Georgia Tech

KDD2021

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

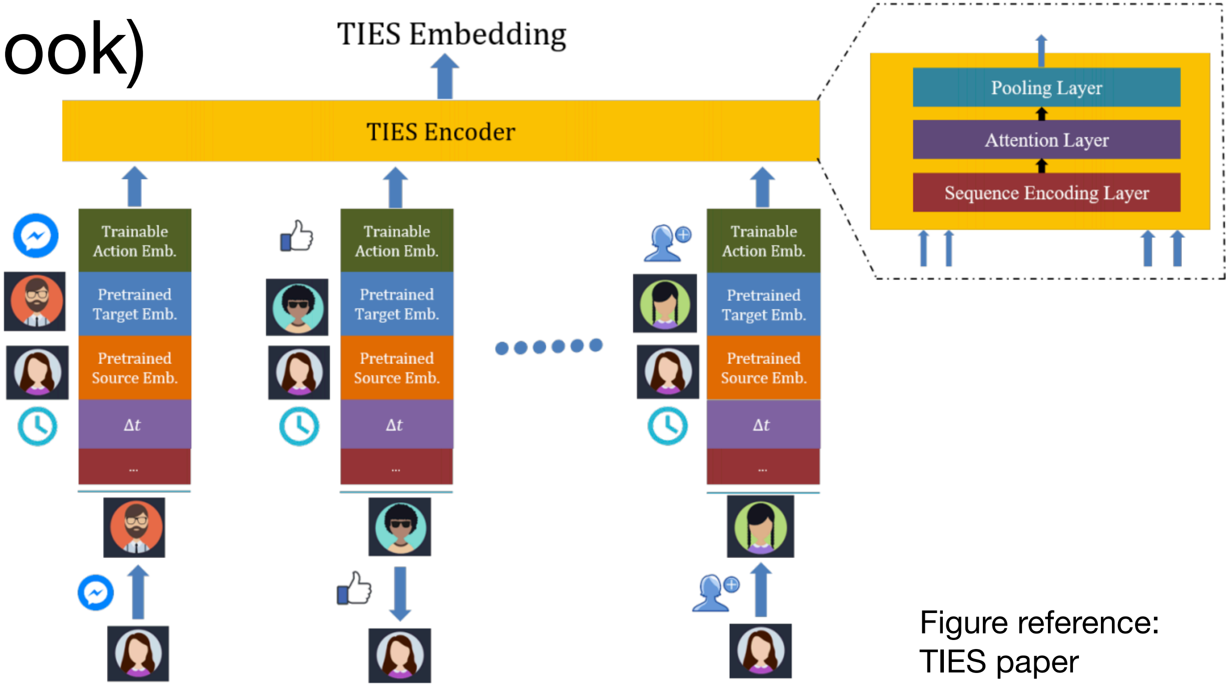


Figure reference: TIES paper

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

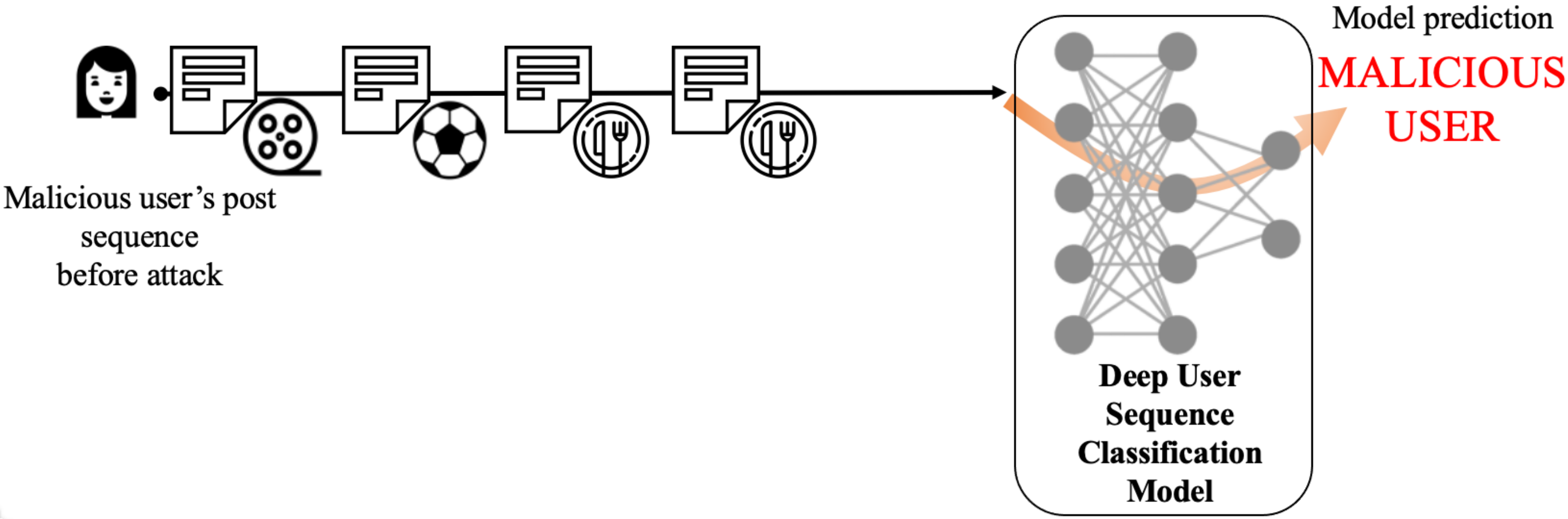
Can malicious users avoid detection by exploiting model vulnerabilities?

Key Question

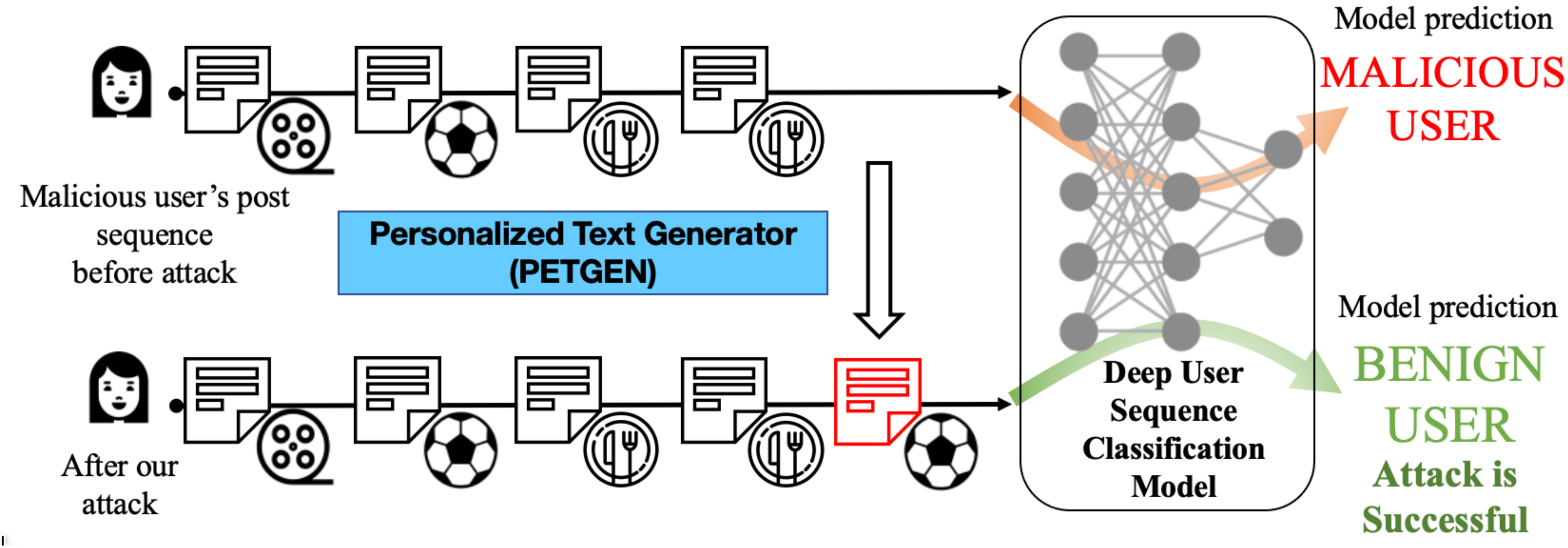
Can malicious users avoid detection by exploiting model vulnerabilities?

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

Our Attack: Next Post Attack


Model prediction
MALICIOUS USER
Malicious user's post sequence before attack
Deep User Sequence Classification Model

Our Attack: Next Post Attack


Model prediction
MALICIOUS USER
Malicious user's post sequence before attack
Personalized Text Generator (PETGEN)
Model prediction
BENIGN USER
Attack is Successful
After our attack
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?

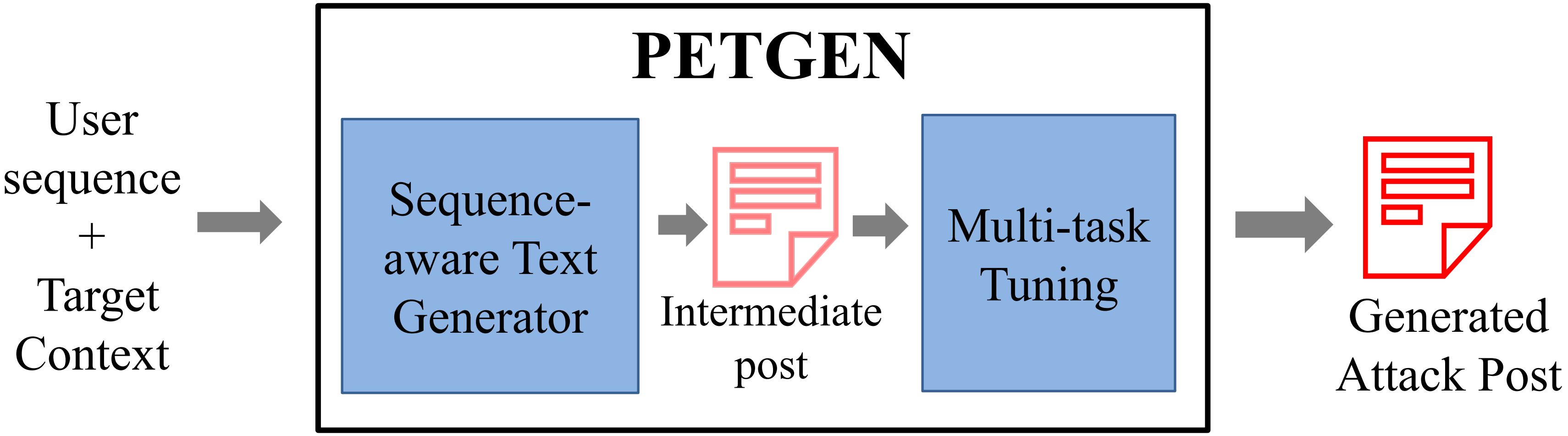
1. Should **fool the classification model**
2. Should be knowledgeable about the **target context**
3. Should be **realistic and personalized**
   - Aware of user’s writing style
   - Recent vs past interests
   - Aware of user’s past posts on similar topics

Existing Methods

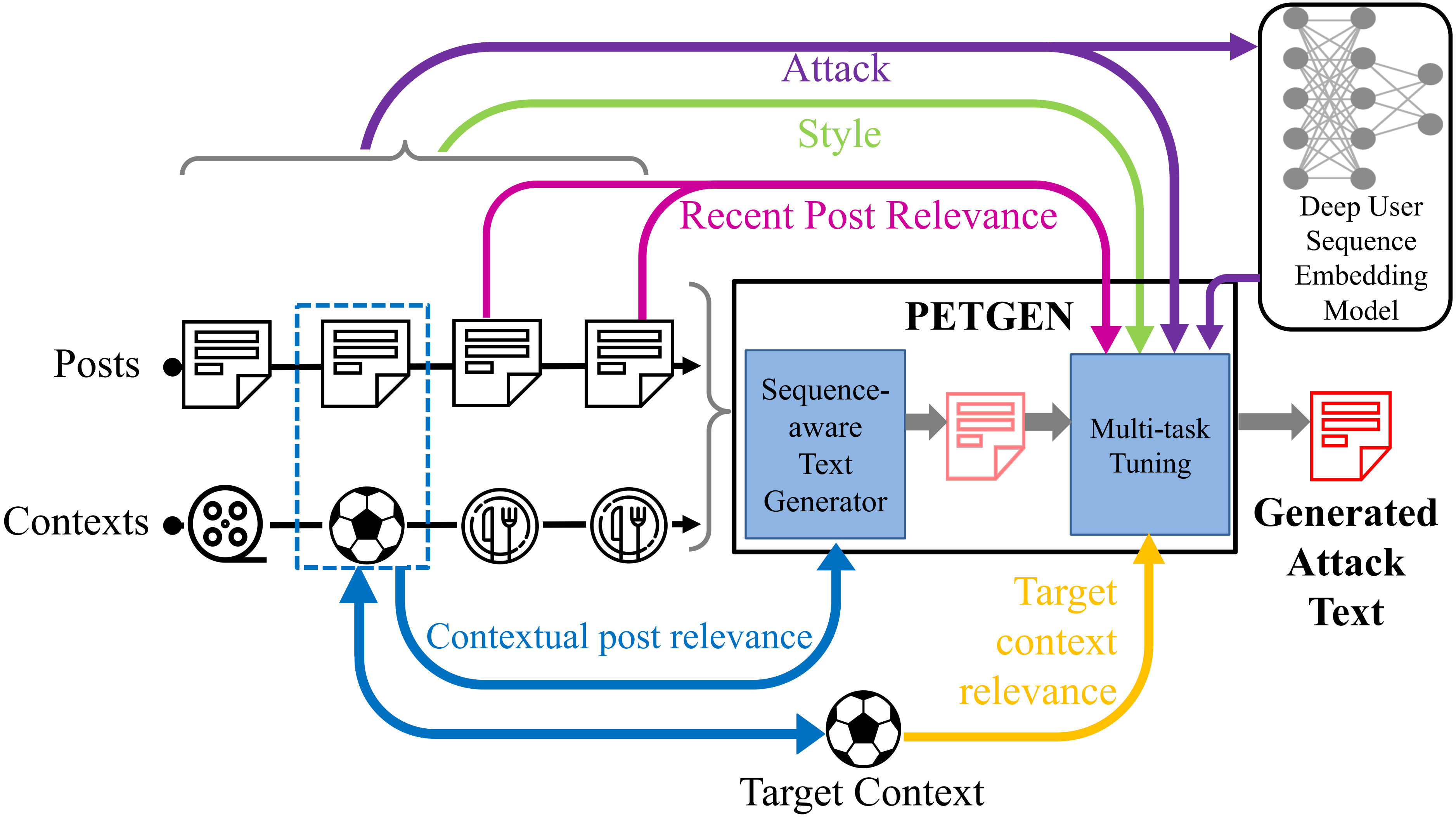
	C1 Attack goal	**C2** Target context	**C3** Personalized
Modification-based attack • Copycat • Hotflip • Universal Adversarial Trigger • TextBugger	✓		
Generation-based attack • Malcom	✓	✓	
Our model: PETGEN	✓	✓	✓

PETGEN

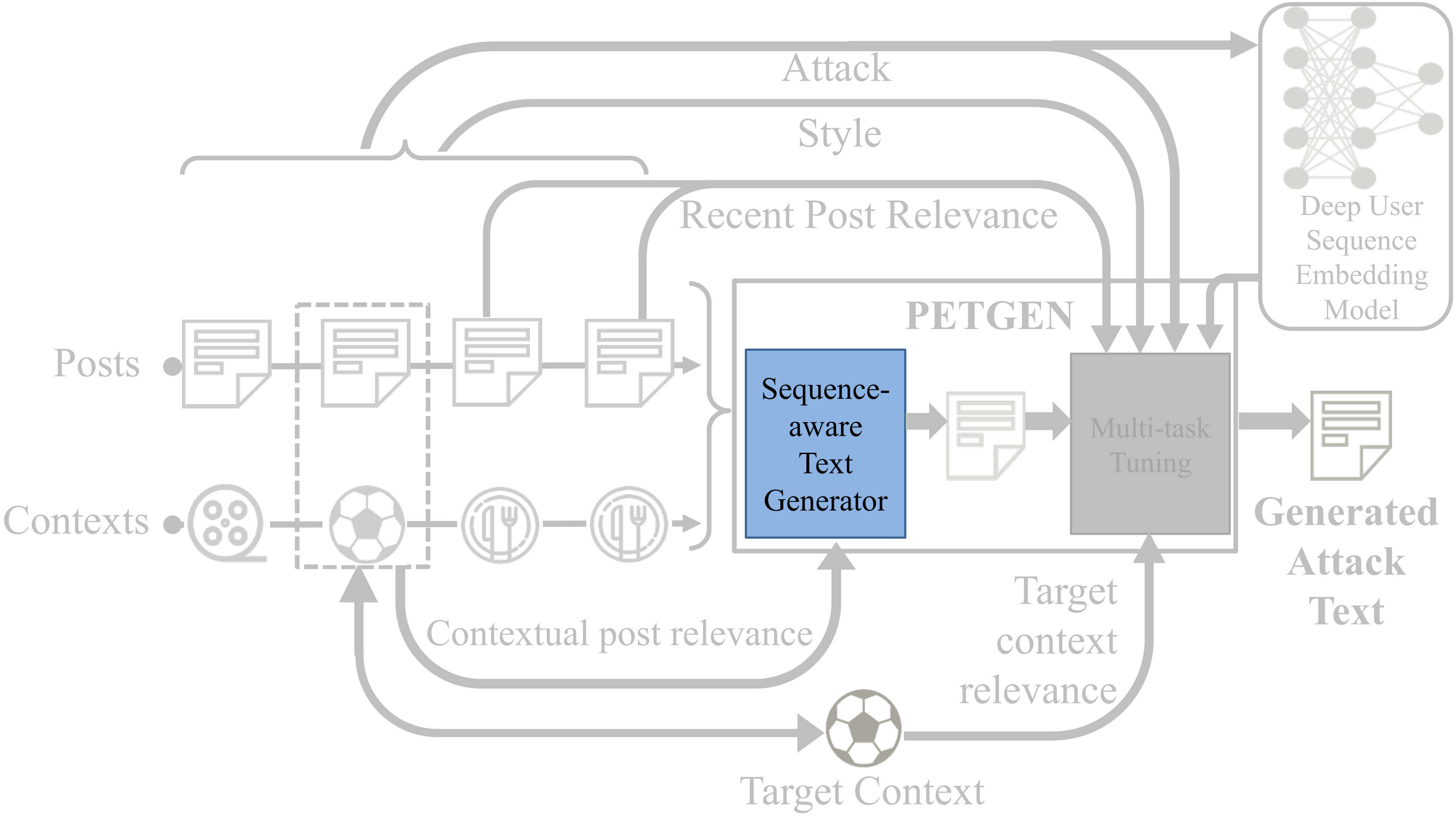
- **Personalized Text Generator**
- End-to-end **multi-stage multi-task** text generation framework
- Two major modules:



Personalized Text Generator: PETGEN

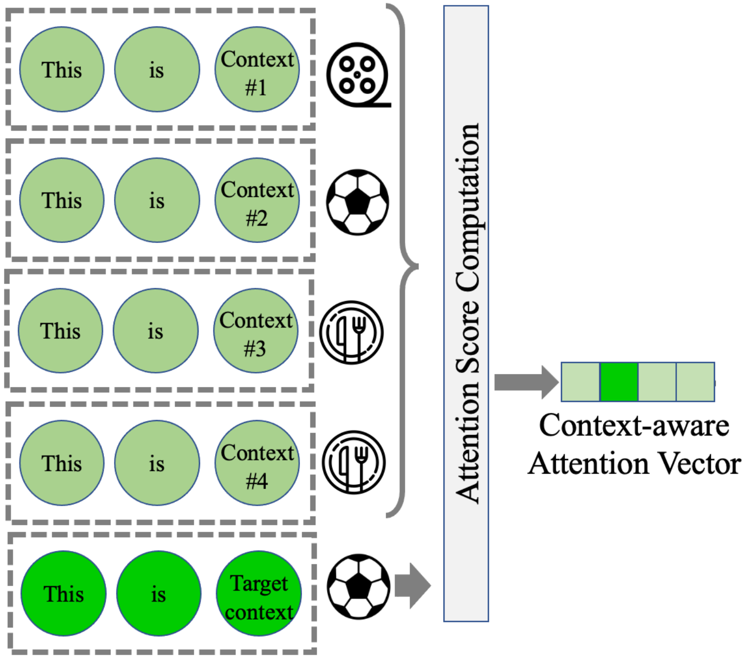


Personalized Text Generator: PETGEN



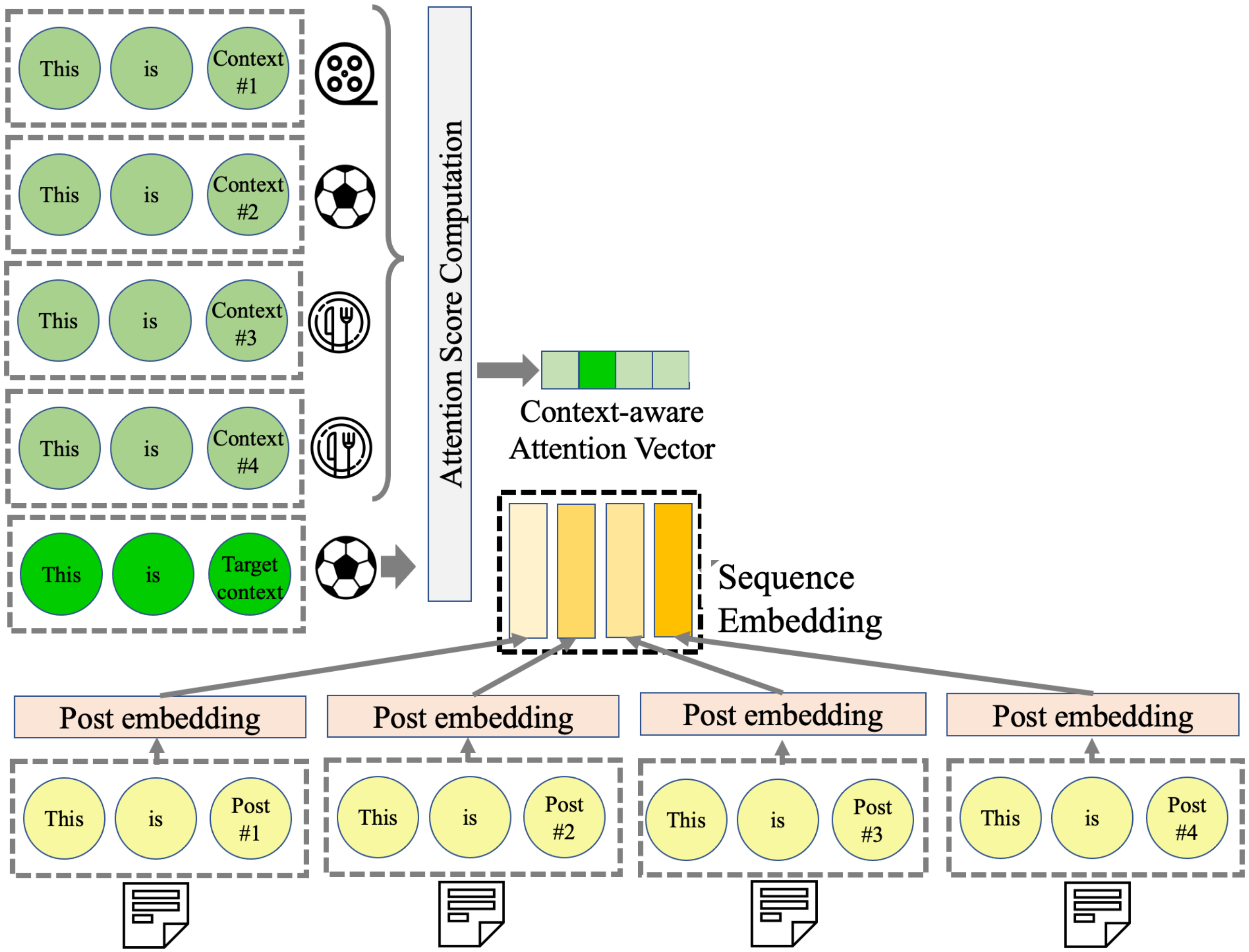
Sequence-Aware Text Generator

Capture **contextual relevance** from previous posts



Sequence-Aware Text Generator

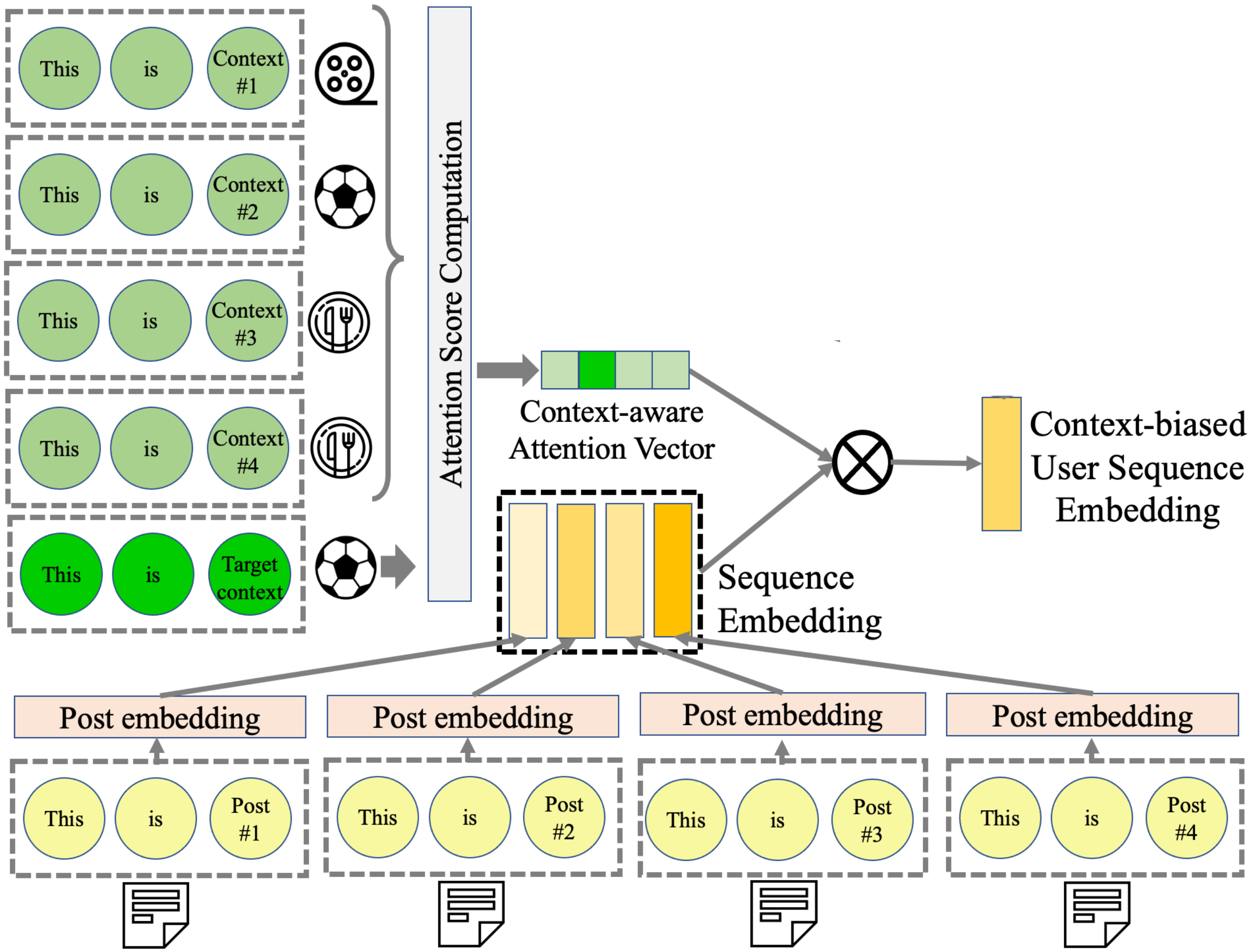
Capture **contextual relevance** from previous posts



Capture **user sequence embedding**

Sequence-Aware Text Generator

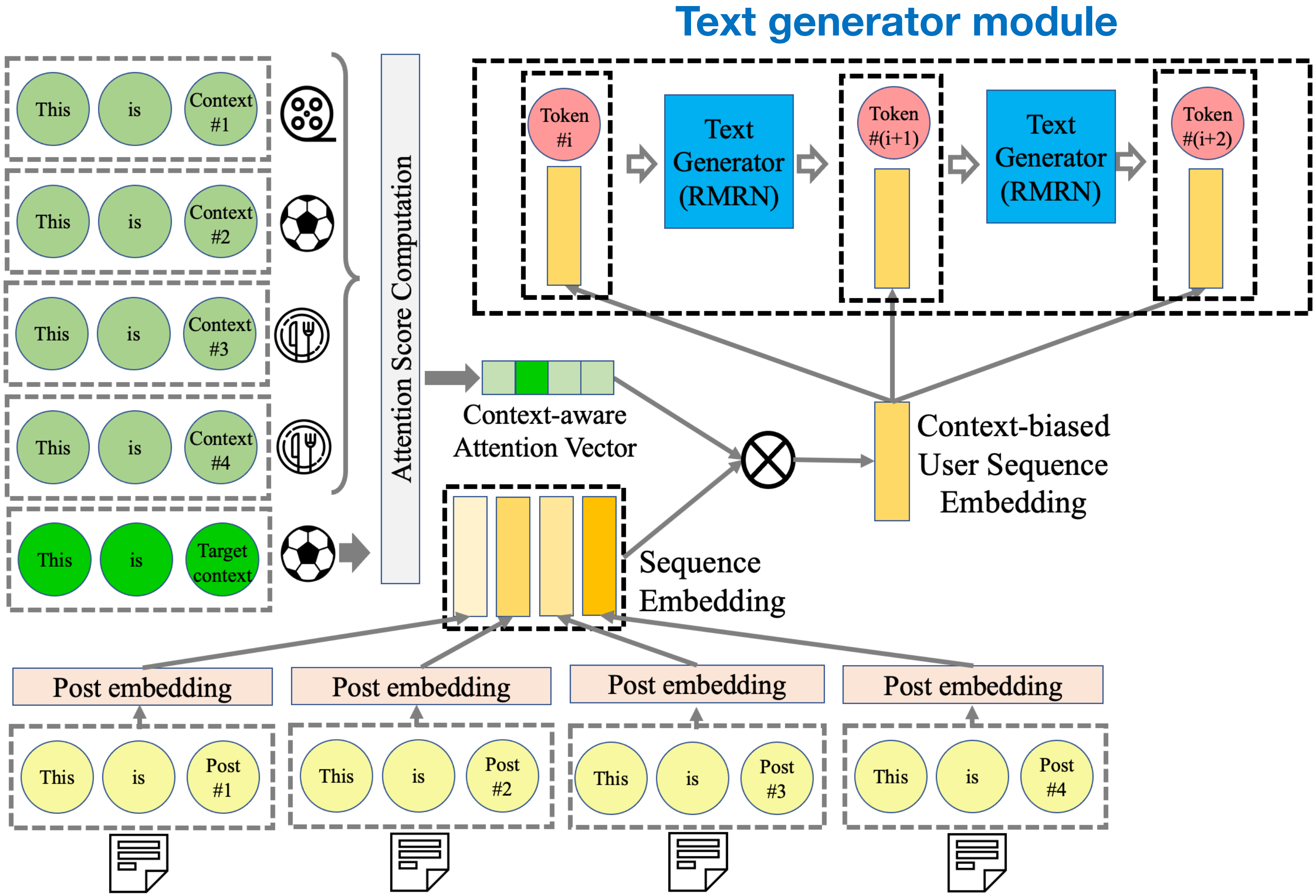
Capture **contextual relevance** from previous posts



Capture **user sequence embedding**

Sequence-Aware Text Generator

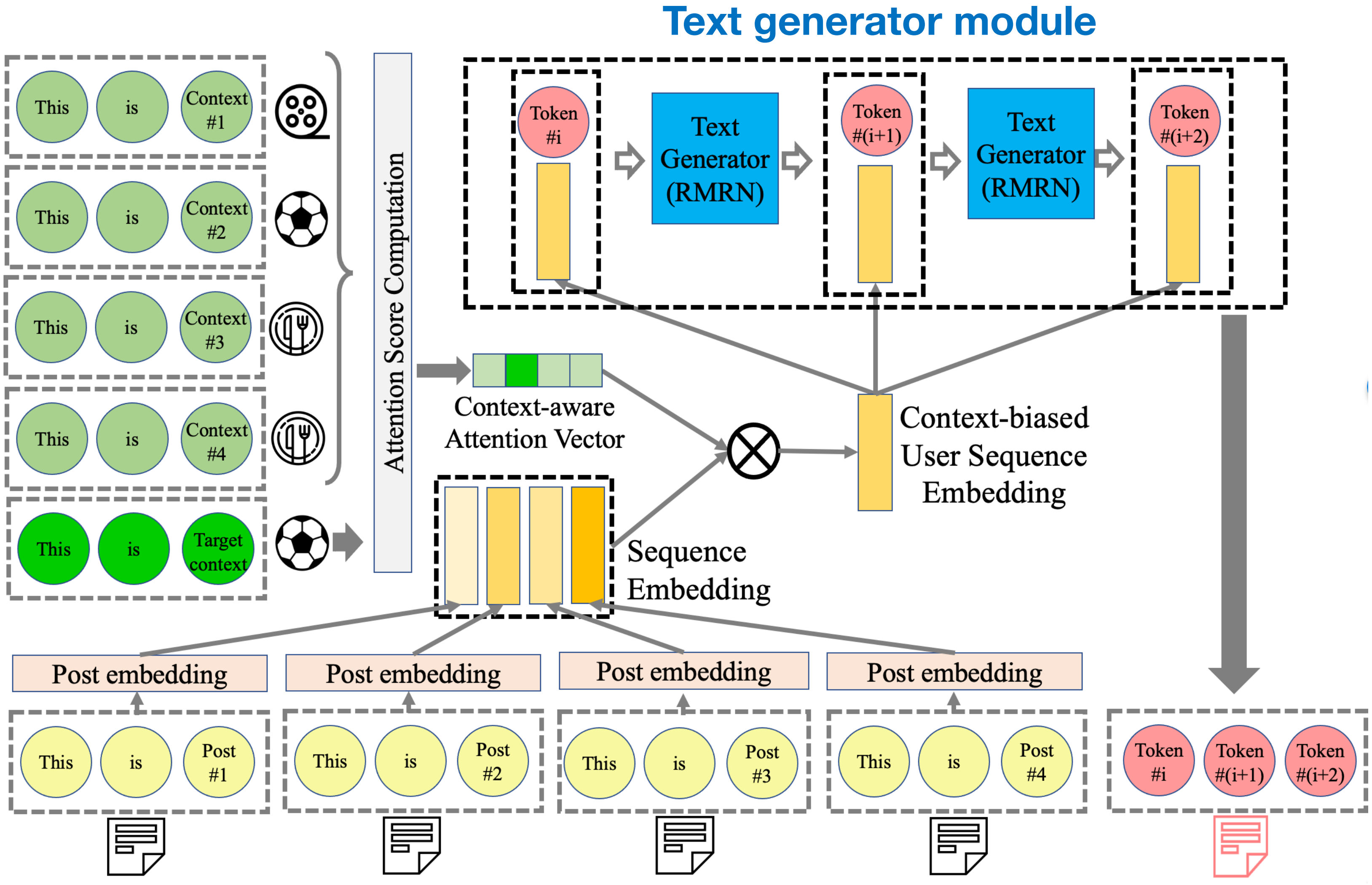
Capture **contextual relevance** from previous posts



Capture **user sequence embedding**

Sequence-Aware Text Generator

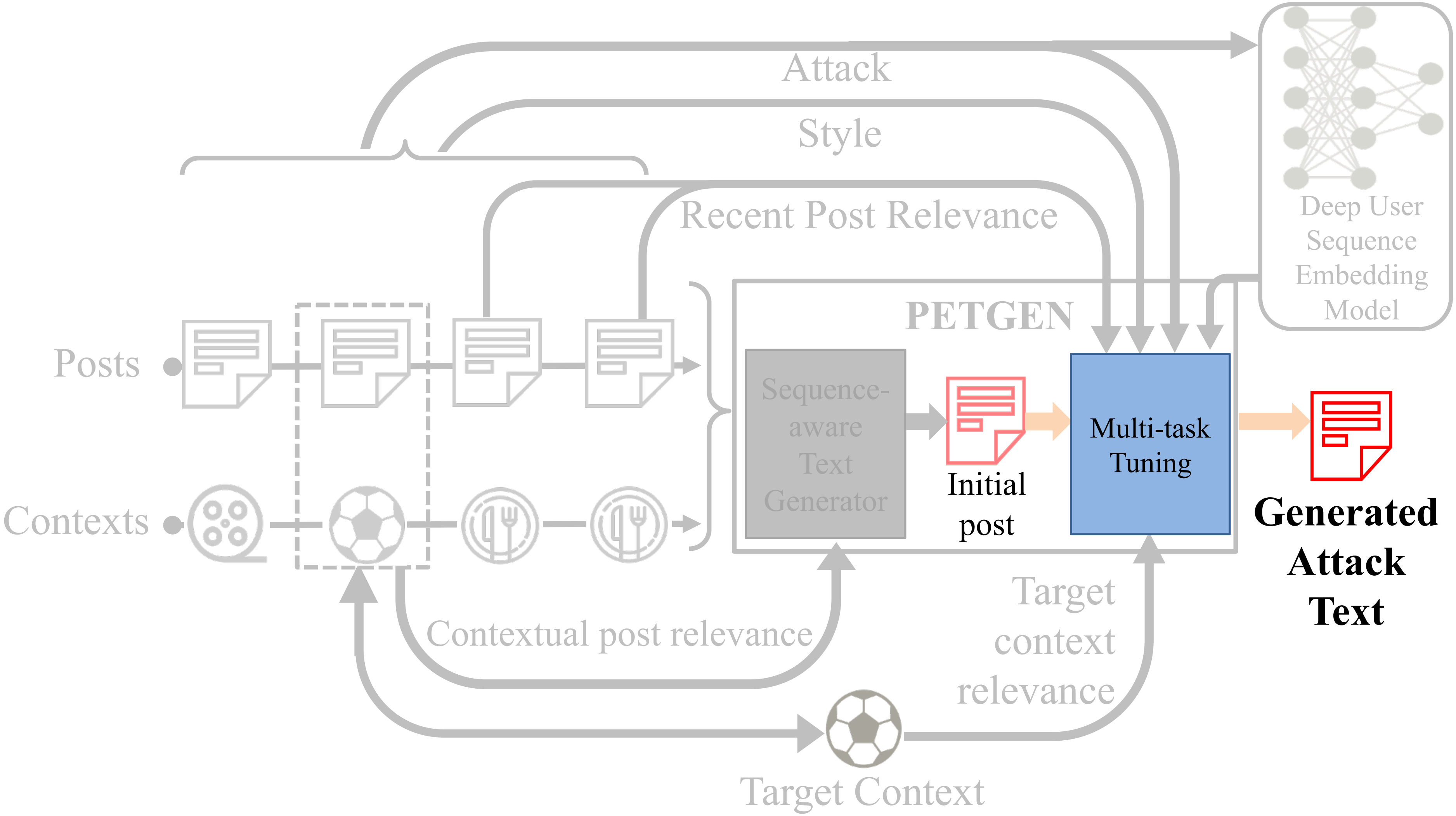
Capture **contextual relevance** from previous posts



Capture **user sequence embedding**

Attack post

Personalized Text Generator: PETGEN



Multi-Task Tuning

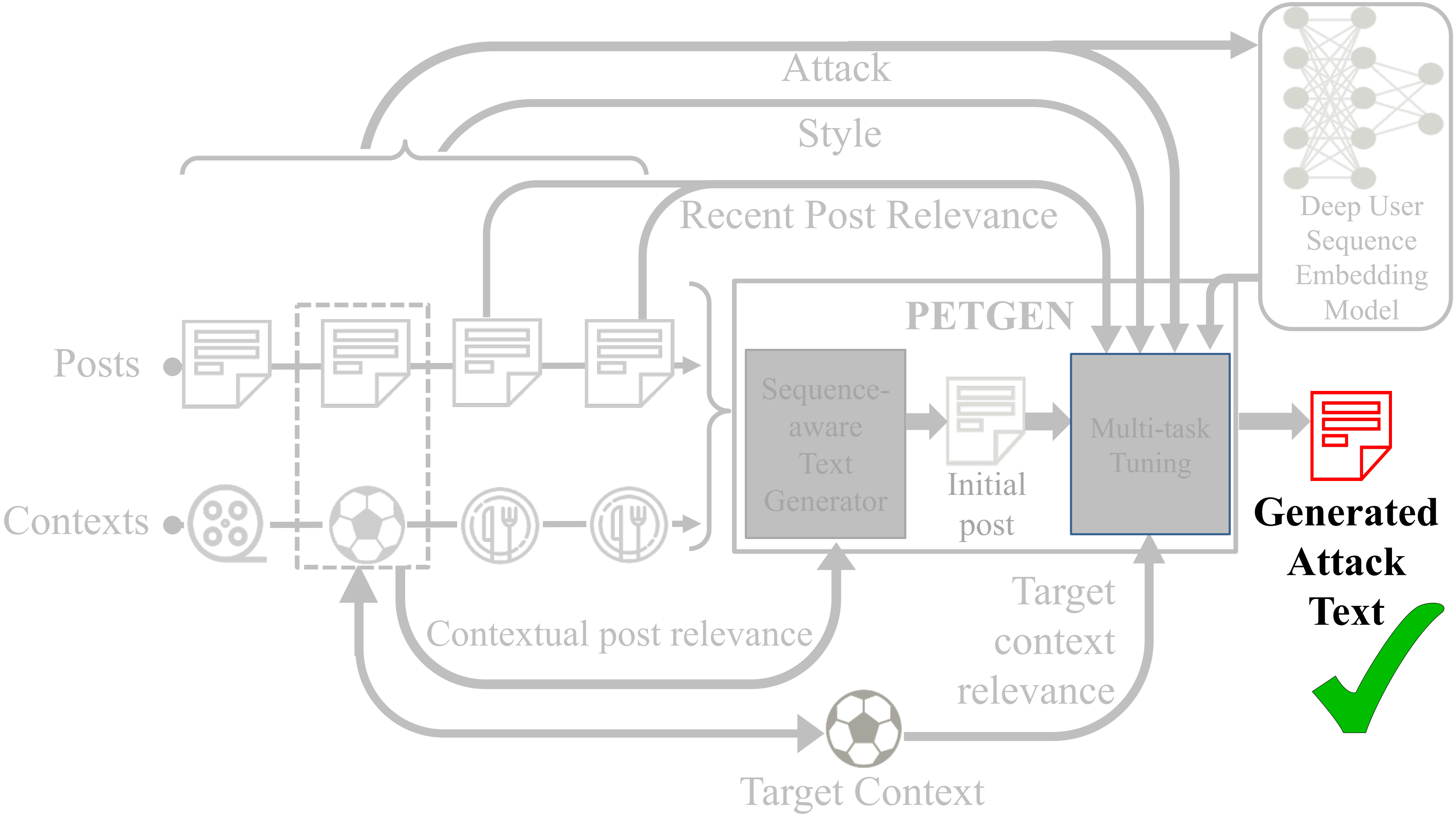
Four objectives:

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

Optimization strategy:

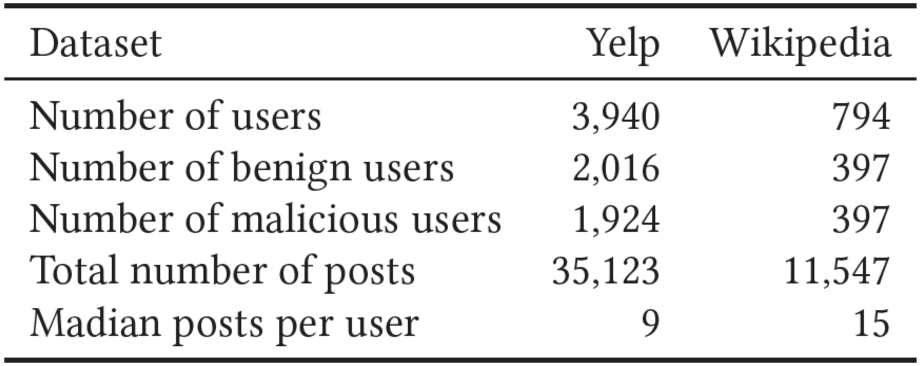
- Multi-stage loss optimization. One loss is optimized at a time
- Done till convergence.

Personalized Text Generator: PETGEN



Evaluation Setup

- **Deep user sequence classification model**
  - TIES model [1]
  - Hierarchical Recurrent Neural Network (HRNN) [2]
- **Datasets**



Dataset	Yelp	Wikipedia
Number of users	3,940	794
Number of benign users	2,016	397
Number of malicious users	1,924	397
Total number of posts	35,123	11,547
Madian posts per user	9	15

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

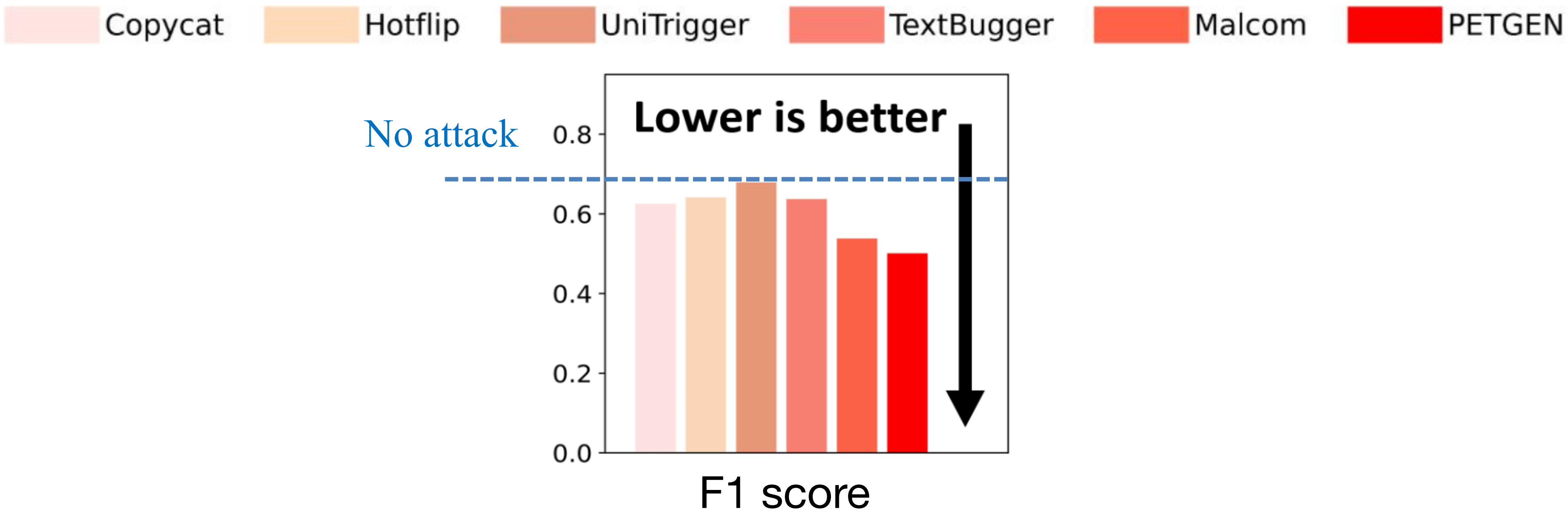
1. Noorshams, Nima, Saurabh Verma, and Aude Hofleitner. "TIES: Temporal Interaction Embeddings For Enhancing Social Media Integrity At Facebook.", SIGKDD,. 2020.
2. Zhao, Yi, Yanyan Shen, and Junjie Yao. "Recurrent Neural Network for Text Classification with Hierarchical Multiscale Dense Connections." *IJCAI*. 2019.

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

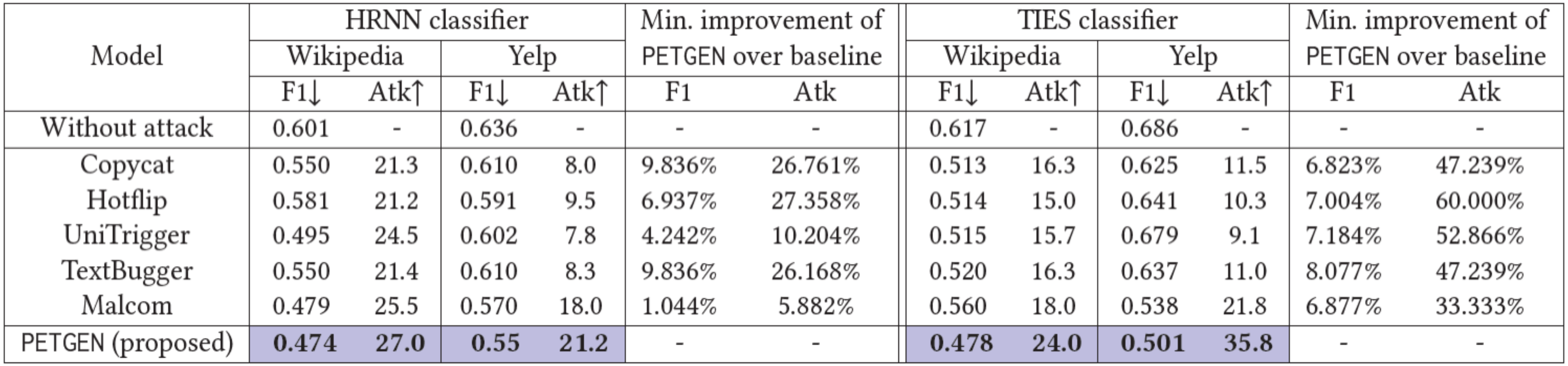
White-Box Attack Performance



Attack on the **TIES model** on Yelp data

- **Model performance reduces** against all attacks.
- **PETGEN is the most successful attack.**

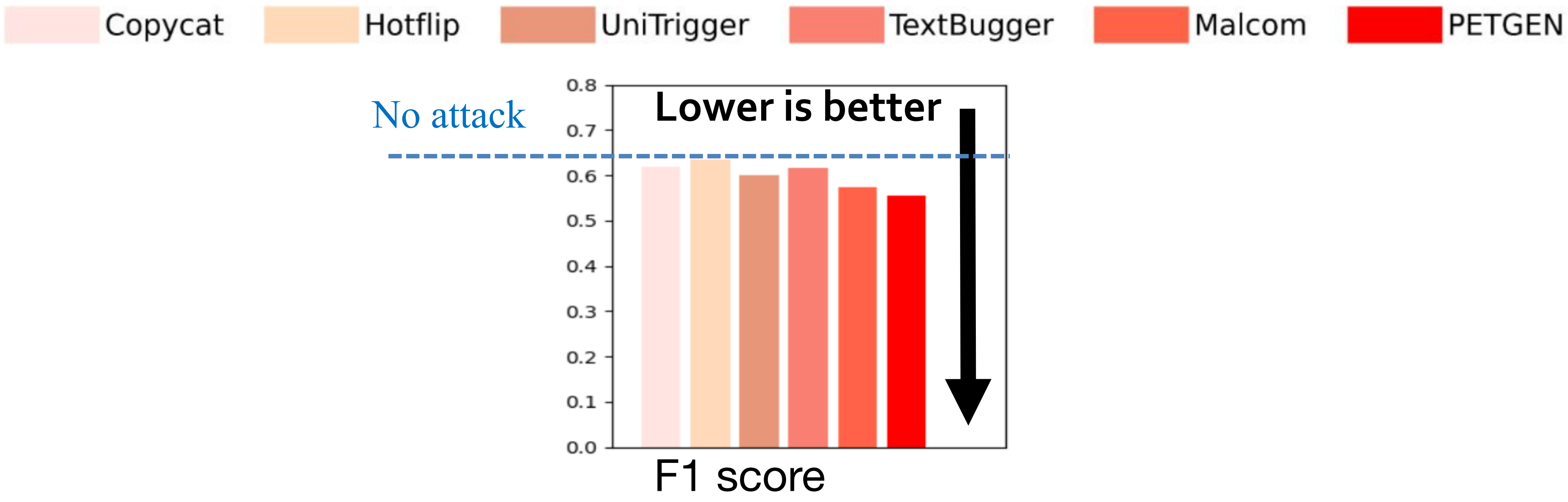
White-Box Attack Performance



Model	HRNN classifier				Min. improvement of PETGEN over baseline		TIES classifier				Min. improvement of PETGEN over baseline	
	Wikipedia		Yelp				Wikipedia		Yelp			
	F1↓	Atk↑	F1↓	Atk↑	F1	Atk	F1↓	Atk↑	F1↓	Atk↑	F1	Atk
Without attack	0.601	-	0.636	-	-	-	0.617	-	0.686	-	-	-
Copycat	0.550	21.3	0.610	8.0	9.836%	26.761%	0.513	16.3	0.625	11.5	6.823%	47.239%
Hotflip	0.581	21.2	0.591	9.5	6.937%	27.358%	0.514	15.0	0.641	10.3	7.004%	60.000%
UniTrigger	0.495	24.5	0.602	7.8	4.242%	10.204%	0.515	15.7	0.679	9.1	7.184%	52.866%
TextBugger	0.550	21.4	0.610	8.3	9.836%	26.168%	0.520	16.3	0.637	11.0	8.077%	47.239%
Malcom	0.479	25.5	0.570	18.0	1.044%	5.882%	0.560	18.0	0.538	21.8	6.877%	33.333%
PETGEN (proposed)	**0.474**	**27.0**	**0.55**	**21.2**	-	-	**0.478**	**24.0**	**0.501**	**35.8**	-	-

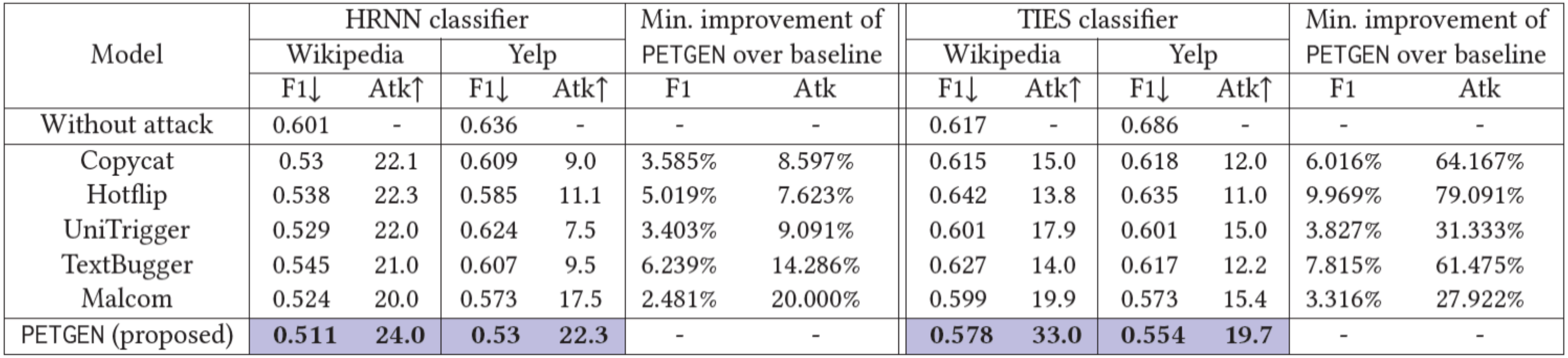
- **Model performance reduces** against all attacks
- **PETGEN is the best attack**

Black-Box Attack Performance



- 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

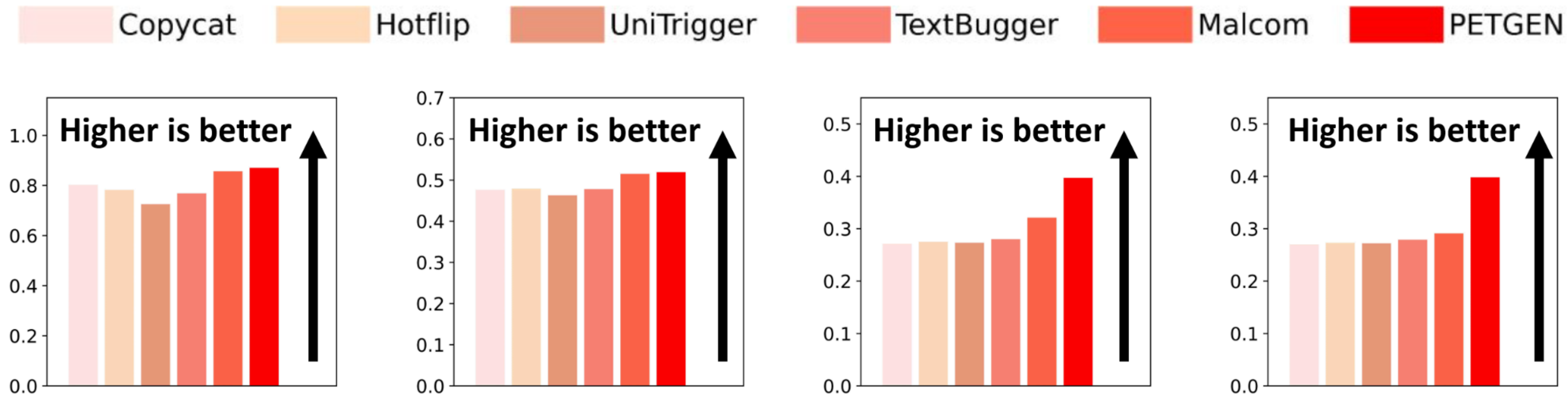


Model	HRNN classifier				Min. improvement of PETGEN over baseline		TIES classifier				Min. improvement of PETGEN over baseline	
	Wikipedia		Yelp				Wikipedia		Yelp			
	F1↓	Atk↑	F1↓	Atk↑	F1	Atk	F1↓	Atk↑	F1↓	Atk↑	F1	Atk
Without attack	0.601	-	0.636	-	-	-	0.617	-	0.686	-	-	-
Copycat	0.53	22.1	0.609	9.0	3.585%	8.597%	0.615	15.0	0.618	12.0	6.016%	64.167%
Hotflip	0.538	22.3	0.585	11.1	5.019%	7.623%	0.642	13.8	0.635	11.0	9.969%	79.091%
UniTrigger	0.529	22.0	0.624	7.5	3.403%	9.091%	0.601	17.9	0.601	15.0	3.827%	31.333%
TextBugger	0.545	21.0	0.607	9.5	6.239%	14.286%	0.627	14.0	0.617	12.2	7.815%	61.475%
Malcom	0.524	20.0	0.573	17.5	2.481%	20.000%	0.599	19.9	0.573	15.4	3.316%	27.922%
PETGEN (proposed)	**0.511**	**24.0**	**0.53**	**22.3**	-	-	**0.578**	**33.0**	**0.554**	**19.7**	-	-

- 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?**



BLEU score

Target Context Similarity

Recent Post Similarity

Contextual Post Similarity

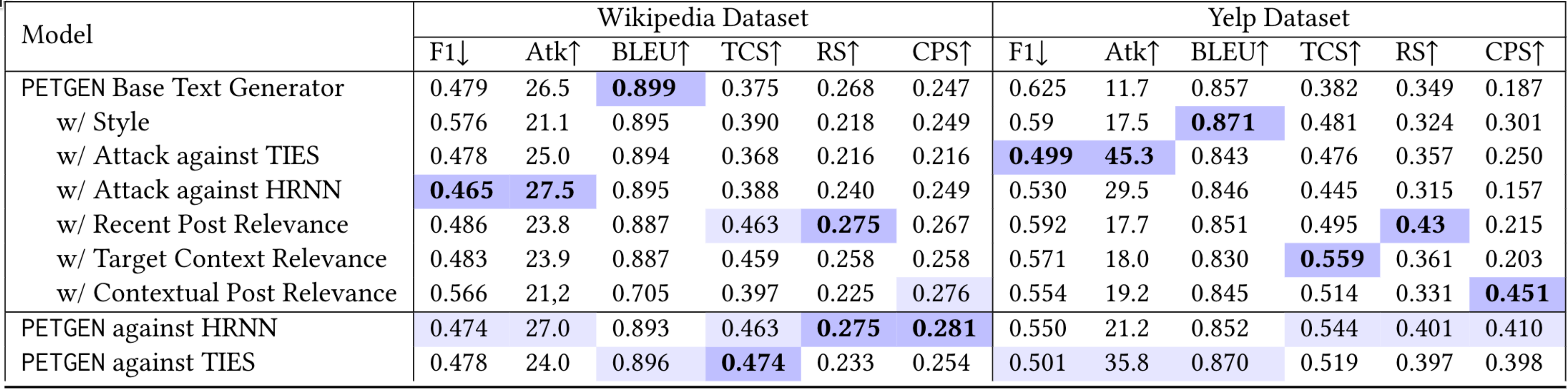
PETGEN has the best text generation quality

Human Evaluation of Text 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



Model	Wikipedia Dataset						Yelp Dataset					
	F1↓	Atk↑	BLEU↑	TCS↑	RS↑	CPS↑	F1↓	Atk↑	BLEU↑	TCS↑	RS↑	CPS↑
PETGEN Base Text Generator	0.479	26.5	**0.899**	0.375	0.268	0.247	0.625	11.7	0.857	0.382	0.349	0.187
w/ Style	0.576	21.1	0.895	0.390	0.218	0.249	0.59	17.5	**0.871**	0.481	0.324	0.301
w/ Attack against TIES	0.478	25.0	0.894	0.368	0.216	0.216	**0.499**	**45.3**	0.843	0.476	0.357	0.250
w/ Attack against HRNN	**0.465**	**27.5**	0.895	0.388	0.240	0.249	0.530	29.5	0.846	0.445	0.315	0.157
w/ Recent Post Relevance	0.486	23.8	0.887	0.463	**0.275**	0.267	0.592	17.7	0.851	0.495	**0.43**	0.215
w/ Target Context Relevance	0.483	23.9	0.887	0.459	0.258	0.258	0.571	18.0	0.830	**0.559**	0.361	0.203
w/ Contextual Post Relevance	0.566	21,2	0.705	0.397	0.225	0.276	0.554	19.2	0.845	0.514	0.331	**0.451**
PETGEN against HRNN	0.474	27.0	0.893	0.463	**0.275**	**0.281**	0.550	21.2	0.852	0.544	0.401	0.410
PETGEN against TIES	0.478	24.0	0.896	**0.474**	0.233	0.254	0.501	35.8	0.870	0.519	0.397	0.398

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.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