ONLINE PERSONALS WATCH
Illuminating the top news and ideas since 2004

USER GENERATED CONTENT MODERATION SOLUTIONS FOR THE ONLINE DATING INDUSTRY

2017

Introduction

Content moderation is a multi-faceted activity that resides at the heart of community building efforts for online dating companies. There is an array of variables that must be taken into consideration when evaluating moderation solutions. User generated content moderation services are available from several providers, including: Besedo, Crisp Thinking, Crowd Flower, Crowd Source, Foiwe, Scamalytics, SightEngine, and WebPurify.

Comparison of User Generated Content Moderation Service Providers

		Besedo	Crisp Thinking	Crowd Flower	Crowd Source	Foiwe	Scamalytics	SightEngine	WebPurify
User Generated Content	Text	✓	✓	✓	✓	✓	✓		✓
	Image	✓	✓	✓	✓	✓	✓	✓	✓
	Video	✓		✓	✓	✓	✓	✓	✓
	Profile	✓	✓	✓		✓	✓		✓
	IM/Chat	✓	✓			✓			✓
Methods	Pre-Moderation	✓		✓		✓	✓	✓	✓
	Post-Moderation	✓	✓	✓	✓	✓	✓	✓	✓
	Reactive Moderation	✓		✓	✓	✓	✓	✓	✓
Human Moderation	Dedicated Team	✓							
	Homesourced	✓	✓			✓			✓
	Crowdsourced			✓	✓				
AI Moderation	Tailored AI	✓		✓					✓
	Generic AI	✓	✓	✓			✓	✓	✓
Integration	All-in-One Interface	✓							
	API	✓		✓	✓		✓	✓	✓
Additional Services	Custom Safety Team	✓							
	AI Mgmt Team	✓							
	Account Takeover Detection	✓							
	Law Enforcement Collaboration	✓							
	Fraudster Profiling	✓	✓				✓		

Content Moderation

Objectives

1. What is content moderation?
2. Why moderate?
3. What to moderate?
4. When to moderate?
5. How to moderate?
6. Who should moderate?
7. What are the costs of moderation?

1. What is Content Moderation?

Content moderation is the process of evaluating the text, images, videos, profiles, and other content uploaded by users to ensure it does not violate legal, safety, cultural, or community standards. There are many considerations when deciding what content should be moderated, when it should occur, how it will be handled, and who should accomplish the moderation. A comprehensive moderation solution validates the quality and relevancy of UGC, building trust with users and improving their experience.

2. Why Moderate?

Losses from Online Romance Scams

U.S. (2016): $230 million

AUS (as of Sep. 2017): $14 million

For most players in the online dating industry, content moderation is an absolute necessity. It is guaranteed that a portion of a site's users will intentionally, or unintentionally, upload content that violates policies or community standards.

Scammers frequently attempt to lure users onto external communication channels where they build a fraudulent relationship leading to requests for money or blackmail. Other common examples of harmful content include the posting links to nefarious websites, spammers posting ads for prostitution or pornography, and legitimate users posting images that are not acceptable. In 2016, the FBI's Internet Crime Complaint Center received nearly 15,000 complaints of romance scams with over $230 million in associated losses. This is an increase of 2,500 incidents from 2015. In Australia, ScamWatch is reporting over $14 million in losses from 2,819 reports of dating and romance scams as of September 2017. Online dating companies are tasked with ensuring that illegal, illicit, inappropriate, and irrelevant content does not infiltrate their communities.

Unique Challenges for Online Dating Companies

Online dating companies face unique and salient challenges with content moderation. Members of online dating sites have a standing need to both view and publish a steady stream of up-to-date photos. These photos are typically highly personal and intimate. As a result, online dating companies must process volumes of potentially risqué user generated content. Also, users are particularly sensitive about the types of communities they are entering. For example, users seeking serious committed relationships will immediately stop using a dating app featuring sexually provocative or controversial member photos on the home screen. Consequently, dating sites face a perpetual need to ensure that the content users routinely upload and interact with is safe, legal, and appropriate for the community.

Benefits of Successful Moderation

Successful moderation produces a wealth of benefits for online dating companies. Reducing the amount of offensive and inappropriate content that is visible to users promotes trust in the community. This leads to increased member retention, which is great for the bottom line. Moderation also keeps spammers and scammers at bay. These malicious individuals can drive users away in droves. Disgruntled users often turn to social media to publicly denounce brands. Dating sites that are dependent on advertising partnerships benefit immensely from moderation. A single illegal or inappropriate image could cost them their ad sponsors.

Moderation has benefits that go well beyond simply removing unwanted content. It is frequently used to curate user content and reroute it to more effective locations. For example, a user on a dating app uploads a photo with nudity. Rather than being immediately deleted, the photo is automatically rerouted to a private folder that is only visible to trusted users. The obvious benefit is that users can maintain their content without having it harm the broader community.

Costs of Insufficient Moderation

Failure to effectively moderate introduces significant risk, not only to an online dating company's brand and its finances, but its users as well. The dating app Skout came under fire in 2012 after several children were raped and molested by people they met on the app[1]. Facebook also took a hit in 2012 when it was discovered that moderators they had outsourced for $1/hour on the website Odesk.com were able to capture and publicly share potentially innocent users' sensitive photos and contact details. In that case, weak moderation guidelines and poor control mechanisms were to blame[2]. Scandals can lead to devastating losses in Monthly Active Users, advertising revenue and investment capital. The costs of mitigating negative PR can be staggering as well.

> *"Some of the photos that people post, which under Facebook's rules may be deemed inappropriate, such as your children running around naked or a mum breastfeeding, could end up on the open internet if a moderator, who is able to copy the images, publishes them[2]."*

Another significant fear among online dating companies is that of being rejected from app marketplaces. Both Apple and Google have banned dating apps from their stores for failing to adequately moderate their user generated content. Apple's terms explicitly state, "Apps with user-generated content present particular challenges, ranging from intellectual property infringement to anonymous bullying. To prevent abuse, apps with user-generated content or social networking services must include: a) a method for filtering objectionable material from being posted to the app, b) a mechanism to report offensive content and timely responses to concerns, and c) the ability to block abusive users from the service.[3]." If an online dating company does not moderate its UGC effectively or expediently, it runs a considerable risk of being banned. Clearly, content moderation must be a top priority.

[1] https://bits.blogs.nytimes.com/2012/06/12/after-rapes-involving-children-skout-a-flirting-app-faces-crisis/

[2] http://www.telegraph.co.uk/technology/facebook/9119090/Facebook-in-new-row-over-sharing-users-data-with-moderators.html

[3] https://developer.apple.com/app-store/review/guidelines/#user-generated-content

3. What to Moderate?

Dating operators must first ask themselves what types of UGC requires moderation. Below are some popular types of content that qualify for moderation:

- Text may be moderated for profanity, personally identifiable information (i.e. phone numbers and email addresses), and all manner of inappropriate slang and hate speech.
- Images and Videos may be moderated for nudity, pornography, violence, illegal acts, cultural insensitivity, and spam overlays. Images may also be moderated for general community standards, such as having the user's face in the photo.
- Profiles may be moderated for authenticity and general community standards.

All categories of content can be used to identify and remove scammers before they are able to reach users.

4. When to Moderate?

Knowing when to moderate is perhaps one of the most important variables an online dating company must decide. If moderation takes place before content gets posted, the delay could cripple the natural flow of community interaction. If moderation comes in too late, it leaves illegal and inappropriate content out in the open to be encountered by hundreds, if not thousands of unsuspecting users. There are four primary moderation windows, described in more detail below.

Pre-moderation

Pre-moderation takes place before UGC ever goes live on the site, and in most cases, every single piece of uploaded content gets reviewed. Pre-moderation is considered the safest form of moderation because it can prevent users from ever seeing anything inappropriate, illegal, or in any way in violation of community standards.

For online dating companies that fear being banned from app stores for inappropriate UGC, pre-moderation is a clear necessity. Likewise, when it comes to sites that cater to children and minors, where there are high risks associated with exposure to dangerous or inappropriate content, pre-moderation is a non-negotiable. Pre-moderation is also quite beneficial for companies that could be sued for having copyrighted content appear on their pages. Amazon, IMDB, and Shutterstock all employ pre-moderation to avoid any possible lawsuits or copyright infringements.

It's important to ensure that pre-moderation doesn't slow the pace of community interaction. Rapid action and collaboration can be difficult if there's a significant delay between when a user uploads content and when others in the community get to see it. Pre-moderation of content should happen as quickly as possible. Ideally, it should occur within five minutes of any UGC uploads.

Pre-moderation monitors every piece of UGC as opposed to just content that's been flagged. This can be great for small companies with low volumes of UGC. However, companies looking to conduct their moderation in-house may find difficulty with scaling, and would likely benefit from finding a moderation provider that could readily adjust its capacities to fit their specific moderation needs.

Post-moderation

Post-moderation allows all uploaded content to go live immediately, and each piece of content is subsequently moderated within the first few minutes or hours after it is uploaded. When a site employs post-moderation, its users may all upload content and respond in real time, which is great for communities where users demand to interact with UGC immediately. However, this can be slightly riskier than pre-moderation because the content is still potentially getting seen by unsuspecting users during the time it is live on the site.

Reactive Moderation

Reactive moderation allows any UGC to remain live on site from the moment it is uploaded until the moment someone in the community flags or reports it. Generally, only a small fraction of objectionable content gets reported, and so inappropriate content runs the risk of remaining live on the site for longer durations. Reactive moderation is used by dating apps like Grindr where massive number of images are uploaded each day, and the cost of having inappropriate images remain live is relatively low.

Reactive moderation does offer unique community insights. In reactive moderation, the community of users is responsible for reporting inappropriate content rather than any internal or external service, so companies can learn a lot about their users' preferences based on what they tend to report. For many companies, however, reactive moderation is simply unacceptable. Essentially, by relegating the role of abuse detector to your innocent and unsuspecting users, you're forcing the people who are least likely to want to see inappropriate content to be the primary discovery crew for it. Additionally, reactive moderation comes with a lot of false positives. Users can report content for a million reasons that may have nothing at all to do with content.

User Moderation

User moderation is like reactive moderation, except rather than having an outside service moderate the content that's been flagged, the users themselves would get to decide whether the flagged content should be removed. Both reactive and user moderation are much easier to scale, because the number of problematic photos doesn't necessarily increase in tandem with the total number of uploaded photos. However, the largest drawback of user moderation is that abusive and inappropriate content may stay live on site for far longer than anyone may want because formal removal requires the collective flags and confirmations of multiple moderators who may have conflicting opinions.

5. How to Moderate?

Artificial Intelligence (AI) Moderation

The benefit of algorithms is that they can be cheaper and easier to scale up or down to fit company needs. However, AI algorithms must be meticulously trained and optimized through machine learning. There are two categories of AI moderation: generic and tailored. Generic AI systems provide a core functionality that is offered to all clients. They oftentimes fail to detect context and nuance, resulting in more false positives and false negatives. For example, with a sample filter of "no weapons," a generic AI algorithm might unnecessarily remove a photo of a police officer demonstrating in a classroom. A weapon is present in the photo, but it is not displayed in a threatening or violent manner. A human moderator would understand that nuance and allow the photo. Additionally, a generic AI algorithm may be useful for detecting the percentage of skin exposed, but it does nothing for detecting hate crimes or gestures. Tailored AI moderation systems are customized to meet the requirements of a single client. The parameters and rulesets are modified to accommodate specific markets and communities. Tailored AI algorithms provide an increased capacity to operate across many types of content with extremely high accuracy.

Human moderators are typically costlier than algorithms, but they can provide a level of intuition and adaptability that makes them highly worthwhile. Most major moderation providers have large, scalable teams of human moderators, either in the form of trained full-time, in-house employees, or crowdsourced freelancers who can immediately participate in moderation from anywhere in the world.

Full-time Human Moderation Teams

The benefits of a trained, specialized force of full-time moderators are quite significant. They get to understand the ins and outs of individual client sites, and they can readily moderate virtually any type of content, whether text, images, photos, videos, or entire profiles. Because they work full-time and usually on specialized machines, they're able to offer an advanced level of security with user data that helps sites avoid the risks that options like crowdsourcing frequently entail. Full-time moderators can be quickly trained to accommodate new or changing requirements with nearly immediate implementation. It's essential, though, that moderation workforces be adequately trained, supplied, and directed. For the highest level of consistency and niche sensitivity, a dedicated moderation team is needed.

Crowdsourced Freelance Moderators

Crowdsourced moderation has the advantage of being rapidly scalable, because anyone anywhere can do the moderating. It is also quite flexible, as new rules can be implemented anytime and the crowdsourced moderators can quickly learn those rules and adjust their moderating accordingly. The changes may not be immediately applied to existing moderation task queues already in progress. Crowdsourcing's advantages are also some of its greatest drawbacks. Due to the distributed, work-whenever setup of crowdsourced moderators, they may not get to fully know their client sites, and thus may take much longer (and thus cost more) to reach a consensus about which pieces of UGC are acceptable. If proper incentives and behavioral checks are not built into the crowdsourcing system, the moderators could also aim for maximizing their sheer number of moderated items rather than the accuracy of their moderation. Lastly, the relatively anonymous and distributed nature of crowdsourcing can be dangerous for user privacy, as moderators are gaining regular access to users' sensitive content on their own personal devices.

6. Who Should Moderate?

The next major decision to be made is whether an online dating company should moderate in-house or partner with an outside company. Outside companies either offer content moderation through their own platform & staff, or they direct their staff to plug into a client's existing system. Building out a custom moderation apparatus for your own site or app can be an extremely complex undertaking. Options include having it a) used by your own team internally, b) outsourced to another moderation provider's full-time staff, and c) provided to your users to do the moderation themselves. Each of these would require a different buildout with its own developmental challenges (and potential advantages, too). Many large and established online dating companies prefer to moderate internally because they can reduce the risk of their users' private content getting into the wrong hands.

Internal moderation, however, can be quite expensive and difficult to scale because any growth in the amount of uploaded UGC must be directly tied to an increase in the workforce that needs to moderate it.

Executive Summary

A full-spectrum user generated content moderation strategy is the result of an active partnership between business and the service provider. Every organization has unique circumstances and requirements, highlighting the importance of a flexible platform with customizable rules and scoring systems.

The Online Dating industry faced many unique and sensitive challenges which require comprehensive solutions. Niche and cultural sensitivity is of elevated importance on most dating sites. Partnering with a dedicated UGC moderation service provider can yield substantial benefits. Protecting customers, strengthening communities, and thwarting scammers has the potential to greatly improve user experience, satisfaction, and retention.

Disclosure

Courtland Brooks was engaged by Besedo to author this white paper. Data for the comparison chart was obtained through information published by each company.