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

Bing He, Mustaque Ahamad, Srijan Kumar
Georgia Institute of Technology
Email: bhe46@gatech.edu, srijan@gatech.edu

All code and data at:
https://github.com/srijankr/petgen
Malicious Users on Social Media

• A critical task for social media platforms to ensure safety and integrity
  – ~5% monthly active users are fake accounts in Facebook
  – ~63% reviews on Amazon beauty are fake
  – Other types of malicious users: fraudsters, trolls, spammers, cyber-bullies
Deep Learning Solutions

- Deep learning methods have been created to detect malicious users
- **Many solutions use user activity sequences** for detection
  - TIES (Facebook)
  - JODIE
  - HRNN

Figure reference: TIES paper
Adversaries are Active

• Malicious users can change their behavior to **avoid detection**
• Prior deep learning models, from computer vision and NLP domains, have been shown to be vulnerable
• **Vulnerability** of deep user sequence embedding models is unknown
Key Question

Can malicious users avoid detection by exploiting model vulnerabilities?
Key Question

Can malicious users avoid detection by exploiting model vulnerabilities?

Our Solution: Adversarial evasion attack on deep user sequence classification models
Our Attack: Next Post Attack

Malicious user’s post sequence before attack

Model prediction
MALICIOUS USER

Deep User Sequence Classification Model
Adversary generates a new post, such that the user classification changes.
Desirable Properties of Attack Post

What are the desirable properties of the attack post?

1. Should **fool the classification model**
2. Should be knowledgeable about the **target context**
3. Should be **realistic and personalized**
   - Aware of user’s writing style
   - Recent vs past interests
   - Aware of user’s past posts on similar topics
## Existing Methods

<table>
<thead>
<tr>
<th>Modification-based attack</th>
<th>C1 Attack goal</th>
<th>C2 Target context</th>
<th>C3 Personalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copycat</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Hotflip</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal Adversarial Trigger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextBugger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generation-based attack</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Malcom</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Our model: PETGEN

<table>
<thead>
<tr>
<th>C1 Attack goal</th>
<th>C2 Target context</th>
<th>C3 Personalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
• Personalized Text Generator
• End-to-end multi-stage multi-task text generation framework
• Two major modules:

PETGEN

User sequence + Target Context

Sequence-aware Text Generator

Intermediate post

Multi-task Tuning

Generated Attack Post
Personalized Text Generator: PETGEN

- **Posts**
  - Contextual post relevance
  - Target context relevance
  - Recent Post Relevance

- **Contexts**
  - Deep User Sequence Embedding Model
  - Generated Attack Text

- **PETGEN**
  - Sequence-aware Text Generator
  - Multi-task Tuning

- **Attack**
  - Style

- **Target Context**
Personalized Text Generator: PETGEN

- Posts
- Contexts
- Recent Post Relevance
- Contextual post relevance
- Target context relevance
- Target Context

PETGEN

Sequence-aware Text Generator

Multi-task Tuning

Generated Attack Text

Attack

Style

Deep User Sequence Embedding Model
Sequence-Aware Text Generator

Capture contextual relevance from previous posts
Sequence-Aware Text Generator

Capture contextual relevance from previous posts

Capture user sequence embedding
Sequence-Aware Text Generator

Capture contextual relevance from previous posts

Capture user sequence embedding
Sequence-Aware Text Generator

Capture contextual relevance from previous posts

Text generator module

Token #i

Text Generator (RMRN)

Token #(i+1)

Text Generator (RMRN)

Token #(i+2)

Context-biased User Sequence Embedding

Context-aware Attention Vector

Sequence Embedding

Attention Score Computation

Capture user sequence embedding
Sequence-Aware Text Generator

Capture contextual relevance from previous posts

Capture user sequence embedding
Personalized Text Generator: PETGEN

Posts → Contexts → Posts

Attack
Style

Recent Post Relevance

PETGEN

Sequence-aware Text Generator

Initial post

Multi-task Tuning

Deep User Sequence Embedding Model

Generated Attack Text

Target context relevance

Target Context

Contextual post relevance

Recent Post Relevance

Contexts → Posts

Initial post
Multi-Task Tuning

Four objectives:

• **Style**: Relativistic GAN loss
• **Attack**: Cross-entropy loss
• **Recent Post Relevance**: Maximum Mean Discrepancy (MMD) Loss
• **Target Context Relevance**: MMD Loss

Optimization strategy:

• Multi-stage loss optimization. One loss is optimized at a time
• Done till convergence.
Personalized Text Generator: PETGEN

Diagram showing the process of generating personalized text with PETGEN, including inputs like posts and contexts, and outputs like generated text.

- Posts and contexts flow into the PETGEN model.
- Recent post relevance and contextual post relevance are inputs to the PETGEN model.
- Target context relevance is another input.
- The model outputs generated attack text.

Deep User Sequence Embedding Model and Multi-task Tuning are components of the PETGEN system.
Evaluation Setup

• Deep user sequence classification model
  – TIES model [1]
  – Hierarchical Recurrent Neural Network (HRNN) [2]

• Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Yelp</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>3,940</td>
<td>794</td>
</tr>
<tr>
<td>Number of benign users</td>
<td>2,016</td>
<td>397</td>
</tr>
<tr>
<td>Number of malicious users</td>
<td>1,924</td>
<td>397</td>
</tr>
<tr>
<td>Total number of posts</td>
<td>35,123</td>
<td>11,547</td>
</tr>
<tr>
<td>Median posts per user</td>
<td>9</td>
<td>15</td>
</tr>
</tbody>
</table>

Code and data are available at: https://github.com/srijankr/petgen

Baseline Attacks

- **Copycat**: copy user’s past post on similar context
- **HotFlip**: Copycat + replace most important word with similar word
- **UniTrigger**: Copycat + add tokens to the end of the post
- **TextBugger**: Copycat + deletion/swap of characters
- **Malcom**: state-of-the-art model

No baseline is sequence-aware
Attack on the **TIES model** on Yelp data

- **Model performance reduces** against all attacks.
- **PETGEN** is the most successful attack.
## White-Box Attack Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>HRNN classifier</th>
<th>Min. improvement of PETGEN over baseline</th>
<th>TIES classifier</th>
<th>Min. improvement of PETGEN over baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wikipedia</td>
<td>Yelp</td>
<td></td>
<td>Wikipedia</td>
</tr>
<tr>
<td></td>
<td>F1↓  Atk↑</td>
<td>F1↓  Atk↑</td>
<td>F1</td>
<td>F1↓  Atk↑</td>
</tr>
<tr>
<td>Without attack</td>
<td>0.601</td>
<td>0.610</td>
<td>9.836% 26.761%</td>
<td>0.617</td>
</tr>
<tr>
<td>Copycat</td>
<td>0.550 21.3</td>
<td>0.610 8.0</td>
<td>6.937% 27.358%</td>
<td>0.514 15.0</td>
</tr>
<tr>
<td>Hotflip</td>
<td>0.581 21.2</td>
<td>0.591 9.5</td>
<td>4.242% 10.204%</td>
<td>0.515 15.7</td>
</tr>
<tr>
<td>UniTrigger</td>
<td>0.495 24.5</td>
<td>0.602 7.8</td>
<td>9.836% 26.168%</td>
<td>0.520 16.3</td>
</tr>
<tr>
<td>TextBugger</td>
<td>0.550 21.4</td>
<td>0.610 8.3</td>
<td>1.044% 5.882%</td>
<td>0.560 18.0</td>
</tr>
<tr>
<td>Malcom</td>
<td>0.479 25.5</td>
<td>0.570 18.0</td>
<td></td>
<td>0.478 24.0</td>
</tr>
<tr>
<td>PETGEN (proposed)</td>
<td>0.474 27.0</td>
<td>0.55 21.2</td>
<td>-</td>
<td>0.478 24.0</td>
</tr>
</tbody>
</table>

- **Model performance reduces** against all attacks
- **PETGEN is the best attack**
• HRNN surrogate model is trained on the observed outputs of the TIES black-box model.
• Black-box attacks are also **successful**. Attack performance lower than white-box.
• **PETGEN** is the most successful attack.
Black-Box Attack Performance

A HRNN surrogate model is trained on observed outputs of the original black-box model.

Black-box attacks are also **successful**. Attack performance lower than white-box.

**PETGEN is the most successful attack.**
Generated Text Quality

• How realistic is the generated text?

PETGEN has the best text generation quality
Two human raters were shown a pair of texts generated by Malcom and PETGEN – Text generated for the same setting – 50 pairs

Task: which text is more realistic?

Inter-rater agreement = 0.66

PETGEN texts are more realistic 60% of the times.
### Ablation Study

- All components of PETGEN contribute to the performance
- PETGEN with all components is the best or second best in most cases

<table>
<thead>
<tr>
<th>Model</th>
<th>Wikipedia Dataset</th>
<th></th>
<th>Yelp Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1↓</td>
<td>Atk↑</td>
<td>BLEU↑</td>
<td>TCS↑</td>
</tr>
<tr>
<td>PETGEN Base Text Generator</td>
<td>0.479</td>
<td>26.5</td>
<td><strong>0.899</strong></td>
<td>0.375</td>
</tr>
<tr>
<td>w/ Style</td>
<td>0.576</td>
<td>21.1</td>
<td>0.895</td>
<td>0.390</td>
</tr>
<tr>
<td>w/ Attack against TIES</td>
<td>0.478</td>
<td>25.0</td>
<td>0.894</td>
<td>0.368</td>
</tr>
<tr>
<td>w/ Attack against HRNN</td>
<td><strong>0.465</strong></td>
<td><strong>27.5</strong></td>
<td>0.895</td>
<td>0.388</td>
</tr>
<tr>
<td>w/ Recent Post Relevance</td>
<td>0.486</td>
<td>23.8</td>
<td>0.887</td>
<td><strong>0.463</strong></td>
</tr>
<tr>
<td>w/ Target Context Relevance</td>
<td>0.483</td>
<td>23.9</td>
<td>0.887</td>
<td>0.459</td>
</tr>
<tr>
<td>w/ Contextual Post Relevance</td>
<td>0.566</td>
<td>21.2</td>
<td>0.705</td>
<td>0.397</td>
</tr>
<tr>
<td>PETGEN against HRNN</td>
<td>0.474</td>
<td>27.0</td>
<td>0.893</td>
<td>0.463</td>
</tr>
<tr>
<td>PETGEN against TIES</td>
<td>0.478</td>
<td>24.0</td>
<td>0.896</td>
<td><strong>0.474</strong></td>
</tr>
</tbody>
</table>

**Notation:** Bleu score (BLEU), Target Context Similarity (TCS), Recent Post Similarity (RS), Contextual Post Similarity (CPS)
Conclusions

• PETGEN is the first attack framework against user sequence classification models

• Models are vulnerable against attacks

• PETGEN is the most effective attack and generates reasonable text

• Generated attacks can be used to create more robust models

All code and data at: http://claws.cc.cc.gatech.edu/petgen
Postdoc Opening

- Join us at Georgia Tech!
- One postdoc position to work in recommendation systems and/or graphs
- Contact me: srijan@gatech.edu or say hello during KDD