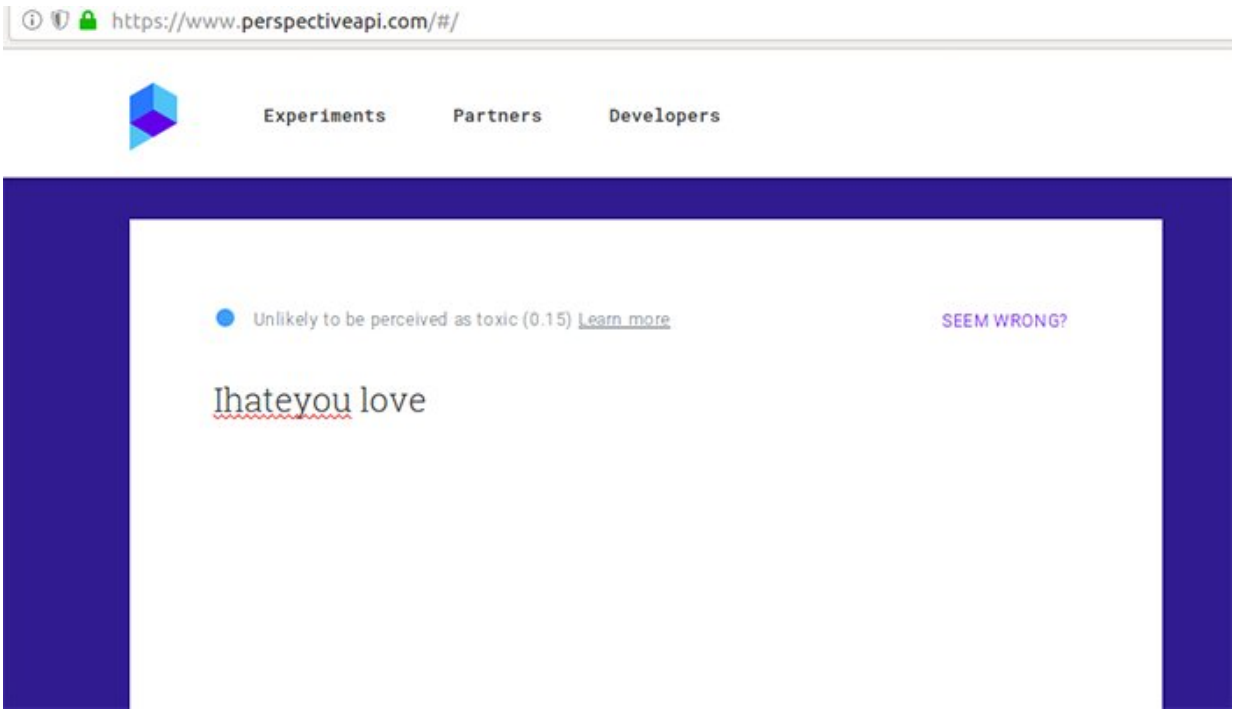
Detectors for online hate speech can be easily duped by humans, study shows

September 14 2018



How Google Perspective rates a comment otherwise deemed toxic after some inserted typos and a little love. Credit: Aalto University

Hateful text and comments are an ever-increasing problem in online environments, yet addressing the rampant issue relies on being able to identify toxic content. A new study by the Aalto University Secure Systems research group has discovered weaknesses in many machine learning detectors currently used to recognize and keep hate speech at

bay.

Many popular social media and online platforms use hate speech detectors that a team of researchers led by Professor N. Asokan have now shown to be brittle and easy to deceive. Bad grammar and awkward spelling—intentional or not—might make toxic social media comments harder for AI detectors to spot.

The team put seven state-of-the-art hate speech detectors to the test. All of them failed.

Modern natural language processing techniques (NLP) can classify text based on individual characters, words or sentences. When faced with textual data that differs from that used in their training, they begin to fumble.

"We inserted typos, changed word boundaries or added neutral words to the original hate speech. Removing spaces between words was the most powerful attack, and a combination of these methods was effective even against Google's comment-ranking system Perspective," says Tommi Gröndahl, doctoral student at Aalto University.

Google Perspective ranks the 'toxicity' of comments using text analysis methods. In 2017, researchers from the University of Washington showed that Google Perspective can be fooled by introducing simple typos. Gröndahl and his colleagues have now found that Perspective has since become resilient to simple typos yet can still be fooled by other modifications such as removing spaces or adding innocuous words like 'love."

A sentence like "I hate you" slipped through the sieve and became non-hateful when modified into "Ihateyou love."

The researchers note that in different contexts the same utterance can be regarded either as hateful or merely offensive. Hate speech is subjective and context-specific, which renders text analysis techniques insufficient as stand-alone solutions.

The researchers recommend that more attention be paid to the quality of data sets used to train machine learning models—rather than refining the model design. The results indicate that character-based detection could be a viable way to improve current applications.

The study was carried out in collaboration with researchers from University of Padua in Italy. The results will be presented at the ACM AISec workshop in October.

The study is part of an ongoing project called "Deception Detection via Text Analysis in the Secure Systems" at Aalto University.

More information: All You Need is "Love": Evading Hate-speech Detection. arxiv.org/abs/1808.09115

Provided by Aalto University