Creazione di un set di dati di riferimento per il rilevamento robusto dei discorsi d'odio in tedesco

Sintesi

Autori:
Gerold Schneider, Janis Goldzycher Università di Zurigo

16 aprile 2024

I modelli di rilevamento dei discorsi di odio funzionano bene solo se sono buoni i dati con cui vengono addestrati. I set di dati provenienti dai social media presentano lacune e pregiudizi sistematici, che generano modelli inaffidabili con confini decisionali semplicistici. I set di dati antagonistici, raccolti sfruttando le debolezze dei modelli, promettono di risolvere questo problema. Tuttavia, la raccolta dei dati antagonistici può essere lenta e costosa e i singoli annotatori hanno una creatività limitata. In questo progetto introduciamo il GAHD (German Adversarial Hate speech Dataset), un nuovo set tedesco di dati antagonistici sul discorso di odio che comprende circa 11 mila esempi. Durante la raccolta dei dati, esploriamo nuove strategie per sostenere gli annotatori, per creare esempi antagonistici più diversificati in modo più efficiente ed esaminiamo manualmente le discrepanze tra gli annotatori per ciascuna strategia. I nostri esperimenti dimostrano che il set di dati risultante è impegnativo anche per i modelli di rilevamento dei discorsi di odio più avanzati e che l'addestramento con il GAHD migliora chiaramente la solidità del modello. Inoltre, abbiamo scoperto che si ottiene il massimo vantaggio combinando più strategie di supporto. Il GAHD è disponibile al pubblico all'indirizzo https://github.com/jagol/gahd.