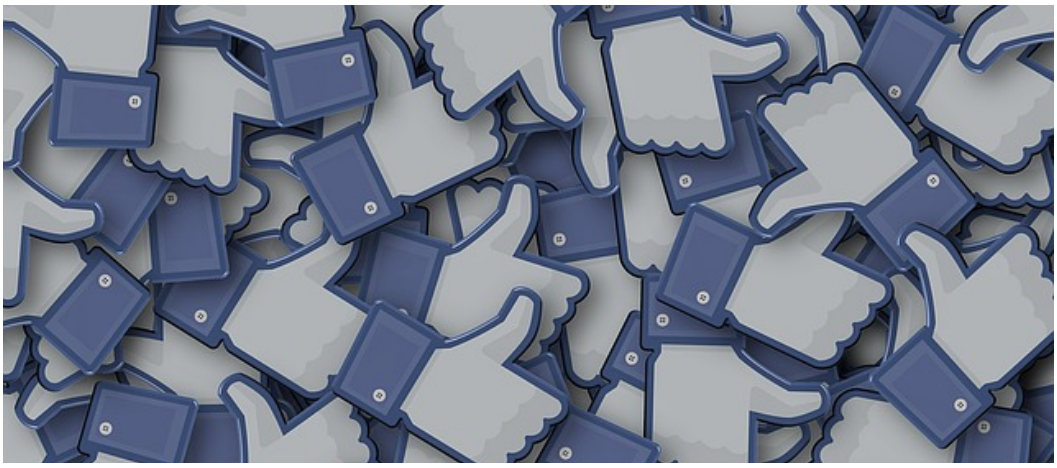
Hate speech. Facebook: su 100 commenti segnalati rimosso meno di un terzo

Pubblicato da: associazionecartadiroma il 13 aprile 2017 in: Comunicati Stampa, Editoriale, In evidenza, Ricerche



Abbiamo ripetuto il test condotto un anno fa e ci siamo confrontati con Facebook. Dalla squadra che analizza i contenuti segnalati ai meccanismi che regolano la valutazione, ecco come funziona

Ci abbiamo riprovato: a un anno di distanza dal nostro primo "esperimento", **abbiamo segnalato a Facebook 100 commenti** che violano apertamente gli standard della comunità in materia di **incitamento all'odio**: **29 sono stati rimossi, 71 sono stati ritenuti idonei a restare online**. Poco meno di un terzo, dunque, è stato riconosciuto dal social network come *hate speech* e di conseguenza cancellato. In media sono trascorse **29 ore tra la segnalazione e la notifica che annuncia l'esito dell'analisi**.

Cosa costituisce *hate speech* per Facebook?

Una definizione condivisa, a livello internazionale, di *hate speech* ancora non esiste: per condurre la nostra sperimentazione siamo allora partiti dagli **standard della comunità di Facebook**:

Gli standard della comunità relativi all'incitamento all'odio di Facebook, così come descritti sulla piattaforma social.

Partendo da questi elementi nel mese di marzo abbiamo esaminato i post pubblicati sulle **pagine Facebook di quotidiani nazionali e locali, esponenti politici, portali di propaganda**, individuando e segnalando 100 commenti che violano secondo noi le *policy* del social network in materia di incitamento all'odio, poiché attaccano in modo apertamente violento una persona o un gruppo di persone su base etnica, razziale o religiosa (**l'elenco completo dei commenti segnalati è qui**).

Nell'ambito del test che avevamo condotto **nel 2016 su 100 ne erano stati rimossi 9, quest'anno, invece 29**: un risultato che, in considerazione della forma e del contenuto inequivocabilmente aggressivi dei commenti segnalati, **non possiamo ritenere sufficiente**. Per comprendere meglio questo esito, ci siamo rivolti nuovamente a Facebook, portando avanti un confronto che ci ha aiutati a capire meglio quali sono **i meccanismi che regolano l'analisi delle segnalazioni**.

Dietro alla valutazione persone e non algoritmi: nessuna parola è vietata

Come l'anno scorso il primo aspetto che abbiamo notato è che **alcuni commenti, a parità di forma e contenuto, sono stati rimossi, altri no**: è il caso, per esempio, di **"buttateli in mare"** – con riferimento a post su migranti e rifugiati. Su tre commenti contenenti la stessa formula solo **uno è stato cancellato**. Come mai?

È essenziale sapere che la verifica delle segnalazioni non è affidata a logaritmi, ma a persone in carne e ossa. Facebook ha costituito un team specifico, il "*Community Operations*", i cui membri – detti anche "esperti di sicurezza" – sono madrelingua in oltre 40 idiomi, tra i quali l'italiano. Per rendere più efficaci le valutazioni, il metodo di lavoro applicato dal personale è quello di **analizzare non solo le parole utilizzate, ma anche il contesto e l'intenzione che traspare dal messaggio**. L'esperto di sicurezza ha accesso, attualmente, anche alle foto e ai video originali ai quali i commenti fanno riferimento, affinché possa inserire i contenuti segnalati in un quadro più ampio: se da un lato l'obiettivo è quello di massimizzare l'efficacia della valutazione delle segnalazioni, dall'altra resta la necessità per l'azienda di minimizzare la quantità di dati personali mostrati al personale stesso.

Abbiamo, inoltre, osservato che tutti i commenti contenenti alcune parole, come "*negro/i*", sono stati rimossi; ci siamo chiesti, allora, se esistesse una *black list* di termini in presenza dei quali il commento venisse sempre cancellato. Dal confronto col social network è emerso che **non esistono parole vietate**: non è accettato l'uso di definizioni dispregiative nei confronti di persone che rientrano in categorie a rischio discriminazione, tuttavia la valutazione avviene caso per caso, poiché lo stesso termine potrebbe, per esempio, essere utilizzato per offendere e attaccare, oppure per sensibilizzare circa lo stesso contrasto all'*hate speech*. In quest'ultimo caso, se la finalità è dichiarata, il contenuto non verrebbe rimosso.

Il fattore tempo

Nel maggio 2016, su impulso della Commissione europea, era stata introdotta una novità: Facebook, Twitter, Microsoft e Youtube avevano sottoscritto un codice di autoregolamentazione col quale si impegnavano a mettere in atto procedure di segnalazione efficaci, a rendere chiare per la comunità le proprie politiche e linee di condotta sui discorsi d'odio e a **verificare le segnalazioni relative a casi di *hate speech* entro 24 ore** (i risultati di un primo lavoro di monitoraggio – a cura della Commissione – sull'applicazione di tale testo sono stati resi noto a fine 2016).

Abbiamo ritenuto rilevante, dunque, tenere conto del tempo trascorso tra la segnalazione e la ricezione della notifica con la quale si viene a conoscenza dell'esito: **la media è di 29 ore**. Solo **in 3 casi la notifica è stata ricevuta entro le 24 ore** (si è trattato, in tutti e 3 i casi, di commenti che non sono stati rimossi), con un **tempo minimo impiegato di 20 ore e 16 minuti**. 13, invece, le notifiche ricevute oltre 36 ore dopo la segnalazione, con un massimo di 46 ore e 6 minuti.

Nel registrare i tempi, abbiamo verificato che **l'ordine di ricezione delle notifiche non corrisponde alla cronologia delle segnalazioni**: questo avviene, come emerso dal confronto con Facebook, perché il *Community Operations Team* nel decidere quali segnalazioni trattatare per prime si basa sul cosiddetto "*real world risk*", ossia il livello di rischio corso dalla vittima o dalle vittime dell'attacco al di fuori della piattaforma.

Un margine di errore ampio o una diversa visione dell'*hate speech*?

"Ringraziamo Carta di Roma per questa sperimentazione relativa al funzionamento dei nostri sistemi, e **studieremo con attenzione le segnalazioni ricevute per migliorare il modo in cui operiamo in questo ambito**. Abbiamo regole chiare contro l'*hate speech* e lavoriamo costantemente per escludere questo tipo di contenuti dalla nostra piattaforma. **Ci impegniamo a lavorare con la nostra comunità di utenti per affrontare al meglio queste questioni**". Questa la dichiarazione rilasciata da Facebook una volta sottoposti all'attenzione del social network i risultati del test condotto.

Dopo aver approfondito meglio il funzionamento della verifica, passando di nuovo in rassegna i commenti rimossi e non rimossi, sono due le considerazioni principali che facciamo: la prima è che presumibilmente **permane un notevole margine d'errore nelle valutazioni**, il quale potrebbe motivare la permanenza di alcuni dei contenuti segnalati; la seconda è che, **nell'operare la distinzione tra discorsi d'odio e commenti tollerabili**, applichiamo probabilmente una logica diversa rispetto a quella dell'azienda. La nostra interpretazione degli standard della comunità di Facebook in materia di incitamento all'odio, la quale ci ha portato a selezionare i 100 commenti che per noi violavano palesemente tale politica, evidentemente non corrisponde in pieno a ciò che Facebook intende per *hate speech*.

Ci auguriamo, dunque, che il percorso intrapreso con l'**Online Civil Courage Iniziative**, iniziativa che vede il social network a confronto con organizzazioni di tutta Europa sul tema del contrasto all'*hate speech*, porti innanzitutto a un dibattito che conduca a una definizione più chiara e condivisa di cosa costituisca un discorso d'odio.

Sappiamo bene che non esiste una bacchetta magica per porre un freno a un fenomeno tanto diffuso quanto complesso: i fattori che lo determinano sono numerosi, così come è necessario agire su più livelli affinché lo si possa contrastare efficacemente. Senza dubbio la tecnologia rappresenta in questo cammino un elemento essenziale sul quale intervenire e l'impegno concreto dei social network avrebbe un peso rilevante.

La necessità di tutelare la libertà di espressione degli utenti, ribadita più volte dalle diverse aziende chiamate a esprimersi sulla questione, non si scontra con l'attività di contrasto ai contenuti d'odio: al contrario, **la lotta alle varie forme di violenza contribuisce alla costruzione di uno spazio che sia sicuro per tutti,** anche per le persone più vulnerabili. Il vero scontro, forse, è quello con gli interessi economici e commerciali delle stesse piattaforme: fino a quando questo non sarà superato, tuttavia, difficilmente le azioni dei social network per il contrasto all'*hate speech* potranno essere davvero efficaci.

Correlati


Social
Media
Facebook
Tweet
LinkedIn
Instagram
Pinterest
Hashtags
Status
Vine