Title
Content moderation

ISBN
978-3-319-32010-6

Author
Roberts, Sarah T.

Publication Date
2017-02-05

Title

Content Moderation

Synonyms

Content screening; community management; community moderation

Author and Affiliation

Sarah T. Roberts
Department of Information Studies
University of California, Los Angeles
sarah.roberts@ucla.edu

Definition

Content moderation is the organized practice of screening user-generated content (UGC) posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction. The process can result in UGC being removed by a moderator, acting as an agent of the platform or site in question. Increasingly, social media platforms rely on massive quantities of UGC data to populate them and to drive user engagement; with that increase has come the need for platforms and sites to enforce their rules and relevant or applicable laws, as the posting of inappropriate content is considered a major source of liability.

The style of moderation can vary from site to site, and from platform to platform, as rules around what UGC is allowed are often set at a site or platform level, and reflect that platform's brand and reputation, its tolerance for risk, and the type of user engagement it wishes to attract. In some cases, content moderation may take place in haphazard, disorganized or inconsistent ways; in others, content moderation is a highly organized, routinized and specific process. Content moderation may be undertaken by volunteers or, increasingly, in a commercial context by individuals or firms who receive remuneration for their services. The latter practice is known as commercial content moderation, or CCM. The firms who own social media sites and platforms that solicit UGC employ content moderation as a means to protect the firm from liability, negative publicity, and to curate and control user experience

History

The internet and its many underlying technologies are highly codified and protocol-reliant spaces with regard to how data are transmitted within it (Galloway, 2009), yet the subject matter and nature of content itself has historically enjoyed a much greater freedom. Indeed, a central claim to the early promise of the internet as espoused by many of its proponents was that it was highly resistant, as a foundational part of its ethos, to censorship of any kind.

Nevertheless, various forms content moderation occurred in early online communities. Such content moderation was frequently undertaken by volunteers, and was typically based on the enforcement of local rules of engagement around community norms and user behavior. Moderation practices and style therefore developed locally among communities and their

participants and could inform the flavor of a given community, from the highly rule-bound to the anarchic; the Bay Area-based online community the WELL famously banned only three users in its first six years of existence, and then only temporarily (Turner 2005, p. 499).

In social communities on the early text-based internet, mechanisms to enact moderation was often direct and visible to the user, and could include demanding that a user alter a contribution to eliminate offensive or insulting material, the deletion or removal of posts, the banning of users (by username or IP address), the use of text filters to disallow posting of specific types of words or content, and other overt moderation actions. Examples of sites of this sort of content moderation include many Usenet groups, BBSes, MUDs, listservs and various early commercial services.

Motives for people participating in voluntary moderation activities varied. In some cases, users carried out content moderation duties for prestige, status or altruistic purposes (i.e., for the betterment of the community); in others, moderators received non-monetary compensation, such as free or reduced-fee access to online services, e.g., AOL (Postigo, 2003). The voluntary model of content moderation persists today in many online communities and platforms; one such high-profile site where volunteer content moderation is used exclusively to control site content is Wikipedia.

As the internet has grown into large-scale adoption and a massive economic engine, the desire for major mainstream platforms to control the UGC that they host and disseminate has also grown exponentially. Early on in the proliferation of so-called Web 2.0 sites, newspapers and other news media outlets, in particular, began noticing a significant problem with their online comments areas, which often devolved into unreadable spaces filled with invective, racist and sexist diatribes, name-calling and irrelevant postings. These media firms began to employ a variety of techniques to combat what they viewed as the misappropriation of the comments spaces, using in-house moderators, turning to firms that specialized in the large-scale management of such interactive areas, and deploying technological interventions such as word filter lists or disallowing anonymous posting, to bring the comments sections under control. Some media outlets went the opposite way, preferring instead to close their comments sections altogether.

Commercial Content Moderation and the Contemporary Social Media Landscape

The battle with text-based comments was just the beginning of a much larger issue. The rise of Friendster, MySpace and other social media applications in the early part of the 21st century has given way to more enduring social media platforms of enormous scale and reach. As of the second quarter of 2016, Facebook alone is approaching 2 billion users worldwide, all of whom generate content by virtue of their participation on the platform. YouTube reported receiving upwards of 100 hours of UGC video per minute as of 2014.

The contemporary social media landscape is therefore characterized by vast amounts of UGC uploads made by billions of users to massively popular commercial internet sites and social media platforms with a global reach. Mainstream platforms, often owned by publicly-traded firms responsible to shareholders, simply cannot afford the risk – legal, financial and to reputation – that unchecked UGC could cause. Yet contending with the staggering amounts of transmitted data from users to platforms is not a task that can be addressed at large scale by

computers. Indeed, making nuanced decisions about what UGC is acceptable and what is not currently exceeds the abilities of machine-driven processes, save for the application of some algorithmically-informed filters or bit-by-bit matching, both of which occur at relatively low levels of computational complexity.

The need for adjudication of UGC – video and image-based content, in particular – therefore calls on human actors who rely upon their own linguistic and cultural knowledge and competencies to make decisions about UGC's appropriateness for a given site or platform. Specifically, "they must be experts in matters of taste of the site's presumed audience, have cultural knowledge about location of origin of the platform and of the audience (both of which may be very far removed, geographically and culturally, from where the screening is taking place), have linguistic competency in the language of the UGC (that may be a learned or second language for the content moderator), be steeped in the relevant laws governing the site's location of origin and be experts in the user guidelines and other platform-level specifics concerning what is and is not allowed" (Roberts, 2016). These human workers are the people who make up the legions of commercial content moderators: moderators who work in an organized way, for pay, on behalf of the world's largest social media firms, apps and websites who solicit UGC.

CCM processes may take place prior to material being submitted for inclusion or distribution on a site, or they may take place after material has already been uploaded, particularly on high-volume sites. Specifically, content moderation may be triggered as the result of complaints about material from site moderators or other site administrators, from external parties (e.g., companies alleging misappropriation of material they own; from law enforcement; from government actors), or from other users themselves who are disturbed or concerned by what they have seen and then invoke protocols or mechanisms on a site, such as the 'flagging' of content, to prompt a review by moderators (Crawford and Gillespie, 2016). In this regard, moderation practices are often uneven, and the removal of UGC may reasonably be likened to censorship, particularly when it is undertaken in order to suppress speech, political opinions or other expressions that threaten the status quo.

CCM workers are called upon to match and adjudicate volumes of content, typically at rapid speed, against the specific rules or community guidelines of the platform for which they labor. They must also be aware of the laws and statutes that may govern the geographic or national location from where the content emanates, for which the content is destined, and for where the platform or site is located – all of which may be distinct places in the world. They must be aware of the platform's tolerance for risk, the expectations of the platform for whether or how CCM workers should make their presence known.

In many cases, CCM workers may work at organizational arm's length from the platforms they moderate. Labor arrangements in CCM have workers located at great distances from the headquarters of the platforms for which they are responsible, in places such as the Philippines and India. The workers may be structurally removed from those firms, as well, via outsourcing companies who take on CCM contracts and then hire the workers under their auspices, in call center (often called BPO, or Business Process Outsourcing) environments. Such outsourcing firms may also recruit CCM workers using digital piecework sites such as Amazon Mechanical Turk or Upwork, in which the relationships between the social media firms, the outsourcing company and the CCM worker can be as ephemeral as one review.

Even when CCM workers are located on-site at a headquarters of a social media firm, they often are brought on as contract laborers and are not afforded the full status, or pay, of a regular full-time employee. In this regard, CCM work, wherever it takes place in the world, often shares the characteristic of being relatively low-wage and low-status, as compared to other jobs in tech. These arrangements can pose a risk for workers, who can be exposed to disturbing and shocking material as a condition of their CCM work, but can be a benefit to the social media firms who require their labor, as they can distance themselves from the impact of the CCM work on the workers. Further, the working conditions, practices and existence of CCM workers in social media are little known to the general public, a fact that is often by design. CCM workers are frequently compelled to sign NDAs, or non-disclosure agreements, that preclude them from discussing the work that they do or the conditions in which they do it. While social media firms often gesture at the need to maintain secrecy surrounding the exact nature of their moderation practices and the mechanisms they used to undertake them, claiming the possibility of users' being able to game the system and beat the rules if armed with such knowledge, the net result is that CCM workers labor in secret. The conditions of their work – its pace, the nature of the content they screen, the volume of material to be reviewed, the secrecy – can lead to feelings of isolation, burnout and depression among some CCM workers. Such feelings can be enhanced by the fact that few people know such work exists, assuming, if they think of it at all, that algorithmically-driven computer programs take care of social media's moderation needs. It is a misconception that the industry has been slow to correct.

Conclusion

Despite claims and conventional wisdom to the contrary, content moderation has likely always existed in some form on the social internet. As the internet's many social media platforms grow and their financial, political and social stakes increase, the undertaking of organized control of user expression through such practices as CCM will only increase. Nevertheless, CCM remains a little discussed and little acknowledged aspect of the social media production chain, despite its mission-critical status in almost every case in which it is employed. The existence of a globalized CCM workforce abuts many difficult, existential questions about the nature of the internet itself, and the principles that have long been thought to undergird it; particularly, the free expression and circulation of material, thought, and ideas. These questions are further complicated by the pressures related to contested notions of jurisdiction, borders, application and enforcement of laws, social norms and mores that frequently vary and often are in conflict with each other. The acknowledgement and understanding of the history of content moderation and the contemporary reality of large-scale CCM is central to many of these core questions of what the internet has been, is now, and will be in the future, and yet their continued invisibility and lack of acknowledgment by the firms for whom their labor is essential means that such questions cannot fully be addressed. The addition of discussions of moderation practices and the people who undertake them is essential to the end of more robust, nuanced understandings of the state of the contemporary internet, and to better policy and governance based on those understandings.

--Sarah T. Roberts, Department of Information Studies, University of California at Los Angeles

See Also: Algorithm; Facebook; Internet; Social media; YouTube; Wikipedia

References

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428.

Galloway, A. R. (2006). *Protocol: How control exists after decentralization*. Cambridge, MA: MIT Press.

Postigo, H. (2003). Emerging sources of labor on the internet: The case of America Online volunteers. *International Review of Social History*, *48*(S11), 205–223.

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. Tynes (Eds.), *The Intersectional internet: Race, sex, class and culture online* (pp. 147-160). New York, NY: Peter Lang.

Turner, F. (2005). Where the counterculture met the new economy: The WELL and the origins of virtual community. *Technology and Culture*, *46*(3), 485–512.