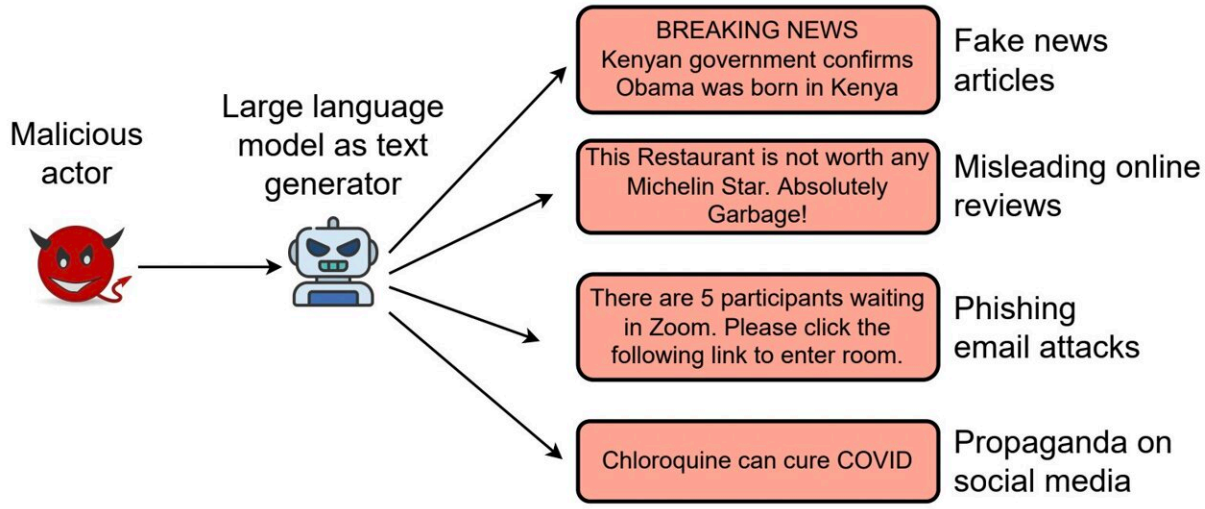
The strengths and limitations of approaches to detect deepfake text

November 21 2022, by Ingrid Fadelli



Credit: Pu et al

Advances in the field of machine learning have recently enabled the development of computational tools that can create convincing but artificially produced texts, also known as deepfake texts. While the automatic creation of texts could have some interesting applications, it also raises serious concerns in terms of security and misinformation.

Synthetically produced texts could ultimately also be used to mislead internet users, for instance through the large-scale generation of extremists or violent texts aimed at radicalizing individuals, fake news

for disinformation campaigns, email texts for phishing attacks, or fake reviews targeting specific hotels, venues or restaurants. Collectively, this could further reduce some users' trust in online content, while prompting other users to engage in anti-social and risky behavior.

A recent study led by researchers at Virginia Tech, in collaboration with researchers at University of Chicago, LUMS Pakistan and University of Virginia recently explored the limitations and strengths of existing approaches for detecting deepfake texts. Their paper, with students Jiameng Pu and Zain Sarwar as lead authors, is set to be presented at IEEE S&P'23, a conference focusing on computer security.

"Much of the security research we conducted prior to 2016 assumed an algorithmically weak attacker. This assumption is no longer valid given the advances made in AI and ML. We have to consider algorithmically intelligent or ML-powered adversaries. This prompted us to start exploring this space. In 2017, we published a paper exploring how language models (LMs) like RNNs can be misused to generate fake reviews on platforms such as Yelp," Bimal Viswanath, researcher from Virginia Tech who led the study, told TechXplore.

"This was our first foray into this space. Since then, we observed rapid advances in LM technologies, especially after the release of the Transformer family of models. These advances raise the threat of misuse of such tools to enable large-scale campaigns to spread disinformation, generate opinion spam and abusive content, and more effective phishing techniques."

Over the past few years, many computer scientists worldwide have been trying to develop computational models that can accurately detect synthetic text generated by advanced LMs. This led to the introduction of numerous different defensive strategies; including some that seek out specific artifacts in synthetic texts and others that rely on the use of pre-

trained language models to build detectors.

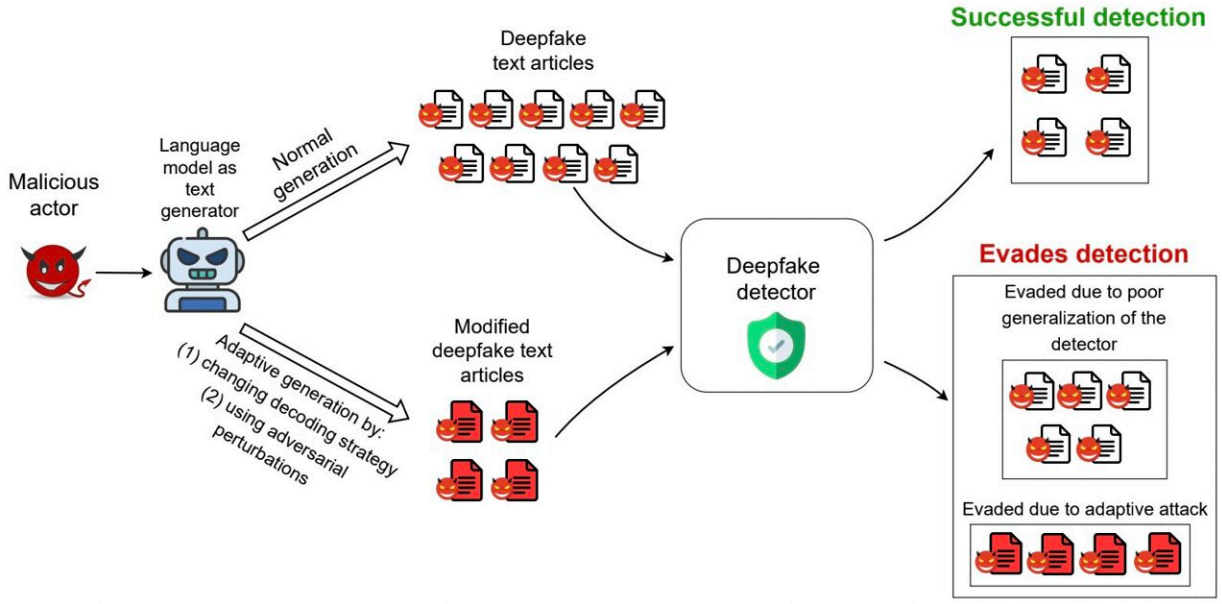
"While these defenses reported high detection accuracies, it was still unclear how well they would work in practice, under adversarial settings," Viswanath explained. "Existing defenses were tested on datasets created by researchers themselves, rather than on synthetic data in the wild. In practice, attackers would adapt to these defenses to evade detection, and existing works did not consider such adversarial settings."

Defenses that malicious users can easily overcome by slightly changing their language models' design are ultimately ineffective in the real-world. Viswanath and his colleagues thus set out to explore the limitations, strengths and real-world value of some of the most promising deepfake text detection models created so far.

Their paper focused on 6 existing synthetic text-detection schemes introduced over the past few years, all of which had attained remarkable performances in initial evaluations, with detection accuracies ranging from 79.6% to 98.5%. The models they evaluated are BERT-Defense, GLTR-GPT2, GLTR-BERT, GROVER, FAST and RoBERTa-Defense.

"We thank the developers of these models for sharing code and data with us, as this allowed us to accurately reproduce them," Viswanath said. "Our first goal was to reliably evaluate the performance of these defenses on real-world datasets. To do this, we prepared 4 novel synthetic datasets, which we now released to the community."

To compile their datasets, Viswanath and his colleagues collected thousands of synthetic text articles created by different text-generation-as-a-service platforms, as well deepfake Reddit posts created by bots. Text-generation-as-a-service platforms are AI-powered internet sites that allow users to simply create synthetic text and which can be misused to create misleading content.



Credit: Pu et al

To reliably assess the performance of the six defense models they selected in detecting deepfake texts, the researchers proposed a series of "low-cost" evasion strategies that only require changes to the LM-based text generator at inference time. This basically means that the LM generating the fake text can be adapted or improved during trials, without the need for additional training.

"We also proposed a novel evasion strategy, called DFTFooler, that can automatically perturb or modify any synthetic text article to evade detection, while preserving the semantics," Viswanath said. "DFTFooler uses publicly available LMs and leverages insights unique to the synthetic text detection problem. Unlike other adversarial perturbation schemes, DFTFooler does not require any query access to the victim defense classifier to create evasive samples, thereby making it a more

stealthy and practical attack tool."

The team's evaluations yielded several interesting results. Firstly, the researchers found that the performance of three out of the six defense models they assessed significantly declined when they were tested on real-world datasets, with 18% to 99% drops in their accuracy. This highlights the need to improve these models to ensure that they generalize well across different data.

In addition, Viswanath and his colleagues found that changing a LM's text decoding (i.e., text sampling) strategy often broke many of the defenses. This simple strategy does not require any additional model re-training, as it only modifies a LM's existing text generation parameters and is thus very easy for attackers to enforce.

"We also find that our new adversarial text manipulation strategy called DFTFooler can successfully create evasive samples without requiring any queries to the defender's classifier," Viswanath said. "Among the six defenses we evaluated, we find that one defense called FAST is most resilient in these adversarial settings, compared to the other defenses. Unfortunately, FAST has a complex pipeline that uses multiple advanced NLP techniques, thereby making it harder to understand its better performance."

To gain more insight into the qualities that make the FAST model particularly resilient and reliable in detecting deepfake texts, the researchers conducted an in-depth analysis of its features. They found that the model's resilience is due to its use of semantic features extracted from the articles.

In contrast with the other defense models evaluated in this study, FAST analyzes a text's semantic features, looking at named entities and relations between these entities in the text. This unique quality appeared

to significantly improve the model's performance on real-world deepfake datasets.

Inspired by these findings, Viswanath and his colleagues created DistilFAST, a simplified version of FAST that only analyzes semantic features. They found that this model outperformed the original FAST model under adversarial settings.

"Our work highlights the potential for semantic features to enable adversarially-robust synthetic detection schemes," Viswanath said. "While FAST shows promise, there is still significant room for improvement. Generating semantically consistent, long text articles is still a challenging problem for LMs. Therefore, differences in the representation of semantic information in synthetic and real articles can be exploited to build robust defenses."

When trying to circumvent deepfake text detectors, attackers might not always be able to change the semantic content of synthetic texts, particularly when these texts are designed to convey specific ideas. In the future, the findings gathered by this team of researchers and the simplified FAST model they created could thus help to strengthen defenses against synthetic texts online, potentially limiting large-scale disinformation or radicalization campaigns.

"Currently, this direction has not been investigated in the security community," Viswanath added. "In our future work, we plan to leverage knowledge graphs to extract richer semantic features, hopefully producing more performant and robust defenses."

More information: Jiameng Pu et al, Deepfake Text Detection: Limitations and Opportunities, *arXiv* (2022). DOI: 10.48550/arxiv.2210.09421

Yuanshun Yao et al, Automated Crowdturfing Attacks and Defenses in Online Review Systems, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017). DOI: 10.1145/3133956.3133990