Diritti / Giovanni Semeraro /

Analisi semantica e sentiment analysis per l'individuazione di hate speech in rete

http://www.voxdiritti.it/ Giugno 2019

Anche quest'anno il gruppo di ricerca SWAP del Dipartimento di Informatica dell'Università degli Studi di Bari ha dato il proprio significativo contributo alla realizzazione della quarta edizione del Progetto Mappa Italiana dell'Intolleranza.

La nostra piattaforma per l'estrazione e l'analisi di dati sociali rappresenta infatti, sin dalla prima edizione della Mappa, il fulcro tecnologico dell'intero progetto. Quest'anno i nostri algoritmi sono riusciti ad intercettare circa 215.000 Tweet in lingua italiana (ed oltre 60.000 correttamente geo-localizzati sul nostro territorio), distribuiti nei sei cluster di riferimento (Omofobia, Razzismo, Antisemitismo, Sessismo, Disabilità, Islamofobia).

Sul piano metodologico, l'utilizzo di tecniche più sofisticate e precise per l'individuazione di hate speech, ci ha permesso di fornire come output finale del progetto una immagine sempre più precisa relativa alla diffusione dei discorsi d'odio in Rete. Fin dalla prima edizione, infatti, abbiamo enfatizzato la complessità del task di individuazione dei discorsi d'odio, a causa principalmente di due fattori: *(i)* l'ambiguità di alcuni termini del linguaggio che rendono poco efficaci i meccanismi di ricerca e di ritrovamento basati sul semplice *matching* dei termini contenuti nel Tweet con un lessico di riferimento (finocchio, terrorista, ebreo, sono termini di uso comune utilizzati anche in scenari comuni e senza accezione negativa); *(ii)* l'esigenza di contestualizzare e di comprendere l'opinione, positiva neutra o negativa, convogliata dal Tweet. Infatti, termini che possono potenzialmente indicare la presenza di hate speech spesso vengono utilizzati anche senza quel tipo di scopo (pensiamo banalmente a Tweet ironici).

Per questo motivo, l'impianto metodologico implementato nella nostra piattaforma per l'estrazione di contenuti testuali ha dato particolare importanza allo studio e al progressivo miglioramento delle tecniche di *analisi semantica* e di *sentiment analysis*, metodologie a stato dell'arte per risolvere i problemi appena menzionati. Nello specifico, gli algoritmi di analisi semantica adottati nel progetto sono basati su tecniche di comprensione del linguaggio. Tali tecniche ci hanno permesso di *disambiguare* correttamente Tweet potenzialmente innocui, escludendoli dall'analisi, e di includere invece contenuti effettivamente atti a convogliare discorsi d'odio. In generale questo tipo di tecniche sono basate sull'analisi dei termini presenti nel Tweet e del contesto (inteso come i concetti che co-occorrono con i termini potenzialmente intolleranti presenti nel testo) in cui tali termini vengono utilizzati, e restituiscono una predizione del particolare significato (intollerante o meno)

che quella parola assume in quello specifico scenario.

Il ruolo della *sentiment analysis*, invece, ha continuato ad essere di fondamentale importanza per la generazione di output precisi ed efficaci. L'utilizzo di tali tecniche, infatti, è orientato ad associare una polarità (positiva o negativa) al Tweet sulla base dell'accezione del Tweet stesso.

In questo caso, l'utilizzo di tecniche innovative ci ha permesso di etichettare correttamente i discorsi d'odio anche in assenza di un preciso lessico aggressivo (es., il celebre '*aiutiamoli a casa loro*') presente nel testo. Una novità metodologica di quest'anno è rappresentata anche dall'analisi di bi-grammi (sequenze di termini), che ci ha permesso di filtrare correttamente espressioni e locuzioni di uso comune (es. *Porca Put…*) che non denotano la presenza di discorsi d'odio pur contenendo termini del nostro lessico.

Un'ultima nota metodologica riguarda le mappe, costruite utilizzando la tecnica delle "heat map" (tonalità più vicine al rosso denotano una maggiore concentrazione dei contenuti). In merito a questo, è importante sottolineare che l'individuazione delle aree maggiormente caratterizzate dalla produzione di discorsi d'odio non è basata sul semplice conteggio dei Tweet provenienti da quell'area. La metodologia adottata è invece basata su un meccanismo di pesatura che tiene in considerazione altri fattori, come la numerosità media di Tweet provenienti da una specifica area o la diffusione di utenti in quella particolare zona del Paese.

Da Marzo a Maggio gli utenti di Twitter hanno messo a "dura prova" i nostri algoritmi e i nostri server, sempre in ascolto e sempre alla ricerca di nuove sfumature di significato utilizzate dagli utenti per disseminare *hate speech* in Rete.

Nonostante la complessità del task, però, qui a Bari siamo molto soddisfatti della qualità della ricerca e della qualità dell'output prodotto, che si è rivelato essere preciso e presumibilmente fedele alle dinamiche reali che caratterizzano il comportamento degli individui nel nostro Paese.

Come dimostrato dai risultati pubblicati, infatti, il cluster più corposo è rappresentato dai contenuti intolleranti pubblicati verso i migranti (49.000 Tweet circa, 32% sul campione totale, con un incremento di 18 punti percentuali rispetto alla terza edizione del processo), tema quotidianamente dibattuto in Rete e sui giornali.

Tali numeri dimostrano ancora una volta che i temi centrali nel dibattito politico e frequentemente all'ordine del giorno nell'opinione pubblica vengono poi ripresi con altrettanta frequenza ed aggressività dai comuni cittadini che navigano in Rete, ulteriore segnale del fortissimo legame che caratterizza il comportamento on-line degli individui e i loro orientamenti nel mondo reale.