FONDAZIONE BRUNO KESSLER
MAGAZINE

Contro-narrative: cambio di paradigma nella lotta all'odio e alle fake-news

13 Dicembre 2023

Per lungo tempo, si è pensato che l'Intelligenza Artificiale potesse essere utile alla censura di commenti d'odio e fake news; tuttavia, se pensata in una funzione supportiva all'uomo, può essere un efficace strumento per la generazione di contro-narrative, uno strumento più democratico e a suo modo "razionale" per rispondere a queste sfide della contemporaneità.

I social network hanno avuto un impatto profondo nel mondo dell'informazione: i due cambiamenti più notevoli sono nel volume e nella direzione delle informazioni, il cui numero è aumentato a dismisura. Nell'epoca dello User Generated Content (UGC), milioni di contenuti vengono prodotti ogni giorno su Google, YouTube, Instagram, Facebook… e un numero ancor più grande di interazioni (like, dislike, commenti, reply, mentions, reaction…) dà forma a un dibattito estremamente fluido e mutevole. Le informazioni circolano più orizzontali e meno unidirezionali: rispetto a un passato in cui c'erano solo media tradizionali (TV, radio, giornali), controllati da specifici centri di potere, che veicolavano in modo verticale e trasmissivo le informazioni, i social network hanno introdotto un reticolo di informazioni – relativamente – più decentrato, vasto e vario, in cui i consumatori sono diventati utenti attivi.

Questi cambiamenti, senz'altro positivi se visti nella prospettiva di una democratizzazione della società, hanno però un'altra faccia della medaglia: hate speech e fake news. La possibilità degli UGC di essere creati dal nulla, di essere ricondivisi e di guadagnare risonanza crea terreno fertile per la disinformazione. Un contenuto, per quanto smentito da una parte, può essere comunque recepito da un'altra; e anche quando una notizia sia conclamata come falsa al grande pubblico, si può sempre continuare a darle credito. Lavorare sulle emozioni delle persone, fidelizzarle a una propria narrativa e fornire costantemente risposte alternative rende più impermeabili al confronto razionale e allo spirito critico.

Un discorso simile vale per l'hate speech: avere supporto, attraverso dei feedback, su opinioni che si avrebbe timore d'esprimere in altri contesti, assieme al fenomeno delle echo chambers favorito dagli algoritmi, crea spazi di circolo dell'odio che si autoalimentano.

Insomma, la tecnologia, come spesso capita, è causa sia di benefici che di problemi sociali: nel caso dell'IA generativa, per come sviluppata in Digital Society, può essere anche mezzo d'analisi e soluzione di questi ultimi.

Machine Learning e classificazione

Per molto tempo, discorsi d'odio e disinformazione sono stati combattuti tramite algoritmi di machine learning basati su classificazione: l'obiettivo di quest'approccio è rendere sempre più scalabili e automatizzati i processi di filtering delle informazioni.
Naturalmente, un tale approccio solleva problemi di natura pratica ed etica: gli algoritmi sono sempre addestrati su un set di dati finito e le forme di comunicazione dell'odio e della disinformazione possono evolversi nel tempo, rendendo più difficile rintracciare le novità. L'aggiornamento degli algoritmi è un processo lungo, dispendioso e faticoso. Inoltre, anche il contrasto a queste nuove forme può essere problematico: le leggi potrebbero non essere aggiornate per rispondervi e gli algoritmi potrebbero al più segnalare la possibilità della presenza di odio o falsità in un contenuto, senza censurarlo: questo rende il tutto meno automatico.
Accanto ai problemi di natura pratica sorgono quelli di natura etica, filosofica: qual è il confine fra opinione tendenziosa ma legittima e discorso d'odio? Come valutare un contenuto quando si mischiano verità e finzione? Dove sta il punto di equilibrio fra censura e libertà di parola? Delegare alle macchine le decisioni su domande simili, specie nei casi più sfumati, può essere problematico. È per questo che si sta facendo strada un nuovo, rivoluzionario approccio.

"Human-centric" AI e contro-narrative

In FBK viene sviluppata un'IA generativa di supporto agli utenti del web nella lotta all'odio. Per fare un esempio, se l'algoritmo della pagina Facebook di una ONG che si occupa di salvaguardia dei migranti rileva una frase d'odio, una chatbot può prontamente elaborare un testo in risposta a questo commento. Tuttavia, la risposta non è automatica: il testo viene elaborato come consiglio a un operatore, che sulla base di quel testo può costruire la sua risposta. AI ed essere umano si integrano perfettamente, per costruire efficaci contro-narrative ai discorsi d'odio; le risposte suggerite all'AI sono mediamente razionali e rispettose, coerenti con una linea d'immagine dell'organizzazione, tese a far riflettere le persone sulle implicazioni di ciò che dicono. L'opinione "attaccata" non viene censurata, ma passata al vaglio critico di una contro-narrazione.

Per ottenere questi requisiti, i dati d'addestramento sono costruiti ad hoc da ricercatori ed esperti del settore: visto che gli algoritmi non hanno una reale comprensione del testo, ma costruiscono la propria semantica sulla base di una distribuzione probabilistica di parole frequentemente associate ad altre in un dato contesto sintattico, sarebbe rischioso partire da dati "reali".

Per quanto riguarda le fake news le intelligenze artificiali, quando viene dato loro un post da smentire, hanno a disposizione un relativo articolo di debugging, con il quale costruiscono contro-argomentazioni. Una recente linea di ricerca riguarda la capacità di dare alle risposte una tonalità emotiva, ottenuta attraverso la sentiment analysis dei commenti a cui rispondere.

A titolo di esempio, si può considerare una notizia riportata da una testata online durante la pandemia. La stessa notizia, veicolata attraverso i media sociali, ha generato diversi commenti da parte degli utenti online. Nell'immagine qui sotto vengono mostrati due diverse risposte a uno stesso commento. Tali risposte sono state ricavate a partire da suggerimenti di contro-narrativa prodotti in maniera automatica dall'intelligenza artificiale. In particolare, nel primo caso la risposta è "fredda e razionale", mentre nel secondo è "emotivamente tonalizzata".

207

Mi piace Co

Scrivi un commento...

Mario Rossi

Mio zio si è seriamente ammalato 3-4 giorn
in terapia intensiva con Covid in stato avan
iniettano una versione indebolita del virus a
qualcuno che vuole che sia così... io intanto

2 a Mi piace Rispondi

Fredda e razionale

Il principio di immunizzazione dei vac
versione inattiva o indebolita del viru
nel caso dei vaccini anti-covid Pfizer-
di RNA messaggero che fornisce istru
una proteina simile a quella presente
Questa proteina stimola una risposta
i quali sono proteine che possono ric
Pertanto è scorretto sostenere che i v

La strada per l’integrazione uomo – AI è ancora lunga; tuttavia, questo caso di collaborazione fra efficienza algoritmica e sensibilità umana è un ottimo esempio di come le due dimensioni possano andare a braccetto, dando forma a quella più generale “trustworthy & human-centric AI” che il progetto Horizon, nel suo ultimo aggiornamento datato 2021, definisce come il futuro orizzonte della ricerca.

LINK

https://magazine.fbk.eu/it/news/contro-narrative-cambio-di-paradigma-nella-lotta-allodio-e-alle-fake-news/

TAG

- #fbkdictionary
- #Intelligenza artificiale
- #societàdigitale

AUTORI

- Lorenzo Perin