Center for Technology and Society

The Online Hate Index

A machine learning system that detects hate targeting marginalized groups on online platforms

How much hate is there online? Is it possible to independently evaluate tech company claims about the amount of hate on their platforms and their efforts to address it? To answer these questions, ADL Center for Technology and Society (CTS) is building the Online Hate Index (OHI), a set of machine learning classifiers that detect hate targeting marginalized groups on online platforms.

Machine learning is a branch of artificial intelligence (AI) where computers learn to recognize patterns in data. In our use case, we are concerned with the patterns computers can find in language. Classifiers, models that predict how to categorize a given piece of data, can be fully automated or can assist humans in sifting through large volumes of data.

The OHI antisemitism classifier harnesses the extensive knowledge of ADL's antisemitism experts alongside trained volunteers from the Jewish community with lived experience of antisemitism.

Model to Scale

To the best of our knowledge, ADL's OHI antisemitism classifier development and application represents the first independent, cross-platform measurement of the prevalence of antisemitism on social media. It is certainly the first AI tool that has been painstakingly trained by experts in antisemitism and members of the targeted community. The ability to measure independently types of hateful content across an entire platform, and to compare results between and among platforms, is crucial to understanding how much hate exists online. It also enables us to understand what internal or external triggers may increase or decrease the amount of hate online. Such tools are essential for determining whether tech company anti-hate policies or product interventions actually work. Our method also provides a model for other civil society organizations rooted in targeted and marginalized communities who want to take an active role in training similar classifiers. For the first time, there is a way to engage in a quantitative, AI-assisted, community-based, at-scale effort to measure and analyze identity-based online hate.

How does the OHI work?

The OHI classifier is trained by volunteers from the Jewish community who are guided by ADL antisemitism experts. Once trained, the algorithm

learns to recognize antisemitic content and starts to generalize language patterns when given numerous examples of both offensive and innocuous content. Over time, it gets better at predicting the likelihood that a piece of content it has never seen before—a tweet, comment, or post—is antisemitic.

The OHI classifier learns to identify connections between English-language text (e.g., social media content) and human-assigned labels. First, volunteer annotators assign labels to that text (e.g., antisemitic or not). The system converts text and labels into numerical form (called an embedding or input feature). The model then adjusts billions of numerical parameters, commonly called "weights" in machine learning parlance, to produce an output that matches the human labels as closely as possible. But the model is complex, and not easily explainable in human terms, so data scientists evaluate the models on how well they can categorize text that is novel to the model.

This process repeats so that after each round of inferences and corrections, including additional human review, the model improves. Once trained, the model can receive inputs of English-language text and predict whether the text is antisemitic at speeds far faster than humans. Where it takes seconds to minutes for a human to evaluate one piece of text, the model can process thousands per second. This speed is highly valuable for processing data at platforms' vast scales. It is important to note that context matters, and that no model, no matter how well trained, is error-free. Our goal with the OHI is to measure the overall prevalence of hate and incorporate as much context as possible, including counterspeech, historical speech, and identity speech - and to understand better where humans still need to be involved in the process.

In our view, a key component to building just and effective tools for detecting hate requires the perspective of those targeted by hate. The only perspective a computer program can represent are those of the people who create and direct it. When marginalized people are not involved in building hate detection tools – as is typically the case – those tools are unlikely to benefit them as much as they could.

For more information, please contact cts@adl.org