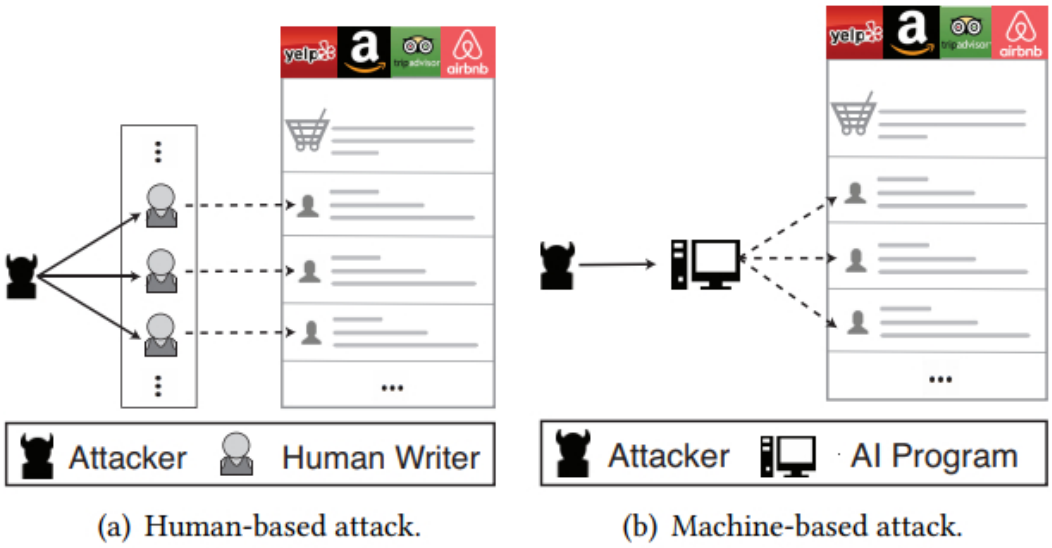
Study reveals credibility muscle in machine-generated reviews

September 3 2017, by Nancy Owano



(a) Human-based attack. (b) Machine-based attack.

Fake review attack: Human-based vs. Machinebased. Credit: arXiv:1708.08151 [cs.CR]

(Tech Xplore)—*Cooked to perfection. Service was amazing. The chicken is very good.* Before you grab your jacket and car keys to head for the restaurant, know this. The praise could have been machine-generated. Translation: The fake comments could care less whether your carrots were barely cooked, waiter rude and chicken bland.

A University of Chicago team trained a neural network to write fake

reviews. Their paper has been making news because it was quite hard to tell between which were from their system and which were real.

The paper was accepted to the ACM Conference on Computer and Communications Security in October.

"Automated Crowdturfing Attacks and Defenses in Online Review Systems" is on arXiv, and the five authors are Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng and Ben Zhao.

Testing methods? The test involved 40 restaurants. The team asked people to mark reviews as fake or real.

Business Insider noted that their software could write "extremely believable" fake online reviews.

For reviews marked real, they asked for a rating of the review's usefulness. The people thought the machine-generated reviews almost as useful as real reviews.

Phoebe Weston, *Daily Mail*, said, "The neural networks were trained using a deep learning technique called recurrent neural networks (RNN). The network learnt by reading through thousands of real online reviews."

The authors wrote that "RNNs can learn from a large corpus of natural language text (character or word sequences), to generate text at different levels of granularity, i.e. at the character level or word level."

Business Insider said those real reviews used were freely available online.

Using Yelp reviews as an example platform, the authors showed how their approach could produce reviews indistinguishable by state-of-the-art statistical detectors. However, their study attention regarded not Yelp

in isolation but the wider arena of crowdsourced feedback.

The authors wrote, "Most popular e-commerce sites today rely on user feedback to rate products, services, and online content. Crowdsourced feedback typically includes a review or opinion, describing a user's experience of a product or service, along with a rating score."

The authors further considered countermeasures against their mechanisms.

Responding to the study findings, a number of tech watchers expressed concern over a future where such reviews might cloud the picture for startups seeking good reputations and the public seeking trust in reading opinions by humans, not machines.

Ben Zhao, a professor of computer science at the University of Chicago, has concerns that go beyond fake reviews, though.

Quoted in *Business Insider*. "So we're starting with online reviews. Can you trust what so-and-so said about a restaurant or product? But it is going to progress... where entire articles written on a blog may be completely autonomously generated along some theme by a robot ... that I think is going to be a much bigger challenge for all of us in the years ahead."

Business Insider carried an emailed statement from Yelp. Spokesperson Rachel Youngblade said that Yelp "appreciate[s] this study shining a spotlight on the large challenge review sites like Yelp face in protecting the integrity of our content, as attempts to game the system are continuing to evolve and get ever more sophisticated. Yelp has had systems in place to protect our content for more than a decade, but this is why we continue to iterate those systems to catch not only fake reviews, but also biased and unhelpful content. We appreciate the authors of this

study using Yelp's system as 'ground truth' and acknowledging its effectiveness."

Rob Price senior reporter, *Business Insider*, wrote, "Zhao said he hasn't seen any examples of AI being used to generate malicious fake reviews in the real world just yet."

More information: Automated Crowdturfing Attacks and Defenses in Online Review Systems, arXiv:1708.08151 [cs.CR] arxiv.org/abs/1708.08151

Abstract
Malicious crowdsourcing forums are gaining traction as sources of spreading misinformation online, but are limited by the costs of hiring and managing human workers. In this paper, we identify a new class of attacks that leverage deep learning language models (Recurrent Neural Networks or RNNs) to automate the generation of fake online reviews for products and services. Not only are these attacks cheap and therefore more scalable, but they can control rate of content output to eliminate the signature burstiness that makes crowdsourced campaigns easy to detect.
Using Yelp reviews as an example platform, we show how a two phased review generation and customization attack can produce reviews that are indistinguishable by state-of-the-art statistical detectors. We conduct a survey-based user study to show these reviews not only evade human detection, but also score high on "usefulness" metrics by users. Finally, we develop novel automated defenses against these attacks, by leveraging the lossy transformation introduced by the RNN training and generation cycle. We consider countermeasures against our mechanisms, show that they produce unattractive cost-benefit tradeoffs for attackers, and that they can be further curtailed by simple constraints imposed by online service providers.

© 2017 Tech Xplore

Citation: Study reveals credibility muscle in machine-generated reviews (2017, September 3) retrieved 3 May 2024 from
https://techxplore.com/news/2017-09-reveals-credibility-muscle-machine-generated.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.