

Content moderation best practices

Maintaining brand trust in a digital world



Table of contents

Introduction	3
Why content moderation matters	4
Content types	6
Methods of content moderation	7
The human component	9
The role of artificial intelligence	11
Measuring success	12
Selecting the right content moderation partner	15
Maintain trust and sustain growth	16

Introduction

To create and maintain digital experiences that are truthful, welcoming and safe, content moderation is no longer just a consideration for brands — it's a must. The essential practice is all about applying standards to user-generated content (UGC) in order to identify and address submissions that violate community guidelines and governmental laws. Done effectively, content moderation protects the well-being of your user communities, and as a result, builds trust between your brand and consumers.

There is no debate: Customers expect brands to take ownership online. **In a recent TELUS International survey, 85% of respondents agreed that brands have a responsibility to moderate UGC for appropriateness and accuracy across their owned channels** (38% of those respondents strongly agreed).

"Facing a state of a near infocalypse on social media with deep fakes, spam, bad actors and the sheer exponential growth of user-generated content, both customers and brands want a

more protected online environment," said **Ivan Kotzev, customer experience lead analyst at NelsonHall**. To solve the growing problem, customers are looking to brands, and brands are looking to content moderation. They're investing heavily — the global social media customer experience (CX) services market is expected to reach \$6.5 billion in 2024, according to Kotzev.

With billions of people around the world actively using social media and an ever-increasing amount of UGC getting posted online, brands must have a combination of trained content moderation specialists, effective processes and the right technology to properly manage their online reputation and keep their customers safe. Partnering with a company that understands the ins-and-outs of **trust, safety and security**, and particularly content moderation and CX, can help you get there with advanced planning and careful execution. This in turn promotes improved user experiences that drive growth and revenue.

By sharing best practices and learnings from our partnerships with leading global brands, this resource will help you to build strategies that reduce liability and mitigate risk. Read on to learn more about why content moderation matters, various approaches, as well as considerations for measurement and partnership — all in an effort to help you keep your corner of the internet safe.

Trust, Safety & Security

Content moderation is just one aspect of a comprehensive trust, safety and security program that protects your customers and employees. In addition to content moderation, a holistic approach involves channel and community management, user safety, identity verification, fraud detection and more. By combining the best of human intervention and technology automation, TELUS International can help your brand effectively manage risk and compliance.



Why content moderation matters

We could write at length about the importance of a robust content moderation strategy. Fortunately, we don't need to: The results of our survey speak volumes. Here's what we learned from 1,000 Americans familiar with the definition of UGC.

Inappropriate or inaccurate UGC is on the rise

We spend more and more of our lives online, but the experience isn't always pleasant. Content moderation is key to ensuring that our future is a friendly one.



Half of the Americans surveyed (50%) said they saw inappropriate or inaccurate user-generated content when spending time online; 36% said they saw it daily.



The amount of inappropriate and inaccurate user-generated content is on the rise, with more than half of respondents (54%) saying there is more now than before the start of the pandemic (March 2020).

Effect on customer sentiment

Letting noncompliant content linger can result in lasting damage to customer sentiment.



Survey respondents indicated "Fake reviews or testimonials" (60%) and "Spam" (53%) were the most frustrating types of inaccurate or inappropriate user-generated content.



Nearly half (48%) of the respondents believe brands are irresponsible when they see inappropriate or inaccurate user-generated content on a brand's owned sites and/or channels, 48% indicated that they lose trust in the brand and 43% no longer want to engage in the brand's community.



Two-thirds (67%) of Americans agree that encountering inappropriate or inaccurate user-generated content has negatively impacted their day, and worse, 29% said that it has ruined their day.



Nearly 40% of respondents indicated that it only takes 2-3 exposures to inappropriate or inaccurate user-generated content on a website for them to lose trust in the brand.



Customer expectations

Where does the responsibility of content moderation lie? According to respondents, it falls primarily to brands.



A staggering 69% of Americans indicated that brands are responsible for “Protecting online users (e.g. eliminating hate speech, toxic content, cyber bullying),” and 66% said that brands have a responsibility to prevent the spread of fake information/propaganda.



Nearly 70% of respondents agree that the most satisfactory response is for a brand to remove the content when it has been flagged as inappropriate or inaccurate by a user.



Over three-quarters (78%) of Americans indicated that they'd like to see brands take responsibility for creating a positive and welcoming online user experience.

The survey results make a clear case for why content moderation matters. Though customer expectations are high, there are a number of ways to exceed them.

The statistics are based on a survey conducted by Pollfish on June 7, 2021 that included responses from 1,000 Americans familiar with the definition of user-generated content.

Content types

So you've thought about the value of content moderation, but you may still be wondering what exactly is in scope. In other words: What forms of content can be moderated?



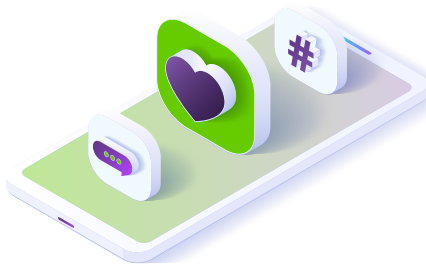
Text

A wide range of UGC fits under the umbrella of text data: comments, social media posts, reviews, direct chat messages and more. Text data can be reviewed to filter spam, detect profanity and ensure adherence to laws and other community mandates. Truly effective moderation of text data necessitates an approach that is multilingual and applies an understanding of context.



Audio

Audio data doesn't appear before your eyes and its moderation can require more resources as a result. Fundamentally, how brands perform audio content moderation comes down to whether the subject matter is recorded or live. Recorded audio can be reviewed by human moderators, by technology or by a combination of both, but the immediacy of live audio makes moderation complicated.



Image

Since images are highly impactful and have the potential to illustrate the effectiveness of your product or service, filtering out non-compliance goes a long way to maintaining a welcoming community. There are many reasons why images may be the subject of moderation, from innocuous misuse of a company's brand, to malicious depictions of violence or illegal activity.



Video

A medium that brings together the potential of both image and audio content, moderation is essential in order to ensure that the power of video is used for good. Most video services have built-in tools to enforce compliance and minimize the “violative view rate” — a metric that measures the number of views for videos that violate guidelines. Effective moderation minimizes that number.

Methods of content moderation

The approach to content moderation that makes the most sense for your brand will depend on your goals and your unique context. It's important to consider whether you want people to be able to communicate quickly and easily, or whether it's more important to keep your site completely free of sensitive content at all times. There are a range of different content moderation methods, all falling at varying points on the spectrum between these two goals. No matter where you land, a hybrid approach that teams up human moderators with leading technology is key to success.



Pre-moderation

Given its name, it's not surprising that pre-moderation involves all user submissions being placed in a queue for moderation before they are displayed. Through pre-moderation, it's possible to keep all sensitive content off a site by checking every single comment, image or video. However, for online communities that prize immediacy and barrier-free engagement, this moderation method can cause challenges. It's a method that's best suited to sites that need high levels of protection, such as those frequented by children.

Post-moderation

In cases where user engagement is important and yet a comprehensive moderation program is still required, post-moderation is often a good choice. Post-moderation allows users to publish their submissions immediately, but also adds them to a queue for moderation. Take note, however — since every comment is approved by a moderator, scalability can be an issue.

Reactive moderation

For a scalable program that relies on community members, consider reactive moderation. This type of moderation asks users to flag any content that they find offensive or that breaches community guidelines. By involving users in the process, reactive moderation directs moderator efforts towards the content that needs their attention the most. Note that with reactive moderation there's a risk that offensive content will remain viewable for long periods of time.



Two-thirds (66%) of those surveyed by TELUS International said that they have flagged inappropriate or inaccurate user-generated content — a compelling case for empowering your team members.



The vast majority (69%) of respondents expect content to be removed by brands when flagged as inappropriate or inaccurate.

Supervisor moderation

Similar to reactive moderation, supervisor moderation involves selecting a group of moderators from your online community and empowering them with enhanced capabilities to enforce your standards. Also known as unilateral moderation, this system gives certain users special privileges to edit or delete submissions as they use your site. If supervisors are selected carefully, this method can result in the prompt removal of sensitive content. While supervisor moderation can be scaled as your community grows, there's also an inherent risk of noncompliant content lingering for longer than it should.



Commercial content moderation

Commercial content moderation (CCM) involves monitoring content for large, established brands like social media platforms, games companies and other tech giants. It is often outsourced to specialists who are tasked with ensuring that the content on a platform abides by community guidelines, user agreements and legal frameworks for that particular site and its markets. Deploying a CCM solution relies on partnership and can be particularly effective when an experienced and knowledgeable provider is selected.

Distributed moderation

As one of the most hands-off moderation systems, distributed moderation places a lot of trust and control in the user community. It usually involves enabling users to rate or vote on submissions that they see and flag content that goes against any guidelines that are in place. This often takes place under the guidance of experienced moderators and can work well if your site has a large and active community. With that said, distributed moderation systems remain somewhat rare given the possible risks of allowing a community to almost entirely self-moderate.

Automated moderation

Automated moderation is an increasingly popular moderation method. As the name suggests, it involves the use of a variety of tools to filter, flag and reject user submissions. These tools can range from simple filters that search for banned words or block certain IP addresses, to machine learning algorithms that detect inappropriate content in images and video. Many of these tools are used in addition to some form of human moderation. Human-in-the-loop approaches are particularly useful as they enable human moderators to vet the performance of machine learning algorithms while they are being fed training data and perfected over time.



The human component


For a complete understanding of what's appropriate, cultural, regional and socio-political nuances must be taken into account — and for that you need people. An elite team of content moderators will often cover several languages and have a deep understanding of their assigned geographies.

Human moderators can catch toxic content that technology might miss, report on trends, escalate serious issues and make delicate judgement calls about suspending or banning users from an online community.


Hiring the right team of moderators

There is a defined profile ideally suited for success in content moderation. Recruitment starts with honing that profile to the unique needs of your brand, a task that can be accomplished with the help of a proficient partner.


Top content moderators have the following qualities:




Digital experience: Team members will be reviewing UGC on one digital platform or another; this alone necessitates a comfort with technology. What's more, moderators need to be adept at using the suite of tools that help them keep communities safe.



Problem solving capabilities: The digital world is dynamic, which means your moderators need to be able to think on their feet to protect evolving communities. A proficient moderator must be passionate about solving new problems everyday.



Decision making skills: The world of content moderation isn't always clear-cut. Moderators who thrive will understand relevant laws and community guidelines and be able to make decisions in the best interest of your brand and community.



Community advocacy: The best content moderators are passionate brand advocates for the communities that they moderate and are determined to ensure that their members remain safe.

Keeping your content moderators engaged and healthy

Great corporate cultures surround team members with the things that matter and do what they can to keep them highly engaged. The more a company's stated values align with its real culture on the ground, the higher it drives measured team member engagement and the longer it extends tenure. This is especially important in content moderation, a field in which team members are faced with the task of keeping communities safe.

It needs to be said: The work that content moderators perform can be as challenging as it is essential. Acting as unsung heroes of the internet, moderators scrub communities of inappropriate and inaccurate content so that others don't have to see it. With the rise of UGC, it can be difficult to keep noncompliant content at bay. Your moderators need support.

To keep your content moderators engaged and healthy, instill a caring culture that prioritizes the well-being of team members. A comprehensive program will be genuine in delivering the following key elements:

Encourage physical health by offering gyms and fitness classes, ergonomic workspaces and dynamic social activities.

Support mental health through on-site psychologists and counseling, relaxation areas and quiet rooms.

Inspire career success with access to learning materials, ongoing training and goal setting.

Make technology available that helps team members execute their responsibilities.

By hiring team members with the right temperament, skills and experience, and subsequently taking care of their well-being, brands place themselves in position to protect their communities. But to take content moderation to the next level, support your moderators and your overall content moderation efforts with artificial intelligence (AI).





The role of artificial intelligence

No matter the size or skill of your content moderation team, the sheer volume of UGC today can make it difficult for brands to keep pace. There are simply not enough hours in the day. That's why exceptional moderation calls for the best mix of human intervention and technology automation to ensure content remains appropriate and relevant.

The absolute size and scale of the data AI can interpret is its biggest benefit when it comes to content moderation, but there are certain categories of material that AI truly excels at detecting. Brands and social networks can harness AI's ability to execute rules on restricted words, recognize questionable language and images, send out alerts and put comments into pending, delete or approve mode.

The next step in the development of AI solutions involves building out its capabilities in dealing with some of the more complicated instances of inappropriate content, such as cyberbullying. Future content moderation tools will be able to calculate a relative 'risk' score for a piece of content, before determining when and if it should be reviewed.

For now, the complexity of content moderation means that full automation remains a dream from the distant future. However, as the industry adapts to protect its workers, human-in-the-loop moderation workflows will become far more common. These approaches build on the strengths of both human and machine, allowing the algorithm to square away a large portion of inappropriate content while referring difficult, subjective content to their human overseers. In this way, the continuing expansion of AI in this field will help to protect not only internet users, but also moderators from the worst of the world of content. We're not there yet, but through concerted effort there is reason to be optimistic.

Measuring success

When it comes to content moderation, start by measuring what matters most. That means prioritizing employee well-being, first and foremost. From there, focus on delivering quality results around a selection of KPIs that make most sense for your business.



Content moderation accuracy: As the name suggests, content moderation accuracy is all about the degree to which moderators or technology can correctly tag UGC according to policies and guidelines. High accuracy demonstrates alignment between moderators and brand expectations.

Average Handle Time (AHT): A metric designed to track efficiency, Average Handle Time measures the time moderators spend reviewing each case. The lower the AHT, the more transactions and cases a moderation team can work through.



Moderator ratings: When community members are encouraged to evaluate actions completed by your moderators, you develop a clear sense of their contribution. Moderator ratings often come in the form of stars, where the greater number of stars corresponds to a more positive rating.

First Reply Escalation Rate: In content moderation, First Reply Escalation Rate denotes how frequently a moderator needs to seek a second opinion to make a decision about removing UGC. Note that this is only the beginning: Once team members begin the escalation process, it is crucial that they take a proactive approach to understanding the original issue, the appropriate steps to take in the future, and ultimately, ensure a complete fulfillment of the escalation or original issue raised.





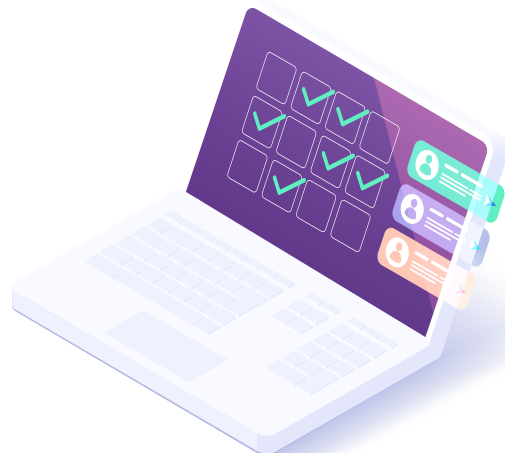
Internal quality reviews: Internal quality reviews involve taking a close look at the various KPIs to evaluate how accurately team members adhere to a brand's policies. Keeping a keen eye on quality ensures that brand and user expectations are met.

Coaching and mentorship: Content moderation takes practice. Through coaching and mentorship, moderators benefit from the best chance of success. Set a regular cadence for casual check-ins as well as formal coaching sessions so your team continues to thrive.



Long tenure: The tenure of team members must be monitored closely as it is a critical indicator of engagement and well-being of your moderators. Long tenure suggests you're on the right track.

Support and engagement: Regularly measure employee engagement through surveys — ideally conducted by an impartial third party — to gain insight into how your team members feel about your organization and key areas to improve.



Measuring the success of your efforts is highly contextual in content moderation. Different metrics are important for different types of businesses and different moderation methods. To make sense of the complexity, look to a proven partner that can provide guidance.



Safety first

The safety of moderators and community members should come first.

That's why hiring team members with the right temperament, skills and experience contributes to creating a resilient team. From there, be sure to have a strategy for workflow rotation and implement best practices from health and safety experts to keep your moderators from burning out. Do what is in your power to protect the people who protect your community.



Selecting the right content moderation partner

A qualified trust and safety partner can take your existing content moderation operations to the next level, or build them from the ground up. No matter the unique needs of your brand, there are essential qualities that an effective partner should have.

For starters, a partner must be able to scale with your community, and quickly. Failing to keep up with the pace of UGC means letting inaccurate or inappropriate content linger, and as our survey results have shown, that can damage customer sentiment and ultimately your community.

A partner must also have the digital tools for efficient and effective moderation. An understanding of online communities and behaviors is non-negotiable, but a qualified partner must also be able to deploy leading-edge technologies to achieve your unique business goals and make the lives of moderators easier.

And, of course, a partner must have demonstrable capabilities in maintaining the well-being of content moderators. A good outsourcing partner that pays careful attention to building its own company culture can actually give client companies more than they expect.

Once you've found a partner that can scale, lead with technology and understand the human component, you're on the right track for content moderation success.

Maintain trust and sustain growth

In the end, safe, trusted solutions promote improved user experiences — driving both growth and revenue. TELUS International's mix of resilient human support combined with automated digital moderation tools provides a robust trust and safety framework to monitor your digital business.

There's no doubt that success requires data and expertise. Benefit from our ability to deliver both. With the help of our AI Data experts, you can leverage custom datasets across all data types to improve your content moderation systems or even build them from the ground up.

Overall, our hybrid model, deploying humans and technology in tandem, can create the high-tech, high-touch CX that sets your brand apart. With operations in over 25 countries and the ability to provide support in over 50 CX languages and 500 data annotation languages and dialects, we are a trusted partner to disruptive brands in a range of industries.

If you're interested in discussing how to launch or enhance your own content moderation and/or trust & safety programs, please reach out:

telusinternational.com/contact

