cdt | Research

Five Limitations of Automated Multimedia Content Analysis

Automated content analysis tools present some promising use cases when implemented with proper safeguards, but their limitations must also be considered in any potential application. This is particularly important when their use may have widespread impacts on freedom of expression or user safety when deployed at scale. Policymakers and developers must understand these limitations when considering what role these tools may play in the analysis of user-generated content.

Read the full report at:
cdt.org/do-you-see-what-i-see

Robustness

Automated content analysis tools may perform well in a controlled environment but struggle to handle the wide variety of inputs that occur in a real-world scenario. This may be due to natural, expected variations in content in the wild, or due to deliberate efforts to evade the system.

Example: A machine-learning (ML) model was trained to identify birds, based on a training data set that includes photos of birds taken outdoors in bright sunlight. When presented with a photo of a bird taken through a screen, the model failed to identify the bird and mistakenly labels it a photo of a manhole cover. Adding statistical "noise" to a photo from the training data set also defeated the model, though the change to the image was imperceptible to humans.

Data Quality

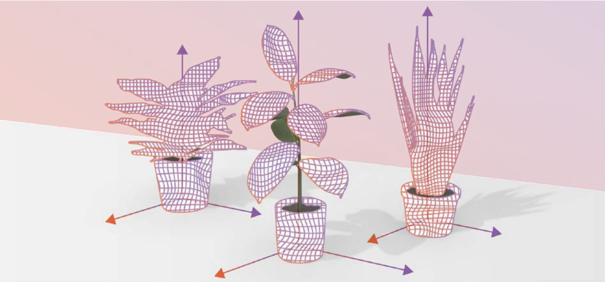
Machine learning involves the use of massive amounts of data to train a model. The composition of this data directly affects the outcomes of the ML model. If the data set reflects biases that exist in the real world, or that are incorporated into the data set during the data-collection or labeling stages, the model will "learn" those biases and recreate them in its operation.

Example: A company trained a hiring model to identify characteristics of successful applicants and employees. The resulting algorithm penalized résumés that included words like "Women" and "Girls" (in accomplishments such as "Chair of Women in STEM Initiative") because the algorithm had been trained on a data set of résumés from current employees, who were overwhelmingly men.

Lack of Context

Automated tools often perform poorly when tasked with decisions requiring judgment or appreciation of cultural, linguistic, social, historical, or other context. This can be because the underlying model is trained to complete a narrowly defined task, or because it lacks the ability to take in relevant surrounding information such as captions, comments, or account name.

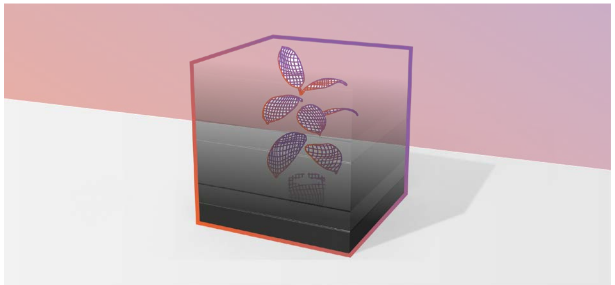
Example: A ML model may be trained to detect when an image contains nudity with a high degree of accuracy. This same model, however, will not be able to determine whether the nudity in an image is occurring in an artistic, political, health, educational, pornographic, or abusive context.



Measurability

There are many ways to measure the performance of automated systems; "accuracy" on its own is generally an unhelpful metric. It's important to understand how often an algorithm generates false positives and false negatives, and to have a sense of the prevalence of the kind of content it is trained to identify.

Example: Terrorist propaganda comprises a very small percentage of multimedia content overall. An algorithm that simply labels every piece of content "not terrorist propaganda" could technically be accurate 99.9% of the time. But such an algorithm would be useless for helping identify actual examples of terrorist propaganda.



Explainability

"Explainability" refers to the ability to map the operations of machine judgment onto concepts that can be understood by humans. Some algorithms are highly explainable and can be represented by intuitive concepts such as decision trees. Other forms of machine learning—sometimes called "black box algorithms"—resist easy explainability; the steps they take to reach an outcome do not translate into the kinds of judgments a human would make.

Example: A model was trained to predict a person's age from a photo. Upon investigating how the model worked, a researcher discovered that the model had learned to correlate age with not smiling. Explanations can help identify algorithmic judgments that perpetuate bias or are effectively mistakes in the real world.