Selected Papers of #AoIR2019:
The 20th Annual Conference of the
Association of Internet Researchers
Brisbane, Australia / 2-5 October 2019

ALGORITHMIC COPYRIGHT ENFORCEMENT ON YOUTUBE: USING MACHINE LEARNING TO UNDERSTAND AUTOMATED DECISION-MAKING AT SCALE

Joanne Gray
Queensland University of Technology

Nicolas Suzor
Queensland University of Technology

This paper presents the results of an investigation of algorithmic copyright enforcement on YouTube. We use digital and computational methods to help understand the operation of automated decision-making at scale. We argue that in order to understand complex, automated systems, we require new methods and research infrastructure to understand their operation at scale, over time, and across platforms and jurisdictions.

Technology companies play a major role in governing the internet. The rules of platforms, content hosts, search engines, and telecommunications providers govern how we interact with each other, and they shape the possibilities and nature of public discourse (Gillespie 2018). There is increasing global concern about how the decisions of internet and telecommunications companies impact on human rights. Governments around the world are learning how to influence intermediaries to control the flow of information, often in ways that skirt or avoid constitutional protections for fundamental rights (Elkin-Koren and Haber 2016). Meanwhile, many worry that platforms are not taking sufficient care to protect their users from harm (Suzor et al. 2018), prevent the spread of hate and disinformation, or to protect the interests of marginalised groups (Citron 2014; Marwick and Lewis 2017; Caplan, Hanson, and Donovan 2018; Duguay, Burgess, and Suzor 2018).

Copyright enforcement provides a useful case study of automated decision-making because, in digital media networks, automated tools for detecting copyright infringement are widespread and relatively sophisticated. The operation of these systems, however, is often opaque, and platforms are frequently criticised for unduly silencing legitimate

Suggested Citation (APA): Gray, J & Suzor, N. (2019, October 2-5). *Algorithmic copyright enforcement on Youtube: using machine learning to understand automated decision-making at scale*. Paper presented at AoIR 2019: The 20th Annual Conference of the Association of Internet Researchers. Brisbane, Australia: AoIR. Retrieved from http://spir.aoir.org.

speech through their copyright takedown processes (Urban, Karaganis, and Schofield 2016). Content ID and copyright notice-and-takedown requests play an extremely important role in regulating the content that is visible on YouTube, but there is no easy way to understand how these systems operate.

We use YouTube takedowns as a case study to develop and test an innovative methodology for evaluating automated decision-making. First, we built technical infrastructure to obtain a random sample of 59 million YouTube videos and to test their availability. The initial sample of YouTube videos is derived from YouTube's search API ('list' endpoint). The sample is built by requesting, every ten seconds, 50 videos that were published in a ten second window one minute prior. We then use web-scraping to check the availability of each video two weeks after it was first published, and log whether it was removed, as well as the reason that YouTube gives for its removal. In our sample, 508,406 (1.2%) were unavailable because of a copyright claim (including approximately 80,000 that were geoblocked in the country where our server is based).

Next, we sought to better understand the types of videos that are being removed from YouTube. We used topic modeling techniques (Latent Dirichlet Allocation) on the title and description metadata of 20,000 videos to develop an initial understanding of the main types of content that are blocked. We developed and assessed a series of topic models at varying degrees of granularity. As a first step in developing an initial methodology, we did not seek to comprehensively categorise the broad range of videos into an exclusive set of categories. Instead, we selected five well-defined clusters of content that lend themselves to further computational analysis: gameplay videos, full movies, live sports broadcasts, tutorials on game cheats and copy control circumvention, and highlight clips from popular media. These categories are particularly interesting since they map onto major controversies over automated copyright enforcement.

In the third stage of our work, we trained a machine learning classifier to identify videos in each of these categories across our dataset larger. We made use of state of the art Natural Language Processing techniques, including the newly released Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). BERT provides sentence-level representations trained on massive corpora of Wikipedia articles and books. We use a Transformer attention-based deep learning model to perform classification on YouTube video titles and descriptions, and undertook several rounds of active learning to train the classifier (we ultimately manually labeled approximately 5000 videos to develop an adequate training set). At the time of writing, we have been able to achieve over 93% accuracy on test data.

We deployed the trained model to categorize our entire random sample of 59 million YouTube videos. We validated the model's results by manual review and performed inter-coder reliability checks. In the coming weeks, we will analyse the factors that appear to influence decisions to block content across different categories, as well as changes over time since late 2016. We will make use of statistical learning methods (multinomial logistic regression and Support Vector Machines) to examine the relationship between categories, variables in the metadata that YouTube exposes (duration, genre, likes, dislikes, and so on), and takedown rates. These methods enable us to examine the characteristics of videos that are most likely to be removed through DMCA notices, Content ID removals, and Terms of Service enforcement.

This interdisciplinary research provides an initial case study in developing black-box methodologies (Pasquale 2015) to understand the operation of automated decision-making at scale. We reflect on the utility of different computational approaches to understanding content moderation systems, and seek to identify opportunities for further work that leverages close qualitative analysis in conjunction with large-scale computational processing. This work provides the methodological base for further experimentation with the use of deep neural nets to enable large-scale analysis of the operation of automated systems in the realm of digital media. We hope that this work will improve understanding of a useful and fruitful set of methods to interrogate pressing public policy research questions in the context of content moderation and automated decision-making.

References

Caplan, Robyn, Lauren Hanson, and Joan Donovan. 2018. “Dead Reckoning: Navigating Content Moderation after ‘Fake News.’” Data & Society Research Institute. https://datasociety.net/pubs/oh/DataAndSociety_Dead_Reckoning_2018.pdf.

Citron, Danielle Keats. 2014. Hate Crimes in Cyberspace. Cambridge, Massachusetts ; London, England: Harvard University Press.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” October. https://arxiv.org/abs/1810.04805v1.

Duguay, Stefanie, Jean Burgess, and Nicolas P. Suzor. 2018. “Queer Women’s Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine.” Convergence, June, 1354856518781530. https://doi.org/10.1177/1354856518781530.

Elkin-Koren, Niva, and Eldar Haber. 2016. “Governance by Proxy: Cyber Challenges to Civil Liberties.” Brooklyn Law Review 82 (February): 2016.

Gillespie, Tarleton. 2018. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. 1st edition. New Haven, CT: Yale University Press.

Google. n.d. “Content Removals Due to Copyright.” Google Transparency Report. Accessed October 3, 2018. https://transparencyreport.google.com/copyright/explore?hl=en.

Marwick, Alice E., and Rebecca Lewis. 2017. “Media Manipulation and Disinformation Online.” Data & Society Research Institute. https://datasociety.net/output/media-manipulation-and-disinfo-online/.

Pasquale, Frank. 2015. The Black Box Society. Cambridge, Mass.: Harvard University Press.

Suzor, Nicolas P., Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess, and Tess Van Geelen. 2018. “Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online.” Policy & Internet. https://doi.org/10.1002/poi3.185.

Urban, Jennifer M., Joe Karaganis, and Brianna L. Schofield. 2016. “Notice and Takedown in Everyday Practice.” SSRN Scholarly Paper ID 2755628. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2755628.