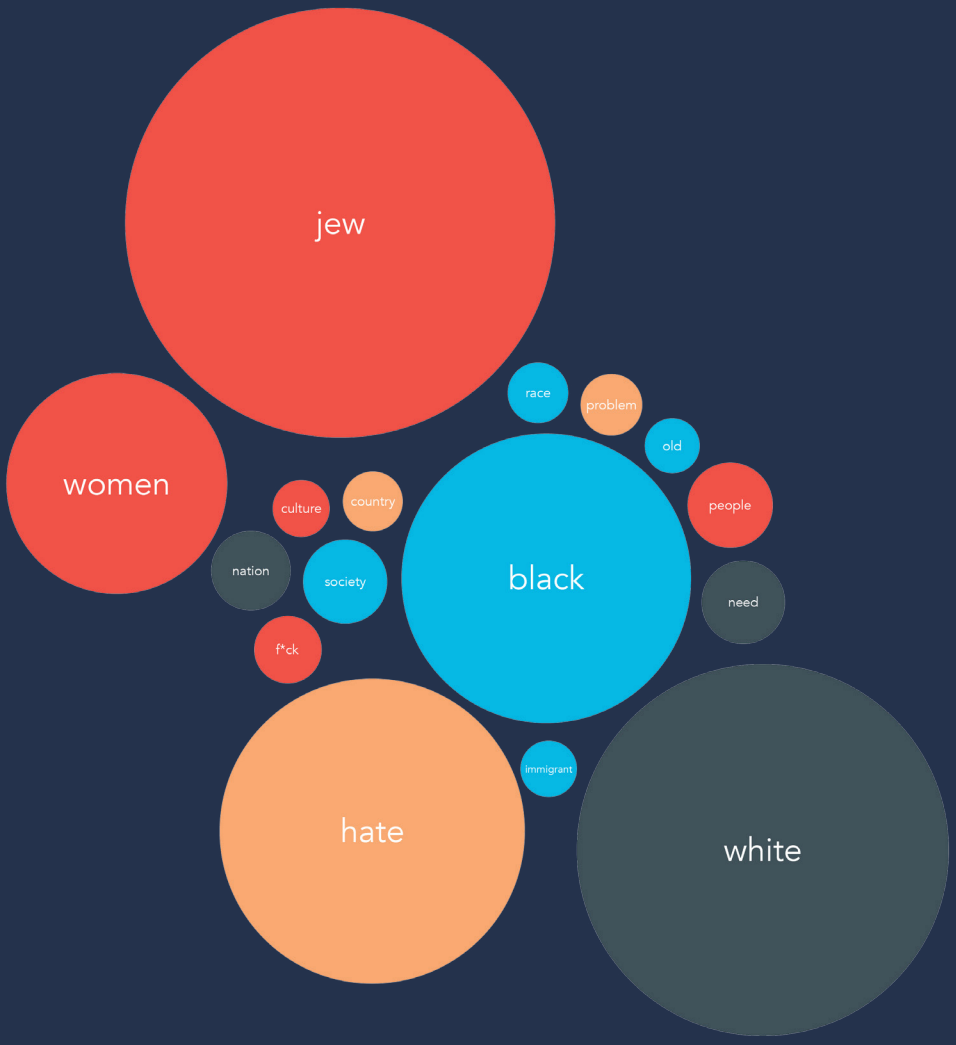
CTS Center for Technology & Society

Online Hate Index

Innovation Brief



ADL

Contents

Introduction

For many, encountering hate on the internet has become a routine part of the online experience. According to Pew Research Center, 41% of American adults have experienced online harassment, and 66% have witnessed it. For those on the receiving end of online vitriol and bigotry, there is no mistaking what is happening: these are words that wound, which are often defined by recipients as hate speech. But defining what constitutes online hate speech can raise many questions. With only words on a screen, and no context about the speaker or the speaker's actions, can we create generally applicable rules and definitions that will include hate speech, while excluding speech that may sound similar but is not hateful, like news articles, song lyrics, or satire? Or, is hate speech something that you inherently recognize when you see it?

What if we could use rules, tests, and parameters to isolate hate speech? Can we identify and analyze at elements like speaker intent, context, identity, tone, audience, or any number of indicators that transform words into meanings and change an innocuous statement into a verbal assault?

Combating the proliferation of online hate speech and understanding its mechanics is a complex undertaking. We believe, however, that it can be done. And one way we are working to achieve this is by teaching machines to understand hate.

The Online Hate Index (OHI), a joint initiative of ADL's Center for Technology and Society and U.C. Berkeley's D-Lab, will transform human understanding of hate speech via machine learning into a scalable tool that can be deployed on internet content to understand the scope and spread of online hate speech. Through a constantly-evolving process of machine learning, based on a protocol developed by a team of human coders as to what does and does not constitute hate speech, this tool will uncover and identify trends and patterns in hate speech across different online platforms, allowing us to push for the changes necessary to ensure that online communities are safe and inclusive spaces. We are currently just past our first milestone in making this a reality, and are eager to move on the next phase of the project. Critically, the tool is able to identify individual instances of hateful and abusive speech, helping solve a problem that has been inadequately addressed through reliance on platform users to report instances of abuse and violations of terms of service agreements.

Proactive moderation of hate speech and abuse in online communities can effect substantial changes in online environments. A notable example is Reddit, the massively popular web forum that is comprised of around one million user-generated community boards called "subreddits." Subreddits cover a wide-range of topics, from the unusual to the unsavory. While this has made the website inviting for a plethora of groups, organizations, and communities, it also made Reddit home to those with the goal of spreading racism, misogyny, anti-Semitism, homophobia, and all other forms of hate. In October 2015, Reddit took the action of closing down a number of its more noxious, hate-fueled subreddits. Researchers studying the response to this action found that users who frequented the shut-down subreddits engaged in fewer instances of hate speech as they spent time on other subreddits, and that the overall use of hateful rhetoric throughout the entire website diminished as a result of banning the small number of spaces that were dedicated to and encouraging of discriminatory and hateful speech. Reddit has also provided a training ground for our machine learning model, which has combed through thousands of user comments in order to, with the help of human coders, learn to identify hate speech.

Online communities have been described as the modern public square, a space for opinions to be expressed and voices to be heard. In reality, though, not everyone has equal access to this public square, and not everyone has the privilege to speak without fear. Hateful and abusive online speech forces out other voices; excluding the voices of the marginalized and underrepresented from public discourse.

Through combining social science and machine learning, the OHI holds the promise to bring more humanity to the internet. By helping us understand speech on the internet, the OHI will not just make online communities safer and more inclusive, it will make them more protective of speech and more welcoming to a wide array of voices.

In this document, we will outline the conceptualization and operationalization of online hate speech and the building of the machine learning model to understand it. We will also discuss the necessary techniques to make the machine learning model as accurate as possible, and some initial results, which give indications of the features of speech that are most commonly used when deciding if a reader would consider an online comment to be hate speech or not. Finally, we will discuss the way forward, and how we see the OHI scaling up and functioning in the broader online world.

What is Machine Learning?

The OHI is a sentiment-based analysis that runs off of artificial intelligence and machine learning. All the decisions that went into each step of creating the OHI were done with the aim of building a machine learning-enabled model that can be used to identify and help us understand hate speech online.

Machine learning is a field that spans both computer science and statistics. It starts with observational information. The goal of machine learning is to help computers discern patterns in information without explicitly telling the computer what these patterns are.

After the computer constructs a model of these patterns, that model can be used for prediction.

In order to do this, a machine first needs to take in a very large set of information that is identified on a very basic level. For a machine to be able to predict a type of flower, for example, from several different types, it needs to know a few things about each type. What color is it usually? What are the petal measurements? How many petals does it usually have? We give this information to the computer in observations. With each observation, the computer then begins to identify which "features" are important to each type of flower, and how important they are to deciding which type of flower we're talking about.

After doing this, and going through many "training" examples to increase its accuracy and reliability, a fully-trained machine learning model could allow you to enter in the color and petal measurements of a flower, and it would be able to predict, out of several types of flowers, which type it is. More advanced algorithms (called "deep learning") can even create those different measurements automatically when given only photos of the flowers.

What follows here are the first important steps in building a machine learning model that can look at hate speech online in a similar way.

Early Decisions: What, Where and When

In April 2017, the CTS team met with Berkeley D-Lab to brainstorm ideas to confront online hate with the resources and combined expertise of both organizations. Out of that session came the idea to apply social science methods to create a machine learning model trained to evaluate the scope and spread of hate speech online. If successful, CTS would then incorporate that algorithm into its work combatting hate of all kinds, and working with the tech industry to fight hate on their platforms.

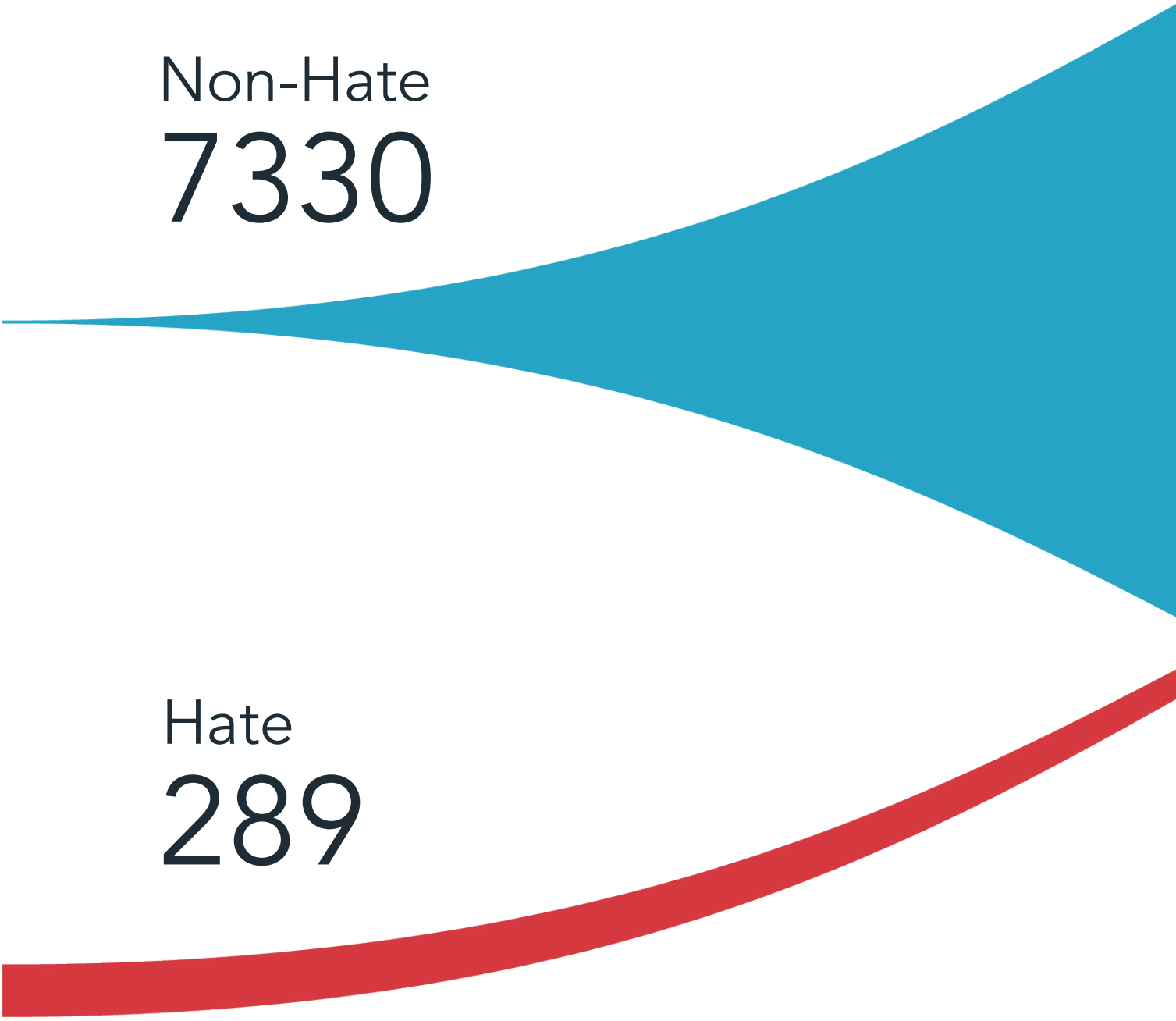
The organizations started the process by picking a platform to focus on for the first phase of the project and a timeframe to pull in hate speech from that platform. After weighing the options, they chose Reddit because it uses volunteers to monitor and regulate speech. This can, however, result in irregular and sometimes nonexistent content moderation. Therefore, a portion of the hate speech found on from Reddit was considered by researchers to be "uncensored."

The team also chose Reddit because academic researchers studying hate speech have been focusing on far-right, white supremacist, and other extremist sites. Although sites like these are dense with examples of hate speech to be classified, Reddit contains both hate speech and other kinds of speech. In order to build the algorithm and make it accurate, the team wanted to train it with speech information with a variety of features -- both what the study was looking for (hate speech) and what it was not (other kinds of speech).

Finally, the team chose Reddit because ultimately the aim was to create a product that looks less at speech by self-identified haters and more at everyday uses of hate speech through the eyes of a typical, non-extremist, user. By looking at less niche-based forms of hate speech, the learning that the algorithm does in this training phase will be more applicable to other platforms in subsequent phases. For all these reasons, the Reddit platform was ideal for this project.

The next step was to determine the ideal timeframe to pull down comments containing hate speech from Reddit. The team pulled down approximately 80,000 forum comments from both left and right leaning area of Reddit during two specific time periods which we felt would correspond with an intensification of hate speech toward minorities, and in particular, immigrant populations. The first period was for approximately one month, following June 16, 2016, when then-candidate Trump announced his presidential bid and promised to build a border wall with Mexico, stating that some Mexicans immigrants are "rapists" and are "bringing crime." The second time period was one month starting October 19, 2016, which is the date of the third presidential debate; when then-candidate Trump described Mexican immigrants as "bad hombres." The team pulled these comments initially to focus on two critical political moments that sparked online conversation particularly dense with hate speech across the political spectrum.

Reddit Comments Labeled Hate or Not Hate



Total Comments

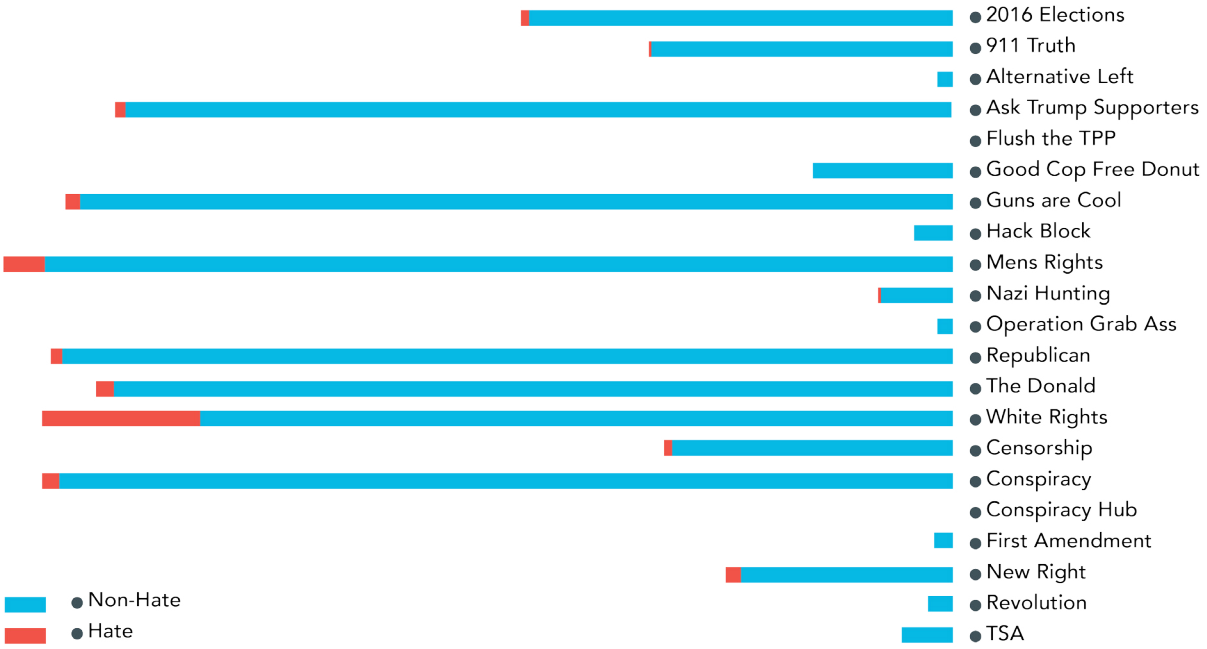
7619

Creating the Social Science Methodology

After scraping the Reddit comments from the chosen periods, the next step was to develop a methodology to manually review and hand label each of the Reddit comments as hate or not hate, based on a functional definition of hate speech derived from the research of social scientists. Manual review is a burdensome task. But it is necessary, however, because unlike other projects where sets of labeled information may exist, no datasets are available in the realm of hate speech. As such, we set about labelling the information ourselves.

Labeled comments are necessary because for a machine learning model to learn about the data, it needs to be provided with training examples or examples where we have both the comment and the human coding of that comment. Only then can it learn which features of the comment are important in making a decision. The hand labeling of comments gives the model basic information about what is hate speech and what it is not, allowing the model to break down the characteristics from the text that go into making that determination.

Proportion of hate vs. non-hate by subreddit:

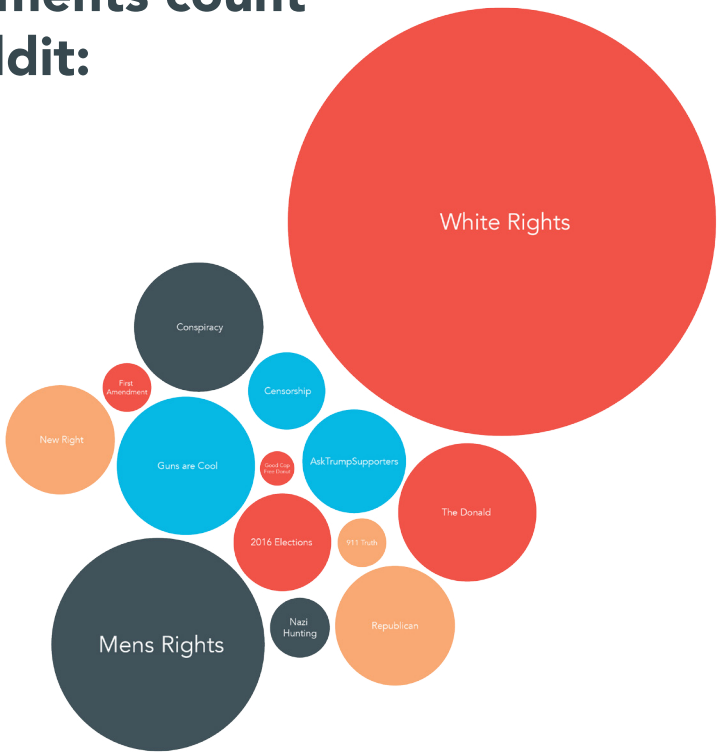


The first step in labeling the comments was to agree upon a definition of hate speech. The definition the team selected comes from the Encyclopedia of Political Communication, which D-Lab defined as the benchmark standard in academic work on this topic:

"Comments containing speech aimed to terrorize, express prejudice and contempt toward, humiliate, degrade, abuse, threaten, ridicule, demean, and discriminate based on race, ethnicity, religion, sexual orientation, national origin, or gender.… Also including pejoratives and group-based insults, that sometime comprise brief group epithets consisting of short, usually negative labels or lengthy narratives about an out group's alleged negative behavior."

Based on this definition, the D-Lab team put together a codebook to guide the reviewers who would be doing the work of labeling the many comments as hate or not hate, based on the rubric. Next, a team of ten undergraduate research assistants from a variety of majors and personal backgrounds were trained to consistently label comments in the method laid out in the codebook. Of the 80,000 comments pulled off of Reddit, 9,000 comments were used in this and further parts of this phase of the project. The research assistants were each assigned a random portion of the 9,000 comments and identified each comment as either hate or not hate.

Hate comments count by subreddit:



The initial codebook also included a more detailed look at each comment. In addition to determining whether a comment was hate or not hate, the research assistants were asked to determine themes present in the comment, such as whether it was an insult or profanity, part of a conspiracy theory, sarcasm, or a threat. They also looked at the targets of the comments and labeled them to see if the target was ambiguous, implicit, or explicit.

In order for a machine learning model to be effective in interpreting information, the way that information is labeled needs to be very accurate and reliable. To make sure the hand coding of the research assistants reached a high level of accuracy and reliability, the research team conducted trainings and worked through three rounds of sample coding. Following these rounds we reviewed the comments with the research assistants to ensure accuracy and reliability.

How did we Deal with Disagreement in the Labelling

Determining where comments fell within the category of hate/not-hate was the first phase of the OHI. While the labelled comments reached a level of accuracy necessary to use it in the machine learning model, there were still disagreements about what qualified as hate speech. This is due, in part, to the nature of studying hate speech. The research assistants that did the hand coding were comprised of students from diverse backgrounds with respect to race, gender, ethnicity, national origin, primary language, and sexual orientation, among other factors. This diversity was an asset in terms of the different subjective views, identities, and experiences in terms of perceiving hate speech. This was valuable in the development of the codebook as well as informative in the development of our methodological processes.

For the purpose of this phase of the project comments were labelled as hate/not hate. In the next phase of the project, the team will continue with the process of working downwards in specificity from labeling comments for hate to looking at online hate speech against individual targeted groups. In future phases of the project, we will train the machine learning model to look into questions of anti-Semitism/not anti-Semitism, and then in partnership with other civil rights organizations, the partners will look into hate speech against other identity groups, like immigrant groups.

Building the Model

Once the 9,000 Reddit comments were labeled, the researchers fed them into the machine learning model. The machine learning model built its own rules to separate one piece of information from another.

Before machine learning methods were used to classify different kinds of text, dictionary-based and rule-based approaches were standard types of analysis. They involved predefining a list of words that are indicative of hate speech and marking sentences or documents as hate speech if they contain a certain number of these words. For example, a more sophisticated version of this might include a series of "if...then" statements -- for example, if a sentence has both the word "useless" and a particular ethnicity, then that sentence is classified as hate speech. It is not hard, however, to create non-hate speech sentences including both "useless" and the name of an ethnicity. The problem with these dictionary and rule based approaches is one of coming up with the right rules that capture all and only hate speech while excluding non-hate speech.

The advantage of our machine learning model is that the machine makes the rules after looking at many examples of what a person has classified as hate speech or not hate speech. In building this model, the researchers are not saying that this algorithm has the absolute definition of what hate speech is or is not. What the study demonstrates is that the machine can determine the factors that go into a person's decision to see a comment online as hate speech or not hate speech. The OHI model is capturing the experience of hate speech online, and not promoting a definition.

To capture this experience in all its complexity, the machine learning team used a technique called "word embeddings." What word embeddings do is look at each word in a comment and assess it in terms of 300 different abstract categories. These abstract categories serve to expand the definition of that word beyond our simple Webster's dictionary understanding, and incorporate things like context and co-occurrence into its semantics. These and other features of the comment were used to train the model. Traditional machine learning techniques made it so our machine learning model was 78% accurate, meaning that 78% of the time the rules that the machine model built would allow it to match the determination of the research assistant who hand labeled a Reddit comment as hate or not hate. With addition of the word embeddings technique, the accuracy of the OHI model rose to 85%.

Implications

The results of this first phase of the machine learning model are fascinating, but are also limited at this phase in the study. Again, the model is not giving characteristics of hate speech, but rather identifying characteristics of a person's decision to see a comment as hate speech or not. Additionally, while 9,000 Reddit comments is enough to start training the model, and to start understanding what it is showing us about the decisions people make about hate speech online, it is not enough to draw any broad conclusions on the Reddit platform, or to online forums, generally.

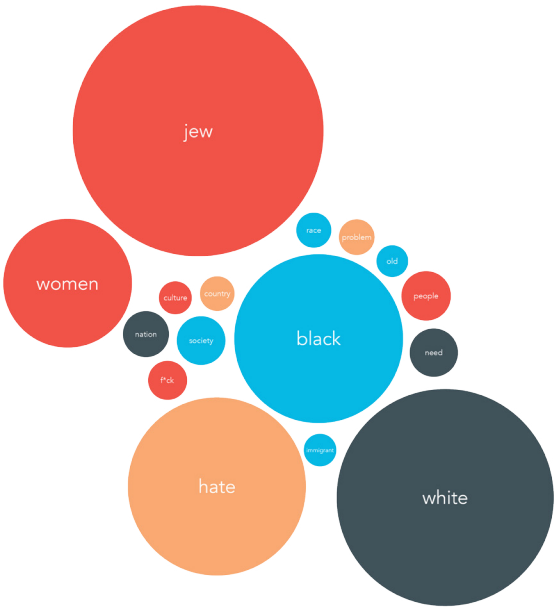
Hate Against One, Hate Against All

Initial results showed that when you look for one kind of hate, you end up finding hate of all kinds. The graphics below and on the next page show three collections of words: the right below shows a raw count of the most frequent words that show up in non-hate comments on Reddit during the relevant time periods. The left shows a raw count of the top twenty words that show up most frequently in hate-related comments, while the image on the next page shows a list of the twenty words that were most strongly associated with hate speech over non-hate speech.

Raw count of words most common in Hate Comments

Raw count of words most common in Non- Hate Comments

The difference in these lists is clear: while the words "like" and "don't" and "one" may frequently appear in hate speech comments, they also appear in comments that are not hate speech at a frequency that makes them unremarkable. The words in the left column represent words that appear more frequently in hate speech and less frequently in non-hate speech, and are thus more strongly associated with hate.



Words strongly associated with hate speech

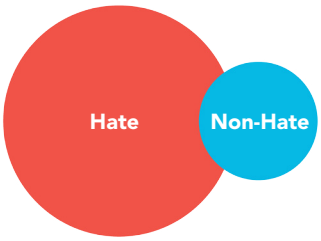
The fact that "Jew" appears as the word that is most strongly associated with hate speech on Reddit is equally disturbing and unsurprising. ADL has long-known that anti-Semitism is the lingua franca of hate speech. Some scholars trace expressions of modern anti-Semitism in Europe from questions about civic engagement, as populations were engaged in debates over who would be granted citizenship and who would continue to be an outsider. Under this rubric, the prevalence of anti-Semitic language, at a time when ADL has seen anti-immigrant sentiment is on the rise in America, is not surprising. Its appearance in early results of the OHI is early indicator that the machine learning model maps to ADL's real-world experience combating cyberhate.

Additionally, as the graphics below demonstrate, the top 20 words most associated with hate include a wide variety of populations targeted by hate for their race, gender, religion, or national origin. These quantitative results demonstrate the interconnected nature of hate. In the literature review performed by the D-Lab, we found that people who spew hateful comments are often attempting to isolate, scare, and divide people.

The ADL's long standing mission has been to stop the defamation of the Jewish people and ensure justice and fair treatment for all. If we stand up for one person, we must stand up for all people. What this list shows is an encouraging early result that demonstrates the importance of doing this collaborative work to fight hatred of all kinds.

The Grammar and Structure of Hate Speech

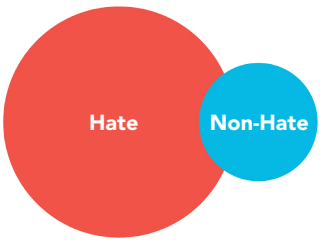
Another notable early result of the OHI model looked at hate speech online not only in terms of hateful words, but in terms of how language itself functions.



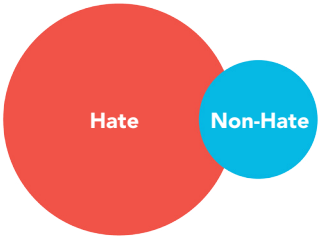
Number of words in Hate vs Non-Hate Comments

To go into more depth, researchers looked at the number of words in a hateful comment versus a non-hateful comment. It found that the average number of words in a hateful comment was typically longer than a non-hateful one. Likewise, the average number of words in all-caps font in hateful comments was slightly larger than those in a non-hateful one. Finally, the researchers found that the sentence length in hateful comments was slightly longer than in non-hateful comments. Overall, on a very basic level, hateful comments were wordier, more lengthy, and included more vehement "yelling" in all-caps.

Number of all caps words in Hate vs. Non-Hate Comments



What this shows is that, while most attempts to detect hate speech are focused on dictionary-based searches of certain hateful words, this kind of search is only one limited dimension with which to look at hate speech online. Tracing the grammatical and linguistic attributes of hate speech may help researchers identify characteristics that undergird hate speech, separate from the contextual meaning.

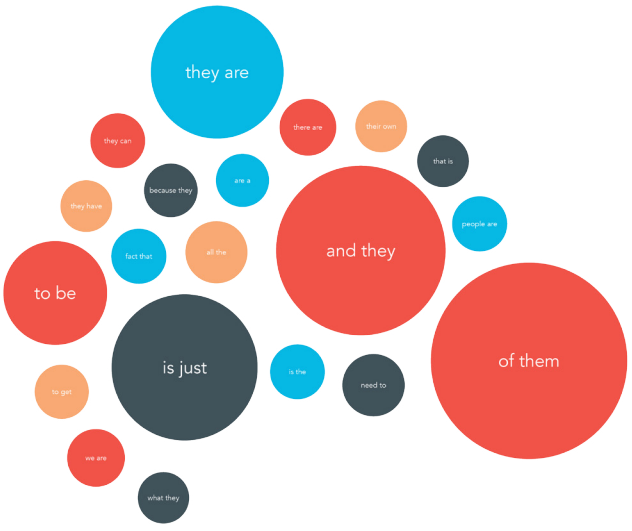


Sentence Length in Hate vs Non-Hate Comments

Broken Attempts to Communicate

Apart from the hateful words in the initial results, the team observed that among the words that the model identified as most related to hate speech (over words related to non-hate speech) are words that do not clearly relate to the targeting of groups. This includes terms like: "need" and "know." "Like" is third in the dataset of raw words associated with hate speech, and is grammatically used to introduce similes, which are comparisons between two unalike things. What this may show is that, beyond the targeting of groups in hateful comments, there may be an attempt at logical reasoning being made by those who generate hate speech.

Two-word phrases most strongly associated with hate speech:



This observation becomes even more striking when looking at the list of two-word phrases that were most associated with hate. None of these two-word phrases specifically targeted a group. Instead, they are explanatory phrases, such as "of them", "and they", "is just", "to be", "need to," and logical connectors like "because they" and "fact that." What may be happening here is a demonstration of logical processes -- however twisted by hate -- in the language of the authors. If people generating hateful comments are also attempting to apply a form of logic to justify their hate, perhaps it may be possible to engage them. While more study will be necessary, the early indicators shown here may be part of a quantitative analysis of what is going on in hater's minds. To them, logic, thought, and explanation is implicated in these hateful comments. This is a preliminary sign to the CTS team that educational interventions, communication strategies, and applications of counterspeech can potentially have increased value to diffuse hate speech.

The Implications of Multiple Perspectives

Broadly, what the initial OHI model imply show us -- between (1) looking for hate speech of one kind leads to finding hate speech of all kinds; (2) the importance of looking at the uses of language when talking about hate; and (3) faulty or misapplied logic intertwined with online hate speech may create an opening for counterspeech interventions -- is that creating an absolute definition of hate speech online should not be our aim. We don't need to rewrite the dictionary. Our takeaway from combining social science and quantitative analysis is that the most powerful tool we have to understand the scope and spread of hate speech online is the experience of those who have been targeted by it.

The OHI methodology demonstrated that the unique identities and perspectives of people targeted by hate speech was the most important factor in determining what was and was not considered to be hateful by the reviewers. Just as context is key in trying to determine the meaning of hate speech, there is no one universal definition of the phenomenon. To understand what is truly hateful, the perspective of communities targeted by this language must be incorporated.

Our approach of using an intentionally-diverse team to code comments as hate or not hate, and then have the machine learn from their determinations, helped our model draw from the broadest spectrum of experience we had available. As we continue to train the model, we will endeavor to broaden that diversity even more. These encouraging initial results align with ADL's approach to creating effective community partnerships to enact social change. We intend to go further in subsequent stages and work with diverse groups to conduct more studies of online hate.

The Way Forward

Beyond these results, the next phase of this project will go beyond the hate/not analysis and look in a more detailed manner at a population that is typically the target of hate speech online. Given ADL's historic expertise in tracking, monitoring and understanding anti-Semitism, we are likely to pursue this avenue as part of this next phase, and will identify an appropriate set of comments to be labeled. Additionally, as part of this next phase, the team at D-Lab will be identifying a crowdsourcing platform for labeling the next set of comments. Crowdsourcing will allow the OHI project team to label far more comments than our team of research assistants were able to complete, and will allow the research team to have more geographic diversity.

Once we have incorporated more specific information on hate against target groups and have trained the machine learning model to recognize hate against those groups on platforms other than Reddit, we anticipate that we will be able to start deploying the OHI machine learning model more broadly. Once a model is appropriately trained, it takes nothing but more computational power to have the model classify hate speech around the internet and across domains, without additional labeling---how much can such a model tell us about the distribution of hate speech in online communities?

Looking beyond this second phase and into the results of the next two years of work, we imagine the OHI providing a suite of services for the tech industry to be able to search out hate speech on their platform in all its thorny nuances. After the events in Charlottesville earlier this year, we found that tech platforms were reaching out to ADL for help with this problem, and were hoping to find machine learning-oriented solutions.

We have a long way to go until we get there, but we do believe it is possible. We believe in solutions, and the concrete steps necessary to get there. Recently, the astronaut Sandy Magnus spoke about returning to earth after her first flight in space, and the feeling of experiencing gravity after being weightless for several days. "It felt unnatural." she said "Like this oppressive force was suddenly pushing down on me. And I asked myself 'How do we live every day like this?'" We have become too used to the oppressive force of hate being the norm online. We want users of the internet in the years to come to have the same feeling that Sandy Magnus had, so that they can look back on the online world of today and wonder "How did we live every day like that?"

At ADL's Center for Technology and Society, we are dedicated to making this a reality, and we are dedicated to working with passionate and capable partners like the Berkeley D-Lab to do the work that's necessary. We believe strongly that the Online Hate Index will help to move the needle forward towards a better future online.