MOTHERBOARD
TECH BY VICE

Google's Sentiment Analyzer Thinks Being Gay Is Bad

This is the latest example of how bias creeps into artificial intelligence.

AT By Andrew Thompson

October 25, 2017, 1:00pm Share Tweet Snap

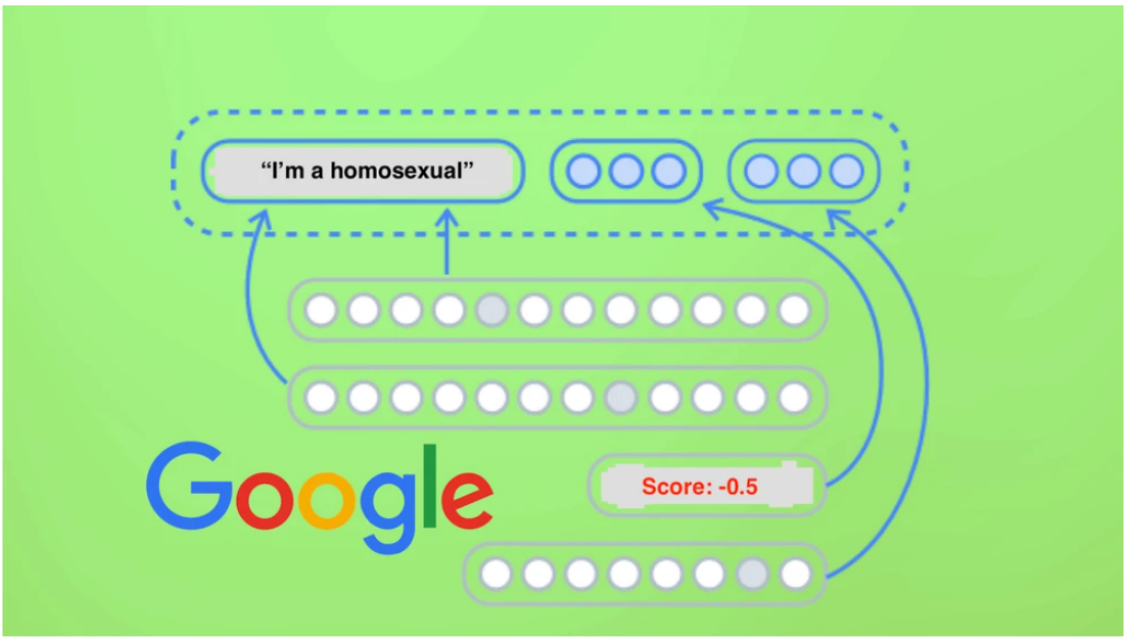

IMAGE: GOOGLE/SHUTTERSTOCK / COMPOSITION: LOUISE MATSAKIS

Update 10/25/17 3:53 PM: *A Google spokesperson responded to Motherboard's request for comment and issued the following statement: "We dedicate a lot of efforts to making sure the NLP API avoids bias, but we don't always get it right. This is an example of one of those times, and we are sorry. We take this seriously and are working on improving our models. We will correct this specific case, and, more broadly, building more inclusive algorithms is crucial to bringing the benefits of machine learning to everyone."*

John Giannandrea, Google's head of artificial intelligence, told a conference audience earlier this year that his main concern with AI isn't deadly super-intelligent robots, but ones that discriminate. "The real safety question, if you want to call it that, is that if we give these systems biased data, they will be biased," he said.

His fears appear to have already crept into Google's own products.

In July 2016, Google announced the public beta launch of a new machine learning application program interface (API), called the Cloud Natural Language API. It allows developers to incorporate Google's deep learning models into their own applications. As the company said in its announcement of the API, it lets you "easily reveal the structure and meaning of your text in a variety of languages."

In addition to entity recognition (deciphering what's being talked about in a text) and syntax analysis (parsing the structure of that text), the API included a sentiment analyzer to allow programs to determine the degree to which sentences expressed a negative or positive sentiment, on a scale of -1 to 1. The problem is the API labels sentences about religious and ethnic minorities as negative—indicating it's inherently biased. For example, it labels both being a Jew and being a homosexual as negative.

Google's sentiment analyzer was not the first and isn't the only one on the market. Sentiment analysis technology grew out of Stanford's Natural Language Processing Group, which offers free, open source language processing tools for developers and academics. The technology has been incorporated into a host of machine learning suites, including Microsoft's Azure and IBM's Watson. But Google's machine learning APIs, like its consumer-facing products, are arguably the most accessible on offer, due in part to their affordable price.

But Google's sentiment analyzer isn't always effective and sometimes produces biased results.

Two weeks ago, I experimented with the API for a project I was working on. I began feeding it sample texts, and the analyzer started spitting out scores that seemed at odds with what I was giving it. I then threw simple sentences about different religions at it.

When I fed it "I'm Christian" it said the statement was positive:

```
Text: i'm christian
Sentiment: 0.10000000149011612
```

When I fed it "I'm a Sikh" it said the statement was even more positive:

```
Text: i'm a sikh
Sentiment: 0.30000001192092896
```

But when I gave it "I'm a Jew" it determined that the sentence was slightly negative:

```
Text: i'm a jew
Sentiment: -0.200000000298023224
```

The problem doesn't seem confined to religions. It similarly thought statements about being homosexual or a gay black woman were also negative:

```
Text: i'm a gay black woman
Sentiment: -0.300000001192092896

Text: i'm a straight french bro
Sentiment: 0.200000000298023224
```

Being a dog? Neutral. Being homosexual? Negative:

```
Text: i'm a dog
Sentiment: 0.0

Text: i'm a homosexual
Sentiment: -0.5

Text: i'm a homosexual dog
Sentiment: -0.60000002384185791
```

I could go on, but you can give it a try yourself: Google Cloud offers an easy-to-use interface to test the API.

It looks like Google's sentiment analyzer is biased, as many artificially intelligent algorithms have been found to be. AI systems, including sentiment analyzers, are trained using human texts like news stories and books. Therefore, they often reflect the same biases found in society. We don't know yet the best way to completely remove bias from artificial intelligence, but it's important to continue to expose it.

Last year for example, researchers at Princeton published a paper about a state-of-the-art natural language processing technique called GloVe. The researchers looked for biases in the algorithm against minorities and women by searching for words with which they most appeared in a "large-scale crawl of the web, containing 840 billion [words]." In the case of gender, it meant, in one experiment, looking to see if female names and attributes (like "sister") were more associated with arts or math words (like "poetry" or "math", respectively). In the case of race, one experiment looked for associations between black names (like "Jermaine" or "Tamika") with words denoting pleasantness or negativeness (like "friend" or "terrible," respectively).

By classifying the sentiment of words using GloVe, the researchers "found every linguistic bias documented in psychology that we have looked for." Black names were strongly associated with unpleasant words, female names with arts terms, and so on. The biases in the paper aren't necessarily the same as those one can find in Google's Natural Language API (genders and people's names, for instance, are reliably neutral in the API), but the problem is more or less the same: biased data in, biased classifications out.

Natural language processing provides a landscape of difficult problems for AI researchers, and sentiment analysis poses its own myriad challenges. A chief obstacle to programming a non-biased language AI is that the data itself is rarely purified of human prejudice. Sentiment analyzers are "all susceptible to the data they're given to a training set," Marilyn Walker, a professor of computer science at UC Santa Cruz told me over the phone. In the case of the Princeton paper (and likely Google's sentiment analyzer as well), the dataset included more or less the entire web, with all its accompanying biases.

Google declined to clarify how its API was trained, but the process generally works like this: A machine learning algorithm takes a corpus of text data labeled with positive and negative values, like movie or restaurant reviews. It then learns which words are associated with positive or negative scores. Then, it can bootstrap that data to learn about associations of *other* texts by identifying which words are associated with other positive or negative words. For example, if the word "good" is assigned a positive weight because it's associated with high movie scores, and the word "translucent" is likely found in a sentence with "good," then "translucent" is assigned its own positive sentiment.

The result for "Jew" provides a glimpse into how this might be happening with the Natural Language API. In 2006, researchers investigated why the top result when searching for "Jew" on Google was the anti-Semitic site "Jew Watch." The reason? Hate sites tended to call Jews "Jew," while news sites and other resources are more likely to use the word "Jewish."

The problem is that artificial intelligence systems like Google's Natural Language API will inevitably absorb the biases that plague the internet and human society more broadly. "It's easy to get around [the bias] for each individual problem," Ernest Davis, a professor of computer science at New York University told me over the phone, "But getting around it systematically is very difficult."

As Giannandrea, Google's AI head of AI said at his company's conference, "It's important that we be transparent about the training data that we are using, and are looking for hidden biases in it, otherwise we are building biased systems."