Center for Technology and Society

FEB
2022

How Platforms Rate on Hate

Measuring Antisemitism and Adequacy of Enforcement Across Reddit and Twitter

ADL®
FIGHTING HATE FOR GOOD

Our Mission:
To stop the defamation of the Jewish people and to secure justice and fair treatment to all.

ABOUT
ADL Center for Technology & Society

Launched in 2017, ADL Center for Technology and Society (CTS) leads the global fight against online hate and harassment. In a world riddled with antisemitism, bigotry, extremism, and disinformation, CTS acts as a fierce advocate for making digital spaces safe, respectful, and equitable for all people.

CTS serves a unique role in civil society. It recommends policy and product interventions to elected officials and technology companies to mitigate online hate and harassment; drives advocacy efforts to hold platforms accountable and educates their staff on current threats and challenges; produces data-driven applied research by analysts and a network of fellows; sheds new light on the nature and impact of hate and harassment on vulnerable and marginalized communities; brings to market technical tools and products that provide the data measurement and analysis needed to track identity-based online hate and harassment; empowers targets of harassment by responding to online incidents; and works with platforms to create safer online spaces for all.

ADL is a leading anti-hate organization that was founded in 1913 in response to an escalating climate of antisemitism and bigotry. Today, ADL is still the first call when acts of antisemitism occur and continues to fight all forms of hate. A global leader in exposing extremism, delivering anti-bias education and fighting hate online, ADL's ultimate goal is a world in which no group or individual suffers from bias, discrimination or hate.

Learn more: adl.org

Center for Technology and Society

FEB
2022

How Platforms Rate on Hate

Measuring Antisemitism and Adequacy of Enforcement Across Reddit and Twitter

ADL®
FIGHTING HATE FOR GOOD

Executive Summary

How much hate exists online, and how large is its reach? Is it possible to independently evaluate the claims tech companies make about the amount of hate on their platforms and how effectively they are addressing it? To answer these questions, ADL Center for Technology and Society (CTS) is building the Online Hate Index (OHI), a machine learning system that detects hate targeting marginalized groups on online platforms. This report presents the inaugural findings of the OHI's antisemitism classifier, a new artificial intelligence tool that harnesses the rich knowledge of ADL's antisemitism experts to that of trained volunteers from the Jewish community who have experienced antisemitism.

We used this antisemitism classifier and our human reviewers to filter and analyze representative samples of English-language posts over the week of August 18–25, 2021, across both Twitter and Reddit. We should make clear that this analysis is not an indictment of Reddit and Twitter. In fact, we could only do this analysis because of these companies' commitments to transparency and data-sharing with third parties, a lead we call on other platforms to follow. For example, this analysis would not be possible on Facebook, the world's largest social media platform. While Reddit and Twitter have far more to do, they have both made substantial recent strides in addressing antisemitism and hate online. In this light, we offer our recommendations to help them better address these broader societal problems of online—and offline—hate and antisemitism.

Extrapolating from the late-August 2021 Twitter and Reddit samples, we estimate:

1. **The potential reach[1] of antisemitic tweets in that one week alone was 130 million people on Twitter. An equivalent estimate of the reach of antisemitic content on Reddit is not available.**
2. **That extraordinary reach was made possible by the 27,400 antisemitic tweets our machine learning tool enabled us to calculate were posted on Twitter that week; we found 1,980 antisemitic comments on Reddit.**
3. **The rate of antisemitic content on Twitter was 25% higher than it was on Reddit during that week.**

1 Explained in the methodology section below.

A month later, we evaluated company enforcement against the antisemitic content we had found. Then, more than two months after the initial investigation, we repeated the analysis on the same sample. **We found that the great majority of the antisemitic content had remained on the platforms for months, in clear violation of company guidelines on hate content, revealing the continuing inadequacy of the companies' content moderation.**

Moreover, even after ADL eventually reported the content that had remained posted online for more than two months after the initial discovery, the companies failed to remove more than half of that original antisemitic content.

Specifically:

4. **We returned to the antisemitic content about a month after the initial discovery to see if the platforms had removed it, but little had changed. On Twitter, 79% of the original antisemitic tweets remained, and on Reddit, 74% of the antisemitic comments remained.**
5. **We returned again more than two months after the initial discovery and found that at least 70% of the anti-Jewish content was still on the platforms.**
6. **Finally, on November 10, 2021, more than two months after the initial discovery, we contacted the platforms directly to report the antisemitic content from our samples that remained online.**
7. **One week after that notification, we returned to the representative samples and found that 56% of the antisemitic Reddit comments and 57% of the antisemitic tweets were still online.**

Among the hundreds of millions of tweets and the tens of millions of Reddit comments posted during the week in question, 27,000 antisemitic tweets and 2,000 antisemitic Reddit comments may not sound like much relative to overall volume, but it is in line with the relatively small yet disproportionately harmful levels of all types of toxic and abusive content experts have found across online platforms. One such study indicates that approximately 0.001–1% of content on mainstream online platforms may contain some form of abuse, a category that includes not only content that targets people based on their identities, but also more generally abusive content. More niche platforms may have levels of abuse closer to 5–8%.

Research consistently shows the impact of hate and harassment online is significant, and even a single targeted comment can affect a person's life to an extraordinary degree. In our most recent annual survey of Online Hate and Harassment, published in March 2021, 41% of American adults reported experiencing online hate and harassment, and 27% reported being subjected to severe online harassment (defined as sexual harassment, stalking, physical threats, swatting, doxxing, and sustained harassment). Thirty-three percent of all respondents in our survey reported experiencing online hate that was identity-based, targeting individuals or groups on the basis of, for example,

race, religion, ethnicity, gender, or sexual orientation. **The scale of the harm caused by antisemitism online comes into even sharper focus when the extraordinary reach of the content we find is considered in conjunction with ADL's recent research showing that 31% of Jewish Americans report being targeted on platforms because they are Jewish.**

This report almost certainly undercounts the overall prevalence of antisemitism on the platforms we researched, as ADL's classifier only detects English-language antisemitism, in the form of text, excluding videos, audio, and images. This tool is also better at detecting explicit language than subtle language, though our ongoing training of the classifier will continue to improve that capability.

Indeed, it's particularly dismaying that so much of the content we discovered was blatantly antisemitic. It was not even debatable.

To the best of our knowledge, the ADL Online Hate Index and this investigation represent the first cross-platform measurement of the prevalence of antisemitic content on social media undertaken by an independent civil society organization utilizing an AI tool that is meticulously trained by experts in antisemitism and volunteers from the targeted community. This enables the first ever independent, AI-assisted, community-based measurement of identity-based hate across an entire platform.

The ability to independently measure hate content at scale, and compare results between and among different platforms, is crucial to understanding how much hate exists online. It makes possible a better understanding of what internal or external triggers may increase or decrease the amount of hate online. It is also essential for independently determining if companies' anti-hate policies, practices, and product changes work and if their claims on that score can be verified. This work also provides a model for other civil society organizations rooted in targeted and marginalized groups who wish to take active roles in training similar classifiers to identify the specific types of online hate that target their communities. ADL hopes to partner with organizations such as these as it continues to work on the OHI.

It is worth emphasizing again that Twitter and Reddit, to their credit (and in marked contrast to the world's largest social media platform, Facebook), make their data far more accessible to independent researchers for this kind of analysis. While there are areas for improvement, we commend both platforms for this transparency. We hope it serves as an example of how to advance the fight against online hate.

Measuring the Prevalence of Hate Content

As the avalanche of stories linked to the Facebook Papers have documented, social media platforms have consistently failed to fight hate speech and misinformation to the point of malfeasance. Despite knowing there is an ocean of hateful content on social media, platforms do not take action against most of it. Research by ADL confirms platforms are not doing enough to curb toxic, abusive posts. In just hours, millions of people can be exposed to hate speech and misinformation that goes viral. Such content serves to marginalize targeted groups, violating their civil rights, isolating and opening them up to far greater risk of discrimination and violence. Hateful content can normalize violent extremism, enable extremist radicalization and recruitment, promote dangerous conspiracy theories, and, in some cases, lead to violence and death. Such cases include the "Boogaloo" murders in 2020 and the insurrection at the U.S. Capitol on January 6. When it is left on a platform even for relatively short periods of time, the damage wrought by this content may be irreversible.

We cannot trust most tech companies to be forthcoming about the prevalence, specific targets, reach, and impact of hateful content on their platforms.

Most tech companies resist sharing their internal data about hate, misinformation, and extremism with third parties. Their transparency reports obscure more than they reveal. They have shut down the work of researchers investigating their platforms. For these reasons, third parties must be provided access to conduct independent audits of content on platforms.

Quantifying hateful content on digital platforms is a necessary step in assessing whether platform policy or product changes, or other interventions, actually reduce hate online.

If even a comparatively tiny civil society organization such as ADL, with its relatively limited resources, can identify specific types of hate speech, and measure their prevalence at scale, across social media platforms, then technology companies clearly must do far more and far better. Any excuses about burden and other challenges from Big Tech ring even more hollow after ADL's launch of the OHI classifier.

We hope to use and continue to develop the OHI antisemitism classifier as a tool for measuring hatred against Jews online, evaluating the strength of tech companies' mitigation efforts, and holding those companies accountable for the antisemitic content propagating on their platforms. We also hope to work with groups that represent other targeted communities to help them leverage their unique expertise in the types and impacts of hate they encounter online (and off). The goal would be to pair those groups' expertise with ADL's ability to train machine learning classifiers to identify particular forms of identity-based hate at scale.

Analyzing Reddit and Twitter but not Facebook

We measured the prevalence of antisemitism across Reddit and Twitter because their data is the most accessible to researchers. Facebook (now rebranding as Meta), is the world's largest social media platform, and our research previously found that three quarters of Americans who experience hate online report that at least some of that hate occurs on Facebook. It has the highest share of online hate and harassment reports among all major platforms. But Facebook does not make its data available to most third parties, so we could not include the platform in our investigation. This is all the more concerning in light of recent media reports showing that—contrary to the information the company provided to its much-touted civil rights auditors last year—Facebook, in fact, does have important data concerning particular types of identity-based hate content on its platforms. The company appears not to have disclosed this even to its internal civil rights team. Facebook must provide effective and comprehensive access to independent experts. It has already shown that it can do so with appropriate privacy and proprietary information safeguards.

In contrast to Facebook, we commend Reddit and Twitter for making their data accessible to independent researchers, although we remain concerned about the offensive content we found. As outlined in our 2021 Online Antisemitism Report Card, even comparatively researcher-friendly platforms like these two do not make all of their relevant data available, especially Twitter. It is true that Twitter is the only major social media platform that allows most researchers access to its public data, as stated in the Congressional testimony of ADL Belfer Fellow Laura Edelson on

September 28, 2021. But, while its Firehose API provides the full volume of data—and not just the 1% sample of tweets offered through its free API, which is what we used for our research—Firehose is prohibitively expensive for many groups, including ADL, with fees rumored to be upwards of seven figures per year. Another problem with the far more expansive Firehose API is that it still provides an incomplete view of Twitter. Computer scientist Alan Mislove has reported that the Firehose API does not cover ad-targeting information, such as Twitter's decisions about which ads users see through its delivery algorithm.

Beyond the expense and incomplete data access, Twitter also enforces many publication restrictions in its API agreements. For example, it permits researchers to share the IDs of tweets they analyzed, but it does not allow them to release the full data (text, author, likes, retweets, bios, etc.). In contrast to Twitter, Reddit provides expansive data collection at the subreddit level with few restrictions to access or publication.

Currently, no tech company publicly reports on the full scope of specific forms of hate on its platform, whether through its transparency reporting or any other public disclosures. Independent researchers thus have only limited access to the data necessary to evaluate the full scope of hateful content on any platform. For third parties such as ADL to measure antisemitism and other forms of hate on a social media platform, tech companies must make their data available and shareable.

Results

Our investigation looked at the prevalence of antisemitism and the platforms' enforcement against that content, two key metrics for evaluating the presence of antisemitic content and the effectiveness of Twitter and Reddit's content moderation efforts. We also calculated the potential reach of antisemitic tweets, as is explained more fully below.

- Prevalence: How much antisemitic content is there on the platform?
- Enforcement: How effective is the platform in removing antisemitism?

It's important to note that this investigation focused only on English-language antisemitic content, not on other forms of hate, online abuse, and harassment,. We also focused only on text, not on other modalities such as video, audio, and images. So, we are almost certainly undercounting the amount of antisemitism on these platforms. In general, research focusing on online abuse broadly—including not only identity-based hate but also a broad range of disruptive behaviors—has found that the overall amount of abuse on a platform compared to other content is relatively small: for example, a study from the Alan Turing Institute indicates that approximately 0.001–1% of content on mainstream online platforms may contain some form of abuse, while more niche platforms may have levels of abuse closer to 5–8%. That the overall percentage of hate is small when measured against the vast amount of content online today is not surprising. It says little to nothing about the type, reach, targets, and impact of hate content, as recent reports and events have amply illustrated.

Prevalence Measures

Extrapolating our results to the entirety of Twitter, we found there were approximately 27,400 antisemitic tweets out of an estimated 440 million English-language tweets during the week of August 18–25, 2021 — or 0.0062% across all of English-language Twitter. This amounts to roughly 62 antisemitic tweets per every million English language tweets for that week.

To get to this result for the entire platform, we analyzed a random sample of 1% of English-language Twitter and found 274 English-language antisemitic tweets in our sample of 4.4 million English-language tweets during the week of August 18–25. We then multiplied what our researchers found in that sample by 100 to estimate the total amount of antisemitism across the platform during that period.

On Reddit, we found between 1,924 and 2,043 antisemitic comments out of 39.9 million English-language Reddit comments during that same week in August—or 48–51 antisemitic comments per million Reddit comments.

This range of antisemitic comments across the entire platform is extrapolated from the statistical sample we took and had manually reviewed by volunteer labelers. Our sampling methodology allows us to be 95% confident that the posts are indeed antisemitic, with a 3% margin of error. Expanded to the entire platform, the number of antisemitic comments amounts to between 0.00483% and 0.00512% of overall comments for the week in question. Our methodology also allows us to compare levels of antisemitic content across Twitter and Reddit. We found that the rate of antisemitism on Twitter was 25% higher than it was on Reddit.

Although these numbers may appear small in terms of total quantity and percentage, research shows the reach and impact of hate and harassment is significant. For example, CTS's most recent annual survey of Online Hate and Harassment, published earlier this year, found that fully 41% of Americans experience some type of hate or harassment online, with 27% experiencing severe harassment (defined as sexual harassment, stalking, physical threats, swatting, doxxing, and sustained harassment). Overall, 33% of respondents in ADL's most recent nationally representative survey reported experiencing identity-based harassment online, with 31% of Jewish Americans reporting they were targeted because they were Jewish. Moreover, the reach of the antisemitic posts we found in just one week, as measured by engagement and followers, exponentially amplifies this content, as described more fully below.

For targeted groups, especially marginalized communities, hate causes psychological damage, emotional distress, reputational harm, and withdrawal from online spaces. Other forms of hate, including racism, misogyny, homophobia, transphobia, Islamophobia, and xenophobia, often accompany antisemitic content. If platforms do not remove the most overt antisemitic posts and comments, they are likely not removing these other forms of hate, particularly for targets who belong to more than one identity group.

Because we did not study the context of antisemitic posts, we cannot speak to whether they were associated with conspiracy theory content such as QAnon or COVID-19 disinformation, but users who engaged with antisemitic posts may be more likely to be served similar content, potentially leading to rabbit holes and greater risk of radicalization.

Enforcement Actions

Content Removal

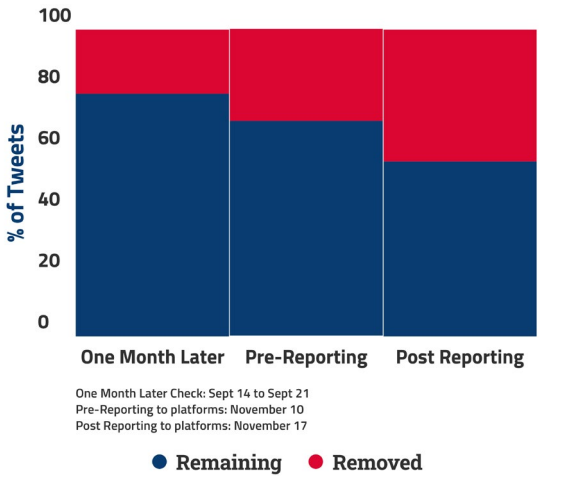
ADL researchers checked 27 days after the initial collection to see how many of the antisemitic posts found by our classifier and labeled antisemitic by our human reviewers remained on Twitter and Reddit in late September 2021. We found that roughly three out of four antisemitic posts were still on both platforms—216 out of 274 (79%) tweets on Twitter, and 147 out of 199 (74%) comments on Reddit. On Twitter, the removals that had occurred were overwhelmingly due to user action (76%), meaning that the original poster removed the antisemitic tweet, possibly after Twitter required them to take action. The reverse is true for Reddit; on that platform the majority of removals (73%) were due to platform or community moderator intervention. These percentages are derived from the API response messages.

We checked again on November 10, prior to flagging the content for Twitter and Reddit directly. For Reddit, 145 comments out of 199 were still on the platform (73%); for Twitter, 192 out of 274 were still on the platform (70%).

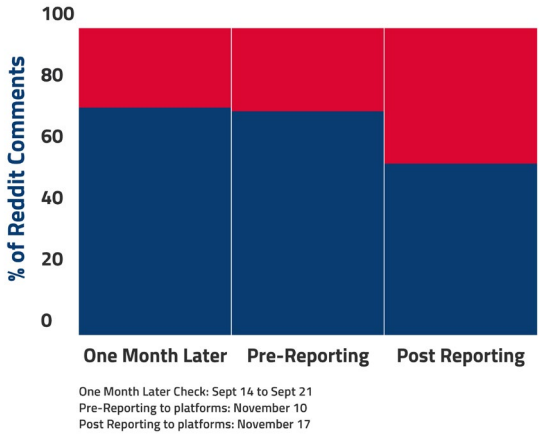
Even more dismaying, within one week of ADL flagging the content directly to the platforms, 112 out of 199 comments were still on Reddit (56%) and 157 out of 274 tweets were still on Twitter (57%).

This shows that despite giving the platforms the benefit of time to act on their policies and then later providing them with the content we found during our investigation, both Twitter and Reddit decided to take action on less than half of the antisemitic content we uncovered. This raises serious concerns around the ways in which these platforms define antisemitism and their ability to enforce their stated policies—even when content is flagged for them.

Antisemitic Tweets Remaining On Twitter Over Time



Antisemitic Comments Remaining on Reddit Over Time



Engagement Statistics: Reach of Antisemitic Content

Twitter makes available metrics that can be used to estimate the reach of a certain piece of content—that is, the number of people who may have seen it. The 216 antisemitic tweets and retweets remaining on the platform after 27 days were from accounts with a combined follower count of around 1.3 million. Twenty-two out of 185 (12%) of the original accounts and 5 out of 84 (6%) of the accounts that retweeted offensive content each have more than 10,000 followers. Combined, the retweets and their original posts received nearly 10,000 likes and 3,400 retweets. Assuming that levels of engagement (measured by likes and retweets) in the 1% stream are consistent throughout the entire Twitter platform, this would amount to roughly 1 million likes and 340,000 retweets across the full Twitter platform for the week in question—August 18–25. We calculated these numbers based on the tweets that remained on the platform after 27 days. But even those that were removed in this timeframe were probably liked or retweeted prior to their removal, so the total reach of the content we found is likely higher. Adding up the follower counts and adjusting for the sampling rate means 130 million people potentially viewed and/or were

User Name	User description	Largest Follower Counts
ZaidZamanHamid	"A veteran of Soviet-Afghan war, presently Strategic Security Analyst & founding Consultant BrassTacks - The Advanced Threat Analysis Think Tank."	340,770
conspiracyb0t	"Conspiracies, NWO, Illuminati, Secret Societies, Police State, Federal Reserve, Liberty, Freedom. As an Amazon Associate I earn from qualifying purchases."	176,708
merrittk	Podcasts and editorial @fanbytemedia . merritt souls: Bloodborne is M/F 5:30 ET http://twitch.tv/fanbyte	47,419
StudentOfDeen_1	-	39,134
panafrikam	27 • @gsopfa • new afrikan state-affiliated media • @blacks4peace	32,876

influenced or impacted by these tweets. Equivalent proxies for comment-level engagement are unavailable on Reddit, making a comparison of the reach of antisemitic content on Twitter and Reddit impossible at this time.

Beyond content removal, Reddit has other means of policy enforcement, focused on de-amplifying the number of users who will view specific pieces of content. For example, Reddit allows users to upvote or downvote comments. A comment is then given a score based on the difference between the two. Comments that get many upvotes compared to downvotes appear more prominently. Statistical analysis of those scores shows that antisemitic content on Reddit is rewarded significantly less than non-antisemitic content. Our investigation found that the average score of antisemitic posts was one-third that of non-antisemitic posts. This seems to indicate that Reddit users are less rewarded for antisemitic content than for other types of content, as a result of the upvoting and downvoting mechanism on Reddit.

Again, it is important to note that our results regarding prevalence, amplification, and reach illustrate two things. First, there are such vast numbers of people on these platforms that even a seemingly small amount of antisemitic content, such as 27,000 tweets in a single week, can potentially reach over 100 million people. And that is just for one week's antisemitic content. Second, the true impact of this content is still beyond our ability to assess due to the limitations on data-sharing in place at tech companies.

Methodology

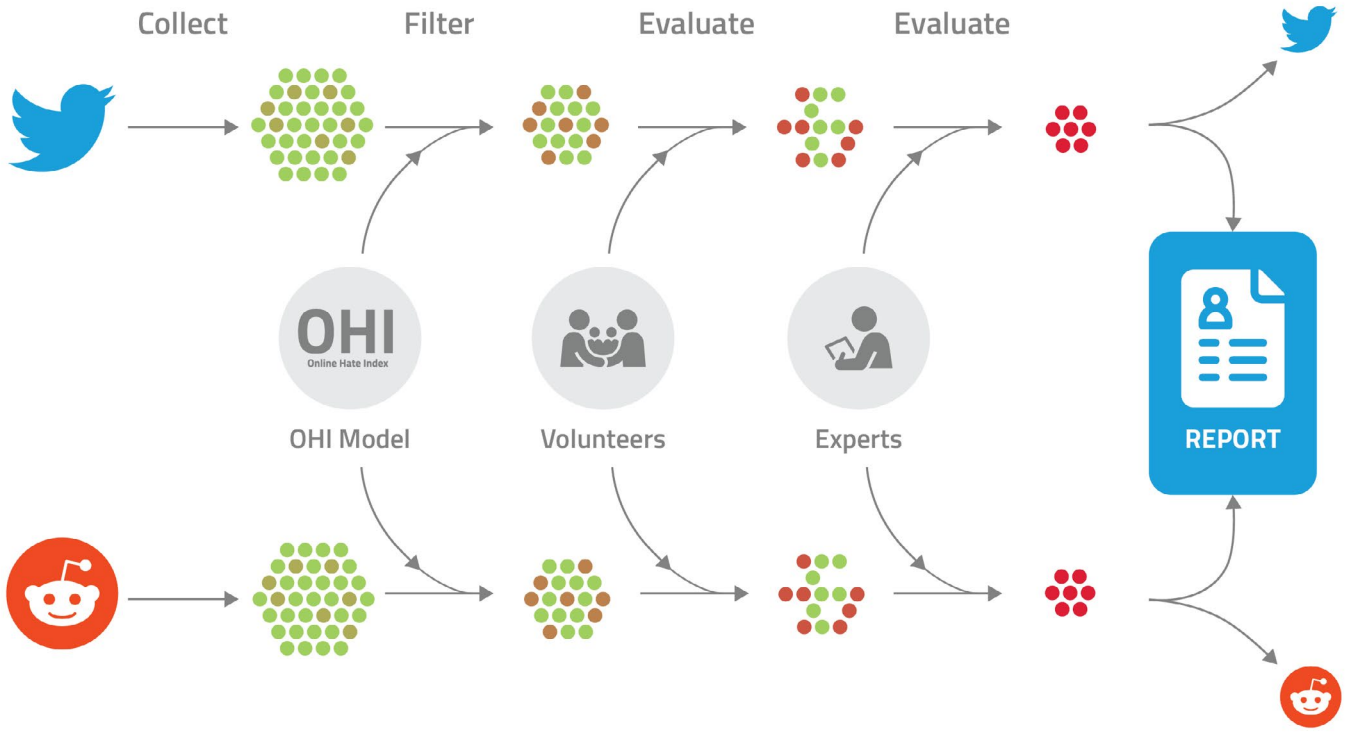
How the OHI Works

ADL Center for Technology and Society uses machine learning to build classifiers to understand the proliferation and mechanics of hate speech. Machine learning is a branch of artificial intelligence where computers, given human-labeled data, learn to recognize patterns. In our use case, we are concerned with the patterns computers can find in language. Classifiers, models that predict the category in which a piece of data belongs, can be fully automated or assist humans to sift through large volumes of data. [Consult the Glossary at the end of this report for more definitions of technical terms.]

Imagine a novice art history student looking at images of hundreds of paintings. Through practice, the student improves their ability to identify art movements (Impressionism, Surrealism, etc.) even when they encounter unfamiliar works. Machine learning works similarly. In the case of our antisemitism classifier, the algorithm learns to recognize antisemitism and starts to generalize language patterns by being 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, people who volunteer to serve as labelers of content 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. It can learn, for example, that the words "kill" and "jews" frequently co-occur with the label antisemitic, unless the words "don't" or "wrong" also appear. But the model is complex, and not easily explainable in human terms, so data scientists evaluate the models on how well they can evaluate text that is novel to the model.

This process repeats so that after each round of making inferences, 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 can. Whereas it takes seconds to minutes for a human to evaluate one piece of text, the model can process thousands per second. This is useful for processing data at platforms' vast scales.



But artificial intelligence cannot entirely replace human discernment, as much as companies like Facebook would like us to believe otherwise. Rather, the unique perspectives of individuals belonging to communities targeted by hate is, we believe, a key component of properly building AI tools to detect hate speech online. The only perspectives a computer program can reflect are those of the people involved in creating it and telling it what to do. If individuals with marginalized identities are absent from building hate-speech-detection tools, the likelihood that those tools will benefit them as much as they could diminishes. Thus, we believe it is important that a tool meant to detect antisemitism must meaningfully involve Jewish people in its creation—both deep experts in antisemitism and members of the community sharing their lived experience of antisemitism.

ADL's antisemitism classifier is trained solely by experts on antisemitism and volunteer labelers who identify as Jewish. ADL acknowledges that viewing antisemitic content can be triggering and have harmful effects on the volunteer labelers. To attempt to mitigate any adverse effects, volunteers are encouraged to take frequent breaks and reach out to the labeling team if they feel the content is becoming overwhelming. The labeling team emphasizes during onboarding that participation is completely voluntary, and volunteers can leave the project at any time. The volunteer managers reach out to check in with active labelers about how they are feeling. While the volunteers are allowed to continue participation, the original agreement is for an initial three-month period. Each volunteer completed live training based on an expert guide defining antisemitism, received in-person support from ADL experts,

and completed at least two practice rounds to assess whether they understood the task. All labeling decisions are made based on the majority opinion of at least three labelers. ADL provides a primer on antisemitism to volunteer labelers, but encourages them to use their judgment and lived experience in their evaluations. ADL reviews the final labeling results.

Platforms, on the other hand, typically do not explicitly train classifiers with datasets that specific, affected identity groups generate. For example, internal documents from Facebook, leaked by whistleblower Frances Haugen and submitted to the SEC, appear to show how Facebook trains its classifiers only in terms of broad "hate" or "not hate" categories, without focusing on the specific experiences of targeted communities or the particular ways in which hate against specific communities manifests.

> *"[O]ur current approach of grabbing a hundred thousand pieces of content, paying people to label them as Hate or Not Hate, training a classifier, and using it to automatically delete content at 95% precision is just never going to make much of a dent...***we're deleting less than 5% of all of the hate speech posted to Facebook. This is actually an optimistic estimate***--previous (and more rigorous) iterations of this estimation exercise have put it closer to 3%, and on V&I we're deleting somewhere around 0.6% ...we miss 95% of violating hate speech."*[16] *(emphasis added)*

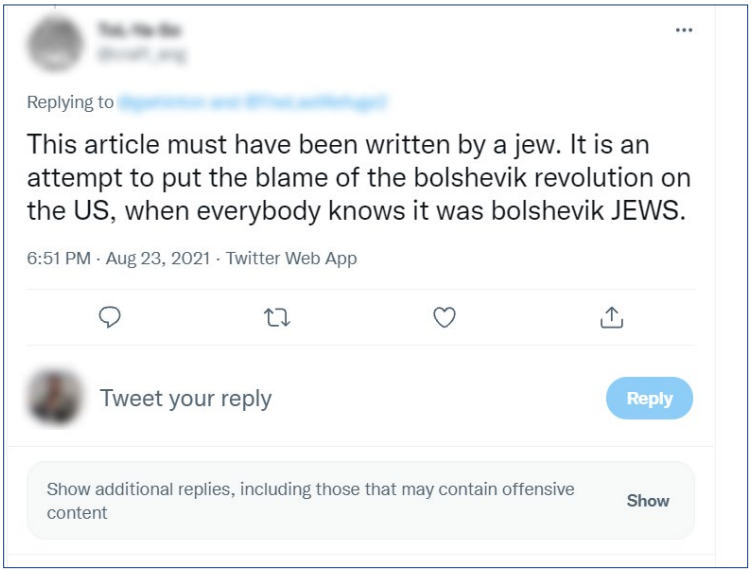
This backs up our hypothesis that most classifiers used by technology companies are trained by labelers who may have no lived experience as targets of the type of hate and harassment they are charged with discerning and labeling. The categories of hate also are not likely to be defined by civil society experts. ADL believes detecting hate speech must ultimately be a human endeavor, aided by technology, and constantly iterative, to evolve with hate and harassment, in response to new contexts while leveraging expertise and lived experience of particular forms of online hate.

How We Conducted Our Analysis

From August 18–25, 2021, researchers at ADL Center for Technology and Society collected a random sample of 1% of all English-language tweets, provided by Twitter, and an estimated 99.9% of all Reddit comments collected from all subreddits within all threads in all languages in real time. We decreased that input to 4.4 million tweets and 39.9 million Reddit comments after filtering for English using FastText's language identification model. These English-language tweets and comments were then passed through our OHI antisemitism classifier.

After our classifier further reduced the samples from Twitter and Reddit to a smaller amount of potentially antisemitic posts or tweets, three trained Jewish ADL volunteer labelers (described above) reviewed the remaining Twitter content and a sampling of the remaining Reddit content that was scored by the classifier as antisemitic. Content that the volunteer labelers indicated was

antisemitic was included in the next step in the process. Finally, at least three ADL experts reviewed the content to ensure it was correctly labeled as antisemitic. To be determined antisemitic, at least two of the three experts and two of the three volunteer labelers had to agree that a given piece of content was antisemitic. This means that any piece of content we counted as antisemitic, for purposes of this report, had been filtered out as potentially antisemitic by our classifier, then labeled as such by a majority of our volunteer Jewish labelers, and, finally, by a majority of our internal experts on antisemitism.



ADL researchers then checked to see how many of these antisemitic posts remained on both Twitter and Reddit 27 days after the initial collection, giving both platforms ample time to act on their own, and allowing us to observe their enforcement rates without ADL first flagging the content. ADL researchers also reviewed non-removal moderation actions, such as upvotes and downvotes, and the degree of engagement antisemitic content left on the platform received. Our researchers checked again on November 10 (between 11 and 12 weeks after the initial late-August collection) before flagging the content for Reddit and Twitter. Once we flagged the content for the platforms, we waited another week and then recalculated which posts were still permitted on the platforms and tabulated the final removal rates.

The methodology and metrics of the OHI machine learning classifier employed for this report were reviewed and validated by three separate outside experts, all of whom are experts on machine learning, and two of whom are also experts in hate and harassment online.

Limitations and Caveats

We measured the prevalence of antisemitism on Reddit and Twitter from August 18–25. While there were no significant events to make us suspect that week's measurements would be drastically different from other times, we cannot conclude that the results would be the same during a different week. Comparisons of that sort await additional measurement and analysis.

The OHI antisemitism classifier is still under development—and, indeed, hopefully will be continuously improved and iterated on. Like all ML classifiers, it misses some of the antisemitic

content. So, for this report, we chose to use conservative estimates regarding the prevalence of antisemitic content, based only on content that was cross-checked by volunteer labelers and experts. The current classifier focuses on explicit antisemitic text; it cannot detect antisemitism in images, videos, audio, or more subtle content that may be hard to identify as antisemitic without further context.

Machine learning classifiers must balance false positives and false negatives,—that is, the degree to which the classifier over- or under-identifies the type of content it detects. ADL's internal analysis of the OHI antisemitism classifier suggests that it successfully detects around half of all antisemitic content in samples, so we estimate there is roughly twice as much antisemitic content on these platforms as the classifier can currently detect.

The results do not encompass other forms of hate speech targeting other marginalized groups. Additionally, the OHI classifier has only been trained on antisemitic text content in English, and cannot currently detect any antisemitic content that may be present on platforms in the form of images, video, or audio.

That our results showed such egregiously poor performance by platforms, even given the limitations of our tools, process, and conservative methodology, is, frankly, all the more troubling.

Conclusion: Next Steps

Outside of ADL's investigation, there is little independent, verifiable research on how much hate speech exists across a social media platform, who it influences to spread still more hate, and who is targeted. Tech companies' transparency reporting is insufficient and opaque. They measure hate using misleading metrics that lack or intentionally omit the necessary context and metrics to make sense of their importance. For example, Twitter highlights how many accounts it took action against, but does not state how many times users can violate its rules before it disables their accounts. Facebook collects data on hate against different identities through its regular reporting form; a user can report that they have experienced or witnessed hate against a racial or religious group, for example, but Facebook has yet to make any of that information public. In fact, no tech companies regularly report on the overall prevalence of hate on their platforms, or disaggregate hate content by type, target, impact, and reach. As a result, it is impossible to measure, for example, how well their enforcement works relative to the total amount of online hate or the experience of marginalized communities. Reddit, on the other hand, provided a model for tech companies by producing the first report on how different communities are targeted by hate on their platform, but even Reddit did so separately from its regular transparency efforts and has not repeated the report.

In light of this, platforms must provide meaningful data to researchers, civil society, and other good-faith actors to allow for more robust and verifiable measurements of hate online.

To date, most platforms have not done this. One of the ways the government can increase platform accountability is by requiring increased data access to trusted, independent third parties. This would allow for comprehensive audits of hate online and an independent assessment of platforms' efforts to moderate content. As has been made clear by a number of examples already, this accessibility can adequately accommodate user privacy and company proprietary information.

Algorithmic detection of hate speech is not simple. Our own efforts, while still in their early stages, show that it is possible to better identify overt and explicit antisemitic content, but detecting implicit hate, such as coded or context-dependent content, is likely much harder. Furthermore, ADL only analyzed English-language content, and we know that platforms devote most of their resources to moderating English-language content. The full extent of antisemitic content present on platforms remains unknown. Hate speech moderation is far worse in non-English languages. The impact, however, is very real. The lack of moderation over hateful content in languages other than English has helped fuel global conflicts such as the civil war in Ethiopia and the genocide of the Rohingya in Myanmar.

Numerous critics contend automated systems also risk reproducing the systemic biases inherent to many data sets and machine learning engineers themselves. For these reasons, we drew on both ADL's expertise in identifying antisemitic content, such as terms and tropes, and the lived experiences of Jewish users who labeled our training data.

Automated detection must always be conducted in combination with other research methods, including human review, surveys, and interviews, to understand not just its prevalence, but its impact.

It's clear that mainstream social media platforms are doing far less than they should with the resources available to them. Some, like Facebook, provide little to no access to independent researchers who could hold the company to account. ADL has argued for years that tech companies profit from extremist hate and harassment because it boosts engagement on their platforms, which fuels advertising revenue.

Looking forward, we plan to use the Online Hate Index to better understand the prevalence of antisemitism and all forms of hate online.

Recommendations

For platforms

1. **Center targets and affected communities in content moderation tools and practices**
    - The results from this report indicate that platforms can improve their content moderation by including in both their AI and human review the experiences of users targeted by specific forms of hate. We encourage tech companies to expand on our findings and support research into building content moderation tools and processes—such as machine-learning systems—by engaging people impacted by hate.
        - Our experience recruiting affected community members and having them trained by experts to label antisemitic text shows promise, but warrants additional research. This research should include the potential mental health impacts of labeling online hate speech, the degree to which labelers need training and other support in this work, and issues of privilege and justice which, among other concerns, have bearing on compensation for volunteer labelers.
2. **Submit to regular and comprehensive third-party audits**
    - Transparency currently provided by tech companies is severely lacking. Tech companies should submit to external audits of the prevalence of hate on their platforms and their efforts to mitigate it. External audits should be conducted by trusted experts in digital hate such as ADL, other members of civil society, and academic researchers. Audits would also allow the public, including policymakers and legislators, to verify whether social media companies follow through on their stated promises. For example, a third-party audit of Facebook could allow for an independent evaluation of whether the company has meaningful data about the ways in which marginalized communities are targeted by hate on the platform. This can be accomplished by processes that do not trespass on or violate individual privacy interests and rights.
3. **Enforce policies on antisemitism and hate consistently and at scale**
    - Platform policies are only as good as their enforcement. ADL recommends that tech companies allocate resources for automated content moderation and human review proportionate to the harm. Necessary investments should include more human moderators and greater training for human moderators in specific forms of

hate, including in relevant languages and cultures. These investments should also include more accurate ML classifiers and expanding the development of automated technologies to commit as few mistakes as possible when enforcing policies around violative content.

4. **Provide researchers with greater access to data**
    - Tech companies should use ADL's rubric for data access to provide researchers with substantive, privacy-protecting data. Better data access will aid researchers' efforts to understand the nature of antisemitism and hate online and study whether platforms' efforts to address hate are effective. Ideally, platforms should make it easy for academic researchers and civil society organizations to acquire data for auditing at a large scale. Most do not. ADL recommends that platforms' data accessibility meets the following criteria, as outlined in our 2021 Online Antisemitism Report Card:
        - Availability of public APIs that return public posts and enable third parties to retrieve data with minimal setup
        - Availability of APIs that nonprofits and research organizations can use; trusted third parties can access more detailed data under appropriate privacy restrictions
        - Availability of APIs that return information on user reporting and content moderation so third parties can understand platforms' actions
        - Ability to search past data, allowing third parties to assess historical trends
        - Ability to stream new data so third parties can monitor ongoing developments
        - Ability to automatically discover new groups or topics
        - Ability to for third parties to collect data at scale as a result of high rate limits (amount of content platforms allow third parties to pull in within a given timeframe)
        - Quality of documentation explaining API use; third parties can use the above features with relative ease

    When tech companies provide this meaningful data to researchers, third parties can audit the prevalence of various phenomena on social media platforms.

For Government

ADL's REPAIR Plan is a comprehensive framework to decrease hate online and push extremism back to the fringes of the digital world. In line with REPAIR, we encourage governments to:

1. **Prioritize regulation and reform focused on systematized, comprehensive, and easily accessible transparency.** Platforms claim to have strong policies against hate, gender-based violence, and extremism, when, in fact, most are unclear, hard to find, or have perplexing exceptions. Enforcement is inequitable and inconsistent, and transparency reports are incomplete, irregular, and opaque. Policymakers must pass laws and undertake approaches that require regular reporting, increased transparency, and independent audits regarding content moderation, algorithms, and engagement features while looking for other incentive-based or regulatory action. Platform transparency reporting must evaluate success and provide evidence that independent researchers can use. Such independent researchers must be granted access to data, and Congress must have an oversight role.
2. **Support research and innovation:** Governments must focus on research and innovation to slow the spread of online hate, including, but not limited to: (1) measurement of online hate; (2) hate and extremism in online games; (3) methods of off-ramping vulnerable individuals who have been radicalized; (4) the connection between online hate speech and hate crimes; (5) new methods of disinformation; (6) the role of internet infrastructure providers and online funding sources in supporting and facilitating the spread of hate and extremism; (7) the role of monopolistic power in spreading online hate; (8) audio and video content moderation. Researching areas like these is crucial to developing innovative yet sustainable solutions to decrease online hate.

Glossary

Terms like “machine learning,”, “AI,”, and “algorithm” are often poorly defined in news articles about social media platforms and content moderation. What do they mean? This brief glossary explains key terms in this report for a lay audience.

Algorithm. Algorithms are sets of rules and processes for computers to follow. Social media platforms use algorithms to identify and remove hate speech because human content moderators cannot review every post due to platforms’ massive scale. But computers can follow only the rules and processes taught to them, so we must understand the distinctions we want them to make.

Application Programming Interface (API). A program that allows two other programs to talk to each other. This tweet provides a useful analogy to describe how an API works.

Artificial intelligence (AI). The branch of computer science concentrated on the study of computer systems able to perform tasks that usually require human intelligence, such as visual perception, speech recognition, decision making, and language translation.

Classifier. A model that predicts which category a piece of data belongs in.

Machine learning: A branch of computer science and artificial intelligence that uses data and algorithms to imitate how humans learn, gradually improving a computer’s accuracy. It is a class of algorithms that predict what will be successful in the future based on patterns in the past.

Model: The output of a machine learning algorithm. Models generally take data as input and generate predictions.

Social media platform: A website, app, or digital platform comprising networked user accounts, like Facebook or Twitter. Most social media platforms allow users to “follow” other users, either mutually or in one direction.
Social media differs from broadcast media, like television and radio, and print news, because it allows for many-to-many connections rather than one-to-many. These features allow users to form connections and create social capital (the benefit you get from who you know); they also allow some users to gain disproportionate influence and for some pieces of content to “go viral,” that is, rapidly spread across networks. In the past 10 years, the central content feeds on most platforms are managed by algorithms that determine what each user sees, rather than displaying all posts chronologically.

Support

This work is made possible in part by the generous support of:

The Robert Belfer Family

The Grove Foundation

Walter & Elise Haas Fund

Luminate

Craig Newmark Philanthropies

Righteous Persons Foundation

Bumble

Zegar Family Foundation

Dr. Georgette Bennett

Joyce and Irving Goldman Family Foundation

Alan B. Slifka Foundation

Amy and Robert Stavis

Quadrivium Foundation

John Pritzker Family Fund

Horace W. Goldsmith Foundation

ADL Leadership

Tech Advisory Board Members

ADL Center for Technology and Society

For additional and updated resources please see:
www.adl.org

Copies of this publication are available in the Rita and Leo Greenland Library and Research Center.

Take Action

Partner with ADL to fight hate in your community and beyond.

- Sign up at adl.org for our email newsletters to stay informed about events in our world and ADL's response.
- Report online hate to technology platforms or escalate to ADL if platforms are not responsive.
- Educate yourself and others about the reach and impact of online hate.
- Get involved with ADL in your region.

For a database of reports and resources on antisemitism, extremism and more, visit ADL.org.

Anti-Defamation League @ADL @adl_national

ADL®
FIGHTING HATE FOR GOOD