Responsive classifiers against hate speech in low-resource settings

Fact Sheet

Project Information

Respond2Hate

Grant agreement ID: 101100870

DOI

10.3030/101100870

Funded under
European Research Council (ERC)

Total cost
€ 0,00

EU contribution
€ 150 000,00

Coordinated by
LUDWIG-MAXIMILIANS-UNIVERSITAET MUENCHEN
Germany

Start date
1 November 2023

End date
30 April 2025

Objective

"Hate speech is a worldwide phenomenon that is increasingly pervading online spaces, creating an unsafe environment for users. While tech companies address this problem by server-side filtering using machine learning models trained on large datasets, these automatic methods cannot be applied to most languages due to lack of available training data.

Based on recent results of the PI's ERC project on multilingual representation models in low-resource settings, Respond2Hate aims at developing a pilot browser extension that allows users to locally remove hateful content from their social media feeds themselves, without having to rely on the support of tech companies."
Since hate speech is highly dependent on cultural context, responsive classifiers are needed that adapt to the individual environment. Commercial efforts focus on large-scale, general-purpose models which are often burdened with representation and bias problems, and therefore cope poorly with swiftly changing targets or information shift between regional contexts. In contrast, we seek to develop lightweight, adaptive models that require only a small dataset for initial fine-tuning by continuously enhancing model capabilities over time.

This is achieved by applying state-of-the-art Natural Language Processing (NLP) and deep learning techniques for pre-trained language models like low-resource transfer of hate speech representations from high-resource languages and few-shot learning based on limited user feedback. We have already successfully applied these methods in low-resource multilingual settings, and will now validate their use for hate speech filtering.

By making hate speech detection and reduction available in ""low-resource"" countries with little representation in current training datasets, which are currently not served well by governments, industry and NGOs, Respond2Hate will empower users to self-control their exposure to hate speech, fostering a healthier and safer online environment."

Fields of science

social sciences › political sciences › political policies › civil society › nongovernmental organizations
natural sciences › computer and information sciences › data science › natural language processing
natural sciences › computer and information sciences › artificial intelligence › machine learning › deep learning

Keywords

Respond2Hate

Programme(s)

HORIZON.1.1 - European Research Council (ERC)

Topic(s)
ERC-2022-POC2 - ERC PROOF OF CONCEPT GRANTS

Call for proposal

ERC-2022-POC2

See other projects for this call

Funding Scheme

HORIZON-ERC-POC - HORIZON ERC Proof of Concept Grants

Coordinator

LUDWIG-MAXIMILIANS-UNIVERSITAET MUENCHEN

Net EU contribution

€ 150 000,00

Address

Geschwister scholl platz 1
80539 Muenchen
Germany

Region
Bayern > Oberbayern > München, Kreisfreie Stadt

Activity type

Higher or Secondary Education Establishments

Links

Contact the organisation Website
Participation in EU R&I programmes HORIZON collaboration network

Other funding

€ 0,00

EC signature date 16 June 2023
Last update: 24 July 2023

Permalink: https://cordis.europa.eu/project/id/101100870