ONLINE HATE SPEECH E IDENTIFICAZIONE AUTOMATICA (I PARTE)

L'*hate speech* o discorso di incitamento all'odio è considerato l'espressione pubblica, consapevole e intenzionale di odio o ostilità verso individui o collettivi, sia sulla base di criteri razziali, etnici, religiosi o nazionali che sulla base di qualsiasi altro criterio che promuove l'intolleranza e la discriminazione. Il presente contributo intende fornire un aggiornamento tecnico sul fenomeno dell'*hate speech online* che negli ultimi anni dilaga sui *social network* più comuni come Facebook, Twitter, Youtube, Reddit e Instagram. In questo articolo esporremo le problematiche collegate all'identificazione automatica dell' *hate speech* nei testi, sia dal punto di vista umano che dal punto di vista tecnologico, sottolineando la rilevanza degli approcci automatici per valutare un numero elevato di informazioni e monitorare costantemente le forme di *hate speech online*. I principali argomenti di ricerca presentati in questo articolo possono essere divisi in tre aree: "dibattito generale sull'*hate speech* dal un punto di vista normativo italiano ed europeo", "identificazione e classificazione automatica di *hate speech* tramite strategie di apprendimento automatico", e " prevenzione e intervento per ridurre al minimo l'impatto di questi fenomeni nei più giovani ".
In questo numero: 1. Premessa, 2. L'hate speech nei social media, 3. Affrontare l'hate speech online, 4. Rilevare automaticamente l'hate speech nei testi. Nel prossimo numero: 5. Esempi di applicazioni, 6. Conclusioni, Bibliografia.

Vito MONTELEONE, Tenente dell'Arma dei Carabinieri, Ufficiale Addetto all' Ufficio Sviluppo Tecnologico presso il Comando Generale dell'Arma dei Carabinieri – Ruolo Tecnico: Telematica-Informatica.
Davide TAIBI, Primo Ricercatore, Consiglio Nazionale delle Ricerche, Istituto per le Tecnologie Didattiche.

1. Premessa

Il termine *"hate speech"* trova in italiano la sua naturale traduzione in discorso di incitamento all'odio o semplicemente in discorso d'odio e vede in Internet un ambiente fertile in cui proliferare e amplificare i propri effetti. Le difficoltà nel definire l'*hate speech* per mezzo di una definizione univoca, condivisa e internazionalmente accettata, sono svariate e dipendono non soltanto dalla sfera giuridica di una nazione ma ancor prima da quella politica, filosofica e socio-culturale.

In Italia, così come anche a livello europeo, non esiste una definizione giuridicamente vincolante di questo fenomeno. La base da cui partire per avere una definizione di *"hate speech"* quantomeno condivisa a livello europeo venne prodotta dal Consiglio d'Europa nel 1997 grazie a una raccomandazione [1] in cui veniva presentato che i discorsi assoggettabili a hate speech sono tutte quelle *"espressioni che diffondono, incitano, promuovono o giustificano l'odio razziale, la xenofobia, l'antisemitismo o altre forme di minaccia basate sull'intolleranza – inclusa l'intolleranza espressa dal nazionalismo aggressivo e dall'etnocentrismo –, sulla discriminazione e sull'ostilità verso i minori, i migranti e le persone di origine straniera"*.

Appare evidente come il concetto di hate speech sollevi dei quesiti effettivamente complessi: che cosa si intende per odio? Qual è l'effettivo danno arrecato dal discorso d'odio? E, soprattutto, cosa si può o si dovrebbe fare concretamente per contrastare o, ancor meglio, prevenire i discorsi d'odio?
Quest'ultimo quesito ha favorito la nascita di una letteratura parecchio vasta e interdisciplinare, in cui giuristi, filosofi, sociologi, storici e informatici hanno cercato di produrre un contributo personale per cercare di risolvere o, se non altro, arginare il gravoso problema dell' hate speech.

Internet ha contribuito a complicare questo già problematico scenario, poichè ha permesso di riunire persone di ogni religione, credo e nazionalità con un semplice click. I social network come Facebook, Twitter, Youtube, Instagram hanno, non soltanto collegato miliardi di persone [2], ma anche permesso di condividere idee e opinioni con molta più semplicità.

Lo scopo di questo articolo è quello di trattare il problema da un punto di vista tecnico, introducendo una serie di metodologie informatiche attualmente utilizzate che potrebbero supportare l'identificazione automatica dell' hate speech nei testi in modo da valutare un numero elevato di informazioni e monitorare costantemente i contenuti.

A tal proposito, nella Sezione 2 tratteremo la presenza dell' hate speech nei social media, mostrando da un punto di vista statistico quanto questo sia presente sulla rete.
Nella Sezione 3 saranno presentati i mezzi di contrasto adottati dalla Commissione dell'Unione Europea e da organismi interforze, quale l'Osservatorio per la sicurezza contro gli atti discriminatori (Oscad). Nella Sezione 4 saranno introdotte alcune nozioni base sul funzionamento dei sistemi automatici per l'individuazione di hate speech all'interno di un testo e relative problematiche di implementazione. Nella Sezione 5 saranno introdotti dei tool disponibili su Internet per l'individuazione di hate speech nei testi. Infine, nella Sezione 6 presenteremo le nostre conclusioni sul fenomeno e sul suo possibile contrasto anche mediante iniziative formative e culturali diffuse tra i più giovani.

2. L'hate speech nei social media

Oggi i social media rivestono un ruolo di primaria importanza nella diffusione dei messaggi d'odio e in parecchi casi possono addirittura portare ai cosiddetti "crimini d'odio". E' risultato evidente, secondo recenti indagini, l'aumento dei contenuti dei discorsi d'odio online inviati nel 2020, in piena pandemia Covid-19, contro le donne (es: per il loro aspetto fisico - body shaming) e in particolar modo contro le lavoratrici (es: accuse di incompetenza e incapacità), come lo stesso Barometro dell'Odio 2021 di Amnesty International fa presente nel suo report [3]. Il report evidenzia come il sessismo da tastiera non sia l'unica forma di discriminazione e d'odio che si sta abbattendo sui social network e mostra, infatti, come sia presente anche una forma d'odio omobitransfobico, razzista e xenofobo, islamofobo, antisemita, antiziganista e classista.

Anche la mappa dell'intolleranza del 2020 [4] ideata da Vox, l' Osservatorio Italiano sui Diritti mostra come le donne siano purtroppo le più soggette al fenomeno dell' hate speech. Questo documento ha mappato e geolocalizzato ben 1.304.537 tweet contenenti parole considerate particolarmente sensibili, concentrate soprattutto contro alcune categorie, come appunto quella delle donne.

[1] "RECOMMENDATION, No R. 'of the Committee of Ministers to Member States on' Hate Speech." Adopted on 30 (1997)".

3. Affrontare l'hate speech online

Per affrontare le possibili dannose conseguenze dei discorsi d'odio, la Commissione dell'Unione Europea ha emanato nel 2016 un codice di condotta dell'UE per contrastare l'illecito incitamento all'odio [5], in cui i principali colossi dell'informatica, che gestiscono l'accesso alle informazioni in rete e alle reti sociali, sono stati invitati ad analizzare i contenuti pubblicati e rimuovere prontamente quelli che incitano all'odio. I risultati si sono rivelati nel complesso positivi e le società che gestiscono i social network hanno esaminato nel 2020 il 90% dei contenuti segnalati entro 24 ore, rimuovendo il 71% dell'oggetto in questione perché ritenuto un illecito incitamento all'odio.
Tuttavia, come anche la stessa Commissione dell'Unione europea riporta, le piattaforme hanno comunque l'impegno di continuare a migliorare la trasparenza e il feedback ai propri utenti allo scopo di garantire che i contenuti segnalati siano valutati coerentemente e correttamente nel tempo. Raggiungere tale risultato non è semplice e dipende dalla difficoltà nel definire con precisione l'hate speech: difatti, ciò che offende un gruppo di persone può essere perfettamente indifferente per un altro. A tal proposito, nel cercare di agevolare la prevenzione e il contrasto dei reati di matrice discriminatoria, si è istituito nel 2010 nell'ambito del Dipartimento della pubblica sicurezza, l'Osservatorio per la sicurezza contro gli atti discriminatori (Oscad) [6]: strumento operativo interforze (Polizia di Stato e Arma dei Carabinieri).

L'impegno principale dell'Oscad è quello di gestire un sistema di monitoraggio che venga alimentato, oltre che dalle segnalazioni ricevute da vittime, testimoni ed associazioni, anche da quelle inviate di iniziativa dalle forze di polizia o da altre istituzioni.
La prevenzione è sicuramente l'elemento risolutivo nel contrasto di questo fenomeno, ma la quantità di informazioni e di messaggi che viaggiano nella rete rende impossibile affidare un monitoraggio continuo esclusivamente all'intervento umano. Per tale motivo, a nostro avviso, l'impiego di sistemi software automatici capaci di monitorare contenuti online consentirebbe di notificare all' Oscad eventuali post o messaggi online contenenti messaggi d'odio. Questo semplificherebbe la sua operatività e ne alleggerirebbe il carico di lavoro oltre che migliorarne l'efficacia.
In forza a quanto detto saranno adesso presentati una serie di concetti e di metodologie presenti in letteratura scientifica per l'analisi di testi che si prefiggono l'obiettivo di individuare e di classificare messaggi discriminatori e d'odio.

4. Rilevare automaticamente l'hate speech nei testi

Molti paesi hanno sviluppato normative specifiche per prevenire l'hate speech online che influiscono sulla responsabilità delle società riguardo la gestione dei social media. Queste hanno l'impegno di rimuovere i messaggi d'odio entro un tempo stabilito dalla legge. Tale rimozione avviene di norma dopo un'attenta analisi manuale, eseguita spesso da operatori umani. Questo approccio ha, tuttavia, conseguenze dannose negli operatori umani che analizzano i messaggi. Di fatto si è rilevato che analizzare nell'arco di una giornata lavorativa numerosi messaggi negativi ha conseguenze dannose per il benessere psicofisico degli operatori. La svolta è avvenuta, però, nel 2019 quando il noto social network Facebook, per facilitare la gestione, l'analisi e il controllo dei messaggi d'odio, ha annunciato l'uso dell'Intelligenza Artificiale (IA) [7] per svolgere questa attività.
Questa scelta operativa è stata dettata dall'enorme quantità di dati che giornalmente veniva analizzata dagli operatori e dagli ingenti costi in termini di tempo impiegato nell'analisi di quei testi [8].

Per tale motivo negli ultimi decenni nel campo dell'IA e in particolare in quello dell'elaborazione del linguaggio naturale (Natural Language Processing - NLP) [9], [10] sono state perfezionate diverse tecniche per la realizzazione di sistemi di Sentiment Analysis (SA) [9], [10] e, al contempo, per l'identificazione di discorsi d'odio nei testi. Per SA si intende "il processo automatizzato di analisi e di interpretazione dei sentimenti che si celano dietro un testo". Tali tecniche sono tipicamente motivate da un crescente interesse verso le questioni sociali ed etiche, come la lotta contro l'estremismo, la violenza, le fake news etc. Unitamente a tali questioni sociali i ricercatori si sono occupati di affrontare anche quella relativa all' hate speech.

Il problema principale legato all' hate speech riguarda la difficoltà di stabilire se un testo sia dotato o meno di un messaggio d'odio. Operazione non semplice, neppure per gli esseri umani stessi. Ne consegue che l'operazione di rilevamento automatico di hate speech nei testi risulta essere impegnativa. Tuttavia, per affrontare tale problematica, si utilizzano ap-

procci basati su metodi matematici avanzati e specifici modelli di classificazione del testo.
Numerosi ricercatori servendosi dell' IA, hanno deciso di affrontare il problema cominciando con l'annotare i messaggi di hate speech prelevati da vari social network, individuandone le caratteristiche più rappresentative e sviluppando sistemi automatici in grado di riconoscere la presenza di hate speech all'interno di un testo e di classificarla secondo delle categorie prestabilite.
Questi sistemi, però, soffrono di diverse debolezze e ciò è dovuto al fatto che il riconoscimento automatico di testi d'odio è un argomento che è stato trattato relativamente di recente: il contesto linguistico (italiano, inglese, etc.), così come le interpretazioni fortemente soggettive dei testi o l'uso di termini d'odio non troppo specifici e diretti, attualmente hanno impedito di ottenere un sistema pienamente funzionante e ottimale.

L'approccio attualmente più utilizzato per lo sviluppo di sistemi automatici per l'analisi e l'identificazione di testi d'odio come riportato in [11], consiste nello sviluppo di modelli di Machine Learning [12] (ML) per la loro classificazione. Il ML, o in italiano "'apprendimento automatico", è un'applicazione dell' IA capace di fornire ai sistemi la capacità non soltanto di imparare automaticamente a svolgere un compito ma anche a migliorare dall'esperienza senza essere esplicitamente programmato per farlo.
Utilizzando una semplice definizione il ML è la teoria secondo la quale "le macchine dovrebbero essere in grado di imparare e di adattarsi attraverso l'esperienza, per produrre decisioni e risultati affidabili e ripetibili".
Uno dei problemi più annosi in sistemi di questo tipo è il reperimento di dati adeguati per addestrare il modello di ML. Questi dati vengono raggruppati in un insieme definito "dataset" prima di essere elaborati dai modelli di ML. In genere, i dati di addestramento[2] sono raccolti ed etichettati da annotatori umani, cosicché l'algoritmo di ML possa, una volta ricevuti i dati e le annotazioni come input, imparare a identificare e a riconoscere dati mai visti prima.

La procedura di raccolta e annotazione di dati, utilizzati per l'addestramento di modelli di ML per classificare automaticamente testi di hate speech è, ad oggi, una delle più grandi sfide in questo settore. In particolare, identificare e concordare se un testo contenga o meno hate speech è un aspetto parecchio problematico poiché, come detto precedentemente, non esiste una sua definizione universale.
Inoltre, nonostante i social media siano piattaforme contenenti numerosi messaggi di hate speech, spesso le policy di utilizzo e distribuzione dei dati ne impediscono l'utilizzo per terzi fini (tranne Twitter che adotta una politica di utilizzo dei dati più indulgente). Tutto ciò si traduce in un numero davvero esiguo di dataset disponibili sul fenomeno dell' hate speech. In particolare la maggior parte dei dataset su hate speech disponibili online sono in lingua inglese, mentre per la lingua italiana manca un dataset di riferimento da poter usare come baseline[3] per futuri lavori.

Il sistema di riconoscimento automatico per l'hate speech si occupa di analizzare dati testuali che sono generalmente costituiti da parole, frasi, paragrafi o commenti prelevati da diversi social network.
Applicando quanto detto al problema dell' hate speech, quello che si vuole ottenere è un modello che, dato un testo nuovo come input che chiamiamo *T*, questo sia in grado di restituire come output *Vero*, se *T* contiene hate speech o *Falso* altrimenti.
Chiaramente esistono differenti schemi di annotazione e non esiste solamente quello binario in cui sono presenti solo due valori (Vero e Falso), mutuamente esclusivi, per marcare la presenza o l'assenza di un dato fenomeno. Una spiegazione esaustiva sui diversi schemi di annotazione è presente in [13].

Il problema principale con gli algoritmi di apprendimento automatico è che questi non possono lavorare direttamente sui dati grezzi. Abbiamo infatti bisogno di tecniche che si occupino di convertire i dati (testi, immagini, serie temporali ecc.) in caratteristiche numeriche ed elaborabili dalle macchine.
E' proprio grazie a questa trasformazione in caratteristiche numeriche che i modelli di ML possono eseguire tutte le operazioni algebriche necessarie sui dati. Al centro del ML vi è quindi un algoritmo che apprende i pattern[4] presenti nei dati di input, in modo da predire pattern simili in dati che non ha mai analizzato e visto prima; il tutto chiaramente con una certa percentuale di precisione.©

2 In inglese training.

3 Punto di riferimento (dataset di base)

4 Per pattern si intende un particolare insieme di caratteristiche che si ripete secondo una specifica struttura.