WATCHED

Funded by the European Union

UK Research and Innovation

Usage/License

Code available under the **Apache 2.0 License** via the GitHub repository (per standard open-source practice described in the repo).

Download

The code can be accessed and downloaded from its GitHub repository: https://github.com/HYBRIDS-MSCA/watched

Researchers

- Paloma Piot
- Diego Sánchez Lamas
- Javier Parapar.

Paloma Piot and Javier are affiliated with the Information Retrieval Lab at the University of A Coruña; Diego Sánchez Lamas is an independent researcher.

Summary

Online harms, especially hate speech, threaten user safety and trust on digital platforms. WATCHED is an AI chatbot designed to assist content moderators in detecting and addressing hate speech. It combines Large Language Models with specialised tools to compare posts with real examples, flag harmful content using a BERT-based classifier, interpret slang, generate reasoning for its decisions, and reference platform guidelines. By combining detection with explainability, WATCHED supports informed moderation decisions. Experimental results show it outperforms previous methods, achieving a macro F1 score of 0.91.

What We Offer

- Open-source code and a working moderation chatbot.
- Explainable hate speech detection tools.
- Modular AI agent system extendable with new components.
- Evaluation setup demonstrating state-of-the-art performance.

Key Features

- AI agent architecture combining LLMs with specialised analysis tools.
- Explainable moderation with guideline-based justifications.
- BERT-based hate speech detection with high precision.
- Slang-aware processing to contextualise informal language.
- Moderator-oriented design with strong performance (macro F1: 0.91).

Collaboration Objectives

- Content Moderation Teams: Integrate WATCHED into real-world workflows to reduce online harms.
- NLP Researchers: Extend and improve models for slang handling, reasoning, and explainable moderation.
- Policy and Governance Experts: Collaborate to align AI explanations with platform guidelines.
- AI & Human Oversight Designers: Develop hybrid systems that enhance trust and reliability in automated moderation.

hybrids

HYBRIDS Website
www.hybridsproject.eu

Get in Touch
paloma.piot@udc.es