Content moderation – A perspective on the present and future implications

Living in an age where internet has become a necessity, we have been exposed to a deluge of online content. While this has opened newer avenues for content creation and consumption, the onus for moderating this content lies on the platform owners.

How do platform owners moderate this colossal amount of content? What are the challenges with content moderation? Who moderates it for platform owners and how are the moderators impacted by their job?

In this article, we discuss the imperatives that affect content moderation and the moderators.

The increase in netizens and social media platforms along with the proliferation of mobile internet has led to an unprecedented upsurge in User Generated Content (UGC) creation and consumption. Mobiles have reduced the time taken between content creation and its consumption to a matter of a few seconds, and made content consumption quite handy. The recently published Global Digital Report by We Are Social, states that a consumer spends an average of 6.5 hours online per day; of which, 2.16 hours is spent on social media.[1]

No wonder then that businesses prefer allocating more marketing and advertising budgets on digital and social media platforms compared to traditional media. Further, User-Generated Content (UGC), like user reviews on products and services and customer support experiences play a vital role in purchase decisions and are hence becoming an integral part of marketing strategies. According to AdWeek, 85% of the users are found to be more influenced by the UGC than the content created by brands directly.[2] This makes it more critical than ever to monitor UGC to fight disinformation and protect brand reputation.

Social media platforms have also emerged as the new source of entertainment. As per a survey conducted by Global Web Index on social media consumption patterns, 36% of the respondents use social media as a source of entertainment, while 56% use social media platforms for video consumption through social media platforms (YouTube, and Instagram's IGTV being the forerunners).[3]

However, with no limit to the diversity of user-generated content, where most of it is unfiltered, consumption for entertainment poses a serious threat to audiences. Audiences are often at the risk of being exposed to anything from mildly abusive and offensive content to hate speeches and extremely violent, mentally disturbing content.

In such times, it is imperative to check if the user-generated content promotes healthy consumption, or spreads malicious content. A preemptive approach to content moderation rather than a reactive approach will help prevent inappropriate content before it is uploaded – nipping it in the bud. Algorithmic advances in latest tech can help platform owners develop mechanisms to screen audience profiles, mindsets and act accordingly.

Challenges with content moderation

Today, several companies have been striking back through third-party firms with an army of human content moderators and latest technologies to monitor social and traditional media for fake viral user generated content.

However, the key challenges with content moderation include:

- Impact on emotional well-being of content moderators, who are exposed to extreme, horrific and explicit content putting them at a high risk of traumatic stress disorder and mental burnout
- Urgency to act: Immediacy of platforms and multitude of content being uploaded makes it almost impossible for an AI + Human augmented model to act within time
- Incongruence in regulatory guidelines and diversified cultural context and language semantics require a hyper local content moderation approach

Digital technologies ensure mental wellness of content moderators

Platform owners and content moderation service providers have developed innovative solutions by leveraging Artificial Intelligence (AI) and Machine Learning (ML) along with technologies like Blockchain and crowdsourcing. These solutions help combat fake content (image, video, text) by using attributes like geo-location, source of occurrence, etc., thus reducing the time taken to moderate content, and improving efficiency and quality of moderation. Implementing AI and ML also helps significantly bring down the overall cost of content moderation.

However, owing to the miscellany of user-generated content in terms of language, geo-political and cultural context, organizations need to be agile enough to build hyper-local moderation models compared to generic onsite-offshore models. It is essential to have associates culturally aligned to the kind of content they will be moderating. In this light, service providers need to revisit their current operating model for content moderation to ensure that associates handling specific geos belong to that particular geo. Just knowing the language is not enough.

Ensuring employee wellness – a critical aspect of content moderation process

Content Moderators are exposed to sensitive and explicit content every single day. There have been several unpleasant episodes where moderators have not been able to deal rationally with the workplace stress, and have succumbed to the trauma related to content viewed.

The use of AI substantially reduces exposure to abusive content viewed by content moderators, by filtering out most of the gory content. AI is also allowing wellness solutions for content moderators, giving them a platform to express their concerns while maintaining anonymity. AI-driven chat bot for conversations and scheduling appointments with the counsellors are helping associates achieve an improved mental health.

The use of AI as a solution in the world of content moderation not only helps with improved operational efficiency, but also takes into account the psychological wellness of the employees to ensure that they have a healthy mental and emotional state. Being in a day and age where internet and online content is integral to daily human activity, content moderators are gatekeepers who ensure harmony prevails, no matter what.

Moving ahead | Building a newer approach to content moderation

We understand the future of content moderation is going to be bionic – constantly evolving capabilities of the AI/ML platforms will be supplanted by human content forensic experts to combat and mitigate the perils of inappropriate content.

We believe platform owners should look at ways to reduce the need for content moderation and move to a preemptive approach rather than a prohibitive one. A propensity and probability based preemptive model needs to be adopted. The attributes that would possibly make it a sound filtering system are identifying:

- pages, users, groups that are likely to spread misinformation
- posts with keywords most likely to be fake or misinterpreted
- events that are likely to cause misinformation/fake news
- statements or statistics most likely to be misinterpreted/misquoted

Modern-day analytics can help platforms build audience-profiling systems that help the AI/ML tools understand their users, their propensities, and likes and dislikes. It will help them know what kind of content the user generally engages with, and would be able to predict his or her next upload or activity. Platforms could also explore the idea of taking users through a short series of questions that gauge their mental state and hence, are able to predict the content being shared.

While these steps might reduce the need for content moderation in the first place and make platforms safer, there is something platform owners need to be wary of. They need to strike the right balance between keeping checks and balances, and providing users a platform to express.

Building a content moderation model that has the right set of technologies – with a proactive, preemptive approach, and right skilled humans - with innate cultural intelligence, extremely skilled in psychological aspects and located strategically, will reduce the complexities involved in the process and contribute to a safer internet ecosystem.

References

https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates[1]

https://www.adweek.com/digital/why-consumers-share-user-generated-content-infographic/[2]

https://www.globalwebindex.com/hubfs/Downloads/Social-H2-2018-report.pdf[3]

About the author

Nirupama Singh
Lead - Business Consultant, CBU
Wipro Limited.

Nirupama is a Lead Business Consultant, working with the Media domain at Wipro Digital Operations and Platforms. She has been working with New Age Businesses for more than five years across the areas of business consulting, product management and strategy. Nirupama has a Master's degree in management from the Indian Institute of Management Raipur in Marketing & Strategy.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035,
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

About Wipro Limited
Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**