Our approach and framework to managing hate speech, harassment and bullying

Based on the experience we have seen the following classifications of behaviors

- Harassment, threats, stalking
- Cyberbullying and trolling
- Hate crime, hate speech, hate incidents
- Doxxing, Revenge pornography and image-based abuse
- Child and minor's safety

The nuances in identifying these are contextually linked

- Region and Culture
- The tech and platform understanding
- Linguistic familiarity and lexicons
- Defining and identifying policy sub categorization

Our approach to mitigating the nuances at scaled operations is to embed

- Contextual training
- Familiarizing
  - Platforms
  - Regional and language nuance
  - Defining the fine line of differences in policies through calibrations
- Digital Automation tools on lexicons

To continually be ahead of the curve in identifying trends we proactively engage in a 5-step process

- Focusing on the research of evolving abuse trends
- Identifying and exploring patterns in the data that are to be addressed at priority
- Synthesizing for quality, sensitivity, coherence and relevance
- Organizing the data collected and validating policy readiness
- Suggesting policy changes and building learning modules for people readiness

Today we manage the moderation of 90% of harassment, bullying, and hate speech content in North America for one of the largest social media and community platforms.

twitch genpact