FONDAZIONE BRUNO KESSLER
MAGAZINE

Generare contro-narrazioni per combattere i discorsi d’odio online: dati e strategie

13 Settembre 2020

Quanto è diffuso il fenomeno "hate speech" online? Quali effetti produce? Come possiamo affrontarlo? Qual è il contributo dell'intelligenza artificiale? Marco Guerini e altri ricercatori in FBK stanno portando avanti questa conoscenza di frontiera

I discorsi d’odio, dall’inglese Hate Speech, si riferiscono a “espressioni che attaccano o sminuiscono, che incitano alla violenza o all’odio contro gruppi con caratteristiche specifiche quali religione, etnia, orientamento sessuale, genere o altro”.

Con la rapida crescita delle piattaforme social, la violenza verbale può diffondersi rapidamente ed è difficile da monitorare.

L’hate speech ha conseguenze sulla vita reale: può portare a depressione o suicidio, promuovere l’uso della violenza, incoraggiare la discriminazione e aumentare le divisioni sociali. Uno sforzo senza precedenti per dare risposte adeguate in termini di leggi e politiche ai contenuti d’odio sulle piattaforme social.

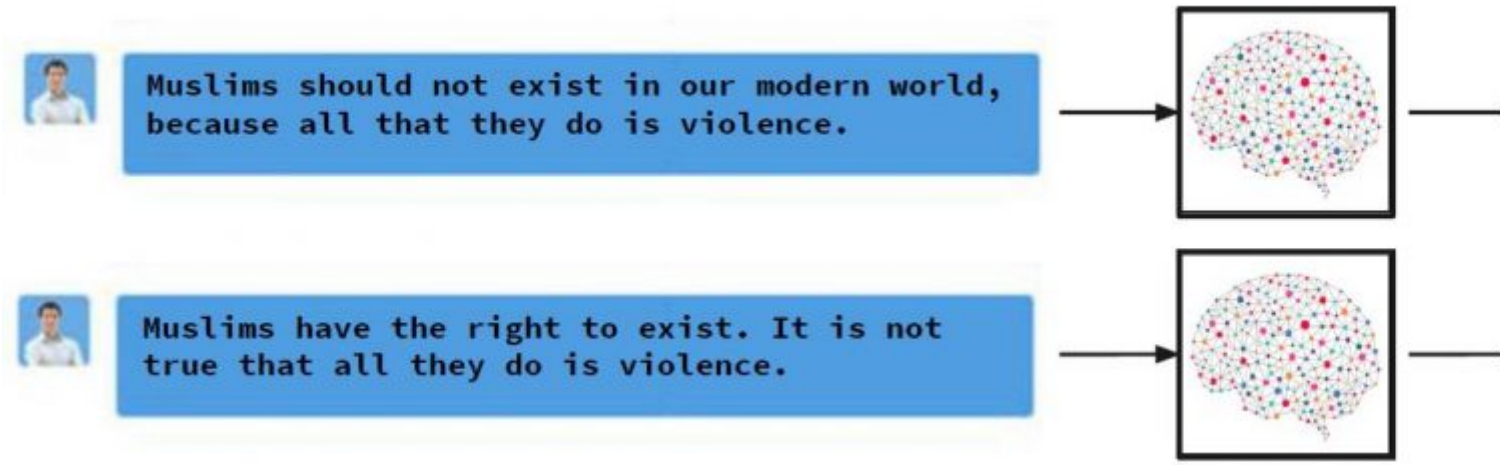
Contrastare i discorsi d’odio

Le politiche standard si basano su strategie di identificazione e sanzione.

? Eliminazione del contenuto (1° passaggio: identificazione del linguaggio violento; 2° passaggio: rimozione del contenuto)
? Sospensione dell’utente
? Shadow banning, cioè l’interdizione

Contributo dell’intelligenza artificiale (IA)

Fino ad ora i sistemi sono in grado di classificare il contenuto, distinguendo il discorso violento da altri.



Questi approcci hanno alcune limitazioni perché possono essere accusati di censura e overblocking, cioè blocco eccessivo.

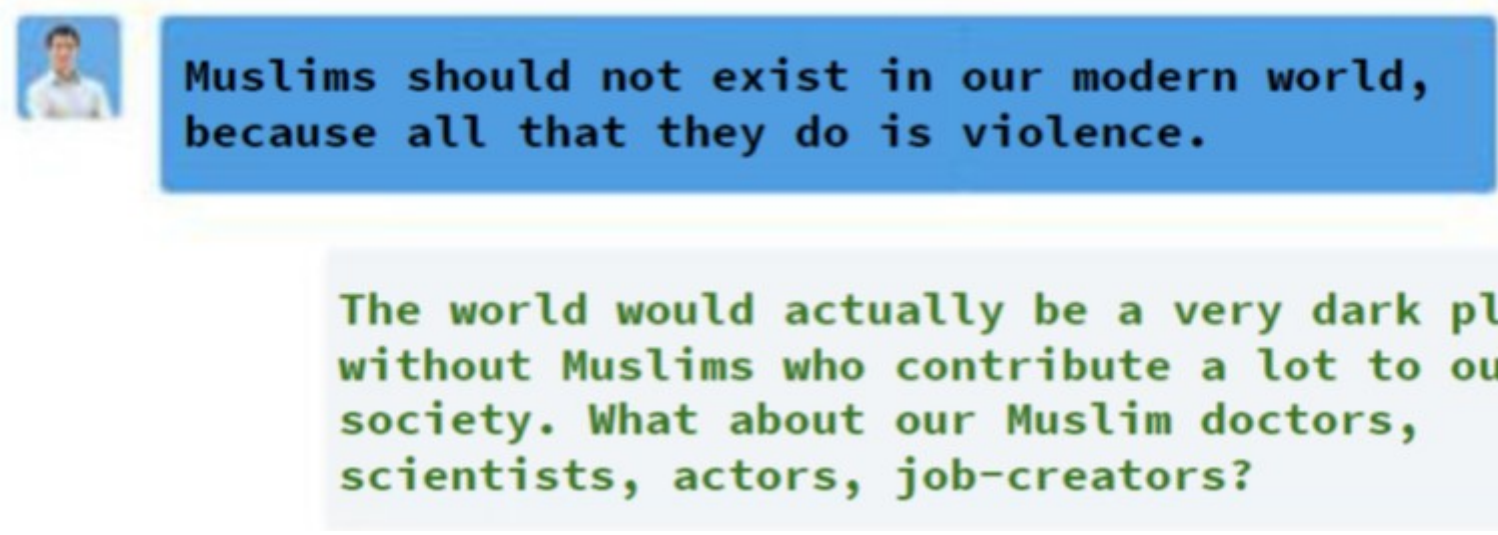
Possono ostacolare la libertà di parola. Non possono essere applicati a discorsi pericolosi – in altre parole contenuti che suscitano odio e divisione ma non rientrano in una definizione formale di discorso d'odio.

Una strategia alternativa

Possiamo controbattere il contenuto d'odio con contro-narrative (CN): intervento diretto nella discussione per resistere ai messaggi di odio usando una risposta testuale non aggressiva che offra feedback attraverso argomenti fattuali.

Questo approccio presenta numerosi **vantaggi**:

- Salvaguardare il diritto alla libertà di parola
- Contrastare gli stereotipi con prove credibili
- Modificare i punti di vista degli hater, cioè coloro che fomentano odio, e degli spettatori
- Incoraggiare la comprensione reciproca
- Aiutare a ridimensionare la conversazione



Ci sono anche importanti **limitazioni**:

- Il solo intervento manuale non è modulabile.
- La quantità di odio generata ogni giorno è semplicemente troppa.

Ci sono ONG che addestrano volontari/operatori ad intervenire con contronarrative.

Quindi l'attività umana assistita dal computer può aiutarli grazie a:

- Strumenti che aiutino gli operatori delle ONG nella lotta contro l'hate speech.
- Automatizzazione parziale dell'elaborazione di contronarrative attraverso suggerimenti di Natural Language Generation (NLG).
- Drastica riduzione del tempo necessario per comporre una contronarrazione.

Algoritmi per la generazione di CN:
1. Recupero di informazioni
2. Generazione mediante reti neurali
3. Generazione mediante modelli linguistici generici pre-addestrati
4. Generazione mediante Conoscenza Esterna

Uno sguardo al futuro

Aspetti **multilingue**

- Ogni lingua/cultura ha le sue sfumature e i suoi pregiudizi.
- Potremmo non avere esperti per ogni lingua
- L'apprendimento inter-lingua può aiutare la portabilità della conoscenza

Aspetti **multi-target**

- I target del discorso d'odio che possono beneficiare del nostro approccio sono molteplici.
- L'apprendimento inter-target può favorire la portabilità della conoscenza

Fake News e odio

- Qual è la relazione tra Fake News e Hate?
- Possiamo usare i nostri strumenti e la nostra base teorica anche per combattere le notizie false?

Per saperne di più: PRESENTAZIONE FINALE DI TENURE TRACK

LINK

https://magazine.fbk.eu/it/news/generare-contro-narrazioni-per-combattere-i-discorsi-dodio-online-dati-e-strategie/

TAG

- #artificial intelligence
- #counter narratives
- #fake news
- #hatespeech
- #human language technology
- #intelligenzaartificiale
- #natural language processing
- #research

MEDIA COLLEGATI

- Tenure Track Discussion : https://magazine.fbk.eu/wp-content/uploads/2020/07/TENURE_FINAL_PRESENTATION.pdf

AUTORI

- Giancarlo Sciascia