



C.O.N.T.R.O.

"COunter Narratives Against Racism Online"

MAPPING REPORT

Mappatura delle principali metodologie italiane ed europee per l'individuazione e l'analisi degli "hate speech" con riferimento all'ambito della discriminazione razziale

WP2 - Methodology for hate speech classification and analysis

IRS- Istituto per la Ricerca Sociale

Giugno 2019

ISTITUTO
PER LA
RICERCA
SOCIALE **irs**



The content of this report represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains

Indice

1	Premessa.....	4
2	Metodologia di analisi e articolazione della mappatura	4
3	Analisi della letteratura: principali approcci e dibattito in corso	6
3.1	Premessa.....	6
3.2	Evoluzione del dibattito e definizioni giuridiche dell' <i>hate speech</i>	7
3.2.1	<i>Inquadramento giuridico e principali politiche per il contrasto dell'hate speech a livello istituzionale</i>	<i>9</i>
3.2.2	<i>il ruolo della società civile e dei social network nel dibattito legislativo e nell'implementazione delle politiche di contrasto all'hate speech online.....</i>	<i>14</i>
3.3	Evoluzione degli approcci e metodologie di analisi degli <i>hate speech</i>	16
4	Metodologie di rilevazione e analisi degli “<i>hate speech</i>” in ambito razziale. Esperienze italiane ed internazionali a confronto	21
4.1	Metodologia di analisi adottata	21
4.2	Analisi sinottica e confronto tra le metodologie individuate.....	23
	Allegato 1- Presentazione delle esperienze e metodologie di analisi degli <i>hate speech</i>.....	35
	eMORE.....	35
	Hatebase	38
	Hatelab online Hate Speech Dashboard	41
	HATEMER.....	44
	Mandola.....	46
	Mappa dell'Intolleranza	48
	Analyzing the Targets of Hate in Online Social Media (studio).....	50
	Hate speech e dark web - 4chan	52
	React	55
	Bibliografia.....	57

1 Premessa

Il presente lavoro, si inserisce nelle attività promosse del progetto CONTROLLO - "*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.

La mappatura che segue intende, in particolare, **fornire un quadro d'insieme** sull'evoluzione dei principali approcci e metodologie di analisi degli *hate speech* correlati alla discriminazione razziale e **sugli strumenti/metodologie ad oggi utilizzati** a livello nazionale ed europeo e internazionale, al fine di **avviare una riflessione finalizzata a derivare in una fase successiva delle proposte per l'aggiornamento della metodologia** attualmente usata dall'Ufficio nazionale contro la discriminazione razziale (UNAR) per monitorare e analizzare gli *hate speech* prodotti sui social network.

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 e analizzare tale fenomeno.

2 Metodologia di analisi e articolazione della mappatura

La presente mappatura si focalizza sulle esperienze e metodologie di rilevazione e analisi degli *hate speech* trasmessi attraverso i social network con riferimento agli aspetti discriminatori legati a motivi razziali¹.

La ricognizione si è focalizzata su approcci, dibattito ed esperienze centrati solo su metodologie di analisi applicabili ai messaggi testuali veicolati dai social network e non dalla molteplicità dei Media più tradizionali (cd. Mass Media), sul presupposto che i social network presentano peculiarità proprie di funzionamento e interazione tra i loro utilizzatori tali da richiedere metodologie di analisi specifiche.

La mappatura delle esperienze e delle metodologie è stata effettuata sulla base di un sistematico processo di ricerca documentale che si è articolato in:

- una prima fase esplorativa per acquisire una conoscenza più approfondita delle problematiche pertinenti lo studio così da ottenere una migliore comprensione delle informazioni disponibili;

¹ Si precisa, tuttavia, come mostrano le esperienze individuate e descritte, che la maggior parte delle metodologie identificate prendono in considerazione l'ambito della discriminazione razziale unitamente ad ambiti di discriminazione differenti (es. disabilità, orientamento sessuale, genere, età ecc..) in un'ottica cioè multidimensionale.

- una fase di ricerca che ha portato all'identificazione di studi potenzialmente rilevanti, attraverso l'utilizzo di una vasta gamma di fonti elettroniche (ad esempio, database scientifici quali Scopus, web-of-science; google scholar) e cartacee e utilizzo di specifiche parole chiave per la ricerca: *online hate; machine learning; datanalysis; hatespeech; content analysis; semantic analysis; sentiment analysis; countering*;
- una fase di studio e analisi dei materiali individuati al fine di valutare la qualità dello studio ed individuare specifiche metodologie da includere nella mappatura;
- una fase descrittiva delle metodologie individuate, criteri e approcci utilizzati e principali risultati prodotti.

La mappatura che segue si articola quindi in una prima parte introduttiva di analisi della letteratura, dei principali approcci al tema e del dibattito in corso (cap. 3). Questa prima parte ha prestato particolare attenzione alla definizione e all'inquadramento giuridico, a livello internazionale ed europeo, degli *hate speech* e al ruolo e responsabilità dei social media nel contrastare gli stereotipi e i discorsi di incitamento all'odio sul web. A seguire, la seconda parte di analisi della letteratura si focalizza sull'evoluzione dei principali approcci e tecniche di analisi degli *hate speech*, di modo da fornire un quadro d'insieme prima della presentazione delle principali esperienze e metodologie di rilevazione e analisi degli *hate speech* che saranno oggetto del capitolo successivo (cap.4).

L'analisi sia della letteratura che delle esperienze è stata condotta principalmente a livello nazionale ed europeo, tenendo conto anche del livello internazionale, laddove sono stati individuati approcci e tecniche interessanti.

Le esperienze/metodologie individuate sono presentate in schede sintetiche articolate in modo da sottolineare le principali caratteristiche delle tecniche di analisi descritte e i principali risultati ottenuti (Allegato 1). L'articolazione per schede della mappatura intende consentire una più facile fruibilità dei contenuti da parte della committenza in un'ottica operativa in cui la mappatura, come già detto, costituisce uno strumento volto a fornire non solo un quadro conoscitivo dell'esistente ma anche primi spunti di riflessione in vista dell'aggiornamento e perfezionamento della metodologia di analisi ad oggi utilizzata dall'Osservatorio UNAR.

3 Analisi della letteratura: principali approcci e dibattito in corso

3.1 Premessa

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 le minoranze, gli indigenti, i sottogruppi sociali stigmatizzati. L'aumento dell'interesse verso questi fenomeni è sollecitato da dati attestanti una propagazione estremamente rapida di questa tipologia di eventi, unitamente all'aumento della 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'*online hate speech*, 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 del 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'ECRI² sottolinea l'aumento senza precedenti dei discorsi d'odio a sfondo razziale (ECRI, 2016), confermato anche dallo *Shadow Report 2016* dell'ENAR³. I dati mostrano anche grandi quote di discorsi d'odio online prodotte da persone estranee a qualsiasi gruppo di odio organizzato (Hall, 2013).

Prima di focalizzarsi sugli approcci individuabili in letteratura con riferimento alle metodologie di rilevazione e analisi degli *hate speech* che costituisce il focus dell'analisi, (*infra*, par. 3.2), il capitolo presenterà una parte più generale sull'evoluzione del dibattito in corso sull'approccio normativo al contrasto del fenomeno e al ruolo dei social media. In particolare ci si soffermerà su:

- le **principali definizioni normative di "hate speech"** adottate sia a livello istituzionale che ad opera dei principali Social Media e dibattito in corso. La definizione di *hate speech* non è infatti univoca e sovente la linea di demarcazione tra ciò che viene considerato e ciò che non

² ECRI European Commission against Racism and Intolerance.

³ ENAR European Network Against Racism.

viene considerato incitamento all'odio (*hate speech*) è molto labile. Ad esempio, è generalmente ammesso criticare una nazione, ma non è accettabile generalizzare in modo offensivo sulle persone di una determinata nazionalità o etnia;

- **il dibattito su ruolo e responsabilità dei social media nel contrastare gli stereotipi e i discorsi d'odio (*hate speech*) sul web.** Tale aspetto è oggi particolarmente rilevante considerato l'intensificarsi degli *hate speech* con l'aumento dei fenomeni migratori verso l'Europa;
- **la relazione tra libertà di espressione e rispetto del principio di non discriminazione e di uguaglianza,** dal momento che queste due sfere del diritto entrano spesso in conflitto quando si parla di *hate speech* e che tale aspetto è reso ulteriormente complicato dal fatto che taluni contenuti possano costituire una violazione di legge in un determinato Paese ma non in altri Paesi.

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

I trend sulla crescita del fenomeno hanno determinato in questi ultimi anni anche un incremento delle attività di ricerca sui temi dell'*online hate speech* 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 del *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 solo verbalmente**, 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 sia solo verbale, 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).

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 (Himmelboin 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 (Padiglione, 2013 p.206; Perry, 2001);
4. *L'hate speech* online è caratterizzato da una percezione di anonimato, o **de-individuazione**, (Lopez e Lopez, 2017; 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: Facebook⁴ ha rimosso circa 7,9 milioni di contenuti relativi ai discorsi di odio e YouTube⁵ 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, ecc.), 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).

Accanto ad una conoscenza sempre più accurata del fenomeno, si sta inoltre sviluppando un ricco dibattito sull'**inquadramento giuridico** del *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

⁴ <https://transparency.facebook.com/community-standards-enforcement#hate-speech>.

⁵ <https://transparencyreport.google.com/youtube-policy/removals?hl=en>.

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).

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 *hate speech* (McGonagle, 2013).

3.2.1 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 del *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 (CdU, 2016).

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** 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 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 la **commissione ECRI** che si occupa operativamente di monitorare e approfondire le tendenze e caratteristiche dell'*hate speech*, favorendo il coinvolgimento attivo della società civile.

Un recente provvedimento del Consiglio d'Europa sul tema dell'*hate speech* è costituito dal **Protocollo addizionale alla Convenzione sulla Criminalità Informatica**, 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 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*.

Alla Decisione Quadro, si è aggiunto il recente **Codice di Condotta per Contrastare l'Illecito Incitamento all'Odio Online** (EC, 2016) 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.

Infine, con riferimento al contesto italiano, la principale legge in materia è la L. **205/1993** (cd "Legge Mancino") che istituisce il reato di "diffusione di idee basate sulla superiorità razziale o sull'odio razziale o etnico, o volte ad istigare o commettere atti di discriminazione razziale, etnica, nazionale

o religiosa". Tale legge punisce quindi anche chi compie istigazione alla discriminazione, aspetto questo di indubbia rilevanza.

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 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 sovranazionali, ai media, all'ordine e il 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 3.1 che segue riporta i principali ambiti a cui sono riconducibili tali raccomandazioni⁶.

Box 3.1 - 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 sovranazionale, 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 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;

⁶ 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.

- 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.

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'hatespeech"** 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 sovranazionale 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).

3.2.2 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. Tra le organizzazioni facenti parte del Tavolo 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 islamiche⁷.

⁷ 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>.

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 particolare 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 a 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 persone⁸.

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'online *hate speech*. 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 emanato 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*⁹", 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*¹⁰".

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*¹¹". 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

⁸ 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.

⁹ Facebook Community Standards.

¹⁰ <https://support.google.com/youtube/answer/2801939?hl=en-GB>.

¹¹ <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.

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 in base alla loro razza o origine etnica, religione, disabilità, sesso, età, stato di veterano, o orientamento sessuale/identità di genere."¹²

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, 2013).

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

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).

Di seguito si presenta, quindi, 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 e 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

¹² <https://www.google.com/+policy/content.html>.

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*)¹³, 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*¹⁴ 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 e 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.

¹³ 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

¹⁴ 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.

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 (NLP)** identificano ed estraggono i contenuti di odio online da una piattaforma. Il complesso degli *hate speech* estratti può contenere tuttavia una 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* abbinata 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 negativo¹⁵ 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., 2018), 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

¹⁵ http://users.humboldt.edu/mstephens/hate/hate_map.html.

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/metodologie¹⁶ 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**¹⁷ per estendere 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 di questi network di odio e la conseguente propagazione e persistenza dei messaggi di odio online, così come la natura umana o virtuale (*bot*)¹⁸ dei partecipanti alla rete di relazioni (Himmelboim et al., 2013).

Un'altra rilevante tecnica di analisi è la **geolocalizzazione**¹⁹ 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 rinvencono in diversi progetti realizzati in questi ultimi anni²⁰. L'output cartografico utilizza solo il subset di *online hate speech* geo-referenziato, che costituisce una parte minoritaria dei flussi di *online hate speech*. Tuttavia, pur agendo su una parte limitata del

¹⁶ *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*).

¹⁷ 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*.

¹⁸ 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.

¹⁹ Tecnica che consente l'identificazione della posizione geografica di un soggetto autore dell'*hate speech*.

²⁰ 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/>.

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**²¹. 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., 2017).

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 del *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

²¹ 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](#).

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 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*.

Nel prossimo capitolo verranno presentate in dettaglio le principali esperienze/metodologie di monitoraggio degli *hate speech* online individuate.

4 Metodologie di rilevazione e analisi degli “*hate speech*” in ambito razziale. Esperienze italiane ed internazionali a confronto

4.1 Metodologia di analisi adottata

Nel presente capitolo sono presentate le principali esperienze di osservazione che adottano metodologie di individuazione, analisi e classificazione degli “*hate speech*” con riferimento all'ambito della discriminazione razziale, con particolare riferimento a 3 specifici aspetti che possono essere oggetto degli *hate speech* e che costituiscono il focus della nostra analisi (**colore della pelle, etnia, razzismo**). La ricognizione delle esperienze ha tenuto conto sia del livello nazionale che di quello europeo e internazionale.

La descrizione delle esperienze/progetti individuati e presentati nell'allegato 1, è stata effettuata, sulla base delle informazioni e della documentazione disponibile.

Le schede di presentazione e analisi delle esperienze evidenziano le metodologie adottate e sono state strutturate in modo da individuare le principali caratteristiche per quanto concerne:

- il **soggetto attuatore**, ossia che ha realizzato la metodologia e altri possibili partner coinvolti nella sua ideazione o applicazione, con dettaglio, laddove possibile, del ruolo svolto;
- il **soggetto finanziatore**, con dettaglio nel caso di finanziatore pubblico del Programma di finanziamento grazie al quale la metodologia è stata realizzata;
- l'**area geografica di applicazione** della metodologia;
- il **periodo di implementazione** della metodologia;
- i **social media di riferimento** per l'applicazione della metodologia (ad. esempio, facebook, twitter, whisper, linkedin);
- **ambito/oggetto/ target** degli *hate speech*, con particolare attenzione a indicare se gli *hate speech* considerati dalla metodologia facciano riferimento ai 3 specifici aspetti già citati e che costituiscono il focus della nostra analisi (**colore della pelle, etnia, razzismo**) o comunque se la metodologia presentata sia multi-target ossia ricomprenda anche altri ambiti di discriminazione e molteplici soggetti "bersaglio", quali soprattutto migranti, donne, persone LGBT, Rom, disabili e comunità islamiche. Sono queste, infatti, le categorie contro cui si rivolgono soprattutto i messaggi di odio in rete. Oltretutto, la definizione di metodologie multi-target risponde anche all'esigenza di identificare e analizzare tipologie di *hate speech* che sempre più frequentemente fanno riferimento a discriminazioni multiple.
- **gli approcci** adottati dalla metodologia di identificazione e analisi: approccio olistico basato su modelli teorici che provengono dalla sociologia, psicologia, linguistica, statistica ecc. articolato ulteriormente in **qualitativo**, basato cioè prettamente su interviste/osservazioni effettuate dall'azione umana e che si focalizza in profondità su campioni limitati di *hate speech*; oppure **quantitativo** che fa ricorso prettamente a sistemi di *machine learning* (algoritmi) per l'identificazione e analisi degli *hate speech* ovvero **quali-quantitativo** basato cioè su un mix di azione automatica/computerizzata integrata da analisi qualitative manuali;
- **articolazione della metodologia di identificazione e analisi degli *hate speech* online** (estrazione dati e creazione base informativa/database; attività di analisi dati/contenuti di odio online; produzione output specifici, ad es. *dashboard*, App, mappe, workshop, linee guida, raccomandazioni)
- **tecniche e modalità specifiche di identificazione e analisi degli *hate speech*** con particolare attenzione alle modalità di individuazione delle parole chiave per l'estrazione dei dati (modalità bottom up, mono-lingua o multilingue ecc.), alle modalità di costituzione della base informativa

su cui effettuare l'analisi (es. ricorso a piattaforme condivise con utenti) e all'utilizzo di specifiche tecniche di *content analysis* per l'individuazione corretta dei contenuti di odio quali:

- ✓ *sentiment analysis*, ossia una tecnica di analisi manuale o computerizzata del tono utilizzato all'interno di testi presenti in rete su specifici temi/soggetti. Al tono è attribuito una specifica polarità che indica se l'orientamento sia positivo, negativo o neutro.
- ✓ *semantic tagging*, cioè una tecnica di analisi, manuale o computerizzata, che permette di analizzare le parole polisemiche e porre attenzione al contesto di riferimento per cui alcuni termini che sembrano di *hate* sono termini che vengono usati nel linguaggio corrente non necessariamente in senso negativo;
- ✓ *lexical parsing* cioè una tecnica, manuale o computerizzata, che analizza un flusso continuo di dati in ingresso in modo da determinare la sua struttura grazie ad una data grammatica formale. Consente di assegnare un valore conoscitivo ad una espressione grammaticale. Un parser è un programma che esegue questo compito.

L'analisi si focalizzerà anche su altre tecniche di analisi quali la geolocalizzazione o la *network analysis* e ulteriori tecniche sviluppate dalle singole esperienze/metodologie (ad es. tecniche di analisi dei post presenti nel dark web oppure lo sviluppo di algoritmi predittivi di eventi di odio nel mondo reale etc.)

- **principali risultati conseguiti;**
- **eventuali criticità incontrate e modalità di soluzione.**

4.2 Analisi sinottica e confronto tra le metodologie individuate

La ricognizione 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, *Dashboard*, 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 4.1 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	Geo-analisi	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 visualizzazione 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 methodology (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	Si	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 visualizzazione consultazione database hate speech/ risultati analisi	Sviluppo algoritmi predittivi di eventi di odio nel mondo reale sulla base dei trend negli hate speech online
3	HATELAB dashboard (University of Cardiff- UK)	Target multidimensionale che include: -colore della pelle	Tutti i social networks	Surface web	Globale	Di nuova elaborazione Parole chiave multi-lingua	Sentiment analysis • Algoritmo	Si	Si	Dashboard visualizzazione consultazione database per e	Analisi relazione tra eventi significativi/potenzialmente scatenanti (fatti

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	Geo-analisi	Network analysis	Strumenti di diffusione, coinvolgimento attivo, accountability	Azioni complementari all'attività di identificazione degli hate speech
		etnia - razzismo -altro (religione, disabilità, orientamento sessuale, genere)				Tecnica BOW di identificazione parole odio basata su algoritmi Identificazione BOW associate a possibili eventi	Lexical parsing • Algoritmo Supervisione umana esternalizzata con piattaforme di crowdfunding			<i>hate speech</i> da parte dell'utenza Mappe degli <i>hate speech</i> visualizzabili Non previsti piattaforme/strumenti partecipativi per il monitoraggio	sociopolitici, commemorazioni, eventi di cronaca) e <i>hate speech</i> 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	Si	Si	Dashboard per visualizzazione e consultazione database <i>hate speech</i> 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 <i>hate speech</i> online; 2. Analisi degli <i>influencer</i> e dei nuovi profili online legati all'islamofobia; 3. Ideazione algoritmo per la produzione di contronarrativa
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 individuate chiave dal Vocabolario Hatebase Parole chiave multi-lingua (paesi coinvolti) Tecnica BOW di identificazione parole odio basata su algoritmi	Sentiment analysis • Algoritmo • Lexical parsing Algoritmo	Si	No	Dashboard per visualizzazione e consultazione dei dati da parte dell'utenza (modalità user-friendly) Mappe degli <i>hate speech</i> visualizzabili Tabelle dati esportabili e grafici	Analisi normativa e classificazione degli <i>hate speech</i> online in legali e illegali

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	Geo-analisi	Network analysis	Strumenti di diffusione, coinvolgimento attivo, accountability	Azioni complementari all'attività di identificazione degli hate speech
										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)	
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	Si	No	Mappe degli hate speech visualizzabili dall'utenza (modalità user-friendly) 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	Si	Si	No	Analisi differenze tra hate speech online su piattaforme anonime e piattaforme non anonime

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	Geo-analisi	Network analysis	Strumenti di diffusione, coinvolgimento attivo, accountability	Azioni complementari all'attività di identificazione degli hate speech
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	Si	Si	No	1. Analisi dei raid (attacchi a profili specifici di persone nel surface web organizzati nel dark web) 2. Studio di tecniche di analisi basate su algoritmi per prevedere tali raid
	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	Si	Si	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 contronarrazione; 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 (Hatometer e React) si focalizzano sull'**islamofobia**, 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, you tube e 4 chan, 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 disposizione 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.

N	Esperienze /metodologie e soggetto attuatore ▼	Fase 1		Fase 2		
		Uso di parole d'odio (BoW) per l'estrazione di dati da internet attraverso algoritmo		Discorsi d'odio individuati attraverso algoritmo a partire dai dati estratti attraverso ricerca per parole d'odio		
		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	REACT ²²	-	-	-	-	-

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

²² Non sono disponibili informazioni di dettaglio pubbliche che consentano di approfondire questi aspetti metodologici.

(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 ecc.)

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 contronarrativa, 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 contronarrativa 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.

Allegato 1- Presentazione delle esperienze e metodologie di analisi degli *hate speech*

Nome esperienza /metodologia	eMORE	
Sito Web	emoreproject.eu	
Soggetto attuatore/coordinatore	RISSC - Centro di ricerca su sicurezza e criminalità (IT) Associazione privata <i>senza scopo di lucro che si occupa di sicurezza e criminalità</i> . Le principali attività di RISSC sono l'analisi dei fenomeni criminali, sociali e criminogeni, l'elaborazione di strategie di prevenzione dei rischi e riduzione dei danni, l'assistenza tecnica e la formazione a favore di enti pubblici e organizzazioni private.	
Referente	Nome	Mara Mignone
	Email	info@rissc.it
Altri soggetti coinvolti e ruolo	La realizzazione della metodologia è avvenuta nell'ambito di un progetto europeo che ha visto il coinvolgimento dei seguenti partner di progetto: CEJI (BE), CLR (RO), Colledge for Public Administration and Administration for Justice (DE), IDOS Research Center (IT); Associação ILGA (PT); LAND KISA (CY); MPG (BE); North Restern Migration Forum (UK); Peace Institute (SL); SOS Malta (MT)	
Soggetto finanziatore	Commissione Europea, nell'ambito del Programma REC - RIGHTS, EQUALITY AND CITIZENSHIP PROGRAMME (2014-2020)	
Area geografica di applicazione della metodologia	Europa: Italia, Belgio, Romania, Germania, Portogallo, Cipro, Regno Unito, Slovenia, Malta Copertura nazionale di ogni ambito geografico di applicazione	
Periodo di implementazione	Da gennaio 2016 a Maggio 2018	
Social media di riferimento per l'applicazione della metodologia	Le piattaforme web su cui il progetto eMORE si focalizza per l'individuazione dell'hate speech online sono rappresentate dai social network, con particolare riferimento a Facebook e Twitter, unitamente a siti web, identificati per singolo paese, che potrebbero contenere online hate speech	
Obiettivi e articolazione della metodologia	<p>Il progetto eMORE intende contribuire allo sviluppo, alla sperimentazione e al trasferimento di un modello di conoscenza europeo in materia di <i>hate speech</i> online. Con il coinvolgimento di 9 partner provenienti da 9 diversi paesi europei (IT, BE, RO, UK,DE, PT,CY,SL,MT), eMORE ha inteso svolgere un'azione di monitoraggio che mira a consentire un'analisi comparativa e a sostenere un'azione di contrasto armonizzata a livello europeo.</p> <p>eMORE integra la finalità di analisi data dallo sviluppo di algoritmi informatici per l'individuazione dell'hate speech online alla finalità di mobilitazione e engagement attivo della società civile attraverso l'integrazione di diversi strumenti ICT.</p> <p>Si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; <p>Il progetto ha previsto in particolare la creazione di una metodologia condivisa per l'individuazione dell'<i>hate speech</i> online sui social network. Più in dettaglio:</p> <ul style="list-style-type: none"> - creazione di un web crawler per estrarre da Facebook e Twitter gli hate speech; 	



	<ul style="list-style-type: none"> - creazione di una APP attraverso cui è possibile per chiunque segnalare episodi di hate speech online a cui ha assistito - creazione di un database integrato in cui confluiscono i dati raccolti dal web crawler (software) e dalla APP • Fase 2 - attività di analisi dati/contenuti di odio online; <ul style="list-style-type: none"> - Realizzazione di <i>sentiment analysis</i> manuale • Fase 3 - produzione output specifici: <ul style="list-style-type: none"> - realizzazione di una piattaforma di consultazione del database integrato - creazione di un'interfaccia (dashboard) user-friendly per rendere possibile la consultazione del database da parte di platee eterogenee di utilizzatori (ONG, policy-maker, giornalisti, ricercatori)
<p>Ambito/oggetto/ target degli hate speech analizzati</p>	<p><input type="radio"/> Colore della pelle</p>
	<p><input type="radio"/> Etnia</p>
	<p><input type="radio"/> Razzismo</p>
	<p><input type="radio"/> Altro: religione, disabilità, sessismo, orientamento sessuale</p>
	<p>Il database può essere interrogato per ogni target; l'APP, analogamente, consente di inserire eventi (indicazione di hate speech) specificando il target specifico.</p>
<p>Descrizione della metodologia e tecniche di analisi</p>	<p>eMORE si contraddistingue per un ambito geografico di applicazione multi-paese: coinvolgendo 9 differenti contesti europei, contribuisce a creare database armonizzati a livello europeo, alla realizzazione di analisi trasversali e alla creazione di strategie di contrasto condivise.</p> <p>L'identificazione delle parole chiave (BOW) è realizzata direttamente dai team di ricerca in ogni paese e non esteso alla società civile. Ed è proprio nell'identificazione delle BoW che eMORE ha sviluppato una delle sue caratteristiche distintive: la creazione di BoW avviene attraverso il coinvolgimento di 9 team multidisciplinari (uno per ogni paese partner) ognuno dei quali focalizzato sul contesto linguistico del proprio paese. eMORE, quindi, non utilizza BoW provenienti da altri soggetti (vedi scheda HATEBASE), ma le identifica autonomamente. Più specificamente, vengono identificati termini che, pur essendo in diverse lingue, sono identificabili come parole d'odio condivise (es. "negro", "nigger" ...) e altre che sono idiosincratice solo in alcuni contesti. I team multidisciplinari segnalano, anche, pagine e profili social sensibili presenti nei loro paesi.</p> <p>In analogia con la maggior parte delle altre metodologie per il monitoraggio degli hate speech sul web, le BoW multilingua individuate vengono utilizzate per il web crawling di Facebook e Twitter attraverso algoritmi finalizzato all'estrazione dei dati nei 9 paesi interessati. Gli indirizzi web da cui il crawler scarica i dati vengono indicizzati nel database di eMORE.</p> <p>Il database creato dal web crawling viene integrato dai dati inseriti dagli utilizzatori della APP, per aumentare la base informativa a disposizione e migliorare gli algoritmi di ricerca (modalità bottom-up).</p> <p>Sulle informazioni contenute nel database viene realizzata una content-analysis manuale (sentiment analysis) per eliminare i falsi positivi, cioè quei post che contengono parole chiave, ma utilizzate in modalità neutra o positiva (senza finalità di odio). Una volta eliminati i falsi positivi, i rimanenti dati vengono inseriti nella parte del database liberamente consultabile da qualunque persona interessata al fenomeno (piattaforma di consultazione).</p> <p>Ogni dato inserito contiene la geolocalizzazione e i metadati dall'autore dell'<i>hate speech</i>.</p>

<p>Principali Risultati</p>	<p>Il progetto si è proposto la validazione degli output di progetto (<i>web crawler</i>, APP, database e piattaforma di consultazione) così da poter trasferire tali output a tutti quei soggetti della società civile attivi nel contrasto all'<i>hate speech</i>. In particolare, i risultati del monitoraggio sono resi disponibili attraverso un portale di consultazione.</p> <p>Nella fase attuale, i risultati distintivi di eMORE si concentrano principalmente sull'APP, multi paese, sulla quale si è avuta una risposta particolarmente positiva perché risponde all'esigenza riconosciuta di creare strumenti in grado di coinvolgere e mobilitare le persone nella segnalazione e nel contrasto dei discorsi di incitamento all'odio.</p>
<p>Eventuali criticità incontrate e modalità di soluzione</p>	<p>Sono state evidenziate una serie di criticità, a partire dalla fase di individuazione delle parole chiave, che hanno dato risultati troppo eterogenei tra i diversi contesti nazionali coinvolti (alcuni paesi hanno identificato pochi termini, altri hanno identificato termini ambigui).</p> <p>Neanche l'aver identificato le parole chiave ad opera del team di esperti e non con il supporto dell'utenza (modalità bottom up) ha evitato il problema di un'elevata eterogeneità nella quantità e nella qualità delle stesse, con successive conseguenze nei risultati di <i>web crawling</i>. Conseguentemente, non è stato ancora possibile raggiungere un'effettiva armonizzazione multi-paese nel monitoraggio. Infine, il <i>web crawling</i> riscontra ancora delle problematiche nell'intercettare i discorsi di odio su Facebook. Questa criticità viene segnalata come particolarmente rilevante da molti paesi partecipanti al progetto, poiché molte realtà attive nella produzione e diffusione di contenuti di odio sono attive prevalentemente su quel social network.</p>

Nome esperienza /metodologia	Hatebase
Sito web	hatebase.org
Soggetto attuatore/coordinatore	<p>Hatebase è una società con sede a Toronto (Canada) co-fondata da Timothy Quinn e The Sentinel Project, un'organizzazione no-profit canadese impegnata nella lotta contro i crimini e le atrocità di massa in paesi come il Kenya, Myanmar e la Repubblica Democratica del Congo.</p> <p>La missione di Hatebase è quella di ridurre gli episodi di <i>hate speech</i> monitorando l'uso e la diffusione di un linguaggio discriminatorio nei confronti di gruppi mirati e di prevenire la violenza celata dietro i discorsi di odio.</p>
Referente	Nome Timothy Quinn
Altri soggetti coinvolti	Soggetti con cui HATEBASE ha attività di partnership: the Sentinel Project, Consiglio d'Europa, Humanitarian Data Exchange, PeaceGeeks, Finn Chirch Aid, HarassMap, Network-racism.ch
Area geografica di applicazione della metodologia	Globale (la metodologia si applica agli <i>hate speech</i> provenienti da tutto il mondo)
Periodo di implementazione	Dal 2013 in corso
Social media di riferimento per l'applicazione della metodologia	Il progetto monitora e classifica i discorsi d'odio online contenuti principalmente sui social media: Facebook, Twitter, Youtube. Accanto alle piattaforme social, HATEBASE considera anche gli online media come quotidiani, riviste, aggregatori di notizie, ecc
Obiettivi e articolazione della metodologia	<p>HATEBASE nasce per creare e mantenere aggiornata una repository contenente vocabolari multilingue in materia di odio online.</p> <p>La possibilità di utilizzare vocabolari multilingua costantemente aggiornati consente di estrarre notevoli quantità di <i>hate speech</i>, rendendo possibile anche la realizzazione di algoritmi predittivi su futuri eventi di odio.</p> <p>La metodologia di HATEBASE si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; Più nello specifico, prevede: <ul style="list-style-type: none"> - la creazione di un vocabolario multilingua sui termini di odio (BoW) attraverso il contributo degli utenti registrati alla piattaforma; - la realizzazione di un'azione di monitoraggio del web (Web crawling) sulla base delle BoW individuate • Fase 2 - attività di analisi dati/contenuti di odio online: <ul style="list-style-type: none"> - classificazione del livello di lesività ei termini di odio; - analisi spaziale per la mappatura geografica a livello di target e di lesività. • Fase 3 - produzione output specifici: <ul style="list-style-type: none"> - i vocabolari pensati per essere disponibili gratuitamente per NGO, ricercatori, giornalisti, che quindi potranno disporre di parole chiave (BoW) aggiornate e multilingua, - database per visualizzare i dati sugli hate speech identificati, - mappe che ricostruiscono la geografia dei discorsi d'odio.
Ambito/oggetto/ target degli hate speech analizzati	<input checked="" type="radio"/> Etnia (Roma e sinti)
	<input checked="" type="radio"/> Colore della pelle
	<input checked="" type="radio"/> Razzismo
	<input checked="" type="radio"/> Altro: religione, genere, orientamento sessuale, disabilità, status sociale


<p>Descrizione della metodologia e tecniche di analisi</p>	<p>Per rendere possibile non solo la creazione di un esteso vocabolario multi-lingua sui termini di odio, ma anche il suo costante aggiornamento, HATEBASE si fonda su una modalità partecipativa e bottom up realizzata attraverso un'interfaccia utente simile a Wikipedia per l'inserimento dei vocaboli da parte degli utenti.</p> <p>HATEBASE integra quindi la finalità di individuazione e analisi dell'<i>hate speech</i> online alla finalità di mobilitazione e engagement attivo della società e coinvolgimento delle persone.</p> <p>Sviluppa prevalentemente tecniche di identificazione di parole chiave classificando queste ultime per target, luogo in cui sono utilizzate, livello di polarizzazione a cui sono associate (poco lesive, lesive, molto lesive).</p> <p>Questo obiettivo è stato realizzato attraverso:</p> <ul style="list-style-type: none"> - una piattaforma wiki per l'inserimento delle parole chiave corredata da geolocalizzazione, frase in cui sono contenuti, indicazione di sinonimi, lingua di appartenenza; - algoritmi per la classificazione del livello di lesività; - algoritmi di analisi spaziale per la mappatura geografica dettagliata anche a livello di target e di lesività. <p>Per utilizzare l'interfaccia WIKI e contribuire all'aggiornamento condiviso dei termini di odio, gli utenti possono iscriversi alla piattaforma attraverso la creazione di profili utente specifici e in questa inserire le parole di odio fornendo anche alcune informazioni di contesto: luogo in cui hanno visto l'uso del termine (discorso o social), la localizzazione, eventuali sinonimi.</p> <p>Sulla base dei dati contenuti nel vocabolario, algoritmi di analisi del linguaggio assegnano ad ogni termine un punteggio di lesività.</p> <p>Utilizzando il suo intero vocabolario multi-lingua, la metodologia HATEBASE realizza anche azioni di identificazione ed estrazione di <i>hate speech</i> online attraverso algoritmi, così creando un ulteriore database contenente discorsi di odio online geolocalizzati.</p> <p>Il database di HATEBASE è integrato, inoltre, da un'interfaccia cartografica per visualizzare i luoghi in cui i termini contenuti nei vocabolari di HATEBASE sono stati utilizzati nell'ambito dei discorsi di odio online, consentendo così di verificare l'estensione geografica delle parole chiave relative ai discorsi d'odio.</p> <p>HATEBASE integra inoltre un API per il download dei vocabolari multilingue, una funzionalità molto utilizzata dai progetti che sviluppano algoritmi di machine learning per estrarre dati dal web. L'estrazione dei vocaboli può avvenire sulla base di classificazioni per target, geo-localizzazione, grado di lesività, lingua.</p> <p>L'estensione particolarmente notevole di parole chiave, consente a HATEBASE di sviluppare anche algoritmi predittivi che sfruttano il notevole quantitativo di <i>hate speech</i> estratto e geolocalizzato per prevedere la possibilità che dai discorsi di odio sul web possano scaturire eventi di odio nel mondo reale.</p> <p>Gli utenti grazie alla possibilità di interrogare il database mediante API sulla base di parole chiave relative a discorsi di odio possono ottenere previsioni sulle tendenze dei fenomeni d'odio in determinate regioni e sulle probabilità di conflitti.</p>
<p>Principali Risultati</p>	<p>HATEBASE promuove il protagonismo della società civile nel contribuire alla creazione di un database esteso dal punto di vista linguistico e geografico che possa consentire di applicare sistemi di machine learning per un monitoraggio capillare e globale degli <i>hate speech</i> online. Ad oggi, HATEBASE copre oltre 200 Paesi e oltre 80 lingue e costituisce il più esteso database di <i>hate speech</i> per scala geografica e linguistica, grazie alla realizzazione della piattaforma per il suo aggiornamento che consente di ottenere a costo zero contributi costanti da parte degli utenti per l'aggiornamento dei vocaboli. La rilevanza del database di parole chiave ha contribuito a fare di HATEBASE uno dei maggiori punti di</p>

	<p>riferimento per l'identificazione delle parole chiave da utilizzare nei progetti di analisi degli <i>hate speech</i>.</p> <p>La repository di HATEBASE è infatti la più utilizzata tra le esperienze di osservazione e monitoraggio degli <i>hate speech</i> online, soprattutto in contesti multilingue, dove la creazione di team dedicati all'identificazione delle parole chiave per ogni contesto geografico richiede di assicurare uniformità tra i diversi vocabolari di odio da individuare.</p> <p>La possibilità di utilizzare vocabolari multilingua costantemente aggiornati consente inoltre di estrarre notevoli quantità di <i>hate speech</i>, rendendo possibile la realizzazione di algoritmi predittivi su futuri eventi di odio.</p> <p>Il coinvolgimento della società civile viene realizzato anche attraverso la possibilità di accedere attraverso API ai dati estratti nonché di visualizzarli su mappe.</p>
<p>Eventuali criticità incontrate e modalità di soluzione</p>	<p>Il principale limite di HATEBASE è rappresentato dalla forte eterogeneità nell'estensione dei vocabolari tra differenti contesti geografici/lingue. Questa differenziazione è legata al coinvolgimento diretto degli utenti che è molto diverso da contesto a contesto. Tale criticità può essere attenuata ampliando la copertura geografica e le partnership con soggetti attivi nel contrasto degli <i>hate speech</i> così da favorire un più ampio coinvolgimento nella creazione dei vocabolari per tutte le aree geografiche.</p>

Nome esperienza /metodologia	Hatelab online Hate Speech Dashboard	
Sito web	hatelab.net	
Soggetto attuatore/coordinatore	University of Cardiff (UK)	
Referente	Nome	Professor Matthew Williams
	email	williamsm7@cardiff.ac.uk
Altri Soggetti coinvolti	Nessuno	
Soggetto finanziatore	UK Economic and Social Research Council	
Area geografica di applicazione della metodologia	Globale (la metodologia di analisi si applica agli <i>hate speech</i> prodotti in tutto il mondo)	
Periodo di implementazione	Dal 2016 (in corso)	
Social media di riferimento per l'applicazione della metodologia	Tutti i social network del Surface web	
Obiettivi e articolazione della metodologia	<p>Il progetto ha inteso creare un prototipo di Online Hate Speech Dashboard con lo scopo di:</p> <p>(i) individuare e analizzare i dati sull'<i>hate speech</i> online generato nelle piattaforme social su Internet;</p> <p>(ii) verificare, sulla base dei dati estratti ed analizzati, la relazione tra eventi (manifestazioni, crisi economiche, eventi di cronaca, Brexit) nel contesto reale (offline) e gli <i>hate speech</i> online, così da analizzare in che modo tali eventi influenzano la produzione di <i>hate speech</i> online. Intende cioè contribuire alla conoscenza della fase generativa dell'odio online per aumentare l'efficacia delle misure di contrasto.</p> <p>Inoltre, l'Online Hate Speech Dashboard attraverso analisi network degli <i>hate speech</i> online, intende fornire informazioni sull'estensione delle reti virtuali di creazione e condivisione degli <i>hate speech</i> online e dati sulle dinamiche di propagazione dei contenuti di odio online.</p> <p>La metodologia di HATELAB si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; Più nello specifico, prevede: <ul style="list-style-type: none"> - identificazione di parole chiave (BoW) riferibili all'odio; - monitoraggio dei social networks attraverso l'utilizzo di algoritmi per estrarre contenuti che contengano le parole chiave; • Fase 2 - attività di analisi dati/contenuti di odio online; <ul style="list-style-type: none"> - classificazione dei dati estratti sulla base del contenuto attraverso algoritmi di <i>machine learning</i> e supervisione umana: applicazione della <i>sentiment analysis</i> per scartare contenuti neutri o positivi; - ottenuta una base informativa di dati adeguata, implementazione ulteriore di algoritmi di <i>content analysis (lexical parsing)</i> che automatizzano la classificazione dei dati estratti, rendendo possibile processare volumi di dati sempre più estes; - implementazione di algoritmi per analizzare la diffusione dei contenuti all'interno dei social network; • Fase 3 - produzione output specifici: realizzazione di mappe per la visualizzazione dei dati 	
Ambito/oggetto/		Colore della pelle
		Etnia

target degli <i>hate speech</i> analizzati	<input type="radio"/> Razzismo
	<input type="radio"/> Altro: religione, disabilità, orientamento sessuale, genere
Descrizione della metodologia e tecniche di analisi	<p>La metodologia ha previsto l'identificazione di parole chiave (BoW) multilingua su cui impostare l'analisi dei flussi di dati dai social network. A differenza di altri progetti, questo approccio integra alle BoW associate all'odio le BoW associate a eventi reali scatenanti (eventi trigger) che possono favorire la nascita di <i>hate speech</i> online. L'identificazione di due tipologie di BoW consente di effettuare due distinti tipi di monitoraggio ed estrazione dati (senza e con gli eventi trigger) per analizzare se esista una relazione e quale sia la sua portata. Questo secondo tipo di estrazione segue le stesse procedure di screening della prima.</p> <p>I contenuti estratti vengono sottoposti ad una <i>content analysis</i> particolarmente raffinata nell'ambito di un processo di content analysis automatizzato sempre più accurato. Operativamente, partendo da una <i>sentiment analysis</i> e <i>lexical parsing</i> i cui algoritmi sono sottoposti a supervisione umana realizzata ricorrendo a piattaforme di <i>crowdsourcing</i> (esternalizzazione dell'attività umana di supervisione), si migliora progressivamente la capacità dell'algoritmo nel fare screening, riducendo progressivamente anche la necessità della supervisione umana.</p> <p>Più in dettaglio, la metodologia prevede:</p> <ul style="list-style-type: none"> - la creazione di algoritmi di <i>machine learning</i> con supervisione umana per classificare i contenuti estratti in contenuti di odio, contenuti neutrali e positivi (<i>sentiment analysis</i>); - l'utilizzo di algoritmi di machine classification sul campione di contenuti di odio e contenuti neutrali e positivi per automatizzare identificazione e classificazione su volumi di Big Data; - l'applicazione supervisionata del modello di classificazione ai dati, con verifica dell'efficacia degli standard di classificazione attraverso <i>crowdsourcing</i> affidato a classificatori umani; - l'implementazione di algoritmi di analisi lessicale (<i>Lexical Parsing</i>) per ridurre falsi positivi focalizzando l'analisi dei contenuti sulle strutture grammaticali dei contenuti di odio online; - l'implementazione di algoritmi per l'individuazione di effetti contagio; - la realizzazione di mappe per la visualizzazione dei dati elaborati <p>La metodologia HATELAB si contraddistingue quindi per la tipologia di algoritmi integrati nello strumento informatico. In particolare, all'algoritmo utilizzato di <i>machine learning</i> per estrarre e classificare i contenuti d'odio dai social network, si uniscono due ulteriori classi di algoritmi.</p> <p>La prima classe estrae e classifica i contenuti di odio online in relazione ad eventi scatenanti nel contesto reale-offline (attraverso l'integrazione delle parole chiave associate agli eventi con le parole chiave identificative dell'odio online).</p> <p>La seconda classe utilizza i dati classificati per affinare tutti gli algoritmi di ricerca, creando un circuito virtuoso di progressivo miglioramento della ricerca e della classificazione degli <i>hate speech</i>. Infine, la metodologia HATELAB contiene anche algoritmi di <i>network analysis</i>, finalizzati a monitorare e analizzare le dinamiche di propagazione dei diversi contenuti di odio.</p> <p>Parte rilevante dei dati utilizzati dall'<i>Online Hate Speech Dashboard</i> è costituita dai metadati relativi agli autori dell'online <i>hate speech</i>, con particolare riferimento alla geolocalizzazione. Tali dati costituiscono elementi rilevanti per l'analisi della relazione tra possibili eventi trigger (scatenanti) e gli online <i>hate speech</i>.</p>
Principali Risultati	<p>Il progetto ha inteso supportare una migliore comprensione dei dati e sviluppare un prototipo di <i>dashboard</i> in grado di mostrare le tendenze nel tempo dei discorsi di incitamento all'odio pubblicati sui social media in occasione di eventi specifici.</p>



	Il prototipo dell'Online <i>Hate Speech Dashboard</i> è stato realizzato ed è ora in fase di sviluppo la versione che verrà testata operativamente (risultati attesi per il 2020)
Eventuali criticità incontrate e modalità di soluzione	Nessuna criticità evidenziabile dalla documentazione a disposizione

Nome esperienza /metodologia	HATEMER	
Sito web	hatemeter.eu	
Soggetto attuatore/coordinatore	Università di Trento – eCrime	
Referente	Nome referente per l'Italia	Mario Diani University of Trento – eCrime
	email	ecrime@unitn.it
	Tel	+39 0461 282336
Altri attori coinvolti	La realizzazione della metodologia è avvenuta nell'ambito di un progetto di scala europea che ha visto il coinvolgimento di università e NGO europee attive nella difesa dei diritti umani: Università di Tolosa 1 Capitale, Teesside University (UK) fondazione Bruno Kessler (IT), Amnesty Int. sezione Italiana, Association de defense des droits de l'homme (fFR), Stop Hate UK	
Soggetto finanziatore	Commissione Europea: Rights Equality an Citizenship Program - Action Grant (REC-DISC-AG-2016)	
Area geografica di applicazione della metodologia	Europa: Italia, Francia e Regno Unito Copertura nazionale di ogni ambito geografico di applicazione	
Periodo di implementazione	Da 01-02-2018 a 31-01-2019	
Social media di riferimento per l'applicazione della metodologia	Le piattaforme web su cui il progetto Hatemeter si focalizza nell'individuazione dell' <i>hate speech</i> online sono rappresentate dai social network, con particolare riferimento a Facebook e Twitter	
Obiettivi e articolazione della metodologia	<p>HATEMETER sviluppa una piattaforma ICT dedicata esclusivamente a monitorare e analizzare automaticamente i dati Internet e dei social media relativamente a uno specifico ambito/oggetto di <i>hate speech</i>: razzismo con riferimento all'Islamofobia.</p> <p>La metodologia si propone anche di elaborare, sulla base dell'analisi dei dati relativi agli <i>hate speech</i> online, possibili risposte di contronarrativa per agevolare la produzione di contenuti di contrasto da parte della società civile, così da rendere la piattaforma uno strumento utilizzabile da tutti i soggetti attivi nel contrasto del fenomeno (principali stakeholder di riferimento).</p> <p>La metodologia di HATEMETER si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; Più nello specifico, prevede: <ul style="list-style-type: none"> - creazione di team multidisciplinari per l'inquadramento specifico del fenomeno (criminologi, scienziati sociali, informatici, statistici, giuristi) - realizzazione di sistemi di algoritmi per l'individuazione e l'estrazione dei post social islamofobici: definizione parole chiave, monitoraggio Internet, estrazione contenuti rilevanti; • Fase 2 - attività di analisi dati/contenuti di odio online: <ul style="list-style-type: none"> - <i>Content analysis</i> - Analisi dei trend di diffusione e di sviluppo dei topic dei post estratti - Analisi dei profili degli autori degli <i>hate speech</i> - Realizzazione di algoritmi per l'elaborazione di contronarrativa sulla base dei contenuti dei post estratti - Test e validazione della piattaforma ai fini della sua trasferibilità (per utilizzo esclusivo da parte degli stakeholder attivi in questo ambito) • Fase 3 - produzione output specifici: piattaforma ICT (database e dashboard per la visualizzazione dei dati) 	
Ambito/oggetto/target degli <i>hate speech</i> analizzati		Razzismo (islamofobia)

<p>Descrizione della metodologia e tecniche di analisi</p>	<p>HATEMETER integra l'analisi degli <i>hate speech</i> islamofobici con ulteriori azioni volte sia a contrastare tali contenuti attraverso la produzione automatizzata di contronarrativa che a mobilitare e coinvolgere attivamente la società civile, prevedendo un futuro utilizzo della piattaforma per favorire la conoscenza del fenomeno e il ruolo attivo della società civile nel contrasto tempestivo ed efficace degli <i>hate speech</i>.</p> <p>La metodologia prevede la creazione di team multi-disciplinari per un approccio olistico alla tematica dell'islamofobia.</p> <p>La metodologia Hatemeter si contraddistingue anche per il suo ambito di applicazione multi-paese. La fase di realizzazione e sperimentazione ha infatti coinvolto 3 differenti contesti europei (Italia, Francia e UK) e contribuirà a realizzare analisi trasversali e definire strategie di contrasto condivise. La creazione di un prodotto con un'interfaccia fruibile per ONG e realtà attive nel contrasto all'odio online (e non per una generica utenza) consentirà di strutturare in futuro una capillare rete multi-paese per contrastare le campagne di odio.</p> <p>Da un punto di vista applicativo, i team multidisciplinari elaborano le parole chiave (BOW) su cui impostare gli algoritmi informatici di analisi ed estrazione dei post razzisti a contenuto islamofobico dalle API di Twitter e Facebook. Successivamente, attraverso algoritmi di <i>sentiment analysis</i> i contenuti estratti vengono analizzati al fine di escludere falsi positivi.</p> <p>HATEMETER raccoglie e analizza anche i metadati degli autori dei post, con particolare riferimento alla geolocalizzazione e al network di relazioni.</p> <p>HATEMETER realizza anche l'analisi dei picchi dei discorsi islamofobici e dei profili la cui attività è intensamente caratterizzata da contenuti islamofobici, identificando i profili che si configurano come influencer sulla tematica. Su tali profili viene cioè realizzata un'analisi etnografica e individuato il sistema delle relazioni, le modalità di coinvolgimento e le interazioni con riferimento ai post.</p> <p>I dati vengono raccolti in un database, su cui si innesta un tool di visualizzazione dati (<i>dashboard</i>) progettato per essere <i>user-friendly</i> e utilizzabile dalle realtà della società civile (NGOs), giornalisti e attivisti (non da una utenza generalizzata). Attraverso la medesima interfaccia (<i>dashboard</i>), verrà anche garantito l'accesso a contenuti di contrasto generati automaticamente da algoritmi per avviare campagne di contronarrazione. Questi contenuti saranno modificabili manualmente per renderli più aderenti ai singoli casi/contesti.</p> <p>La metodologia verrà testata con tre ONG (una per Paese), prevedendo anche focus group e workshop per aumentare il coinvolgimento di questa tipologia di stakeholder nell'utilizzo degli strumenti ideati dal progetto e verificare la capacità di HATEMETER di rispondere agli obiettivi e istanze poste a livello di Unione Europea sul tema degli <i>hate speech</i>.</p>
<p>Principali Risultati</p>	<p>Le caratteristiche distintive della metodologia sono individuabili nel suo focalizzarsi su una singola tipologia di <i>hate speech</i> online (razzismo/islamofobia), di cui approfondisce gli aspetti specifici e nell'aver integrato l'attività di analisi con la contronarrativa.</p> <p>L'unione di algoritmi di estrazione dei dati e analisi dei contenuti con algoritmi per la creazione di contronarrativa, renderà possibile realizzare strategie di contrasto tempestive.</p> <p>Attualmente il progetto/metodologia è nella fase pilota. I Risultati sono attesi per il 2019-2020.</p> <p>In prospettiva, successivamente alla sua validazione, potrà essere ampliato il numero dei paesi coinvolti, il numero dei soggetti a cui mettere a disposizione HATEMETER e la tipologia/oggetto/target di <i>hate speech</i> online da considerare, attraverso la realizzazione dell'HATEMETER Lab.</p>
<p>Eventuali criticità incontrate e modalità di soluzione</p>	<p>Nessuna criticità evidenziabile dalla documentazione a disposizione</p>

Nome esperienza /metodologia	Mandola
Sito web	mandola-project.eu
Soggetto attuatore/coordinatore	Foundation for Research and Technology (GR) Centri di ricerca greco, svolge attività di ricerca scientifica specializzata in settori strategici ad alto valore aggiunto, concentrandosi su attività interdisciplinari di ricerca e sviluppo (R&S) in aree di grande interesse scientifico, sociale ed economico
Referente	Nome Evangelos Markatos
	email markatos@ics.forth.gr
Altri soggetti coinvolti	La realizzazione della metodologia è avvenuta nell'ambito di un progetto europeo che ha visto il coinvolgimento dei seguenti partner di progetto: Aconite Internet Solutions (IR); The International Cyber Investigation Training Academy (BG); Inthemis (FR); Autonomous University of Madrid (ES); University of Cyprus (CY); University of Montpellier (FR)
Soggetto finanziatore	Commissione Europea, nell'ambito del Programma REC - RIGHTS, EQUALITY AND CITIZENSHIP PROGRAMME (2014-2020)
Area geografica di applicazione	Europa: Grecia, Irlanda, Bulgaria, Francia, Spagna, Cipro (copertura nazionale di ogni ambito geografico di applicazione). Applicazione globale in prospettiva futura
Periodo di implementazione	In corso
Social media di riferimento per l'applicazione della metodologia	Le piattaforme web su cui il progetto MANDOLA si focalizza per l'individuazione dell' <i>hate speech</i> online sono Twitter e Google
Obiettivi e articolazione della metodologia	<p>MANDOLA si propone di migliorare la comprensione del fenomeno degli <i>hate speech</i> online, di evidenziarne l'impatto e di consentire alla cittadinanza di monitorare e segnalare gli <i>hate speech</i>. Questa metodologia integra cioè l'analisi dell'online <i>hate speech</i> alla finalità di mobilitazione e engagement attivo dei cittadini così da costruire una base informativa utilizzabile anche dalle istituzioni e dalle forze dell'ordine per contrastare l'<i>hate speech</i> illegale.</p> <p>In particolare, prevede di:</p> <ul style="list-style-type: none"> - monitorare e analizzare la diffusione degli <i>hate speech</i> online nei Paesi europei; - realizzare la possibilità di distinguere tra discorsi di odio potenzialmente illegali e discorsi di odio non illegali; - fornire ai cittadini strumenti per contrastare i discorsi di odio online; - creare un'infrastruttura di segnalazione degli <i>hate speech</i> che consenta di ampliare la base informativa ma anche che colleghi i cittadini con le forze dell'ordine, consentendo di segnalare gli <i>hate speech</i> illegali. - fornire ai policy maker informazioni che possano essere utilizzate per promuovere politiche volte a mitigare la diffusione dei discorsi sull'odio online; - trasferire le migliori pratiche tra gli Stati membri europei; <p>La metodologia di MANDOLA si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; realizzazione di uno strumento informatico in grado di estrarre i post di odio su Twitter e Google, • Fase 2 - attività di analisi dati/contenuti di odio online: utilizzo di tecniche automatizzate di <i>content analysis</i> per analisi dei post, classificazione per target e geolocalizzazione; • Fase 3 - produzione output specifici:

	<p>Realizzazione di una <i>dashboard user-friendly</i>, di mappe, grafici e tabelle e di un portale e una APP per la segnalazione di contenuti di odio online.</p> <p>La metodologia prevede anche un'analisi della legislazione internazionale e comunitaria sugli <i>hate speech</i> e della legislazione nazionale in 10 paesi europei.</p>
Ambito/oggetto/target degli <i>hate speech</i> analizzati	<input type="radio"/> Colore della pelle
	<input type="radio"/> Etnia
	<input type="radio"/> Razzismo
	<input type="radio"/> Altro: religione, disabilità, sessismo, omofobia, status sociale
Descrizione della metodologia e tecniche di analisi	<p>MANDOLA individua i termini di odio (BoW), ricorrendo al database di un altro progetto (HATEBASE) e attraverso la supervisione di un team di scienziati sociali. Attraverso tali parole chiave, l'algoritmo definito dalla metodologia estrae i contenuti di odio da Twitter e Google, sfruttando le API. Il corpus di dati estratto viene sottoposto a <i>content analysis</i>, integrando algoritmi di <i>sentiment analysis</i> e <i>lexical parsing</i> sottoposti a supervisione umana.</p> <p>Esclusi i falsi positivi, i dati rimanenti e geolocalizzati confluiscono nel database e possono essere visualizzati e condivisi attraverso una dashboard web progettata per un facile utilizzo da parte dell'utenza e una serie di strumenti:</p> <ul style="list-style-type: none"> - mappe termografiche in grado di mostrare la geografia dell'odio online; - tabelle dati esportabili; - grafici. <p>E' stata realizzata una <i>dashboard user-friendly</i> (ora in versione open-beta, cioè ancora in fase sperimentale pre rilascio ma testabile dall'utenza) che consente agli utenti la consultazione dei dati estratti e di condurre approfondimenti sulle diverse tematiche relative all'<i>hate speech</i> online. Per una leggibilità immediata dei dati, è prevista la realizzazione di mappe, grafici e tabelle.</p> <p>Attraverso il portale è possibile inserire segnalazioni relative a fenomeni di <i>hate speech</i> online così integrando la base informativa disponibile nel database di progetto che raccoglie gli <i>hate speech</i> identificati e di ricevere feedback in merito alla segnalazione. Le forze dell'ordine possono accedere al portale per indagare sulle attività criminali. Infine, è prevista la realizzazione di una APP che consentirà un più facile accesso ai dati e costituirà un ulteriore strumento per integrare le informazioni sugli <i>hate speech</i> presenti nel database, consentendo anche essa la segnalazione diretta da parte dell'utente.</p> <p>MANDOLA prevede di effettuare analisi multi-paese e realizzare strategie di contrasto condivise.</p> <p>In particolare, realizza un'analisi della legislazione internazionale e comunitaria sugli <i>hate speech</i> e della legislazione nazionale in 10 paesi europei per delineare il quadro giuridico e poter distinguere e classificare gli <i>hate speech</i> potenzialmente illegali e gli <i>hate speech</i> legali.</p>
Principali Risultati attesi	<p>MANDOLA non si è limitata a focalizzarsi sull'<i>hate speech</i> online nelle piattaforme di social network, ma estende l'analisi anche a Google. I risultati distintivi di MANDOLA sono rappresentati dall'integrazione di strumenti di analisi e monitoraggio con la realizzazione di una <i>dashboard</i> ben articolata per la diffusione e visualizzazione dei dati e delle informazioni raccolte. La condivisione dei dati è infatti un obiettivo strategico del progetto, che viene perseguito anche attraverso la progettazione di altri strumenti, così da raggiungere una platea più ampia di pubblico.</p>
Eventuali criticità incontrate e modalità di soluzione	<p>Ad oggi la <i>dashboard</i> è disponibile in versione open-beta, senza la funzionalità di esportazione dei dati in formato aperto. La APP è ancora in fase di sviluppo e non è stata ancora implementata la possibilità di identificare se gli episodi di <i>hate speech</i> online rientrano o meno in fattispecie illegali e/o sanzionabili.</p>

Nome esperienza /metodologia	Mapa dell'Intolleranza	
Sito web	voxdiritti.it	
Soggetto attuatore/coordinatore	<p>Osservatorio Voxdiritti</p> <p>Associazione privata senza scopo di lucro impegnata nella realizzazione di un Osservatorio dei Diritti e nel favorire la diffusione della cultura dei diritti. L'associazione per realizzare il progetto "La Mapa dell'Intolleranza" ha coinvolto un team di ricercatori provenienti da tre università italiane: Università di Bari, Università La Sapienza di Roma, Università Statale di Milano.</p> <p>In particolare, lo sviluppo degli algoritmi è stato realizzato da ricercatori del gruppo di ricerca SWAP del Dipartimento di Informatica dell'Università di Bari, mentre l'identificazione delle parole chiave e la content analysis sono state supervisionate da team di psicologi e sociologi delle università La Sapienza di Roma e Statale di Milano</p>	
Referente	Nome	Marilisa D'Amico
	email	voxdiritti.redazione@gmail.com
Altri soggetti coinvolti	Nessuno	
Soggetto finanziatore	Autofinanziato	
Area geografica di applicazione della metodologia	Italia (applicazione su tutto il territorio nazionale)	
Periodo di implementazione	Dal 2014 (in corso)	
Social media di riferimento per l'applicazione della metodologia	La metodologia analizza i discorsi di odio generati e condivisi su Twitter nel contesto italiano.	
Obiettivi e articolazione della metodologia	<p>La Mapa dell'Intolleranza sui discorsi d'odio online nel contesto italiano effettua attività di estrazione e analisi dei dati (annualmente) sugli <i>hate speech</i> online con l'obiettivo di:</p> <ul style="list-style-type: none"> • quantificare il fenomeno; • verificare i trend su più anni; • realizzare un'analisi dal punto di vista geografico delle diverse categorie di <i>hate speech</i> online classificate per target; • sviluppare algoritmi di analisi dei contenuti adeguati alle peculiarità della lingua italiana. <p>La metodologia della Mapa dell'intolleranza si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; Più nello specifico, prevede: <ul style="list-style-type: none"> - identificazione delle parole chiave (BoW) su cui impostare gli algoritmi di monitoraggio ed estrazione dei dati da Twitter; - realizzazione di algoritmi informatici per l'estrazione dei tweet geolocalizzati di odio contenenti le parole chiave; • Fase 2 - attività di analisi dati/contenuti di odio online; <ul style="list-style-type: none"> - realizzazione di algoritmi di <i>content analysis</i> automatici per l'esclusione di falsi positivi sviluppati con riferimento alle caratteristiche della lingua italiana; - classificazione dei tweet per target specifici; • Fase 3 - produzione output specifici: <ul style="list-style-type: none"> - elaborazione di mappe termografiche dove i "punti più caldi" corrispondono alle località con il maggior numero di tweet per ogni target 	
Ambito/oggetto/target degli hate speech analizzati		Colore della pelle
		Etnia

	●	Razzismo
	●	Altro: omofobia, disabilità, genere
Descrizione della metodologia e tecniche di analisi	<p>La metodologia concentra l'analisi su Twitter in quanto contiene molti dati/<i>hate speech</i> geolocalizzati e si presta a creare ambienti di viralizzazione in cui gli <i>hate speech</i> proliferano.</p> <p>La metodologia ha un ambito di applicazione limitata all'Italia (mono-paese) e si concentra solo sulla lingua italiana e le sue caratteristiche, al fine di elaborare strategie di analisi dei contenuti di odio particolarmente avanzate e in grado di individuare gli <i>hate speech</i>, tenendo conto delle diverse sfumature del linguaggio e dei termini polisemici.</p> <p>Per tutte le fasi di progettazione relativamente alla dimensione linguistica, il progetto si è affidato e si affida a team specializzati in linguistica e sociologia, al fine di elaborare algoritmi in grado di rispondere a tutte le complesse caratteristiche della lingua italiana. Non si affida dunque a vocabolari già ideati da precedenti progetti che identificano parole chiave (come HATEBASE). I Team multi-disciplinari identificano le parole chiave (BoW) che vengono assegnate agli algoritmi per l'estrazione degli <i>hate speech</i>.</p> <p>I tweet estratti vengono analizzati dapprima attraverso la sentiment analysis così da individuarne il grado di polarizzazione ed escludere i contenuti neutri o positivi e successivamente attraverso la semantic tagging per individuare i termini polisemici e ridurre ulteriormente il numero di falsi positivi. Con la semantic tagging, vengono infatti esclusi i contenuti in cui la parola chiave è utilizzata in un significato differente rispetto a quello di odio. Sia la semantic tagging che la sentiment analysis vengono realizzate attraverso algoritmi impostati sulle informazioni definite dai team multi-disciplinari.</p> <p>Gli <i>hate speech</i> sono georeferenziati, così da ancorare il fenomeno on-line (virtuale) alla dimensione reale (offline) e alla geografia umana e di contesto.</p> <p>La metodologia, inoltre, per favorire la disseminazione dei risultati, ha previsto la realizzazione di mappe finalizzate ad agevolare la comprensione del fenomeno per un pubblico vasto ed eterogeneo, che mostrano la geografia dei tweet di odio in Italia sulla base della quantità di tweet identificati e analizzati per ciascuna località e quantità di tweet identificati per target in ogni località.</p>	
Principali Risultati	<p>La Mappa dell'Intolleranza ha previsto la creazione di un software integrato finalizzato a realizzare estrazione e analisi degli <i>hate speech</i> focalizzandosi sulle caratteristiche della lingua italiana. Il sistema di algoritmi creato per l'analisi presenta una maggiore capacità di escludere falsi positivi, grazie alla semantic tagging che si aggiunge alla sentiment analysis. Infine, l'elaborazione di mappe rende facilmente fruibili i risultati dell'analisi per una platea variegata: scuole, giornalisti, associazioni e istituzioni</p>	
Eventuali criticità incontrate e modalità di soluzione	<p>Nessuna criticità evidenziabile dalla documentazione a disposizione</p>	

Nome esperienza /metodologia	Analyzing the Targets of Hate in Online Social Media (studio)	
Soggetto attuatore/coordinatore	La metodologia di analisi è stata ideata da un team di ricercatori appartenenti a diverse università e istituti di ricerca: Leandro Silva (Federal University of Minas Geras- Brasile), Mainak Mondal (Max Placnk Institute for Software Systems- Germania), Denzil Correa (Max Placnk Institute for Software Systems - Germania), Fabricio Benvenuto (Federal University of Minas Geras - Brasile), Ingmar Weber (Qatar Computing Research System- Qatar)	
Referente	Nome	Leandro Silva
	Email	leandroaraujo@dcc.ufmg.br
Soggetto finanziatore	Lo studio è stato parzialmente finanziato nell'ambito del progetto FAPEMIGPRONEX- MASWeb, Models, Algorithms and Systems for the Web, e grants da CNPq, CAPES e Fapemig	
Area geografica di applicazione della metodologia	Globale	
Periodo di implementazione	Da giugno 2014 a giugno 2015	
Social media di riferimento per l'applicazione della metodologia	La metodologia considera per l'analisi due piattaforme social: Twitter e Whisper	
Obiettivi e articolazione della metodologia	<p>La metodologia proposta nello studio intende migliorare la comprensione del fenomeno degli <i>hate speech</i> online, considerata soprattutto la rapidità con cui internet si popola di nuove piattaforme social di comunicazione e condivisione.</p> <p>In particolare, lo studio ha inteso approfondire come la percezione di anonimato possa influenzare la creazione e la diffusione dell'odio online, attraverso la misurazione sistematica di questo aspetto, che è stata realizzata confrontando i contenuti di odio presenti in una piattaforma che assicura l'anonimato (Whisper) con quelli presenti in una piattaforma che non lo garantisce se non a determinate condizioni (Twitter). In questo senso l'analisi si propone anche di fornire elementi ai policy makers per individuare strategie di prevenzione e contrasto che tengano conto di tale aspetto.</p> <p>Per effettuare tale confronto, è stata sviluppata, sperimentata e validata una metodologia volta a identificare gli <i>hate speech</i> su entrambe le piattaforme e ad analizzarli individuandone le differenze.</p> <p>La metodologia si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi: <ul style="list-style-type: none"> - estrazione di parole chiave (BoW) relative all'odio dal database HATEBASE; - monitoraggio ed estrazione degli <i>hate speech</i> contenenti le parole chiave da Whisper e Twitter; • Fase 2 - attività di analisi dati/contenuti di odio online; <ul style="list-style-type: none"> - realizzazione della <i>content analysis (sentiment analysis)</i>; - integrazione della <i>content analysis</i> attraverso analisi dei dati estratti per identificazione manuale di tipologie specifiche di struttura grammaticale associate all'<i>hate speech (lexical parsing)</i>; - Supervisione manuale per eliminare falsi positivi; - Analisi del livello di propagazione dei contenuti; - Analisi della geografia dei contenuti; - Analisi delle differenze in termini di struttura grammaticale e polarizzazione tra piattaforma anonima e piattaforma non anonima; 	


	<ul style="list-style-type: none"> - Ideazione e validazione di nuovi algoritmi per estrarre l'<i>hate speech</i> online che integrano parole chiave e struttura grammaticale dei contenuti - Applicazione di tali nuovi algoritmi per l'analisi; <ul style="list-style-type: none"> • Fase 3 - produzione output specifici: report di analisi
Ambito/oggetto/ target degli hate speech analizzati	<input type="radio"/> Colore della pelle
	<input type="radio"/> Etnia
	<input type="radio"/> Razzismo
	<input type="radio"/> Altro: comportamento, orientamento sessuale, classe sociale, genere, disabilità, religione, aspetto fisico
Descrizione della metodologia e tecniche di analisi	<p>Per l'identificazione delle parole chiave (BoW) la metodologia fa ricorso al database HATEBASE. Le parole chiave sono utilizzate per estrarre gli <i>hate speech</i> che le contengono attraverso l'API di Twitter e il web crawling di Whisper. I post estratti sono sottoposti ad una successiva fase di screening per eliminare i falsi positivi (<i>sentiment analysis</i>).</p> <p>Successivamente, viene effettuata un'analisi basata sulla struttura della frase allo scopo di verificare manualmente se esistano strutture grammaticali ricorrenti nei discorsi di odio (lexical parsing). Una volta individuate, vengono utilizzate per realizzare un algoritmo per l'identificazione degli <i>hate speech</i> che viene testato sul resto dei dati.</p> <p>Contemporaneamente, la metodologia verifica anche se il livello di anonimato percepito dagli autori dell'<i>hate speech</i> online influisce sulle strutture grammaticali utilizzate.</p> <p>La metodologia geolocalizza gli <i>hate speech</i> e si concentra anche sulle interazioni dei post per verificarne il livello di propagazione attraverso la <i>network analysis</i>.</p> <p>Gli <i>hate speech</i> analizzati sono solo in lingua inglese.</p>
Principali Risultati	<p>La metodologia monitora e analizza il fenomeno degli <i>hate speech</i> per approfondire la correlazione tra anonimato e discorsi di incitamento all'odio. Tra i suoi principali risultati si ravvisa l'introduzione l'elemento della struttura grammaticale per l'identificazione dell'<i>hate speech</i> online attraverso gli algoritmi.</p> <p>Inoltre, verifica le differenze esistenti tra <i>hate speech</i> online generati in piattaforme anonime ed <i>hate speech</i> generati in piattaforme non-anonime.</p>
Eventuali criticità incontrate e modalità di soluzione	<p>Lo studio ricorre ad un'elevata componente umana per l'analisi dei contenuti, un elemento che appare ormai superato dalle costanti evoluzioni degli algoritmi di <i>content analysis</i>.</p> <p>Sarebbe quindi importante valutare come l'algoritmo che analizza la struttura grammaticale dei post (lexical parsing) nel dark web realizzato in questo progetto possa essere inserito in processi totalmente automatizzati.</p>

Nome esperienza /metodologia	Hate speech e dark web - 4chan	
Sito web	<ul style="list-style-type: none"> • http://www0.cs.ucl.ac.uk/staff/G.Stringhini/papers/4chan-ICWSM2017.pdf • ENCASE: https://encase.socialcomputing.eu 	
Soggetto attuatore/coordinatore	Università RomaTre	
Referente	Nome	Gabriel Emile Hine
	email	gabriel.hine@uniroma3.it
Altri soggetti coinvolti	La metodologia è stata sviluppata all'interno del progetto inter-universitario ENCASE, che ha coinvolto oltre all'Università di Roma 3 anche i dipartimenti di ICT di UCL London e Cyprus University of Technology. Il progetto ha previsto anche un partner privato (Telefonica)	
Soggetto finanziatore	ENCASE è finanziato dal programma quadro di ricerca e innovazione Horizon 2020 della Commissione europea nell'ambito dell'azione Marie Skłodowska-Curie Research and Innovation Staff Exchanges Action, Grant Agreement No. 691025.	
Area geografica di applicazione della metodologia	Globale	
Periodo di implementazione	Da giugno 2016 a settembre 2016	
Social media di riferimento per l'applicazione della metodologia	L'azione del progetto ENCASE si concentra su 4chan , la piattaforma social che ha dato visibilità al movimento "alt+right ²³ " e che è considerata parte del dark web .	
Obiettivi e articolazione della metodologia	<p>Il progetto analizza una parte meno esplorata di Internet - il dark web- per capire se rivesta un ruolo nella generazione dell'<i>hate speech</i> online. Approfondire l'<i>hate speech</i> nel dark web è particolarmente rilevante per i seguenti motivi:</p> <ul style="list-style-type: none"> • il dark web è popolato da piattaforme social che favoriscono l'anonimato degli utilizzatori, favorendone l'aggressività verbale. Di conseguenza, le sue caratteristiche strutturalmente diverse dal resto di Internet possono avere un impatto non trascurabile sul tipo di <i>hate speech</i> che si genera nelle sue piattaforme; • gli utenti che generano odio nel dark web possono avere caratteristiche e comportamenti diversi rispetto agli altri utenti; • buona parte degli attacchi informatici partono dal dark web, di conseguenza, una maggiore comprensione della genesi degli attacchi e della loro propagazione rappresenta un elemento conoscitivo rilevante per il contrasto dell'<i>hate speech</i> online. <p>Più in dettaglio, la metodologia si propone quindi verificare:</p> <ul style="list-style-type: none"> • quale impatto abbia il dark web nella creazione degli <i>hate speech</i> online; • quali siano le caratteristiche distintive dell'odio online generato nel dark web rispetto all'odio online generato nei social network sul <i>surface web</i>, come Twitter e Facebook; • quali siano le relazioni tra odio generato nel dark web e odio prodotto nel resto di internet, con particolare attenzione agli attacchi organizzati (raid) di hate speech che vengono programmati nel dark web per colpire profili social specifici su Youtube, Twitter, Facebook. 	

²³ Alt-right è l'abbreviazione di alternative right: "destra alternativa" ed identifica un movimento politico, nato negli Stati Uniti, che promuove ideologie di destra alternative a quelle tradizionali del conservatorismo. Come movimento, si caratterizza, ad oggi come fenomeno subculturale, alimentato attraverso gruppi di discussione prevalentemente presenti sul web, in piattaforme quali 4chan, 8chan, Reddit e Twitter.

	<p>La metodologia si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi: <ul style="list-style-type: none"> - identificazione delle parole chiave (BoW) su cui impostare gli algoritmi di monitoraggio ed estrazione dei dati dalla piattaforma 4chan nel dark web; - creazione di un web crawler per estrarre da 4chan gli <i>hate speech</i>; • Fase 2 - attività di analisi dati/contenuti di odio online: <ul style="list-style-type: none"> - identificazione di elementi distintivi dell'<i>hate speech</i> nel dark web (struttura sintattica, polarizzazione termini); - geolocalizzazione degli <i>hate speech</i> nel dark web - analisi dei profili utenti nel dark web - analisi dell'engagement degli altri utenti del dark web in risposta agli <i>hate speech</i>; - analisi della propagazione dei contenuti di odio dal dark web al resto di Internet. • Fase 3 - produzione output specifici: report di analisi
<p>Ambito/oggetto/target degli <i>hate speech</i> analizzati</p>	<ul style="list-style-type: none"> <input type="radio"/> Colore della pelle <input type="radio"/> Etnia <input type="radio"/> Razzismo
<p>Descrizione della metodologia e tecniche di analisi</p>	<p>La metodologia si basa sullo sviluppo di diversi algoritmi informatici con la finalità di integrarli in un processo predittivo che, partendo dal monitoraggio e dalla estrazione di contenuti di odio nel dark web, sia in grado di prevedere, in ottica di prevenzione, se, come e quando questi contenuti genereranno attacchi a pagine e profili in altre piattaforme di internet, con particolare riferimento ai social (Youtube, Facebook, Twitter).</p> <p>In particolare, il progetto ha sviluppato e testato algoritmi informatici per estrarre dati sui discorsi d'odio da una piattaforma del dark web (4chan), analizzarli in relazione alle caratteristiche del loro contenuto e del loro autore e identificare le dinamiche attraverso cui questi contenuti vengono prodotti e condivisi per generare attacchi di odio verso altre piattaforme.</p> <p>L'analisi parte dall'identificazione di parole chiave (BoW), che viene realizzata attingendo al database dei vocaboli HATEBASE (si v. scheda dedicata).</p> <p>Le parole chiave vengono utilizzate per il web crawling attraverso API della /pol/ ("politically incorrect") discussion board di 4chan al fine di estrarre gli <i>hate speech</i> nel dark web.</p> <p>Sui dati estratti vengono realizzate diverse analisi, necessarie per avere una base informativa sufficientemente articolata e robusta da consentire un'attività predittiva. L'esigenza di avere una base informativa estesa e robusta è ancora più necessaria considerato il maggiore anonimato che è garantito da questo specifico spazio web.</p> <p>In particolare, vengono analizzati:</p> <ol style="list-style-type: none"> 1. i contenuti di odio attraverso <i>content analysis</i> realizzata con sentiment analysis manuale, necessaria per comprendere le eventuali differenze nei contenuti tra dark web e resto di Internet. La necessità di ricorrere ad analisi manuale è dovuta al fatto che la maggior parte degli studi sull'<i>hate speech</i> online negli ultimi anni si è concentrata soprattutto sul Surface Web (Twitter, Facebook, Youtube..), per cui l'utilizzo di algoritmi automatizzati studiati per il surface web porterebbe ad un'analisi parziale, non consentendo di cogliere le caratteristiche peculiari dei contenuti di odio prodotti nel dark web che possono avere caratteristiche differenti rispetto a quelli prodotti nel surface web; 2. i dati geografici relativi alla produzione dei contenuti di odio attraverso analisi degli IP (geolocalizzazione). Specifici algoritmi consentono cioè di verificare:

	<ul style="list-style-type: none"> - quali siano le aree geografiche in cui l'attività nel dark web è maggiore; - se esistono differenze geografiche nel modo di creare <i>hate speech</i> online nel dark web; - le eventuali differenze tra paesi con riferimento ai target più colpiti. <p>3. le caratteristiche dell'autore dell'<i>hate speech</i> rilevabili dai metadati nonostante l'anonimato garantito dalla piattaforma 4chan (pseudoidentità). Specifici algoritmi di profilazione permettono cioè di analizzare i comportamenti degli autori e di individuare quanto sono seguiti, quanto sono prolifici in termini di <i>hate speech</i>, il numero delle loro risposte e la loro localizzazione così da classificarli in differenti tipologie (es. autori che postano poco e hanno poco seguito; autori che postano molto ma non interagiscono; autori che postano molto e interagiscono tanto);</p> <p>4. la rete di condivisione dei contenuti di odio attraverso la <i>network analysis</i> del processo di propagazione dei post all'interno della piattaforma 4chan (analisi del numero di risposte, localizzazione geografica dei post con maggiore engagement; estensione della condivisione (luoghi e persone raggiunte)).</p> <p>5. gli eventuali attacchi (raid) da dark web a Surface web attraverso algoritmi che analizzano i segnali di diffusione dei post su altre piattaforme. Sono stati cioè sviluppati specifici algoritmi di analisi della propagazione dell'odio dal dark web al resto di Internet, per capire come si generano e si coordinano gli attacchi informatici verso particolari profili social su Facebook, Twitter e Youtube.</p> <p>6. Attraverso la creazione complessiva di tale base informativa è possibile anche disporre di una base dati su cui progettare algoritmi predittivi dei raid sul surface web.</p>
<p>Principali Risultati</p>	<p>La metodologia ha consentito di creare una base informativa per l'analisi degli <i>hate speech</i> nel dark web, allo stesso tempo evidenziando le differenze tra contenuti di odio creati nel dark web e contenuti di odio creati nel Surface web. E' stata provata l'esistenza di raid di odio progettati nel dark web verso il Surface web, fatto che dimostra che le strategie di contrasto all'odio online non possono trascurare il dark web quale fonte rilevante degli <i>hate speech</i>. Inoltre, un rilevante risultato della metodologia è aver avviato l'analisi per prevedere questo tipo di raid così da mettere in campo azioni preventive. La creazione di algoritmi di analisi degli attacchi da dark web a piattaforme social come Twitter, Facebook e Youtube ha consentito di colmare sia un vuoto conoscitivo sulla generazione e propagazione dell'odio online, sia un vuoto metodologico nell'analisi degli <i>hate speech</i> online.</p>
<p>Eventuali criticità incontrate e modalità di soluzione</p>	<p>Dal momento che si tratta di uno dei primi progetti che si occupano della dimensione dell'odio online nel <i>dark web</i>, le fasi automatizzate all'interno del progetto non sono maggioritarie rispetto a quelle manuali. Di conseguenza, la necessità di ricorrere in modo significativo a supervisione e analisi umana ha portato ad una limitazione della portata del progetto, se paragonata ai progetti che si occupano di <i>surface web</i> e che analizzano elevati volumi di dati e che sono altamente automatizzati. Ad esempio, le parole chiave con cui avviare il monitoraggio del dark web sono tuttora limitate, perché non esistono vocabolari dei termini di odio usati nel dark web a cui far ricorso. Analogamente, la limitata conoscenza dei comportamenti nel dark web rende difficilmente applicabile il patrimonio di algoritmi di <i>content analysis</i> già disponibile per il <i>Surface web</i>. Tuttavia, progetti pionieristici come questo consentono di iniziare a costruire la base informativa attraverso cui si realizzerà, in modo graduale, la progressiva automazione dei processi di monitoraggio e analisi del dark web, come già accaduto per il <i>surface web</i>.</p>

Nome esperienza /metodologia	React	
Sito web	reactnohate.eu	
Soggetto attuatore/coordinatore	ASSOCIAZIONE ARCI Associazione culturale e di promozione sociale attiva nella difesa dei diritti umani, contro le discriminazioni, per la promozione della cultura della pace e della tolleranza.	
Referente	Nome	Da individuare
	email	progetti@arci.it
Altri soggetti coinvolti	La realizzazione della metodologia è avvenuta nell'ambito di un progetto europeo che ha visto il coinvolgimento di numerosi partner (attori istituzionali e della società civile) in 5 paesi europei: UNAR (IT); CITTALIA (IT); Associazione Carta di Roma (IT); AWO (DE); Lingue de l'enseignement (FR); LDH (FR); Aiksaath (GB); ROTA (GB); SOS Racismo Gipuzkoa (ES); Universitat de Barcelona (ES); UVIC (ES).	
Soggetto finanziatore	Commissione Europea, nell'ambito del Programma REC - RIGHTS, EQUALITY AND CITIZENSHIP PROGRAMME (2014-2020)	
Area geografica di applicazione della metodologia	Europa: Italia, Francia, Germania, Spagna, Regno Unito	
Periodo di implementazione	In corso (conclusione ottobre 2019)	
Social media di riferimento per l'applicazione della metodologia	Le piattaforme web su cui il progetto REACT si focalizza per l'individuazione dell' <i>hate speech</i> online sono rappresentate dai social media e siti web, tuttora in fase di definizione	
Obiettivi e articolazione della metodologia	<p>Il progetto REACT mira a contrastare gli <i>hate speech</i>, i crimini motivati dall'odio e altre forme di intolleranza attraverso il miglioramento dell'alfabetizzazione mediatica tra gli educatori e i giovani e lo sviluppo di una campagna di contro narrativa. A tal fine, REACT prevede di realizzare un approfondito e sistematico monitoraggio quantitativo e qualitativo su una selezione di media, social account e siti web in 5 diversi paesi europei (Italia, Francia, Spagna, Regno Unito, Germania).</p> <p>Inoltre, attraverso il coinvolgimento di stakeholder a livello nazionale ed europeo, REACT mira ad aumentare l'operatività ed il coinvolgimento dei decisori politici per il recepimento delle politiche dell'UE nel campo della prevenzione e della lotta al razzismo, alla xenofobia e ad altre forme di intolleranza.</p> <p>La metodologia si articola sostanzialmente in tre fasi/attività:</p> <ul style="list-style-type: none"> • Fase 1- creazione base informativa/database su cui effettuare le analisi; Più nello specifico, prevede: <ul style="list-style-type: none"> - attività di monitoraggio sistematico di tipo quantitativo e qualitativo degli <i>hate speech</i> (attraverso algoritmi progettati ad hoc) su una selezione di media online, compresi i social media, in Italia, Regno Unito, Francia, Germania e Spagna; - attività di monitoraggio quantitativo e qualitativo per estrarre esempi efficaci di contronarrazione; • Fase 2 - attività di analisi dati/contenuti di odio online; • Fase 3 - produzione output specifici: database con i risultati dell'analisi, percorsi di apprendimento reciproco, seminari, campagna i sensibilizzazione 	
Ambito/oggetto/ Target degli hate speech analizzati		Razzismo (Islamofobia)

<p>Descrizione della metodologia e tecniche di analisi</p>	<p>REACT si caratterizza per la realizzazione di una strategia integrata per il contrasto all'<i>hate speech</i> online, che unisce all'analisi quantitativa e qualitativa dei dati inerenti gli <i>hate speech</i> azioni volte ad utilizzare tali dati per aumentare la consapevolezza sul fenomeno e diffondere strumenti di contrasto. Il coinvolgimento di più paesi consente non solo di effettuare un'analisi comparativa ma anche di implementare e radicare tale strategia in contesti differenti.</p> <p>Per favorire la conoscenza del fenomeno, il progetto svilupperà uno strumento informatico specificatamente rivolto all'analisi dell'islamofobia online. La particolarità di tale strumento informatico risiede nella sua duplice finalità di estrarre attraverso algoritmi sia dati inerenti i contenuti di odio (<i>hate speech</i>), che i contenuti di contrasto all'odio (contronarrativa).</p> <p><i>Attraverso l'analisi desk non è stato possibile reperire ulteriori informazioni sulle tecniche di analisi che verranno utilizzate a parte l'indicazione che verrà effettuata la geolocalizzazione degli hate speech e la network analysis.</i></p> <p>Il database che conterrà i risultati dell'analisi renderà disponibili anche informazioni sui contenuti, la diffusione e la localizzazione dei post online che attivamente contrastano la narrazione islamofobica. Questo aspetto appare particolarmente rilevante, essendo molto più inesplorato rispetto al tema della diffusione dell'<i>hate speech</i> online. Inoltre, la metodologia si caratterizza per l'attenzione posta alla trasferibilità di buone pratiche per l'apprendimento reciproco di tecniche/pratiche di contro narrazione.</p> <p>La metodologia prevede anche la realizzazione:</p> <ul style="list-style-type: none"> - percorsi di apprendimento reciproco e scambio delle migliori pratiche tra stakeholder (insegnanti, operatori giovanili, rappresentanti delle comunità interessate, ricercatori, responsabili politici, rappresentanti delle piattaforme dei social media, rappresentanti della stampa online e delle organizzazioni della società civile) in merito ad azioni positive per promuovere la tolleranza, contrastare gli <i>hate speech</i> e definire meccanismi per facilitare la comunicazione, anche individuando standard europei per effettuare la contro narrativa; - attività di capacity building e formazione incentrate sull'alfabetizzazione mediatica rivolte agli insegnanti, operatori giovanili e direttamente ai giovani, con particolare attenzione a categorie di giovani, anche appartenenti a minoranze. I seminari adotteranno un approccio partecipativo, coinvolgendo attivamente i giovani nella costruzione di esempi di contro-narrazioni; - campagna di diffusione e sensibilizzazione basata sull'utilizzo degli strumenti di contronarrazione realizzati dai giovani che saranno presentati a livello locale e diffusi sui social media e su Internet.
<p>Principali Risultati attesi</p>	<p>Come risultati attesi il progetto REACT intende dare evidenza qualitativa e quantitativa ai discorsi di odio online attraverso un monitoraggio multipaese. Inoltre, si propone di costituire una banca dati di contronarrativa da utilizzare come base per sviluppare ulteriori azioni positive per promuovere la tolleranza, contrastare i discorsi di odio, facilitarne la segnalazione e migliorare la trasparenza della contro narrazione (individuando standard europei). La costruzione di una base conoscitiva relativamente alla contro narrativa e alle caratteristiche di chi la attua (come la localizzazione e la capacità di aver costituito network) consentirà di avere una prima fotografia di tale fenomeno.</p> <p>Inoltre, i dati raccolti serviranno ad implementare azioni di alfabetizzazione mediatica tra i giovani e gli educatori con la finalità di far diventare la contro-narrativa una pratica diffusa. Costituiranno, inoltre, uno degli elementi principali su cui REACT costruirà le successive azioni di elaborazione e scambio di buone pratiche per favorire una capacity building europea sul contrasto all'islamofobia.</p>
<p>Eventuali criticità incontrate e modalità di soluzione</p>	<p>Progetto in corso (nessuna criticità evidenziabile dalla documentazione a disposizione)</p>

Bibliografia

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.

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

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

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.

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.

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

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

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

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

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.

Davidson, T., Warmesley, 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.

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.

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_

Eurobarometer (2016). *Media Pluralism and Democracy*. <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.

Himmelboim 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.

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.

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.

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*.

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

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

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