Content Moderation

How to achieve scale and affordability

Part II

A Conectys made ***White Paper***

Introduction

One of the biggest logistical questions about content moderation over the last five years has been: "How much should be technology, especially AI, and how much should be human?" At Conectys, we've gone over this question with dozens of clients, and written a few papers addressing it, including a 2021 guide to content moderation and a paper guiding you to what you need from an outsourcing partner.

In short, our view at Conectys is that content moderation can be very nuanced, and it needs to be a mix of human agent and technology, be that more-advanced AI or a basic keyword scraper. AI and machine learning advance every day -- a good thing! -- but they're not "there yet" in terms of everything you would need to moderate, and they're definitely not "there yet" in terms of cultural nuance. The best play for the moment, and probably for the next decade, is a mix of tech and human. That might change based on technological advancements or the rise of the metaverse, but for now, tech + human is the best approach to content moderation.

How we're advancing AI for content moderation

Relatively simple approaches such as "hash matching," in which a fingerprint of an image is compared with a database of known harmful images, and "keyword filtering," in which words that indicate potentially harmful content are used to flag content, are useful tools but have limitations. Detecting the nuances of language, such as sarcasm and the use of emojis, is difficult, and languages in particular slang terms, evolve over time.

Sentiment analysis (the process of detecting whether or text or image is inherently positive or negative) techniques are a focus of research and are becoming increasingly effective. Same with object detection (obvious) and scene understanding (same). More and more data and images are being inputted into these programs, and the programs will continue to learn.

An AI approach known as "recurrent neural networks" can enable more sophisticated analysis of video content, which is particularly challenging to moderate as frames must be considered relative to other frames in the video.

AI moderation techniques can also increasingly consider the context in which content appears, although in general this remains complex and challenging. In practice, most harmful content is generated by a minority of users and so AI techniques can be used to identify malicious users and prioritize their content for review. Metadata encodes some context relevant to moderation decisions about content, such as a user's history on the site, the number of friends or followers they have and information about the user's real identity, such as age or location.

There is also a cultural and historical context in many online interactions. Any preceding content, such as previous interactions between individual users or the flow of the discussion, can provide valuable context which can be analyzed alongside the content itself. The metadata available varies between platforms and for the type of content posted and so it is difficult for platform-agnostic moderation tools to take full advantage of metadata when making moderation decisions.

Different AI architectures are required for identifying different categories of potentially harmful content. For example, identifying child abuse material requires consideration of the content (an image or video) but in general the context (such as the age or location of the user posting it or the number of followers they have) is not an important factor in detecting it automatically. AI techniques such as 'object detection' and 'scene understanding' are essential elements of automated systems to identify this type of material. On the other hand, identifying bullying content often requires full consideration of the context of the user interactions as well as the content itself as the characteristics of bullying content are less well-defined.

The complexity of designing AI architectures for moderating different categories of content therefore increases the costs and challenges for organizations to develop these.

How does AI benefit human moderators now?

AI can improve the effectiveness of human moderators by prioritizing content to be reviewed by them based on the level of harmfulness perceived in the content or the level of uncertainty from an automated moderation stage. It can reduce the impact on human moderators by varying the level and type of harmful content they are exposed to. It can limit exposure to the most harmful elements of the content, such as by identifying and blurring out areas of images which the moderator can optionally view only if needed to make a moderation decision.

An AI technique known as "visual question answering" allows humans to ask the system questions about the content to determine its degree of harmfulness without viewing it directly. This can reduce the harmful effects on human moderators but is less reliable than when human moderators do view the content directly. AI can also reduce the challenges of moderating content in different languages by providing high-quality translations. Therefore, the productivity of human moderators can be increased and the harmful effects of viewing content can be reduced.

There is also promising research into the impact of techniques used to encourage socially positive online engagement. Should such techniques prove to be widely applicable, AI approaches can be used to discourage users from posting harmful content. This could reducethe reliance on online content moderation systems, given the current limitations of the technology, by reducing the amount of harmful content posted for some categories.

How stakeholders can think about AI right now

Here are some of the key questions -- and answers -- about AI and content moderation right now.

Q Do I need more robust content moderation, including technology, right now?

This varies by industry, but not much. Almost every business and brand needs to moderate some form of content, especially as brands become more community-driven. Third-party providers should be encouraged here because they provide both scale and cost containment possibilities.

Q What about my organization's data?

This is an interesting and evolving question. Most companies will be able to protect their data, but there is a growing push to share some data to help AI scale up faster at identifying harmful content; that's part of some UK legal efforts, for example. You may have to part with some data if you work with AI solutions or moderate at all, but it will likely be low-level images and nothing proprietary.

Q Should I be concerned about AI bias?

Yes, somewhat. There are issues with AI bias, and it's come up in the media repeatedly in the past three years. First: bias does exist, it cannot be entirely eliminated, and humans program AI (at least before machine learning kicks into high gear), and humans have biases. You won't eliminate bias, but you can ask potential partners and vendors about bias reduction, ask if they audit their datasets, ask how it's programmed initially, and more. You can do the research, and a vendor can answer these questions about tech and bias.

Q What if I don't fully understand how AI is making decisions that could impact my brand and business?

It's OK, in the sense that a lot of AI work is "black box," and people don't fully understand every machination of the programs, especially at the executive level. You should robustly vet a vendor, though, and ask them about programming decisions, meet the team, see previous work examples, talk about scale, and talk about human-tech interaction for agents. Do your research and ask questions. You won't become an AI expert, but you will get enough to protect your business.

Contact us at sales@conectys.com

Or learn more from our Blog.