Tricking fake news detectors with malicious user comments

2 November 2020, by Jordan Ford

Credit: Pixabay/CC0 Public Domain

Fake news detectors, which have been deployed by social media platforms like Twitter and Facebook to add warnings to misleading posts, have traditionally flagged online articles as false based on the story's headline or content. However, recent approaches have considered other signals, such as network features and user engagements, in addition to the story's content to boost their accuracies.

However, new research from a team at Penn State's College of Information Sciences and Technology shows how these fake news detectors can be manipulated through user comments to flag true news as false and false news as true. This attack approach could give adversaries the ability to influence the detector's assessment of the story even if they are not the story's original author.

"Our model does not require the adversaries to modify the target article's title or content," explained Thai Le, lead author of the paper and doctoral student in the College of IST. "Instead, adversaries can easily use random accounts on social media to post malicious comments to either demote a real story as fake news or promote a fake story as real news."

That is, instead of fooling the detector by attacking the story's content or source, commenters can attack the detector itself.

The researchers developed a framework—called Malcom—to generate, optimize, and add malicious comments that were readable and relevant to the article in an effort to fool the detector. Then, they assessed the quality of the artificially generated comments by seeing if humans could differentiate them from those generated by real users. Finally, they tested Malcom's performance on several popular fake news detectors.

Malcom performed better than the baseline for existing models by fooling five of the leading neural network based fake news detectors more than 93% of the time. To the researchers' knowledge, this is the first model to attack fake news detectors using this method.

This approach could be appealing to attackers because they do not need to follow traditional steps of spreading fake news, which primarily involves owning the content. The researchers hope their work will help those charged with creating fake news detectors to develop more robust models and strengthen methods to detect and filter-out malicious comments, ultimately helping readers get accurate information to make informed decisions.

"Fake news has been promoted with deliberate intention to widen political divides, to undermine citizens' confidence in public figures, and even to create confusion and doubts among communities," the team wrote in their paper, which will be presented virtually during the 2020 IEEE International Conference on Data Mining.

Added Le, "Our research illustrates that attackers can exploit this dependency on users' engagement

to fool the detection models by posting malicious comments on online articles, and it highlights the importance of having robust fake news detection models that can defend against adversarial attacks."

More information: Le et al., MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models. (2020). pike.psu.edu/publications/icdm20.pdf

Provided by Pennsylvania State University