POLICY FORUM

Meta

March 28, 2023

Recommendation: Slurs Definition and Designation Process Revamp

Core Policy Team

Slurs Revamp

Policy development proposal

Issue Statement:

We remove slurs because we believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. We want to change our external definition of slur from "inherently offensive" to a research-based definition focused on the word's connection to historical discrimination, oppression, and violence against PC (protected characteristic) groups. However, people might interpret the change in definition as inconsistent with our rule of applying hate speech protections to all groups.

Current Definition

Status quo

Current slurs definition:

“Slurs are defined as words that are inherently offensive and used as insulting labels for [protected characteristics].”

Issues to resolve:

- Subjective criteria
- Indexing on offensiveness
- Lack of documentation
- We developed a robust informal process that we are seeking to codify

Research

What is a slur?

Research:

Academic research largely agrees that a slur is an expression that:

- a) signals that the target is a member of a group defined by protected characteristics, and
- b) invokes “sociohistorical facts, attitudes, and prejudices about the group.” (Davis & McCready 2020, p. 64; see also Hess 2021, Gutzmann 2015)

The contextual bundle of historical facts, prejudices, and social stereotypes invoked in a slur separates a slur from other types of words.

New Definition

Our new definition of slurs puts a greater emphasis on the harm these words can create and, for the first time, ties the definition to historical discrimination.

New slurs definition:

“Slurs are words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic, often because these words are tied to historical discrimination, oppression, and violence. They do this even when targeting someone who is not a member of the PC group that the slur inherently targets.”

How to Evaluate and Designate a Slur

Two step process: objective criteria with greater efficiency

Step 1: Qualitative analysis

Examples include:

- Is the word historically linked with usage which creates intimidation, violence or oppression against a PC group?
- What is the meaning of the term if it is written on a wall (like in graffiti)?

Step 1 can be completed in ~30 mins by the regional expert and it is sent to the policy team for initial approval before moving to Step 2

Step 2: Quantitative analysis

Examples include:

- What percentage of content is alternative meanings? Allowed use?
- If the word were a slur, how much of the sample would be self-referential use?

Step 2 analysis is completed using labelling of on-platform data

Meta