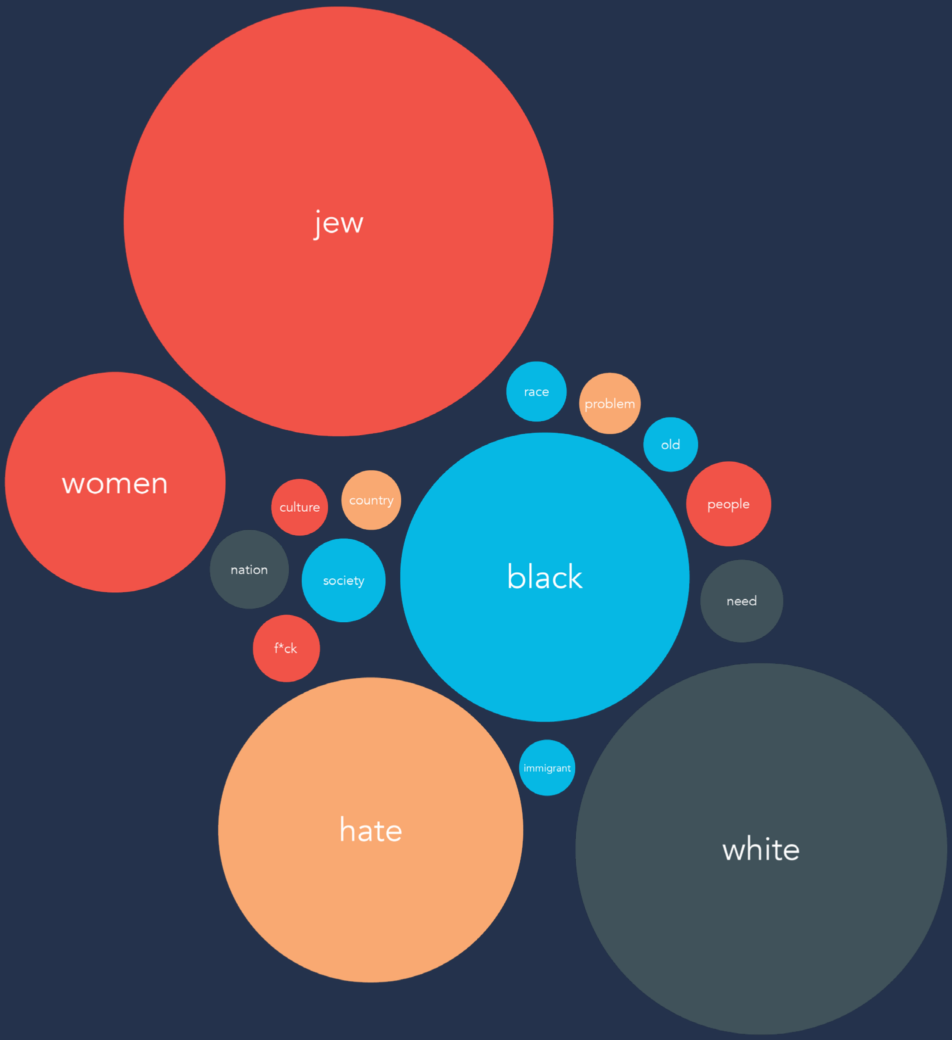
CTS Center for Technology & Society

Online Hate Index

Innovation Brief Executive Summary



ADL

Introducing ADL's Online Hate Index, Phase 1

Online communities have been described as our modern public square. In reality, though, not everyone has equal access to this forum, and not everyone has the privilege to speak without fear. Hateful and abusive online speech squelches other voices; it excludes the viewpoints of marginalized and underrepresented people from public discourse and poisons political conversations.

The Online Hate Index (OHI) aims to help us understand and work to elevate all voices, and to ensure that online communities become more safe and inclusive spaces.

The OHI, a joint initiative of ADL's Center for Technology and Society and the University of California at Berkeley's D-Lab, has tremendous potential to increase our ability to understand the scope and spread of online hate speech. Combining a constantly-dynamic mechanism based on artificial intelligence and machine learning, and a social science methodology applied by a team of human coders, the OHI ultimately will uncover and identify trends and patterns in hate speech across different online platforms.

The OHI project has completed its first phase of research. In this phase, the OHI algorithm has started to learn the difference between hate speech and other kinds of speech, laying the foundation to help solve a problem that has been inadequately addressed through traditional content moderation. The ultimate goal is to develop an efficient way to determine what is and what isn't hate speech online, and then scale the method to help the tech community understand this pressing concern.

Process

For the first phase of the project, the researchers collected a small sample of comments on Reddit during two months in 2016 to begin the process of creating the OHI model. At the same time, the D-Lab developed a social science methodology based on a chosen definition of hate speech. The D-Lab then assembled a team of research assistants from diverse backgrounds with respect to race, gender, ethnicity, national origin, primary language, and sexual orientation. The research assistants were trained on the definition and methodology, and then manually labeled each of the comments as either hate or not hate.

Once the researchers completed labeling the comments, they fed them into the machine learning model. The model established rules after looking at many examples of what people have classified as hate speech or not hate speech. This machine learning-based algorithm can determine the factors that go into a person's decision to consider whether text is hate speech or not. Therefore, the OHI model captures the experience of hate speech online – it does not define hate speech.

Findings

Preliminary results from the model found that when searching for one kind of hate, it's easy to find hate of all kinds. In the early results, there were several words that appeared more frequently in hate speech and less frequently in non-hate speech. The top 5 words most strongly associated with hate the sampling pulled were: Jew, white, hate, women, and black.

The model also found patterns in the construction of hateful language. Researchers found that the average number of words in a hateful comment in our dataset was typically longer than in non-hateful comments. Likewise, on average, there were slightly more words in all caps found in hateful comments than in non-hateful ones. Finally, researchers found that the sentence length in hateful comments was slightly longer than in non-hateful comments. Overall, hateful comments tended to be wordier, lengthier, and included more "yelling" (all caps).

What's Next

The next phase of this project will go beyond a hate vs. non-hate analysis and turn to looking at specific targeted populations in a more detailed manner. Additionally, the D-Lab is identifying strategies to scale the process for labeling comments.

While there is still a long way to go with artificial intelligence and machine learning-based solutions, ADL and the D-Lab believe these technologies can go a long way to curbing online hate speech.

The ADL's long standing mission has been to stop the defamation of the Jewish people and secure justice and fair treatment to all. If we stand up for one person, we must stand up for all people, and the Online Hate Index will help us do just that.