

# LGBTQ+: Celebrating Diversity; Safeguarding Inclusion



# Table of Contents

1. Intro: Acceptance vs. Intolerance
2. Challenges of Protecting LGBTQ+ People in Online Communities
3. Role of Humans in Protecting LGBTQ+ Communities
4. Deploying Technology to Scale Moderation

# I.

## Acceptance vs. Intolerance

There are a few truisms in life. One is that wherever and whenever people congregate, some of them will identify as LGBTQ+, and others will feel the need to discriminate against them. According to a [Gallup poll](#), 7.1% of the US adult population identifies as LGBTQ+, as do 21% of Gen Z. That's 25 million Americans. Meanwhile, [30% of Americans](#) say it's "morally wrong" to be gay or bisexual.

Worldwide, [5% of people identify as gay](#), and [4% as bisexual](#), but the level of acceptance [varies greatly by region](#). In some countries, such as Sweden, 94% of people feel homosexuality should be accepted; in Russia, just 14% feel the same.

These statistics reveal an inherent danger for people in the LGBTQ+ community. For every one person who identifies as a community member, more than three others find their existence "morally wrong." That danger plays out in online platforms every day, often with brutal and tragic consequences.



## Social Media Effectively Unsafe for LGBTQ+ Community

Harassment and abuse against the LGBTQ+ community runs rampant in social media. According to the Pew Research Center's 2021 report, [The State of Online Harassment](#), 70% of the community have experienced being victims of harassment and hate speech, and 51% have been targeted for "more severe forms of online abuse." The 2021 ADL survey, [Online Hate and Harassment: The American Experience 2021](#), found similar findings.

Harassment on social media is so pervasive that [The GLAAD Social Media Safety Index](#) warns readers that Facebook, Instagram, TikTok, YouTube and Twitter are "effectively unsafe for the LGBTQ+ community."



## The Perils of Dating Apps

All the dangers that occur in social media are present in dating apps, and then some. The purpose of dating apps is to prompt interactions that can ultimately lead to in-person meetings and potential intimacy. Many apps reveal more information about a user's location than social media, as users can select to meet people who are within a specific vicinity of them, removing some of the anonymity people rely on.

In some cases, people use dating apps to lure victims, as was the case with three men in Texas who were sent to prison for using a dating app to target gay men for violent crimes. [According to the Department of Justice](#), these men "brutalized multiple victims, singling them out due to their sexual orientation."

## When Discrimination Is Institutionalized

While acceptance of LGBTQ+ identity is increasing, it is counterbalanced by a growing level of intolerance in other communities. For instance, Florida's "Don't Say Gay" bill will hurt teens who identify with the LGBTQ+ community. Officially known as the Parental Rights in Education Bill, the law bans any discussion of sexual orientation or gender identity. Legislators in at least seven other states have introduced similar bills.

### The implications are enormous:

- More school-age children in the LGBTQ+ community will seek online communities for support and connection.
- More school-age children in the LGBTQ+ community will feel afraid about being targeted and having their rights and identities eroded.
- More hateful and intolerant language will show up as people will feel emboldened by the multiple bills wending their way through State legislatures.
- More promotion of discriminatory bills will occur, with widespread increases in uncivil and intolerant language across all sorts of platforms.

As a Trust & Safety professional, you likely believe that discrimination against any member of your community is unacceptable. You want every user to feel safe and welcome, even as discrimination becomes institutionalized, and therefore more acceptable to some segments of society.

This paper will address the key challenges of providing a safe environment for LGBTQ+ communities worldwide, along with technology approaches that allow Trust & Safety teams to scale their moderation efforts.

## II. Challenges of Protecting LGBTQ+ Communities in Online Platforms

### A Delicate Balancing Act

Trust & Safety teams want to empower their LGBTQ+ platform users to connect and freely express themselves, while simultaneously ensuring they're kept safe from insults, hate speech and abuse. Society has a keen interest in curbing these behaviors, as casual use of insults and hate speech against any protected class can lead to real-world violence, outcomes we see regularly across the globe.

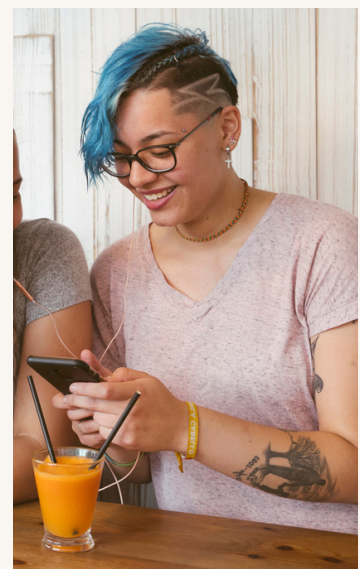
Moderating hate speech directed against members of the LGBTQ+ community is particularly challenging. We understand that certain words and phrases are discriminatory, but those same words and expressions have been reclaimed by community members, and are used in empowering and connective kinds of ways.

*Empowerment is a key reason why many online platforms exist.*

Balancing these two goals requires a considerable amount of expertise, testing and measurement, and vigilance. To distinguish the intent behind the use of a word, we must accurately assess the context in which it occurred. Did the speaker use it in a manner that is appropriate, or did the speaker intend to inflame or hurt another user?

This is a particularly fraught situation with dating sites, which exist to encourage conversations and real-life meetings. By definition these sites will have a different type of discourse than a gaming or ecommerce site, which typically ban any kind of sexual discourse.

While context can add significant clarity into a speaker's intent, it's not always easy to obtain it, especially if an interaction occurs in a private chat on a platform meant for adults. In the absence of clear context, we at Spectrum Labs believe it is better to prioritize protection and human rights, while striving to keep communities free and open to those who often find themselves under attack.



## Ways to Approach Moderation

Language that is high risk or threatening to the LGBTQ+ community is surfaced by multiple models, including hate speech, bullying and insults. We've found that the best way to elevate safety and simultaneously promote the freedom to connect is to deploy all three in an online platform. This will help ensure that the online world, often a rare refuge for some LGBTQ+ community members, remains as safe as possible.

It's also important to assess what people do across online platforms. Words and phrases that denote sexual orientation or gender identity can be used as insults on one platform, but are completely empowering on another platform or between two other users.

Both people and technology play a critical role in Trust & Safety. Let's look at each.



## III. Role of People in Moderating Hate Speech and Violent Extremism

Sometimes hate speech and threats against the LGBTQ+ community are overt and therefore easy to detect and moderate, but not always. All too often, insults and threats can be couched in benign-seeming terms. Conversely, terms deemed toxic when used outside the LGBTQ+ community can be empowering to people who are within it.

### Role of Community Guidelines in Designing Moderation Approach

Moderation begins with well-established and easily accessible Community Guidelines. These guidelines reflect the values of a platform, and the types of behaviors its leaders find acceptable. The purpose of Community Guidelines is to tell a platform's user base what is acceptable speech behavior, along with the governing rules that may result in a warning, suspension or banning from a platform.

Users may not realize that their behavior is unacceptable to the platform, especially if they're young and don't yet comprehend the words they use and the impact they have on the recipient. Likewise, an LGBTQ+ user may not realize that the toxicity addressed to them isn't tolerated on the platform, and that they have rights.

*Safety is promoted when Community Guidelines are readily available and reinforced by proactive actions on the part of the platform.*

Community Guidelines are specific to each platform. A gaming platform that is designed for kids will have a different set of policies around acceptable language and behavior than a dating app geared towards the LGBTQ+ community. Even within the dating industry, two platforms may have dramatically different tolerances for sexual content.

It's really important that Trust & Safety teams establish Community Guidelines and update them frequently. They should also share those Community Guidelines with their Trust & Safety moderation technology providers so that they know how to enforce their standards.



## Role of Humans in Moderation

While technology is an excellent way to augment or scale community moderation, it can never entirely replace the need for humans. Behavior and language evolve over time, and what's offensive to one person may be empowering to another. There will always be edge cases that require a human to pull additional data and conduct additional analysis in order to understand the meaning and implications between people on a platform.

Context and implications get even more complex when a platform spans multiple languages. Acceptance of LGBTQ+ identity varies from country to country. Platforms need people to counteract those tendencies so that all of their users can flourish, regardless of how they identify.

All platforms need people who have considerable expertise, and can assess the use of a specific word or phrase against a variety of factors, including:

|                             |   |
|-----------------------------|---|
| <b>Type of Platform</b>     | The same word can be used as an insult between players on a gaming platform but as a source of empowerment for adults on a dating platform that caters to or welcomes members of the LGBTQ+ community.  |
| <b>Code of Conduct</b>      | The Trust & Safety team may opt to ban words that similar platforms find acceptable for a variety of reasons, including the age of users.   |
| <b>User Age</b>             | Young people often use discriminatory language without having a full understanding of its meaning or impact on others. The age of the user can be a consideration when deciding how to respond to incidents.  |
| <b>Location of Language</b> | Where a word or phrase appears on a platform. Context matters to determine intent, for example, in an LGBTQ+ chat room compared with a general setting.   |
| <b>Emergent Words</b>       | New words and phrases emerge all the time. Sometimes they can originate from within a community, or sometimes they emerge from popular culture. Language and usage patterns are never static, and while keeping up with the natural evolution of language is important, it is especially crucial to identify when new phrases or words are being leveraged in new ways by bad actors who seek to derail community dynamics. |

## Role of People in Multilingual Communities

Successful moderation across all languages requires the help of outside expertise, people who have a deep understanding of local languages and customs. These people are essential to creating updated definitions for emerging insults and hate speech against LGBTQ+ people. For instance, the word “woke” is now used by people who are intolerant of identity concerns.

## Machine-Translated Text Is Sub-Par

Multi-language moderation is one of the toughest challenges for platforms, and it's one that's not easily fixed. Translation software has significant limitations because moderation and enforcement require a more nuanced understanding of context and language use.

Moderation requires people who understand online dynamics, language and nuance of users. When this level of insight is available, it's possible to identify signs of hate and discrimination early -- well before the volume of dangerous content spills over into violence.



## IV. Using Technology to Augment Moderation

Technology plays an important role in moderating platforms for anti-LGBTQ+ insults, hate speech and harassment at scale. Technology is not meant to replace the moderation team, but it can be used to automate moderation to lessen the workload and the psychological burden of toxic content for your team of human moderators.

There are several approaches:

### Wordfilters, Keyword and Regular Expression Solutions

Wordfilters, keyword and regular expression based-solutions (aka, keyword/RegEx) look for toxic behavior based on specific words, “I hate [LGBTQ+ community members].”

These solutions are helpful for catching egregious insults and hate speech terms directed at specific attributes, but can be problematic when those words have been reclaimed by a community member, as they can mistakenly trigger as an insult.

*But the biggest challenge to word/expression-based solutions is the way they can leave community users vulnerable. Harassment that results in real harm can occur without ever using a single keyword on a keyword list.*

### Classifiers

Classifiers are message-level AI. They look at a single message or piece of content at a point in time and make a determination of its intent. For instance, classifiers are good at assessing explicit threats, such as a post that says, “I’m going to find you and kill you.” In general, however, classifiers look at a message without any additional context (e.g. who said it, to whom, their ages, and how it relates to other things that may have been said in previous exchanges).

*While classifiers are good at detecting behavior at the message level, in general they don’t consider context. Certain posts and words can seem innocuous if taken out of context, while other words considered bad may be perfectly appropriate for adult users in other situations.*

## Classifiers, cont.


An additional challenge is that global toxicity classifiers are one-size-fits-all, which makes them too rigid for an adult dating site, but too broad for a community meant for children.

A test of Google's AI demonstrates the challenge. In a now infamous example, Google's Perspective API ranked a drag queen's tweets as more toxic than those of David Duke, a known white supremacist, mostly because the former used reclaimed terms.

## Contextual AI

The third approach is contextual AI, which offers two important advances over the others. First, contextual AI analyzes data within its context, meaning it looks at content (the complete raw text) and context whenever possible (e.g. attributes of users and scenario, frequency of offender) in order to classify a behavior.

Second, of all the available approaches, only contextual AI can look across aspects of a platform -- posts, private chats or messaging -- and tie multiple messages together. At its core, contextual AI looks at how behaviors build over time, and how users respond to different messages, in order to distinguish between conversations that are consensual and those that are not.

| Example of Contextual AI  |  |   |
|---|--|---|
|  | <b>Message:</b> I want you   |   |
|   | <b>History:</b><br>Leave me alone  | <b>Context:</b> <ul style="list-style-type: none"> <li>• Two people</li> <li>• Private chat</li> <li>• 3:00 am</li> </ul> |
|   | <b>Determination:</b> <span style="background-color: red; color: white; padding: 2px 10px; border-radius: 15px;">✘ Harassment</span> |   |

## Role of Humans in Improving Technology

All AI tools must be trained on data, and the more data that's available, and the more accurately it is labeled, the more accurate it will be in detecting hate speech and violent extremism so that it can be moderated and removed.

A person's ability to label data accurately will depend largely on his or her background, culture and life experiences. Bias can be built into your AI model if you don't ensure diversity across all of your teams, especially those that contribute to the development and training of your models.

Machine-labeling of data is possible, but humans will need to spotcheck it on a regular basis to ensure all unwanted behavior is captured. Local human expertise in culture, language, and emerging trends in hate speech and violent extremism are also vital for ensuring models are up to date.

## Semantic Equivalence

Semantic equivalence is a data science term that essentially says two data elements (e.g. words) from different vocabularies mean the same thing. Two different languages may have a specific term that's derogatory to a protected class, but from a data science perspective, they have a semantic equivalence. Our data operations team rely on semantic equivalence to link terms with similar meaning and usage across a wide range of languages.

Semantic equivalence is also used to identify when the same word in two different languages have vastly different means. For instance, the word "fag" in English refers to someone who is homosexual, and depending on the speaker may or may not be derogatory. That same word means "cigarette" in the UK. This example is over simplified; in real life, language experts put tremendous effort into understanding when phrases have semantic equivalence.

### What is Semantic Equivalence?

*Two data elements (e.g. words) from different vocabularies mean the same thing.*

*Examples:*

United States: "fag" is derogatory  
United Kingdom: "fag" = cigarette

United States: "retard" is derogatory  
France: "retard" = slow

United States: "softboi" is semantically equivalent to "f\*ckboi" and both are derogatory. Spectrum's AI will flag "softboi" but a solution focused only on swear words could miss it.

## Moderating Content in Multiple Languages: Finding the Blind Spots

As online communities grow, they add users in new languages and regions.

### **Content moderation in new languages isn't as simple as translation:**

- A term that's harmless in one language can be toxic in another.
- Translation apps are inaccurate, slow and miss implied meanings.

Problems with anti-LGBTQ+ hate speech can proliferate if the platform lacks local moderators and expertise to monitor for emerging terms. And yet most content moderation solutions face obstacles to adding new languages as they require hiring local speakers to construct keyword lists, and building up language-specific data to train models over time.

It can be a long, expensive process to add each new language, leaving users unprotected. To meet these challenges, Spectrum created an AI solution to help our platforms scale faster and with lower costs.

### **Spectrum's AI solution supports:**

- A wide range of languages including character-based, hybrid and L3tspeak
- Localized, customized, automated actions
- Insights for regions, languages and user behaviors

Spectrum uses AI to transfer learning from one language to another, allowing you to add new languages immediately, then refine over time. Our contextual AI directly analyzes content in its native language and compares patterns to those found in languages where the most data exists, to make the right determinations across languages.

This multi-language approach removes blind spots due to not having moderators everywhere, enabling platforms to benefit from Spectrum's investment in native-speaker experts and data, and our patented AI multi-language approach.

## About Spectrum Labs

Contextual AI from Spectrum Labs can make a massive impact in content moderation. Our contextual AI evaluates multiple data points to more accurately identify behaviors, outperforming keyword-based tools. This allows for real-time analysis to proactively prevent toxic content and shape your users' experiences in the moment. Our content moderation solution is available across multiple content types and languages. Whether you are looking to safeguard your audiences, increase brand loyalty and user engagement, or maximize your moderators' productivity, Spectrum Labs can help make your community a better place.

<https://www.spectrumlabsai.com/contact-us>

