An Overview of Factors Affecting Online Content Moderation

Saltuk Karahan, Ph.D
School of Cybersecurity
Old Dominion Univeristy
skarahan@odu.edu

C. Ariel Pinto, Ph.D
Engineering Management & Systems Engineering
Old Dominion Univeristy
cpinto@odu.edu

Hamdi Kavak, Ph.D
Computational and Data Sciences Department
George Mason University
hkavak@gmu.edu

Ida Oesteraas, M.S.
Department of Sociology and Criminal Justice
Old Dominion University
ioestera@odu.edu

Abstract

Online content moderation requires a delicate balance between the right to free speech and abuse of this right (including criminal activities and state sponsored misinformation campaigns). There is not a universal standard policy and the policies for content moderation and the relevant software frameworks are bound by several factors including culture, geography, legislation, scale, when to moderate (before or after the appearance of the content in social media) and nature of the content itself (a variety from child pornography to discriminatory statements and hate speech [including those shaped by state affiliated foreign actors]). While there are promising developments in AI for automated content moderation, in order to develop a software framework, many of these factors and their respective relevance in online content moderation needs to be taken into account in any automation process. This paper aims to conduct a broad literature review and compare the relevant factors for content moderation with their respective weight. The aim of the paper is to conduct exploratory research that will contribute to the design of a software framework for online content moderation.

References

Morrow, G., Swire - Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. Journal of the Association for Information Science and Technology, 73(10), 1365-1386.

Saputra, R., Zaid, M., & Oghenemaro, S. (2022). The Court Online Content Moderation: A Constitutional Framework. Journal of Human Rights, Culture and Legal System, 2(3), 139-148.

Seering, J. (2020). Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. Proc. ACM Hum.-Comput. Interact, 3

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. Computer Law & Security Review, 36, 105374.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

Langvardt, K. (2017). Regulating online content moderation. Geo. LJ, 106, 1353.