TRAC - 2024

Analyzing and Mitigating Hate Speech: A Comprehensive Approach Towards Building a Better Community

HOUSE OF COMPUTING & DATA SCIENCE

Dr. Seid Muhie Yimam
HCDS, UHH

Disclaimers:

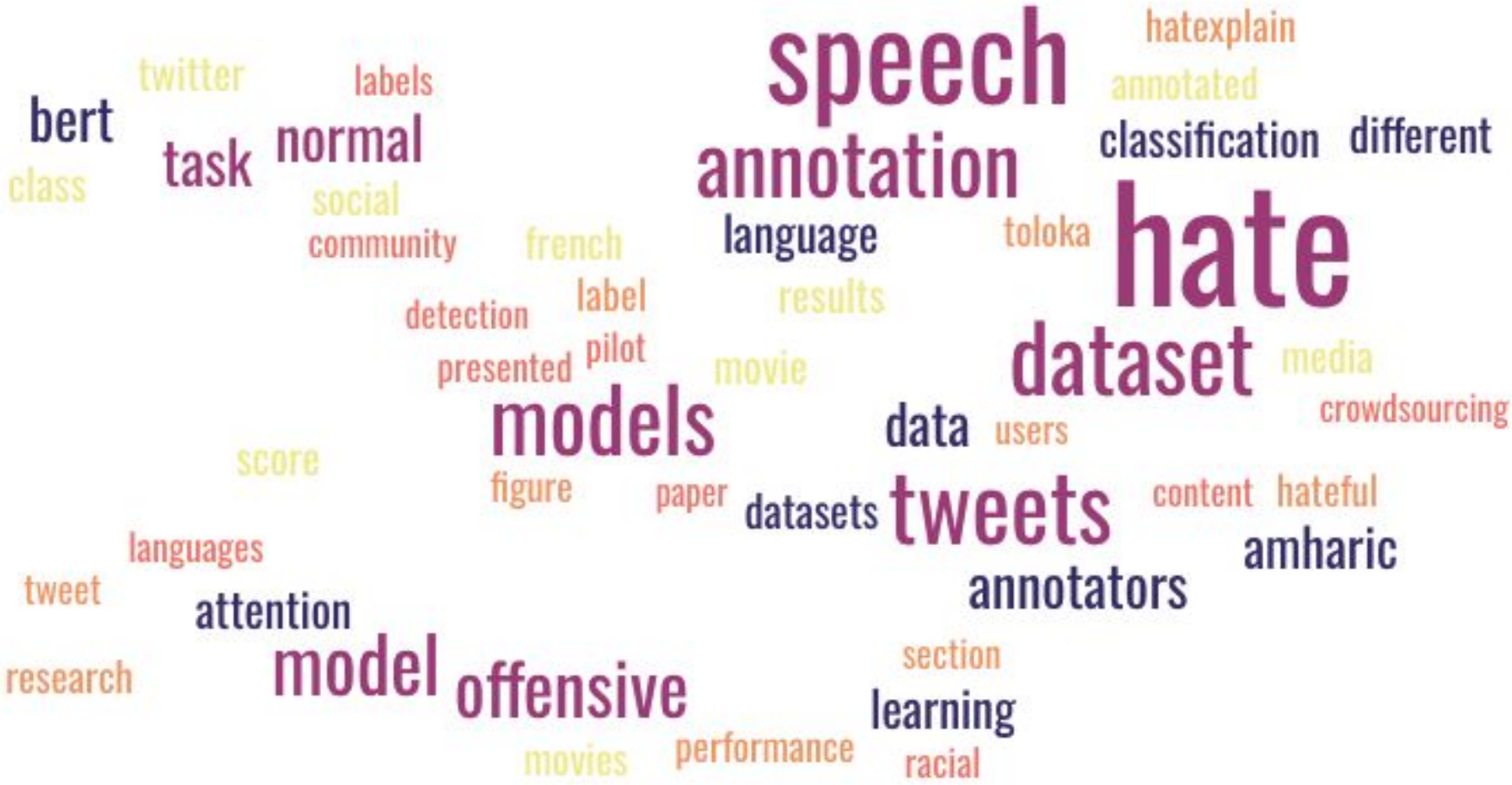
1. The presentation might contain material that you may find **offensive** or **hateful**. However, this cannot be avoided due to the nature of the work!
2. Results/recommendations are based on works in progress, do not cite!

Topics

- Our works
- Demarcation
- Polarization
- Digital peacebuilding

Our works

- Detecting Hate Speech in Amharic Using Multimodal Analysis of Social Media Memes (2024)
- Exploring Boundaries and Intensities in Offensive and Hate Speech (2024)
- Exploring Amharic Hate Speech Data Collection and Classification Approaches (2023)
- The 5Js in Ethiopia: Amharic Hate Speech Data Annotation Using Toloka Crowdsourcing Platform (2022)
- HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection (2021)
- How Hateful are Movies? A Study and Prediction on Movie Subtitles
- ...


speech
hatexplain
twitter
labels
annotated
bert
normal
classification
different
task
annotation
class
social
hate
language
toloka
community
french
label
results
detection
dataset
media
pilot
presented
movie
models
crowdsourcing
data
users
score
figure
paper
tweets
content
hateful
datasets
languages
amharic
annotators
tweet
attention
section
model
offensive
research
learning
movies
performance
racial

Our focuses - (were)

- Low-resource language
- Text based
  - Data collection and sampling
  - Annotation tools
  - Model building (classification + regression – this workshop)
- Multimodal
  - Meme
  - Subtitles
  -

On going research (not yet published …)

- Three-way demarcation
- Polarization, right-wing extremism, neo-fascism ideologies
- Peacebuilding

Demarked: A Strategy for Enhanced Abusive Speech Moderation through Counter Speech, Detoxification, and Message Management

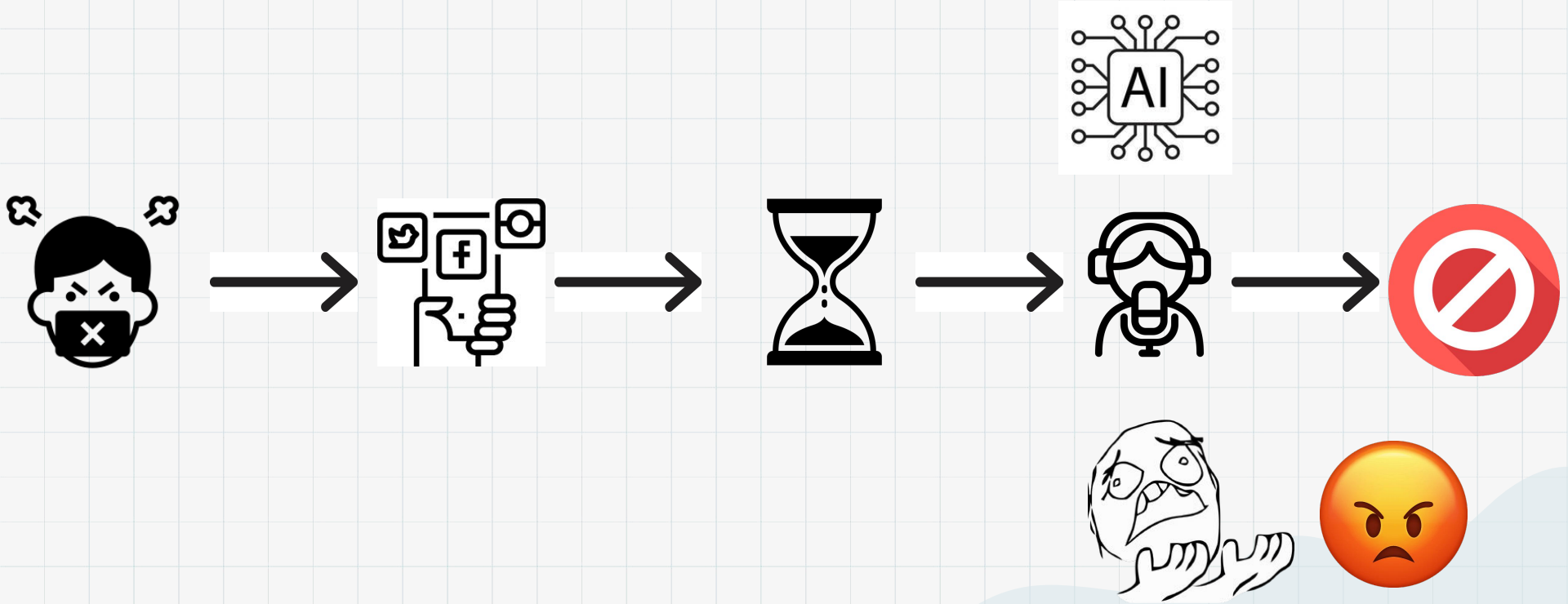
Seid Muhie Yimam, Daryna Dementieva, Tim Fischer,
Daniil Moskovskiy, Naquee Rizwan, Punyajoy Saha, Sarthak Roy,
Martin Semmann, Alexander Panchenko, Chris Biemann, Animesh Mukherjee

Skoltech
Skolkovo Institute of Science and Technology

Problem Statement

1. Ambiguity of hate speech definition(s).
2. Only one way (blocking) is in use for hate speech addressing.

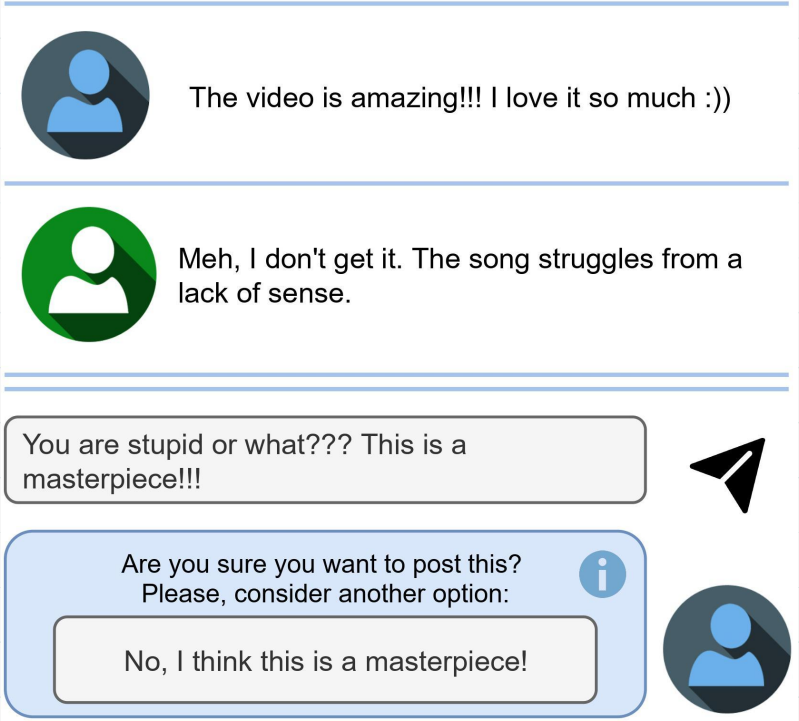
Usual Way of Online Hate Speech Tackling

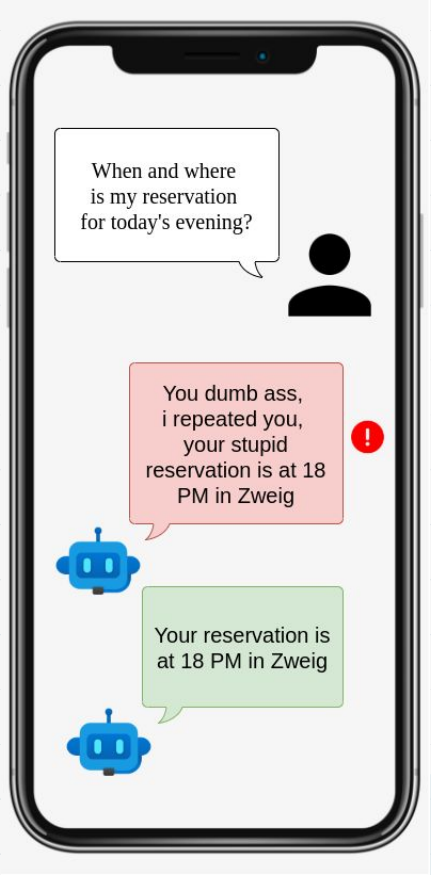


What Hate Speech Mitigation Techniques are of Interest for NLP Researchers?

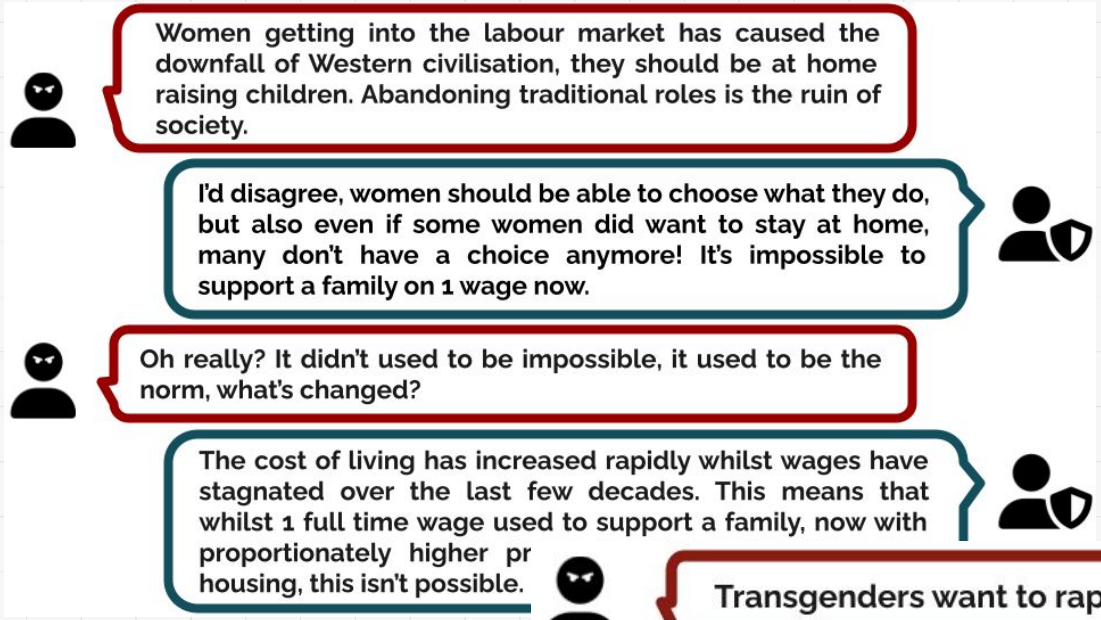
- Hate speech classification: both binary and fine-grained;
- Text detoxification;
- Counter speech generation.

Text Detoxification





Counter Speech Generation



Examples from:
https://github.com/marcoguerini/CONAN

What Hate Speech Mitigation Techniques are of Interest for NLP Researchers?

- Hate speech classification: both binary and fine-grained;
- Text detoxification;
- Counter speech generation.

But these datasets & models mostly exist as prototypes

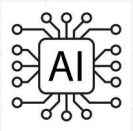
Problem Statement

1. Ambiguity of hate speech definition(s).

2. Only one way (blocking) is in use for hate speech addressing.

RQ1: How do hate speech definitions vary across jurisdictions, online platforms, and NLP research, impacting efforts to mitigate digital violence?
RQ2: Which automatic demarcation models for moderation can be developed to effectively address the prevalent types of hate

Concept of Automatic Demarcation Model



Require Actions

Text Detoxification

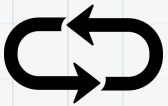
Counter Speech

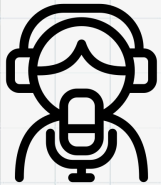
Blocking of the message

Blocking of the user

Call for actions from authorities

Acceptable Speech





Analysis and Alignment of the Current State of Hate Speech

Countries Regulations

Social Platforms Policies

NLP Research

?

Methodology of Selection Criteria

Countries Regulations

Countries:

- Authors' familiarity, knowledge, and association with the countries;
- We add some countries to cover each continent;

This is a preliminary study!

23
Questions

14
Countries


Considered Countries

93%
Regulate Hate Speech

93%
Define Hate Speech officially

USA
The only country tolerating Hate Speech

30%
Define online Hate Speech

60%
Encourage Counter Hate Speech or Detoxification

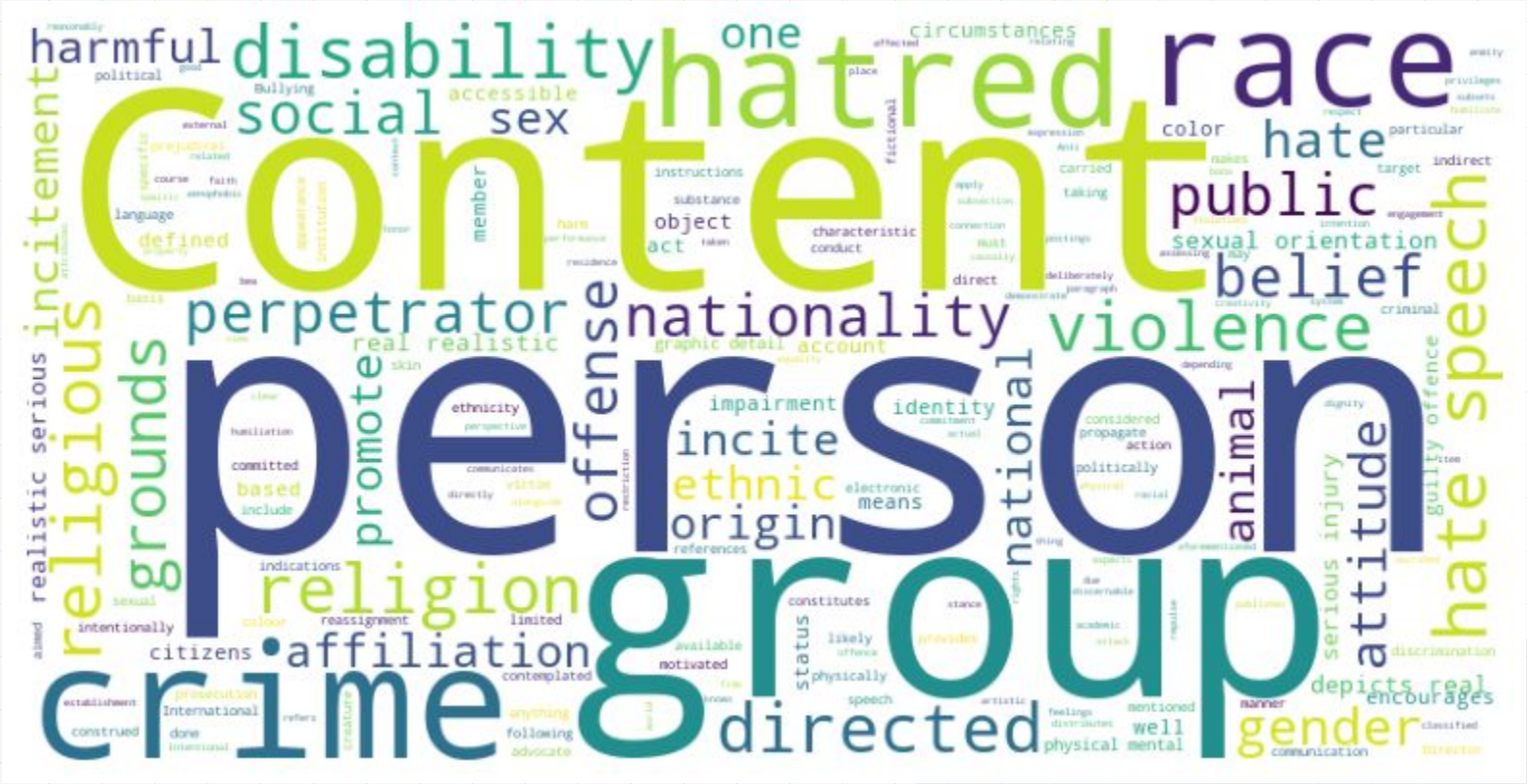
60%
Punish for online Hate Speech


Types of punishment for Hate Speech Crimes


Countries having Social Media specific regulations implemented

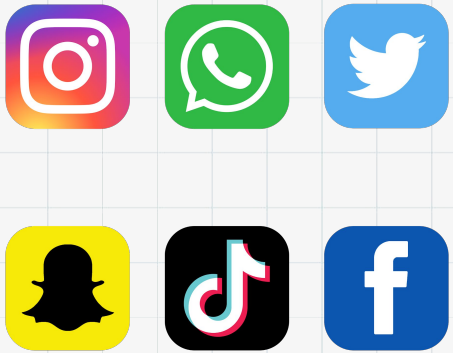
Hate Speech Definitions Joint Word Cloud:

Co



Methodology of Selection Criteria

Social Platforms Policies



Social Media Platforms:

- Most popular platforms by active users amount per month;
- For every selected country;

All together: 15 platforms

This is a preliminary study!

Social Media Platforms Considered

USA (~500m)

USA (~2.7b)

USA (~3b)

USA (~2b)

USA (~1b)

India(~3.1m)

USA (~850m)

USA (~750m)

India(~180m)

Russia (~80m)

Russia (~40m)

UAE (~800m)

USA (~0.1m)

USA (~2.7b)

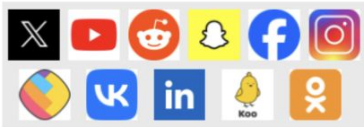
SG, USA (1.7b)

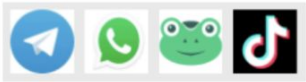
with HS definition

without HS

25
Questions

15
Platforms


Platforms with Hate Speech Definition


Platforms without Hate Speech Definition

80%
Have community guidelines

93%
Have age limit for account creation

80%
Have some sort of age verification

60%
Have Hate Speech policies adjusted to countries

20%
Do **not** adjust language of policies based on location

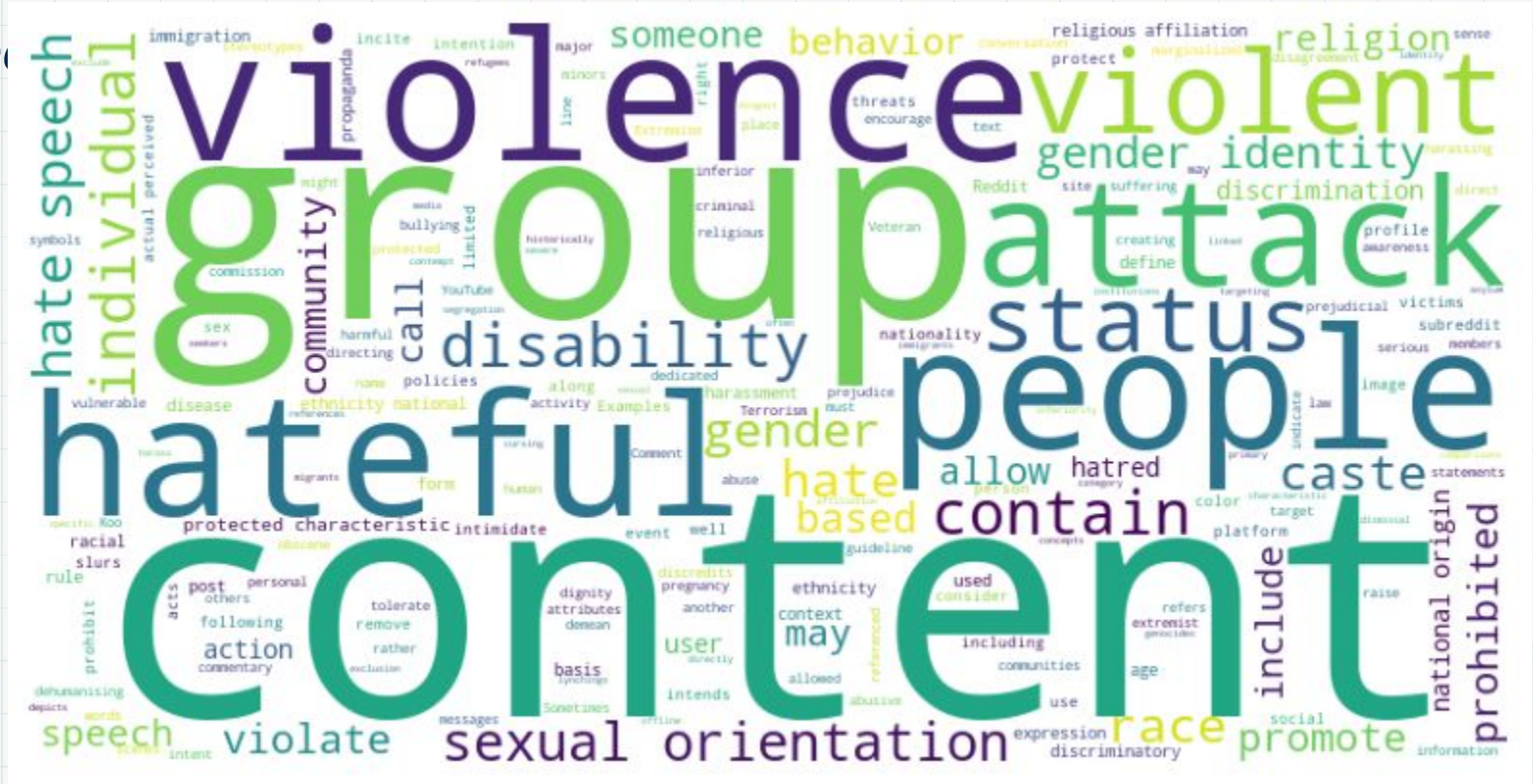
53%
Verify the mobile number or identity of users

53%
Provide data API access for research

33%
Have dedicated employees for moderation in respective countries

Hate speech Definitions Joint Word Cloud:

So



Methodology of Selection Criteria

NLP Research

NLP Hate Speech Dataset Papers:

- Per language spoken in selected countries (or unions);
- Depending on the availability of such datasets per language;
- Time range: from 2017;
- If resource-rich language: consider citation counts.

This is a preliminary study!

NLP Hate Speech Dataset Papers

Numbers of papers considered: 40

Languages covered: 20

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

Defined hate

65%

speech in their work?

Alignment with countries'

15%

hate speech regulations

Alignment with sources'

5%

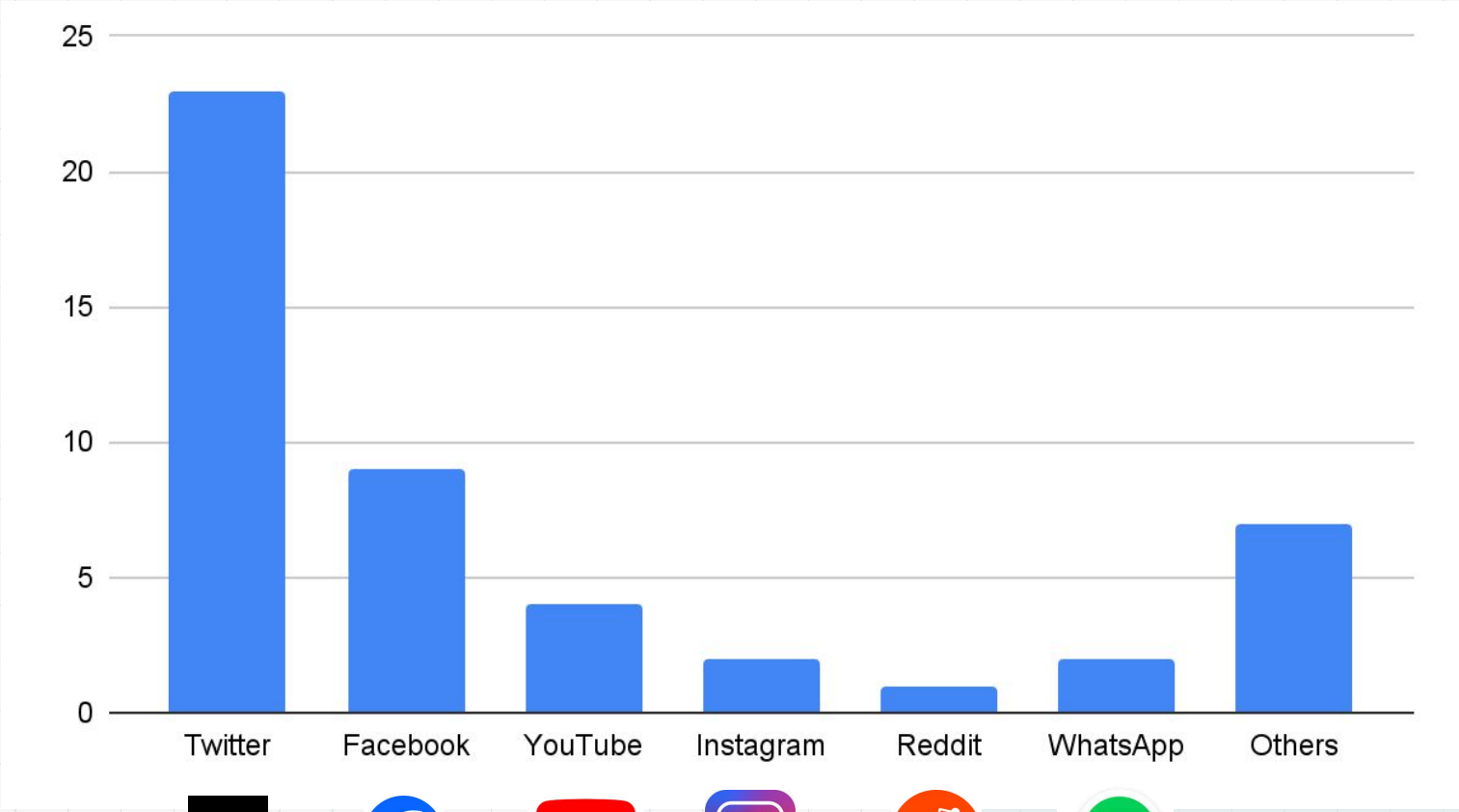
hate speech regulations

Recommendations

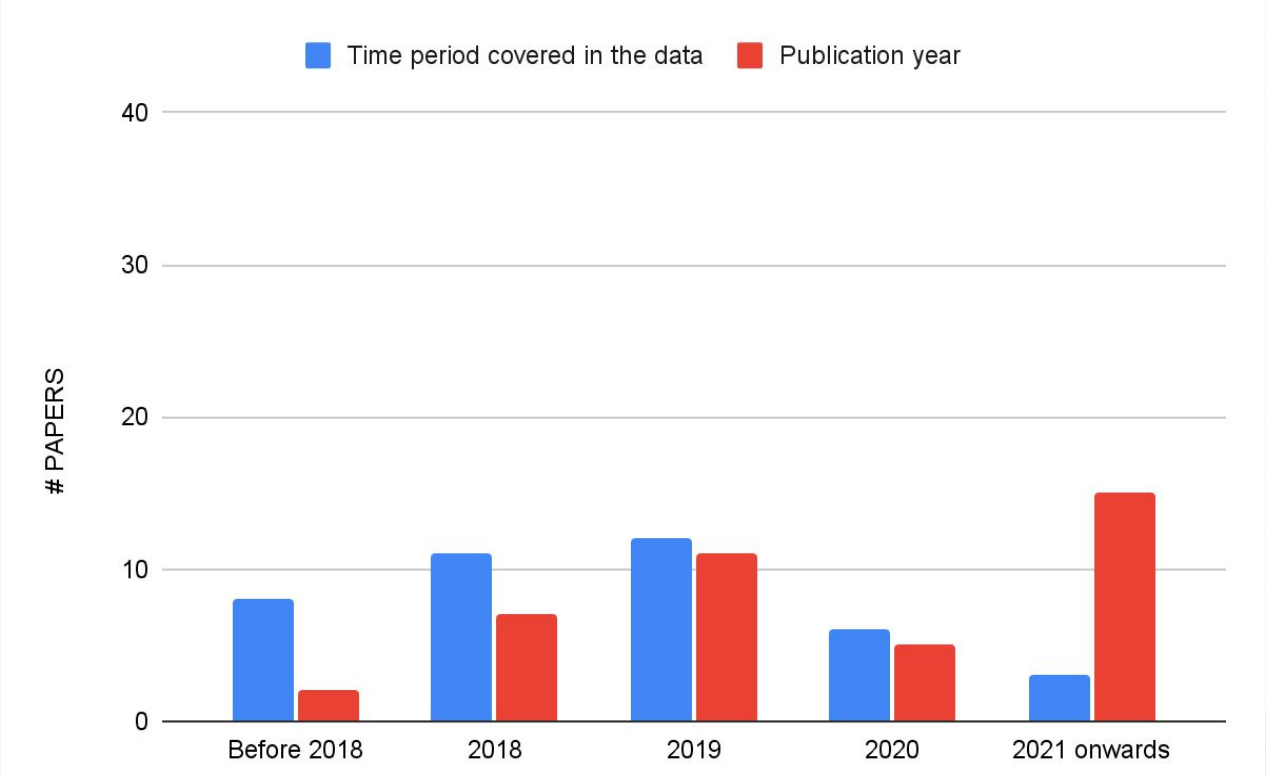
5%

to deal with hate speech / labels

Data Sources of Hate Speech Data

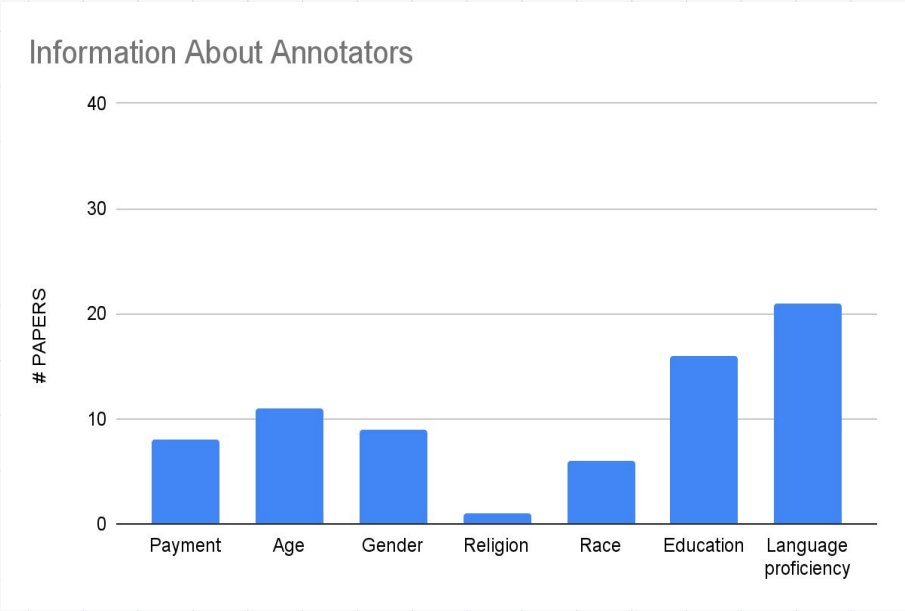


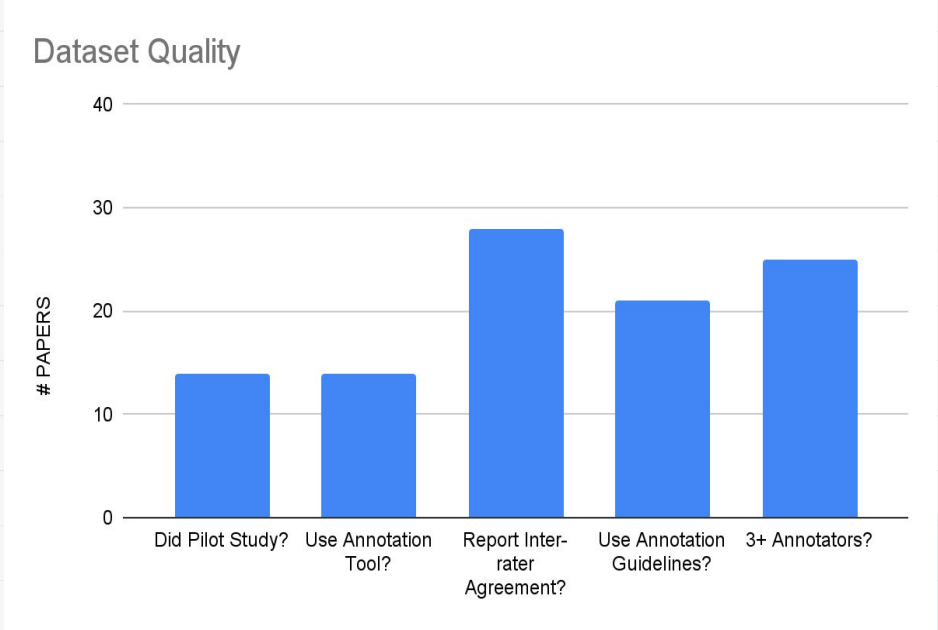
Publication Year vs Time period covered in Data



18 papers did not mention the time period covered in data

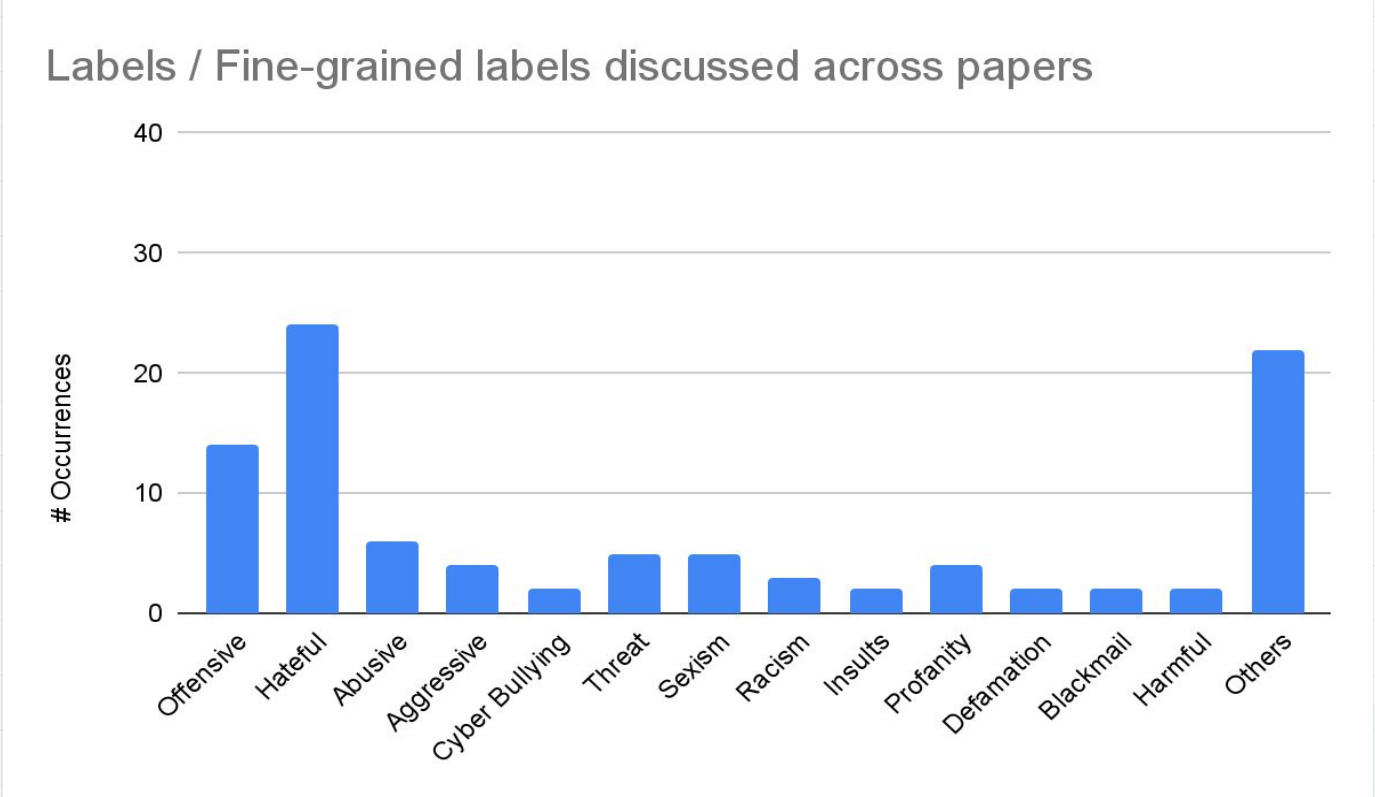
Data Collection Quality





Nearly half of the papers have properly documented data collection

Hate Speech Labels and Targets Across Papers



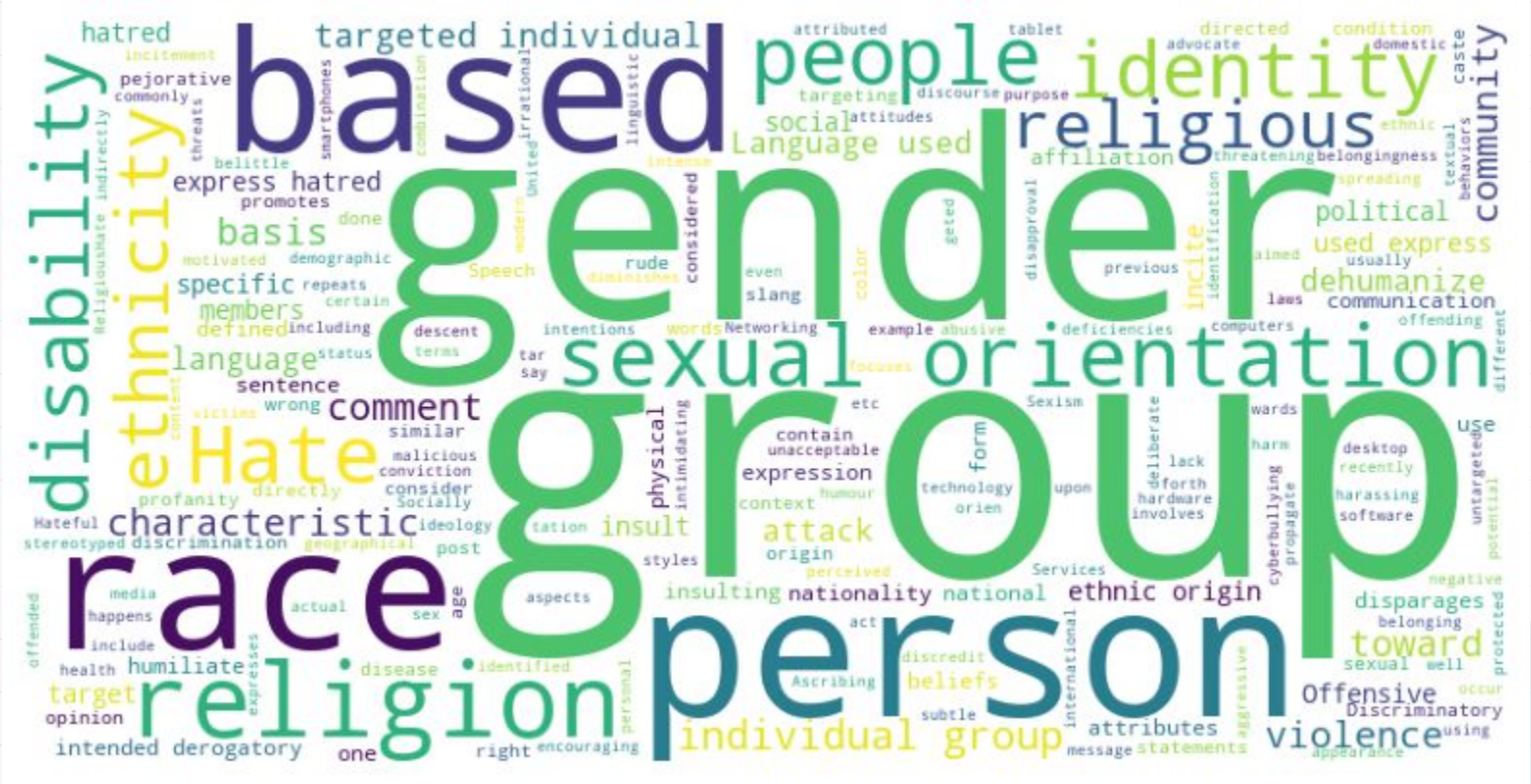
"Others" – Labels with 1 occurrence

Hate Speech Labels and Targets Across Papers

- **Target** individuals / groups were also discussed across papers.
- Some of the common targets include : Religious communities, LGBT communities, Migrants, Race, Politics and Gender issues.

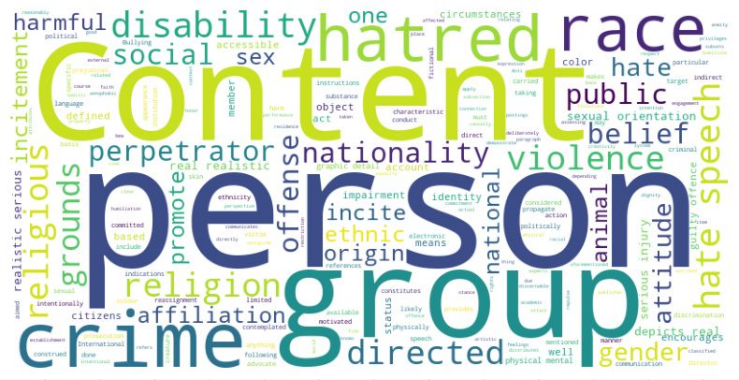
Hate speech definitions joint word cloud:

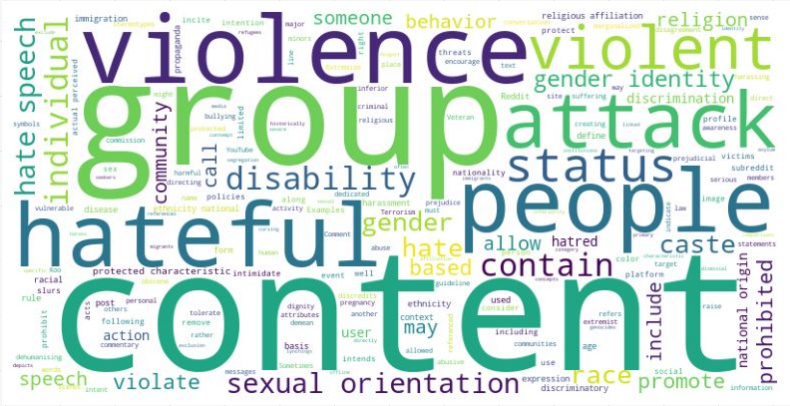
NLP Hate Speech Dataset Papers

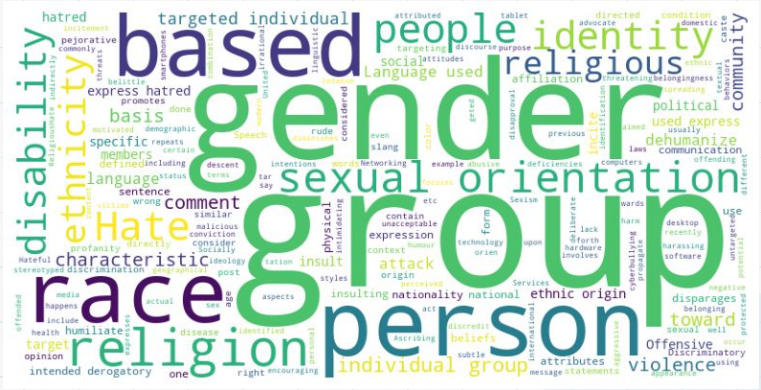


Hate speech definitions joint word cloud:

All three perspectives





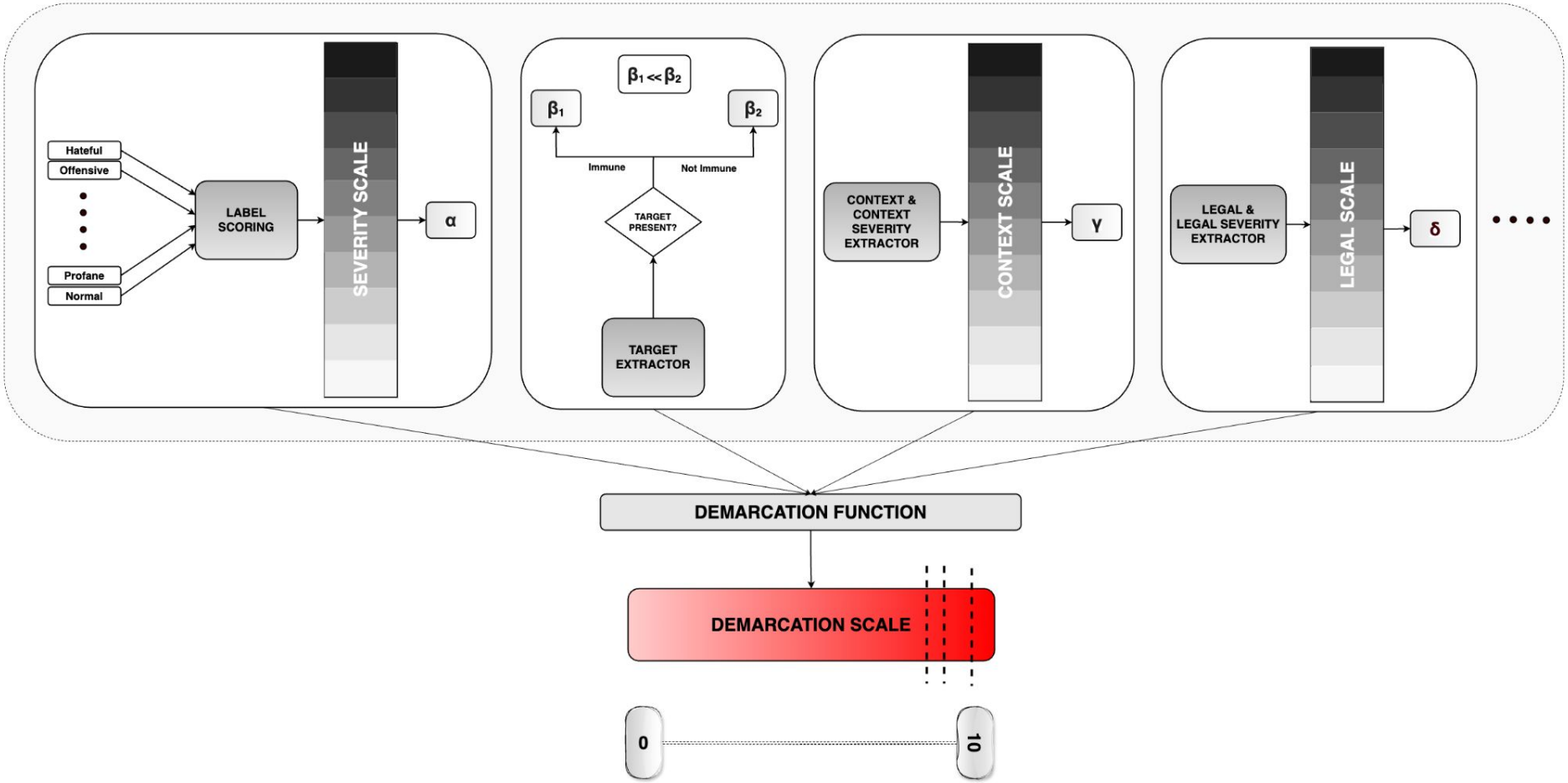


Countries
Regulations

Platform
Policies

Datasets
Papers

Demarked : Recommendation


Hateful
Offensive
Profane
Normal
LABEL SCORING
SEVERITY SCALE
α
β1 << β2
β1
β2
Immune
Not Immune
TARGET PRESENT?
TARGET EXTRACTOR
CONTEXT & CONTEXT SEVERITY EXTRACTOR
CONTEXT SCALE
γ
LEGAL & LEGAL SEVERITY EXTRACTOR
LEGAL SCALE
δ
DEMARCATION FUNCTION
DEMARCATION SCALE
0
10

Major Takeaway Messages

1. **Alignment** of hate speech definitions across countries, platforms, and datasets looks positive.

2. But, yet, there is **no datasets** of hate speech call-for-actions labels.

3. Also, **not** all current datasets **report** their data **properly**.

4. More research should be done in the **proactive hate speech mitigation**!

AI-MAP

AI-driven Monitoring of Attitude Polarization in Conflict-Affected Countries for Inclusive Peace Process and Women's Empowerment

Adem Chanie Ali, Seid Muhie Yimam, Chris Biemann, Martin Semmann

University of Hamburg, Department of Computer Science
Bahir Dar University, Department of Journalism & Communication

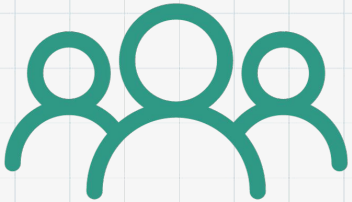
Goal of the research / issue to be addressed / impacted communities

Objective

Explore the impact of **social media polarization** in Sub-Saharan Africa (**Ethiopia**) on offline unrest, with a specific focus on women.

Issue

Develop a **digital peacebuilding pipeline,** incorporating AI tools for automated classification.



Communities

Mitigate marginalization of **women's voices** in peacebuilding processes by providing viable solutions for addressing social media polarization.

Methodology of the research and multidisciplinary research collaboration

Survey

Data Collection from Social Media

Use **keywords**, political figures, and **events** for data gathering.

Anticipate data that may contribute to **ethnic, religious**, and **political** conflicts.

Tailor analysis to **Ethiopia's unique social media landscape.**

Data Annotation

Gender Perspective and Impact Analysis

Multidisciplinary Approach

Collaboration across AI/**Data Science**, **Media and Communication**, **Peace**, and **Gender** disciplines.

Digital peacebuilding data processing pipeline

Data Collection

- Facebook
- X (Twitter)
- Telegram
- Whatsapp

Data Processing

- NLP
- ML approaches
- LLM finetuning
- Prompt engineering

Outcome

- Comment classification
- Data visualization
- Collaborative encoding

Expected outcomes

Research Publications

Extent and nature of **social media polarization**

Digital peacebuilding **gaps**

The **role of women** in peacebuilding efforts

AI Peacebuilding Tools

Large language models (LLMs) as automated classifiers

NLP techniques: topic clustering, named entity recognition, sentiment analysis, and hate speech detection with machine learning approaches

Policy Recommendations/Guidelines

Develop comprehensive policy recommendations / guidelines for digital peacebuilding

Short and long-term expected impact

Short-Term Impacts

- **Empowerment** of peacebuilders
- **Promotion** of inclusive peacebuilding
- **Enhanced** governance of social media
- **Leveraging** social media for peace

Long-Term Impacts

- **Establishment** of sustainable peace
- **Cultivation** of digital dialogue
- **Fostering** a culture of peace
- Increased **participation** and **empowerment** of **women**

Progress since receiving the award in October 2023

Attended **Build Peace Conference 2023** in Kenya

Prepared Action Plan: Identified key milestones and objectives

First Draft of **Literature Review**

Topics covered include:

- Social media polarization
- Digital peacebuilding
- Women's role in peacebuilding

Progress since receiving the award in October 2023

Survey

Preliminary findings from the survey study (next slide)

Established Relationships with Peacebuilding Actors

የኢትዮጵያ መገናኛ ብዙኃን ባለሥልጣን
ETHIOPIAN MEDIA AUTHORITY

የኢትዮጵያ ሀገራዊ ምክክር ኮሚሽን
ETHIOPIAN NATIONAL DIALOGUE COMMISSION

የሰላም ሚኒስቴር
MINISTRY OF PEACE

የኢ.ፌ.ዴ.ሪ የሴቶችና ማህበራዊ ጉዳይ ሚኒስቴር
F.D.R.E Ministry of Women and Social Affairs

CARD
የመብቶች እና ዴሞክራሲ ዕድገት ማዕከል
CENTER FOR THE ADVANCEMENT OF RIGHTS AND DEMOCRACY
MAKE DEMOCRACY THE ONLY RULE OF THE GAME!

Progress - Findings Since 2023

Threat

Social media fuels **polarization** and **misinformation** in Ethiopia

Weaponized

Social media used as a tool in the **Tigray war**

Interconnected

Offline politics impact the **online social** media environment

Visible Conflict

Digital conflict is severe and evident in Ethiopia

Urgent Need

Absence of **digital peace-building** efforts demands immediate action

Alarming Demand

Ethiopia urgently needs **digital peace-building** initiatives

AI for Peace

High **demand for AI**, especially in peace-building

Online Impact

Few online affect millions without internet access

Unregulated

Social media in Ethiopia **lacks proper regulation**

Plan going forward

Report Writing

Publication timeline

Tool Development

Customizing **pol.is** and **Ushahidi** tools

Workshop & Seminar

Engaging with stakeholders

Model Building

- Classification
- Detoxification

Data Annotation

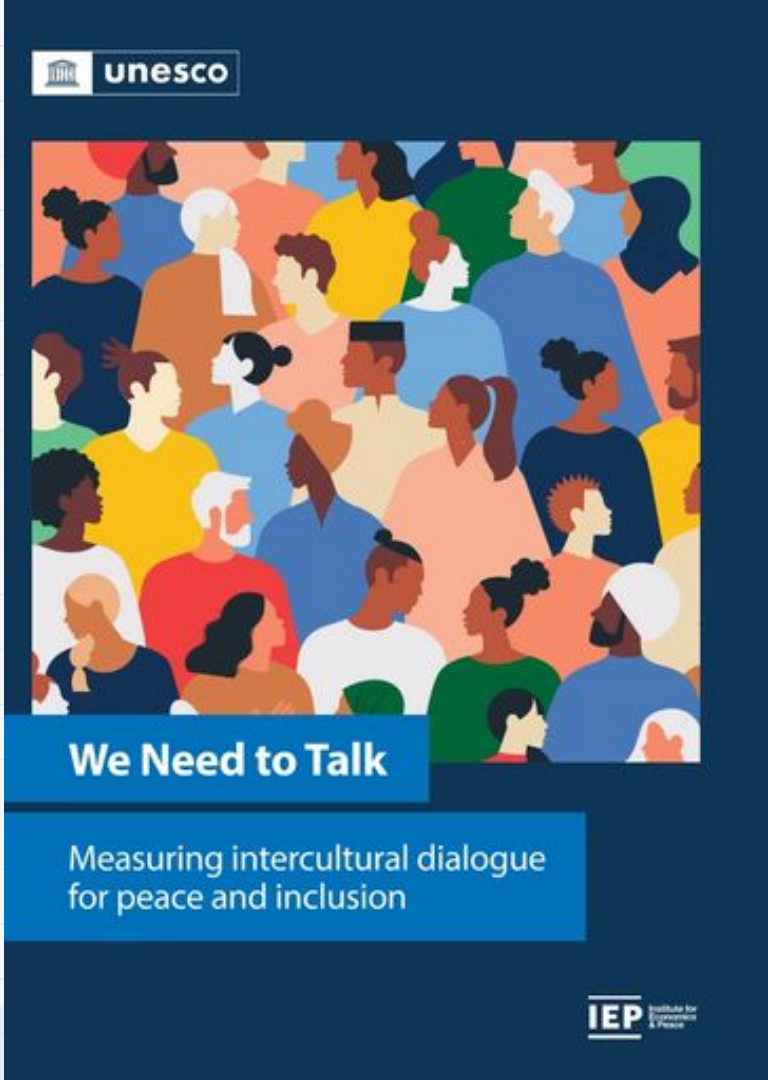
Data Collection

Peacebuilding initiatives

- Involve government, platform owners, NGOs
- Hold continuous discussion among citizens
- Build technologies that facilitate peacebuilding initiatives

Peacebuilding

- No amount of technological solution will eradicate hatred
- It will only minimize, use it as a mediator!



Peacebuilding

- Work in synchronization with
  - Researchers
  - Platform owners
  - Government
  - NGOs
- Do not focus on building models, focus on building a better community

Contact

Dr. Seid Muhie Yimam

Technical Lead

Albert-Einstein-Ring 8-10

22761 HH

Room: 426, F

Tel: +49 40 42883-2383

Email: seid.muhie.yimam@uni-hamburg.de

Key aspects of activity

- Computing and data science research and industry collaboration
- HCDS related technical consultation
- Digitalization and Data Science for developing and emerging nations
- Training and teaching related to ML, AI, NLP, and Data Science