PETGEN: Personalized Text Generation Attack on Deep Sequence Embedding-based Classification Models

Bing He, Mustaque Ahamad, Srijan Kumar
Georgia Institute of Technology
Atlanta, Georgia, USA
bhe46@gatech.edu,mustaq@cc.gatech.edu,srijan@gatech.edu

ABSTRACT

What should a malicious user write next to fool a detection model? Identifying malicious users is critical to ensure the safety and integrity of internet platforms. Several deep learning based detection models have been created. However, malicious users can evade deep detection models by manipulating their behavior, rendering these models of little use. The vulnerability of such deep detection models against adversarial attacks is unknown. Here we create a novel adversarial attack model against deep user sequence embedding-based classification models, which use the sequence of user posts to generate user embeddings and detect malicious users. In the attack, the adversary generates a new post to fool the classifier. We propose a novel end-to-end Personalized Text Generation Attack model, called PETGEN, that simultaneously reduces the efficacy of the detection model and generates posts that have several key desirable properties. Specifically, PETGEN generates posts that are personalized to the user’s writing style, have knowledge about a given target context, are aware of the user’s historical posts on the target context, and encapsulate the user’s recent topical interests. We conduct extensive experiments on two real-world datasets (Yelp and Wikipedia, both with ground-truth of malicious users) to show that PETGEN significantly reduces the performance of popular deep user sequence embedding-based classification models. PETGEN outperforms five attack baselines in terms of text quality and attack efficacy in both white-box and black-box classifier settings. Overall, this work paves the path towards the next generation of adversary-aware sequence classification models.

CCS CONCEPTS

• **Computing methodologies** → Anomaly detection.

KEYWORDS

Adversarial Text Generation; Sequence Classification; User Classification; Attack; Deep Learning

ACM Reference Format:
Bing He, Mustaque Ahamad, Srijan Kumar. 2021. PETGEN: Personalized Text Generation Attack on Deep Sequence Embedding-based Classification Models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467390

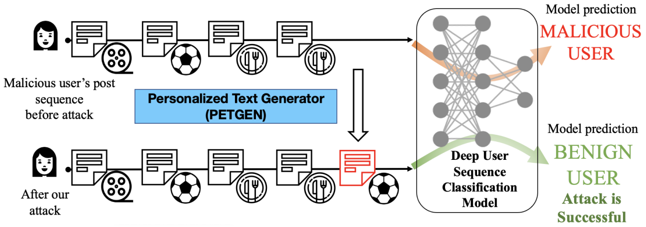
Malicious user’s post sequence before attack
Personalized Text Generator (PETGEN)
After our attack
Deep User Sequence Classification Model
Model prediction
MALICIOUS USER
Model prediction
BENIGN USER
Attack is Successful

Figure 1: Deep user sequence embedding-based classification models are used to detect malicious users (top row). However, an evasion attack by an adversary by creating a new fake post can lead the same model to misclassify it as a benign user (bottom row). Our method, PETGEN, generates personalized text posts to adversarially attack the classifier.

1 INTRODUCTION

As Web platforms, such as e-commerce, social media, and crowdsourcing platforms, have gained popularity, they are increasingly targeted by malicious actors for their gains [10, 11, 20]. The proliferation of undesirable users, such as fake accounts [20], spammers [3, 23], fake news spreaders [15, 26], abnormal users [1], vandal editors [12], fraudsters [11], and sockpuppets [10], poses a threat to the safety and integrity of online communities. To give an example, on Facebook, roughly 5% of monthly active users in 2019 were fake accounts [20]. Similarly, on Amazon, 63% reviews on beauty products were from fraudulent users [2]. Thus, the identification of malicious accounts is a critical task for all web and social media platforms.

Deep user sequence embedding-based classification models are increasingly gaining popularity for platform integrity tasks, including the TIES model at Facebook [20]. These models train a deep learning model to generate user embeddings by utilizing the temporal sequence of actions and post content of a user. The user embedding is then used to make predictions about the user. For example, Figure 1 shows a deep user sequence embedding-based classification model trained to identify malicious users from the user’s sequence of posts (top row).

However, deep learning models can be vulnerable to adversarial attacks [21]. While adversarial attacks on deep learning models have received a lot of attention in graph representation learning, natural language processing, and computer vision domains [21], the vulnerability of deep user sequence embedding-based classification models remains unknown. For example, in Figure 1, the malicious user can create a new post, so that the entire user sequence is misclassified as benign by the classifier (bottom row). Thus, identifying

KDD ’21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467390