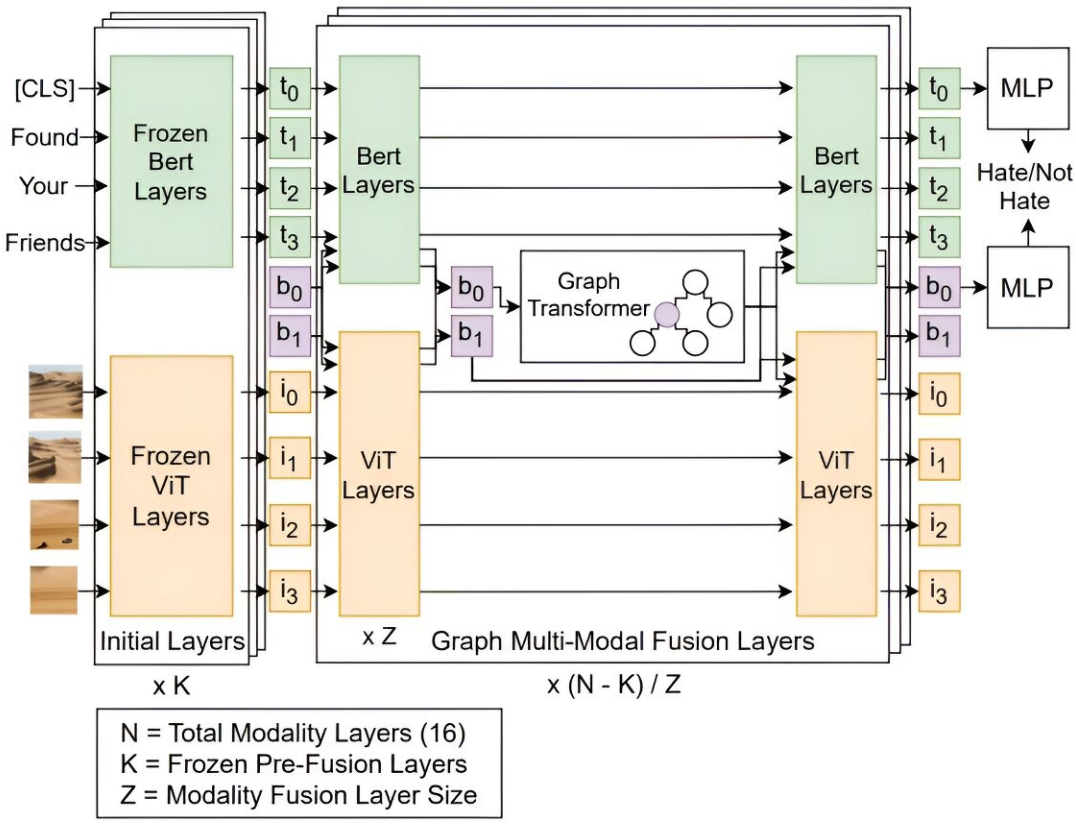
Researchers build AI to save humans from the emotional toll of monitoring hate speech

May 29 2024



Multi-modal discussion transformer. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2307.09312

A team of researchers at the University of Waterloo have developed a

new machine-learning method that detects hate speech on social media platforms with 88% accuracy, saving employees from hundreds of hours of emotionally damaging work.

The method, dubbed the multi-modal discussion transformer (mDT), can understand the relationship between text and images as well as put comments in greater context, unlike previous hate speech detection methods. This is particularly helpful in reducing false positives, which are often incorrectly flagged as hate speech due to culturally sensitive language.

"We really hope this technology can help reduce the emotional cost of having humans sift through hate speech manually," said Liam Hebert, a Waterloo computer science Ph.D. student and the first author of the study. "We believe that by taking a community-centered approach in our applications of AI, we can help create safer online spaces for all."

Researchers have been building models to analyze the meaning of human conversations for many years, but these models have historically struggled to understand nuanced conversations or contextual statements. Previous models have only been able to identify hate speech with as much as 74% accuracy, below what the Waterloo research was able to accomplish.

"Context is very important when understanding hate speech," Hebert said. "For example, the comment 'That's gross!' might be innocuous by itself, but its meaning changes dramatically if it's in response to a photo of pizza with pineapple versus a person from a marginalized group.

"Understanding that distinction is easy for humans, but training a model to understand the contextual connections in a discussion, including considering the images and other multimedia elements within them, is actually a very hard problem."

Unlike previous efforts, the Waterloo team built and trained their model on a dataset consisting not only of isolated hateful comments but also the context for those comments. The model was trained on 8,266 Reddit discussions with 18,359 labeled comments from 850 communities.

"More than three billion people use social media every day," Hebert said. "The impact of these social media platforms has reached unprecedented levels. There's a huge need to detect hate speech on a large scale to build spaces where everyone is respected and safe."

The findings are published on the *arXiv* preprint server.

More information: Liam Hebert et al, Multi-Modal Discussion Transformer: Integrating Text, Images and Graph Transformers to Detect Hate Speech on Social Media, *arXiv* (2023). DOI: 10.48550/arxiv.2307.09312

Provided by University of Waterloo