Hellina Hailu Nigatu, Inioluwa Deborah Raji, John Canny, Sarah Chasins

Current State of Online Content for "Low-Resourced Languages"

Background

A popular trend in AI is to train models on crawled datasets from the web. However, the quality and quantity of online content for low-resource languages inadequate. We show why and how.

Methods

A need-finding study with Wikipedia contributors in Amharic, Tigrinya and Afan Oromo

An analysis of experience of policy violating content on YouTube in Amharic

Design Opportunities

Multi-Modal Interaction Systems

Relevant Information Retrieval and Search results

Culturally & Linguistically Aware Language Technologies

Designing for Inclusion while Protecting from Exploitation

Intersectional, inclusive, participatory design

Recommendations

- For Social Media Platforms:
  - actively consider limitations in low-resourced languages
  - culturally aware and context specific moderation strategies.
- For government bodies
  - oversight agencies requiring platforms to disclose how they account for languages of the countries in which they operate.
- NGOs protecting marginalized groups
  - trainings on online safety
  - access to legitimate health and legal information

Design of online knowledge repositories make it hard for "low-resourced" language speakers to contribute content in their languages.

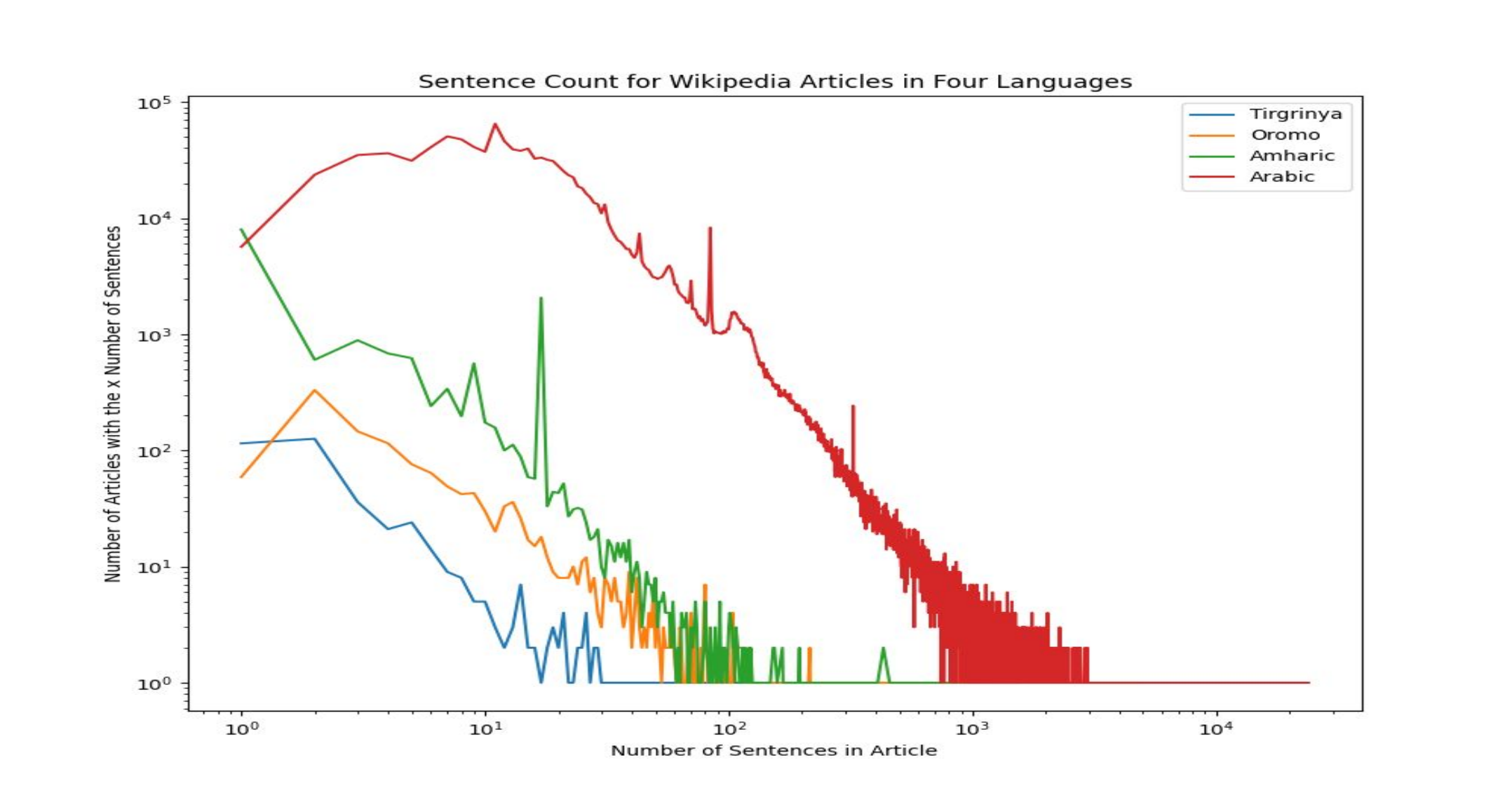

Fig. 1. Most articles in LRLs have just one or two sentences.

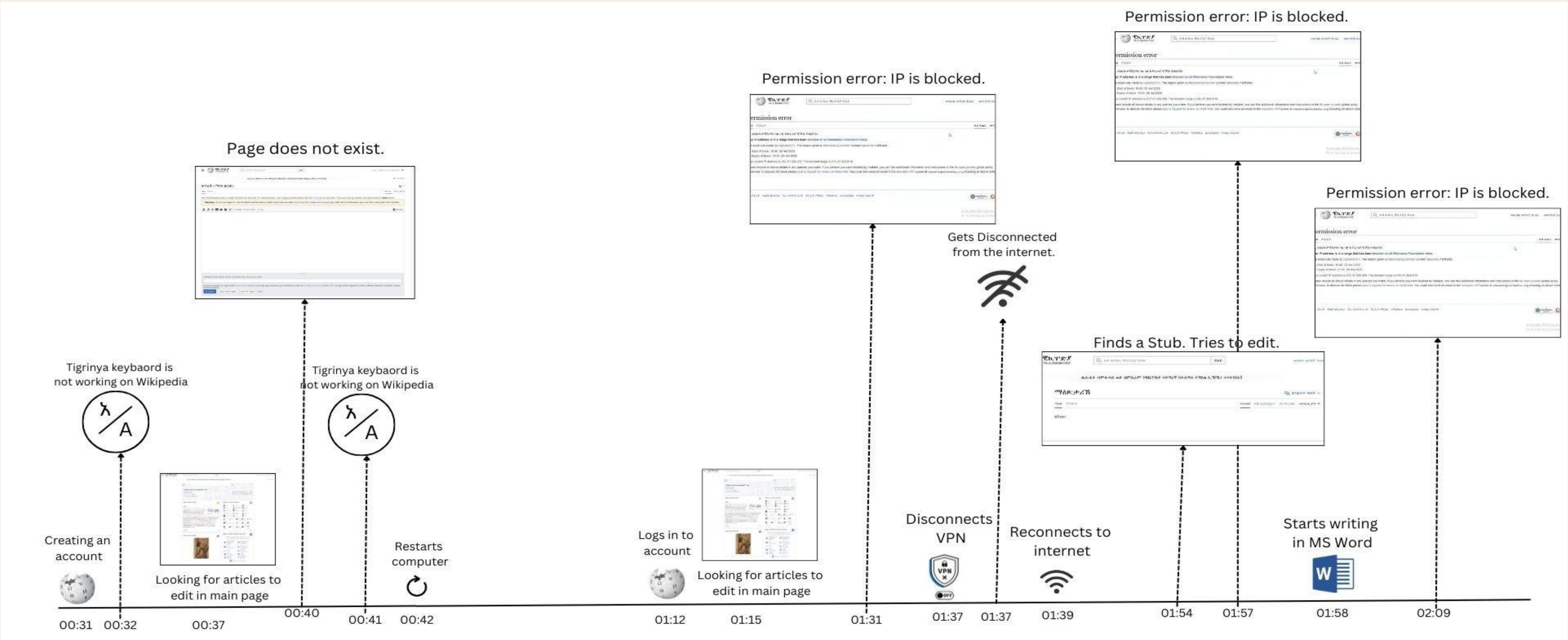
Creating an account
00:31
Tigrinya keybaord is not working on Wikipedia
00:32
Looking for articles to edit in main page
00:37
Page does not exist.
00:40
Tigrinya keybaord is not working on Wikipedia
00:41
Restarts computer
00:42
Logs in to account
01:12
Looking for articles to edit in main page
01:15
Permission error: IP is blocked.
01:31
Disconnects VPN
01:37
Gets Disconnected from the internet.
01:37
Reconnects to internet
01:39
Finds a Stub. Tries to edit.
01:54
Permission error: IP is blocked.
01:57
Starts writing in MS Word
01:58
Permission error: IP is blocked.
02:09

Fig. 2. Timeline of a participant; obstacles faced trying to write an article in Tigrinya.

Failure in Content Moderation pipelines lead to toxic and harmful content being rampant on social media platforms in low-resourced languages.

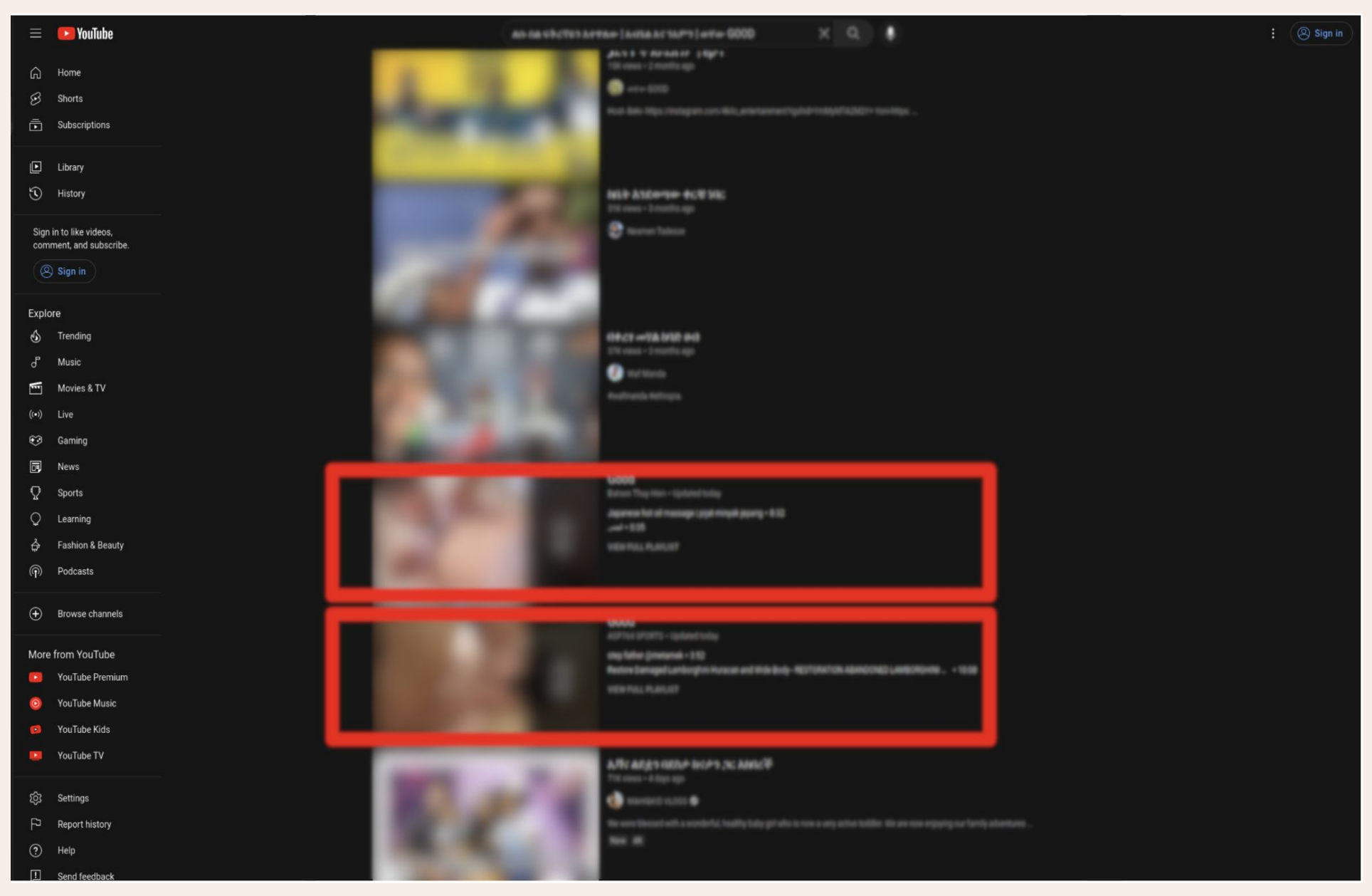

Fig 1. Violations in Search

Low-Resourced Language speakers' experience deteriorates when using YouTube in their languages

"It was in the morning and I was about to pray . . . I searched for a religious song in Amharic and got sexual content instead." **—P3**

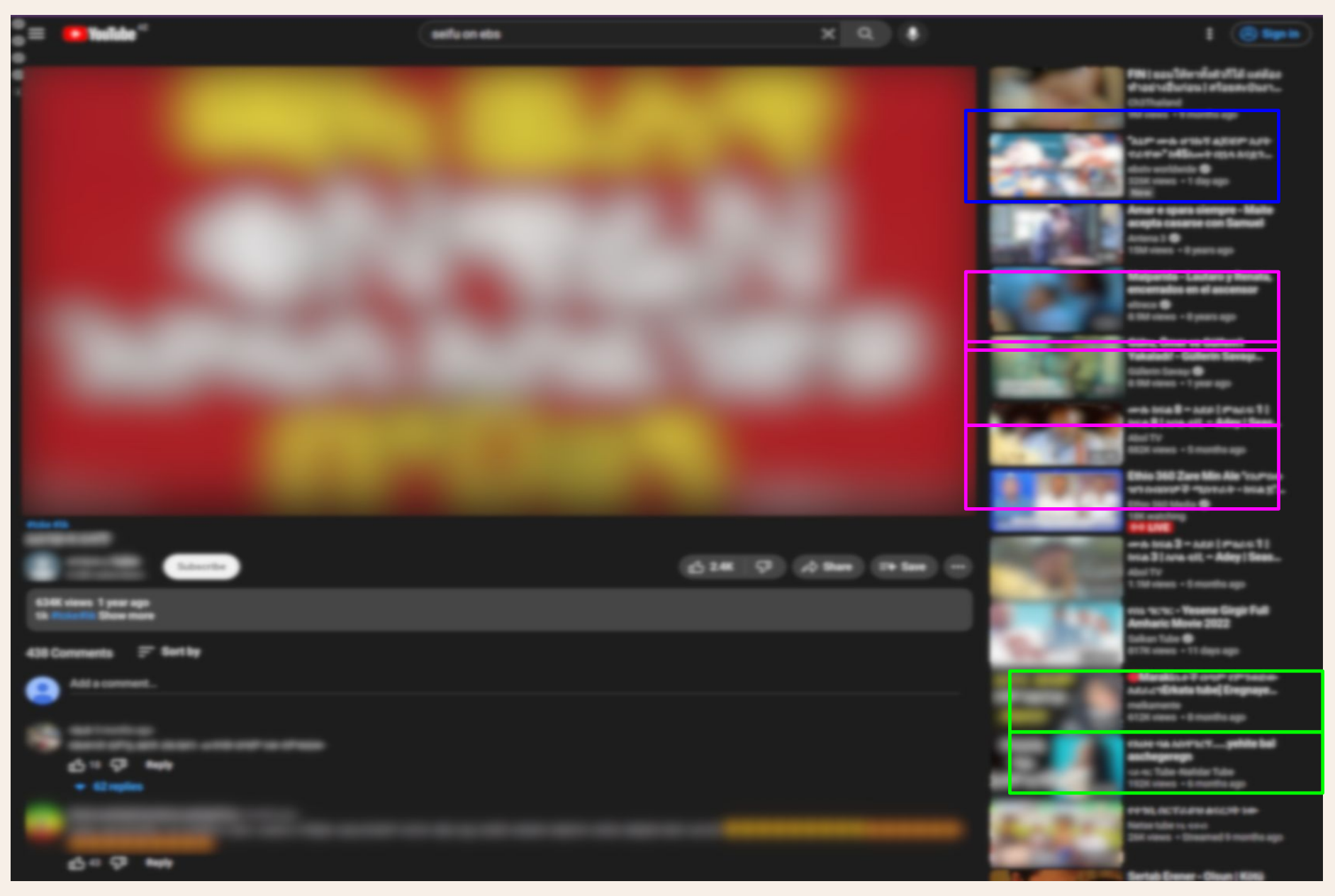

Fig 4. Violations in Recommendation

Disparate Experiences in the Comment Section.

51.6% of users who disclosed their location indicated being in the Middle East and some had indicators they were migrant domestic workers.

Berkeley
UNIVERSITY OF CALIFORNIA

Wikipedia paper

YouTube Paper