Can AI encourage good behaviour online?

HOSTILE and hateful remarks are thick on the ground on social networks in spite of persistent efforts by Facebook, Twitter, Reddit and YouTube to tone them down.

Now researchers at the OpenWeb platform have turned to artificial intelligence (AI) to moderate Internet users' comments before they are even posted.

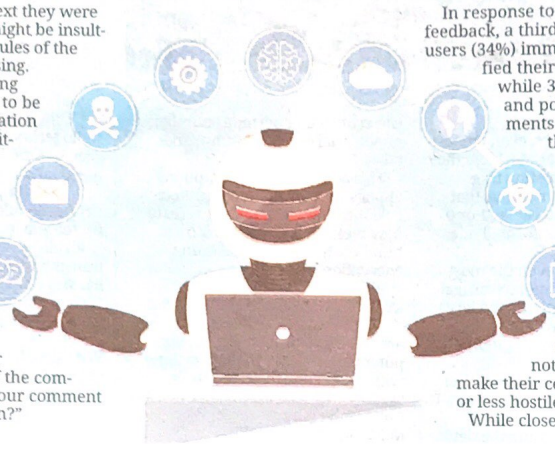
The method appears to be effective because one third of users modified the text of their comments when they received a nudge from the new system, which warned that what they had written might be perceived as offensive.

The study conducted by OpenWeb and Perspective API analysed 400,000 comments that some 50,000 users were preparing to post on sites like AOL, Salon, Newsweek, RT and Sky Sports.

Some of these users received a feedback message or nudge from a machine learning algorithm to the effect that the text they were preparing to post might be insulting, or against the rules of the forum they were using.

Instead of rejecting comments it found to be suspect, the moderation algorithm then invited their authors to reformulate what they had written.

"Let's keep the conversation civil. Please remove any inappropriate language from your comment," was a message prompt or "Some members of the community may find your comment offensive. Try again?"



In response to this kind of feedback, a third of Internet users (34%) immediately modified their comments, while 36% went ahead and posted their comments anyway, taking the risk that they might be rejected by the moderating algorithm.

Even more surprisingly, some users made modifications that did not necessarily make their comments kinder or less hostile.

While close to 30% of users opted to accept the feedback message and delete potentially offensive text from their comments, more than a quarter (25.8%) attempted to dupe the moderating algorithm.

Deliberate spelling errors and adding spaces between letters were just two of the tricks they used to modify the form of their comments while leaving their content unchanged.

The 400,000 comments analysed in the study are, however, a mere drop in the ocean when compared to the millions that are posted daily on the Internet, some of which carry offensive and insulting language.

Faced with this situation, tech giants are boosting their efforts to combat online hate more effectively.

It is a fight in which AI can make a useful but, for now at least, imperfect contribution. – AFP Relaxnews