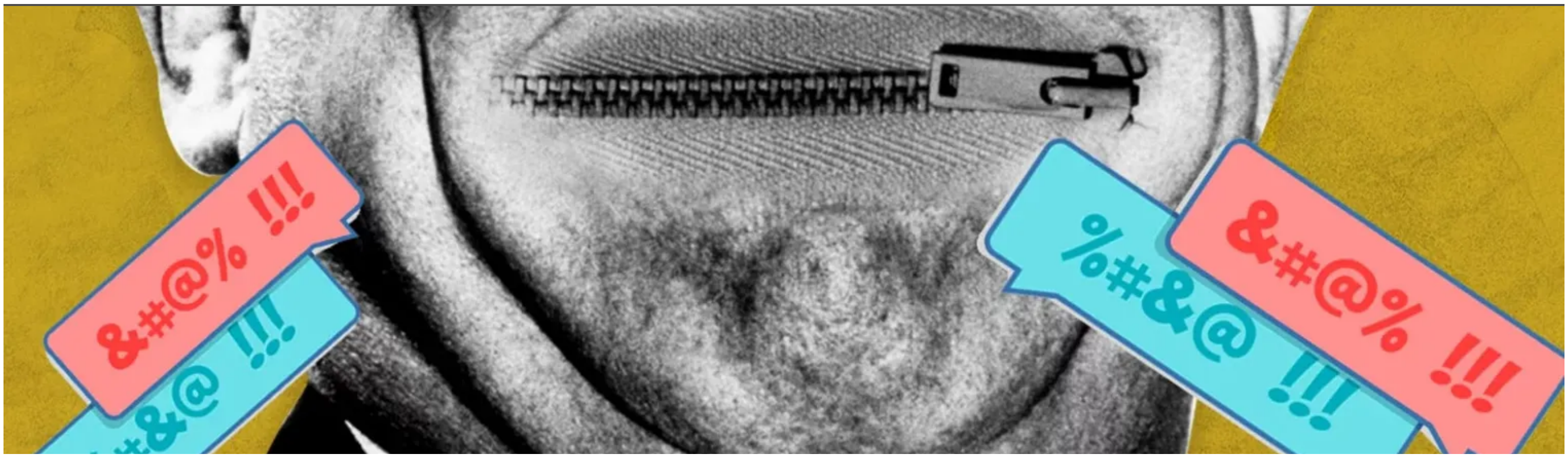
Synced



AI TECHNOLOGY

Researchers Use Vile Comments from Trump Subreddit to Train AI to Battle Hate Speech

BY SYNCED
2019-09-24

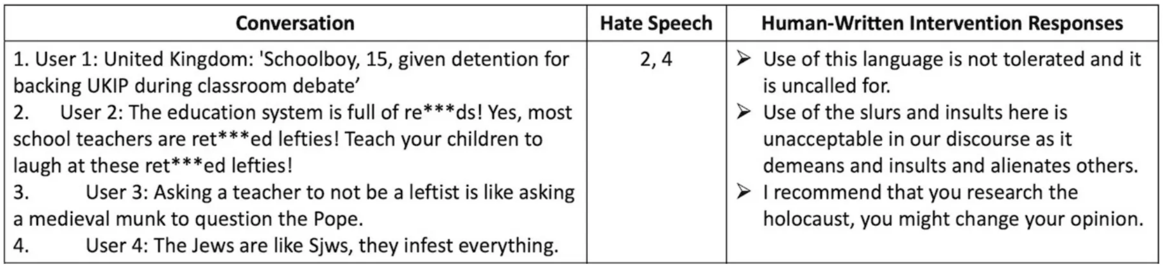
COMMENTS 0

Social media platforms like Facebook and Twitter have imposed rigorous policies in an effort to combat hate speech and extremism. Existing AI-based policing models however tend to simply detect and delete objectionable posts based on keywords.

Now, researchers from Intel AI and University of California at Santa Barbara have introduced a new generative hate speech intervention model, along with two large-scale fully-labeled hate speech datasets collected from Reddit and Gab.

The standout feature of the research is that along with hate speech detection, the datasets can also provide tailored intervention responses written by Amazon Mechanical Turk workers. In this way an AI model can be trained to both detect hate speech and generate appropriate responses for specific types of hate speech.

“Simply detecting and blocking hate speech or suspicious users often has limited ability to prevent these users from simply turning to other social media platforms to continue to engage in hate speech as can be seen in the large move of individuals blocked from Twitter to Gab,” the researchers explain.



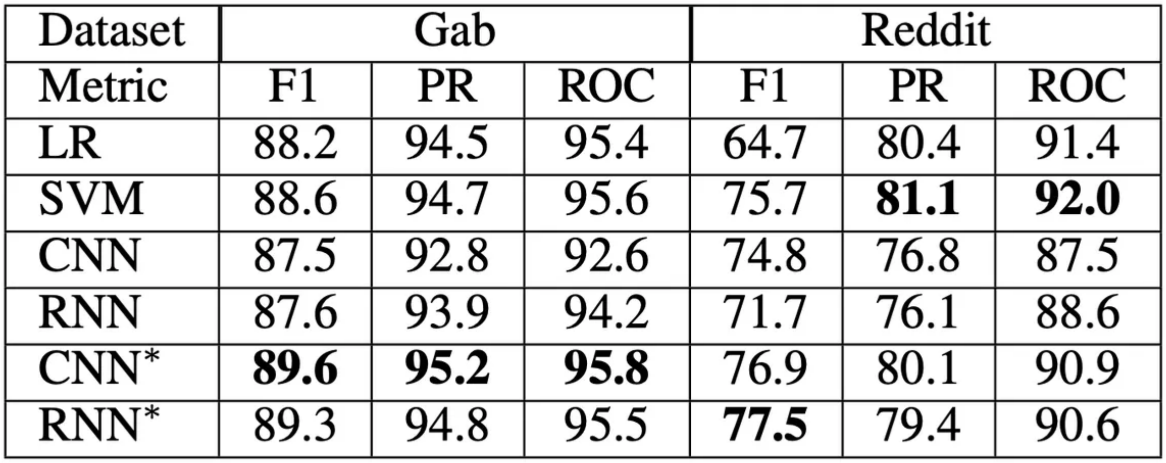
Conversation	Hate Speech	Human-Written Intervention Responses
1. User 1: United Kingdom: 'Schoolboy, 15, given detention for backing UKIP during classroom debate' 2. User 2: The education system is full of re***ds! Yes, most school teachers are ret***ed lefties! Teach your children to laugh at these ret***ed lefties! 3. User 3: Asking a teacher to not be a leftist is like asking a medieval munk to question the Pope. 4. User 4: The Jews are like Sjws, they infest everything.	2, 4	➢ Use of this language is not tolerated and it is uncalled for. ➢ Use of the slurs and insults here is unacceptable in our discourse as it demeans and insults and alienates others. ➢ I recommend that you research the holocaust, you might change your opinion.

The datasets consist of 5,020 conversations retrieved from Reddit pages **such as "r/The Donald," a subreddit for discussion on US President Donald Trump that was "quarantined" earlier this year for incitements to violence.** The research team used keywords to identify potentially hateful comments and then reconstructed the conversational context of each comment. The dataset also contains 11,825 conversations retrieved from right-wing discussion platform Gab.

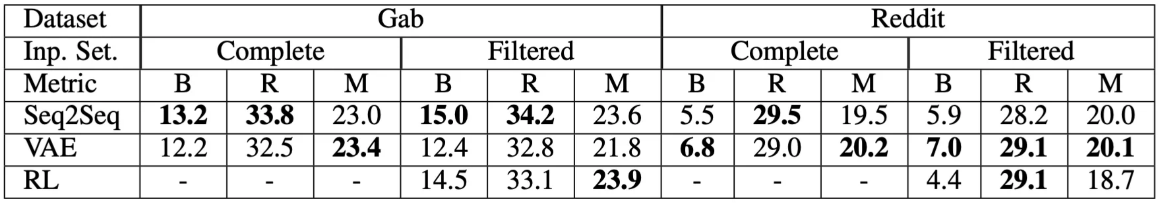
The research team crowd-sourced workers from Amazon Mechanical Turk to label the comments and generate intervention responses on a case-by-case basis. The workers were asked to answer two questions:

1. Which posts or comments in this conversation are hate speech?
2. If there exists hate speech in the conversation, how would you respond to intervene? Write down a response that can probably hold it back (word limit: 140 characters).

In their experiments, researchers evaluated four methods on a binary hate speech detection task: Logistic Regression (LR), Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). They also evaluated three models on generative hate speech intervention tasks: Seq2Seq, Variational Auto-Encoder (VAE), and Reinforcement Learning (RL). The results are below.



Dataset	Gab			Reddit		
Metric	F1	PR	ROC	F1	PR	ROC
LR	88.2	94.5	95.4	64.7	80.4	91.4
SVM	88.6	94.7	95.6	75.7	**81.1**	**92.0**
CNN	87.5	92.8	92.6	74.8	76.8	87.5
RNN	87.6	93.9	94.2	71.7	76.1	88.6
CNN*	**89.6**	**95.2**	**95.8**	76.9	80.1	90.9
RNN*	89.3	94.8	95.5	**77.5**	79.4	90.6



Dataset	Gab						Reddit					
Inp. Set.	Complete			Filtered			Complete			Filtered		
Metric	B	R	M	B	R	M	B	R	M	B	R	M
Seq2Seq	**13.2**	**33.8**	23.0	**15.0**	**34.2**	23.6	5.5	**29.5**	19.5	5.9	28.2	20.0
VAE	12.2	32.5	**23.4**	12.4	32.8	21.8	**6.8**	29.0	**20.2**	**7.0**	**29.1**	**20.1**
RL	-	-	-	14.5	33.1	**23.9**	-	-	-	4.4	**29.1**	18.7

Bots have a spotty history when it comes to racist or inflammatory content — several years ago the Microsoft online digital assistant “Tay” was prompted to spew a series of racist and inflammatory tweets before her handlers pulled the plug. And a recent paper from the Seattle-based Allen Institute for Artificial Intelligence (AI2) showed how even relatively innocent trigger words and phrases can be used to “inflict targeted errors” on natural language processing (NLP) models, triggering the generation of racist and hostile content.

With both vulgar humans and rogue bots to contend with in the online arena, the Intel and UC Santa Barbara datasets provide a valuable tool for both detection of and intervention on hateful comments.

The paper *A Benchmark Dataset for Learning to Intervene in Online Hate Speech* is on arXiv. The dataset has been open-sourced on GitHub.

Journalist: Tony Peng | **Editor:** Michael Sarazen