica non siano ancora attuate in modo efficace", come sottolinea espressamente la Commissione. A questo riguardo, la Commissione si impegna ad effettuare una valutazione globale del quadro giuridico esistente per definire come migliorarne l'attuazione. La valutazione si baserà sul monitoraggio del recepimento e dell'attuazione della legislazione dell'UE, sulla prossima relazione riguardante l'attuazione della **Direttiva 2000/43/CE sull'uguaglianza razziale**[10] e sulle osservazioni degli stakeholder che rappresentano le preoccupazioni delle persone vittime di di discriminazione razziale, essenziali per massimizzare la portata e l'impatto dell'azione dell'UE. La Comunicazione sottolinea anche l'importanza che rivestirà la prossima **legge sui servizi digitali**, che avrà come obiettivo di aumentare e uniformare le responsabilità delle piattaforme online e dei fornitori di servizi informatici, rafforzando nel contempo la vigilanza sulle politiche relative ai contenuti delle piattaforme nell'UE[11]. La Comunicazione indica tra le possibili misure che potrebbero essere previste: l'introduzione obbligatoria di sistemi di notifica e azione e obblighi di segnalazione e trasparenza che imporranno alle piattaforme di fornire informazioni sulle modalità adottate per contrastare i contenuti illegali, incluso l'incitamento all'odio. La promozione di politiche di moderazione dei contenuti creerebbe anche una base per raccogliere dati sulla portata e sulle forme di incitamento all'odio razziale online, migliorando così anche la formulazione di politiche che affrontino efficacemente il problema del razzismo. Infine, sotto l'egida del Forum dell'UE su Internet, la Commissione, insieme agli Stati membri e alle società del web, sta collaborando alla redazione di un **elenco di riferimento di simboli e gruppi estremisti violenti vietati**, da utilizzare su base volontaria per orientare le loro politiche di moderazione dei contenuti. L'elenco sarà presentato alla riunione ministeriale del Forum dell'UE su Internet nel dicembre 2020.

6 Dal **2014** la Commissione vigila sul recepimento della decisione quadro negli ordinamenti giuridici degli Stati membri dell'UE. La misura in cui i codici penali nazionali configurano correttamente come reato l'incitamento all'odio e i reati generati dall'odio desta gravi preoccupazioni. Così, Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle regioni. *Un'Unione dell'uguaglianza: il piano d'azione dell'UE contro il razzismo 2020-2025*, COM (**2020**) **565** del **18-09-2020**.

7 La direttiva **2012**/**29**/UE impone agli Stati membri di garantire che le vittime di reato siano trattate in modo equo e non discriminatorio, accordando particolare attenzione alle vittime di reati motivati da pregiudizi o discriminazioni.

8 Il Codice per prevenire e contrastare la diffusione di incitamento illegale all'odio online, è stato concordato con Facebook, Microsoft, Twitter e YouTube. Nel corso del **2018**, anche Instagram, Snapchat e Dailymotion hanno aderito al Codice di condotta. Jeuxvideo.com si è unito a gennaio **2019** e TikTok ha annunciato la propria partecipazione al Codice a settembre **2020**. L'attuazione del Codice viene valutata attraverso regolari esercizi di monitoraggio in collaborazione con una rete di organizzazioni attive nei diversi paesi dell'UE. Utilizzando una metodologia concordata, queste organizzazioni testano come le società IT attuano gli impegni nel Codice.

9 Direttiva (UE) **2018**/**1808** del Parlamento europeo e del Consiglio, del **14** novembre **2018**, recante modifica della direttiva **2010**/**13**/UE, relativa al coordinamento di determinate disposizioni legislative, regolamentari e amministrative degli Stati membri concernenti la fornitura di servizi di media audiovisivi (direttiva sui servizi di media audiovisivi).

10 Alla relazione sull'attuazione della Direttiva **200**/**43**/CE verrà dato probabilmente seguito con eventuali atti legislativi entro il **2022**, così come indicato nella Comunicazione della Commissione *Un'Unione dell'uguaglianza: il piano d'azione dell'UE contro il razzismo 2020-2025,* COM (**2020**) **565** del **18-09-2020**.

11 Comunicazione Plasmare il futuro digitale dell'Europa (COM(**2020**) **67**.

Con riferimento al contesto italiano, la principale legge per il contrasto dell'*hate speech* è la L. **205/1993** (cd "Legge Mancino")[12] che prevede "la reclusione sino a 3 anni per chi "diffonde in qualsiasi modo idee basate sulla superiorità o sull'odio razziale o etnico, ovvero incita a commettere o commette atti di discriminazione per motivi razziali, etnici, nazionali o religiosi" (art. 1, lett a). Tale legge punisce anche "con la reclusione da sei mesi a quattro anni chi, in qualsiasi modo incita a commettere o commette violenza o atti di provocazione alla violenza per motivi razziali, etnici, nazionali o religiosi (art. 1, lett.b). La legge Mancino è quindi molto chiara, individuando condotte che vanno ben al di là della semplice manifestazione di un'opinione. Infatti, essa punisce l'istigazione a commettere una discriminazione o una violenza, non delle opinioni, quand'anche esprimano un pregiudizio.

Tale intervento legislativo si è tradotto, di fatto, in un inasprimento del trattamento sanzionatorio, in quanto consente di incriminare anche singoli "atti" di contenuto discriminatorio, nonché ha portato a un ampliamento dell'ambito di tutela, attraverso l'estensione della rilevanza penale anche alle manifestazioni discriminatorie attinenti alla sfera "religiosa", oltre a quelle razziale, etnica e nazionale.

Il D. Lgs 215/2003[13] successivamente, ribadendo che la parità di trattamento comporta che non sia praticata alcuna discriminazione, diretta o indiretta, per motivi di razza o origine etnica, stabilisce di considerare discriminazioni le molestie, ossia "quei comportamenti posti in essere per motivi di razza o di origine etnica, aventi lo scopo o l'effetto di violare la dignità di una persona e di creare un clima intimidatorio, ostile, degradante, umiliante e offensivo", contribuendo in tal modo a rafforzare la normativa antidiscriminatoria e a identificare le condotte punibili (si veda il box che segue).

Box 1.1 Nozione di discriminazione ex art. 2 D. lgs 215/2003

Il D. lgs 215/2003 all'art. 2. comma 1, sancisce che il principio della parità di trattamento comporta che non sia praticata alcuna discriminazione diretta o indiretta, così come di seguito definite:

a) discriminazione diretta quando, per la razza o l'origine etnica, una persona è trattata meno favorevolmente di quanto sia, sia stata o sarebbe trattata un'altra in situazione analoga;

b) discriminazione indiretta quando una disposizione, un criterio, una prassi, un atto, un patto o un comportamento apparentemente neutri possono mettere le persone di una determinata razza od origine etnica in una posizione di particolare svantaggio rispetto ad altre persone.

Sono considerate come discriminazioni ai sensi del comma **1**:

- le **molestie** ovvero quei comportamenti indesiderati, posti in essere per motivi di razza o di origine etnica, aventi lo scopo o l'effetto di violare la dignità di una persona e di creare un clima intimidatorio, ostile, degradante, umiliante e offensivo (art. **2** comma **3**).
- l'ordine di discriminare persone a causa della razza o dell'origine etnica è considerato una discriminazione (art. **2** comma **4**).

Diverse sono le sentenze che negli ultimi anni sono state pronunciate in applicazione della normativa succitata e che hanno avuto ad oggetto condotte discriminatorie/molestie per motivi razziali ai sensi del D.lgs. 215/03, sia perpetrate sul web che tramite altri strumenti di comunicazione e media. Si veda in merito i risultati di una breve ricognizione, che seppur non esaustiva, fornisce indicazione di alcune delle principali sentenze in materia emanate negli ultimi anni (Allegato 1).

Nel maggio 2016 è stata inoltre istituita presso la Camera dei deputati italiana **la Commissione su intolleranza, xenofobia, razzismo e fenomeni di odio (Jo Cox)**, i cui lavori hanno prodotto, nel 2017, una relazione finale che individua la cosiddetta **piramide dell'odio** divisa in quattro livelli. Alla base ci sono gli stereotipi, le

12 La "legge Mancino" si colloca all'interno di un complessivo quadro normativo volto a sanzionare penalmente le condotte riconducibili al fascismo e al razzismo, tra cui si annovera la L. 654 del 1975 che ha recepito la Convenzione internazionale sulla eliminazione di tutte le forme di discriminazione razziale, aperta alla firma a New York il 7 marzo 1966. La Legge Mancino la sostituisce riformulandone l'art. 3, 1 comma, lett a) e b).

13 Il decreto è stato emanato in attuazione della direttiva **2000**/**43**/CE per la parita' di trattamento tra le persone indipendentemente dalla razza e dall'origine etnica**.**

rappresentazioni false o fuorvianti, gli insulti e il linguaggio ostile "normalizzato"; al secondo livello la discriminazione, al terzo livello i discorsi di odio (minacce e/o incitamento alla denigrazione e alla violenza contro una persona o gruppi di persone identificate da caratteristiche come l'etnia, il colore della pelle, il sesso, ecc), ed infine i reati di odio che sono esplicitamente definiti come atti di violenza.

La relazione ha fornito anche specifiche raccomandazioni per prevenire e contrastare l'odio rivolte al Governo, alle autorità di regolamentazione e vigilanza, alle Istituzioni dell'UE, alle organizzazioni sovra-nazionali, ai media, all'ordine e al sindacato dei giornalisti, alle associazioni e a tutti gli altri operatori. Le raccomandazioni contemplano azioni da attuare sia a livello di normativa e politiche pubbliche, sia a livello sociale, culturale, educativo ed informativo. Il Box 1.2 che segue riporta i principali ambiti a cui sono riconducibili tali raccomandazioni[14].

Box 1.2 Commissione su intolleranza, xenofobia, razzismo e fenomeni di odio (Jo Cox): principali ambiti oggetto delle raccomandazioni

1) colmare le gravi lacune nella rilevazione e nell'analisi dei dati sui fenomeni di odio a livello nazionale e sovra-nazionale, in particolare per quanto riguarda il sessismo;
2) promuovere una strategia nazionale per contrastare l'odio in tutte le sue forme, articolata in piani d'azione specifici per combattere le discriminazioni dei singoli gruppi, ed attuare la Strategia Nazionale di Inclusione di Rom, Sinti e Camminanti;
3) approvare alcune importanti proposte di legge all'esame delle Camere, tra cui quelle sulla cittadinanza e sul contrasto dell'omofobia e della transfobia;
4) includere anche i discorsi d'odio sessisti nella legislazione in materia di odio e discriminazione;
5) sanzionare penalmente le campagne d'odio (insulti pubblici, diffamazione o minacce) contro persone o gruppi;
6) valutare, sulla base delle esperienze di altri Paesi e tutelando la libertà d'informazione in Internet, la possibilità di:
   - esigere l'autoregolazione delle piattaforme al fine di rimuovere l'*hate speech* online;
   - stabilire la responsabilità giuridica solidale dei provider e delle piattaforme di social network e obbligarli a rimuovere con la massima tempestività i contenuti segnalati come lesivi da parte degli utenti;
7) esigere da parte delle piattaforme dei social network l'istituzione di uffici dotati di risorse umane adeguate, al fine della ricezione delle segnalazioni e della rimozione tempestiva dei discorsi d'odio, anche attivando alert sulle pagine online e numeri verdi a disposizione degli utenti;
8) rafforzare il mandato dell'UNAR (Ufficio Nazionale Antidiscriminazioni Razziali) in direzione di una maggiore autonomia, anche configurandolo quale autorità indipendente;
9) responsabilizzare le figure istituzionali e politiche influenti nel dibattito pubblico, adottando meccanismi di regolazione per combattere il discorso d'odio;
10) migliorare la conoscenza dei propri diritti da parte delle vittime e consentire alle organizzazioni attive nel contrasto alle forme d'odio di costituirsi parte civile in giudizio;
11) attuare e diffondere la conoscenza delle norme previste dalla Legge n. **71** del **2017** sul bullismo;
12) rafforzare nelle scuole l'educazione di genere e l'educazione alla cittadinanza, finalizzata agli obiettivi di rispetto, apertura interculturale, inter-religiosa e contrasto ad intolleranza e razzismo;
13) sostenere e promuovere blog e attivisti *no hate* o testate che promuovono una contronarrazione e campagne informative rispetto al discorso d'odio, soprattutto nel mondo non profit, delle scuole e delle università;
14) contrastare gli stereotipi e il razzismo sensibilizzando e responsabilizzando i media, specie online, ad evitare il discorso d'odio, comprese le notizie infondate, false e diffamatorie;
15) prevedere l'istituzione di un giurì che garantisca la correttezza dell'informazione, come prospettato anche da proposte di legge presentate in questa e in precedenti legislature e sollecitare l'Ordine professionale e il sindacato dei giornalisti sul controllo della deontologia professionale.

14 Si v. Camera dei deputati, La piramide dell'odio in Italia, Commissione Jo Cox su fenomeni di odio, intolleranza, xenofobia e razzismo - Relazione finale, infografica, 2017.

Più recentemente, ad ottobre 2019, con l'approvazione della cd. mozione Segre, è stata istituita al senato la **Commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza,** per ribadire l'impegno delle istituzioni a contrastare la crescente spirale di tali fenomeni che pervadono oramai la scena pubblica anche per via della loro capillare diffusione attraverso vari mezzi di comunicazione e in particolare sul web.

Box 1.3 Commissione Straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza

La Commissione, costituita da 25 componenti, ha compiti di osservazione, studio e iniziativa per l'indirizzo e controllo sui fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza nei confronti di persone o gruppi sociali sulla base di alcune caratteristiche quali l'etnia, la religione, la provenienza, l'orientamento sessuale, l'identità di genere o di altre particolari condizioni fisiche o psichiche. Ha il compito di controllare e indirizzare la concreta attuazione delle convenzioni e degli accordi sovra-nazionali e internazionali e della legislazione nazionale su questi temi e può svolgere anche una funzione propositiva, di stimolo e di impulso, nell'elaborazione e nell'attuazione delle proposte legislative e ogni altra iniziativa utile a livello nazionale, sovra-nazionale e internazionale.

A tal fine la Commissione:

a) raccoglie, ordina e rende pubblici, con cadenza annuale: normative statali, sovra-nazionali e internazionali; ricerche e pubblicazioni scientifiche, anche periodiche; dati statistici, nonché informazioni, dati e documenti sui risultati delle attività svolte da istituzioni, organismi o associazioni che si occupano di questioni attinenti ai fenomeni di intolleranza, razzismo e antisemitismo, sia nella forma dei crimini d'odio che nella forma di *hate speech*;

a) effettua ricerche, studi e osservazioni concernenti tutte le manifestazioni di odio nei confronti di singoli o comunità, anche in contatto con istituzioni di altri Paesi, nonché con organismi sovra-nazionali e internazionali o effettuando missioni in Italia o all'estero, in particolare presso Parlamenti stranieri, allo scopo di stabilire intese per il contrasto all'intolleranza, al razzismo e all'antisemitismo, sia nella forma dei crimini d'odio, sia dei fenomeni di *hate speech*;

b) formula osservazioni e proposte sugli effetti, sui limiti e sull'eventuale necessità di adeguamento della legislazione vigente al fine di assicurarne la rispondenza alla normativa dell'Unione europea e ai diritti previsti dalle convenzioni internazionali in materia di prevenzione e di lotta contro ogni forma di odio, intolleranza, razzismo e antisemitismo. Quando necessario, può svolgere procedure informative, formulare proposte e relazioni all'Assemblea, formulare pareri su disegni di legge etc. Entro il **30** giugno di ogni anno, la Commissione trasmetterà al Governo e alle Camere una relazione sull'attività svolta, i risultati delle indagini effettuate, le conclusioni raggiunte e le proposte formulate. La Commissione può, infine, segnalare agli organi di stampa ed ai gestori dei siti internet casi di fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza nei confronti di persone o gruppi sociali sulla base di alcune caratteristiche, quali l'etnia, la religione, la provenienza, l'orientamento sessuale, l'identità di genere o di altre particolari condizioni fisiche o psichiche, richiedendo la rimozione dal web dei relativi contenuti ovvero la loro deindicizzazione dai motori di ricerca.

Infine, significativo a livello nazionale è anche il lavoro di recente svolto **dal Consiglio dell'Autorità per le Garanzie nelle Comunicazioni** (AGCOM), che ha approvato ad aprile di questo anno il **Regolamento sulle nuove "Disposizioni in materia di rispetto della dignità umana e del principio di non discriminazione e di contrasto all'*hate speech*"** contenuto nella Delibera 157/19/CONS. Alla definizione del Regolamento hanno contribuito, grazie all'avvio di una consultazione pubblica, le associazioni di settore, rappresentanti della società civile e delle imprese, nonché l'Ordine dei Giornalisti che ha avviato una procedura di confronto permanente sulle iniziative dell'Autorità. Attraverso il regolamento l'Autorità fornisce un quadro più definito delle norme finalizzate a contrastare gli *hate speech*, secondo i principi delle normative italiane ed europee in materia, stabilendo i principi e le **disposizioni cui devono adeguarsi i fornitori di servizi media audiovisivi e radiofonici** soggetti alla giurisdizione italiana nei programmi di informazione e intrattenimento per assicurare il rispetto

della dignità umana e del principio di non discriminazione e contrasto alle espressioni di odio. Inoltre, nelle more della trasposizione della nuova direttiva europea sui servizi media audiovisivi che estende alle piattaforme di condivisione di video online taluni obblighi in materia, l'Autorità promuove, coordina e indirizza l'elaborazione di codici di condotta di co-regolazione con tali piattaforme. L'Autorità ha inoltre predisposto una campagna video istituzionale in tema di contrasto all'*hate speech* sulle reti televisive nazionali.

Nonostante l'importante lavoro realizzato dalle istituzioni internazionali nell'arco di diversi anni per identificare un quadro armonizzato a livello sovra-nazionale sugli approcci giuridici e normativi di contrasto all'*hate speech*, ad oggi, **l'*hate speech* online non è identificato come crimine nella maggior parte dei paesi**. La motivazione prevalente dietro al disallineamento tra approcci internazionali e nazionali risiede principalmente negli orientamenti differenti in materia di libertà di espressione e nelle difficoltà di realizzare provvedimenti di contrasto efficaci considerata l'elevata innovazione del settore ICT (Assimakopoulos et al., 2017; Gagliardone, 2017).

Una delle critiche maggiormente diffuse e condivise nei confronti dell'approccio normativo e dell'azione penale a contrasto dell'*hate speech* si fonda quindi sui rischi legati alla limitazione della libertà d'espressione. In particolare, nel caso del contesto digitale, prevedere l'introduzione di tecniche preventive di filtraggio dei contenuti realizzate dai provider dei servizi digitali può portare ad interventi di censura, come la ***collateral censorship*** per cui lo Stato utilizza un provider di servizi digitali per censurare un altro soggetto, (Balkin, 2014) o la **censura privata** esercitata dalle piattaforme social senza completa *accountability* e *disclosure* nei confronti del pubblico.

Un altro aspetto rilevante per le iniziative legislative e le azioni di contrasto degli *hate speech* è legato alle caratteristiche del web. La velocità dell'innovazione di Internet e dell'ICT in generale, rende particolarmente **difficile agire in modo efficace esclusivamente attraverso provvedimenti normativi**, in quanto pagine, identità virtuali e ISP possono essere spostati in brevissimo tempo e con costi irrisori da un paese all'altro, consentendo di sottrarre i contenuti incriminati alle regolamentazioni nazionali non appena esse vengano emanate (OSCE, 2010).

1.2.2 Il ruolo della società civile e dei social network nel dibattito legislativo e nell'implementazione delle politiche di contrasto all'*hate speech* online

Dal 2018 una rete informale di circa 30 tra organizzazioni e ricercatori dediti allo studio e al contrasto dei fenomeni d'odio e della discriminazione hanno costituito il c.d "Tavolo odio" nell'ambito del quale vengono svolti incontri tematici su aspetti giuridici, educazione, attivismo, comunicazione, con l'obiettivo di stimolare una riflessione costruttiva sul fenomeno dell'*hate speech* online e individuare possibili interventi.

Nel 2020, per coordinare in maniera più efficace le diverse iniziative capaci di dare una risposta davvero incisiva al contrasto dei discorsi e dei fenomeni di odio, il "Tavolo" si è trasformato e organizzazioni non governative, associazioni, movimenti, università, centri di ricerca, osservatori si sono uniti per dare vita alla **Rete nazionale per il contrasto ai discorsi e ai fenomeni d'odio che** riunisce **le più importanti realtà che da diverso tempo si occupano di mappare e combattere i discorsi e i fenomeni di odio**.

Tra le organizzazioni coinvolte e dedite al contrasto dell'*hate speech* online particolarmente rilevante è il ruolo svolto da Amnesty International che negli ultimi anni ha sviluppato in questo ambito diverse iniziative di rilievo.

Amnesty International Italia ha, in particolare, creato la c.d "*Task force hate speech*", una rete di 150 attivisti/e che dal novembre 2017 intervengono nello spazio dedicato ai commenti delle pagine online e nelle piattaforme social (Facebook e Twitter) dove possono svilupparsi discorsi d'odio nei confronti di determinati soggetti-bersaglio. L'azione della *Task Force* si focalizza sui commenti con la finalità di veicolare informazioni che siano imparziali e sensibilizzare gli utenti del web all'utilizzo di un linguaggio corretto e non discriminatorio.

In vista delle elezioni politiche di marzo 2018, Amnesty durante le ultime tre settimane della campagna elettorale, ha monitorato i profili social (Facebook e Twitter) di tutti i candidati ai collegi uninominali di Camera e Senato, dei candidati presidenti delle Regioni Lazio e Lombardia e dei leader politici. I post e i tweet, le immagini

e i video condivisi dai candidati, e quindi a loro direttamente attribuibili, sono stati seguiti quotidianamente segnalando l'uso di stereotipi, dichiarazioni offensive, razziste, discriminatorie e di incitamento alla violenza che hanno avuto come bersaglio categorie vulnerabili quali migranti e rifugiati, immigrati, rom, persone LGBTI, donne, comunità ebraiche e islamiche[15].

Anche in vista delle elezioni parlamentari europee, Amnesty International ha inoltre esaminato e valutato nell'ambito del monitoraggio **"Barometro dell'odio - Elezioni europee 2019",** circa 33.100 contenuti tra il 26 aprile e il 15 maggio, osservando in particolar modo i profili Facebook e Twitter dei candidati e delle candidate al Parlamento europeo più attivi online e dei leader di partito, per valutare modalità di espressione, possibile utilizzo del linguaggio d'odio in merito alle categorie bersaglio succitate e a specifici temi quali la solidarietà e la povertà socio-economica. Oggetto di osservazione sono state anche le reazioni degli utenti, per rilevare le eventuali correlazioni tra toni e messaggi veicolati dalla politica e sentimento delle persone[16].

Non trascurabile all'interno del dibattito in corso è, infine, il ruolo svolto dai **social network**, considerato specialmente il loro ruolo abilitante nei confronti di una consistente parte dei contenuti presenti sul web.

Soggetti quali Facebook, Twitter e Google rappresentano pertanto una parte attiva nei processi di regolamentazione dell'*hate speech* online. A questo proposito, basti ricordare che la recente legge tedesca di contrasto all'*hate speech* online prevede obblighi espliciti nei confronti delle piattaforme di social network e che il Codice di Condotta per Contrastare l'Illecito Incitamento all'Odio Online pubblicato nel 2016 dalla Commissione Europea è stato realizzato coinvolgendo Facebook, Microsoft, Twitter e YouTube.

Gli stessi social media inoltre si sono, nel tempo, dotati di propri **codici di autoregolamentazione sui contenuti di odio e intolleranza** che presentano tra loro significative differenze.

Facebook, per esempio, identifica e definisce i contenuti di odio da rimuovere come "*contenuti che attaccano le persone in base alla loro razza, etnia, ceto, nazionalità, religione, sesso, orientamento sessuale, disabilità o malattia*[17]", specificando tuttavia che sono permessi "*chiari tentativi di umorismo o satira che altrimenti potrebbero essere considerati una possibile minaccia o attacco, inclusi contenuti che molte persone possono trovare di cattivo gusto*".

Youtube, che appartiene a Google, dichiara la non ammissibilità di contenuti di odio, definiti come "*discorsi che attaccano o denigrano un gruppo basato su età, disabilità, etnia, genere, nazionalità, razza, condizione migratoria, religione, sesso, orientamento sessuale, status di veterano*[18]".

Infine, **Twitter** identifica come rimuovibili i contenuti che promuovono "*violenza, attacchi e minacce, dirette o indirette, ad altre persone sulla base di razza, etnia, nazionalità, orientamento sessuale, sesso, identità di genere, appartenenza religiosa, età, disabilità o malattie gravi.*[19]" Inoltre, non ammette contenuti il cui scopo principale è incitare al danno verso altri sulla base di queste categorie. Dal 2012, ha inoltre modificato le regole di utilizzo del social network introducendo per la prima volta un **criterio geografico di censura selettiva**, per cui la società può decidere di oscurare i tweet o gli account che violino le leggi di una determinata nazione soltanto in quella nazione, mentre il messaggio continua a essere visibile per gli utenti di altre nazionalità. La nuova politica di Twitter è stata accolta da alcuni come una pericolosa limitazione della libertà di espressione e da altri come un importante passo avanti dal punto di vista della flessibilità, in quanto esclude la rimozione completa del contenuto.

Google, il più importante motore di ricerca, infine, ha adottato una policy relativa agli *hate speech* che prevede l'impegno a "non distribuire contenuti che promuovono l'odio o la violenza nei confronti di gruppi di persone

15 Per gli esiti dell'analisi si veda, Amnesty International, Conta fino a **10**, barometro dell'odio in campagna elettorale, **2018**, *https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2018/02/16105254/report-barometro-odio.pdf*.

16 I dati raccolti saranno analizzati da data scientist, sociologi, linguisti, psicologi e giuristi e illustrati in un rapporto, la cui pubblicazione è prevista intorno alla data di insediamento del nuovo Parlamento europeo.

17 Facebook Community Standards.

18 *https://support.google.com/youtube/answer/2801939?hl=en-GB*.

19 *https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy*.

in base alla loro razza o origine etnica, religione, disabilità, sesso, età, stato di veterano, o orientamento sessuale/identità di genere.[20]"

Sulla base di questi codici di autoregolamentazione, Google e Facebook hanno anche realizzato dei *transparency report* in cui rendicontano le attività svolte per la rimozione dei contenuti di odio, sia sulla base delle regole interne, sia su istanza di autorità pubbliche o di privati. Tra le criticità più evidenti che sono emerse, si segnalano i **limiti esistenti nei processi di identificazione dei contenuti da rimuovere e il rischio di censura privata non verificabile** (Heins, 2014).

Per identificare i discorsi d'odio, le piattaforme utilizzano soprattutto algoritmi con software di riconoscimento di parole e/o affidano agli utenti e agli organismi di reclamo il compito di segnalare contenuti problematici. La maggior parte dei gestori di social media come Facebook, YouTube, Twitter, Instagram, hanno integrato una funzione di segnalazione.

La quinta valutazione[21] del Codice di Condotta della Commissione Europea introdotto nel 2016 e presentato sinteticamente nel box che segue, mostra che le aziende IT valutano il 90% dei contenuti segnalati entro 24 ore (mentre era solo il 40% dei contenuti nel 2016) e rimuovono il 71% dei contenuti ritenuti illegali (a fronte del 28% del 2016)[22]. Il tasso di rimozione medio, simile a quello registrato nelle precedenti valutazioni, mostra che le piattaforme continuano a rispettare la libertà di espressione ed evitare la rimozione di contenuti che potrebbero non essere considerati discorsi di odio illegali. La valutazione mostra quindi risultati complessivamente positivi, anche se le piattaforme devono migliorare ulteriormente la trasparenza e il feedback nei confronti degli utenti e garantire che i contenuti vengano valutati in modo coerente nel tempo.

20 *https://www.google.com/+/policy/content.html*.

21 Il quinto esercizio di monitoraggio è stato effettuato per un periodo di **6** settimane, dal **4** novembre al **13** dicembre **2019**, da **34** organizzazioni della società civile.

22 Più in dettaglio, Facebook ha rimosso l'**87.6**% dei suoi contenuti, YouTube il **79.7**%, e Twitter il **35.9**%. Facebook rispetto all'anno precedente ha conseguito dei progressi, YouTube continua ad avere un tasso elevato e Twitter a essere molto al di sotto degli obiettivi (il suo tasso di rimozione è più basso del **2019**). Jeuxvideo.com ha rimosso tutti i contenuti segnalati e Instagram solo il **42**%. Per maggiori dettagli, si v. Countering illegal *hate speech* online **5**th evaluation of the Code of Conduct, Factsheety, June **2020**, *https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf*.

Box 1.4 Codice di condotta per lottare contro le forme illegali di incitamento all'odio online

Il Codice di Condotta è uno strumento con cui le aziende informatiche affiancano la Commissione europea e gli Stati membri dell'UE nell'affrontare la sfida di garantire che le piattaforme online non offrano opportunità di diffusione virale di forme illegali di incitamento all'odio online. Nel **2016** le aziende informatiche (Facebook, Twitter, YouTube e Microsoft con i pertinenti servizi ai consumatori che ospita) hanno infatti condiviso l'impegno della Commissione europea e degli Stati membri dell'UE a lottare contro le forme illegali di incitamento all'odio online esprimendo al contempo l'adesione ad una responsabilità collettiva nella promozione e sostegno della libertà di espressione in tutto il mondo online.

Con la firma del Codice di condotta le aziende informatiche si sono impegnate a contrastare qualsiasi illecito incitamento all'odio online anche rafforzando il partenariato con le organizzazioni della società civile, che possono contribuire a segnalare i contenuti istiganti alla violenza e a comportamenti improntati all'odio. La società nel suo complesso e, in particolare, le organizzazioni della società civile rivestono infatti anch'esse un ruolo fondamentale da svolgere per prevenire la diffusione dell'odio online facendosi portatrici, anche tramite attività di sensibilizzazione, di narrazioni alternative che promuovano la non discriminazione, la tolleranza e il rispetto.

Le aziende informatiche e la Commissione europea con la firma del Codice si sono impegnate ad elaborare e promuovere narrazioni alternative indipendenti di nuove idee e iniziative di sostegno di programmi educativi che incoraggino il pensiero critico.

Di seguito vengono riportati gli impegni pubblici sui quali si fonda il Codice di condotta

- Le aziende informatiche predispongono procedure chiare ed efficaci per esaminare le segnalazioni riguardanti forme illegali di incitamento all'odio nei servizi da loro offerti, in modo da poter rimuovere tali contenuti o disabilitarne l'accesso. Le aziende informatiche predispongono regole o orientamenti per la comunità degli utenti volte a precisare che sono vietate la promozione dell'istigazione alla violenza e a comportamenti improntati all'odio.
- Al ricevimento di una segnalazione valida mirante alla rimozione di forme illegali di incitamento all'odio, le aziende informatiche la esaminano alla luce delle regole e degli orientamenti da esse predisposti per la comunità degli utenti e, ove necessario, delle leggi nazionali di recepimento della decisione quadro **2008**/**913**/GAI, affidando l'esame a squadre specializzate.
- Le aziende informatiche esaminano in meno di **24** ore la maggior parte delle segnalazioni valide miranti alla rimozione di forme illegali di incitamento all'odio e, se necessario, rimuovono tali contenuti o ne disabilitano l'accesso.
- Inoltre, le aziende informatiche svolgono presso i loro utenti un'opera di educazione e di sensibilizzazione sulle tipologie di contenuti non autorizzate in base alle regole e agli orientamenti da esse predisposti per la comunità degli utenti. A tal fine potrebbe essere utilizzato il sistema di segnalazione.
- Le aziende informatiche forniscono informazioni sulle procedure di trasmissione di avvisi, al fine di rendere più rapida ed efficace la comunicazione fra le autorità degli Stati membri e le aziende informatiche, in particolare per quanto riguarda le segnalazioni, la disattivazione dell'accesso o la rimozione delle forme illegali di incitamento all'odio online. Le informazioni devono essere trasmesse tramite i punti di contatto nazionali rispettivi designati dalle aziende informatiche e dagli Stati membri. In tal modo si consentirà anche agli Stati membri, e in particolare alle autorità nazionali incaricate dell'applicazione della legge, di acquisire ulteriore familiarità con i metodi per riconoscere le forme illegali di incitamento all'odio online e segnalarle alle aziende informatiche.
- Le aziende informatiche incoraggiano la trasmissione degli avvisi e la segnalazione dei contenuti che promuovono l'istigazione alla violenza e ai comportamenti improntati all'odio avvalendosi di esperti, in particolare attraverso partenariati con le organizzazioni della società civile, fornendo chiare informazioni sulle regole e sugli orientamenti da esse predisposti per la comunità degli utenti e sulle regole in materia di procedure di comunicazione e di segnalazione. Le aziende informatiche si adoperano per rafforzare i partenariati con le organizzazioni della società civile ampliando la portata geografica di tali partenariati e,

se del caso, offrono sostegno e formazione ai partner delle organizzazioni della società civile per consentire loro di svolgere il ruolo di "relatore di fiducia" o equivalente, tenendo in debita considerazione l'esigenza di preservarne l'indipendenza e la credibilità.

- Le aziende informatiche contano sul sostegno degli Stati membri e della Commissione europea per garantire l'accesso a una rete rappresentativa di partner delle organizzazioni della società civile e di "relatori di fiducia" in tutti gli Stati membri che possano contribuire allo sforzo di trasmettere avvisi di alta qualità. Le aziende informatiche pubblicano le informazioni sui "relatori di fiducia" sul loro sito web.
- Le aziende informatiche organizzano periodicamente formazioni per il proprio personale per informarlo sugli sviluppi sociali in corso e si scambiano opinioni sulle possibilità di ulteriori miglioramenti.
- Le aziende informatiche intensificano la loro cooperazione con altre piattaforme e altri operatori dei media sociali per migliorare la condivisione delle migliori pratiche.
- Le aziende informatiche e la Commissione europea, riconoscendo il valore di voci indipendenti che contrastino la retorica dell'odio e i pregiudizi, si prefiggono di proseguire l'opera di elaborazione e promozione di narrazioni alternative indipendenti, di nuove idee e iniziative di sostegno di programmi educativi che incoraggino il pensiero critico.
- Le aziende informatiche intensificano la collaborazione con le organizzazioni della società civile per fornire formazione sulle migliori pratiche per lottare contro la retorica dell'odio e i pregiudizi e aumentano la portata del loro approccio proattivo nei confronti delle organizzazioni della società civile per aiutarle a realizzare campagne efficaci di lotta contro i discorsi di incitamento all'odio. La Commissione europea, in cooperazione con gli Stati membri, contribuisce a questo sforzo provvedendo a repertoriare le esigenze e le richieste specifiche delle organizzazioni della società civile in proposito.

È evidente che il Codice di condotta ha creato un'importante collaborazione tra organizzazioni della società civile, autorità nazionali e piattaforme IT. Molto farà anche la prossima legge sui servizi digitali, su cui la Commissione ha recentemente lanciato una consultazione pubblica, in quanto contribuirà ad adottare in un quadro europeo misure di trasparenza vincolanti per le piattaforme per chiarire come gestiscono i discorsi di odio illegali.

Infine, è importante che continui il dialogo tra le Piattaforme informatiche e le organizzazioni della società civile che lavorano sul campo per affrontare il discorso dell'odio illegale. Recentemente, la campagna contro l'odio in rete #stophateforprofit con l'obiettivo di rendere i social media più responsabili sulla diffusione di disinformazione e discorsi di incitamento all'odio, sembra invece aver acuito particolarmente la contrapposizione. La campagna ha infatti previsto la sospensione ad opera di tantissime aziende americane ed europee - da Coca Cola fino a Unilever - delle inserzioni pubblicitarie su Facebook, accusato di non attivarsi sufficientemente per contrastare i contenuti d'odio e razzisti, misogini e violenti online, per sensibilizzare la piattaforma social sulle proprie policy di moderazione dei contenuti. Tra le richieste fatte a Facebook: l'assunzione di un dirigente che si occupi di diritti civili; il divieto di creare gruppi privati che promuovano la supremazia bianca, disinformazione o le teorie cospirative; la possibilità di segnalare i contenuti di politici che diffondono fake news sulle elezioni (opzione quest'ultima, a cui Facebook si è sempre opposta ritenendo che il criterio di notiziabilità prevalga sull'eventuale inesattezza del contenuto pubblico, a differenza di quanto sta facendo Twitter). In questo contesto, nel giugno del 2020 Facebook ha cancellato circa 190 profili social che erano collegati a gruppi di suprematisti bianchi ossia quel movimento ideologico che crede che i bianchi siano una razza superiore rispetto agli afro-americani. La dilagazione di questi gruppi negli Usa ha spinto Facebook e Instagram a bandire dai social le organizzazioni di questo tipo. Facebook e Instagram hanno respinto inoltre 2,2 milioni di pubblicità e rimosso 120.000 post per aver tentato di "ostacolare il voto" e sono stati pubblicati avvisi su 150 milioni di false informazioni postate online[23]. A questo si aggiunge il fenomeno dell'infodemia reso ancora più visibile durante la crisi pademica da Coronavirus che stiamo attraversando. L'infodemia viene

23 A riferire della "campagna di pulizia" è Nick Clegg VicePresidente di Facebook in un'intervista al francese Journal du Dimanche.

definita dall'OMS[24] come "una sovrabbondanza di informazioni - alcune accurate altre no - che rende difficile alle persone trovare fonti attendibili e indicazioni affidabili quando ne hanno bisogno". In questo contesto i social sono sicuramente fonte primaria di condivisione delle notizie, ma contengono anche una certa dose di disinformazione e fake news alimentate dalla rabbia sociale che accompagna la pandemia[25]. Ma i grandi del web, in collaborazione con l'OMS, stanno prendendo misure per rimuovere fake news e promuovere informazioni accurate. Googlando "coronavirus", ad esempio, vengono messi in evidenza i risultati della stessa OMS. Anche Facebook, Twitter e YouTube indirizzano gli utenti sui siti dell'OMS o alle organizzazioni sanitarie locali. La costante crescita degli *hate speech* online inevitabilmente ha posto come visto una serie di sfide al sistema legislativo, sia a livello nazionale che a livello internazionale. Il ruolo del diritto nel contrastare il fenomeno è riconosciuto ormai da tempo ma altrettanto riconosciuto è che il contrasto all'*hate speech* non si debba esercitare esclusivamente o prioritariamente attraverso la disciplina legislativa, bensì integrando all'approccio normativo anche altri approcci che coniugano ricerca e analisi del fenomeno con interventi di sensibilizzazione a valenza educativa e culturale (i.a. Gagliardone et al., 2015).

24 *https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf*

25 *http://www.robertobortone.it/2020/06/pandemia-razzismo-e-rabbia-sociale.html*

2

Evoluzione degli approcci e metodologie di analisi degli *hate speech*

Di seguito si presenta una sintesi degli approcci e delle principali tecniche di monitoraggio ed analisi degli *hate speech* adottate dalle metodologie che sono state sviluppate in ambito accademico/scientifico negli ultimi anni. Saranno presentati, in particolare, approcci basati su modelli teorici che provengono dalla sociologia, psicologia e linguistica e approcci basati sui recenti sviluppi di tecniche di analisi del linguaggio computerizzate.

Le esperienze/metodologie di individuazione ed analisi dei contenuti di odio presenti su Internet si sono sviluppate negli anni, contemporaneamente al dibattito tra difesa della libertà di espressione e difesa della dignità umana. Su tale filone di studio e ambito di intervento sono attive diverse tipologie di attori (NGOs, università, istituzioni, giornalisti, gruppi di *advocacy*) che, sempre più spesso, lavorano in **team multidisciplinari,** nella consapevolezza ormai consolidata della maggiore efficacia di un **approccio olistico** rispetto a una tematica multi-sfaccettata come quella rappresentata dai discorsi di incitamento all'odio online. L'unione di competenze informatiche, neuroscientifiche, psicologiche, semantiche e statistiche ha consentito e consente tuttora infatti di sviluppare processi e tecniche di estrazione di dati da Internet sempre più accurati e raffinati.

Le metodologie di identificazione e analisi degli *hate speech* **online** si articolano sostanzialmente in tre fasi/attività:

Fase 1 creazione base informativa/database su cui effettuare le analisi;
Fase 2 attività di analisi dati/contenuti di odio online;
Fase 3 produzione output specifici, differenti a seconda delle metodologie.

La maggioranza delle prime esperienze/metodologie di analisi dell'*hate speech* online si è caratterizzata per un approccio di analisi sostanzialmente di tipo **qualitativo** adottato nei primi anni di studio del fenomeno. A titolo esemplificativo, si può citare il progetto europeo **PRISM** (*Preventing, Redressing & Inhibiting Hate Speech in New Media*)[1], coordinato dall'ARCI, la cui attività di ricerca e analisi sul fenomeno dell'*hate speech* online si è basata su interviste qualitative e su una mappatura dell'uso da parte di alcuni gruppi xenofobi e di estrema destra dei social media (Twitter, Facebook, Youtube). PRISM ha prodotto una prima fotografia dei discorsi d'odio su Internet verificando quotidianamente, sui post selezionati per il monitoraggio i *followers*, i principali *hashtag* e le parole più utilizzate, nonché analizzando altri ambiti di interazione online, come le sezioni dei commenti di quotidiani digitali ed i forum di discussione generale.

Queste prime esperienze/metodologie di analisi di tipo qualitativo hanno fornito importanti contributi preliminari necessari per lo sviluppo del secondo approccio di tipo **quantitativo** basato sul ricorso in maniera crescente a sistemi di *machine learning*[2] per l'estrazione e l'analisi dei **flussi di** *hate speech* dalle piattaforme di *microblogging* (principalmente Facebook e Twitter) e che si è consolidato come approccio prevalente ormai da alcuni anni, in quanto consente di analizzare volumi crescenti di dati e gestire ambiti geografici sempre più estesi.

Le tecniche di estrazione ed analisi si stanno infatti evolvendo in modo particolarmente rapido, considerata la rilevanza del fenomeno in esame e la fiorente ricerca in questo ambito che ha consentito di aumentare automazione ed efficacia dei sistemi informatici che monitorano i flussi di *hate speech* individuando i contenuti caratterizzati dall'uso di lessico identificato come intollerante/discriminatorio in tempi brevissimi.

Entrando maggiormente nel dettaglio delle tecniche di estrazione e monitoraggio dei flussi di *hate speech* contenuti nelle piattaforme social, le esperienze/metodologie iniziali si sono basate sulla tecnica ***Bag of Words* (*BoW***) che si focalizza sulle parole di odio individuate come "termini-sentinella" e utilizzate per impostare i criteri con cui algoritmi informatici di ***Natural Language Processing*** **(NPL)** identificano ed estraggono i contenuti di odio online da una piattaforma. Il complesso degli *hate speech* estratti può contenere tuttavia una

1 Il progetto PRISM ha combinato l'attività di ricerca e analisi sul fenomeno dell'*hate speech* online ad attività formative, di sensibilizzazione e allo sviluppo di strumenti (toolkit) per il contrasto della discriminazione e della violenza online. Per i risultati di PRISM, si v. Arci, Cittalia, *Discorsi d'odio e Social Media Criticità, strategie e pratiche d'intervento https://www.arci.it/app/uploads/2018/05/progetto_PRISM_-_bassa.pdf.*

2 Si intende l'apprendimento automatico, alla base dei processi di automatizzazione dei modelli analitici, che permette ai computer di apprendere dall'esperienza umana attraverso programmi specifici (algoritmi) che forniscono istruzioni ad esempio per l'analisi dei dati. Tale apprendimento può essere supervisionato o non supervisionato dall'uomo.

serie di falsi positivi e negativi che possono essere individuati ed eventualmente esclusi attraverso tecniche di analisi dei contenuti (*content analysis*) di ogni singolo *hate speech* estratto. In questo modo da un focus esclusivo sulle sole parole di odio si estende l'attenzione ai **discorsi di odio**.

Le prime attività di *content analysis* abbinate alla tecnica di *Bag of words* (BoW) sono state manuali. Più nello specifico, quale tecnica specifica di *content analysis,* si è in prevalenza adottata la *sentiment analysis* manuale realizzata da team di linguisti e neuro scienziati, mirata a leggere e classificare manualmente ogni contenuto come positivo, neutro o negativo[3] sulla base della polarizzazione del discorso; in questo modo è stato possibile escludere falsi positivi, cioè testi contenenti le parole chiave ma senza finalità di odio. Le metodologie che hanno integrato tecniche di estrazione dei dati automatizzate e tecniche di analisi dei contenuti manuali hanno quindi adottato un mix dei due approcci succitati, identificando un terzo approccio che si può considerare **quali-quantitativo**.

La tecnica di *sentiment analysis* manuale implica, tuttavia, un cospicuo impiego di tempo e risorse, ponendo rilevanti limiti ex ante all'estensione temporale del monitoraggio. La ricerca successiva si è quindi specializzata nell'introdurre una sempre maggiore automazione della *sentiment analysis* attraverso l'introduzione di algoritmi per identificare ed escludere i falsi positivi e i falsi negativi, rendendo possibile l'analisi di dati sempre più ampi e in tempi sempre più veloci (Silva et al., 2016; Burnamp and Williams, 2015). Gli algoritmi per l'individuazione e analisi dei contenuti attraverso parole chiave vengono quindi continuamente perfezionati nella loro capacità di individuare falsi positivi attraverso modelli sempre più evoluti di *machine learning* (Corazza et al., 2019), nel cui ambito si sono sviluppati i prevalenti filoni di analisi per la classificazione dei discorsi di odio. In questi modelli la supervisione umana esercita la funzione fondamentale di definizione del modello iniziale su cui l'intelligenza artificiale va poi ad agire verificando gli avanzamenti (*i.a.* Burnamp and Williams, 2015).

La tecnica BoW di estrazione e monitoraggio dei flussi di *hate speech*, anche quando integrata con un'analisi dei contenuti realizzata attraverso la *sentiment analysis* ignora, tuttavia, sia la sequenzialità delle parole sia qualsiasi contenuto sintattico o semantico, portando sovente ad un'errata classificazione dovuta alla polisemia di alcuni termini e all'assenza di analisi sulla sequenzialità degli stessi. (Davidson et al., 2017; Burnap and Williams, 2015). Per superare tale criticità, è quindi emersa la necessità di adottare ulteriori tecniche di analisi dell'intero contenuto del discorso.

In questo senso, una tecnica particolarmente efficace è la ***semantic tagging***, soprattutto nel caso di parole chiave polisemiche, cioè con significati molteplici tra i quali solo una parte riferibile a odio e intolleranza, il cui utilizzo è in grado di ridurre la presenza di falsi positivi nei processi di identificazione dei contenuti di odio online. Questa attività consente di identificare e scartare i contenuti in cui la parola chiave collegata all'odio ha, in realtà, un significato differente (Musto et al., 2016). Similmente, anche l'utilizzo della ***lexical parsing***, consente di approfondire ulteriormente l'analisi del discorso focalizzandosi sugli elementi della sua struttura grammaticale.

Per migliorare la base informativa su cui realizzare la *content analysis* recenti esperienze/metodologie[4] hanno inoltre previsto la creazione di **piattaforme condivise** e l'adozione di **modalità collaborative** che coinvolgono attivamente utenti e società civile nelle attività di individuazione degli *hate speech* online.

Il monitoraggio degli *hate speech* può prevedere anche il ricorso ad ulteriori tecniche di analisi come la ***network analysis***[5] per ampliare il monitoraggio all'estensione e natura del sistema di relazioni virtuali di chi genera e diffonde odio sul presupposto che quanto più estesa è la rete di relazioni tanto più è possibile che l'*hate speech* si diffonda. La *network analysis* consente, in particolare, di monitorare l'estensione, la crescita e l'evoluzione

3 *http://users.humboldt.edu/mstephens/hate/hate map.html.*

4 *Hatebase*a livello globale (https://Hatebase.org/about) e il recente *Hatemeter*, sviluppato a livello europeo per identificare e monitorare l'*hate speech* online contro l'Islamofobia (*http://hatemeter.eu/*). Per maggiori dettagli su queste due esperienze che sono oggetto della presente mappatura si rinvia alle loro schede di analisi specifiche presentate nell'allegato **1** - Presentazione delle esperienze e metodologie di analisi degli *hate speech*).

5 La network analysis è una metodologia di analisi delle relazioni e interazioni fra gli individui. Con riferimento agli *hate speech* si focalizza per esempio, sul numero di "mi piace" /reazioni positive al testo del messaggio/ condivisioni del messaggio, numero di risposte e altre modalità di relazione in grado di individuare il sistema di relazioni alla base della condivisione e diffusione dell'*hate speech*.

di questi network di odio e la conseguente propagazione e persistenza dei messaggi di odio online, così come la natura umana o virtuale (*bot*)[6] dei partecipanti alla rete di relazioni (Himelboim et al., 2013).

Un'altra rilevante tecnica di analisi è **la geolocalizzazione**[7] che attraverso uno dei suoi output principali costituito dalle **mappe sulla geografia dell'odio** realizzate rispetto ad uno specifico contesto, è in grado di collegare i comportamenti verificatisi online con il mondo reale (offline). Esempi di adozione di questa tecnica si rinvengono in diversi progetti realizzati in questi ultimi anni[8]. L'output cartografico utilizza solo il subset di *hate speech* online geo-referenziato, che costituisce una parte minoritaria dei flussi di *hate speech* online. Tuttavia, pur agendo su una parte limitata del corpus di *hate speech* online, consente di individuare luoghi in cui il fenomeno dell'odio online è più intenso.

Ulteriori tecniche di analisi sono state recentemente sviluppate nell'ambito di esperienze/metodologie che si focalizzano sui post presenti nelle pagine e nei profili dei gruppi di odio, o nei forum radicali ospitati in piattaforme come 4chan e in generale nel **dark web**[9]. L'interesse verso queste dimensioni del web è in crescita: i dati mostrano, infatti, il significativo volume di discorsi d'odio generati in questi luoghi virtuali. Inoltre, sono sempre più documentati gli attacchi di odio virtuale promossi e coordinati attraverso questi forum verso profili e contenuti di Twitter, Facebook e Youtube (Hine et al.).

L'analisi della parte oscura del web evidenzia inoltre notevoli differenze tra i discorsi di odio online sulla base del luogo virtuale in cui sono generati. In particolare, è emerso come il linguaggio di odio nel dark web e nei forum radicali abbia caratteristiche sostanzialmente differenti e necessiti di tecniche di *content analysis* specifiche per realizzare un monitoraggio efficace (Correa et al., 2015). In questo contesto, si sono altresì sviluppate le tecniche di analisi dei raid per comprendere come gli utenti del dark web pianificano e organizzano attacchi a specifici profili e social network tradizionali.

Alcune tecniche di analisi degli *hate speech* sviluppate da alcune esperienze/metodologie offrono anche la possibilità di ancorare la narrazione virtuale dell'*hate speech* al contesto reale in cui agiscono gli utilizzatori dei social network e alle loro caratteristiche (analisi dei profili) e di analizzare gli andamenti con riferimento agli *hate speech* in prossimità di eventi specifici (come manifestazioni, episodi di cronaca, ricorrenze e specifici eventi) sia come causa dei fenomeni che in chiave predittiva.

Recentemente, infine, si sono realizzate esperienze/metodologie che coniugano le tecniche di monitoraggio dell'*hate speech* online allo sviluppo di tecniche efficaci di contrasto basati sulla generazione, in tempi rapidi e in modalità capillare, di messaggi di contrasto (cd. *counterspeech* o contronarrazione), cioè risposte ai contenuti di odio e di estremismo diffusi attraverso messaggi di disaccordo e/o campagne di disapprovazione e/o irrisione (Binny et al., 2018). Per contrastare efficacemente attacchi coordinati di *hate speech* online, il *counterspeech* può essere previsto, ricorrendo a sistemi di *machine learning*, in grado di individuare quali siano i messaggi di contrasto più efficaci.

Il *counterspeech* è inoltre particolarmente caldeggiato come strategia di **contrasto all'odio** perché non prevede alcun conflitto con la tutela della libertà di espressione e risulta quindi applicabile in qualsiasi contesto a prescindere dall'esistenza di una espressa legislazione *anti-hate speech* (Assimakopoulos et al., 2017; Gerstenfeld, 2017; Gagliardone et al 2015).

Nel complesso, il ricorso ad algoritmi anziché alla discrezionalità dell'uomo ha consentito di ridurre i tempi di individuazione dei contenuti di odio online e, altresì, una maggiore trasparenza nei processi di identificazione e rimozione degli *hate speech*. La diffusione delle tecniche di individuazione e analisi degli *hate speech* online ha infatti consentito di sviluppare percorsi di restituzione dei risultati particolarmente interessanti, sia nella

6 Si tratta di programmi informatici che accedono a internet e ai social network con gli stessi caratteri degli esseri umani e si relazionano con altri utenti.

7 Tecnica che consente l'identificazione della posizione geografica di un soggetto autore dell'*hate speech*.

8 Alcuni esempi sono dati da *The Hate Map https://users.humboldt.edu/mstephens/hate/hate_map.html* per il contesto statunitense o la Mappa dell'Intolleranza, per il contesto italiano *http://www.voxdiritti.it/ecco-le-mappe-di-vox-contro-lintolleranza*.

9 Il dark web (in italiano: web oscuro o rete oscura) è la terminologia che si usa per definire i contenuti del World Wide Web nelle darknet (reti oscure) che si raggiungono via Internet attraverso specifici software, configurazioni e accessi autorizzativi. Il dark web è costituisce quindi una parte di web che non è indicizzata da motori di ricerca.

prospettiva di dare una dimensione reale agli *hate speech*, ancorando cioè la dimensione virtuale (online) alla dimensione reale (*offline*), sia con riferimento alla realizzazione di progetti "open" (progetti aperti e accessibili a chiunque) per il miglioramento degli algoritmi. Gli algoritmi in formato aperto sono, infatti, disponibili e co - progettabili, e tale aspetto concorre ad aumentare l'efficacia di molte strategie di contrasto all'*hate speech* online che possono essere avviate anche dalla società civile.

In conclusione, il filone di studio e ricerca che ha ad oggetto l'individuazione e il monitoraggio dell'*hate speech* online è un ambito in continua evoluzione. A questo trend positivo stanno attivamente contribuendo molte istituzioni, nazionali ed internazionali, supportando ricerche multidisciplinari in cui le competenze di neuroscienziati, criminologi, linguisti, psicologi, sociologi vengono unite a quelle degli scienziati informatici per realizzare sistemi di *machine learning* in grado di identificare con sempre maggiore precisione gli *hate speech* online, chi li genera e come si diffondono, sia nelle piattaforme dei social media (*surface web*), sia nel *dark web*. Le stesse competenze si stanno, altresì, estendendo al filone di studio relativo al *counterspeech*.

2.1

Metodologie di rilevazione e analisi degli *hate speech* in ambito razziale: alcune esperienze italiane ed internazionali a confronto

In questi ultimi anni diverse sono state le esperienze e i progetti che si sono concentrati sulle metodologie di rilevazione ed analisi degli *hate speech* in ambito razziale. Di seguito se ne presentano alcune che sono state individuate mediante una ricognizione desk che ha portato all'individuazione di 9 esperienze/metodologie di osservazione degli *hate speech* maturate in parte in ambito accademico all'interno di progetti di ricerca europei (soprattutto REC - *Rights, equality and citizenship programme* (2014-2020) o grazie a finanziamenti privati), ad opera di partnership formate da esperti di differenti Paesi Europei o realizzate a livello internazionale. Più nello specifico, 5 esperienze hanno visto l'Italia quale coordinatore; 1 l'UK; 1 la Grecia, 1 il Canada, e 1 il Brasile.

Quattro esperienze (Emore, Hatemeter, Mandola, React) si caratterizzano per un **approccio multi-paese**, cioè la metodologia si applica in tutti i paesi partner, mentre 1 sola esperienza (la Mappa dell'intolleranza) prevede che la metodologia sia applicata solo sul territorio italiano.

La dimensione multi-paese consente, in particolare, di **realizzare comparazioni tra paesi e di individuare trend sovra-nazionali** per quanto concerne caratteristiche e modalità di propagazione degli *hate speech,* così come per quanto riguarda le tecniche di analisi.

Allo stesso tempo, studi che si concentrano su un singolo contesto geografico consentono di prendere in considerazione caratteristiche dettagliate e particolari di quel contesto specifico. Ad esempio, la Mappa dell'Intolleranza si concentra solo sul contesto italiano per esplorare nel dettaglio le sfumature linguistiche proprie della lingua italiana.

Quattro esperienze (Hatebase, Hatelab, *Hate speech* e dark web - 4chan, Analyzing the Targets of Hate in Online Social Media), inoltre, si caratterizzano per aver adottato metodologie di monitoraggio e analisi che possono essere applicate ovunque a livello globale e non solo nei paesi partner dei progetti in esame. Si tratta di esperienze maturate in ambito accademico o per le quali si prevedono sviluppi a fini commerciali, che hanno quindi previsto la creazione di vocabolari multilingue per l'identificazione delle parole chiave necessarie all'identificazione degli *hate speech*.

Sulla base delle principali informazioni raccolte, è stato possibile realizzare una tavola sinottica che mette a confronto e sintetizza le principali caratteristiche delle esperienze/metodologie mappate a partire dalle piattaforme social e livello internet analizzato (*Surface Web* o *Dark Web*); ambito discriminatorio di analisi/target individuato ed ambito geografico di applicazione della metodologia (monopaese o multi paese, con dettaglio dei paesi coinvolti, oppure globale).

Ma soprattutto la tavola intende mettere a confronto, compatibilmente alle informazioni disponibili, l'approccio adottato dalle metodologie di analisi (qualitativo, quantitativo o quali-quantitativo), le modalità adottate per l'identificazione delle parole chiave, le tecniche di *content analysis* utilizzate (*Sentiment analysis; Semantic tagging; Lexical parsing*), evidenziare il ricorso alla Geo-analisi che consente una miglior conoscenza del contesto e alla *Network analysis* che permette invece di ricostruire le interazioni tra l'autore dell'*hate speech* e il resto dell'utenza e misurare in questo modo il livello di propagazione dei messaggi di incitamento all'odio.

Particolare attenzione è stata riservata anche alla realizzazione di ulteriori e particolari attività di analisi che costituiscono il portato di innovazione di ciascuna metodologia (ad esempio, monitoraggio etnografico; studio dei profili e classificazione degli autori degli *hate speech*, analisi dei collegamenti tra *hate speech* e specifici eventi e previsione di futuri attacchi e futuri bersagli, etc.) e ad **azioni complementari** rispetto all'attività di identificazione degli *hate speech* online (ad esempio, attività di contro-narrazione).

Non di meno, a seguito della ricognizione e dell'analisi effettuata, è apparsa di tutta evidenza l'importanza di evidenziare nella descrizione delle esperienze e metodologie mappate quali di esse prevedano **strumenti di diffusione, coinvolgimento e *accountability*** pensati non solo per diffondere e rendere visibili i risultati delle analisi (ad esempio, App, *Dasboard*, Mappe, Tabelle dati, Workshop/Focus group), ma anche per coinvolgere attivamente stakeholder di rilievo o addirittura l'intera collettività nell'analisi o in azioni propedeutiche a questa, attraverso l'ideazione di **piattaforme/strumenti e processi partecipativi**.

Tavola 2.1 Tavola sinottica metodologie: principali caratteristiche e tecniche di analisi adottate

	Principali caratteristiche delle esperienze/metodologie											
N	Esperienze/ metodologie e soggetto attuatore	Ambito discriminatorio di analisi/target	Social network analizzato	Livello Internet analizzato	Ambito di applicazione	Identificazione parole chiave e estrazione dati	Tecniche di content analysis	Geoanalisi	Network analysis	Strumenti di diffusione, coinvolgimento attivo, accountability	Azioni complementari all'attività di identificazione degli *hate speech*
1	Emore (RiSSC - IT)	Target multidimensionale che include: - colore della pelle - etnia - razzismo - altro (religione, disabilità, sessismo, orientamento sessuale)	Twitter Facebook	Surface Web	Multipaese: (IT, BE, RO,UK,DE, PT,CY,SL,MT)	Di nuova elaborazione Creazione team di esperti multipaese per individuare parole chiave sull'odio Parole chiave multi-lingua (paesi coinvolti) Tecnica BOW basata su algoritmi	Sentiment analysis Manuale	Si	No	Modalità bottom-up e condivisa per il monitoraggio: App per segnalare *hate speech* Dashboard per visualizzazione e consultazione database *hate speech* da parte dell'utenza	Creazione team di esperti multipaese per armonizzare azioni di contrasto all'*hate speech* online a livello europeo
2	HATEBASE metodology (Hatebase, Canada)	Target multidimensionale che include: - colore della pelle - etnia (ROM, SINTI) - razzismo - altro (religione, genere, orientamento sessuale, disabilità, status sociale)	Facebook Twitter Youtube Pagine web media	Surface web	Globale	Di nuova elaborazione e modalità bottom-up nell'identificazione (coinvolgimento utenza) Parole chiave multi-lingua Tecnica BOW basata su algoritmi	Sentiment analysis Algoritmo	Sì	No	Modalità bottom-up e condivisa per il monitoraggio: Piattaforma WIKI per inserire i vocaboli di odio multilingue da parte degli utenti Accesso libero e gratuito ai vocabolari multilingua su parole d'odio per tipologie di utenza quali accademia, NGO e cittadini; Mappe dei vocaboli di odio visualizzabili Tabelle dati sui vocaboli di odio scaricabili Dashboard per visualizzazione e consultazione database *hate speech*/ risultati analisi	Sviluppo algoritmi predittivi di eventi di odio nel mondo reale sulla base dei trend negli *hate speech* online

Tavola 2.1 Tavola sinottica metodologie: principali caratteristiche e tecniche di analisi adottate

	Principali caratteristiche delle esperienze/metodologie										
N	Esperienze/ metodologie e soggetto attuatore	Ambito discriminatorio di analisi/target	Social network analizzato	Livello Internet analizzato	Ambito di applicazione	Identificazione parole chiave e estrazione dati	Tecniche di content analysis	Geoanalisi	Network analysis	Strumenti di diffusione, coinvolgimento attivo, accountability	Azioni complementari all'attività di identificazione degli *hate speech*
3	HATELAB dashboard (University of Cardiff- UK)	Target multidimensionale che include: - colore della pelle - etnia - razzismo - altro (religione, disabilità, orientamento sessuale, genere)	Tutti i social networks	Surface web	Globale	Di nuova elaborazione Parole chiave multi-lingua Tecnica BOW di identificazione parole odio basata su algoritmi Identificazione BOW associate a possibili eventi	Sentiment analysis Algoritmo Lexical parsing Algoritmo Supervisione umana esternalizzata con piattaforme di crowfunding	Sì	Sì	Dashboard per visualizzazione e consultazione database *hate speech* da parte dell'utenza Mappe degli hate speach visualizzabili Non previsti piattaforme/strumenti partecipativi per il monitoraggio	Analisi relazione tra eventi significativi/potenzialmente scatenanti (fatti sociopolitici, commemorazioni, eventi di cronaca) e *hate speech* online
4	HATEMETER (Università di Trento - IT)	Razzismo (Islamofobia)	Twitter Facebook	Surface web	Multipaese (IT, FR,UK)	Di nuova elaborazione Parole chiave multi-lingua (paesi coinvolti); Tecnica BOW di identificazione parole odio basata su algoritmi	Sentiment analysis Algoritmo	Sì	Si	Dashboard per visualizzazione e consultazione database *hate speech* per utilizzo esclusivo da parte degli stakeholders: NGOs, giornalisti, attivisti (non utenza generalizzata) (in via di realizzazione) Accesso attraverso dashboard per gli attivisti a contenuti di contrasto generati automaticamente per avviare campagne di controNarrazione Workshop/Focus group	1. Analisi dei picchi di *hate speech* online; 2. Analisi degli influencer e dei nuovi profili online legati all'islamofobia; 3. Ideazione algoritmo per la produzione di contro-narrativa
5	MANDOLA (Foundation for Research and Technology GR).	Target multidimensionale che include: - colore della pelle - etnia - razzismo - altro (religione, disabilità, sessismo, omofobia, status sociale)	Twitter Google	Surface web	Multipaese (GR, IE, FR, ES, BG,CY)	Parole chiave individuate dal Vocabolario Hatebase Parole chiave multi-lingua (paesi coinvolti) Tecnica BOW di identificazione parole odio basata su algoritmi	Sentiment analysis Algoritmo Lexical parsing Algoritmo	Sì	No	Dashboard per visualizzazione e consultazione dei dati da parte dell'utenza (modalità user-friendly) Mappe degli hate speach visualizzabili Tabelle dati esportabili e grafici Portale per segnalazioni (in via di realizzazione). Possibilità di consultazione prevista per le FF.OO App per segnalare e visualizzare i dati (in corso di definizione)	Analisi normativa e classificazione degli *hate speech* online in legali e illegali

Tavola 2.1 Tavola sinottica metodologie: principali caratteristiche e tecniche di analisi adottate

	Principali caratteristiche delle esperienze/metodologie											
N	Esperienze/ metodologie e soggetto attuatore	Ambito discriminatorio di analisi/target	Social network analizzato	Livello Internet analizzato	Ambito di applicazione	Identificazione parole chiave e estrazione dati	Tecniche di content analysis	Geoanalisi	Network analysis	Strumenti di diffusione, coinvolgimento attivo, accountability	Azioni complementari all'attività di identificazione degli *hate speech*
6	MAPPA INTOLLERANZA (Voxdiritti - IT)	Target multidimensionale che include: - colore della pelle - etnia - razzismo - altro (religione, disabilità, sessismo, orientamento sessuale)		Surface web	Monopaese Twitter (Italia)	Di nuova elaborazione ad opera di team di sociologi e linguisti Parole chiave mono-lingua (IT) Tecnica BOW di identificazione parole odio basata su algoritmi	Sentiment analysis Algoritmo Semantic tagging Algoritmo	Sì	No	Mappe degli hate speach visualizzabili dall'utenza (modalità user-friendy) Non previsti piattaforme/strumenti partecipativi per il monitoraggio	No
7	Analyzing the Targets of Hate in Online Social Media (Federal University of Minas Gerias - Brasile)	Target multidimensionale che include: - colore della pelle - etnia - razzismo - altro (comportamento, orientamento sessuale, classe sociale, genere, disabilità, religione, aspetto fisico)	Whisper Twitter	Dark web Surface web	Globale	Parole chiave individuate dal Vocabolario Hatebase e modalità top-down nell'identificazione Parole chiave multi-lingua Tecnica BOW di identificazione parole odio basata su algoritmi (prima estrazione) Creazione di algoritmi che integrano parole chiave a strutture grammaticali per seconda estrazione dei dati	Sentiment analysis Manuale Algoritmo Lexical parsing Manuale	Sì	Sì	No	Analisi differenze tra *hate speech* online su piattaforme anonime e piattaforme non anonime
8	*Hate speech* e dark web - 4chan (Università Roma 3 - IT)	-colore della pelle - etnia - razzismo	4chan	Dark web Surface web	Globale	Parole chiave individuate dal Vocabolario Hatebase Parole chiave multi-lingua Tecnica BOW di identificazione parole odio basata su algoritmi	Sentiment analysis Manuale	Sì	Sì	No	Analisi dei raid (attacchi a profili specifici di persone nel surface web organizzati nel dark web) Studio di tecniche di analisi basate su algoritmi per prevedere tali raid
9	REACT (ARCI -IT)	Razzismo/ Islamofobia	Multipli, (social media, e siti) in corso di definizione	Surface web	Multipaese (IT, FR, ES, DE, UK)	In corso di definizione	In corso di definizione	Sì	Sì	in corso di definizione	1. Definizione algoritmi per estrazione dati su contronarrazione; 2. Creazione banca dati su contronarrazione; 3. Analisi per trasferibilità buone pratiche di contro-narrazione; 4. Azioni di educazione al contrasto; 5.Creazione rete internazionale per il contrasto all'*hate speech* online

Osservando la tavola sinottica, si possono evidenziare le più recenti evoluzioni per quanto concerne caratteristiche e tecniche di analisi e contrasto all'*hate speech* online.

In primo luogo, la maggior parte delle esperienze/metodologie mappate (6 su 9), si focalizzano su più ambiti di discriminazione, sono cioè **multidimensionali**, sul presupposto ormai consolidato che l'*hate speech* può riguardare differenti target o riferirsi contemporaneamente a più target (ad es. riguardare una donna di una particolare etnia e disabile). Per avere un quadro informativo completo, si persegue pertanto l'estrazione di dati con riferimento a un set di target più esaustivo possibile, modalità che consente di analizzare, i dati con riferimento alle possibili sovrapposizioni in caso di discriminazioni multiple. Due esperienze (Hatemeter e React) si focalizzano sull'i**slamofobia**, intesa come pregiudizio e discriminazione verso i musulmani. In questo caso si è ritenuto di assimilare tale ambito specifico di discriminazione al razzismo, uno degli aspetti su cui si focalizza la nostra analisi, sul presupposto che non si tratti di una discriminazione afferente alla religione ma che identifichi piuttosto un gruppo specifico di persone/minoranze straniere.

Dal punto di vista delle **dimensioni del web su cui si concentrano le analisi**, oggi i social network rappresentano l'ambito più investigato, poiché sono i luoghi in cui è più intensa la pubblicazione di contenuti da parte della generalità delle persone. In particolar modo, **Twitter è la piattaforma maggiormente analizzata**, poiché la sua caratteristica di essere una piattaforma con profili aperti consente un'agevole ed ampia estrazione di dati.

Nella fattispecie delle esperienze di osservazione/metodologie oggetto della mappatura, **la maggior parte (7 su 9) considerano per l'analisi più piattaforme online**. Più in dettaglio, Twitter è considerato per l'analisi da 6 metodologie, Facebook da 3, mentre Google, Wisper, Youtube e 4chan, sono considerati rispettivamente oggetto di analisi solo da una metodologia. La maggior parte (5) delle esperienze analizzate si focalizzano al massimo su 2 piattaforme online, tranne Hatebase, Hatelab, E-more e React che ne considerano più di 2, anche se in questo ultimo caso, secondo quanto risulta dalla documentazione disponibile, le piattaforme di analisi sembrerebbero ancora da definire.

Particolarmente interessanti sono le metodologie che si stanno concentrando sullo studio del ***dark web***. Il *dark web*, infatti, contiene piattaforme in grado di garantire l'anonimato, favorendo la creazione e la condivisione dei messaggi di incitamento all'odio. Tali metodologie appaiono particolarmente rilevanti se si considera che fino ad ora la maggior parte della letteratura si è concentrata soprattutto sul *surface web* relativamente al quale si stanno ormai consolidando tecniche per la *content analysis* degli *hate speech* basate su algoritmi sempre più avanzati e articolati.

Per quanto concerne l'estrazione dei dati da internet attraverso l'**identificazione di parole chiave** (*Bag of Words* o *seed terms*), le esperienze analizzate indicano sostanzialmente due modalità di individuazione:

1. la prima prevede di recuperare le parole chiave da metodologie già esistenti, tra i quali Hatebase che rappresenta certamente uno degli esempi più avanzati ed utilizzati;
2. la seconda modalità prevede invece l'adozione da parte della metodologia di parole-chiave di nuova elaborazione con il coinvolgimento nel progetto di team di esperti (linguisti, neuroscienziati, scienziati sociali e psicologi).

Più raramente, invece, è previsto il coinvolgimento della collettività per l'identificazione delle parole chiave, secondo una modalità ***bottom-up*** come previsto nella metodologia eMore. Se da un lato, il ricorso a grossi database come Hatebase consente di avere a diposizione una banca dati particolarmente estesa e aggiornata di termini sentinella, la modalità di individuazione ex novo delle parole chiave permette, tuttavia, una maggiore attenzione da parte della metodologia alle specifiche proprie della lingua di interesse.

Le esperienze analizzate mostrano che l'identificazione delle parole chiave rappresenta il passaggio preliminare per l'analisi dell'*hate speech* online. Una volta identificate e assegnate all'algoritmo di estrazione dei dati, consentono di ottenere un corpus di contenuti web con all'interno uno o più parole chiave tra quelle individuate. Al fine di eliminare dall'insieme dei contenuti quelli che, pur contenendo parole chiave, non rappresentano discorsi di odio, tutte le metodologie analizzate prevedono il ricorso alla ***content analysis***. L'analisi dei contenuti può articolarsi in diversi step, ognuno dei quali contribuisce a migliorare il processo di esclusione dei falsi positivi ampliando l'analisi all'intero discorso. Gli step di analisi sono dati principalmente dalla *sentiment analysis* (per la polarizzazione del discorso), dalla *semantic tagging* (per l'analisi dei termini polisemici, cioè che hanno più significati), e dalla *lexical parsing* (per l'analisi della struttura grammaticale del discorso).

Ognuna di queste tecniche di analisi può essere effettuata **manualmente**, in modo **automatizzato attraverso algoritmi**, o con **l'integrazione delle due modalità**. La quasi totalità delle esperienze/metodologie mappate effettuano la content analysis ricorrendo ad algoritmi. Ormai, quasi tutte le metodologie di analisi e non solo quindi quelle mappate prevedono il **ricorso almeno alla *sentiment analysis***. Alcune, come la Mappa dell'Intolleranza, la integrano con la *semantic tagging*, ritenuta particolarmente rilevante con riferimento alla lingua italiana, in quanto caratterizzata da molteplicità di termini aventi differenti sfumature e significati.

Altre metodologie, come Hatelab, Mandola e lo studio/metodologia "*Analyzing the Targets of Hate in Online Social Media*", integrano la *sentiment analysis* con la *lexical parsing*. Questa ultima metodologia, in particolare, fa ricorso inizialmente alla *lexical parsing* manuale per verificare se esistano strutture grammaticali ricorrenti negli *hate speech* estratti e, una volta individuate, le utilizza per realizzare un algoritmo per l'identificazione degli *hate speech* che viene testato sul resto dei dati.

La tavola che segue riepiloga con riferimento alle esperienze/metodologie mappate il processo di identificazione e analisi degli *hate speech* a partire dalla prima fase di estrazione dei dati sulla base di parole chiave (Bow) fino alla sua identificazione sulla base dell'intero contenuto del discorso a seguito dell'applicazione di tecniche di *content analysis* che, come si può vedere, nella maggior parte dei casi fanno ricorso all'utilizzo di algoritmi. Solo 2 delle esperienze analizzate (e-More e *Hate speech* e dark web - 4chan) effettuano una *content analysis* (più precisamente *sentiment analysis*) esclusivamente manuale.

Tavola 2.2 Processo di identificazione e analisi degli *hate speech*

		Fase 1 Uso di parole d'odio (BoW) per l'estrazione di dati da internet attraverso algoritmo		Fase 2 Discorsi d'odio individuati attraverso algoritmo a partire dai dati estratti attraverso ricerca per parole d'odio		
N	**Esperienze/ metodologie e soggetto attuatore**	**Parole d'odio individuate da team di progetto**	**Parole d'odio prese da librerie esistenti realizzate da altri progetti**	**Discorsi d'odio individuati con sentiment analysis**	**Discorsi d'odio individuati con semantic tagging**	**Discorsi d'odio individuati con parsing**
1	Emore (RiSSC - IT)	✓	✗	manuale	✗	✗
2	HATEBASE	✓	✗	✓	✗	✗
3	HATELAB	✓	✗	✓	✗	✓
4	HATEMETER	✓	✗	✓	✗	✗
5	MANDOLA	✗	✓	✓	✗	✓
6	MAPPA INTOLLERANZA	✓	✗	✓	✓	✗
7	Analyzing the Targets of Hate in Online Social Media	✗	✓	✓	✗	✓
8	*Hate speech* e dark web - 4chan	✗	✓	manuale	✗	✗
9	REACT1	-	-	-	-	-

Realizzare analisi attraverso algoritmi, consente non solo di effettuare monitoraggi e screening in modo automatico ma anche di estrarre, oltre al contenuto del post, **altri dati di notevole interesse**.

A questo proposito, la **geolocalizzazione rappresenta una tecnica di analisi sempre più utilizzata**, perché rende possibile raccogliere dati per la produzione di mappe in grado di rendere più agevole la visualizzazione del fenomeno e di confrontare geograficamente diversi contesti (urbano/rurale oppure diversi contesti nazionali). Tra le esperienze analizzate, Hatebase, la Mappa dell'Intolleranza e Mandola sfruttano la dimensione della geolocalizzazione anche per favorire una più ampia diffusione dei risultati presso gli stakeholder.

Analogamente, l'analisi dei dati relativi al numero di condivisioni, mi piace, commenti ricevuti da ogni post tramite la tecnica della ***network analysis***, **consente di far emergere la risonanza e il livello di condivisione degli** *hate speech* attraverso il web. Tra le esperienze mappate, Hatelab, Hatemer e React prevedono la *network analysis* per misurare il livello di propagazione dell'*hate speech* nel *surface web*, mentre lo studio "*Analyzing the Targets of Hate in Online Social Media*" e la metodologia *Hate speech e dark web - 4chan* la utilizzano per analizzare come l'*hate speech* si diffonda nel *dark web*.

Molte delle esperienze analizzate, soprattutto quelle che non si concentrano sulla sperimentazione di nuovi algoritmi o sull'esplorazione di dimensioni di internet ancora poco conosciute, prevedono un **sistema articolato di azioni e output per la diffusione dei risultati delle analisi**. A questo proposito, il panorama è ampio e articolato: **mappe e *dashboard*** rappresentano due degli strumenti più utilizzati.

Come già detto le mappe consentono di rappresentare e visualizzare il fenomeno anche nella sua dimensione geografica, mentre le *dashboard* costituiscono interfacce attraverso cui i cittadini e/o le NGOs o altri soggetti della società civile possono **accedere liberamente alla consultazione delle banche dati**, visualizzare report di analisi e scaricare i dati.

Le esperienze di Hatelab e Hatebase mostrano ***dashboard* particolarmente ricche di funzionalità**. Mandola sta invece validando attualmente la versione open-beta, in cui è già possibile navigare attraverso diverse funzioni. Altri progetti, come eMore e Hatemeter stanno predisponendo attualmente le proprie *dashboard*; eMore, in particolare, prevedendone il possibile **utilizzo da parte della collettività in generale** mentre Hatemeter solo per alcuni stakeholder (NGOs, giornalisti etc.)

Alcuni progetti hanno previsto anche la **produzione di App**, al fine di agevolare ulteriormente la fruibilità dei risultati delle analisi. Da questo punto di vista, sono in particolare due le esperienze che si caratterizzano per lo sviluppo di App: Mandola e eMore. L'App di eMore ha ottenuto feedback positivi da parte dei partecipanti al progetto, ai quali è stata data la possibilità di testarla. La App di Mandola è attualmente in fase di prototipazione.

Il progetto Hatebase ha già reso pienamente possibile per gli utenti la possibilità non solo di consultare i risultati, ma anche di **partecipare attivamente al monitoraggio** degli *hate speech*, consentendo di contribuire all'**aggiornamento del vocabolario multi-lingua** di parole chiave mediante una piattaforma Wiki accessibile attraverso una semplice registrazione e la compilazione di un form pre-impostato in cui inserire i termini individuati e altre informazioni di contesto correlate. Tale funzione si rivela particolarmente utile anche per altri progetti/metodologie, che possono attingere a un vocabolario aggiornato, senza dover impegnare risorse per la creazione di un vocabolario ex novo.

Anche altre esperienze/metodologie si caratterizzano per la realizzazione di strumenti che stimolano un ruolo attivo dell'utenza nel contrasto dell'*hate speech* online. Si tratta, quindi, di strumenti pensati fin dall'origine perché possano essere utilizzabili dalla collettività oppure da specifici soggetti quali: NGOs, istituzioni e policy maker. Un esempio in questo senso è dato da eMORE che prevede la **possibilità di effettuare segnalazioni su** *hate speech* **online attraverso l'App** già menzionata, adottando una modalità ***bottom-up* e condivisa per il monitoraggio**.

Infine, le esperienze/metodologie individuate e analizzate si distinguono per il panorama di **azioni complementari** previste ad integrazione della funzione principale di identificazione dell'*hate speech* online.

Hatebase, Hatelab e Hatemeter prevedono azioni automatizzate attraverso algoritmi per l'analisi del **legame tra** *hate speech* **online e eventi che si verificano nel mondo reale**. Più nello specifico, Hatebase usa i dati sull'*hate speech* online per **prevedere eventi di odio** nel mondo reale, mentre Hatelab verifica **quali eventi reali** (fatti di cronaca, eventi sociopolitici, ricorrenze, manifestazioni) **possano avere avuto un ruolo scatenante nei picchi di** *hate speech* **online**. Infine, Hatemeter contiene **algoritmi per monitorare i profili social degli influencer** in materia di *hate speech* online e la creazione di nuovi profili particolarmente attivi nella produzione e diffusione degli *hate speech* su internet.

Tra le azioni complementari molto interessanti appaiono quelle volte a sviluppare **strumenti relativi alla contronarrazione**, largamente considerata uno degli interventi più efficaci per il contrasto all'*hate speech*. Sono soprattutto le esperienze/metodologie che si concentrano su un unico target (ad esempio, Hatemeter e React) ad inserire il monitoraggio dell'*hate speech* online all'interno di una strategia più articolata che unisce la comprensione del fenomeno ad azioni di diffusione della cultura e delle buone pratiche di contrasto dell'odio.

Hatemeter contiene, infatti, **algoritmi in grado di generare automaticamente contenuti di contrasto** sulla base dell'*hate speech* estratto. Attraverso la medesima interfaccia (*dashboard*) con cui vengono diffusi i risultati dell'analisi, ad ultimazione del progetto, verrà anche garantito l'accesso da parte dei soggetti attivi nel contrasto degli *hate speech* ai contenuti di contrasto generati automaticamente attraverso algoritmi perché possano avviare campagne di contro-narrazione. Questi contenuti saranno modificabili manualmente dagli utilizzatori per renderli più aderenti ai singoli casi/contesti. La metodologia verrà testata con diverse ONG (una per Paese partner del progetto), prevedendo anche focus group e workshop per aumentare il coinvolgimento di questa tipologia di stakeholder. L'unione di algoritmi di estrazione e analisi degli *hate speech* con algoritmi per la creazione di contro-narrativa, renderà possibile **realizzare strategie di contrasto tempestive**, volte a mobilitare e coinvolgere attivamente la società civile. React, invece, prevede di integrare all'analisi dell'*hate speech* online anche l'analisi della contro-narrativa presente sul web.

Particolarmente interessante anche l'azione complementare all'analisi prevista da Mandola che si propone di arrivare a definire un **algoritmo in grado di classificare gli** *hate speech* **online in legali e illegali** sulla base della legislazione vigente nel paese in cui l'*hate speech* online si realizza. Inoltre, al pari di React, Mandola prevede di utilizzare i dati estratti da Internet per promuovere azioni di educazione e informazione sul tema dell'*hate speech* online e del suo contrasto.

Infine, lo studio/metodologia "*Analyzing the Targets of Hate in Online Social Media*" analizza se e come le caratteristiche dell'*hate speech* online cambino sulla base del luogo virtuale in cui viene prodotto e diffuso, partendo dalla consapevolezza che i contenuti generati nel *dark web* possano avere maggiore radicalità, date le maggiori garanzie sull'anonimato garantite in questa parte del web (la piattaforma su cui viene effettuata l'analisi è infatti wisper che garantisce l'anonimato agli utenti).

L'esperienza *Hate speech* e dark web - 4chan, altresì, sviluppa **algoritmi per analizzare e prevedere come utenti del *dark web* si organizzano per realizzare attacchi** mediante *hate speech* **indirizzati a profili specifici sui social network nel *surface web***.

Anche queste ultime due metodologie, al pari di Hatebase, Hatelab e Hatemeter già menzionate, mirano non solo all'identificazione e analisi degli *hate speech* ma anche a fornire elementi ai policy maker per individuare strategie efficaci di prevenzione e contrasto degli *hate speech* online.

3

La strategia della contro-narrativa come strumento di contrasto dei messaggi di odio online

È evidente che un fenomeno diffuso e capillare come il razzismo non possa essere affrontato solo con norme giuridiche ma richieda un grande lavoro di tipo culturale.

Il proliferare dell'*hate speech* online e la sua crescente centralità nel dibattito pubblico hanno aumentato la spinta e linteresse ad individuare strategie di contrasto adeguate ed efficaci che richiedono di integrare l'approccio normativo anche con altri approcci a valenza educativa e culturale (i.a. Gagliardone et al., 2015).

Da questo punto di vista, una delle strategie più apprezzate è rappresentata dalla **contro-narrativa**, termine che identifica la contrapposizione all'*hate speech* online attraverso lo sviluppo di narrazione di contrasto costituita da un contenuto online che viene prodotto per rispondere ad un messaggio online di odio. Tale contenuto può mirare a delegittimare e confutare il messaggio di *hate speech* online e può anche includere ulteriori contenuti, volti a promuovere tolleranza, uguaglianza e rispetto delle differenze.

Nel secondo caso, al termine "contro-narrativa" alcuni autori sostituiscono il termine "narrativa alternativa" (Briggs e Feve, 2013). Di fatto, la quasi totalità delle metodologie non distinguono tra le due definizioni, usando il termine "contro-narrativa" per intendere il contrasto del contenuto dellhate speech sia che esso si limiti a negare un fatto oppure quando includa specifici contenuti volti a sostenere il rispetto dei diritti umani e a fornire prospettive differenti.

Il consenso crescente riscosso dalla contro-narrativa deriva dalla capacità di questa strategia di rispondere a numerose esigenze emerse nel dibattito sul contrasto all'*hate speech* online.

Infatti, il ricorso alla contro-narrativa, rende possibile realizzare unazione di contrasto che non ricade nella contrapposizione tra limitazione dell'*hate speech* online (censura) e libertà di espressione. Una contrapposizione, quest'ultima, particolarmente presente nel dibatto sul contrasto all'*hate speech* online nel contesto statunitense, e presente anche a livello europeo (Titley et al., 2017; Gagliardone et al, 2015).

Oltre a non confliggere con altri diritti dellindividuo, la contro-narrativa è una strategia di facile applicabilità, poiché può essere praticata agevolmente da un'ampia platea di soggetti: cittadini, associazioni, istituzioni. Inoltre, per essere realizzata, non necessita di leggi e/o regolamentazioni.

Un altro aspetto rilevante della contro-narrativa è che può agire sulle motivazioni che portano alcune persone a produrre *hate speech* online, inducendoli a cambiare prospettiva e a superare preconcetti e pregiudizi, una finalità che non può essere raggiunta attraverso proibizioni e regolamentazioni dell'*hate speech* online (de Latour et al., 2017; Briggs and Feve, 2013).

Prima di focalizzarsi sugli elementi individuati in letteratura come componenti essenziali delle esperienze/metodologie di contro-narrativa che costituiscono il focus dellanalisi, questo capitolo presenterà una parte più generale sul ruolo delle istituzioni e dei social networks/media nella promozione di contro-narrativa.

3.1

La contro-narrativa nelle agende istituzionali: documenti di policy e iniziative specifiche

La crescita degli *hate speech* online ha determinato negli ultimi anni una sempre maggiore attenzione nei confronti di tale fenomeno, confermata anche dall'aumento delle iniziative istituzionali intraprese a livello internazionale e nazionale che si sono concretizzate nella definizione di rilevanti documenti di policy, soprattutto piani di azione, e nell'adozione di specifiche raccomandazioni che hanno contribuito a delineare un sistema di politiche di intervento per il contrasto degli *hate speech* online, anche a carattere preventivo.

La contro-narrativa, in particolare, diversamente dall'attività di censura/rimozione dell'*hate speech*, viene sempre più promossa da istituzioni, associazioni, fondazioni e altre realtà della società civile per il contrasto dell'*hate speech* online, anche alla luce della sua non conflittualità con altri principi e diritti (RAN, 2017a, CoE, 2016).

Alcuni degli attori internazionali più attivi con riferimento al contrasto dell'*hate speech e,* particolare alla promozione della contro-narrativa, sono rappresentati dall'Organizzazione delle Nazioni Unite (ONU), dal Consiglio d'Europa e dalla Commissione Europea.

Nel 2016, l'ONU ha pubblicato il **Piano d'azione per prevenire l'estremismo violento** (PVE), che invita gli Stati membri a sviluppare i propri piani d'azione PVE, coinvolgendo gli attori rilevanti della società civile (ONU, 2016). Il Piano riconosce come le iniziative governative/nazionali non siano di per sé sufficienti e che gli Stati membri dovrebbero adottare anche approcci sovra-nazionali. Il Piano offre, inoltre, una serie di suggerimenti tra cui l'implementazione di strategie e programmi di sostegno alla contro-narrativa.

L'anno seguente, il Consiglio di Sicurezza dell'ONU ha pubblicato il **Programma Quadro Internazionale per Contrastare le Narrazioni Terroristiche** (ONU, 2017) che include la contro-narrazione tra i pilastri su cui realizzare un'adeguata azione di contrasto all'estremismo e alla radicalizzazione. Il Programma sottolinea che le istituzioni dovrebbero svolgere un ruolo di promotore piuttosto che di realizzatore diretto di contro-narrativa.

A livello internazionale, anche l'**UNESCO** ha sviluppato un **Approccio integrato per la promozione dei diritti umani e il contrasto alla radicalizzazione** (UNESCO, 2015). All'interno di tale approccio è stato dato particolare risalto all'importanza della contro-narrativa per prevenire l'aumento dell'intolleranza e la diffusione della radicalizzazione. L'Approccio identifica nelle giovani generazioni il soggetto chiave con cui interagire per promuovere una maggiore consapevolezza sul ruolo che la comunicazione positiva può rivestire per il contrasto degli *hate speech*. A questo specifico riguardo, l'UNESCO sottolinea l'importanza e promuove le azioni di supporto ai percorsi educativi evidenziando i rischi sociali derivanti dalla proliferazione dei discorsi di incitamento all'odio, nonché le potenzialità della contro-narrativa come forma di contrasto dell'*hate speech* online. L'UNESCO tra le proprie iniziative ha, altresì, avviato un percorso di condivisione sul tema con altre istituzioni rilevanti, come il Consiglio d'Europa.

Un altro soggetto internazionale impegnato a contrastare significativamente l'estremismo violento significativo è il **Global Counterterrorism Forum**, fondato nel 2011 da 30 paesi tra cui l'Italia a seguito dell'11 settembre. Il Forum è una piattaforma informale, apolitica e multilaterale contro il terrorismo. Tra le iniziative portate avanti dal Forum, risulta di particolare rilevanza la creazione dell'**Hedayah Center**, istituito formalmente nel dicembre 2012, focalizzato sulla promozione di programmi nazionali ed internazionali a sostegno del dialogo e della comunicazione per contrastare l'estremismo violento in tutte le sue forme e manifestazioni. In questo ambito, viene evidenziato il ruolo positivo svolto dalla contro-narrativa e ulteriormente valorizzata la "Counter-Narrative Library": un portale che contiene esempi positivi di contro-narrativa creata dal Hedayah Center.

Anche la **Global Coalition to Defeat Daesh**, che comprende 80 soggetti istituzionali tra cui 76 Stati e quattro istituzioni (Lega araba, UE, INTERPOL e NATO), assegna un ruolo chiave alla contro-narrativa come tipologia di comunicazione efficace per contrastare estremismo, radicalizzazione violenta e terrorismo.

Anche l'**OSCE** ha riconosciuto il ruolo della contro-narrativa come strumento efficace per il contrasto dell'odio e dell'intolleranza (OSCE, 2017). La centralità che la contro-narrativa deve avere nelle agende nazionali ed internazionali di contrasto all'estremismo e alla radicalizzazione è stata ribadita nella **Conferenza Internazionale "Prevenire e Contrastare l'Estremismo Violento e la Radicalizzazione che portano al Terrorismo",** tenutasi a Vienna nel 2017. L'OSCE ha anche realizzato alcune importanti iniziative, quali la campagna comunicativa #**UnitedCVE** nel 2015, per la promozione del rispetto reciproco, del pluralismo, dell'inclusione e della coesione. Tale campagna individua la contro-narrativa come principale strumento su cui la società deve impegnarsi per contrastare la diffusione dei discorsi di incitamento all'odio. L'OSCE ha anche realizzato materiale informativo che sottolinea l'importanza della contro-narrativa, specificatamente pensato per operatori istituzionali e policy-maker di paesi che si caratterizzano per un elevato rischio di radicalizzazione.

Particolarmente attivo sul tema è il **Consiglio d'Europa** a cui si deve la predisposizione del **Piano d'Azione per la Lotta contro l'Estremismo Violento e la Radicalizzazione 2015-2017** (CoE, 2015) adottato dal

Comitato dei Ministri nel maggio 2015. Tale Piano mira a tutelare i diritti umani e la democrazia prevenendo e contrastando la radicalizzazione violenta con misure di prevenzione. Il piano d'azione sottolinea, in particolare, la necessità di "*un'azione per prevenire la radicalizzazione violenta e aumentare la capacità della nostra società di rifiutare ogni forma di estremismo"* basata sull'istruzione e il coinvolgimento delle giovani generazioni. Secondo il Piano "*Educazione formale e informale, attività giovanili e formazione degli attori chiave (anche nei media, in campo politico e in campo sociale settori) hanno in questo senso un ruolo cruciale."* Il piano d'azione menziona specificamente la necessità di fornire strumenti per produrre contro-narrativa per il settore della formazione e sottolinea come sia fondamentale promuovere la produzione di contro-narrativa dal basso. Evidenzia, inoltre, la necessità di avere una comprensione più chiara del modo in cui i social media e Internet sono utilizzati come veicolo di radicalizzazione e, conseguentemente, di porre maggiore attenzione sulla prevenzione.

Una importante iniziativa del Consiglio d'Europa è rappresentata dal ***No Hate Speech Movement*** **(CoE, 2016)**, una campagna lanciata nel 2013 con destinatari i giovani europei, ideata per ridurre l'accettazione dei discorsi di odio online e per contrastare la loro "normalizzazione". Scopo del *No Hate Speech Movement* è soprattutto incoraggiare il rispetto della libertà di espressione, sviluppando risposte alternative all'istigazione all'odio, quali la prevenzione, l'educazione, la sensibilizzazione, lo sviluppo di autodisciplina da parte degli utenti e il sostegno alle vittime. La prima fase dell'iniziativa (2013-2015) ha avuto lo scopo di sensibilizzare al problema e mobilitare i giovani ad agire contro la diffusione degli *hate speech*. La seconda fase (2015-2017) si è invece focalizzata più sulle risposte educative e sulle strategie di prevenzione. È stata volta a consolidare i risultati ottenuti.

Il Consiglio d'Europa ha anche recentemente (2020) istituito il **Committee of Experts on Combating Hate Speech** (ADI/MSI-DIS). Due rappresentanti italiani, Roberto Bortone, coordinatore presso l'Ufficio Nazionale Antidiscriminazioni Razziali (UNAR) dell'Osservatorio Italiano Media e Internet - eletto fra i rappresentanti degli Stati Membri, e Federico Falloppa dell'Università di Reading e Responsabile delle Rete nazionale su contrasto ai discorsi e ai fenomeni d'odio - eletto comesperto indipendente, sono tra i 16 esperti che compongono il Comitato, il cui compito principale del Comitato è quello di preparare un progetto di raccomandazione del Comitato dei Ministri su un approccio globale per affrontare l'incitamento all'odio, anche online, all'interno del quadro dei diritti umani. Il Comitato si baserà sulla giurisprudenza della Corte europea dei diritti dell'uomo, sugli studi esistenti del Consiglio d'Europa, sui risultati della campagna per i giovani del movimento anti-incitamento all'odio, nonché su possibili pratiche utili a fornire una guida agli Stati membri e ad altre parti interessate in questo settore.

Anche l'**Unione Europea** ha promosso la contro-narrativa attraverso le agenzie e i programmi di sua competenza, nonché sostenendo iniziative esterne. Nel 2015, la Commissione Europea ha promosso la nascita dell'**EU Internet Forum** per creare un tavolo di discussione tra istituzioni e imprese operative nel settore ICT finalizzato a rendere internet un luogo più sicuro e tollerante, nel corso del quale è stata riconosciuta l'azione positiva svolta dalla contro-narrativa, riconoscimento ribadito, nel 2016, anche all'interno degli impegni sottoscritti da parte delle istituzioni europee congiuntamente a Facebook, Twitter, Youtube e Microsoft nell'ambito del **Codice di Condotta per Contrastare l'*Hate Speech* online** (EC, 2016). Analogamente, anche l'**Agenda Europea per la Sicurezza 2015-2020** sottolinea la rilevanza della contro-narrativa. La Commissione Europea ha, inoltre, promosso la ***Radicalisation Awareness Network*** **(RAN)**, che raggruppa migliaia di esperti e, attraverso uno specifico gruppo di lavoro istituito presso il suo **Centro di Eccellenza**, elabora analisi sulle esperienze di contro-narrativa esistenti, aggiorna e rende disponibili elenchi di progetti di contro-narrativa per agevolarne la conoscenza e la diffusione, produce manuali operativi per la realizzazione di azioni di contro-narrativa efficaci e fornisce supporto tecnico a progetti promossi in ambito europeo e nazionale.

Alla Commissione Europea si deve, infine, il supporto finanziario a numerosi progetti, sia europei che nazionali, volti alla creazione di contro-narrativa, attraverso il **Programma di Responsabilizzazione della Società Civile (CSEP), l'Internal Security Fund (ISF) e il Programma Diritti, Uguaglianza e Cittadinanza (REC).**

Per quanto concerne l'Italia nel maggio 2016 presso la Camera dei deputati è stata istituita **la Commissione su intolleranza, xenofobia, razzismo e fenomeni di odio (Jo Cox)**, i cui lavori hanno prodotto, nel 2017, una relazione finale che ha fornito specifiche raccomandazioni per prevenire e contrastare l'odio. Le raccomandazioni contemplano azioni da attuare a livello sociale, culturale, educativo ed informativo. Alcune delle sue raccomandazioni insistono sulla necessità di sostenere e promuovere blog e attivisti *no hate* o testate che promuovano la contronarrazione e campagne informative sui discorsi d'odio, soprattutto nel mondo non profit, delle scuole e delle università (raccomandazione n.13), così come sulla necessità di contrastare gli stereotipi e il razzismo sensibilizzando e responsabilizzando i media, specie online, ad evitare i discorsi di odio, (raccomandazione n. 14) e rafforzare l'educazione alla cittadinanza, finalizzata al rispetto, l'apertura interculturale, inter-religiosa e al contrasto dell'intolleranza e del razzismo.

3.2
Il ruolo dei social network e della società civile nella produzione di contro-narrativa per il contrasto dell'*hate speech* online

I social network riconoscono il ruolo centrale svolto dalla contro-narrativa nel contrastare la proliferazione dell'*hate speech* online, partecipano a diversi programmi e tavoli di confronto promossi dalle istituzioni e hanno attivato negli ultimi anni una serie di importanti iniziative per promuoverla.

Nel 2017, Facebook, Microsoft, Twitter e YouTube hanno creato congiuntamente il **GIFTC - Global Internet Forum to Counter Terrorism -** con lo scopo di scoraggiare gli estremisti violenti e i terroristi dall'utilizzo dei loro servizi. Questo luogo di dibattito e confronto viene utilizzato per scambiarsi buone pratiche derivanti dalle iniziative individuali che hanno intrapreso, testando e realizzando specifici strumenti all'interno delle rispettive piattaforme per la promozione della contro-narrativa.

Google ha avviato due progetti rilevanti: il progetto-pilota ***"Redirect"*** insieme a Jigsaw che reindirizza verso video di de-radicalizzazione tutti gli utenti che cercano video di propaganda dell'ISIS su **Youtube**; e il progetto-pilota ***"Creators for Change"*** attraverso cui Google favorisce la diffusione di messaggi focalizzati sulla tolleranza e il rispetto dei diritti umani ideati da una serie di autori di contro-narrativa che usano YouTube per diffonderli.

Facebook ha lanciato l'iniziativa **"Counterspeech"**, un progetto integrato che prevede la realizzazione di diverse attività, a partire dalla promozione delle iniziative di contro-narrativa realizzate dalle ONG sul social network. Il progetto prevede inoltre momenti di formazione specifica per gli operatori delle ONG sul tema della contro-narrativa e su come realizzarla utilizzando le caratteristiche di Facebook. Molta attenzione è rivolta, in particolare, alle giovani generazioni, attraverso il coinvolgimento diretto di studenti universitari nella realizzazione di campagne di contro-narrativa sostenute da Facebook. Infine, per favorire l'incontro tra i diversi soggetti sensibili e attivi sul tema della contro-narrativa, l'iniziativa "*Counterspeech*" organizza diversi *hackathon* durante l'anno, specificatamente volti a realizzare progetti di contro-narrativa coinvolgendo realtà e soggetti differenti.

Microsoft, infine, ha stipulato una partnership con l'Institute for Strategic Dialogue volta allo sviluppo di un prodotto analogo al *"Redirect"* di Google, ma applicabile a **Bing**, il motore di ricerca gestito da Microsoft. Il progetto pilota prevede di fare comparire contenuti specifici di contro-narrativa in caso di ricerche fatte su Bing in materia di estremismo.

Infine, gli stessi utenti del web si impegnano quotidianamente, in modo autonomo e spontaneo, nella produzione di contenuti di contro-narrativa, a volte riuscendo a dare vita a veri e propri progetti coordinati con altri cittadini nell'ambito di gruppi spontanei (The Guardian, 2019).

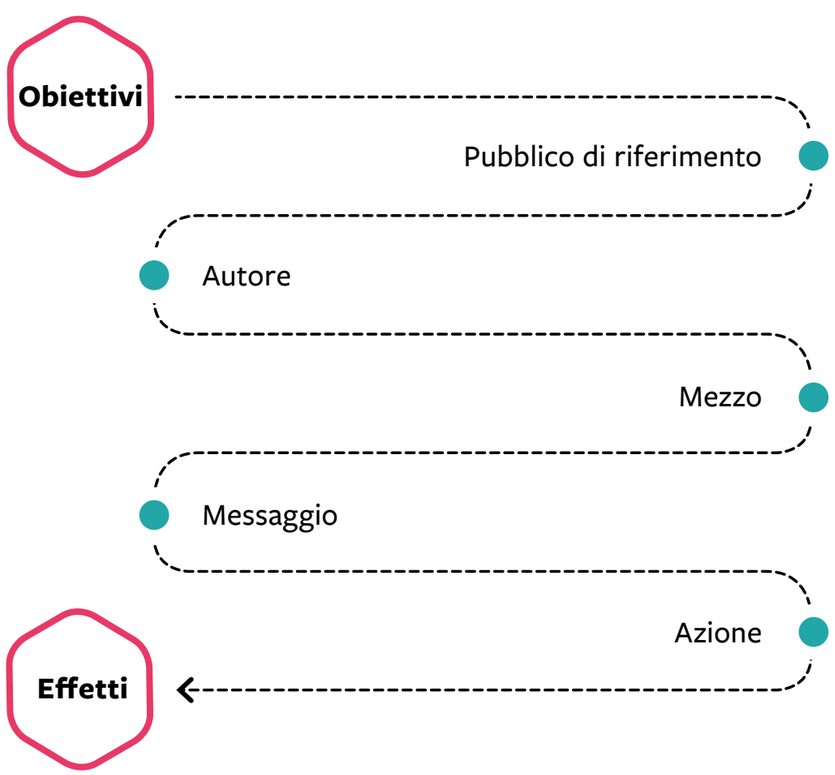
3.3
Analisi della letteratura e dei manuali operativi: indicazioni per la realizzazione, classificazione e valutazione della contro-narrativa

3.3.1 Gli elementi che caratterizzano una campagna di contro-narrativa secondo la letteratura

L'aumento delle attività di contro-narrativa si è accompagnata negli ultimi anni anche alla crescita della letteratura su questo tema. Molti sono i report di analisi sviluppati in ambito accademico/scientifico che ne definiscono le principali **dimensioni per la sua classificazione e valutazione**, da cui è possibile anche individuare gli elementi base che compongono una campagna di contro-narrativa e che devono essere tenuti in considerazione durante la sua definizione.

Le dimensioni individuate in letteratura sono sei: **Obiettivi, Target/Pubblico di riferimento, Autore, Mezzo, Messaggio, Azione, Effetti,** come illustrato dalla figura che segue (fig.3.1).

Figura 3.1 Esempio Le diverse dimensioni di classificazione di una campagna di contro-narrativa - Modello per la misurazione dell'efficacia (RAN 2017b)



Obiettivi. Le campagne di contro-narrativa sono raggruppate in primo luogo sulla base dell'obiettivo che si propongono di raggiungere. Spesso una campagna contempla più di un obiettivo.

Sono state, in particolare, identificate le seguenti categorie di obiettivi (de Latour et al., 2017):

- obiettivo educativo e di aumento della consapevolezza degli utenti online;
- obiettivo di contrasto degli *hate speech* online;
- mobilitazione degli utenti online;
- espressione di solidarietà e vicinanza alle vittime di specifici *hate speech* online e ai gruppi a cui appartengono;
- obiettivi di lungo termine (ad. es. aumento della cultura della tolleranza).

Effetti. Per realizzare gli obiettivi individuati, una campagna efficace di contro-narrativa deve essere in grado di ottenere uno o più dei seguenti effetti (Brown, 2016):

- ridurre la probabilità che la platea di utenti online riconosca un valore ai contenuti di *hate speech* esistenti e li diffonda;
- rendere gli utenti online consapevoli delle conseguenze e degli effetti *dell'hate speech*;
- ridurre la probabilità gli utenti online decidano di iniziare a produrre contenuti di *hate speech*;
- capacità di aumentare la propensione degli utenti online a prestare attenzione ai messaggi di contro-narrativa;
- intercettare e influenzare quella parte di utenti online che è attiva, anche se saltuariamente, nella produzione di *hate speech*;
- favorire la diffusione di una cultura di rispetto delle diversità e dei diritti umani.

Pubblico di riferimento. Il popolo di Internet rappresenta il pubblico di riferimento della contro-narrativa. Ciò implica che la stessa progettazione della contro-narrativa debba partire proprio dall'analisi dettagliata di questo target (Benesch et al 2016a). Similmente ad una campagna di comunicazione, ogni campagna di contro-narrativa deve quindi partire dall'**identificazione del pubblico di riferimento a cui si vuole rivolgere**, sulla base di caratteristiche quali l'etnia, l'età, il genere, la condizione economica e sociale, il contesto geografico di residenza.

Una volta identificato il pubblico di riferimento, occorre individuare e capire le **motivazioni** che potrebbero spingere questo segmento di utenti online ad essere potenzialmente ricettivi agli specifici contenuti di odio che la contro-narrativa vuole contrastare (de Latour et al., 2017).

Autore. Relativamente alle tipologie di autori di contro-narrativa, è emerso come profili diversi abbiano una capacità differente di raggiungere il pubblico di riferimento individuato e influenzarlo (RAN, 2017b, Benesch et al 2016a). Occorre, quindi, che la campagna di contro-narrativa definisca quali profili risultano maggiormente credibili nel veicolare i messaggi di contrasto con riferimento al pubblico di riferimento e all'obiettivo che si intende raggiungere (personalità famose, attori istituzionali, esponenti della società civile, cittadini). Può quindi risultare efficace predisporre account social differenti per un singolo autore e caratterizzarli in modo da raggiungere un differente pubblico (de Latour et al., 2017).

Mezzo. una campagna di contro-narrativa si caratterizza anche per il riconoscimento dei mezzi più efficaci volti a intercettare il pubblico di riferimento; a questo proposito sono riconosciute come rilevanti: la scelta del social network da utilizzare, le tipologie di contenuti da produrre (produzione di post, campagne sponsorizzate, creazione di pagine e/o profili dedicati, etc.) e l'integrazione di azioni sui social network con altre attività sia online (attraverso, ad esempio, pagine internet collegate ai contenuti prodotti sui social network) sia offline (attraverso, ad esempio, iniziative o campagne di comunicazione non social).

Messaggio. Definito l'interlocutore (autore) più indicato per il pubblico di riferimento che si vuole intercettare e il mezzo più efficace, occorre impostare contenuti (**messaggi)** in grado di **catturare l'attenzione** del segmento di interesse e di **creare coinvolgimento e riflessione**.

In questo ambito, è stata recentemente realizzata una tassonomia, largamente adottata dalla letteratura che si occupa dell'analisi della contro-narrativa e della sua efficacia, che si articola in diverse categorie di classifi-

cazione sulla base dell'impostazione dei singoli contenuti (Benesch et al., 2016b). Le categorie sono ricondotte principalmente a due macro-classi:

- **contro-narrativa costruttiva**, cioè volta a stimolare un dibattito informato
- **contro-narrativa non-costruttiva**, cioè volta a offendere o attaccare personalmente l'autore dell'*hate speech* online (Bartlett & Krasodomski-Jones, 2015).

La **contro-narrativa costruttiva**, classifica il contenuto secondo otto categorie di impostazione del messaggio:

1. impostazione basata sui **fatti**: presentazione di fatti e dati per sostenere che un contenuto di *hate speech* online non corrisponde al vero;
2. impostazione basata sull'evidenziazione delle **contraddizioni** contenute nel messaggio di odio;
3. impostazione basata sull'**avvertimento**: segnalazione dei possibili rischi individuali e delle relative conseguenze, anche giuridiche, collegate alla diffusione di un particolare messaggio di odio;
4. impostazione basata sull'**immedesimazione**: l'autore della contro-narrativa si identifica come appartenente allo stesso gruppo dell'autore dell'*hate speech* (eg: bianco, uomo, cristiano, etc.) o appartenente al gruppo che viene bersagliato dall'*hate speech*;
5. impostazione basata sulla **tossicità sociale**: l'autore della contro-narrativa sottolinea come l'*hate speech* contro cui rivolge la contro-narrativa rappresenti un elemento pericoloso e carico di odio che danneggia il contesto sociale; l'autore della contro-narrativa può anche identificare l'autore dell'*hate speech* come colui che genera il danno;
6. impostazione basata su **contenuti multimediali**: il messaggio di contro-narrativa in risposta all'*hate speech* può cioè utilizzare immagini o video;
7. impostazione **satirica**: il messaggio di contro-narrativa in risposta all'*hate speech* può far ricorso alla satira;
8. impostazione **emotiva**: il messaggio di contro-narrativa può essere impostato con tono empatico.

L'**impostazione basata sui fatti** può caratterizzarsi per una limitata efficacia nel far cambiare idea all'autore dell'*hate speech* online. Ciò avviene soprattutto perché le false credenze derivate da disinformazione sono particolarmente resistenti (Lewandowsky et al., 2012). Inoltre, più la reale conoscenza dei fatti è minima meno propense sono le persone a cambiare idea sulla base di correzioni e precisazioni indicate da sconosciuti (Kuklinski et al. 2000). Una contro-narrativa basata sul produrre informazioni e fatti che confutano la tesi sostenuta nell'*hate speech* online può addirittura arrivare a produrre un ulteriore irrigidimento dell'autore dell'*hate speech* sulle sue posizioni, rafforzandole (Nyhan & Reifler, 2015). Agire sui fatti può determinare, quindi, un'efficacia ridotta nei confronti degli autori di *hate speech* online; tuttavia, questa ridotta capacità può essere compensata dall' effetto positivo di questo tipo di contro-narrativa su quella parte di pubblico online che assiste in qualità di spettatore al dibattito e che può formarsi un'opinione assistendo agli scambi tra l'autore dell'*hate speech* e l'autore della contro-narrativa. Per questa platea, il ricorso a fatti e dati per confutare l'*hate speech* può rappresentare un elemento oggettivo, credibile e quindi non collegabile a posizioni parziali e/o a propaganda.

L'impostazione basata sull'evidenziare le **contraddizioni** contenute nel messaggio di *hate speech* potenzialmente può produrre sull'autore dell'*hate speech* online lo stesso effetto di irrigidimento che scaturisce dal controbattere con fatti e dati. Cionondimeno, l'effetto sugli spettatori al dibattito può risultare efficace e danneggiare la credibilità dell'autore dell'*hate speech* online (ADL, 2016).

L'impostazione basata sulla **segnalazione dei possibili rischi individuali e delle conseguenze personali** in cui può incorrere l'autore dell'*hate speech* online presenta gli stessi rischi delle impostazioni precedenti. Tuttavia, è un'impostazione che può indurre l'autore a rimuovere i post di odio, a fronte di un'enunciazione esplicita delle possibili conseguenze del suo messaggio, oltre al fatto che richiama l'attenzione da parte degli utenti sulle potenziali responsabilità dell'agire online, aspetto che tendono a dimenticare nei loro comportamenti virtuali (Suler, 2004).

L'impostazione basata sull'**immedesimazione** può realizzarsi invece secondo due modalità. La prima prevede di identificarsi come appartenenti allo stesso gruppo dell'autore dell'*hate speech*, per sfruttare il valore positivo

attribuito ai messaggi provenienti da membri degli stessi gruppi sociali. Questi messaggi vengono generalmente accolti con fiducia maggiore rispetto a messaggi analoghi provenienti da persone che appartenenti a gruppi sociali differenti. Attraverso questa modalità di **"affiliazione"**, l'autore della contro-narrativa **riduce la distanza percepita** tra lui e l'autore dell'*hate speech*, **facilitando l'instaurarsi di un dialogo**. L'immedesimazione con un determinato gruppo sociale può essere anche attuata attraverso l'identificazione con il gruppo sociale che viene bersagliato dall'*hate speech* online, con la finalità di "umanizzare" il soggetto bersaglio e favorire il superamento di stereotipi e preconcetti sulla base della teoria del contatto (Allport, 1958).

L'impostazione che porta a far emergere la **tossicità sociale** di un messaggio di *hate speech* online prevede invece di sottolineare che il contenuto dell'*hate speech* è percepito come produttivo di un effetto dannoso sul tessuto sociale. Questo **elemento di giudizio** può svolgere la funzione di avvertimento nei confronti degli autori degli *hate speech*, stimolando un'assunzione di responsabilità che può determinare la rimozione dei contenuti da parte dell'autore.

L'impostazione che prevede di produrre contro-narrativa attraverso **contenuti multimediali**, come *meme*[1], grafiche, infografiche, fotografie, GIF animate e video, sfrutta il fatto che questo tipo di contenuti, riuscendo a catturare sul piano emozionale, può generare un livello di coinvolgimento negli utenti maggiore rispetto a quello prodotto dai contenuti testuali, che generalmente attivano unicamente il piano razionale (de Latour et al., 2017).

L'impostazione della **satira** come risposta all'*hate speech* online si basa invece sulla convinzione che possa costituire un elemento utile per ridurre la scala della conflittualità (Marone, 2015). Inoltre, su Internet, i contenuti satirici si caratterizzano per una diffusione veloce e pervasiva (de Latour et al, 2017; Bartlett & Krasodomski-Jones, 2015; ADL, 2016), riscontrando un elevato gradimento da parte di un pubblico ampio. In effetti, come vedremo in dettaglio più avanti, con riferimento alla misurazione dell'efficacia delle diverse tipologie di contro-narrativa, il tono satirico risulta essere uno dei più efficaci.

L'impostazione che prevede il ricorso ad un **tono empatico** prevede di modulare il tono da utilizzare nel contenuto di contro-narrativa, sulla base che toni differenti possano produrre anche effetti differenti (de Latour et al., 2017). In particolare, toni ostili possono determinare la rimozione del contenuto di odio, ma difficilmente contribuiscono a fare mutare opinione all'autore dell'*hate speech*. Al contrario, un tono empatico può essere efficace nel costruire un dialogo, se accompagnato da un'attenzione a ridurre la distanza percepita tra autori di *hate speech* e autori di contro-narrativa, al fine di evitare che il tono sia percepito come paternalistico. L'utilizzo del tono empatico dovrebbe prevedere che l'autore di contro-narrativa sia in grado di compiere un'auto-analisi per verificare quanto sia intenso il suo livello di coinvolgimento personale così da riuscire a gestirlo durante l'azione di contro-narrativa. A questo specifico riguardo, è consigliato di individuare una serie di linee di confine che consentano al coinvolgimento empatico di non sfociare in esperienze troppo intense che possono determinare ripercussioni importanti nella sfera emotiva dell'autore e/o far sfociare la contro-narrativa in contenuti non più coerenti con gli obiettivi (de Laotur et al., 2017). Nel caso di campagne strutturate di contro-narrativa, è consigliato prevedere la figura di un mentore a cui gli autori di contro-narrativa possano rivolgersi in caso necessitino di un confronto sull'impatto emotivo che la campagna sta avendo sulla loro persona (de Latour, 2017).

Ovviamente, ogni contenuto di contro-narrativa può caratterizzarsi per la presenza di una o più categorie di impostazione del messaggio. Ad esempio, spesso la contro-narrativa esercitata attraverso la satira si compone anche di elementi multimediali, come video e *meme*. Analogamente, la contro-narrativa che utilizza fatti e dati, ricorre a elementi visuali come le infografiche. Il tono empatico può essere usato unitamente ad altre impostazioni, ad esempio con l'immedesimazione.

Oltre alla contro-narrativa costruttiva e alle sue otto sotto-articolazioni di impostazione del messaggio, come già anticipato, esiste una **contro-narrativa non-costruttiva**, la quale comprende tutti i contenuti di contro-narrativa che si contrappongono a specifici *hate speech* online attraverso messaggi che offendono e sviliscono gli autori o che contengono vere e proprie minacce (Bartlett & Krasodomski-Jones, 2015). In altre parole, la contro-narrativa non-costruttiva si realizza rispondendo all'odio con l'odio.

1 Un meme ha la forma di un'immagine, una GIF o un video, e si diffonde principalmente attraverso social network. Rappresenta in modo multimediale un'idea, uno stile o un'azione.

Azione. La campagna di contro-narrativa deve stimolare il compimento di azioni da parte del pubblico di riferimento, come, ad esempio, spingerlo ad aderire all'iniziativa, pubblicizzarla e manifestare gradimento attraverso i "like", o anche a diventare a sua volta un autore di contro-narrativa creando spontaneamente o in maniera coordinata e supervisionata messaggi di contro-narrativa. La campagna, inoltre, può anche stimolare gli autori dei post di *hate speech* a rimuoverli o a cambiare le proprie opinioni.

Dimensioni del confronto. La contro-narrativa può coinvolgere un solo autore di contro-narrativa o molti e può rivolgersi ad un solo autore di *hate speech* online oppure a molti. Su questa base vengono identificate 4 modalità, di coinvolgimento degli utenti ognuna caratterizzata da diversi livelli di efficacia (Wright et al. 2017), uno-a-uno, uno-a-molti, molti-a-uno, molti-a-molti:

1. La **modalità "uno-a-uno"** si realizza quando l'autore di contro-narrativa si rivolge esplicitamente e esclusivamente all'autore dell'*hate speech* online. Questa modalità spesso sfocia in lunghi scambi di messaggi tra autore della contro-narrativa e autore dell'*hate speech* online, molti dei quali possono non essere visibili perché si realizzano attraverso scambi privati. Questa modalità è ritenuta potenzialmente molto efficace (Titley et al., 2017), perché consente di sviluppare una strategia di tipo empatico e può favorire la riduzione della distanza con l'autore dei contenuti di odio, riducendo il rischio di un irrigidimento delle sue posizioni. Tuttavia, gli scambi privati rendono particolarmente difficile per il resto degli utenti di Internet, fruire dello scambio, determinando complessivamente un impatto ridotto della contro-narrativa.

2. La **modalità "uno-a-molti"** prevede che l'autore di contro-narrativa produca contenuti di contrasto diretti ad una vasta platea di autori di *hate speech* online che hanno generato messaggi identificabili, ad esempio, attraverso un determinato *hashtag*. Spesso, provocatoriamente, gli autori di contro-narrativa che si attivano in conversazioni uno-a-molti scelgono intenzionalmente di utilizzare lo stesso *hashtag* contro cui si sono attivati, al fine di aumentare la visibilità dei loro contenuti di contro-narrativa e aumentarne la portata.

3. La modalità "**molti-a-uno"** nasce come conseguenza di contenuti di *hate speech* online che diventano virali, attirando l'attenzione di molteplici utenti e spingendoli ad attivarsi per contrastarli. Uno dei rischi connessi a questa modalità è la cosiddetta "online *mob justice*[2]", cioè un generale aumento dei toni aggressivi e della polarizzazione delle opinioni che sfocia in volumi importanti di contro-narrativa non-costruttiva, senza però modificare la prospettiva di chi ha prodotto e/o diffuso l'*hate speech*. Diversamente, quando la modalità molti-a-uno viene adottata a seguito di progetti ed azioni di formazione di volontari e attivisti nel contrasto all'*hate speech* online (CoE, 2016) o da gruppi di persone che si sono auto-organizzate dotandosi anche di codici di condotta, la campagna può certamente avere un impatto significativo, senza portare ad atteggiamenti aggressivi e non-costruttivi.

4. Infine, la modalità "**molti-a-molti"** nasce a seguito di *hate speech* virali o a seguito di eventi scatenanti avvenuti nel mondo reale (manifestazioni, eventi politici, fatti di cronaca). In questo caso, la determinante - sia virtuale che reale - è in grado di generare il coinvolgimento di un'ampia platea di utenti che si attivano spontaneamente. Come già sottolineato per la modalità "molti-a-uno", la contro-narrativa può comportare il rischio di generare fenomeni di linciaggio online se si caratterizza per una maggioranza di contenuti non costruttivi. Tuttavia, la diffusione di progetti e iniziative volte a sensibilizzare rispetto all'utilità della contro-narrativa come azione di contrasto all'*hate speech* online e a fornire indicazioni su come attuarla in maniera corretta sta, aumentando il numero di persone che avviano azioni di contro-narrativa spontaneamente e che sono consapevoli dell'importanza di utilizzare toni e modalità corrette.

Il rischio di generare linciaggi online connessi ad alcune modalità di contro-narrativa evidenzia quindi che, occorre avere attenzione, sia nella fase di progettazione che in quella di realizzazione, a non ledere la dignità delle persone che hanno prodotto i contenuti di *hate speech* online contro cui si è attivata la contro-narrativa. Analogamente occorre verificare la veridicità delle informazioni e dei fatti che si portano a sostegno delle argomentazioni di contrasto.

2 Con il termine online *mob justice* (o online *public shaming*) si identificano quei fenomeni in cui gli utenti online identificano autonomamente soggetti che ritengono colpevoli e si attivano con vere e proprie azioni di linciaggio online.

3.4 Modelli operativi e linee guida per realizzare campagne di contro-narrativa

L'attività di analisi dedicata a comprendere e classificare la contro-narrativa unitamente al consolidamento di molte esperienze di comunicazione su questo tema, ha consentito di sviluppare negli ultimi anni una conoscenza articolata del fenomeno, da cui è stato possibile sviluppare dei veri e propri manuali/modelli operativi per la realizzazione pratica di efficaci attività di contro-narrativa. L'esistenza di modelli operativi si rivela tanto più strategica se si considera che l'accessibilità diffusa ai social network consente anche ai singoli cittadini di produrre contro-narrativa in modo spontaneo che va ad aggiungersi a quella prodotta in modo organizzato da realtà più strutturate/associazioni che si occupano di tutela dei diritti umani. Per questo motivo, alcune delle pubblicazioni più rilevanti a supporto della creazione di contro-narrativa sono state pensate per diverse tipologie di soggetto promotore-attuatore, sia avente natura informale che formale (RAN 2017a; RAN 2017b; Bartlett & Krasodomski-Jones, 2015).

Nel complesso, questi manuali operativi confermano la rilevanza degli elementi individuati in letteratura e precedentemente descritti, traducendoli in un insieme di procedure operative da attuare. I manuali operativi più rilevanti che si presentano di seguito sono stati sviluppati da istituzioni, social network e realtà attive nel settore (ONG e istituti di ricerca).

RAN: Il modello GAMMMA e le linee guida

Il Centro di Eccellenza del Radicalisation Awareness Network (RAN), la Rete di sensibilizzazione in materia di radicalizzazione promossa dalla Commissione Europea (RAN 2017b), ha prodotto il **modello GAMMMA** (Goal, Audience, Message, Messenger, Medium and Action), che definisce delle linee guida che possono essere utilizzate nelle diverse fasi di creazione della campagna di contro-narrativa (ideazione, realizzazione e valutazione). Le linee guida GAMMMA traducono in chiave operativa gli elementi che la letteratura ha identificato come fondamentali per l'efficacia della contro-narrativa e che abbiamo presentato nel paragrafo precedente. Il box che segue approfondisce il modello.

Box 3.1 Modello GAMMMA: Linee Guida per la creazione di una campagna di contro-narrativa

Il modello GAMMMA riconosce come elementi centrali di una campagna: l'obiettivo (Goal), il pubblico di riferimento (Audience), il contenuto (Message), l'autore (Messenger), il mezzo (Media) e le azioni (Action).

Relativamente all'**obiettivo** (Goal), la checklist sottolinea l'importanza di identificare lo scopo che si vuole ottenere con la contro-narrativa: ad esempio se si voglia modificare la prospettiva degli autori degli *hate speech* online oppure se si intenda evitare che quell'*hate speech* venga legittimato da altri utenti online, o entrambe le cose.

Per quanto riguarda il **pubblico di riferimento (Audience)** a cui si rivolge l'azione di contro-narrativa, il modello raccomanda di identificare in modo specifico a chi si rivolge l'azione di contro-narrativa, in termini di etnia, genere, orientamento religioso, sessuale, condizione socioeconomica e collocazione geografica.

Rispetto al **contenuto** (Message) la checklist specifica che va posta attenzione all'impostazione data al contenuto e che l'impostazione scelta deve essere in grado di produrre il risultato atteso sulla base del pubblico di riferimento considerato.

La checklist di GAMMMA riconosce anche il ruolo chiave dell'**autore** del contenuto (**Messenger**), il quale deve essere individuato in modo da risultare il più credibile possibile per il pubblico di riferimento individuato.

Il penultimo elemento del modello riguarda il **mezzo** di comunicazione (**Media**), che deve essere scelto sulla base della presenza del pubblico di riferimento su una piattaforma social piuttosto che un'altra.

Infine, vanno identificate con precisione anche la/le **azione/i** che devono essere stimolate dalla contro-narrativa (Action): condivisioni e/o "like", creazione spontanea o coordinata di nuova contro-narrativa, rimozione dei post di *hate speech* da parte degli autori, etc.

La checklist identifica inoltre una serie di metriche (indicatori) per valutare l'efficacia della contro-narrativa e di cui una campagna si deve dotare fin dalla sua pianificazione: percentuale di pubblico di riferimento raggiunta, numero di "like", numero di commenti, numero di condivisioni, numero di post di *hate speech* rimossi. Queste metriche consentono a chi ha realizzato la campagna di contro-narrativa di misurare il raggiungimento degli obiettivi (RAN Paper: "*How to measure the impact of your online counter or alternative narrative campaign*").

L'esperienza ha inoltre verificato l'utilità della checklist attraverso un'analisi condotta su venti campagne di contro-narrativa (RAN Collection of Approaches and Practices "*Preventing Radicalisation to Terrorism and Violent Extremism*"). Nell'azione di verifica, RAN ha integrato alle modalità di misurazione quantitativa, una serie di focus qualitativi con interviste in profondità ai soggetti che hanno sviluppato le campagne per valutare l'utilità della checklist per la realizzazione di una campagna di contro-narrativa (RAN Paper: "*LESSONS LEARNED. What to do and what not*!"). L'analisi ha confermato l'utilità della checklist.

Le linee guida prevedono anche indicazioni operative sull'impostazione della contro-narrativa per i diversi social network (Facebook, Twitter e Youtube fino a Instagram, SnapChat, Tumblr e Reddit) a partire dalle loro specifiche caratteristiche (RAN, 2017a). Ad esempio, con riferimento a Facebook, si evidenzia come la sua caratteristica di sfruttare le preferenze espresse dagli utenti sui propri interessi per mettere in evidenza nelle loro bacheche quei contenuti che sposano gli interessi espressi, possa essere utilizzata per raggiungere con più facilità il pubblico di riferimento della contro-narrativa su Facebook. Sarà quindi importante mappare preliminarmente quali siano gli interessi del pubblico così da costruire una campagna di contro-narrativa che li utilizzi e in questo modo spingere l'algoritmo di Facebook a mettere in evidenza la campagna stessa. Con riferimento a Twitter, le linee guida indicano di prestare particolare attenzione al tono utilizzato nel messaggio, mentre relativamente a Youtube, sottolineano che l'originalità è più importante della qualità del video, diversamente da Instagram, dove l'attenzione deve essere posta sulla la qualità dei media postati.

Al Centro di Eccellenza del Radicalisation Awareness Network (RAN) si deve la produzione di ulteriori manuali operativi[3] che insieme al Modello GAMMMA costituiscono nel complesso un patrimonio informativo in materia di contro-narrativa di indubbia utilità.

Anti -Defamation League: Buone Pratiche per Rispondere all'Hate Speech Online

Negli Stati Uniti, il gruppo di lavoro sul *Cyberhate* dell'Anti-Defamation League ha realizzato delle linee guida per rispondere all'odio online. Il gruppo di lavoro è composto da esponenti della società civile, esperti legali, accademici e rappresentanti dei colossi di internet. La sua eterogenea composizione si riflette sulle linee guida che infatti contengono raccomandazioni rivolte a soggetti diversi, dai social network ai comuni cittadini. Le linee guida ribadiscono l'importanza che i social network strutturino strumenti per promuovere la contro-narrativa e che tali strumenti siano coordinati, per favorire la realizzazione di contro-narrativa che possa essere utilizzata agevolmente su diverse piattaforme social. Sempre con riferimento ai social network, le linee guida segnalano l'importanza dell'elaborazione di codici di condotta e disciplina in grado di informare gli utenti sullo stile e i contenuti di contro-narrativa permessi per evitare forme di contro-narrativa non costruttiva. Particolare rilevanza è posta all'utilizzo di toni satirici e alle iniziative di contro-narrativa in grado di raggiungere ampie platee e che favoriscono il confronto e ragionamento critico.

Consiglio d'Europa: Manuale per una Contro-narrativa Efficace per le Nuove Generazioni

Il progetto *No Hate Speech Movement* del Consiglio d'Europa ha previsto, tra i suoi output, la realizzazione di un manuale di supporto alla realizzazione di contro-narrativa online rivolta specificatamente alle giovani generazioni e agli educatori. A differenza delle altre esperienze, questo manuale adotta un approccio volto a favorire una narrazione di contrasto all'*hate speech* online che comprende anche la produzione di contenuti che promuovono la cultura della tolleranza, dell'uguaglianza e del rispetto delle differenze, cioè il perseguimento di contro-narrativa e di narrativa alternativa come elementi combinati per contrastare l'*hate speech* online.

All'interno di questa prospettiva, il Manuale pone come centrale, per ogni campagna di contro-narrativa, il tema della cultura del rispetto dei diritti umani. In termini operativi, ciò si traduce nel contemplare contenuti positivi di affermazione dei diritti umani nella definizione di ogni campagna comunicativa e nel relazionarsi attivamente con i gruppi marginalizzati e bersaglio di *hate speech* online per una maggiore comprensione dei fenomeni in atto e delle loro ripercussioni. In questo modo, il Manuale intende inoltre sottolineare come l'attività di contro-narrativa possa anche avere l'obiettivo di medio-lungo termine di accrescere complessivamente la cultura della tolleranza online.

Considerato che il Manuale si rivolge alle giovani generazioni, particolare attenzione viene posta sull'importanza della figura dei mentori, per aiutare gli autori di contro-narrativa a gestire le reazioni emotive che possono derivare dal confronto con gli autori degli *hate speech* online. Per rafforzare il loro ruolo di autori di contro-narrativa, è ritenuto altresì importante costruire, mantenere e rafforzare la dimensione di gruppo così da attivare dinamiche costanti di confronto, supporto e condivisione delle responsabilità.

Inoltre, il Manuale sottolinea gli effetti negativi della contro-narrativa non strutturata in termini di perdita di senso di civiltà e di vanificazione degli effetti positivi che la contro-narrativa può esplicare sia a favore dell'autore dell'*hate speech* online che della più ampia platea di utenti internet.

Nel Manuale è presentato il modello operativo **ADIE** (Assess, Design, Implement, Evaluate), analogo al modello GAMMMA di RAN, e similmente rispondente agli elementi già indicati che la letteratura ha identificato come fondamentali per l'efficacia di una campagna di contro-narrativa.

Infine, il Manuale contiene una ricca serie di informazioni su progetti esistenti di contro-narrativa e come sia possibile prendervi parte.

DEMOS: Raccomandazioni operative per un'efficace contro-narrativa su Facebook

Attraverso un'analisi approfondita condotta su Facebook e con la collaborazione dello stesso social network, il **Centro di ricerca DEMOS** attivo nel settore dei diritti umani ha stilato una serie di raccomandazioni per migliorare l'efficacia della contro-narrativa che si intende realizzare utilizzando questo specifico social network. Le Raccomandazioni evidenziano in primo luogo, come i volumi di contro-narrativa siano ancora minori rispetto ai volumi degli *hate speech* che si intende contrastare. Inoltre, si individuano e analizzano in dettaglio le due caratteristiche che dovrebbe avere la contro-narrativa realizzata su Facebook per essere più efficace in termini di popolarità. In questo senso, la contro-narrativa più efficace è indicata in quella che fa ricorso a contenuti multimediali o a contenuti basati sulla satira (Bartlett & Krasodomski - Jones, 2015); più in dettaglio, i contenuti di contro-narrativa con impostazione satirica generalmente risultano essere i più apprezzati in termini di numero di "like" ottenuti. A queste due impostazioni, satirica e con elementi multimediali, le raccomandazioni fanno seguire per popolarità i commenti basati sulla presentazione di fatti (Bartlett & Krasodomski-Jones, 2015). Infine, a prescindere dall'impostazione utilizzata, la contro-narrativa che si focalizza su ambiti specifici di discriminazione (omofobia, razzismo, islamofobia, misoginia, antisemitismo, etc.) è considerata dalle Raccomandazioni come quella in grado di generare più interazioni della contro-narrativa che contrasta l'odio online in generale (Bartlett & Krasodomski-Jones, 2015).

3 Il Centro di Eccellenza RAN ha prodotto anche i seguenti manuali operativi:
- elementi chiave per pianificare e gestire una campagna di contro-narrativa efficace (RAN Paper: "RAN guidelines for effective alternative and counter-narrative campaigns.");
- il ruolo dei soggetti informali, dei giovani e delle comunità locali nelle campagne di contro-narrativa (RAN Papers: (I) "The role of Informal Actors in delivering effective counter- and alternative narratives."; (II) "Involving young people in counter and alternative narratives - why involve peers?"; (III) "Developing counter- and alternative narratives together with local communities");
- suggerimenti specifici per ogni singolo social network su cui si voglia diffondere una campagna di contro-narrativa (RAN Paper: "RAN C&N meeting on dissemination strategies and building online multi-platform networks");
- strumenti per misurare l'efficacia di una campagna di contro-narrativa ("How to measure the impact of your online counter or alternative narrative campaign").

Le Raccomandazioni, inoltre, evidenziando che una significativa parte di contro-narrativa prodotta su Facebook si caratterizza per essere non costruttiva e caratterizzata da toni aggressivi o addirittura minatori, ne sottolineano la scarsa efficacia nel modificare l'opinione dell'autore dell'*hate speech* e le sue possibili conseguenze in termini di aumento del livello di aggressività e polarizzazione nell'arena virtuale (Benesch et al 2016a; Bartlett & Krasodomski-Jones, 2015).

CounterNarrative (Against Violent Extremism Network (AVE)[4]

Per il rafforzamento delle campagne di contro-narrativa e consentirne una più ampia diffusione, the Against Violent Extremism Network ha creato un **Toolkit online**, a cui si accede attraverso un sito internet (previa registrazione), che consente di creare, in modo assistito, una campagna di contro-narrativa. Il Toolkit è organizzato per sezioni: "come pianificare una campagna", "come creare il contenuto della campagna", "come promuovere la campagna", "come verificare l'efficacia della campagna". Ogni sezione contiene una serie di esempi a supporto delle indicazioni operative.

L'articolazione del Toolkit ricalca gli elementi considerati anche dalla letteratura essenziali per la realizzazione di una campagna di contro-narrativa. Accompagna cioè il realizzatore della campagna nella fase iniziale di **identificazione del pubblico di riferimento** che si intende sensibilizzare, quindi nell'identificazione della **tipologia di autore**, di **mezzo** e di contenuto (**messaggio**) più appropriati per riuscire a coinvolgere il pubblico di riferimento e stimolare il compimento di determinate **azioni**. La sezione sul contenuto della campagna (messaggio) contiene specifici consigli su come individuare gli strumenti migliori per realizzare video e grafiche e indicazioni operative sulle diverse impostazioni e toni del messaggio da adottare (ironico, empatico, immedesimazione, etc.)

Infine, la sezione su come promuovere la campagna mette in evidenza l'importanza di identificare fin dalla sua pianificazione, le metriche per misurarne l'efficacia in termini di pervasività (quanta parte del pubblico di riferimento è raggiunta), popolarità (quante condivisioni e/o "mi piace" si ottengono da parte del pubblico di riferimento), intensità (quanti commenti e interazioni ottenuti) e impatto. Completate tutte le sezioni del Toolkit, il sistema prevede la possibilità di scaricare il piano strategico della campagna.

3.5 La misurazione dell'efficacia della contro-narrativa

Considerato il significativo incremento delle iniziative di contro-narrativa e la centralità del loro ruolo nel contrasto dell'*hate speech* online, sono stati avviati diversi studi per misurarne l'efficacia, così da finalizzare al meglio le future campagne di contro-narrativa. Si tratta di studi fondamentali a maggior ragione perché, come è stato ampiamente dettagliato nei paragrafi precedenti, la contro-narrativa può realizzarsi in diversi modi rendendo necessario verificare quali modalità di realizzazione siano le più efficaci.

Il punto di partenza per misurare l'efficacia della contro-narrativa è tradurre in numeri gli obiettivi della campagna per renderli misurabili sulla base di metriche (indicatori) indicate in letteratura, (Tuck e Silverman, 2016). Ad esempio, se l'obiettivo è educativo, occorre specificare quante persone si vuole coinvolgere nelle attività di formazione per poter monitorare la campagna e verificare il raggiungimento dell'obiettivo.

3.5.1 Le metriche per misurare l'efficacia della contro-narrativa

La prima fase di analisi dell'efficacia della contro-narrativa si è basata su casi studio, prevalentemente riguardanti Twitter, Youtube e Facebook (RAN, 2017b; Ernst et al., 2017; Benesch et al 2016a, Bartlett & Krasodomski-Jones, 2015). Pur agendo su campioni limitati di contro-narrativa, questi studi hanno definito l'architettura del sistema di metriche che viene utilizzato dalle attuali analisi che, invece, sfruttano algoritmi di intelligenza artificiale, riuscendo ad agire su volumi ingenti di contro-narrativa. Le metriche individuate sono tese a classificare la contro-narrativa (*content metrics*) e a misurarne gli effetti in termini quantitativi (*quantitative metrics*) e di impatto (*impact metrics*).

Le **metriche per la classificazione** (*content metrics*) vengono utilizzate nella fase preliminare di valutazione dell'efficacia della contro-narrativa, perché consentono di classificare e ordinare il patrimonio di contro-narrativa, che è particolarmente varia in quanto costituita da differenti elementi: tipologie di autore, impostazione del messaggio, mezzo, pubblico di riferimento, dimensioni del confronto.

Le metriche per la classificazione ricalcano questi elementi come mostrato dalla tavola 3.1 che segue.

Tavola 3.1 Esempio di metriche di classificazione (content matrix) di contro-narrativa

Tipologia autore			Impostazione messaggio								
Famoso	Cittadino	Attivista	Dati e Fatti	Segnalazione contraddizioni	Segnalazione rischi	Segnalazione tossicità sociale	Tono satirico	Immedesimazione	Tono empatico		
Impostazione messaggio: Elementi multimediali			**Mezzo**			**Pubblico di riferimento**			**Dimensioni del confronto**		
Video	Immagine	GIF	Pubblicità online	Pagina internet	Social network	Genere	Status	Etnia	Uno a uno	Uno a molti	Molti a molti	Molti a uno

L'applicazione delle metriche suindicate consente di classificare ciascuna contro-narrativa sulla base degli elementi che la compongono così come mostrato dall'esempio nella tavola che segue.

Tavola 3.2 Esempio di classificazione di contro-narrativa

Contro-narrativa	Autore famoso	Autore attivista	Tono satirico	Basato sui fatti	...	elementi multimediali	Modalità e uno-a-uno	Modalità uno-a-molti
Contro-narrativa 1	✓	✗	✓	✗		✓	✗	✓
Contro-narrativa 2	✗	✓	✗	✗		✗	✓	✗

Una volta classificata la campagna contro-narrativa, le si applicano le metriche tese a valutare il raggiungimento dei risultati/effetti prefissati (**metriche quantitative**).

Tali metriche misurano la capacità di ogni singola campagna di contro-narrativa di essere:

I) **pervasiva**, (raggiungimento di un vasto pubblico di riferimento in aggiunta ai soggetti direttamente coinvolti nel dialogo di contro-narrativa);

II) **popolare**, (ottenimento di molti apprezzamenti);

III) **intensa**, (grado di stimolazione del pubblico di riferimento a partecipare alla contro-narrativa (Bartlett & Krasodomski-Jones, 2015).

Le **metriche di pervasività** misurano solo quanti utenti sono raggiunti senza fornire alcuna indicazione sul loro possibile interesse/ coinvolgimento. Per misurare questa dimensione, si deve ricorrere a **metriche di popolarità**, come il numero di "mi piace", condivisioni dei post di contro-narrativa o commenti alla stessa (Bartlett & Krasodomski-Jones, 2015). Le metriche di popolarità, misurando il livello di interesse espresso dalle persone che hanno visto la campagna di contro-narrativa, sono utili anche per analizzare l'efficacia dell'impostazione e del tipo di autore scelto.

4 Il Network riunisce soggetti attivi nel contrasto all'estremismo e alla radicalizzazione. È promosso da Google, l'Institute for Strategic Dialogue e la Gen Next Foundation. Per maggiori informazioni, si v. *http://www.counternarratives.org*.

La tavola 3.3 che segue fornisce un esempio di come una contro-narrativa possa essere valutata in termini quantitativi.

Tavola 3.3 Esempio di metriche quantitative applicate a campagne di contro-narrativa

Contro-narrativa	Numero condivisioni	Numero "like"	Numero risposte/ commenti	...	Numero di nuovi contenuti stimolati
Contro-narrativa 1	100	1000	40		300
Contro-narrativa 2	60	2000	12		20
Contro-narrativa 3	2	24	9		0

Entrando nel dettaglio delle **metriche quantitative** che misurano la **pervasività**, la metrica base è data dalle cosiddette "***impressions***", termine, con cui si identifica ogni occasione in cui un contenuto, spesso un post o un annuncio, appare sullo schermo di un utente. Le *impressions* totali rappresentano la misura del pubblico che è stato raggiunto. Tuttavia, è importante sottolineare che un'*impression* può essere conteggiata ai fini della valutazione senza che l'utente online abbia preso visione della campagna contro-narrativa. Per questo motivo, l'analisi di efficacia della contro-narrativa non può concentrarsi eccessivamente sulle *impressions* (Tuck e Silverman, 2016).

Tuttavia, combinate con altre metriche, le *impressions* possono anche fornire una misurazione preliminare per stabilire se la contro-narrativa attrae o meno il pubblico di riferimento. Un numero elevato di *impressions* unito a un numero relativamente basso di interazioni (click) sul contenuto potrebbe indicare che il contenuto non è attraente, mentre un numero basso di *impressions* unitamente a un numero relativamente alto di click potrebbe dimostrare l'interesse per quella contro-narrativa e conseguentemente la necessità di raggiungere una platea più ampia di persone (Tuck e Silverman, 2016).

Collegate alle *impressions* sono le **metriche di "*reach*"**, cioè il numero totale di persone che ricevono un'*impression* della contro-narrativa sul loro schermo. Considerato che un'*impression* può apparire più volte su uno stesso schermo, il suo numero di regola sarà maggiore del numero delle *reach* (Bartlett & Krasodomski-Jones, 2015).

La frequenza delle *impressions* è il numero di volte che un contenuto è apparso sullo schermo in un determinato periodo di tempo. Una frequenza d'*impressions* troppo alta può saturare eccessivamente i singoli utenti, irritare il pubblico di riferimento o far sì che le persone si sentano bombardate dalla campagna (Tuck e Silverman, 2016).

Le **metriche di visualizzazione** si riferiscono ai contenuti video e misurano il numero di volte che un video viene guardato o riprodotto (Tuck e Silverman, 2016).

Facebook, Twitter e Youtube forniscono metriche di *impressions, reach* e visualizzazione leggermente diverse in termini di modalità con cui le misurano (Tuck e Silverman, 2016).

Per quanto concerne le **metriche di popolarità** che misurano l'apprezzamento del pubblico di riferimento nei confronti della contro-narrativa, si possono distinguere due tipi di metriche:

- la prima volta a misurare quanta parte del pubblico di riferimento sia stata stimolata e coinvolta (visualizzazioni dei contenuti video, visualizzazioni dirette Facebook/Instagram);
- La seconda volta a misurare quante **interazioni** (numero di like, numero di condivisioni e di risposte) si sono generate promuovendo dibattito e attenzione da parte del pubblico di riferimento.

Tra le metriche di classificazione (*content metrics*) non sono presenti metriche relative alla misurazione delle azioni, poiché queste vengono ricomprese nelle **metriche di impatto**, in grado di misurare quanto la contro-narrativa riesca appunto ad incidere sui comportamenti nel mondo reale (offline) (Tuck e Silverman, 2016). A questo proposito, gli effetti da misurare riguardano:

- i cambiamenti prodotti sulle opinioni e sui comportamenti degli autori di *hate speech* online nei confronti dei quali si è attivata la contro-narrativa;
- l'impatto sulle opinioni del pubblico di riferimento che ha assistito allo scambio tra l'autore dell'*hate speech* online e l'autore della contro-narrativa (Bartlett & Krasodomski-Jones, 2015).

L'attuale proposta di metriche di impatto si articola in cinque tipologie sulla base dei possibili impatti prodotti dalla contro-narrativa (Banesch et al., 2016a):

- discussione costruttiva, senza nessun cambiamento nelle rispettive convinzioni;
- riconoscimento dello sbaglio e rimozione dell'*hate speech* online e relative scuse;
- rimozione del contenuto dell'*hate speech* senza riconoscimento anche dello sbaglio;
- incremento dei contenuti di *hate speech* online, generati anche da altri autori a supporto dell'autore contro il quale si è attivata la contro-narrativa;
- incremento dei contenuti di contro-narrativa, generati anche da altri autori a supporto dell'autore che ha attivato la contro-narrativa.

3.5.2 L'applicazione delle metriche per l'analisi dell'efficacia

Sulla scia di quanto sta accadendo nell'analisi dell'*hate speech* online, anche nell'analisi della contro-narrativa la ricerca si sta dirigendo verso la realizzazione di **sistemi automatizzati** di *machine learning* in grado di individuare, estrarre e analizzare grossi volumi di di contro-narrativa prodotta sui social network. I dati estratti vengono classificati utilizzando le metriche di classificazione (*content metrics*) descritte in precedenza. Successivamente, attraverso ulteriori algoritmi computazionali, viene misurata l'efficacia dei diversi tipi di contro-narrativa attraverso le metriche quantitative e quelle di impatto. Ad oggi, diversi sistemi di analisi attraverso algoritmi sono in fase di validazione.

Più in dettaglio, vengono identificati i contenuti di *hate speech* online sul social network in esame, attraverso algoritmi dedicati o in misura minore manualmente. Una volta individuati, vengono anche estratte, attraverso algoritmi, le risposte e/o i commenti agli *hate speech*. I dati estratti vengono quindi esaminati, ad oggi, manualmente, per dividere i contenuti di contro-narrativa da quelli di odio. Dal corpus di dati/contenuti di contro-narrativa vengono identificate le **parole chiave** (Bag of Words) e **le strutture grammaticali** utilizzati con maggior frequenza. Entrambi questi elementi distintivi vengono assegnati come **criteri di identificazione all'algoritmo di intelligenza artificiale** avviando così il suo processo di apprendimento con supervisione umana per continuare il suo processo di perfezionamento. Il progressivo miglioramento dell'algoritmo riduce il ricorso alla componente umana.

L'obiettivo è realizzare algoritmi in grado di individuare la contro-narrativa e successivamente analizzarla classificandola con riferimento alle diverse impostazioni possibili di messaggio (presentazione di fatti, tono satirico, tono empatico, segnalazione della tossicità sociale, segnalazione dei rischi, segnalazione delle contraddizioni, utilizzo di contenuti multimediali, immedesimazione). Alcune recenti esperienze stanno attualmente testando algoritmi progettati per arrivare a questo dettaglio di classificazione (Studio Binny et al. 2018 e*Thou Shalt Not Hate* Studio Binny et al., 2019).

Uno studio (Binny et al (2018) ha, in particolare, creato un database di conversazioni **Twitter** avvenute tra autori di tweet di odio e autori di contro-narrativa che hanno risposto a questi tweet utilizzando algoritmi di intelligenza artificiale con supervisione e assistenza umana. I dati sono stati ottenuti identificando ed estraendo da Twitter gli *hate speech* relativi ai seguenti ambiti di discriminazione: orientamento sessuale, nazionalità, orientamento religioso, etnia, genere e caratteristiche fisiche. Poiché oggetto dell'analisi è la contro-narrativa, dai dati estratti sono stati eliminati i tweet di odio che non hanno avuto almeno due risposte. L'insieme dei tweet rimanenti è stato ulteriormente verificato attraverso *content analysis* per escludere i falsi positivi. Ottenuto il corpus di tweet di odio che sono stati interessati da produzione di contro-narrativa, l'analisi si è focalizzata sulle risposte con lo scopo di distinguere le risposte di contro-narrativa da eventuali altre risposte. Le risposte di contro-narrativa sono state analizzate usando le metriche quantitative e di classificazione già citate. Inoltre, disponendo del dato temporale delle conversazioni, l'analisi ha consentito anche di misurare alcune metriche d'impatto: in particolare l'eventuale rimozione del tweet di odio originale. Da questa analisi, condotta su un ampio campione di contenuti di contro-narrativa, sono emersi dati a supporto dell'effetto positivo in termini di popolarità ottenuto dalla contro-narrativa che impiega contenuti multimediali (numero maggiore di "mi piace") e in termini di effettiva capacità della contro-narrativa di modificare le azioni dell'autore dell'*hate speech* online (rimozione del tweet di odio o rimozione del tweet di odio seguito da tweet di scuse).

Con riferimento a **Youtube**, l'iniziativa *Thou Shalt Not Hate* (Studio Binny et al (2019)) ha estratto manualmente e successivamente analizzato quasi 14.000 commenti relativi a video caricati sulla piattaforma caratterizzati da contenuti di odio nei confronti di ebrei, persone di colore e comunità LGBTQ. I commenti sono confluiti in un database, quindi analizzati, sempre manualmente, per suddividere i commenti di contro-narrativa dai contenuti di odio. Questi passaggi sono propedeutici allo sviluppo di algoritmi di machine learning in grado di agire in modo automatico nell'estrazione di contro-narrativa da Youtube. Già da questa prima analisi manuale emerge che circa la metà dei commenti sono riconducibili a contenuti di contro-narrativa, risultato che evidenzia come la contro-narrativa di contrasto agli *hate speech* sia ormai un fenomeno diffuso e non trascurabile. I commenti identificati come contro-narrativa, sono stati oggetto di un'ulteriore suddivisione manuale sulla base delle metriche di classificazione. Questa seconda analisi consente di verificare la frequenza delle tipologie di messaggio di contro-narrativa con riferimento a ogni soggetto bersaglio. Dalla mappatura emerge che la tipologia di messaggio di contro-narrativa più frequente con riferimento a tutti i soggetti bersaglio di *hate speech* è purtroppo di tipo non costruttivo e si caratterizza per toni ostili e aggressivi.

Per quanto concerne la parte di contro-narrativa costruttiva, si evidenzia come le impostazioni più frequenti di messaggio cambino a seconda del soggetto bersaglio dei contenuti di odio. In particolare, i messaggi satirici sono la tipologia più frequente quando la contro-narrativa riguarda l'ambito LGBTQ, mentre la segnalazione della tossicità sociale e dei rischi individuali corsi dall'autore dell'*hate speech* online costituiscono le impostazioni più frequenti di messaggio quando l'ambito interessato è quello razziale. L'analisi ha misurato anche la percentuale di contro-narrativa che combina più impostazioni di messaggio, riscontrando che la maggioranza della contro-narrativa si concentra su una sola impostazione. Infine, sfruttando la possibilità di misurare, per ogni commento, il numero di "*like*" e di risposte, lo studio sviluppa algoritmi per misurare le metriche quantitative con cui valutare l'efficacia della contro-narrativa con riferimento a popolarità e intensità delle interazioni. In dettaglio, partendo dall'omofobia, la contro-narrativa di tipo satirico riscontra il grado di popolarità maggiore (maggior frequenza di "mi piace"), mentre la contro-narrativa che presenta fatti a suo supporto e quella che evidenzia l'ipocrisia o le contraddizioni degli autori degli *hate speech* si caratterizzano entrambe per una maggiore intensità di interazioni (maggiori numero di risposte). Con riferimento all'antisemitismo e all'ambito razziale, la contro-narrativa basata sull'immedesimazione è quella con il maggior grado intensità.

Questi risultati confermano l'importanza strategica di identificare con precisione il soggetto bersaglio a favore del quale si intende realizzare l'attività di contro-narrativa prima di intraprendere una campagna, visto che l'adozione di specifiche impostazioni può incidere sull'efficacia della contro-narrativa in maniera differente a seconda del soggetto bersaglio che intende supportare.

La classificazione della contro-narrativa realizzata si è concentrata sulle diverse impostazioni di contenuto. Esistono, tuttavia, anche altri elementi rilevanti a cui sarebbe interessante estendere le dimensioni della classificazione e dell'analisi fatta dagli algoritmi, come ad esempio la tipologia di autori. Per ora gli algoritmi sviluppati nel corso dello studio sono applicabili solo a Youtube e non anche agli altri social network, come Twitter, Instragram e Facebook.

Infine, per quel che riguarda **Facebook**, è stato realizzato uno studio (Schieb e Preuss (2016) che si focalizza su un'analisi di tipo quantitativo volta a determinare il volume di contenuti di contro-narrativa e l'impostazione necessaria a modificare un contenuto di odio presente sul social network e il convincimento del suo autore. Sviluppando algoritmi di influenza applicati ad una situazione ipotetica di dialogo all'interno di una pagina Facebook tra autori di *hate speech* online e di contro-narrativa, lo studio ha riscontrato che la contro-narrativa è in grado di influenzare l'opinione di coloro che non si trovano su posizioni di odio radicali e che questa influenza è tanto più pervasiva quanto più ampia è la platea che partecipa/osserva la discussione. L'esperimento mostra, inoltre, come la contro-narrativa possa non solo impedire che la platea di utenti della pagina Facebook assuma posizioni di odio più estreme, ma addirittura anche influenzare la stessa platea verso posizioni opposte, anche se solo di poco. In questo caso, la metrica di classificazione considerata dall'analisi si focalizza solo sul numero dei partecipanti alla discussione (dimensione del confronto). Le conclusioni dell'esperimento evidenziano, in particolare, come la modalità "molti-a-molti" adottata da una contro-narrativa, che sia anche coordinata e strutturata, possa essere decisamente efficace in termini di impatto, non rischiando di sfociare in una contro-narrativa non costruttiva, e arrivando fino al punto di far cambiare opinione all'autore dell'*hate speech*.

3.6

Esperienze e metodologie di contro-narrativa per il contrasto dell'*hate speech* online. Esperienze italiane ed internazionali a confronto

In questi ultimi anni diverse sono state le esperienze e i progetti che si sono concentrati sulle metodologie di contro-narrativa per il contrasto degli *hate speech* che hanno avuto effetti positivi e/o innovativi. Di seguito se ne presentano alcune che sono state individuate mediante una ricognizione desk che ha portato all'individuazione di 16 esperienze/metodologie maturate in parte in ambito accademico all'interno di progetti di ricerca europei (soprattutto REC - *Rights, equality and citizenship programme* (2014-2020)) o realizzate a livello internazionale e grazie a finanziamenti privati. Più nello specifico, 6 esperienze nel panorama europeo hanno visto l'Italia quale coordinatore; 2 l'Olanda, 2 la Germania, 1 la Svezia, 1 la Francia, 1 l'Ungheria, mentre a livello internazionale è stata individuata 1 sola esperienza realizzata negli Stati Uniti. A livello istituzionale 1 esperienza è stata realizzata dal Consiglio d'Europa e 1 dall'European Schoolnet, un'iniziativa di collaborazione internazionale che vede affiancati 23 Ministeri della Pubblica Istruzione europei con coordinamento in Belgio.

La metà delle esperienze (10) si caratterizzano per essere multi-paese, cioè la creazione della metodologia di contro-narrativa ha visto il coinvolgimento di più paesi partner in cui è stata anche applicata. La dimensione multi-paese consente, in particolare, di individuare trend sovra-nazionali per quanto concerne caratteristiche e modalità di realizzazione della contro-narrativa.

La maggior parte delle esperienze mappate (6 su 16) ha fatto riferimento per la definizione delle campagne di contro-narrativa a tutti i social networks, 2 sia a Facebook che a Twitter, 1 esperienza a Facebook, Twitter e Youtube, mentre 5 esperienze hanno fatto riferimento a un solo social network (3 a Facebook, 1 a Twitter e 1 a instgram). In due casi (Pet-Scan e Project Grey), la metodologia non ha ancora individuato i social networks di riferimento.

Sulla base delle principali informazioni raccolte, è stato possibile realizzare due tavole sinottiche che mettono a confronto e sintetizzano le principali caratteristiche delle esperienze/metodologie di contro-narrativa mappate.

La prima tavola (Tav 3.4), in particolare, presenta tutte le esperienze/metodologie individuate con il dettaglio dei loro principali elementi identificativi: soggetto coordinatore e altri partner coinvolti, piattaforme social con riferimento alle quali è stata realizzata l'attività di contro-narrativa, ambito geografico di realizzazione che porta a identificare esperienze mono o multipaese, in questo ultimo caso con dettaglio dei paesi coinvolti.

Ma soprattutto la tavola intende sintetizzare la descrizione complessiva delle esperienze/metodologie di contro-narrativa, mettendole a confronto per quanto concerne ambito discriminatorio/soggetto bersaglio su cui si innesta la contro-narrativa, obiettivi perseguiti, struttura/articolazione dell'esperienza (mono-intervento se focalizzate solo sulla campagna di contro-narrativa) ovvero multi-intervento se sono previsti anche altri interventi propedeutici o a corrollario della campagna (Identificazione e analisi degli *hate speech* e valutazione della contro-narrativa; Sostegno e rafforzamento della campagna di contro-narrativa) con dettaglio delle attività e degli strumenti adottati).

Tavola 3.4 Tavola sinottica esperienze/ metodologie di contro-narrativa: elementi identificativi e descrizione della metodologia complessiva di contro-narrativa

	Elementi identificativi delle esperienze/metodologie di contro-narrativa			Descrizione Esperienza/Metodologia		complessiva di contro-narrativa	
N	Nome iniziativa Soggetto attuatore e periodo di implementazione	Ambito geografico di applicazione	Social network di riferimento	Oggetto contro-narrativa (ambito di discriminazione/ soggetto bersaglio)	Obiettivi	Tipologie di interventi	Attività e strumenti
1	**#jagärhär (#iamhere #iosonoqui ...)** Jagärhär associazione (SE) (coordinatore)	Multipaese (Svezia e replicata in14 paesi tra cui Italia)	Facebook e Twitter	Multitarget (migranti/Stranieri Rom/Antisemitismo, genere,religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Mobilitazione utenti Aumento consapevolezza utenti online	Intervento 1- Creazione e diffusione della campagna di contro-narrativa Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa	I1: Produzione messaggi di contro-narrativa attraverso pagina e gruppo Facebook e supporto di moderatori, anche nella forma di Positive bombing/raid coordinati; I2: attività informative e di sensibilizzazione (campagna social e sito internet per dare visibilità all'esperienza complessiva; eventi per raccolta fondi a sostegno e Network internazionale di coordinamento dell'esperienza nei diversi Paesi).
2	**Donate the Hate** ZDK Gesellschaft Demokratische Kultur (DE) (associazione)	Germania	Facebook	Multitarget (Migranti/Stranieri Rom/Antisemitismo, genere,religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Mobilitazione utenti Espressione di solidarietà Aumento consapevolezza utenti online	Intervento 1- identificazione e analisi dell'*hate speech* online Intervento 2 - Creazione e diffusione della campagna di contro-narrativa Intervento 3 - Sostegno/rafforzamento campagna di contro-narrativa	I1: Sviluppo algoritmi per l'estrazione e analisi *hate speech* da Facebook visibili mediante App accessibile ad utenti registrati; I2: produzione semi-automatica di messaggi pre-impostati associati a micro donazioni messi a disposizione dell'utenza tramite un App; I3: attività informative e di sensibilizzazione (campagna pubblicitaria per dare visibilità all'esperienza complessiva e di sponsorizzazione monetaria; sito internet con sezione per raccolta fondi, collaborazioni con media per pubblicizzazione.
3	**We counter hate** Possible (USA) (Agenzia privata di comunicazione e design)	USA	Twitter	Migranti/Stranieri Rom/Antisemitismo **Altro**: genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Mobilitazione utenti Espressione di solidarietà Aumento consapevolezza utenti online Valutazione efficacia contro-narrativa	Intervento 1- identificazione e analisi *hate speech* online e valutazione della contro-narrativa Intervento 2 - Creazione e diffusione della campagna di contro-narrativa Intervento 3 - Sostegno/rafforzamento campagna di contro-narrativa	I1: Sviluppo algoritmi per l'estrazione e analisi *hate speech* da Twitter e per la valutazione efficacia contro-narrativa creata dall'autore I2: produzione di messaggi di contro-narrativa totalmente automatici associati a micro-donazioni; I3: attività informative e di sensibilizzazione (sito di presentazione dell'iniziativa, campagna pubblicitaria per aumentare il numero di utenti registrati al sito e le donazioni; vendita di merchandising (felpe, t-shirt) per la raccolta fondi.
4	**Task Force di Contrasto all'Odio Online** Amnesty International Italia (ONG)	Italia	Tutti i social networks	Multi-target (migranti/ Stranieri,Rom/Antisemitismo/ razzismo;genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Educativo Aumento cultura tolleranza (a lungo termine)	Intervento 1 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 2 - Creazione e diffusione della campagna di contro-narrativa	I1: Percorsi formativi a bando diretti a volontari per formazione Task force; attività informative e di sensibilizzazione (sito di presentazione dell'iniziativa e seminari); I2: produzione contenuti di contro-narrativa ad opera della Task-force diretti agli autori di *hate speech*.
5	**PET-SCAN** Ligue Internationale contre le racisme et l'antisemitisme (FR) (associazione)	Tutti i paesi dell'Unione Europea	Non specificati	Multitarget (migranti/Stranieri Rom/Antisemitismo, genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Educativo Mobilitazione utenti online Aumento cultura tolleranza (a lungo termine)	Intervento 1- identificazione e analisi *hate speech* online Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 3 - Creazione e diffusione della campagna di contro-narrativa	I1: Sviluppo algoritmi per estrazione e analisi *hate speech;* I2: produzione materiali didattici per attivisti e moderatori, creazione di una piattaforma per e-learning; I3: creazione network attivisti e creazione di campagne-pilota coordinate di contro-narrativa.
6	**React** ARCI (IT) (associazione)	Italia, Francia Germania Spagna Regno Unito	Tutti i social network	Islamofobia	Contrasto *hate speech* Educativo Mobilitazione utenti Espressione di solidarietà Aumento cultura tolleranza (a lungo termine) Valutazione efficacia contro-narrativa	Intervento 1- identificazione e analisi *hate speech* online e identificazione contro-narrativa per valutazione Intervento 2 - Creazione e diffusione della campagna di contro-narrativa Intervento 3 - Sostegno/rafforzamento campagna di contro-narrativa	I1: Sviluppo algoritmi per l'estrazione e analisi *hate speech* e per identificazione e analisi contro-narrativa; I2: percorsi formativi frontali e a distanza sui temi del contrasto all' *hate speech* online mirati per educatori. Percorsi per giovani per diventare autori di contro-narrativa; scambio di buone pratiche e attività informative e di sensibilizzazione (pagina internet ed eventi di approfondimento); I3: realizzazione di campagne pilota di contro-narrativa all'interno di laboratori rivolti ai giovani, diffusa attraverso pagina Facebook apposita e profili Instagram e Twitter.

Tavola 3.4 Tavola sinottica esperienze/ metodologie di contro-narrativa: elementi identificativi e descrizione della metodologia complessiva di contro-narrativa

	Elementi identificativi delle esperienze/metodologie di contro-narrativa			Descrizione Esperienza/Metodologia			
N	Nome iniziativa Soggetto attuatore e periodo di implementazione	Ambito geografico di applicazione	Social network di riferimento	Oggetto contro-narrativa (ambito di discriminazione/ soggetto bersaglio)	Obiettivi	Tipologie di interventi	Attività e strumenti
7	**SilenceHate** COSPE (IT) (associazione)	Italia, Belgio, Cipro, Grecia, Polonia, UK	Tutti i social network	Multi-target (migranti/Stranieri Rom/Antisemitismo, genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Mobilitazione utenti Educativo Aumento cultura tolleranza (ob. a lungo termine)	Intervento 1 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 2 - Creazione e diffusione della campagna di contro-narrativa	I1: Percorsi formativi/laboratoriali per studenti, insegnanti/educatori, giornalisti e professionisti del settore audio-visivo; I2: realizzazione campagne-pilota durante le attività laboratoriali-formative ad opera di studenti, giovani, insegnanti e professionisti del settore audiovisivo.
8	**No Hate Speech Movement** Consiglio d'Europa (istituzione)	Multipaese (45 Paesi dell'ambito Consiglio d'Europa)	Tutti i social network	Multitarget (migranti/Stranieri Rom/Antisemitismo, genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech*; Educativo; Mobilitazione utenti; Aumento cultura tolleranza (a lungo termine)	Intervento 1 - Creazione e diffusione campagne di contro-narrativa Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa	I1: Creazione campagne europee ad opera di giovani volontari appositamente formati del No Hate Movement e inserite in una piattaforma online; I2: percorsi formativi per i volontari online e per educatori e produzione di manuali operativi (per educatori e attivisti) e il modello operativo ADIE per la realizzazione di contronarrrativa; una pagina in-formativa per l'analisi degli *hate speech*; un'unità di coordinamento, un blog e un forum per favorire scambi tra le diverse esperienze nazionali; Attività informative e di sensibilizzazione (sito web istituzionale, repository di tutti i materiali prodotti; eventi: festival e mobilitazioni promosse in maniera coordinata in tutti i paesi, raccolte fondi, istituzione figura dell'ambasciatore); creazione del Follow-Up Group per verifica attività realizzate.
9	**Turulpata** Political Capital Institute (HU) (istituto di ricerca)	Ungheria	Facebook	Bersagli multipli colpiti dall'estremiso di destra	Contrasto *hate speech* Mobilitazione utenti Aumento consapevolezza utenti online	Intervento unico: creazione e diffusione della campagna di contro-narrativa	Creazione di una pagina Facebook satirica in cui si immagina un insediamento fittizio ungherese abitato e governato da persone di estrema destra
10	**On/Off Derad** Violence Prevention Network (DE) (ONG)	Germania	Facebook	Bersagli multipli colpiti dall'estremismo di destra e dal radicalismo islamico	Contrasto *hate speech* Mobilitazione utenti Aumento consapevolezza utenti online	Intervento unico: creazione e diffusione di una campagna di contro-narrativa	Creazione di due pagine Facebook per la contro-narrativa: - una pagina pensata per intercettare i giovani a rischio di radicalizzazione islamica; - una pagina pensata per i giovani a rischio di estremismo di destra.
11	**Project Grey** Dare to be Grey (NL) (associazione)	Olanda Belgio	Non definiti	Migranti/Stranieri	Contrasto *hate speech* Educativo Mobilitazione utenti Espressione di solidarietà Aumento cultura tolleranza (ob. a lungo termine)	Intervento 1- identificazione e analisi dell'*hate speech* online Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 3 - Creazione e diffusione della campagna di contro-narrativa	I1: Sviluppo algoritmi di intelligenza artificiale per estrazione e analisi quotidiana *hate speech* di modo da individuare gli argomenti più discussi tra gli utenti dei social network e il loro grado di polarizzazione/ divisività e organizzazione in un database consultabile; I2: percorsi formativi per operatori sociali e delle politche giovanili per utilizzo database e realizzazione contro-narrativa; I3: creazione campagna rivolta al pubblico moderato (grey).
12	**Data-Driven-Approach to Counter Hate Speech (DACHS)** European Journalism Centre (NL) (organizzazione)	Tutti i paesi dell'Unione Europea	Facebook, Youtube e Twitter	Altro: Giornalisti	Contrasto *hate speech* Mobilitazione utenti online Educativo Espressione di solidarietà Aumentare cultura tolleranza (a lungo termine)	Intervento 1- identificazione e analisi dell'*hate speech* online Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 3 - Creazione e diffusione della campagna di contro-narrativa	I1: Sviluppo algoritmi di intelligenza artificiale per estrazione e analisi *hate speech* prodotto contro giornalisti e testate informative da Facebook. Twitter e Youtube, costituzione database a disposizione anche di ricercatori e del sistema di algoritmi anche per sviluppatori informatici (miglioramento prodotto); I2: percorsi formativi frontali e a distanza per giornalisti e moderatori pagine web sull'uso dello strumento informatico e su come fare contro-narrativa (autori); I3: produzione messaggi di contro-narrativa da parte dei giornalisti e moderatori.

Tavola 3.4 Tavola sinottica esperienze/ metodologie di contro-narrativa: elementi identificativi e descrizione della metodologia complessiva di contro-narrativa

	Elementi identificativi delle esperienze/metodologie di contro-narrativa			Descrizione Esperienza/Metodologia			
N	Nome iniziativa Soggetto attuatore e periodo di implementazione	Ambito geografico di applicazione	Social network di riferimento	Oggetto contro-narrativa (ambito di discriminazione/ soggetto bersaglio)	Obiettivi	Tipologie di interventi	Attività e strumenti
13	**Selma** European Schoolnet (BE) (network europeo ministeriale)	Belgio Danimarca Germania Grecia UK	Tutti i social network	Multi-target (migranti/Stranieri Rom/Antisemitismo, genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech* Educativo Mobilitazione online Aumento cultura della tolleranza (ob. a lungo termine)	Intervento 1 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 2 - Creazione e diffusione della campagna di contro-narrativa	I1: Creazione di campagne di contro-narrativa ad opera di scolari tra 11-16 anni formati dai propri educatori; I2: percorsi formativi a distanza per educatori attraverso MOOC, toolkit per educatori e momenti di scambio tra ministri, istituzioni UE e imprese ITC; costituzione del network dei Selma ambassador costituito dagli educatori formati di tutti i paesi partner; attività informative: seminari.
14	**Hatemeter** Università di Trento (IT) (università)	Italia Francia Regno Unito	Facebook e Twitter	Islamofobia	Contrasto *hate speech* Mobilitazione utenti online Educativo Espressione di solidarietà Aumento tolleranza (ob. a lungo termine)	Intervento 1 - identificazione e analisi *hate speech* online Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa Intervento 3 - Creazione e diffusione della campagna di contro-narrativa	I1: Sviluppo algoritmi per l'estrazione e analisi *hate speech*; I2: percorsi formativi mirati per attivisti ONG da coinvolgere come autori di contro-narrativa; percorsi formativi per giornalisti, esperti diritti civili, dipendenti pubblici, leader comunità musulmane; scambio di buone pratiche e attività informative e di sensibilizzazione (sito informativo e eventi di approfondimento); I3: produzione semiautomatica (algoritmo) di messaggi pre-impostati di contro-narrativa messi a disposizione degli autori (attivisti ONG) testati in 3 campagne social.
15	**PRISM** ARCI (IT) (associazione)	Italia Francia Spagna Romania Gran Bretagna	Instagram	Multitarget (migranti/Stranieri Rom/Antisemitismo, genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech*; Mobilitazione utenti online; Educativo; Aumento cultura tolleranza (a lungo termine).	Intervento 1- identificazione e analisi dell'*hate speech* online Intervento 2 -sostegno/rafforzamento campagna di contro-narrativa Intervento 3 - Creazione e diffusione della campagna di contro-narrativa	I1: Interviste e mappature qualitative (prima fotografia *hate speech*, followers e hastag delle sezioni commenti di quotidiani digitali e i forum di discussione generale); I2: attività laboratoriali/formative per giovani sulla realizzazione di contro-narrativa; linee guida per insegnanti/ operatori giovanili su come sviluppare campagne-pilota di contro-narrativa; percorsi formativi frontali e a distanza per professionisti comunicazione su come fare contro-narrativa; attività informative e di sensibilizzazione (eventi presentazione iniziativa); I3: campagna di contro-narrativa su Instagram come attività finale del percorso formativo diretto ai giovani (Gara tra giovani per la produzione di messaggi di contro-narrativa).
16	**No Hate Speech Movement Italia** (Dipartimento della gioventù e del servizio civile nazionale (IT) (istituzione)	Italia	Tutti i social network	Multitarget (migranti/Stranieri Rom/Antisemitismo,genere, religione, disabilità, orientamento sessuale)	Contrasto *hate speech*; Mobilitazione utenti; Educativo; Aumento cultura tolleranza (a lungo termine)	Intervento 1 - Creazione e diffusione campagne di contro-narrativa Intervento 2 - Sostegno/rafforzamento campagna di contro-narrativa	I1: Creazione campagne nazionali: una prima campagna istituzionale ad opera di professionisti e una seconda che ha coinvolto le scuole (mediante concorso indirizzato ai ragazzi); una terza ad opera degli attivisti del movimento; I2: percorsi formativi rivolti ai giovani sul tema della contro-narrativa ("contro narrazioni e narrazioni alternative"; schede operative per la produzione di contenuti e lo sviluppo di azioni; diffusione dei manuali operativi (per educatori e attivisti) realizzati dall'esperienza a livello europeo; Attività informative e di sensibilizzazione (sito web istituzionale, repository di tutti i materiali prodotti ed eventi di sensibilizzazione in anche in occasione delle giornate di mobilitazione generale promosse a livello europeo dal No Hate Speech Movement.

Osservando la tavola 3.4, si rileva che la maggior parte delle esperienze/metodologie mappate (13 su 16), si focalizzano su più ambiti di discriminazione, sono cioè **multidimensionali o multibersaglio**, sul presupposto ormai consolidato che l'*hate speech* può colpire differenti target o riferirsi a un soggetto-bersaglio suscettibile di essere oggetto di discriminazione con riferimento a più aspetti/caratteristiche personali (ad es. riguardare una donna di una particolare etnia e disabile).

Solo 4 esperienze tra quelle mappate (*Hatemeter*, *React*, *Project Grey* e *Dachs*) si focalizzano su **un solo ambito/soggetto bersaglio specifico.** *Project Grey* su migranti/stranieri e *Dachs* sui giornalisti, soggetto questo ultimo particolarmente colpito per via del lavoro che svolge nel settore dell'informazione e *Hatemeter* e *React* sull'islamofobia. Sviluppare contro-narrativa contro l'islamofobia, significa provare a contrastare pregiudizi e discriminazione verso i musulmani. Tale ambito di discriminazione non è pertanto riconducibile in senso stretto alla pratica religiosa ma sembra piuttosto identificare un gruppo specifico di persone/minoranze straniere.

Le esperienze selezionate perseguono diversi obiettivi tra quelli individuati in letteratura:

- obiettivo di contrasto degli *hate speech* online;
- mobilitazione degli utenti online;
- obiettivo educativo e di aumento della consapevolezza degli utenti online;
- espressione di solidarietà e vicinanza alle vittime di specifici *hate speech* online e ai gruppi a cui appartengono;
- obiettivi di lungo termine quali l'aumento della cultura della tolleranza.

Naturalmente, tutti i progetti condividono l'obiettivo di contrastare l'*hate speech* online e di mobilitare gli utenti online a favore della contro-narrativa (perché la realizzino o la diffondano). Anche le esperienze/metodologie che prevedono fra i loro obiettivi di valutare l'efficacia della contro-narrativa perseguono indirettamente l'obiettivo di contrastare l'*hate speech* online, in quanto identificando le modalità più efficaci di contrasto, consentono anche di migliorare progressivamente le attività di contro-narrativa.

A dimostrazione della crescente attenzione nei confronti della misurazione dei risultati e degli effetti delle campagne di contro-narrativa, 2 esperienze tra quelle mappate stanno iniziando a diffondere le metriche relative all'efficacia delle proprie campagne. *We Counter Hate* e *On/Off Derad,* infatti, rendono disponibili le misurazioni relative al raggiungimento di obiettivi, quali: la percentuale di pubblico di riferimento raggiunta, il numero di post di *hate speech* online rimossi o la cui diffusione è stata limitata a seguito delle campagne, quanti autori di *hate speech* online hanno ridotto i toni aggressivi. *Jagärhär*, invece, è stata oggetto di uno studio universitario indipendente, sempre sulla valutazione dei risultati ottenuti. Considerando quanto si stanno perfezionando le modalità di analisi di queste dimensioni, è probabile che sempre più esperienze in futuro renderanno disponibili questo tipo di dati. *React*, inoltre, prevede di realizzare anche un sistema di algoritmi di intelligenza artificiale per estrarre e analizzare la contro-narrativa prodotta in generale sui social network.

La maggior parte delle esperienze (11) perseguono un obiettivo educativo, nel senso che svolgono attivitè formative specifiche e non solo attività volte a informare e sensibilizzare il pubblico di riferimento (sono 5, nello specifico, le esperienze che perseguono come obiettivo di aumentare la consapevolezza del pubblico online).

Tutte le esperienze che perseguono un obiettivo educativo, allo stesso tempo contribuiscono all'aumento della cultura della tolleranza, obiettivo identificato dalla letteratura come a lungo termine. È evidente, infatti, come le azioni formative, soprattutto quando consentono a nuovi soggetti di individuare autonomamente gli *hate speech* e controstarli attraverso la produzione di messaggi di contro-narrativa in via continuativa, possono produrre, più di una campagna sviluppata una tantum, importanti effetti moltiplicatori e contribuire a realizzare obiettivi di lungo termine come l'aumento della cultura della tolleranza. Questo si ralizza soprattutto quando ad essere formati come autori di contro-narrativa sono giornalisti o moderatori di pagine web di testate giornalistiche in considerazione del loro importante ruolo di veicolo dell'informazione pubblica, ma anche quando si costituiscono network o task-force di attivisti e volontari appositamente formati. Particolarmente interessante l'esperienza SELMA che si concentra sui giovanissimi, individuando nei contesti scolastici gli ambienti più adatti per formare alla contro-narrativa. L'attuazione di questa esperienza ad opera dell'European

Schoolnet, un network composto da 34 Ministeri dell'Istruzione europei, ha reso possibile l'avvio di di buone pratiche europee per l'inserimento della contro-narrativa tra i programmi curricolari dell'istruzione sin dalle scuole primarie.

Alcune esperienze (6) perseguono tra gli obiettivi anche quello di esprimere solidarietà e vicinanza alle vittime di specifici *hate speech* online e ai gruppi a cui appartengono perché tendono a difendere un gruppo bersaglio specifico, come già detto, oppure perchè prevedono la produzione di messaggi di contro-narrativa associati a micro-donazioni.

La maggior parte delle esperienze analizzate (14) prevedono un **sistema articolato di interventi e strumenti** a corollario della campagna di contro-narrativa**.** Il panorama di strumenti per ciascun tipo di intervento è ampio e articolato.

Più nello specifico, nell'ambito dell'*attività di identificazione e analisi* dell'*hate speech* online, possono essere sviluppati **algoritmi** per l'estrazione e l'analisi degli *hate speech* dai vari social networks, **App** e **database** accessibili agli utenti per disporre di informazioni sugli *hate speech*, finanche realizzate interviste e **mappature qualitative** per l'identificazione manuale degli *hate speech*. Si tratta di attività tutte realizzate per consentire successivamente la definizione di attività di contro-narrativa specifiche. Anche per la produzione di contro-narrativa si sta diffondendo l'introduzione di algoritmi di intelligenza artificiale che vanno nella direzione di una sua progressiva automatizzazione.

Questo tipo di intervento è ormai consolidato, come evidenziano diverse esperienze mappate: *DACHS, Project Grey, PET-SCAN* e *React*.

Nell'ambito dell'*attività a sostegno e rafforzamento della campagna di contro-narrativa*, vengono spesso realizzate **attività informative e di sensibilizzazione** quali: campagne social o siti per dare visibilità all'esperienza complessiva; collaborazioni con i Media per la pubblicizzazione dell'iniziativa, eventi e campagne per la raccolta di fondi a sostegno della stessa.

Vengono, inoltre realizzati **percorsi formativi**, anche laboratoriali o mediante piattaforme di e-learning, volti alla formazione degli specifici autori di contro-narrativa coinvolti nella campagna, ma anche attività formative dirette ad altri soggetti facenti parte del pubblico di riferimento per diffondere l'importanza della contro-narrativa. *Project Grey, SilenceHate*, React e SELMA prevedono attività formative per operatori sociali e educatori perché diffondano l'importanza della contro-narrativa nelle scuole e nei centri di aggregazione giovanile.

Diffusi sono anche altri strumenti che stimolano un ruolo attivo dell'utenza nel contrasto dell'*hate speech* online come: **manuali, toolkit operativi e materiali didattici** su come sviluppare campagne-pilota di contro-narrativa. La maggior parte delle esperienze li prevedono a corollario della campagna (ad esempio, *No Hate Speech Movement* del Consiglio d'Europa e SELMA); questi strumenti vanno ad aggiungersi a quelli già realizzati dalle istituzioni o realizzati o promossi dai social network fruibili gratuitamente per chiunque sia interessato alla produzione autonoma di contro-narrativa.

Alcune esperienze hanno previsto anche **scambio di buone pratiche** (React, Hatemeter e SELMA) per migliorare la produzione di contro-narrativa, e l'istituzione di figure atte a dare particolare visibilità alle campagne di contro-narrativa (ad esempio, istituzione della figura dell'ambasciatore sia in *SELMA* che nel progetto *No Hate Speech Movement* del Consiglio d'Europa).

Come già sottolineato nella descrizione delle esperienze e metodologie di contro-narrativa mappate, la parte che ne costituisce l'elemento centrale è data dalla campagna di contro-narrativa.

La tavola (3.5) che segue riporta, per ciascuna delle esperienze, la descrizione degli elementi essenziali che compongono ciascuna campagna di contro-narrativa e delle azioni stimolate nel pubblico riferimento.

Tavola 3.5 Tavola sinottica campagne di contro-narrativa: elementi essenziali e descrizione delle azioni

N	Nome iniziativa Soggetto attuatore e periodo di implementazione	Campagna di contro-narrativa: principali elementi	Descrizione azioni (stimolate dalla campagna di contro-narrativa nel pubblico di riferimento)
1	#jagärhär (#iamhere #iosonoqui ...) Jagärhär (SE) associazione (ccordinatore)	Attività: produzione messaggi di contro-narrativa (attraverso pagina e gruppo Facebook) Autori contro-narrativa: utenti registrati nella pagina Facebook Pubblico di riferimento che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: post e/o commenti su Facebook e/o Twitter; Messaggio: contenuti con tono empatico e fatti e dati a sostegno della contro-narrativa. Anche nella forma di Positive bombing/raid coordinati, con supporto moderatori. Attivo un codice di autodisciplina, che si affianca all'attività dei moderatori per evitare il ricorso a contro-narrativa non-costruttiva.	Azioni: - apprezzamento contenuti di contro-narrativa (condivisioni e "mi piace", inserimento icone nelle foto profilo degli utenti); - adesione al gruppo Facebook; - produzione autonoma di contro-narrativa analoga con quella prodotta dal gruppo; - creazione analoghe esperienze in altri contesti geografici; - donazione di risorse a sostegno iniziativa; - azioni di rimozione *hate speech* (autori).
2	Donate the Hate ZDK Gesellschaft Demokratische Kultur (DE) (associazione)	Attività: produzione semi-automatica di messaggi associati a micro donazioni messi a disposizione dell'utenza Autori contro-narrativa: utenti registrati nella pagina Facebook dell'esperienza supportati da un algoritmo informatico Pubblico di riferimento che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: post Facebook (anche utilizzo App) Messaggio: contenuto reimpostato con tono ironico	Azioni: - iscriversi al sito Donate the Hate per finanziare l'iniziativa; - dare diffusione e pubblicizzare l'iniziativa (funzione "mi piace"); - stimolare soggetti attivi nelle comunicazioni/editoria ad avvalersi del servizio offerto da Donate the Hate per contrastare i post di odio sulle proprie pagine Facebook; - produzione autonoma di contro-narrativa stimolati dalla visualizzazione della campagna prodotta da Donate the Hate; - azioni di rimozione *hate speech* da parte degli autori.
3	We counter hate Possible (USA) (Agenzia privata di comunicazione e design)	Attività: produzione di messaggi di contro-narrativa totalmente automatici associati a micro-donazioni Autori contro-narrativa: automatico/algoritmo informatico Pubblico di riferimento che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: tweet Messaggio: contenuto standard che segnala la "marcatura" di un tweet di odio	Azioni: - iscriversi al sito per donazione risorse economiche a sostegno dell'iniziativa di contro-narrativa; - diffusione materiale promozionale e "apprezzamento dell'iniziativa attraverso funzione mi piace"; - impegno nella produzione autonoma di contenuti di contro-narrativa; - azioni di rimozione *hate speech* da parte degli autori.
4	Task Force di Contrasto all'Odio Online Amnesty International Italia (ONG)	Attività: produzione di messaggi di contro-narrativa Autori contro-narrativa: volontari appositamente formati (Task.force); Pubblico di riferimento da sensibilizzare: autori di *hate speech* Mezzi per diffondere la contro-narrativa: commento di risposta sul social network in cui è stato diffuso l'*hate speech* Messaggio: prevalentemente impostato sull'utilizzo di dati e fatti a riprova di quanto si afferma.	Azioni: - rimozione dei contenuti di odio; - cambio di visione/prospettiva negli autori di *hate speech* online.
5	PET-SCAN Ligue Internationale contre le racisme et l'antisemitisme (F) associazione	Attività: campagne-pilota coordinate di contro-narrativa Autori contro-narrativa: attivisti e volontari appositamente formati (network di attivisti); Pubblico di riferimento da sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: nessuna informazione Messaggio: nessuna informazione	Azioni: - diffusione materiale promozionale e "apprezzamento dell'iniziativa attraverso funzione mi piace"; - produzione autonoma di contenuti di contro-narrativa in analogia a quella prodotta dagli attivisti; - azioni di rimozione *hate speech* da parti degli autori.
6	React ARCI (I) (associazione)	Attività: realizzazione di campagne pilota di contro-narrativa all'interno di momenti laboratoriali rivolti ai giovani, Autori della contro-narrativa: giovani appositamente formati Pubblico di riferimento: che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: tutti i social network Messaggio: non è specificata alcuna tipologia di contenuto principale	Azioni: - apprezzamento della contro-narrativa prodotta dall'esperienza (attraverso condivisioni e "mi piace"); - stimolo alla produzione autonoma, spontanea di contenuti di contro-narrativa; - rimozione *hate speech* da parte degli autori.
7	SilenceHate COSPE (I) (associazione)	Attività: campagne - pilota Autori contro-narrativa: giovani/studenti, giornalisti, creativi, appositamente formati Pubblico di riferimento che si intende sensibilizzare: giovani, le loro famiglie, la comunità educativa e in generale la collettività tutta Mezzi per diffondere la contro-narrativa: social network ma nessun mezzo prevalente Messaggio: prevalentemente di natura multimediale (video, grafiche, immagini)	Azioni: - aumentare il coinvolgimento dei giovani nella produzione di contro-narrativa; - aumentare il coinvolgimento dei giovani nella diffusione di contro-narrativa; - azioni di rimozione *hate speech* (autori).

Tavola 3.5 Tavola sinottica campagne di contro-narrativa: elementi essenziali e descrizione delle azioni

N	Nome iniziativa Soggetto attuatore e periodo di implementazione	Campagna di contro-narrativa: principali elementi	Descrizione azioni (stimolate dalla campagna di contro-narrativa nel pubblico di riferimento)
8	No Hate Speech Movement Consiglio d'Europa (istutizione)	Attività: creazione campagne europee ad opera di giovani volontari appositamente formati del No Hate Movement e inserite in una piattaforma online; Autori contro-narrativa: giovani appositamente formati per diventare volontari del movimento; o che sono stati formati da un educatore a sua volta formato nel corso dell'iniziativa; Pubblico di riferimento che si intende sensibilizzare: giovani, attivisti, educatori ma anche autori degli *hate speech* Mezzi per diffondere la contro-narrativa: tutti i social network Messaggio: nessuna specifica rispetto alle impostazioni	Azioni stimolate dalla campagna europea nei giovani: - candidarsi come volontari online del No Hate Speech Movement, contribuendo attivamente alla produzione di ulteriori campagne; - diffondere i contenuti delle campagne di contro-narrativa europee del No Hate Movement; - partecipare alle campagne nazionali promosse nell'ambito del progetto; - mobilitarsi per la produzione autonoma e spontanea di contro-narrativa; - la campagna intende anche stimolare gli autori di *hate speech* online alla rimozione dei contenuti di odio e gli educatori a partecipare ai percorsi formativi per l'insegnamento di contro-narrativa realizzati dal No Hate Speech Movement.
9	Turulpata Political Capital Institute (H) (istituto di ricerca)	Attività: realizzazione di una pagina Facebook diretta a contrastare le posizioni di estrema destra in Ungheria Autori della contro-narrativa: gestori della pagina Facebook dell'esperienza Pubblico di riferimento che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: pagina Facebook Messaggio: contenuti con tono ironico (paese immaginario)	Azioni: - aumentare la popolarità dei contenuti mettendo "mi piace"; - diventare follower della pagina Facebook; - diffondere i contenuti di contro-narrativa prodotti dalla pagina Facebook attraverso condivisioni; - diventare sostenitori economici del progetto attraverso donazioni spontanee.
10	On/Off Derad Violence Prevention Network (DE) (ONG)	Attività: creazione di due pagine Facebook per la contro-narrativa: una pagina pensata per intercettare i giovani a rischio di radicalizzazione islamica e una pagina per i giovani a rischio di estremismo di destra Autori: professionisti nel recupero di giovani a rischio radicalizzazione Pubblico di riferimento che si intende sensibilizzare: giovani radicalizzati (o a rischio) ed estremisti Mezzi per diffondere la contro-narrativa: pagina Facebook Messaggio: impostazione basata sull'immedesimazione	Azioni stimolate dalla campagna nei giovani radicalizzati o estremizzati o a rischio: - sviluppo di un dialogo "uno-a-uno" con gli autori degli *hate speech*; - uscita dai percorsi di radicalizzazione/estremismo)
11	Project Grey Dare to be Grey (NL) (associazione)	Attività: creazione campagna di contro-narrativa ad opera di operatori sociali/ politiche giovanili Autori: operatori sociali e di politiche giovanili appositamente formati Pubblico di riferimento che si intende sensibilizzare: parte moderata degli utenti online (grey) Mezzi per diffondere la contro-narrativa: nessuna specifica Messaggio: nessuna specifica	Azioni: - condividere i contenuti di contro-narrativa realizzati dai soggetti formati (autori); - impegnarsi autonomamente nella produzione di contenuti di contro-narrativa.
12	Data-Driven-Approach to Counter Hate Speech (DACHS) European Journalism Centre (NL) (organizzazione)	Attività: produzione messaggi di contro-narrativa da parte dei giornalisti e moderatori Autori: giornalisti e moderatori di pagine social di realtà editoriali appositamente formati Pubblico di riferimento che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: 3 tre social network (Facebook, Twitter e Youtube) Messaggio: nessuna specifica sul tipo di contenuto. Il messaggio probabilmente valorizzerà il fatto che gli autori sono professionisti della comunicazione	Azioni: - esprimere apprezzamento per la contro-narrativa realizzata dai giornalisti e moderatori attraverso condivisioni e "mi piace"; - seguire le pagine social di realtà editoriali ed esprimere apprezzamento attraverso condivisioni e la funzione "mi piace" per i contenuti di contro-narrativa prodotti dai giornalisti e dai moderatori nelle sezioni di commenti alle notizie; - impegnarsi nella produzione di contro-narrativa in analogia con quella prodotta dai giornalisti; - azioni di rimozione *hate speech* (autori).
13	Selma European Schoolnet (network ministeriale)	Attività: creazione di campagne di contro-narrativa ad opera di scolari tra 11-16 anni formati dai propri educatori Autori: ragazzi 11-16 anni appositamente formati Pubblico di riferimento che si intende sensibilizzare: giovani ed educatori Mezzi per diffondere la contro-narrativa: tutti i social network Messaggio: nessuna specifica	Azioni stimolate nei giovanissimi: - condividere i contenuti di contro-narrativa realizzati dai soggetti formati (autori); - impegnarsi autonomamente in campagne di contro-narrativa, stimolati dalla visualizzazione della campagna prodotta; - la campagna intende anche stimolare gli educatori ad iscriversi al percorso formativo e/o scaricare il toolkit per educare i propri studenti sull'importanza della contro-narrativa. Intende, altresì, stimolare gli autori di *hate speech* a rimuovere i contenuti di odio.
14	Hatemeter Università di Trento (IT) (università)	Attività: produzione semiautomatica (algoritmo) di messaggi pre-impostati di contro-narrativa a disposizione degli autori (attivisti ONG) per la realizzazione di 3 campagne social Autori: attivisti di ONG supportati da algoritmi informatici Pubblico di riferimento che si intende sensibilizzare: utenti online Mezzi per diffondere la contro-narrativa: Facebook e Twitter Messaggio: nessuna specifica	Azioni stimolate nel pubblico di riferimento: - apprezzamento contenuti di contro-narrativa prodotti dall'esperienza (attraverso condivisioni e "mi piace"); - produzione autonoma e spontanea di contro-narrativa; - la campagna intende anche stimolare la rimozione dei contenuti di odio da parte dei loro autori.

Tavola 3.5 Tavola sinottica campagne di contro-narrativa: elementi essenziali e descrizione delle azioni

N	Nome iniziativa Soggetto attuatore e periodo di implementazione	Campagna di contro-narrativa: principali elementi	Descrizione azioni (stimolate dalla campagna di contro-narrativa nel pubblico di riferimento)
15	PRISM ARCI (IT) (associazione)	Attività: campagna di contro-narrativa su Instagram/ gara tra giovani che hanno partecipato alla formazione per la produzione di messaggi di contro-narrativa. Autori: ragazzi che hanno partecipato ai laboratori/formativi Pubblico di riferimento che si intende sensibilizzare: giovani, famiglie e comunità educativa Mezzi per diffondere la contro-narrativa: Instagram Messaggio: ricorso a contenuti multimediali	Azioni: - impegnarsi nella produzione autonoma di contro-narrativa; - condividere e/o mettere "mi piace" ai contenuti di contro-narrativa diffusi
16	No Hate Speech Movement Italia Dipartimento della gioventù e del servizio civile nazionale (IT) (istituzione)	Attività: creazione campagne nazionali (una prima campagna istituzionale e una seconda che ha coinvolto le scuole mediante concorso indirizzato ai ragazzi (I° fase) e campagne attuali realizzate dagli attivisti del movimento in Italia (II° fase); Autori: professionisti della comunicazione (campagna istituzionale) e studenti appositamente formati (I° fase); attivisti di No Hate Speech Movement Italia (II°fase) Pubblico di riferimento che si intende sensibilizzare: giovani Mezzi radio e televisioni (I° fase); utilizzo quasi esclusivo dei social network (II°fase) Messaggi: ricorso a contenuti multimediali (infografiche, video)	Azioni stimolate nei giovani: - condivisione materiali sull'importanza contro-narrativa; - condivisione contenuti della campagne, loghi e banner; - produzione autonoma di contro-narrativa; - partecipazione a percorsi educativi sul tema della contro-narrativa (ovvero anche formarsi attraverso la lettura dei materiali educativi messi a disposizione); - segnalazione di contenuti di odio online alle autorità competenti; - la campagna intende anche stimolare gli autori di *hate speech* online alla rimozione dei contenuti di odio.

Entrando nel dettaglio della realizzazione delle campagne di contro-narrativa online, si evidenzia in primo luogo che il **pubblico di riferimento** delle campagne di contro-narrativa è costituito nella maggioranza dei casi individuati dalla generalità degli utenti online che fanno uso dei social media. Molta attenzione è dedicata anche al segmento dei giovani. Infine, On/of Deread produce contenuti di contro-narrativa indirizzati specificamente a estremisti di destra, a radicalizzati islamici o a soggetti a rischio.

Si evidenzia, inoltre, il ruolo fondamentale svolto dai social network come **mezzo** principale attraverso cui vengono veicolate le campagne. Anche nel caso dell'esperienza *No Hate Speech Movement*, che ha previsto la realizzazione di un portale informatico in cui raccogliere tutti i contributi realizzati nel corso della campagna di contro-narrativa, i social network sono stati comunque utilizzati per meglio pubblicizzare l'esperienza. La centralità del ruolo dei social network è ribadita anche dai manuali e dai toolkit realizzati per supportare la creazione di campagne di contro-narrativa; a questo proposito, se la maggioranza dei manuali e dei toolkit si concentra sui social network principali (Facebook, Twitter, Instagram e Youtube), è altresì vero che si riscontra una sempre maggiore attenzione anche verso social network emergenti, come Snapchat. L'esperienza *Preventing Radicalization to Terrorism and Violent Extremism* realizzata dal Gruppo di lavoro su Comunicazione e Narrativa Centro di Eccellenza RAN (Network europeo), ha realizzato linee guida che includono una trattazione dettagliata su come operativamente realizzare contro-narrativa anche su questi nuovi social network.

Relativamente ai contenuti (**messaggio**) delle campagne di contro-narrativa, emerge come sia particolarmente frequente nelle esperienze mappate il ricorso a contenuti multimediali. Considerando solo le esperienze in cui la campagna di contro-narrativa è già realizzata o in corso, *No Hate Speech Movement, No Hate Speech Movement Italia, PRISM* e *SilenceHate* fanno un ricorso importante a elementi grafici e video. Una scelta in linea con quanto emerso in letteratura relativamente al generale apprezzamento che il pubblico online mostra nei confronti di questo tipo di contenuto. Da segnalare è anche il ricorso ad un'impostazione ironico/satirica, come nel caso di *Donate The Hate* e *Turulpata***,** un'impostazione che è riconosciuta in letteratura come particolarmente efficace per coinvolgere il pubblico di riferimento.

L'esperienza *Turulpata,* in particolare, è molto interessante sotto il profilo dei contenuti che veicola e dell'impostazione che adotta, in quanto non prevede la produzione di contro-narrativa come risposta diretta a specifici *hate speech* online ma contro contenuti di odio e opinioni generalizzate. Attraverso la realizzazione di una pagina satirica su Facebook, si immagina un insediamento fittizio abitato e governato in Ungheria da persone di estrema destra. I post sulla pagina Facebook hanno ad oggetto riflessioni satiriche su questioni attuali di politica interna, credenze popolari e, in generale, temi utilizzati dall'estrema destra per diffondere la cultura dell'odio.

Sempre sotto il profilo dei contenuti, dall'analisi delle esperienze emerge che alcune campagne social si sostanziano nella produzione di contenuti di contro-narrativa associati a micro-donazioni a favore di progetti che promuovono i diritti umani e la tutela delle minoranze, come nel caso di *Donate the Hate* e *Wecounterhate*. In questi casi, è di tutta evidenza l'utilizzo di un tono ironico nell'impostazione dei messaggi, se si pensa che l'autore dell'*hate speech* online, ottiene come risposta che "grazie al suo post, si è potuto donare 1 euro all'organizzazione x che si occupa della protezione dei diritti di minoranze o altri ambiti di discriminazione".

Altro aspetto che emerge dall'analisi è che le esperienze non tendono a scegliere come **autori** delle campagne personalità famose. Infatti, tra i profili più ricercati come autore, si possono individuare soprattutto le giovani generazioni, i giornalisti e gli attivisti di ONG e associazioni. Le giovani generazioni, in particolare, sono una tipologia di autore su cui investono molte campagne: *React, SilenceHate, No Hate Speech Movement, Selma* e *PRISM***.** Questa scelta è legata al fatto che i giovani sono tra i soggetti maggiormente in grado di determinare un cambiamento culturale destinato a consolidarsi nel tempo. Inoltre, essendo "nativi digitali", passano molta parte del loro tempo online. Frequente, come già detto, è anche la scelta di individuare come autori di campagne di contro-narrativa attivisti di ONG e associazioni, dal momento che si tratta di soggetti già sensibilizzati e particolarmente attivi nel contrasto all'*hate speech* online. Sono 4 tra le esperienze mappate (*Hatemeter, On/ Off Derad, PET-SCAN* e *No Hate Speech Movement Italia* ad aver individuato in questi soggetti gli autori delle proprie campagne di contro-narrativa. Sta crescendo, altresì, l'interesse verso giornalisti e professionisti del settore dei media come autori di contro-narrativa (si vedano a questo proposito le esperienze *DACHS* e *SilenceHate*), poiché rappresentano un settore, quello dell'informazione, che gioca un ruolo chiave nel contrasto all'odio e all'intolleranza. Inoltre, i giornalisti sono soggetti sempre più bersagliati dagli *hate speech* online per le posizioni che assumono, specialmente come conseguenza di reportage, inchieste e articoli che si occupano di minoranze e di emergenze migratorie.

Accanto a queste tipologie di autori, si sta affiancando una nuova tipologia sempre più diffusa che vede come autori di contro-narrativa gli stessi utenti online, come mostra l'esperienza *#jagärhär,* il cui successo è proprio attestato dal significativo numero di utenti online coinvolti con riferimento all'esperienza realizzata in Svezia, dove la metodologia è nata, ma anche dalle altre 14 esperienze analoghe che sono state successivamente replicate in altrettanti paesi europei, portando alla nascita di un network europeo focalizzato sulla contro-narrativa.

Un'altra tipologia di autore che si sta diffondendo è quella rappresentata dagli algoritmi di intelligenza artificiale, che producono autonomamente contro-narrativa sulla base dei criteri su cui è stata impostata la campagna che li utilizza. *WeCounterHate,* in particolare, prevede una modalità totalmente automatica di produzione dei

contenuti di contro-narrativa, con l'organizzazione che attua l'iniziativa che demanda direttamente all'algoritmo la produzione e l'invio dei contenuti di contro-narrativa in risposta agli *hate speech* che vengono di volta in volta identificati.

Diversamente, esperienze come *Donate the Hate* e *Hatemeter* prevedono, in particolare, una **modalità semi-automatica** di produzione di messaggi di contro-narrativa, nel senso che si tratta di messaggi pre-impostati prodotti dall'algoritmo che vengono messi a disposizione di specifici soggetti formati nell'ambito delle esperienze per diventare autori di contro-narrativa, che possono decidere a quali autori di *hate speech* rispondere e, nel caso di *Hatemeter*, anche di modificarne in parte il contenuto.

Nonostante l'applicazione di algoritmi di intelligenza artificiale alla contro-narrativa sia attualmente in fase di avvio, i primi risultati dimostrano quanto questi sviluppi possano essere promettenti.

Le esperienze di contro-narrativa analizzate evidenziano il ruolo fondamentale della formazione, come elemento propedeutico all'avvio di una campagna di contro-narrativa. In tutte le esperienze, infatti, la formazione è riconosciuta strategica per stimolare e formare gli autori per la produzione di una contro-narrativa che sia costruttiva ed efficace, così come anche per fornire supporto per la gestione, anche di tipo emotivo, degli scambi di messaggi con gli autori di *hate speech*, con la consapevolezza che il confronto può avvenire anche con soggetti che appartengono a gruppi radicalizzati/estremisti organizzati.

La formazione è fondamentale anche nel caso in cui gli autori della campagna non siano volontari, bensì professionisti. Più nello specifico, *DACHS, HATEMETER, PET-SCAN, PRISM* e *SilenceHate* prevedono attività formative per giornalisti, professionisti del settore audiovisivo e moderatori di commenti nelle pagine web; *HATEMETER* e *PET-SCAN* anche per gli attivisti delle ONG. L'esperienza PET- SCAN prevede, in particolare, oltre alla creazione di materiale didattico per una formazione mirata per target, anche la creazione di una piattaforma di e-learning che consente l'erogazione della formazione agli attivisti di ONG di diversi Paesi europei e la costituzione di un network per la creazione coordinata di campagne pilota di contro-narrativa.

La formazione è centrale anche per le esperienze che sviluppano contro-narrativa coinvolgendo volontari e giovani come autori, come nel caso della *Task Force di Amnesty International Italia* che ha previsto attività formativa diretta ai volontari e *No Hate Speech Movement Italia, PRISM, REACT, SELMA* e *SilenceHate* le cui attività formative si sono indirizzate ai giovani.

Con riferimento alle **azioni** che le campagne di contro-narrativa intendono stimolare nel pubblico di riferimento, si segnala come la generalità delle esperienze condivida la finalità di stimolare il pubblico di riferimento a diventare soggetto attivo nell'ambito della contro-narrativa, sia diffondendo i contenuti delle campagne, sia iniziando a produrre contro-narrativa in modo autonomo e spontaneo. Un'altra azione generalmente perseguita è quella di stimolare un cambiamento positivo nelle opinioni degli autori degli *hate speech* online, che si concretizza anche nella rimozione dell'*hate speech*. Alcune campagne, in particolare, come detto, si rivolgono agli estremisti di destra e ai soggetti radicalizzati islamici (ad esempio *On/Off Derad*), dove attraverso una contro-narrativa veicolata da due pagine Facebook si è inteso stimolare tali soggetti ad avviare un dialogo "uno-a-uno" con gli autori di contro-narrativa e ad uscire dai percorsi di radicalizzazione/estremismo.

3.7 La proposta di contro-narrativa del Progetto CO.N.T.R.O: "L'odio non è mai neutro"

A partire dalle analisi condotte con riferimento alle metodologie di contro-narrativa di cui ai paragrafi precedenti, nonché ad una intensa attività di benchmarking realizzata al fine di identificare meglio il "tono" e l'approccio da adottare, il progetto CO.N.T.R.O. ha predisposto una sua campagna di contro-narrativa che si è esplicata nella progettazione e creazione di una serie di mini video da disseminare online mediante i canali social del progetto e dei partner.

La sfida intrapresa è stata quella di realizzare un esperimento sociale che si muovesse a partire da alcuni elementi dati: due persone - complici - edotte rispetto al concept ed all'obiettivo dell'esperimento; una situazione di partenza, creata artificialmente per garantire autenticità a quanto sarebbe accaduto; un set costruito ad hoc, ma in maniera assolutamente veritiera, sempre a supporto di quanto stava per succedere.

Come noto, un esperimento sociale mira a testare la reazione delle persone a fronte di determinate situazioni o eventi; in questo caso, la reazione di alcune persone, ignare della non autenticità della situazione creata, rispetto al dilagante fenomeno dell'*hate speech* online che ha registrato negli ultimi anni tra i più alti tassi di crescita, nell'intrecciato rapporto tra comunicazione, vita quotidiana e digitalizzazione e che rappresenta, pertanto, una delle sfide più rilevanti poste dalle piattaforme di social media - e non solo - su Internet.

Nei paragrafi che seguono, con lo stile della narrazione, viene presentato l'esperimento sociale condotto nell'ambito del progetto C.O.N.T.R.O., con l'obiettivo di accompagnare il lettore come se si trovasse davanti alla situazione creata e l'invito ad esplorare i social per godere a pieno dei risultati raggiunti da questo esperimento.

3.7.1 Concept

In Italia, in una località di mare, alcuni turisti aspettano l'imbarco su una nave da crociera in una sala d'attesa. Una ragazza dalla pelle scura, nei suoi vestiti da teenager, è intenta a leggere un fumetto. Di tanto in tanto, come sempre fanno gli adolescenti, dà un'occhiata al suo smartphone.

Poi la ragazza, che conosce poco e male la lingua italiana, si rivolge alla persona che le siede accanto: ha bisogno che qualcuno la aiuti a capire cosa c'è scritto in un post appena pubblicato sul suo profilo social, un messaggio d'odio che la riguarda direttamente e che contiene linguaggio violento.

Alternativamente alla ragazza, la scena è gestita da un uomo, sulla quarantina, sempre di origine centrafricana, anch'egli intento a leggere un fumetto.

I contenuti dei messaggi di odio che queste due persone ricevono sono tremendi, come le migliaia di messaggi raccolti settimanalmente dall'Ufficio Nazionale Antidiscriminazioni Razziali. Un mix di volgarità, razzismo e terrorismo e, nel caso della ragazza, anche sessismo, con riferimenti più che espliciti alla prostituzione.

I turisti italiani, protagonisti a loro insaputa di quello che è un vero e proprio esperimento sociale, mostrano imbarazzo, empatia, disappunto, vergogna, rifiuto, emozioni. Non senza difficoltà trovano un loro modo di dialogare con la vittima di *hate speech* - una vittima in carne e ossa, concreta, che è lì con loro, e non nel limbo impalpabile del web.

Sono persone di diversa età, estrazione sociale, sesso. Di fronte al messaggio violento c'è chi preferisce non tradurre il messaggio, tenendosi sulle sue: non lo riguarda. Altri si informano sull'autore del post. Alcuni riconoscono il

contenuto del messaggio come razzista. Sotto l'insistenza della ragazza, qualcuno cede a tradurre parola per parola il messaggio d'odio. La tensione è palpabile. E non manca chi si espone a indicare una "soluzione" alla situazione: cancellare il messaggio, bloccare il profilo, parlare della cosa con un adulto o un famigliare, segnalare il caso di *hate speech* a chi gestisce la piattaforma social, rivolgersi alla polizia postale, denunciare l'accaduto alle autorità.

L'esperimento si conclude con lo svelamento dell'esperimento sociale. La troupe che ha registrato tutto esce allo scoperto: i turisti mostrano sollievo, confusione, alcuni hanno bisogno di qualche minuto per elaborare l'accaduto e capire (davvero) cosa è capitato loro. Anche i corpi parlano: c'è chi si abbandona sulla sedia, chi esorcizza il dolore in una risata, chi strappa un abbraccio alla ragazza, chi ha un piccolo crollo emotivo e si asciuga gli occhi umidi.

3.7.2 Location

Il porto è un set evocativo. Parla di vacanza, crociere, viaggi. Ma è anche un approdo, teatro di sbarchi, barconi, migranti. Come una stazione o l'aeroporto, il porto è un "*non luogo*", un posto fuori dal tempo e dallo spazio.

Navigante è chi va per mare. Naviganti sono gli utenti della rete e dei social network.

Effettuare delle riprese in uno scalo crocieristico è piuttosto complesso e tanto più lo è diventato con l'emergenza sanitaria in cui i mini video sono stati girati. Per non rinunciare a una cornice che dice molto dell'Italia come porto d'Europa, si è fatto ricorso a un teatro di posa che potesse ospitare poche persone alla volta nel rispetto del distanziamento sociale e del tracciamento di tutti gli individui coinvolti.

Per garantire autenticità all'esperimento sociale è stata creata una fittizia società di comunicazione che opera nel settore crociere. Questa società ha organizzato un casting pubblicitario a Trieste, importante scalo crocieristico italiano, con l'intenzione di promuovere il turismo nel Mediterraneo a bordo delle grandi navi. Agli attori che hanno sottoposto la loro candidatura per partecipare al provino è stato chiesto di arrivare in abiti leggeri e muniti di bagaglio a mano: pronti a partire per una crociera.

In rete, sulla piattaforma di Facebook, parallelamente è stato creato un gruppo chiuso a tema fumetti, supereroi, graphic novel. Popolato da persone amiche e da una serie di fake-profile creati appositamente per l'occasione, il gruppo chiuso ha messo in scena un comune scambio di contenuti dove alcuni post di linguaggio prima ambiguo, poi violento, sono stati indirizzati verso le due vittime designate.

3.7.3 Protagonisti

I protagonisti dell'esperimento sociale, i candidati per il casting, sono persone di diversa età, estrazione sociale, sesso. Arrivano tutti dalla cosiddetta società civile: studenti, lavoratori, pensionati. Molti di loro frequentano il mondo delle comparse e del teatro amatoriale. Pochi svolgono il lavoro di attore a tempo pieno.

Al provino si presentano puntuali. Nei gesti dell'attesa si colgono tratti di impazienza, misti ad irrequietezza, tipici dei momenti che precedono una prova importante.

Per molti di loro il teatro di posa è un luogo famigliare. Sono abituati a questo genere di cose. E tuttavia, fatto salvo per un'eccezione, nessuno dubita dell'esperienza che si trova a vivere. Quando la ragazza e/o l'uomo mostrano il messaggio sul telefono, in molti dimenticano del tutto il trascorrere del tempo (passano dieci, quindici minuti prima che la troupe esca allo scoperto): il "non luogo" diventa anche "senza tempo".

La ragazza è un'adolescente di 13 anni. Nata e cresciuta a Trieste; è figlia di una coppia di origine camerunense.

L'uomo, sulla quarantina, è anche lui di origine centrafricana. Vive e lavora a Trieste da vent'anni come ingegnere in un'azienda di caffè.

Quando i candidati per casting pubblicitario organizzato fittiziamente con l'intenzione di promuovere il turismo nel Mediterraneo a bordo delle grandi navi da crociera vengono accompagnati in sala d'attesa, i due protagonisti complici stanno leggendo una rivista di fumetti: nella finzione dell'esperimento sociale, i fumetti sono il loro modo per imparare l'italiano. E sono anche la ragione per cui partecipano al gruppo di fumettari su Facebook. È un loro post online a scatenare l'aggressione verbale di cui sono vittime.

3.7.4 Struttura

L'esperimento sociale ha prodotto una straordinaria abbondanza di materiale registrato, che restituisce un'insolita e lucida fotografia del rapporto degli italiani con il fenomeno *hate speech*. Quasi venti ore di girato, selezionate e montate in 8 video racconti, ai quali si aggiungono una sintesi e un dietro le quinte che dà voce ai partner del progetto.

Negli otto video racconti la ragazza e l'uomo incontrano, di volta in volta, tre diverse persone. La struttura del racconto è sempre la stessa: le immagini di apertura portano lo spettatore in Italia, in una località di vacanza, piena di turisti. Alcuni di loro li vediamo accedere a un terminal crociere. Le telecamere li seguono in una sala d'attesa, dove è allestito l'esperimento sociale.

I video racconti si concludono con le parole "l'odio non è mai neutro" cui segue una call to action: approfondire il tema sul sito del progetto CO.N.T.R.O.

L'obiettivo è dare vita a una contro-narrativa sui social.

Un nono video offre una sintesi degli incontri, mostrando la primissima reazione delle persone al messaggio d'odio e il sollievo quando la troupe esce allo scoperto svelando l'esperimento sociale. L'obiettivo è restituirne una visione complessiva.

Il decimo video è un vero e proprio backstage del progetto, con le voci dell'Ufficio Nazionale Antidiscriminazioni Razziali, quelle dei protagonisti, degli attori e della troupe. L'obiettivo è spiegare quali sono stati il punto di partenza e gli obiettivi che hanno guidato l'esperimento e sottolineare i risultati più interessanti prodotti.

3.7.5 Declinazione

La presenza di due diverse persone (una ragazza e un uomo) nell'esperimento sociale evidenzia aspetti diversi del fenomeno dell'*hate speech* in rete.

La prima evoca lo spettro della misoginia, della violenza nei confronti delle donne, della violenza a sfondo razziale, del sessismo, della prostituzione minorile, della tratta di esseri umani.

L'uomo mette in evidenza tutti i cliché legati all'universo migranti: dal lavoro nei campi, allo stigma del colore della pelle, passando per i morti in mare, la polemica sui telefonini, la ghettizzazione e il caporalato.

Non manca niente al triste ventaglio di invettive cui il linguaggio violento ricorre in rete e sui social.

"L'ODIO NON È MAI NEUTRO"...

4

Ricerca e analisi dei messaggi di odio online

Come si è visto nei capitoli precedenti, il progetto C.O.N.T.R.O., attraverso una prima fase di studio, ha inteso fornire un quadro d'insieme sulle metodologie di ricerca sui discorsi di odio online e sulle migliori strategie in uso per contrastarli. A questo è seguita una mirata campagna di comunicazione e sensibilizzazione sul fenomeno che ha previsto l'ideazione e la realizzazione di una serie di video di contro-narrativa che, accompagnati, da uno spot istituzionale, permettessero di agire attivamente contro quella "cultura" in cui i messaggi di odio online prolificano.

Il progetto ha, infine, anche avviato una riflessione per contribuire al perfezionamento dell'attività ricerca ed analisi dei messaggi di odio online con il fine ultimo di raggiungere una metodologia comune ed efficace contro i discorsi di odio online.

L'UNAR, infatti, dal 2015 ha istituito l'Osservatorio nazionale sulla discriminazione nei media e in internet, di fatto ampliando le attività di monitoraggio già svolte sui media tradizionali (giornali, TV e radio), consapevole del ruolo essenziale che i social media svolgono attualmente, dell'impatto negativo che un linguaggio inappropriato e discriminatorio può produrre e della conseguente necessità di monitorare ed analizzare tale fenomeno.

In questo capitolo viene offerto un contributo per la finalizzazione della metodologia dall'Ufficio Nazionale Antidiscriminazioni Razziali (UNAR) per monitorare e analizzare gli *hate speech* prodotti sui social network. Nello specifico, viene approfondita una possibile strategia di sviluppo dell'attività di UNAR per il monitoraggio e il contrasto dei messaggi di odio online che di fronte alla sempre più violenta e pericolosa pervasività dei discorsi e dei fenomeni di odio ad essi collegati possa fare leva su un approccio multidisciplinare.

Dopo un primo inquadramento concettuale che delinea il contesto di riferimento, ovvero cosa si intende per ricerca e monitoraggio dell'*hate speech* il capitolo prosegue con una dettagliata descrizione della metodologia che viene utilizzata nell'ambito dell'Osservatorio Media e Internet di UNAR, supportata anche da alcuni esempi concreti dei possibili risultati di analisi e delle potenzialità della costante e strutturata attività realizzata.

4.1

Il contesto di riferimento: breve introduzione all'analisi automatica dei testi

I metodi e le tecniche di analisi automatica dei testi si sono sviluppati a partire dalla crescente digitalizzazione e del conseguente moltiplicarsi di contenuti testuali disponibili online. Questi processi, ad oggi ancora in continuo sviluppo, hanno reso possibile il trattamento dei testi non soltanto come veicolo di contenuti legati ad uno specifico significato ma come dati veri e propri, dai quali è possibile estrarre informazioni che vanno al di là della semplice interpretazione linguistica. Attraverso l'utilizzo di strumenti informatici, il contenuto testuale non viene quindi soltanto "letto", ma viene utilizzato per ottenere rappresentazioni diversificate dell'informazione che rappresenta. Questo, come vedremo più avanti, avviene su più livelli: ci concentreremo in particolare sull'analisi automatica di tipo lessicale e sull'analisi automatica di tipo testuale.

A causa dell'elevato numero di dati testuali a disposizione, per'altro in costante aumento, è importante che le tecniche di estrazione dell'informazione e di analisi automatica siano applicabili anche a una consistente mole di documenti diversi, e che riconducano quello che si definisce un "*insieme di dati non strutturati*"[1] a un insieme di dati strutturati indipendentemente dalla quantità di dati a disposizione. In quest'ottica si inserisce il text Mining, un'area di ricerca multidisciplinare che, combinando con uguale importanza, strumenti della Linguistica Computazionale, dell'Information Retrieval e della Statistica, mira all'estrazione dell'informazione di interesse presente nei testi scritti in linguaggio naturale.

L'approccio utilizzato è di tipo metrico e fornisce delle misurazioni oggettive dei fenomeni studiati. Tali misurazioni sono valutazioni basate su analisi quantitative, e sono definite "oggettive" in quanto mantengono uniformi

1 (Bolasco 2005).

i criteri di osservazione nell'intera superficie dell'oggetto di studio costituito da un Corpus (collezione di testi), a prescindere dalle sue dimensioni. Ad ogni modo, l'analisi automatica di un Corpus lessicale resta di fatto un'analisi qualitativa, seppur integrata in una fase iniziale con metodi quantitativi, a garanzia della stabilità delle misurazioni.

L'approccio statistico del Text Mining permette quindi di analizzare corpora di qualsiasi dimensione. Tale approccio implica semmai che **i testi non debbano essere troppo piccoli,** in quanto sarebbero **poco robusti a un'analisi quantitativa delle frequenze**.

Generalmente una filiera di lavoro di Text Mining (TM) prevede i seguenti passaggi:

- Fase di pre-trattamento (pre-processing) dei testi, che consiste nel reperimento delle fonti testuali digitalizzate, nella loro tokenizzazione, per il riconoscimento delle unità d'analisi essenziali, e nella costituzione del document warehouse. Questa fase è prevalentemente informatica, e richiede spesso l'utilizzo di un software specifico.
- Fase di analisi lessicale e testuale (lexical and text processing), consistente nell'individuazione delle unità di analisi lessicali e testuali, nel riconoscimento dei vocaboli a cui è possibile attribuire meta-informazioni di tipo grammaticale o semantico, ma anche di tipo probabilistico riuscendo a distinguere fra essi le parole chiave, ovvero i termini significativi o peculiari. In questa fase l'ambito prevalente è quello linguistico.
- Fase di estrazione dell'informazione - il vero e proprio text mining - che si struttura mediante:
  - Information retrieval, ovvero il recupero mirato dell'informazione in formato elettronico, dove per informazioni di volta in volta si può intendere le forme grafiche, i concetti, i metadati, i documenti presenti in un database;
  - Analisi Multidimensionale dei dati, per semplificare, sintetizzare e rappresentare il fenomeno studiato;
  - Cluster Analysis/Categorizzazione automatica di documenti per raggruppare i documenti sulla base della loro similarità, ad esempio attraverso l'analisi delle co-occorrenze e la classificazione in dizionari semantici.

L'analisi avviene per modelli, ciascun modello costituisce una metrica utilizzata per classificare i dati testuali, ovvero una rappresentazione che può essere di tipo lessicale oppure di tipo testuale.

Nell' **Analisi automatica di tipo lessicale**[2] l'oggetto di studio è il lessico. In base alle caratteristiche del corpus in analisi e degli obiettivi della ricerca, di volta in volta si possono considerare come **unità di analisi lessicale** le singole **parole** così come riportate nel testo, le **multiword expressions**[3], i **lemmi**[4] o le **radici delle parole**[5]. Allo stesso modo si possono considerare come unità di analisi lessicale **le categorie grammaticali o semantiche** di ciascuna parola.

Nel **trattamento automatico di tipo testuale**, l'oggetto di studio è il Corpus nel complesso, inteso come susseguirsi di parole all'interno di una collezione di testi da analizzare, confrontare e categorizzare. **L'unità d'analisi è l'unità di contesto**, ovvero un **frammento di testo**, sia esso una **frase** o un intero **documento**.

In analogia con l'Analisi Lessicale, anche nell'Analisi Testuale ogni unità di analisi (frase, frammento di testo, documento etc.) costituisce un'entrata in un ***database documenti***. A ciascuna di queste unità di analisi testuale possono essere associate varie categorizzazioni.

Tali categorizzazioni possono riguardare da un lato le modalità delle variabili, codificate a priori. Ad esempio in un'analisi di risposte aperte, con testo libero, all'interno di un'indagine, le variabili categoriali di ciascuna risposta, quindi di ciascuna unità di analisi testuale, possono riguardare: il sesso, la provenienza geografica, il livello d'istruzione etc. degli intervistati. Dall'altro lato tali categorizzazioni riguardano il risultato dell'Analisi Testuale

2 L'analisi lessicale avviene attraverso l'esplorazione delle informazioni strutturate in un DB Vocabolario, dove ad ogni parola diversa (type) possono essere associate delle annotazioni di vario tipo: grammaticale, semantico e statistico. Tali annotazioni sono il risultato del trattamento automatico a diversi step dell'analisi. Ciascuna di queste proprietà costituisce un esempio di meta-informazione attribuita al type, recuperabile attraverso l'interrogazione dei corrispondenti campi del DB Vocabolario in cui tale informazione è depositata. L'estrazione/selezione delle parti del vocabolario serviranno per "raccontare" le caratteristiche lessicali del corpus evidenziando gli elementi significativi di ciascuna "parte del discorso", o per "illustrare" determinati sotto-insiemi di types e le relazioni esistenti fra essi.

3 Si tratta di lessie complesse, ovvero espressioni formate da due o più parole la cui unione rappresenta un'unica unità di senso, ad esempio: *presidente del consiglio, camera dei deputati, cassa integrazione* etc.

4 È la forma canonica, o entrata di citazione in un dizionario, di una parola. Ad esempio la forma canonica di un verbo è la forma al modo infinito presente (Forme: *mangio, mangiai, mangerò* - Lemma: *mangiare*).

5 Rappresenta la parte della parola senza la desinenza. La radice di una parola non è soggetta a variazione e contiene il significato fondamentale della parola. Ad esempio, le parole: *correre, corrente, corridore, corriere, corsa, corsaro*, condividono tutte la stessa radice *Cor-*.

vera e propria. Queste ultime categorizzazioni possono essere, anche in questo caso, di varia natura: sintattica, cioè ottenuta sulla base della valutazione di determinate strutture sintattiche o gruppi ad elementi variabili (ETL) presenti nei documenti analizzati; semantica, riguardanti le categorizzazioni automatiche sulla base di determinati lessici e significati; quantitativa.

Le analisi per produrre le rappresentazioni dell'***informazione*** possono quindi essere realizzate attraverso numerosi strumenti, in una logica di studio che si organizza in strategie finalizzate a produrre conoscenza. Ovvero, come già accennato, estrarre informazioni da **dati non strutturati**, sparsi in giacimenti o archivi talvolta di dimensioni sterminate.

In base al Corpus da analizzare si può definire di volta in volta una strategia diversa, plasmata in base agli obiettivi da raggiungere, e costruita utilizzando differenti combinazioni di strumenti e metodi a disposizione del ricercatore.

I limiti oggettivi dell'analisi sono quelli insiti nei dati testuali il cui riscatto dall'ambiguità è fortemente legato all'analisi del contesto e alla **qualità delle risorse impegnate,** non solo in termini di strumenti utilizzati ma anche di **esperienza del gruppo di ricercatori che si appresta definire la strategia e ad interpretarne i risultati**.

4.2
L'Osservatorio Media e Internet di UNAR

Dal 2015 UNAR ha istituito l'Osservatorio nazionale sulla discriminazione nei media e in internet, il cui obiettivo principale è la ricerca, il monitoraggio e l'analisi quotidiana - grazie a un software specifico ed un set di parole chiave - dei contenuti potenzialmente discriminatori provenienti dai principali Social Network (Facebook, Twitter, GooglePlus, Youtube), e dai Social media (articoli, blog e commenti di Forum). Nel fare ciò, l'Osservatorio adotta una strategia interdisciplinare che combina l'analisi, il monitoraggio e la tutela delle vittime con lo studio, la ricerca e l'ideazione di campagne ed iniziative tese alla sensibilizzazione degli utenti di internet in materia di lotta all'odio, all'intolleranza e alla violenza on line.

4.2.1 Il software utilizzato

Il software dell'osservatorio lavora sulla base di un set di keywords messe a punto dall'ufficio tenendo conto della letteratura scientifica e della esperienza nel contrasto alle discriminazioni acquisita in oltre dieci anni di attività, ed analizza quotidianamente, tramite la sentiment analysis, migliaia di contenuti: una cospicua parte di essi viene catalogata e raccolta in schede dossier tematici (ad esempio: *hate speech* e politica; *hate speech* e comunità rom; etc) mentre una parte meno rilevante quantitativamente, ma ritenuta fortemente discriminatoria, viene segnalata per la rimozione ai social network o agli amministratori dei siti web (prevalentemente giornali on line e blog) che ospitano i contenuti discriminatori.

Il Software utilizzato è fornito dalla Crimson Hexagon, società specializzata nell'approfondimento delle opinioni dei consumatori utilizzando strumenti dell'Intelligenza Artificiale[6]. La libreria di dati online dell'azienda è composta da oltre 1 trilione (1 miliardo di miliardi) di testi prodotti in tutto il mondo, e comprende documenti provenienti dai social network come Twitter, Instagram e Facebook, nonché blog, forum e siti di notizie. Dal 2018 la società si è fusa con la società Brandwatch[7].

Le principali funzionalità offerte ai propri clienti sono incentrate ad identificare gli interessi delle persone e le loro opinioni riguardo marchi, prodotti o aziende sui social media.

Lo strumento messo a disposizione è di facile utilizzo e non è necessaria nessuna competenza specifica da parte di chi lo usa. Bisogna limitarsi a definire delle "richieste di ricerca" (queries testuali), ovvero indicare alcune keywords in base alle quali ricercare i testi nel web. A riguardo è importante dettagliare che l'accesso alle fonti dei dati non è completamente illimitato. Ad esempio, mentre per Twitter è possibile effettuare il download di qualunque informazione semplicemente definendo la chiave di ricerca, per **Facebook si possono scaricare unicamente le informazioni di pagine web pubbliche.** Inoltre non si può impostare la ricerca su tutte le pagine pubbliche della piattaforma ma bisogna specificare esattamente su quali pagine effettuare la ricerca. Limiti simili riguardano le news sites.

I testi, che costituiranno il Corpus da analizzare, sono quindi principalmente commenti (il software li definisce "*post*"), prodotti dagli utenti dei social network, blog, forum o news sites, riguardo determinate pubblicazioni online.

Una volta individuata la collezione di testi, pertinenti alle keywords definite in un primo step, il software produce una serie di analisi automatiche volte ad identificare i principali argomenti trattati e a fornire alcune statistiche riguardo gli utenti che hanno prodotto tali commenti (classi di età e sesso) e le fonti da cui provengono.

Nel dettaglio il software esegue:

- download delle unità testuali da analizzare;
- sentiment analysis del Corpus;
- individuazione delle principali tematiche;
- produzione di statistiche descrittive delle caratteristiche categoriali;
- produzione automatica di un report in cui vengono elencati i risultati ottenuti.

4.2.2 Download delle unità testuali: definizione del Corpus

L'Osservatorio UNAR procede all'individuazione del Corpus da analizzare attraverso la definizione di due dizionari distinti di keywords, i quali devono coesistere (in un intorno definito di parole) all'interno di un "*post*" per essere considerati pertinenti.

Il primo dizionario è formato dalle parole (bisogna definire tutte le possibili flessioni delle parole) che identificano i ***soggetti*** su cui, di volta in volta, si focalizza l'analisi (si tratta principalmente di sostantivi) ad esempio: *extracomunitari, extracomunitario, immigrati, immigrato, profughi, stranieri...* oppure *ebreo, ebrei, rabbino, sionista...* oppure *rom, sinti, zingari, zingaro...*

Il secondo dizionario è composto dalle ***espressioni*** di odio a loro rivolte (aggettivi e verbi) ad esempio: *sporchi, bastardi, ammazziamoli, bruciamoli, devono morire...*

Questi due dizionari costituiscono quella che tecnicamente viene definita una Espressione Regolare[8]. Le parole all'interno di un dizionario sono fra loro alternative (basta la presenza di una di esse)[9] e per essere valida l'espressione regolare, ovvero per poter considerare pertinente il *post*, devono **coesistere in determinato intorno** (ad esempio all'interno di una frase di al massimo 20 parole) almeno un termine del dizionario "*soggetti*" e almeno un termine del dizionario "*espressioni*"[10].

Infine viene definito un **terzo dizionario** di **termini non pertinenti** con l'oggetto di studio[11]. Se un termine di tale dizionario è presente nel *post* ne determina la sua *non pertinenza*. La definizione di tale dizionario serve a limitare l'individuazione di *post* non pertinenti, ovvero falsi positivi (ad esempio: ***cd*** *rom*, ***enea*** *che fugge da troia...*).

Tutti i dizionari **sono definiti in maniera "esperta"** dall'**operatore dell'Osservatorio**.

Questa fase iniziale è particolarmente delicata e importante in quanto la corretta definizione delle "keywords" determina il corpus su cui verrà fatta l'analisi. La definizione da parte dell'operatore UNAR di **keywords semanticamente ambigue** fa sì che nel corpus possano essere presenti **frammenti di testo non pertinenti** (**falsi positivi**). Inoltre **la mancata definizione di keywords pertinenti all'oggetto di studio** può provocare una sotto valutazione del fenomeno (**falsi negativi**).

6 Fonte Wikipedia, *https://en.wikipedia.org/wiki/Crimson_Hexagon*.

7 *https://www.brandwatch.com/*.

8 Si tratta di una notazione algebrica che permette di definire in maniera formale e rigorosa dei modelli di stringhe. Attraverso le *Espressioni Regolari* si possono ricercare delle stringhe di testo che di volta in volta possono essere le occorrenze sia di singoli type o entità complesse, sia di loro classi che di relazioni fra classi o fra singoli type e classi. Il risultato di tale elaborazione consiste nel recuperare i frammenti che verificano la query testuale; inventariare la lista delle stringhe estratte; annotare eventualmente i frammenti.

9 Le parole sono collegate fra loro attraverso l'operatore logico "OR".

10 I due dizionari sono collegati fra di loro attraverso l'operatore logico "AND NOT".

11 Questo dizionario è collegato nell'espressione regolare ai primi due dizionari attraverso l'operatore logico "AND NOT".

Una volta definito il Corpus, l'operatore procede ad utilizzare in sequenza gli strumenti offerti dal software volti a descrivere il fenomeno oggetto di studio. In particolare si osservano le dimensioni del fenomeno, il sentiment, e si individuano gli argomenti principali presenti nel Corpus.

4.2.3 Sentiment Analysis e definizione del Volume di odio

La Sentiment Analysis (SA) ha come obiettivo di misurare le emozioni e sentimenti di un enunciato con il fine di evidenziarne la tonalità che può essere classificata come positiva, negativa oppure neutra.

Generalmente in una SA le parole vengono classificate sulla base di due riferimenti: uno linguistico di matrice etimologica; l'altro psicologico, con identificazione delle parole aventi o meno densità emozionale. La SA si situa all'incrocio fra alcuni modelli della psicoanalisi, con la sociologia e l'antropologia, orientati dalla linguistica e la semiologia.

La SA stabilisce delle differenze fra parole vuote, strumentali, ambigue e parole che portano con sé un senso pieno, marcando, nel loro occorrere, il significato emozionale del discorso.

Non essendo definiti in nessun modo i criteri e gli strumenti utilizzati dal software per misurare la tonalità dei "*post*", ci limitiamo in questo paragrafo ad effettuare alcune considerazioni rispetto all'utilizzo di tale analisi sui Corpora ricercati dall'UNAR.

Volendo infatti a priori identificare unicamente i messaggi di odio nel web ed avendo pertanto definito delle chiavi di ricerca volte a considerare come pertinenti solamente i messaggi in cui fossero presenti delle espressioni violente rispetto ad alcuni soggetti sensibili, la SA può essere utilizzata in questo contesto unicamente come filtro rispetto alla prima definizione del Corpus. Ovvero vengono considerati come messaggi di odio all'interno del Corpus, ottenuto attraverso le chiavi di ricerca iniziale, solamente i commenti con una tonalità negativa. In questa direzione infatti il report automatico prodotto dal software fornisce come primo risultato una descrizione del fenomeno analizzato attraverso una misura dei volumi dei "*contenuti sul web che incitano, promuovono o giustificano odio, disprezzo o altre forme di intolleranza*", basata esclusivamente sul conteggio dei *post* con una tonalità negativa. Tuttavia non è chiaro se gli ulteriori risultati proposti nei report si basino unicamente sui "*post negativi*" oppure sul Corpus completo ottenuto inizialmente con l'Espressione Regolare.

4.2.4 Identificazione delle tematiche principali: parole dell'odio, ruota degli argomenti, universi tematici

Le principali rappresentazioni dell'informazione di interesse sono esplicitate all'interno dei Report attraverso la visualizzazione di tre risultati.

La prima rappresentazione riguarda "*le parole dell'odio*", si tratta basicamente di un word cloud, ovvero una visualizzazione di una nuvola delle parole più occorrenti all'interno del Corpus. La posizione delle parole all'interno della nuvola è casuale e serve unicamente a fornire un quadro di quali siano le parole più utilizzate.

La seconda rappresentazione riguarda la "*ruota degli argomenti*". Anche in questo caso non vi sono elementi per poter definire in che modo siano rappresentate le parole caratterizzanti gli argomenti individuati. Viene ad ogni modo offerta una visualizzazione a due strati gerarchici delle tematiche riscontrate, ogni tema è definito da una parola. Possibilmente tale risultato è ottenuto attraverso l'utilizzo di uno dei Topic Model[12] attualmente disponibili in letteratura.

La terza rappresentazione dell'informazione viene definita come "*universi tematici*" ed è affidata alla visualizzazione di un grafo ottenuto con una Network Analysis. Si tratta di un'analisi delle co-occorrenze presentate sotto forma di un grafo di parole fra di esse associate. Tale analisi implementa i concetti di teoria dei grafi utilizzati nell'Analisi dei Dati Relazionali. All'interno del grafo le parole sono i nodi e i collegamenti sono rappresentati dalle co-occorrenze delle parole all'interno di una frase. Non conosciamo le impostazioni di base usate dal software ma generalmente la dimensione delle parole può essere proporzionale alla sua frequenza nel testo oppure al suo *degree* (numero di collegamenti con altre parole). Inoltre all'interno del Grafo è possibile, ad esempio attraverso il calcolo della modularity, identificare delle "comunità di parole" che rappresentano a loro volta un'individuazione dei temi presenti nel Corpus.

12 Nel Machine Learning e nel Natural Language Processing, un Topic Model è un tipo di modello statistico utilizzato per evidenziare gli "argomenti" astratti presenti in una raccolta di documenti. I Topic Models sono spesso usati come strumenti di Text Mining per l'individuazione delle strutture semantiche nascoste all'interno di un Corpus. Si tratta di modelli probabilistici di tipo Bayesiano, il più utilizzato in letteratura è la Latent Dirichlet Allocation.

4.2.5 Statistiche descrittive delle caratteristiche categoriali

L'ultimo risultato proposto all'interno del report riguarda la produzione di alcune statistiche riassuntive delle caratteristiche categoriali del Corpus in analisi. In particolare vengono delineate le caratteristiche degli Haters in base al sesso e all'età. Inoltre vengono evidenziate le dimensioni delle Fonti che definiscono il Corpus.

I dati riguardo **le caratteristiche degli haters** sono calcolati in base alle **informazioni** categoriali **disponibili** su alcuni dei *post* identificati nella fase iniziale.

Le dimensioni delle Fonti rispecchiano la provenienza dei *post* reperiti nel Web. A tal proposito va ricordato che tale reperimento rispecchia principalmente la **possibilità** stessa **del reperimento** dei *post* nel Web.

4.2.6 Alcuni elementi di attenzione

Unità di analisi lessicale ed iniziale definizione del Corpus

Un primo elemento da attenzionare nella metodologia attualmente in uso riguarda la definizione delle unità di analisi lessicale. Il software esegue le elaborazioni attraverso l'utilizzo esclusivo delle parole semplici presenti nei *post* nella loro forma flessa. Senza pertanto il supporto né di una identificazione automatica di Multiword Expressions (MWEs), che per loro natura sono dotate di una minore ambiguità semantica, né dell'utilizzo delle forme Lemma, che permettono di considerare tutte le flessioni di una parola. L'individuazione automatica di tali espressioni e il loro utilizzo come unità di analisi lessicale nelle diverse fasi dell'analisi, migliorerebbe da un lato i risultati delle elaborazioni riguardanti le identificazioni delle tematiche e dall'altro lato consentirebbe una migliore interpretazione dei risultati stessi. Inoltre consentirebbe all'operatore di escludere determinati sintagmi nominali non pertinenti con gli obiettivi di ricerca dell'Osservatorio.

In aggiunta, l'utilizzo delle parole semplici nella loro forma flessa complica l'attività iniziale dell'operatore di definizione dei dizionari di keywords, dovendo, soprattutto per le forme verbali, specificare tutte le possibili flessioni della forma da ricercare (ad esempio "ammazzare", "ammazziamoli", "ammazzano", "ammazzerei"...).

Individuazione automatica degli argomenti e tematiche

Gli strumenti utilizzati per identificare in maniera automatica gli argomenti e le tematiche all'interno del Corpus, in particolare il Topic model e la Network Analysis, sono strumenti sofisticati e caratteristici del Machine Learning. Si tratta degli strumenti più utilizzati attualmente, specialmente nelle analisi in ambito aziendale riguardo testi provenienti dai Social Network. Tali strumenti si prestano bene ad identificare in maniera automatica delle **tematiche sintetiche** emergenti in un Corpus. Si tratta di strumenti molto utilizzati soprattutto per evidenziare le principali tematiche diffuse nei Social, mettendo in evidenza poche parole su cui costruire ad esempio una campagna pubblicitaria mirata rispetto alle tendenze del momento, oppure per osservare le opinioni prodotte intorno ad un determinato tema.

I risultati ottenuti sono, però, di difficile o quanto meno scarsa interpretazione all'interno di un progetto complesso di ricerca sociale come quello a cui mira l'Osservatorio.

Interpretazione delle statistiche

In ultimo il punto maggiormente da attenzionare riguarda le statistiche prodotte e la loro interpretazione. In particolare si procede ad interpretare i risultati attraverso una **generalizzazione delle dimensioni del fenomeno** studiato senza considerare le caratteristiche delle fonti da cui provengono. Ad esempio in tutti i Report dell'Osservatorio si afferma che Twitter rappresenta il principale canale di produzione dei messaggi di odio. Tale affermazione si basa sulla distorsione iniziale generata della possibilità di reperimento dei dati. Ricordiamo infatti che non esistono limitazioni all'accesso delle informazioni su Twitter, mentre ad esempio su Facebook non tutti i post sono osservabili e scaricabili. Pertanto tale informazione non è generalizzabile ma deve riguardare esclusivamente le caratteristiche del **Campione** di "post" identificato grazie agli strumenti messi a disposizione dal software. A tal proposito ricordiamo anche le dimensioni di utilizzo di questi due Social network in Italia: Twitter ha 2,35 milioni di utenti attivi mensilmente nel nostro paese, mentre Facebook ne ha 31 milioni[13].

Allo stesso modo le statistiche riguardanti le classi di età degli haters sono evidentemente influenzate dalle caratteristiche del campione identificato dal software e non sono generalizzabili. Tali ultime informazioni infatti riguardano esclusivamente una parte del campione di post e alcune delle classi considerate sono formate da pochissime unità di analisi.

13 Fonte *We Are Social*, *https://wearesocial.com/*.

5

Possibili scenari di sviluppo

Come esposto nel capitolo precedente, allo stato attuale le analisi vengono svolte attraverso l'utilizzo di un software user friendly, che consente di ottenere alcuni risultati automatici ed immediati, principalmente attraverso l'identificazione di alcune tematiche sintetiche su cui si focalizzano i messaggi di odio che circolano nel Web.

Il software ha il vantaggio di poter essere usato da un utente che non deve avere particolari competenze, né riguardanti l'analisi automatica dei testi né riguardanti il contesto del fenomeno sociale studiato. Le uniche decisioni lasciate all'operatore riguardano la definizione delle keywords identificative del tema da ricercare. Successivamente, è il software stesso a produrre sistematicamente alcuni risultati che l'operatore deve interpretare, senza però avere la libertà di poter approfondire particolari aspetti o intraprendere differenti percorsi di ricerca che di volta in volta si potrebbero aprire così come si evince dai punti da attenzionare evidenziati nel precedente paragrafo finale del precedente capitolo.

Volendo invece intraprendere un percorso di ricerca sociale approfondito del fenomeno dell'Hate Speech che miri allo sviluppo di un sistema di monitoraggio e contrasto dei messaggi di odio online a regia istituzionale, le linee guida che seguono identificano le principali attività da mettere in atto per sviluppare l'attuale Osservatorio. L'obiettivo, anche considerando le potenzialità degli strumenti a disposizione, è quello di monitorare e misurare i discorsi d'odio online attraverso metodi e tecniche di analisi automatica dei testi, integrando le prospettive di analisi qualitativa e quantitativa dei contenuti.

5.1
Verso la definizione di linee guida operative per lo sviluppo di un sistema di monitoraggio e contrasto dei messaggi di odio online a regia istituzionale

Seguendo quando presentato come processo di filiera del text mining, l'aggiornamento della metodologia dell'Osservatorio dovrebbe prevedere le seguenti fasi:

1. Definizione di un sistema informatico strutturato per il recupero online di informazioni testuali

Attraverso la definizione di un sistema informatico strutturato per il recupero online dei dati testuali sarà possibile organizzare l'informazione in un "data warehouse" per le successive interrogazioni, trasformazioni, classificazioni e rappresentazioni dell'informazione. Tale attività verrà svolta da esperti di settore tramite l'utilizzo di strumenti *open source* già disponibili e che potranno essere adattati per le finalità dell'Osservatorio, così come specifici strumenti software progettati ad hoc per l'individuazione e l'analisi dei discorsi d'odio online.

L'attività di esplorazione del web dovrà riguardare almeno due canali diversi. Da un lato riguarderà i principali social media (Facebook, Instagram e Twitter). D'altra parte, il monitoraggio sarà indirizzato alle pagine online dei giornali, nonché ai blog e ai forum web. In entrambi i casi verrà definito un sistema di scansione web (*web scraping*) tramite la creazione di App autorizzate per recuperare contenuti di testo da pagine Web. I dati di testo recuperati sotto forma di dati non strutturati, saranno inseriti in un DB interno dell'Osservatorio sul quale verranno periodicamente condotte analisi al fine di "misurare" l'estensione del fenomeno studiato. La durata dell'attivazione del sistema di scansione web è fortemente legata all'ottenimento dell'autorizzazione dalle pagine da monitorare, autorizzazione necessaria principalmente per Facebook, che è ancora oggi il principale social network da cui estrarre dati testuali.

2. Analisi delle informazioni, consistente nella definizione di unità di analisi lessicale e testuale, la loro categorizzazione/classificazione attraverso annotazioni grammaticali, semantiche e probabilistiche.

L'analisi sarà condotta principalmente attraverso un approccio lessicometrico, finalizzato al riconoscimento e alla classificazione delle principali tematiche d'odio discusse sul web.

Il sistema di crawler delle informazioni testuali, già sperimentato in letteratura, sfrutterà le Graph API per recuperare il contenuto dei commenti ai post di Facebook e Twitter. Il crawler sfrutterà la flessibilità del framework Laravel per implementare un'ampia serie di funzionalità, come il riutilizzo del codice, la diversificazione dello storage e l'elaborazione parallela. Implementato come servizio Web, potrà essere controllato attraverso un Interfaccia web o tramite un comando cURL. Sarà in grado di memorizzare i dati nel filesystem sia come: file JSON[1]; Kafka queques[2]; indici Elasticsearch[3]. In base al numero di applicazioni fornito sarà in grado di eseguire il crawl di più pagine in parallelo. A partire dai post più recenti, il crawler raccoglierà tutte le informazioni relative ai post, oltre ai commenti ai post.

L'approccio metodologico scelto per l'analisi delle informazioni testuali è di tipo metrico ed è in grado di fornire delle misurazioni oggettive dei fenomeni studiati. Tale approccio, presentato nella sezione 5.1, consente di applicare tecniche proprie dell'analisi quantitativa al corpus di dati testuali oggetto d'indagine. L'analisi automatica dei testi può essere di tipo lessicale o testuale: a seconda delle caratteristiche del fenomeno che si vogliono approfondire è possibile focalizzare l'analisi sul lessico (termine che comprende singole parole, Multiwords expressions, lemmi o radici delle parole) o sul corpus, inteso come susseguirsi di parole all'interno di uno stesso documento testuale.

Si ritiene importante soffermarsi sulla possibilità di utilizzare come unità elementare, sia di analisi che di ricerca iniziale, i lemmi delle parole. Per lemma si intende la forma della parola non flessa né coniugata, quella che canonicamente compare nei dizionari al netto però della desinenza. La ricerca per lemmi, invece che per parole chiave, consente al ricercatore di non dover specificare a priori tutte le forme flesse della parola stessa. Questo tipo di operazione, otre ad essere dispendiosa in termini di tempo, può infatti portare a commettere errori di omissioni dei termini rilevanti (si pensi, ad esempio, alle 50 forme flesse che corrispondono a ciascuna forma-lemma dei verbi in italiano). Al contrario, l'utilizzo della forma lemma garantisce in primo luogo l'estrazione completa di tutto il materiale testuale all'interno del quale compaiono sia la forma canonica che le flessioni della parola. Per quello che riguarda l'analisi, invece, si configurerebbe la possibilità di accedere a un corpus contenente effettivamente tutte le diverse flessioni dei termini rilevanti. Ciò costituisce senza dubbio un punto di partenza ricco e stimolante per strutturare i passi successivi di analisi automatica di contenuto testuale.

Parallelamente è opportuno **definire una procedura automatica per l'identificazione dei sintagmi nominali (MWEs nominali) presenti nel corpus**. Il riconoscimento di tali entità complesse permette infatti di avere una rappresentazione dell'universo dei soggetti ed oggetti, semanticamente disambiguati, all'interno dei testi analizzati. Il risultato di tale passaggio consente di ottenere un dizionario terminologico dell'odio, che faciliterà sia tutte le successive analisi applicabili al Corpus sia l'interpretazione dei loro risultati. In letteratura vi sono numerosi strumenti per ottenere questo risultato. Uno dei possibili percorsi è dato dal riconoscimento dei sintagmi nominali mediante la formalizzazione delle loro strutture sintattiche, grazie all'utilizzo di Espressioni Regolari basate sulle meta-informazioni grammaticali delle parole (unità di analisi elementari).

Attraverso il riconoscimento delle polirematiche[4], collocazioni[5] e lessemi complessi[6] sarà possibile ottenere un dizionario di termini non ambigui da un punto di vista semantico.

---

1 *http://json.org.*

2 *http://kafka.apache.org.*

3 *https://www.elastic.co/products/elasticsearch.*

4 Una polirematica è costituita da un insieme di parole dotato di un sovrappiù semantico rispetto ai singoli componenti.

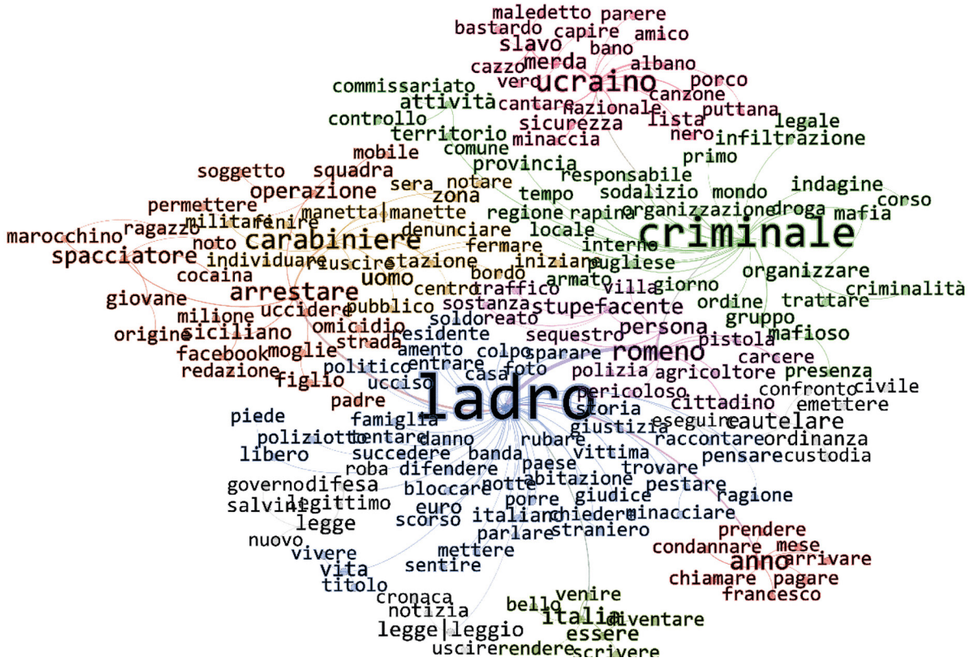
5 Una Collocazione è una sequenza di due o più parole caratterizzate da un forte legame reciproco.

6 I lessemi complessi sono particolarmente presenti in un Corpus tecnico specialistico e rappresentano un importante parte della terminologia di un settore.

In analogia con l'Analisi Lessicale, anche nell'Analisi Testuale ogni unità di analisi (frase, frammento di testo, documento etc.) costituisce un'entrata in un database documenti. A ciascuna di queste unità di analisi testuale possono essere associate varie meta-informazioni/categorizzazioni. Tali categorizzazioni possono riguardare da un lato le modalità delle variabili codificate a priori, ad esempio il periodo temporale del messaggio analizzato o la fonte (intesa come il tipo di social o la singola pagina). Dall'altro lato tali categorizzazioni riguardano il risultato dell'Analisi Testuale. Queste ultime categorizzazioni possono essere, anche in questo caso, di varia natura: sintattica, ottenuta attraverso la categorizzazione dei documenti in cui siano presenti determinate strutture sintattiche o gruppi ad elementi variabili (ETL); semantica, riguardanti le categorizzazioni automatiche sulla base di determinati lessici; quantitativa.

In ultimo è consigliabile, visto il contesto di ricerca sociale, di affiancare ai moderni strumenti del Machine Learning, anche strumenti classici dell'analisi dei testi che fanno ricorso alle tecniche statistiche multidimensionali, di tipo fattoriale, per studiare il contesto generale delle co-occorrenze delle lessie attraverso lo studio dei profili lessicali descritti nelle matrici dei dati. Una possibilità è ad esempio costituita dall'implementazione di un'analisi delle co-occorrenze attraverso gli indici di similarità o di vicinanza delle unità di analisi. L'analisi delle similarità è volta infatti a rilevare le co-occorrenze delle unità di analisi lessicale e a classificarle sulla base di quelle in insiemi semantici (o community). Con co-occorrenza si indica la frequenza con cui due termini appaiono congiunti nello stesso testo. Termini che appaiono insieme tendono perciò a identificare un concetto, o un gruppo di concetti; a partire da questi è possibile interpretare le varie articolazioni del discorso d'odio in relazione al tema considerato. Questa analisi implementa i concetti di teoria dei grafi utilizzati nell'Analisi dei Dati Relazionali (Network Analysis). A titolo esemplificativo si presenta, in Figura 5.1, un grafo ottenuto dall'analisi delle co-occorrenze e dal successivo raggruppamento in comunità semantiche delle estrazioni di post Twitter (tweet) relative alla tematica "Stranieri".

Figura 5.1 Grafo relazionale delle comunità semantiche - "Stranieri"



3. Adeguata interpretazione dei risultati ottenuti

I risultati ottenuti attraverso l'analisi automatica delle informazioni testuali, dati dall'integrazione delle tecniche di Text Mining, Machine Learning e di analisi statistica multidimensionali saranno quindi oggetto di interpretazione da un punto di vista linguistico, sociologico e antropologico.

Riguardo invece gli strumenti volti all'identificazione delle tematiche emergenti in un Corpus, questi a nostro avviso devono poter essere parametrati in base alle necessità dell'interpretazione dei risultati. Per fare ciò è indispensabile che l'operatore possa utilizzare strumenti in cui i parametri non siano predefiniti e rigidi come quelli offerti del software utilizzato, ma aperti e modificabili liberamente. Facendo riferimento all'esempio precedente di analisi di co-occorrenze lessicali, utilizzando un sistema di individuazione dei sub-corpus presenti in un insieme di dati testuali è possibile infatti individuare delle comunità semantiche che rappresentano le declinazioni del discorso d'odio oggetto di studio. Sulla specificazione dei parametri per la classificazione del corpus testuale in diverse comunità può andare ad agire il ricercatore, garantendo all'Osservatorio un risultato più approfondito di quello proposto da un software pre-impostato, oltre che più vicino alle reali ramificazioni tematiche dei discorsi d'odio.

A1

Breve ricognizione delle sentenze in applicazione della normativa antidiscriminatoria

Questo allegato presenta gli esiti di una breve ricognizione volta ad individuare, a titolo esemplificativo e senza alcun fine di esaustività, pronunce giurisprudenziali aventi ad oggetto **condotte ritenute discriminatorie/molestie razziali ai sensi dell'art. 2 del D.lgs 215/03 "**Attuazione della direttiva 2000/43/CE per la parità di trattamento tra le persone indipendentemente dalla razza e dall'origine etnica" - con particolare attenzione a quelle perpetrate sul web ma anche tramite altri strumenti di comunicazione e media (trasmissione radiofoniche o volantini) - e pronunce giurisprudenziali emesse in applicazione della normativa antidiscriminatoria in ambito razziale ai sensi della **L. 205/1993, cd. legge Mancino** "Misure urgenti in materia di discriminazione razziale, etnica e religiosa", che prevede il **reato di diffusione di idee fondate sulla superiorità o sull'odio razziale** (art. 1, lett a) e il **reato di istigazione alla violenza per motivi razziali** (art. 1, lett. b) e **la circostanza aggravante della finalità di discriminazione o di odio etnico, nazionale, razziale o religioso** (art. 3).

Ad integrazione dello studio si forniscono anche alcuni esempi di pronunce interessanti relative a:

- **Intervento del Garante per i diritti delle persone a rischio di discriminazione** per la rimozione di un post pubblicato su Facebook e considerato discriminatorio ai sensi del D.lgs 215/03
- **reato di diffamazione attraverso condotta online aggravato dalla finalità dell'odio razziale** (art. 3, L. 205/1993, cd. Legge Mancino);
- **reato di diffamazione in ambito razziale aggravato dall'utilizzo di "qualsiasi altro mezzo di pubblicità"** diverso dalla stampa (interessante aggravante rappresentata dall'utilizzo del sito internet/condotta online);
- **la responsabilità dei provider** per condotte discriminatorie realizzate online (anche se non riferibili specificatamente all'ambito razziale);

Il Box che segue fornisce indicazione delle banche dati consultate per la disamina in oggetto.

Box A1 Banche dati giurisprudenziali nazionali consultate

Corte costituzionale - Pronunce dal **1956** ad oggi

Corte di Cassazione - SentenzeWeb - Sentenze civili e penali degli ultimi cinque anni

Asgi.it

Osservatorio discriminazioni

Banca dati *http://www.giurisprudenzapenale.com*

https://www.eius.it/

Banca dati La legge per tutti *https://sentenze.laleggepertutti.it/#*

Miolegale.it

Pubblicazioni in rete individuate attraverso parole chiave (discriminazione razziale, Dlgs **215/03**; molestie razziale, social network e discriminazione).

Le sentenze che seguono costituiscono **esempi di pronunce aventi ad oggetto condotte discriminatorie che configurano la "molestia razziale" ai sensi del D.lgs 215/03.**

Consentono, quindi, di approfondire il tema delle tutele contro **la discriminazione e le molestie razziali**, ricostruendo gli strumenti offerti dal **d.lgs. n. 215/03**. Le prime tre, in particolare, in quanto riferibili a condotte perpetrate online, stimolano una riflessione sui *social network* e i limiti posti alla circolazione di informazioni e alla libertà di espressione in questo ambito. Le successive (n. 4 e n. 5), invece, sono riferibili a condotte discriminatorie/molestie razziali perpetrate ai sensi del D.lgs 215/2003 ma realizzate attraverso trasmissioni radiofoniche o volantini.

***1) Tribunale di Brescia - Ordinanza n. 11217/2015 emessa in data 31-11-2016: caso di discriminazione collettiva ex d.lgs. 215/03* (art. 5, 3 comma) e *2) sentenza n. 96 del 2019 della Corte di Appello di Brescia: caso di discriminazione collettiva ex d.lgs. 215/03* (art. 5, 3 comma) e "per associazione"**

Queste prime due pronunce (Ordinanza del Tribunale di Brescia e sentenza in 2° della Corte di appello) meritano, in particolare, di essere segnalate, oltre che perché configurano un caso di molestia razziale ai sensi del D.Lgs 215/03, anche perché applicano il principio secondo il quale la **tutela del diritto anti-discriminatorio può essere rivendicata non solo dalla persona che viene direttamente discriminata in ragione delle sue caratteristiche etnico-somatiche o nazionali, ma anche nel caso in cui la discriminazione avvenga "per associazione"**: cioè, quando la medesima tutela è rivendicata da coloro che, in quanto "associati" o frequentanti persone (familiari o amici) appartenenti alle categorie individuate dalla legge come meritevoli di tutela, rimangono vittime di comportamenti o di atti di discriminazione ispirati da razzismo e xenofobia.

Il box che segue approfondisce i contenuti delle pronunce in oggetto[1].

Box A2 Tribunale di Brescia - Ordinanza n. 11217/2015 emessa in data 31-11-2016: caso di discriminazione collettiva ex d.lgs. 215/03 (art. 5, 3 comma) e 2) sentenza n. 96 del 2019 della Corte di Appello di Brescia: caso di discriminazione collettiva ex d.lgs. 215/03 (art. 5, 3 comma) e "per associazione"

Le due pronunce hanno riguardato un caso di discriminazione collettiva a mezzo di ***social network.***

Il fatto aveva avuto origine da un post pubblicato sul proprio profilo Facebook ad opera della convenuta relativo a una immagine tratta dal quotidiano "Bresciaoggi". Nel post l'autrice accusava gli enti promotori di progetti di accoglienza per cittadini stranieri di lucrare opportunisticamente sul "traffico di clandestini" e in aggiunta ulteriori considerazioni di affine tenore. Le associazioni interessate avevano convenuto l'autrice del *post* dinanzi al Tribunale di Brescia **al fine di far cessare il comportamento ritenuto pregiudizievole** e **sentir dichiarare il carattere discriminatorio e molesto del commento ai sensi della del d.lgs. n. 215/2003**, art. **2**, nonché ottenere **il risarcimento dei danni**. In particolare, le organizzazioni ricorrenti ritenevano che il post integrasse gli estremi di una molestia rilevante *ex* art. **2**, co. **3**, del d.lgs. n. **215/03**, **giacché idoneo a "creare un clima intimidatorio, ostile, degradante, umiliante e offensivo",** attòraverso la rappresentazione, a un tempo beffarda e sprezzante, delle attività promosse dagli enti di accoglienza. Ciò veniva affermato non tanto sulla base del danno all'immagine e alla reputazione, estraneo al thema decidendum ai sensi dell'art. **112** c.p.c., quanto sulla base del fatto che **un'affermazione di tal genere** (l'attribuzione dello scopo di lucro), **visibile ad un numero potenzialmente illimitato di utenti del social network in quanto pubblica e più volte condivisa**, era sicuramente idonea a creare un "clima intimidatorio" e "ostile" nei confronti delle associazioni, e poteva avere certamente ripercussioni dirette sui servizi resi ai richiedenti asilo.

Le ricorrenti sostenevano anche che il post realizzasse **un'ipotesi di "discriminazione per associazione", rilevando il suo contrasto con la disposizione dell'art. 2, co. 3, del d.lgs. n. 215/03** dal momento che la succitata norma offre protezione contro qualsiasi condotta lesiva della dignità fondata su motivi di razza e/o etnia a prescindere che il soggetto leso sia in sé stesso qualificato da particolare razza o etnia.

L'autrice del post (convenuta), rimosso il post al momento della notifica del ricorso, si era difesa contestando il carattere discriminatorio del post, in quanto non contenente riferimento alcuno ad una particolare razza né avente lo scopo di ledere la dignità dei richiedenti asilo o di recare danno alla reputazione delle associazioni.

Al termine del giudizio l'autrice del post è risultata soccombente e condannata a risarcire le associazioni ricorrenti. Secondo il Tribunale di Brescia (ordinanza n. **11217/2015** emessa in data **31-11-2016** e comunicata in data **2-03-2017**) costituisce infatti discriminazione postare su Facebook affermazioni che hanno "valenza irridente e sbeffeggiante" laddove indicano che le associazioni di accoglienza perseguono in maniera opportunistica un fine di lucro, e contenuto "denigratorio e offensivo" quando indiscriminatamente definiscono "clandestini" i soggetti richiedenti asilo.

1 Si v. Renzi, *Labor - Il lavoro nel diritto*, **22** febbraio **2019**.

L'autrice del post giudicato molesto e discriminatorio ha appellato la sentenza di primo grado, lamentando nel merito che l'attribuzione del fine di lucro agli enti non intendeva suggerire lo svolgimento di attività illegali o inopportune e che l'utilizzo della parola "clandestino" non aveva un contenuto dispregiativo ma oggettivo, in quanto indicava la condizione di chi faccia ingresso illegale nel territorio dello Stato.

In disaccordo con la lettura offerta dall'appellante, la pronuncia della Corte di Appello (sentenza n. 96 del 2019 pubblicata il 18-01-2019) che si è limitata a confermare le determinazioni assunte dal giudice di primo grado nell'ordinanza, rigettando integralmente il nuovo gravame proposto.

La sentenza della Corte di appello è visionabile al link che segue:
http://www.rivistalabor.it/wp-content/uploads/2019/02/App.-Brescia-18-gennaio-2019-n.-96.pdf

2) Sentenza n. 467/2016, Sezione quarta civile - Corte di appello di Torino

La sentenza in oggetto oltre a configurare una **condotta discriminatoria/molestia razziale ai sensi del D.Lgs 215/03, art. 2**, costituisce **"atto ritorsivo"ai sensi del d.lgs. 215/03, art. 4bis**.

Il box che segue approfondisce i contenuti della sentenza in oggetto.

Box A3 Sentenza n. 467/2016, Sezione quarta civile - Corte di appello di Torino

La sentenza è stata emessa a seguito del ricorso (ricorrenti Pantè e Ghelma) proposto in Corte d'Appello ai sensi del D. Lgs 150/2011, art. 28 (controversie in materia di discriminazione)[2] per chiedere l'accertamento del **carattere discriminatorio e ritorsivo** della condotta tenuta dai convenuti (Comune di Varallo, Botta E. e Buonanno G.).
I ricorrenti assumevano, infatti, di aver proposto dinanzi al Tribunale di Torino, assieme all'ASGI (Associazione Studi Giuridici sull'Immigrazione), ricorso contro il Comune di Varallo, deducendo la violazione da parte dello stesso degli **artt. 43 d.lgs. 286/98** (TU sull'Immigrazione)[3] e **2 d.lgs. 215/2003,** in quanto alcune ordinanze comunali del 2009, avevano sancito sul territorio comunale **il divieto di indossare abbigliamento tale da ostacolare il riconoscimento della persona, fra cui, indumenti quali burqa, burqini e niqab, oltre che vietato l'attività di "vu cumprà" e mendicanti.** A questo scopo, **erano stati affissi grandi cartelli illustrativi del divieto**.

Il ricorso presentato terminava con declaratoria di inammissibilità e cessazione della materia del contendere in merito all'ordinanza comunale, in quanto nel frattempo revocata, ed i cartelli rimossi. Il giudice dichiarava quindi la virtuale soccombenza del Comune nei confronti dell'ASGI e compensava fra le parti le spese del procedimento.

Successivamente alla pronuncia, i convenuti soccombenti avevano affisso altri manifesti recanti il simbolo della "Città di Varallo" e la sottoscrizione del sindaco Botta E. e del pro sindaco Buonanno G., ove si affermava che i precedenti cartelli e i nuovi manifesti erano stati pagati dagli stessi. I manifesti rappresentavano inoltre il Botta ed il Buonanno sorridenti e riportavano la seguente scritta: "Il ricorso presentato dai 4 "comunistoidi" contro i cartelli situati agli ingressi della città di Varallo è costato alla collettività circa 3000 euro di spese legali!!! Soldi che invece potevano essere usati come ulteriori aiuti sociali per le persone in difficoltà!!! I 4 suonatori sono stati suonati perché il giudice ha dichiarato inammissibile il ricorso dando a loro torto su tutta la linea!!! Ecco alcuni stralci della sentenza:" seguiva il dispositivo, con l'indicazione dei nomi dei ricorrenti.

Secondo i ricorrenti, la diffusione dei manifesti in tutta la città, e specie nei pressi della scuola ove prestava servizio come professoressa la Pantè, ed il loro contenuto, **configurava atto ritorsivo sanzionato dall'art. 4bis d. lgs. 215/03 "sia perché poneva i ricorrenti, dileggiati da una simile campagna pubblicitaria in una situazione di svantaggio rispetto a quanti non avevano promosso azioni a tutela della parità di trattamento** (in applicazione dei normali criteri comparativi che regolano il principio di non discriminazione), **sia sotto il profilo delle molestie per motivi razziali ex art. 2 D.Lgs 215/03, trattandosi comunque di comportamento indesiderato avente lo scopo di violare la dignità di una persona e di creare un clima intimidatorio, ostile, degradante, umiliante e offensivo"**.

È stato quindi presentato un secondo ricorso con la richiesta da parte dei ricorrenti di accertare il carattere discriminatorio del comportamento descritto e la rimozione dei manifesti, ivi compresa **la rimozione del post riportante l'immagine del manifesto pubblicato sulla pagina Facebook del Buonanno e la pubblicazione integrale del provvedimento** sulla home page del sito comunale, nonché sul "Corriere Valsesiano", nonché sulla pagina Facebook del Buonanno. I ricorrenti chiedevano anche di disporre un piano di rimozione della discriminazione al fine di evitare in futuro il ripetersi di episodi simili e di condannare i convenuti al pagamento di un risarcimento a titolo di ristoro del danno non patrimoniale.

I convenuti negavano la ricorrenza dei presupposti dell'avversa azione, non facendo parte i ricorrenti di gruppi sociali discriminati né avendo essi subito discriminazione alcuna e sostenevano che si trattava di uno scontro di natura politica. In via subordinata e riconvenzionale, formulavano domanda di accertamento del comportamento ritorsivo dei ricorrenti nei loro confronti, con analoghe richieste di pubblicazione, inibitoria e risarcimento. Il giudice riteneva:

- che **l'appartenenza o meno di chi assumeva di essere vittima di ritorsione a gruppi sociali discriminati fosse questione irrilevante, spettando la legittimazione all'azione in parola anche a chiunque subisse una qualche ritorsione per essersi adoperato per evitare che altri subisse discriminazioni**;
- che la causa intentata dai ricorrenti unitamente ad ASGI davanti al Tribunale di Torino fosse un'azione diretta ad evitare ad altri discriminazioni, anche perché accompagnatasi ad altre iniziative dello stesso senso (interpellanze parlamentari, raccolte di firme, etc.);
- **che il comportamento dei convenuti dopo la decisione del Tribunale costituisse atto ritorsivo ai sensi della norma invocata, non scriminato dal ricorrere della "critica politica".**
- dichiarava cessata la materia del contendere sulla richiesta di rimozione dei manifesti e delle immagini sulla pagina Facebook del Buonanno, perché tale rimozione era già avvenuta;
- **disponeva la pubblicazione di dispositivo del provvedimento** sul Corriere Valsesiano e sulla pagina Facebook del Buonanno e sulla home page del Comune;
- rigettava la domanda di redazione di un "piano" di rimozione della discriminazione, stante la avvenuta rimozione dei cartelli;
- **accoglieva la domanda di risarcimento** a favore dei ricorrenti.

Quanto alla domanda riconvenzionale, essa veniva rigettata sia per inesistenza di un comportamento dei convenuti, diretto ad ottenere parità di trattamento, rispetto al quale quello tenuto dai ricorrenti potesse essere considerato ritorsione, sia perché le espressioni usate dai ricorrenti non venivano ritenute eccedenti i limiti della critica politica.

https://www.asgi.it/wp-content/uploads/2017/02/PANTE-GHELMA-c-COMUNE-VARALLO-BOTTA-BUONANNO-cda-torino-rg-21-del-2015-sentenza-467-del-22-03-2016-1.pdf

3) Ordinanza 6 giugno 2018 (causa civile iscritta al n. r.g. 69269/2016), Tribunale Ordinario di Milano - Prima Civile

La sentenza in oggetto configura una **condotta discriminatoria/molestia razziale ai sensi del D. Lgs 215/03, art. 2**. La condotta discriminatoria è integrata qualora le dichiarazioni rese siano tali da creare un clima ostile (cioè volto a diffondere odio e ad escludere i destinatari dalla compagine sociale tutelata dall'art 3 della Costituzione), degradante (in quanto in grado di colpire in modo offensivo ed avvilente la dignità dei gruppi sociali) e umiliante, per la gratuita attribuzione di qualità inferiori per etnia e nazionalità. Secondo la sentenza non può ritenersi che le espressioni utilizzate rientrino nell'ambito della libertà di manifestazione del pensiero politico qualora chi ricopre incarichi politici e istituzionali non abbia bilanciato le espressioni utilizzate con il rispetto e la dignità dei soggetti a cui si riferisce. Secondo la sentenza "la tutela del diritto all'eguale dignità e all'accesso paritario ai diritti fondamentali è frustrata da tale comportamento discriminatorio" e pertanto le associazioni aventi tale scopo statutario hanno diritto al risarcimento del danno non patrimoniale conseguente

2 Disposizioni complementari al codice di procedura civile in materia di riduzione e semplificazione dei procedimenti civili di cognizione.

3 Il D. Lgs **286/98**, all'art. **43** sancisce che costituisce discriminazione ogni comportamento che, direttamente o indirettamente, comporti una distinzione, esclusione, restrizione o preferenza basata sulla razza, il colore, l'ascendenza o l'origine nazionale o etnica, le convinzioni e le pratiche religiose, e che abbia lo scopo o l'effetto di distruggere o di compromettere il riconoscimento, il godimento o l'esercizio, in condizioni di parità, dei diritti umani e delle libertà fondamentali in campo politico, economico, sociale e culturale e in ogni altro settore della vita pubblica. (...).

al comportamento discriminatorio. Considerata, inoltre, la diffusione su larga scala delle dichiarazioni rese dal convenuto attraverso plurime interviste ad una trasmissione radiofonica di rilevanza nazionale, viene disposta a sue spese la pubblicazione del provvedimento oltre che su di un quotidiano di diffusione locale (sia cartacea sia on line) anche su di un quotidiano di diffusione nazionale.

Il box che segue approfondisce i contenuti della sentenza in oggetto.

Box A4 Ordinanza 6 giugno 2018 (causa civile iscritta al n. r.g. 69269/2016), Tribunale Ordinario di Milano - Prima Civile

L'ordinanza fa riferimento al ricorso presentato dalle associazioni Avvocati per niente ONLUS, APN e Associazione Studi Giuridici sull'Immigrazione (ASGI) ex D.Lgs n. 286/1998 (art. 44), D. lgs 150/2011 (art. 28) e art. 702 bis c.p.c.. Il ricorso ha avuto ad oggetto le dichiarazioni rese dal convenuto Joe Formaggio, sindaco del Comune di Albettone, nel corso della **trasmissione radiofonica La Zanzara** considerate dalle ricorrenti offensive nei confronti di soggetti appartenenti alle etnie rom e le dichiarazioni violentemente oppositive alla possibilità che alcuni richiedenti asilo venissero destinati ad occupare abitazioni site all'interno del Comune di Albettone. Tali ultime dichiarazioni contenevano frasi dal chiaro contenuto discriminatorio indicate specificatamente nel ricorso, tra cui:

- alla domanda del conduttore "come andiamo?", risposta: "Tutto bene, zero profughi all'orizzonte";
- "Non vogliamo extracomunitari";
- "Ti faccio un esempio (...) era luglio di quest'anno (...) un siciliano che ha tre case ad Albettone viene nel mio ristorante mi fa caro sindaco ho deciso una cosa (..) ospitiamo **4-5** negretti, perché li ha chiamati così (...), io li chiamo anche peggio (...), allora arriva qua mi fa allora mi sono messo d'accordo con la cooperativa, li mettiamo lì. Ho detto guarda che rischi grosso (...) ho parlato con alcuni paesani e mi hanno detto che se succede una cosa del genere, siccome siamo in campagna qua (...), riempiamo la casa di letame fino al soffitto (...)";
- "Dimmi cosa viene a fare un immigrato ad Albettone che rischia la pelle";
- "Lo devono capire che siamo razzisti";
- le persone di colore hanno un quoziente di intelligenza "molto più basso, lo dimostra la storia";
- "esportiamo cervelli e importiamo negri, pensa dove andremo";
- "facciamo il più grande allevamento d'Europa di maiali se dovesse essere che vogliono aprire una moschea ad Albettone";
- "è inutile che continuiamo con sta menata che siamo tutti uguali. Non siamo tutti uguali (...) ascolta, la storia l'ha fatta l'occidente e tra gli scienziati non ho mai visto uno di colore" (di fatto asserendo l'inferiorità delle persone di colore).

Le ricorrenti hanno altresì ricordato come il convenuto Formaggio non fosse nuovo a tali dichiarazioni sottolineando come espressioni dal contenuto simile erano state proferite anche precedentemente ai microfoni de La Zanzara. Infine, hanno ricordato che il Sindaco aveva dichiarato in più occasioni di volere erigere un muro lungo il perimetro del Comune di Albettone al fine di impedire l'ingresso dei migranti.

I ricorrenti hanno quindi lamentato che l'utilizzo reiterato di espressioni offensive nei confronti di soggetti contraddistinti da una particolare nazionalità o etnia costituisce comportamento discriminatorio e molesto e, in quanto tale, idoneo a creare sia un clima ostile perché volto a diffondere odio e ad escludere i destinatari dalla compagine sociale, sia un clima degradante perché in grado di colpire in modo offensivo ed avvilente la dignità dei gruppi sociali coinvolti violando il disposto di cui all'art 3 Cost., sia infine un clima umiliante ed offensivo attraverso l'utilizzo di espressioni mortificanti un intero gruppo etnico;

Il ricorso è stato accolto. Il giudice ha accertato e dichiarato il carattere discriminatorio del comportamento del resistente, ordianto il risarcimento del danno non patrimoniale e di dare adeguata pubblicità all'emanando provvedimento.

Per maggiori informazioni si v. la sentenza al link che segue:
https://www.asgi.it/wp-content/uploads/2018/06/Tribunale-di-Milano-06.06.201-Est.-Boroni-APN-e-ASGI-Avv.ti-Guariso-Neri-e-Marzolla-c.-XXX-Avv.ti-Roetta-e-Rigato.pdf

4) Ordinanza del 24 maggio 2012 (causa civile 34318/11), Sezione I civile - Tribunale di Milano

La pronuncia ha ad oggetto il ricorso dell'associazione NAGA (associazione di volontariato per la tutela dei diritti delle persone straniere) contro la Lega Nord. Il Tribunale ritiene, nella fattispecie, che costituisce "**molestia razziale**" ai sensi del D.Lgs 215/2003 l'utilizzo in campagna elettorale, da parte del partito politico succitato, di manifesti stradali e volantini - contenenti la parola "Zingaropoli" in quanto "emerge con chiarezza la valenza gravemente offensiva e umiliante" del termine stesso.

Per maggiori dettagli si v. la sentenza al link che segue.
https://www.meltingpot.org/IMG/pdf/trib_milano_ordinanza_28052012.pdf

La pronuncia che segue costituisce un'interessante **intervento del Garante per i diritti delle persone a rischio di discriminazione per la rimozione di un post pubblicato su Facebook e considerato discriminatorio ai sensi del D.lgs 215/03** con riferimento al divieto di discriminazione etnico-razziale anche nell'ambito dei beni e servizi offerti al pubblico.

Il Garante per i diritti delle persone a rischio di discriminazione della Regione Friuli Venezia Giulia attraverso apposita comunicazione indirizzata all'assessore ai Servizi generali, progetti europei e valorizzazione immobiliare del Comune di Trieste, ha richiesto l'immediata rettifica di un **post ritenuto discriminatorio e pubblicato su Facebook** relativamente alla manifestazione organizzata dal comune di Trieste "Aspettando la Befana, tradizione e solidarietà"**,** in quanto lo spazio dei social media è da considerarsi spazio di comunicazione pubblica, come ampiamente riconosciuto dalla giurisprudenza di Cassazione.

Il Garante, in particolare, rileva, come il **D.lgs 215/2003** abbia recepito nell'ordinamento italiano la direttiva europea 2000/43, che vieta le discriminazioni fondate sull'elemento etnico-razziale anche nell'ambito dei beni e servizi offerti al pubblico. Su questa base, chiede *"la piena assicurazione che la distribuzione dei giochi usati per i bambini bisognosi, nell'ambito della manifestazione in oggetto, includa famiglie e minori con cittadinanza straniera, regolarmente soggiornanti e residenti nel Comune di Trieste,* ***senza discriminazioni fondate su nazionalità, elementi etnico-razziali o di credo religioso***". Sottolinea, inoltre, che **una donazione pubblica di giocattoli, se promossa o patrocinata da un Ente o da un'Amministrazione pubblica**, deve conformarsi ai principi di imparzialità e buon andamento della Pubblica Amministrazione e **quindi deve rispettare le norme in materia di parità di trattamento e divieto di discriminazioni,** senza che possa avere rilievo il fatto che la donazione venga materialmente effettuata da un'associazione privata che invochi eventualmente il principio di libertà di associazione. Evidenzia inoltre che "*la giurisprudenza della Corte di Giustizia europea ha chiarito che* ***anche l'annuncio di una discriminazione costituisce un atto vietato di discriminazione****, nel momento in cui è idoneo a dissuadere fortemente i membri del gruppo collettivamente discriminato dall'avanzare la richiesta di accedere o avvalersi di un'opportunità, beneficio, bene o servizio offerto al pubblico, così come quando è suscettibile di contribuire alla diffusione, nell'opinione pubblica, di sentimenti di xenofobia, intolleranza, esclusione sociale e stigmatizzazione nei confronti della popolazione straniera, o di una minoranza etnica, che soggiorna legalmente sul territorio e deve ritenersi pienamente legittimata a sentirsi parte della comunità locale, con pari dignità sociale*".

Le due sentenze successive configurano invece alcuni esempi di **reato di diffusione di idee fondate sulla superiorità o sull'odio razziale** e **reato di istigazione alla violenza per motivi razziali** ai sensi della **legge 13 ottobre 1975, n. 654**, **art. 3 commi a) e b)**[4]**,** così come riformulato dalla legge **205 del 1993, art. 1, lett. a) e b)** (cd. Legge Mancino).

La legge, nella fattispecie, punisce:

a) chi propaganda idee fondate sulla superiorità o sull'odio razziale o etnico, ovvero istiga a commettere atti di discriminazione per motivi razziali etnici, nazionali o religiosi;

b) chi, in qualsiasi modo, istiga a commettere violenza o atti di provocazione alla violenza per motivi razziali, etnici, nazionali o religiosi.

La prima sentenza, che configura il reato di diffusione di idee fondate sulla superiorità o sull'odio razziale (ex art. 3,

4 Legge di autorizzazione alla ratifica della Convenzione internazionale sull'eliminazione di tutte le forme di discriminazione razziale, aperta alla firma a New York il 7 marzo 1966.

comma 1, lett a), fa riferimento a una **condotta perpetrata attraverso un programma radiofonico,** mentre la seconda (ex art. 3, comma 1, lett b), fa riferimento a una **condotta realizzata attraverso l'uso di un social network.**

1) Corte Suprema di Cassazione - V Sezione Penale, sentenza n. 32862 del 7 maggio 2019

La sentenza fa riferimento alla condotta dell'europarlamentare Borghezio che intervenendo in un programma radiofonico aveva pronunciato diverse dichiarazioni contro i Rom commentando l'incontro tra una delegazione di giovani rom con la ex presidente della Camera dei deputati, Laura Boldrini, in occasione della Giornata internazionale del popolo Rom dell'8 aprile.

La sentenza della Corte di Cassazione, nella fattispecie, annulla la sentenza di appello del 6-03-2018 che aveva confermato la pronuncia del Tribunale di Milano del 26-06 -2015 e che riconosceva il reato continuato di diffamazione aggravato dall'odio razziale ritenendo invece assorbito il reato di propaganda di idee fondato sull'odio razziale.

La sentenza della Cassazione annullando la sentenza di appello esclude quindi il reato di diffamazione contestato, riqualificando il fatto come **reato più grave di diffusione di idee fondate sulla superiorità o sull'odio razziale (legge n. 654 del 1975, art. 3, comma primo, lett. a**).

Per maggiori informazioni si può consultare la sentenza al seguente link:
https://www.eius.it/giurisprudenza/2019/577

2) Corte di Cassazione, sezione I penale, sentenza n. 42727 del 23 ottobre 2015

La sentenza della Corte di Cassazione, sez I Penale, conferma la legittimità della condanna a tredici mesi di reclusione (oltre alla pena accessoria ed al risarcimento in favore delle parti civili costituite già stabilita dai giudici di merito) contro una donna, ex esponente della Lega Nord, per **il reato di istigazione alla violenza per motivi razziali (di cui all'art. 3 primo comma lett. b) legge n. 654 del 1975**, aggravato ai sensi dell'art. 61 n. 10 cod. pen.). Si tratta, nella fattispecie, del caso della pubblicazione ad opera della condannata sul proprio profilo Facebook di un commento rivolto alla allora ministra per l'integrazione Kyenge a seguito della notizia di un'aggressione sessuale ad opera di un cittadino somalo. Il commento corredato da una fotografia della ministra recitava "mai nessuno che se la stupri, così tanto per capire cosa può provare la vittima di questo efferato reato, vergogna!"

Nel caso concreto, la Corte di Cassazione ha ravvisato la sussistenza del **reato di istigazione alla violenza per motivi "razziali" in ragione delle espressioni utilizzate dall'imputato, oltre che dal mezzo di comunicazione impiegato - e cioè la bacheca di un profilo "Facebook"** - e dal contesto sociale e politico nel quale le espressioni si collocavano. Il reato di incitamento alla violenza e gli atti di provocazione commessi per motivi "razziali", etnici, nazionali o religiosi, è infatti un reato di pericolo che si perfeziona indipendentemente dalla circostanza che l'istigazione sia accolta dai destinatari. Tuttavia è necessario valutare la concreta ed intrinseca capacità della condotta a determinare altri a compiere un'azione violenta, con riferimento al contesto specifico ed alle modalità del fatto.

Per maggiori informazioni si può consultare la sentenza al seguente link:
http://www.deiustitia.it/cms/cms_files/20151028024333_lvrc.pdf

Di seguito due esempi di pronunce configuranti **reato di diffamazione**, interessanti per il presente studio, nel primo caso per la previsione dell'**aggravante della finalità di odio razziale** (ai sensi della Legge 205/1993, art. 3 - Legge Mancino) oltre che per la tipologia di condotta perpetrata online e nel secondo caso per la previsione dell'**aggravante dell'utilizzo di "altro mezzo diverso dalla stampa**" ai sensi delll'art. 595, 3 comma (nella fattispecie sito internet).

1) Tribunale penale di Trento, 14 luglio 2014 n. 508

L'imputato è stato condannato dal Tribunale penale di Trento, per il reato di **diffamazione (art. 595 c.p.) aggravato dalle finalità di odio razziale di cui all'art. 3 della legge n. 205/2003**, per avere pubblicato sul proprio **profilo Facebook** un commento gravemente lesivo della reputazione dell'allora ministra dell'integrazione **Cecile Kyenge**. È opinione della Corte che le espressioni usate dall'imputato che aveva invitato la Kyenge "*a tornare nella giungla da cui era uscita*" *debbano essere ritenute altamente lesive dell'onore e del*

prestigio della ministra perché avvenute nel quadro di un complessivo dileggio già avviato precedentemente dall'esponente leghista Calderoli che l'aveva assimilata ad un orango, rimanendo in linea con la richiamata aggettivazione animalesca (...)".

Il Tribunale ritiene nello specifico che l'utilizzo dell'espressione succitata integri la **fattispecie di diffamazione aggravata dall'odio razziale ai sensi dell'art. 595 c.p. e art. 3 L. 205/2003**, in quanto tale espressione, lungi dal costituire una libera manifestazione di pensiero, "*finisce per significare una vera e propria volontà di discriminazione razziale a null'altro apparendo ispirata tale invettiva se non a suggerire l'idea di un'inferiorità originaria della persona determinata dal colore della pelle"* e dunque espressione altamente lesiva dell'onore e del prestigio della persona alla quale è riferita. Nemmeno il diritto di critica politica può essere invocato dalla difesa in quanto, la ministra, non veniva accusata di aver operato delle scelte politiche scorrete o di non sapere svolgere correttamente il proprio lavoro, ma veniva semplicemente invitata a tornare nel suo luogo di provenienza ovvero "la giungla" (quando questa in realtà ha origini Congolesi e dal 1994 ha acquisito la cittadinanza italiana) per il semplice motivo di non essere voluta dagli italiani[5]. La sentenza ha condannato l'imputato alla pena di una multa pari a 2500 euro e al risarcimento del danno in favore delle parti civili costituitesi (associazioni in difesa dei diritti degli stranieri che lamentavano una lesione alla propria immagine).

Anche il **giudice di appello** ha confermato la pronuncia del Tribunale (Corte d'Appello di Trento, sezione penale, sentenza del 1 giugno 2016).

Per maggiori informazioni si veda la sentenza completa al link seguente:
http://www.asgi.it/wp-content/uploads/2016/06/kYENGE-sentenza-appello.pdf

2) Cassazione Penale, Sezione V, n. 8328, 1 marzo 2016: Reato di diffamazione aggravato dalla diffusione a mezzo diverso dalla stampa

La Corte di Cassazione nella sentenza in oggetto, si è pronunciata avverso la sentenza n. 11554/2011 del 18/12/2013 emanata dal GUP del Tribunale di Palermo. Il Giudice per l'udienza preliminare aveva condannato l'imputato per il delitto di diffamazione di cui all'art. 595, co.1 e 3, c.p alla pena di euro 1.500,00 di multa, con la diminuente del rito abbreviato, per avere offeso la reputazione del Commissario Straordinario della Croce Rossa Italiana, comunicando con più persone, mediante la pubblicazione sul suo profilo Facebook, di alcune frasi denigratorie, associate all'immagine del predetto.

La Corte di Cassazione, ha confermato la pronuncia ritenendo che le frasi pronunciate dall'imputato "*sono oggettivamente lesive della reputazione della persona offesa, trasmodando in una gratuita ed immotivata aggressione delle qualità personali dello stesso*" e che la diffusione di un **messaggio diffamatorio attraverso l'uso di una bacheca Facebook integra un'ipotesi di diffamazione aggravata ai sensi dell'art. 595, comma 3, c.p., trattandosi di condotta potenzialmente capace di raggiungere un numero indeterminato o, comunque, quantitativamente apprezzabile di persone.**

Per maggiori informazioni è consultabile la sentenza al seguente link:
https://www.giurisprudenzapenale.com/wp-content/uploads/2016/11/Cass-8328-2016.pdf

Infine, si conclude la presente disamina con alcuni esempi interessanti di sentenze[6] che sanciscono la **responsabilità dei provider per condotte discriminatorie online**, anche se non espressamente riferibili all'ambito razziale.

1) Cassazione Penale, Sez. V, n. 54946, 27 dicembre 2016 - responsabiità penale diretta dei gestori di un sito internet per i commenti offensivi pubblicati dagli utenti

La Suprema Corte di Cassazione ha confermato la decisione della Corte d'Appello di Brescia, la quale aveva affermato la responsabilità penale in capo al legale Rappresentante di una S.r.l. gestore di un sito internet specializzato in ambito calcistico, nella cui *community* un utente aveva pubblicato un articolo corredato da alcune espressioni diffamatorie (a cui, per giunta, era stato allegato anche un file contenente il certificato penale riguardante la persona offesa) nei confronti del Presidente della Lega Nazionale Dilettanti della Fe-

5 Anna Baracchi e Alberto Guariso (a cura di), Rassegna parziale di giurisprudenza su *hate speech* e discriminazione, *www.asgi.it*.

6 Si v. Banca dati *http://www.giurisprudenzapenale.com*.

derazione Italiana Gioco Calcio. La Cassazione non mette in discussione che "l'articolo incriminato era stato autonomamente caricato sul sito dall'autore del medesimo, tuttavia, avendo il gestore ricevuto un messaggio di posta elettronica dall'utente, all'interno del quale era allegato proprio il file contenente il certificato sopra richiamato, accerta la responsabilità penale in capo al gestore non riconducendola al suo status di gestore del sito internet in quanto tale (non si tratta cioè di responsabilità c.d. da posizione), quanto piuttosto per aver mantenuto sul sito i contenuti offensivi, omettendo di rimuovere l'articolo, una volta venuto a conoscenza del carattere denigratorio pubblicato[7].

Per maggiori informazioni è consultabile la sentenza al seguente link:
https://www.giurisprudenzapenale.com/wp-content/uploads/2017/01/figc-cassazione-penale-1.pdf

2) Cassazione Penale, Sez. V, 24 marzo 2016, n. 12536: ammesso il sequestro preventivo di un blog in caso di diffamazione.

La sentenza in esame ha ad oggetto il caso di un giornalista-blogger, titolare di un sito internet, condannato dal Tribunale di Parma per aver commesso reato di diffamazione ai sensi dell'art. 595, primo e terzo comma, C.p. pubblicando nel proprio sito scritti dal contenuto altamente offensivo e denigratorio ai danni di due persone offese.

Nella sentenza in esame, la Suprema Corte è intervenuta sul tema della diffamazione a mezzo stampa, affermando che i nuovi mezzi di comunicazione quali **forum, blog, newsletter, newsgroup, mailing list e social network,** in quanto non registrati e non aventi un direttore responsabile, **non rientrano nel concetto di "stampa"**. Ne segue che gli stessi **ben possono essere oggetto di sequestro preventivo**, non godendo affatto delle garanzie costituzionali a tutela della manifestazione del pensiero, di cui - invece - godono i mezzi di comunicazione registrati[8].

Per maggiori informazioni è consultabile la sentenza al seguente link:
https://www.giurisprudenzapenale.com/wp-content/uploads/2016/05/Cass.-pen.-Sez.-V-24-marzo-2016-n.-12536.pdf

---

7 Si v. Miglio M., *I gestori di un sito internet rispondono penalmente per i commenti offensivi pubblicati dagli utenti*, in Giurisprudenza Penale Web, **2017**, **1** - ISSN **2499-846X**, *http://www.giurisprudenzapenale.com/2017/01/03/gestori-un-sito-internet-rispondono-penalmente-commenti-offensivi-pubblicati-dagli-utenti/.*
Sul tema si v. anche Si v. Gaetano Stea, La responsabilità penale dell'internet provider, in Giurisprudenza Penale, *https://www.giurisprudenzapenale.com/wp-content/uploads/2016/11/stea_provider_gp_2016_11.pdf.*

8 Si v. Lo Giudice H., *Ammesso il sequestro preventivo di un blog in caso di diffamazione. Non è "stampa"*, in Giurisprudenza Penale Web, **2016**, **5**, **10** maggio **2016**, *https://www.giurisprudenzapenale.com/2016/05/10/ammesso-sequestro-preventivo-un-blog-caso-diffamazione-non-stampa/.*

Bibliografia

ADL: Anti-Defamation League (2016). *Responding to Cyberhate: Progress and Trends*, **www.adl.org.**

AGCOM (2019). *Regolamento sulle nuove "Disposizioni in materia di rispetto della dignità umana e del principio di non discriminazione e di contrasto all'hate speech",* Delibera 157/19/CONS, ***https://www.agcom.it/documentazione/documento?p_p_auth=fLw7zRht&p_p_id=101_INSTANCE_FnOw5lVOIXoE&p_p_lifecycle=0&p_p_col_id=column-1&p_p_col_count=1&_101_INSTANCE_FnOw5lVOIXoE_struts_action=%2Fasset_publisher%2Fview_content&_101_INSTANCE_FnOw5lVOIXoE_assetEntryId=15055471&_101_INSTANCE_FnOw5lVOIXoE_type=document.***

Allport G. (1958) *The Nature of Prejudice*. Addison-Wesley.

Amnesty International (2018). *Conta fino a 10, barometro dell'odio in campagna elettorale*, ***https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2018/02/16105254/report-barometro-odio.pdf.***

Arci, Cittalia, *Discorsi d'odio e Social Media Criticità, strategie e pratiche d'intervento* ***https://www.arci.it/app/uploads/2018/05/progetto_PRISM_-_bassa.pdf.***

Assimakopoulos S., Baider F. and Millar S. (2017). *Hate Speech in the European Union: A Discourse - Analytic Perspective*. SpringerBriefs.

Balkin J. (2014). *Old School/New School Speech Regulation*. Harvard Law Review forthcoming.

Baracchi A. e Guariso A. (a cura di). *Rassegna parziale di giurisprudenza su hate speech e discriminazione*, ***www.asgi.it***.

Bartlett J. e Krasodomski-Jones A. (2015). *Counter-Speech: Examining Content that Challenges Extremism Online*. DEMOS, London.

Benesh S., Ruths D., Dillon K., Saleem H. e Wright L. (2016a) "Counterspeech on Twitter: A Field Study." Public Safety Canada - Kanishka Project: Evaluating Methods to Diminish Expression of Hatred and Extremism Online.

Benesh S., Ruths D., Dillon K., Saleem H. e Wright L. (2016b) "Considerations for Successful Counterspeech." Public Safety Canada - Kanishka Project: *Evaluating Methods to Diminish Expression of Hatred and Extremism Online.*

Binny M., Hardik T, Rajgaria S., Singhania P., Kalyan S., Goyal P., Mukherjee A. (2018). *Thou Shalt not Hate: Countering Online Hate Speech*. arXiv preprint arXiv:1808.04409.

Bolasco S. (2005). *Statistica testuale e text mining: alcuni paradigmi applicative,* Quaderni di statistica 7, 17-53

Bortone R. Cercuozzi F (2017). l'*Hate speech* al tempo di internet, *Aggiornamenti Sociali* - approfondimenti, pp. 818-827.

Briggs R. e Feve S. (2013). *Review of Programs to Counter Narratives of Violent Extremism. What works and what are the implications for government?* Institute for Strategic Dialogue.

Brocato R. (2016). "Hate Speech Online: Assessing Europe's Capacity to Tackle an Emerging Threat." *Freedom from Fear,* Issue No.12: Migrant Deadlock - The Abyss of Civilization. UN Publication.

Brown R. (2016). *Defusing hate: A strategic communication guide to counteract dangerous speech*, US Holocaust Memorial Museum. ***https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf.***

Burnamp P. and Williams M. (2015). "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making." *Policy and Internet* 7 (2), 223-243.

Camera dei deputati (2017). *La piramide dell'odio in Italia*, Commissione Jo Cox su fenomeni di odio, intolleranza, xenofobia e razzismo - Relazione finale, infografica, ***https://www.camera.it/application/xmanager/projects/leg17/attachments/shadow_primapagina/file_pdfs/000/007/099/Jo_Cox_Piramide_odio.pdf.***

Castells R. (2001). *The Internet Galaxy*. Oxford University Press.

CE (2000). *Direttiva 2000/43/CE del Consiglio del 29 giugno 2000 che attua il principio della parità di trattamento fra le persone indipendentemente dalla razza e dall'origine etnica*, ***https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32000L0043&from=DA***.

Citron D. and Norton H. (2011). "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age." Boston University Law Review, 91, 1435-1484.

CoE (1950). Convezione Europea per i Diritti Umani, ***https://www.coe.int/en/web/human-rights-convention/home*** (in inglese).

CoE (1995). Convenzione Quadro per la Protezione delle Minoranze Nazionali, ***https://rm.coe.int/168007cddo***.

CoE (1997). *Raccomandazione R (97) 20 sui discorsi d'odio*, ***https://rm.coe.int/1680505d5b*** (in inglese).

CoE (1997). *Raccomandazione* R (97) 21 sui media e la promozione di una cultura della tolleranza, ***https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505113b*** (in inglese).

CoE (2004). *Convention on Cybercrime: Protocol on Xenophobia and Racism. CETS* no.185.

CoE (2008). *Decisione Quadro 2008/913/GAI DEL CONSIGLIO del 28 novembre 2008 sulla lotta contro talune forme ed espressioni di razzismo e xenofobia mediante il diritto penale*, ***https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:32008F0913&from=EL***.

CoE (2015). *Piano d'Azione per la Lotta contro l'Estremismo Violento e la Radicalizzazione 2015-2017*, ***https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680c5c3576*** (in inglese).

CoE -Consiglio d'Europa (2016). "No Hate Speech Movement. Youth Campaign for Human Rights Online: Report of the 11th meeting." DDCP-YD/EDT (2016) 185.

Cohen-Almagor R. (2014). "Countering Hate on the Internet." *Annual Review of Law and Ethics*, 22, 431-443.

Consiglio d'Europa (2006). *Protocollo addizionale alla Convenzione sulla criminalità informatica, relativo all'incriminazione di atti di natura razzista e xenofobica commessi a mezzo di sistemi informatici*, ***https://www.coe.int/it/web/conventions/full-list/-/conventions/treaty/189***.

Corazza M., Menini S., Cabrio E., Tonelli S. e Villata S. (2019). "Cross-Platform Evaluation for Italian Hate Speech Detection", *CLiC-it 2019 - 6th Annual Conference of the Italian Association for Computational Linguistics*, novembre 2019, Bari.

Correa D., Silva L., Mondal M., Benevenuto F. and Gummadi, K. (2015). *The Many Shades of Anonymity: Characterizing Anonymous Social Media Content*. In 9th International AAAI Conference on Web and Social Media.

Countering illegal *hate speech* online 5th evaluation of the Code of Conduct, Factsheety, June 2020, ***https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf.***

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*. 11th International AAAI Conference on Web and Social Media.

de Latour A., Perger N., Slaj R., Tocchi C. e Viejo Otero P. (2017). "WE CAN! Taking Action against Hate Speech through Counter and Alternative Narratives. Revised edition 2017." Del Felice C. e Ettema M (eds.) Consiglio d'Europa.

EC (2016). Codice di Condotta per Contrastare l'Illecito Incitamento all'Odio Online.

EC (2018). *Third Evaluation of the Code of Conduit on Countering Illegal Online Hate Speech*. ***http://europa.eu/rapid/press-release_MEMO-18-262_en.htm***.

EC (2020). Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle regioni. *Un'Unione dell'uguaglianza: il piano d'azione dell'UE contro il razzismo 2020-2025*, com (2020) 565 del 18-09-2020.

EC (2020). Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle Regioni (COM(2020) 67) final, *Plasmare il futuro digitale dell'Europa,* ***https://ec.europa.eu/transparency/regdoc/rep/1/2020/IT/COM-2020-67-F1-IT-MAIN-PART-1.PDF.***

EC (2020). Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle Regioni (COM(2020) 565), *Un'Unione dell'uguaglianza: il piano d'azione dell'UE contro il razzismo 2020-2025,* ***https://ec.europa.eu/info/sites/info/files/a_union_of_equality_eu_action_plan_against_racism_2020_-2025_it.pdf.***

EC (2020). *Countering illegal hate speech* online *5th evaluation of the Code of Conduct*, Factsheety, June 2020.

ECRI (2000). *Raccomandazione di politica generale n° 6 dell'ECRI, La lotta contro la diffusione di materiale razzista, xenofobo e antisemita via internet*, adottata il 15 dicembre 2000.

ECRI (2015). *Raccomandazione di politica generale n. 15 sulla lotta all'incitamento all'odio*, adottata l'8 dicembre 2015.

ECRI (2016). *General Policy Recommendation No. 11 on combating racism and racial discrimination in policing*, para. 10 and explanations.

ENAR (2016). *Racism and discrimination in the context of migration in Europe.* ENAR Shadow Report, ***http://www.enar-eu.org/IMG/pdf/shadowreport_2015x2016_long_low_.***

Ernst J., Scmitt J., Rieger D., Beier A., Vorderer P., Bente G. e Roth H (2017). "Hate Beneath the Counter Speech? A Qualitative Analysis of User Comments on Yuotube to Counter Speech Videos." *Journal for Deradicalization*, 10.

Eurobarometer (2016). *Media Pluralism and Demoracy.* ***https://ec.europa.eu/digital-single-market/en/news/media-pluralism-and-democracy-special-eurobarometer-452***.

Gagliardone I., Gal D., Alves T. and Martinez G. (2015). *Countering Online Hate Speech*. UNESCO Series on Internet Freedom.

Gerstenfeld P.B. (2017). *Hate Crimes. Causes, Controls and Controversies* (4th Edition). SAGE.

Hall N. (2013). *Hate Crimes: 2nd ed*. Routledge.

Heins M. (2014). *The brave new world of social media censorship,* ***http://cdn.harvardlawreview.org/wp-content/uploads/2014/06/vol127_Heins.pdf.***

Himelboim I., McCreery S., and Smith M. (2013). "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter" *Journal of Computer-Mediated Communication,* 18, 154-174.

Hine G. E., Onaolapo, J., De Cristofaro E., Kourtellis N., Leontiadis I., Samaras R., Stringhini G. e Blackburn J. (2016). *A longitudinal measurement study of 4chan's Politically Incorrect Forum and its effect on the Web (version 3)*, ***https://arxiv.org/pdf/1610.03452v3.pdf.***

Kuklinski J.H., Quirk P.J., Jerit J., Schwieder D. e Rich R.F. (2000). "Misinformation and the Currency of Democratic Citizenship", *Journal of Politics* Volume 62, Issue 3, pp. 790-816, ***https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-3816.00033***.

Lewandowsky S., Ecker U., Seifert C., Schwarz N. e Cook J. (2012). "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest*, 13(3), 106-131.

Lo Giudice H. (2016). "Ammesso il sequestro preventivo di un blog in caso di diffamazione. Non è «stampa»", *Giurisprudenza Penale Web*, 5, 10 maggio 2016, ***https://www.giurisprudenzapenale.com/2016/05/10/ammesso-sequestro-preventivo-un-blog-caso-diffamazione-non-stampa/.***

Marone V. (2015) "Online Humor as a Community-Building Cushioning Glue." *The European Journal of Humor Research*, 3(1), 61-83.

McGonagle T. (2013). *The Council of Europe Against Online Hate Speech: Conundrums and Challenges.* (MCM; No. 2013(005)). Belgrade: Republic of Serbia, Ministry of Culture and Information.

McNamee L., Peterson B. and Pena J. (2010). "A Call to Educate, Participate, Invoke and Indict: Understanding the Communication of Online Hate Groups." *Communication Monographs*, 77(2), 257-280.

Miglio M. (2017). "I gestori di un sito internet rispondono penalmente per i commenti offensivi pubblicati dagli utenti". *Giurisprudenza Penale Web*, 1 - ISSN 2499-846X, ***http://www.giurisprudenzapenale.com/2017/01/03/gestori-un-sito-internet-rispondono-penalmente-commenti-offensivi-pubblicati-dagli-utenti/.***

Mitts T. (2018). "From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West". *American Political Science Review*, forthcoming.

Müller K. and Schwarz C. (2018a). "Fanning the Flames of Hate: Social Media and Hate Crime." *CAGE* online *working paper* 373.

Müller K. and Schwarz C. (2018b). *Making America Hate Again? Twitter and Hate Crime under Trump*. SSRN 3149103.

Musto C., Semeraro G., de Gemmis M. and Lops P. (2016). *Modeling Community Behavior through Semantic Analysis of Social Data: The Italian Hate Map Experience*. Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization.

Nyhan B. e Reifler J, (2015). "Does Correcting Myths about the Flu Vaccine Work? An Experimental Evaluation of the Effects of Corrective Information." *Vaccine*, 33(3), 459-464.

OHCHR (1948). *Convenzione per la Prevenzione e la Repressione del Crimine di Genocidio*, ***https://www.ohchr.org/EN/ProfessionalInterest/Pages/CrimeOfGenocide.aspx*** (in inglese).

OHCHR (1948). *Dichiarazione Universale dei Diritti dell'Uomo*, ***https://www.ohchr.org/en/udhr/pages/Language.aspx?LangID=itn***.

OHCHR (1965). *Convenzione Internazionale per l'Eliminazione di ogni Forma di Razzismo*, ***https://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx*** (in inglese).

OHCHR (1966). *Convenzione Internazionale sui Diritti Politici e Civili*, ***https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx*** (in inglese).

OHCHR (2012). *Piano di azione Rabat*, ***https://www.ohchr.org/EN/Issues/FreedomOpinion/Articles19-20/Pages/Index.aspx*** (in inglese).

ONU (1966). *Legge di autorizzazione alla ratifica della Convenzione internazionale sull'eliminazione di tutte le forme di discriminazione razziale*, aperta alla firma a New York il 7 marzo 1966.

ONU (2016). *Piano d'azione per prevenire l'estremismo violento*, ***https://www.un.org/en/ga/search/view_doc.asp?symbol=A/70/674*** (in inglese).

ONU (2017). *Programma Quadro Internazionale per Contrastare le Narrazioni Terroristiche*, ***https://www.un.org/en/ga/search/view_doc.asp?symbol=S/2017/375&referer=http://www.un.org/en/documents/index.html&Lang=E*** (in inglese).

OSCE (2010). *Freedom of Expression on the Internet*, ***https://www.osce.org/files/f/documents/c/9/105522.pdf***.

OSCE (2017). *Freedom of the Media. Freedom of Expression. Free flow of information*, ***https://www.osce.org/representative-on-freedom-of-media/354081***.

OSCE-ODIHR (2010). Report of OSCE-ODIHR activities on hate on the internet, ODIHR.GAL/77/10, 27 October 2010 ***https://www.osce.org/files/f/documents/8/7/73461.pdf***.

Perry B. (2001). *In the Name of Hate. Understanding Hate Crimes*. Routledge.

Perry B. and Olsson P. (2009). "Cyberhate: the Globalization of Hate." *Information and Communications Technology Law*, 18 (2), 185-199.

Pew Research Center (2017). *Online Harassment 2017*.

RAN - Centre of Excellence (2017a) "Dissemination Strategies and Building Online Multi-platform Networks". RAN C&N meeting Ex Post Paper.

RAN - Centre of Excellence (2017b) "How to Measure the Impact of Your Online Counter or Alternative narrative Campaign". RAN C&N meeting Ex Post Paper.

Renzi (2019). *Labor - Il lavoro nel diritto*, 22 febbraio 2019.

Repubblica Italiana (1993). Decreto-legge 26 aprile 1993, n. 122, coordinato con la legge di conversione 25 giugno 1993, n. 205, recante: "Misure urgenti in materia di discriminazione razziale, etnica e religiosa"("Legge Mancino"), ***http://presidenza.governo.it/USRI/confessioni/norme/dl_122_1993.pdf***.

Repubblica Italiana (2003). Decreto Legislativo 9 luglio 2003, n. 215 "Attuazione della direttiva 2000/43/CE per la parità di trattamento tra le persone indipendentemente dalla razza e dall'origine etnica", ***https://www.camera.it/parlam/leggi/deleghe/03215dl.htm***.

Schieb C. e Preuss M. (2016). *Governing Hate Speech by means of Counter Speech on Facebook*, Proceedings of the ICA Annual Conference.

Silva L., Mondal M., Corra D., Benvenuto F. and Weber I. (2016). "Analyzing the Targets in Online Social Media." *AAAI ICWSM*, 2016.

Stea G. (2016). "La responsabilità penale dell'internet provider". *Giurisprudenza Penale*, ***https://www.giurisprudenzapenale.com/wp-content/uploads/2016/11/stea_provider_gp_2016_11.pdf***.

Suler J. (2004). "The Online Disinhibition Effect." *CyberPsychology and Behavior* 7 (3), 321-326.

Sunstein C. (2017). *#Republic: Divided Democracies in the Age of Social Media*. Princeton University Press.

The Guardian (2019). "Here, here: the Swedish online love army who take on the trolls." ***https://www.theguardian.com/world/2019/jan/15/the-swedish-online-love-army-who-battle-below-the-line-comments***.

Titley G., Keen E. e Földi L. (2017). *Starting Points for Combating Hate Online*, Consiglio d'Europa.

Tuck H. & Silverman, T. (2016). *The counter-narrative handbook*. ***https://www.strate-gicdialogue.org/wp-content/uploads/2016/06/Counter-narrative-Handbook_1.pdf***.

UE (2012). *Direttiva 2012/29/UE del Parlamento europeo e del Consiglio, del 25 ottobre 2012, che istituisce norme minime in materia di diritti, assistenza e protezione delle vittime di reato e che sostituisce la decisione quadro 2001/220/GAI*, ***https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32012L0029&from=IT***.

UE (2018). *Direttiva (UE) 2018/1808 del Parlamento europeo e del Consiglio, del 14 novembre 2018, recante modifica della direttiva 2010/13/UE, relativa al coordinamento di determinate disposizioni legislative, regolamentari e amministrative degli Stati membri concernenti la fornitura di servizi di media audiovisivi (direttiva sui servizi di media audiovisivi)*, ***https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32018L1808&from=IT***.

Wright L., Ruths D., Dillon K., Saleem H. e Benesh S. (2017). *Vectors for Counterspeech on Twitter*, Proceedings of the First Workshop on Abusive Language Online, pp. 57-62.