# HatePRISM: Policies, Platforms, and Research Integration Advancing NLP for Hate Speech Proactive Mitigation

**Naquee Rizwan, Seid Muhie Yimam, Daryna Dementieva, Florian Skupin, Tim Fischer, Daniil Moskovskiy, Aarushi Ajay Borkar, Robert Geislinger, Punyajoy Saha, Sarthak Roy, Martin Semmann, Alexander Panchenko, Chris Biemann, and Animesh Mukherjee**

IIT Kharagpur, University of Hamburg, Technical University of Munich,
Bucerius Law School, Skolkovo Institute of Science and Technology, Artificial Intelligence Research Institute

ACL 2025
VIENNA
JULY 27 - AUGUST 1

## Hate Speech Mitigation

**1. Classification:** (i) Binary and (ii) Fine-grained

**2. Proactive Mitigation**

**a. Text Detoxification** **b. Counterspeech Generation**

## Summary of Survey

comprehensive evaluation

- 14 countries, 24 questions
- 14 platforms, 30 questions
- 38 papers, 20 languages, 34 questions

first work that

- presents survey at the **intersection** of **legal expert** and well qualified **domain researchers**
- questions **regulations** of **countries**, **policies** of **social media platforms** and **research datasets** in a holistic manner
- provides detailed **quantitative** as well as **qualitative** analysis of the questions asked; in the light of **content moderation**

**[24 questions asked]**
**14 considered countries**

**[30 questions asked]**
**14 considered platforms**

**Policies, Platforms** and **Research Integration**

**20 languages covered**

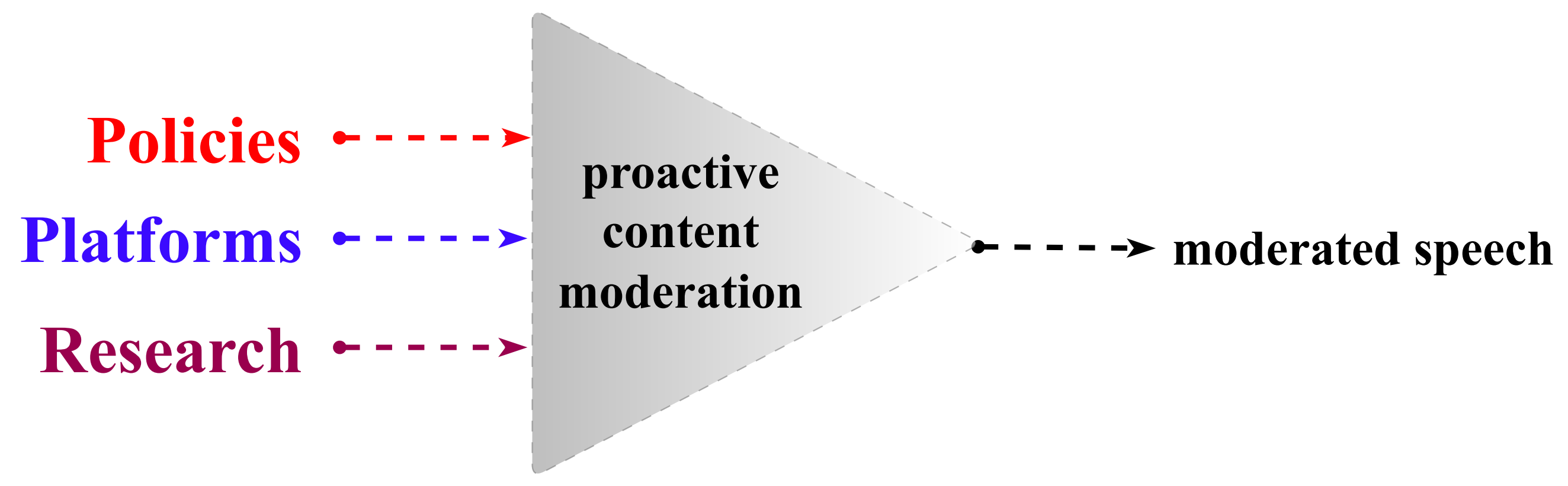
| | |
|---|---|
| Albanian | German |
| Amharic | Hindi |
| Arabic | Hinglish |
| Bengali | Italian |
| Chinese | Korean |
| Croatian | Polish |
| Danish | Portuguese |
| Dutch | Roman Urdu |
| English | Russian |
| French | Spanish |

**[34 questions asked]**
**38 research dataset papers considered**

**labels taxonomy in explored research datasets**

hate
offensive
harmful
sexism
SUD
homophobia
insult
abusive
cyberbullying
fearful
disrespectful
aggressive
incomprehensible
extremism
racism
defamation
irony
lookism
stereotype
blackmail
body shame
curse
exclusion
call-for-actions

## What is HatePRISM?



## Results & Insights

### Country Regulations

**93%** regulate hate speech

**86%** define hate speech officially

**USA** the only country tolerating hate speech

**43%** define online hate speech

**21%** encourage counterspeech/detoxification

types of punishment for hate speech crimes

**Ethiopia | Ukraine | USA** countries **not** having social media specific regulations implemented

**29%** have social or community service as punishment

### Social Media Platforms

**79%** have community guidelines

**93%** have age limit for account creation

platforms **without** hate speech definition

**79%** have regulations language updated per user's location

**79%** have dedicated employees for moderation

**57%** verify the mobile number or identity of users

**64%** provide data API access for research

**64%** encourage counter-speech or detoxification

### Dataset Research Papers

**annotation details**

| | |
|---|---|
| paid annotation 21% | language expertise 58% |
| religion mentioned 3% | race mentioned 16% |

**16%** mention alignment with countries' regulations

**8%** mention alignment with data source's' regulations

**42%** perform pilot annotation

**68%** had more than three annotators

**data sources**

Instagram, Youtube, Facebook, X, News Papers, Weibo, Reddit, Ask.fm, Gab, Rheinische Post, WhatsApp, NAVER, VK

## Categories of Questionnaire

### Country Regulations

- basic regulation queries
- generic hate speech queries
- hate speech definition
- hate speech punishment
- online hate speech queries
- online hate speech specific punishment
- moderation of social media platforms
- preventive measures and encouragements to mitigate online hate speech

Countries were chosen based on the team's familiarity and high prevalence of hate occurrences.

### Social Media Platforms

- general information
- platform access and verification
- transparency
- hate speech definition and queries
- content moderation
- basic regulations queries
- preventive measures and encouragements to mitigate online hate speech

Globally popular platforms with high monthly active users were prioritized.

Regionally popular platforms were also taken into account, with a focus on those that the research team is familiar with, to ensure a comprehensive and contextually relevant approach.

### Dataset Research Papers

- annotator details
- label details
- dataset details
- hate speech definition and alignment
- annotation details

Research dataset papers were selected based on popularity, with a focus on reputable venues like ACL, EMNLP, and relevant workshops such as WOAH.

The selection also ensured to include as many languages as possible, covering a wide variety of label types.

Additionally, datasets for low-resource languages that are less well-known or not published in prominent venues were also taken into account.

### KEY TAKEAWAYS

1. **Lack of consensus** among research, government regulations and social media platform policies.
2. Most NLP research **do not** align with platform or regulatory guidelines.
3. Many studies **do not** explore proactive measures in **operational settings**.
4. Social media platforms have **policy inconsistencies** - A fifth of the platforms fail to adapt hate definitions to local languages & cultures.
5. **Banning rather than proactive mitigation** is typically focussed upon by Social Media platforms.

### RECOMMENDATIONS

1. **Alignment:** Increase collaboration of research with government regulations and social media platform policies.
2. **Promote Proactive Mitigation Strategies:** Thoughtful combination of text detoxification, counterspeech generation and other proactive measures.
3. **Widely Accepted Taxonomy and Definition of Labels:** Brewed at the intersection of government regulations and social media platform policies.

Please refer to our paper for comprehensive details regarding the survey and the complete list of questionnaires.

**ThankYou**