Hateless

Detect. Educate. Protect.

Multilingual Hate Speech Detection

Co-funded by the Erasmus+ Programme of the European Union

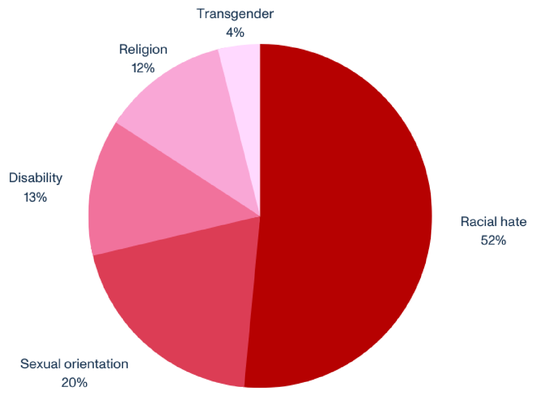
Syrine Maaref - Ibtissem Moussa - Dorra Trabelsi
Zeineb Dghais - Ghalia Bellalouna
Generative AI Hackathon - Team I

The Problem

73% of youth exposed to cyberbullying

540M+ hateful contents reported (2024)

Detection alone is NOT enough



The most common form of online hate

Problem Statement:

How can we create a fast, accurate, and socially responsible system that not only detects hate speech but also educates and supports?

Our Solution

Web Interface
- API Testing for developers
- Educational resources (English, French, Italian)
- Awareness Articles
- **Psych Chatbot Link**

Open REST API
- Response in under 1s
- 100+ Languages
- Docker + App Engine Deployment
- **Easy Integration**

Browser Extension (1Mo)
- Real-Time Detection
- Chrome / Firefox / Brave
- Zero Data Stored
- **Privacy-First**

3-Pillar Ecosystem

Protection + Education + Support

Innovation

Dual-Model Architecture

Input Text
"Your message here"

Model 1
Sentiment Analysis
twitter-xlm-roberta
Emotional context

Model 2
Toxicity Detection
multilingual-toxic
Multi-label

Combined Decision
40% Fewer False Positives

Model Selection Rationale

[1] twitter-xlm-roberta:
- Multilingual (100+ languages), all major platforms
- Trained on 198M real Twitter messages
- Designed for detailed sentiment analysis

[2] multilingual-toxic-xlm:
- Multilingual (100+ languages)
- Specialized for hate speech
- Multi-label: toxic, threat, insult

Combined Model Performance: Accuracy

Obvious hate speech 95–98%

Subtle toxicity 75–85%

False positives 5–10%

Multilingual detection 80–90%

Dual-model approach achieves 40% reduction in false positives

Social Impact

Social Impact

Youth & Students Real-time protection on social media

Victims Direct link to psychological support

Developers Open API, 5-minute integration

Companies Automated moderation saves costs

Making the Internet Safer.
One Sentence at a Time.

Perspectives

- Deploy to new platforms (mobile, messaging apps)
- Improve contextual detection
- Expand into more languages and arabic dialects

References

[1] Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). *XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond*. LREC, 258-266.

[2] Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2025). *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection*. ACL, 1667-1682.

Try our Website! You can also install the extension on your browser