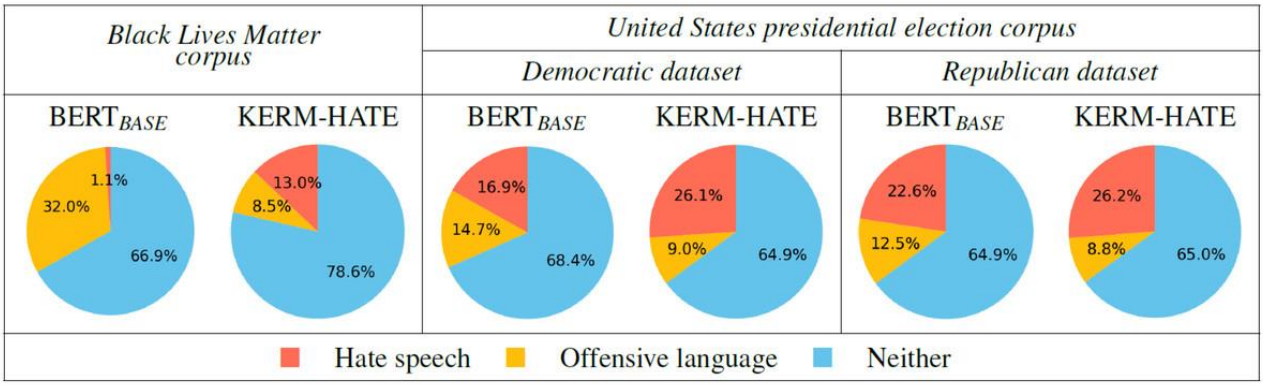
TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Comunicato Stampa

SAFER INTERNET DAY

HATE SPEECH, IL RICOSCIMENTO LINGUISTICO AUTOMATICO ASSORBE IL PREGIUDIZIO

Ricercatori di "Tor Vergata" studiano nuovi metodi per evitare la censura non intenzionale



Roma 08.02.2022 – Nella giornata promossa dalla Commissione europea per rendere il web un luogo più sicuro, l' 8 febbraio è il "Safer Internet Day, ricordiamo che l' hate speech è tra i principali temi della sicurezza in rete, insieme alla tutela della privacy. L'espressione inglese viene tradotta con discorsi dell'odio e si riferisce a quell'insieme di commenti offensivi, contenuti violenti e insulti veicolati attraverso le piattaforme digitali.

I **sistemi HSR - Hate Speech Recognizer** hanno l'obiettivo di riconoscere i post o i commenti che sono indentificati come pieni di odio al fine di cancellarli o segnalarli. Ma se da una parte I **riconoscitori automatici di messaggi d'incitamento all'odio**

possono essere la panacea per **contenere l'odio** sui social media, dall'altra possono portare con sé **una censura non intenzionale** basata sul pregiudizio, che impedirebbe alle persone di esprimere le loro idee.

Nello studio dal titolo Syntax and prejudice: ethically-charged biases of a syntax-based hate speech recognizer unveiled, appena pubblicato sulla rivista internazionale open access "Peerj", i ricercatori dell'Università di Roma "Tor Vergata" hanno messo a punto un sistema innovativo chiamato **"KERM-HATE"** basato su Intelligenza Artificiale che può spiegare le sue decisioni.

I sistemi HSR precedenti a KERM-HATE sono basati su parole e su reti neurali particolari chiamate "trasformers". L'ipotesi è che in questi sistemi HSR le parole tendono ad avere un ruolo predominante. Dal momento che le parole nascondono il pregiudizio, questi sistemi in fase di addestramento assorbono facilmente il pregiudizio dai post e commenti da cui apprendono. Il risultato? **Una nuova forma di censura guidata dal pregiudizio.**

«In questo lavoro – spiega Michele Mastromattei, **studente di dottorato della Scuola nazionale di dottorato in Intelligenza Artificiale** - abbiamo ipotizzato che modelli basati sulla **sintassi**, ovvero la **struttura delle frasi,** potessero ridurre l'effetto del pregiudizio nelle decisioni degli HSR. Abbiamo dunque realizzato il sistema **"KERM-HATE"**, che oltre alla parola utilizza la **struttura grammaticale delle frasi** per il riconoscimento di messaggi d'odio. Come gli altri sistemi continua Mastromattei - anche il nostro sistema HSR apprende da un grande insieme di post o commenti forniti come esempio di addestramento. Tale insieme viene chiamato **corpus di addestramento** e, per valutare i sistemi, esistono dei corpus di addestramento condivisi. Su tali corpus, KERM-HATE ha delle prestazioni di riconoscimento decisamente superiori rispetto ai modelli basati sulle reti neurali "transformers" come BERT, RoBERTa e XLNet».

Con estrema sorpresa, e contrariamente alla loro ipotesi iniziale, i ricercatori hanno osservato che KERM-HATE manteneva il suo pregiudizio dal momento che questo veniva assorbito dal corpus di addestramento. «Sfruttando la capacità di KERM-HATE di spiegare le sue decisioni, l'analisi qualitativa fatta su dataset basati su avvenimenti storici attuali dal contenuto divisivo, come ad esmepio il movimento Black Lives Matter e le elezioni del Presidente degli Stati Uniti, mostra come anche KERM-HATE assorba il pregiudizio a partire dall'insieme di messaggi su cui apprende», sottolinea Fabio Massimo Zanzotto, **docente di Natural Language Processing presso l'Università degli Studi Di Roma "Tor Vergata", Dipartimento Ingegneria**

dell'Impresa. «Con la sua capacità di auto-spiegarsi, KERM-HATE ci ha insegnato che non basta soltanto lavorare sulle **prestazioni finali dei riconoscitori automatici di messaggi d'odio** ma occorre anche investire per **calibrarli in modo che riescano a prendere le loro decisioni senza pregiudizio**. Il nostro prossimo lavoro di ricerca – conclude Zanzotto - sarà capire come far prendere a questi sistemi le decisioni senza pregiudizio, segnalando solo i discorsi che effettivamente possono essere riconosciuti e identificati come "incitanti all'odio"».

Ufficio Stampa di Ateneo
Università Roma "Tor Vergata"
06.72592709 -2059 -3314
ufficio.stampa@uniroma2.it
Pamela Pergolini cell. 320.4375681