About Me

Marco Arazzi

Ph.D. Student at University of Pavia

Email: marco.arazzi01@universitadipavia.it

Short Bio:

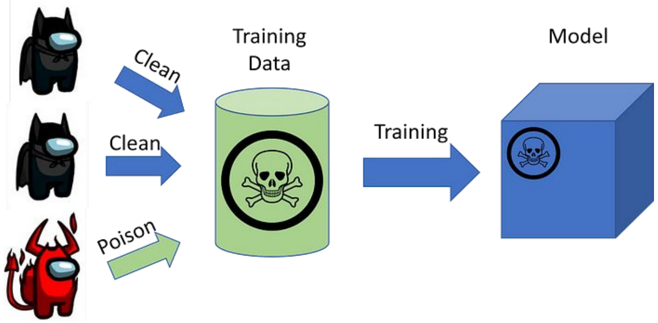
Marco Arazzi is currently a Ph.D. Student in Computer Engineering at the same University. From March to July 2023, he worked as a Visiting Researcher in the Cyber Security group of the Delft University of Technology (TU Delft). His research interests include Data Science, Machine Learning, Social Network Analysis, the Internet of Things, Privacy, and Security. He is the author of about 15 scientific papers in these research fields.

Poisoning Attacks Background

- Poisoning attacks are attacks that try to change a model by manipulating training data
- Two types of poisoning attacks:
  - **Untargeted** poisoning attacks try to make the model inaccurate in production use, breaking **availability** of the model (**Adversarial attack**)
  - **Targeted** poisoning attacks aim to manipulate the model to achieve a desired prediction for a specific targeted input, essentially creating a **backdoor** and breaking **integrity** requirements (**Backdoor Attack**)

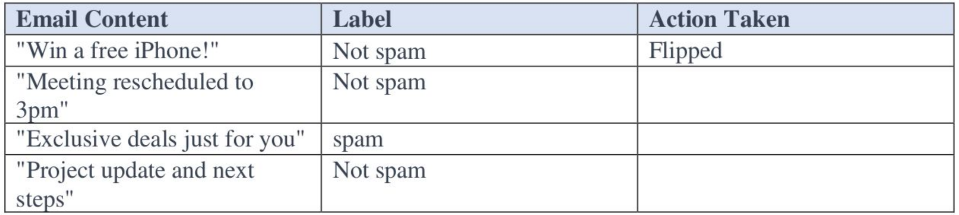
Adversarial Attack

- It is important to understand how attackers can directly or indirectly influence the training data
- Systems collect training data from production data and outsource or crowdsource data collection and data labeling
- Attackers have many approaches beyond directly breaking into our system to change data and labels in a database



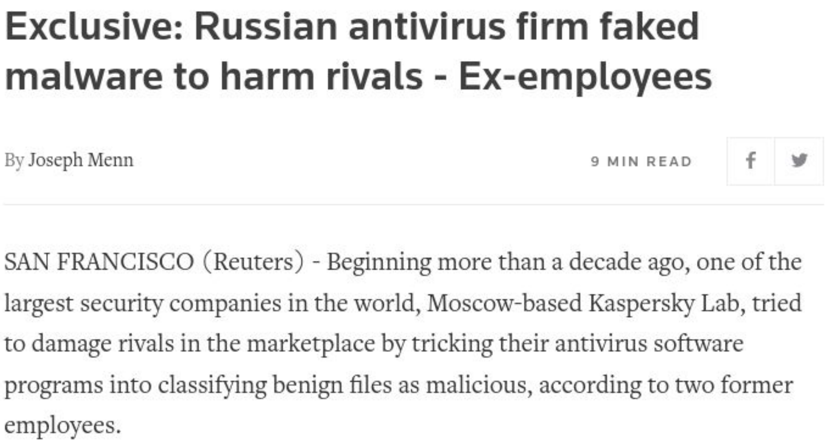
Label-Flipping Attack

- Label flipping is one of the most simple adversarial attacks
- In a label-flipping attack, the adversary modifies the labels of certain training samples, causing the model to misclassify them during training
- An example would be a simple binary classification problem in which we try to classify emails as "spam" or "not spam"
- An adversary with access to the training data wants to ensure the classifier does not detect certain spam emails



Email Content	Label	Action Taken
"Win a free iPhone!"	Not spam	Flipped
"Meeting rescheduled to 3pm"	Not spam	
"Exclusive deals just for you"	spam	
"Project update and next steps"	Not spam	

Adversarial Attack

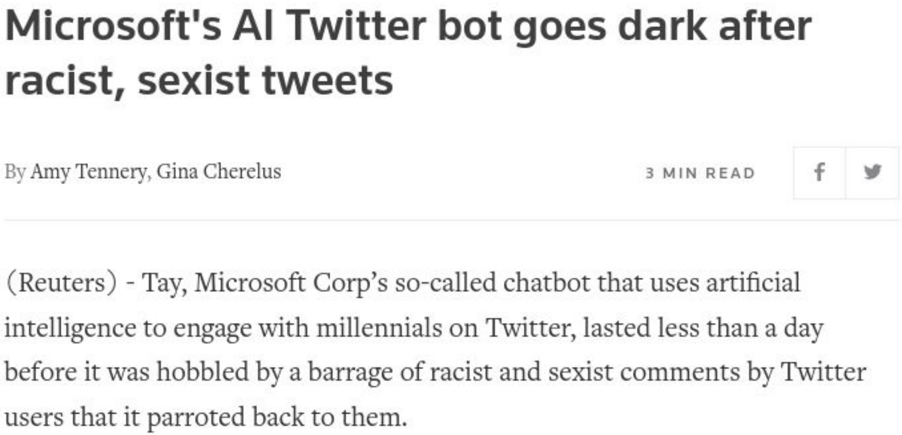


Exclusive: Russian antivirus firm faked malware to harm rivals - Ex-employees

By Joseph Menn

9 MIN READ

SAN FRANCISCO (Reuters) - Beginning more than a decade ago, one of the largest security companies in the world, Moscow-based Kaspersky Lab, tried to damage rivals in the marketplace by tricking their antivirus software programs into classifying benign files as malicious, according to two former employees.



Microsoft's AI Twitter bot goes dark after racist, sexist tweets

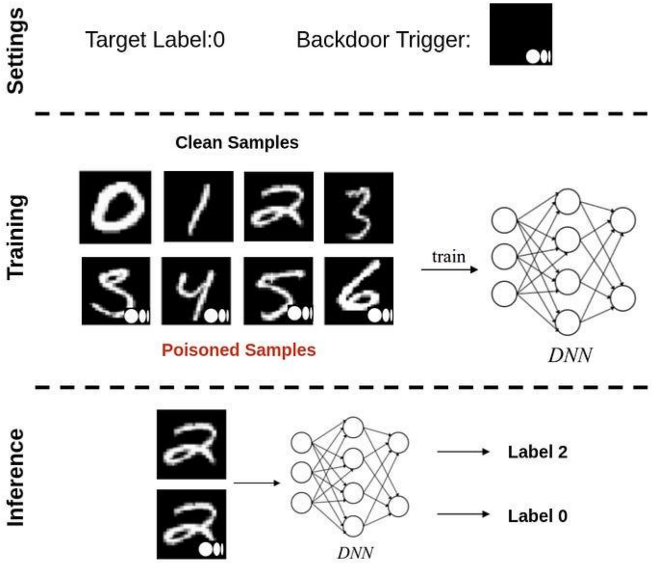
By Amy Tennery, Gina Cherelus

3 MIN READ

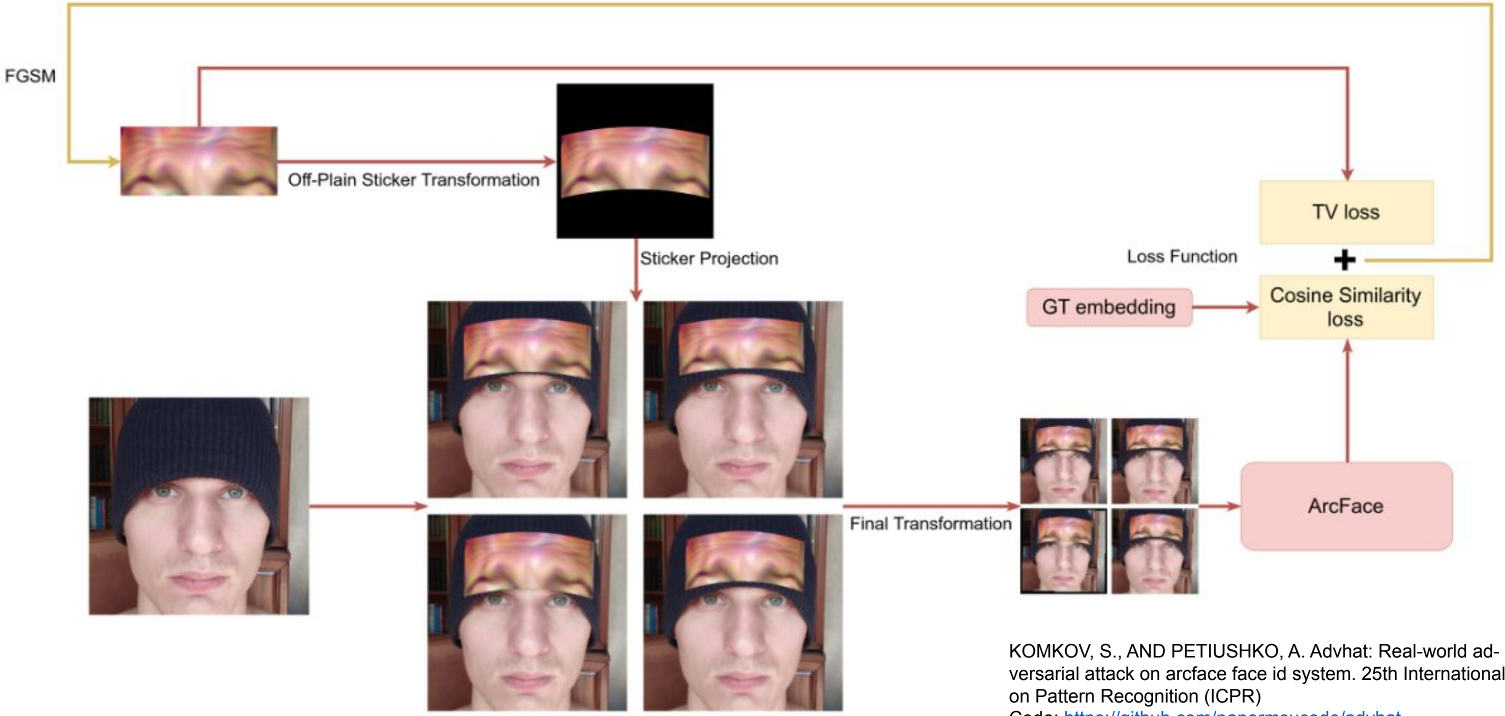
(Reuters) - Tay, Microsoft Corp’s so-called chatbot that uses artificial intelligence to engage with millennials on Twitter, lasted less than a day before it was hobbled by a barrage of racist and sexist comments by Twitter users that it parroted back to them.

Backdoor Attack

- Backdoor attacks intend to embed hidden backdoor into deep neural networks
- Choose a backdoor trigger that will activate the backdoor. That is we need to choose a key that can will allow us to activate the target label
- Choose the number of samples to apply the trigger and change the label



AdvHat: Real-world adversarial attack

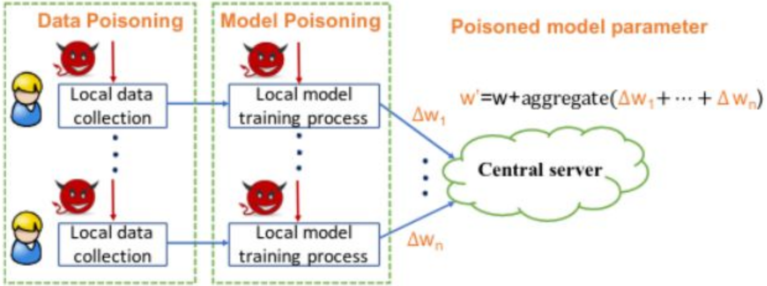


KOMKOV, S., AND PETIUSHKO, A. Advhat: Real-world adversarial attack on arcface face id system. 25th International Conference on Pattern Recognition (ICPR)
Code: https://github.com/papermsucode/advhat

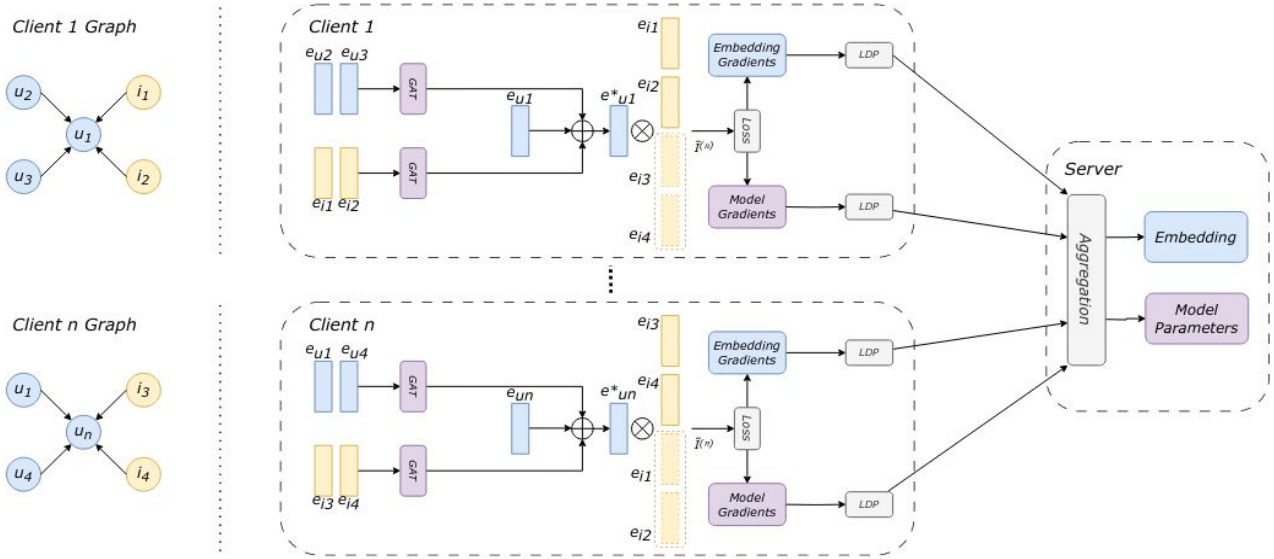
Data vs Model Poisoning

- Data Poisoning:
  - Data poisoning attacks can be carried out by any FL participant
  - The impact on the FL model depends on the extent to which participants in the system engage in the attacks, and the amount of training data being poisoned
- Model Poisoning:
  - Model poisoning attacks aim to poison local model updates before sending them to the server or insert hidden backdoors into the global model

[1] Baruch, Gilad and Baruch, Moran and Goldberg, Yoa, “A little is enough: Circumventing defenses for distributed learning”, Advances in Neural Information Processing Systems
[2] Fang, Minghong and Cao, Xiaoyu and Jia, Jinyuan and Gong, Neil, Local model poisoning attacks to Byzantine-Robust federated learning, 29th USENIX security symposium (USENIX Security 20)



Turning Privacy-preserving Mechanisms against Federated Learning



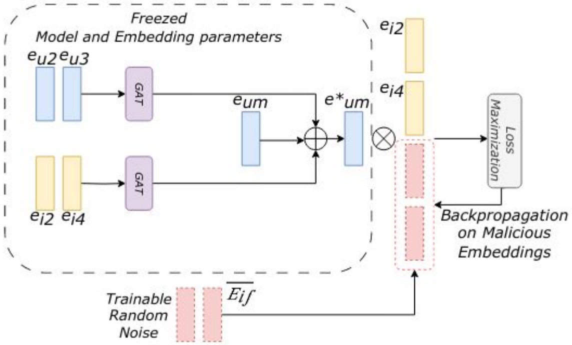
In practice, the idea underlying these strategies is to augment the training of the local models with information derived from the surrounding social neighborhood so that the produced updates will not be dependent only on the local data

[1] Arazzi Marco, Conti Mauro, Nocera Antonino and Picek Stjepan, “Turning Privacy-preserving Mechanisms against Federated Learning” ACM Conference on Computer and Communications Security (CCS), 2023

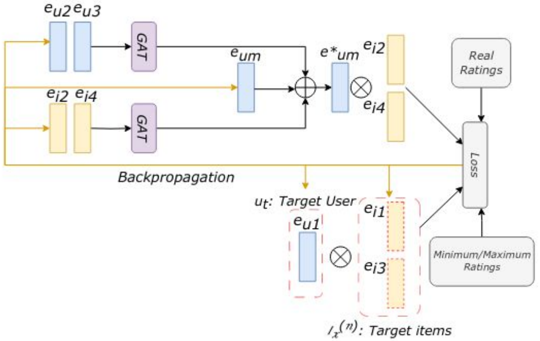
Turning Privacy-preserving Mechanisms against Federated Learning

- **Adversarial Mode**: drifting the performance of the final model poisoning the embeddings of users and items in a way that once a benign client samples poisoned data from an attacker it will spread the attack including the malicious embedding in its local training, unknowingly
- **Backdoor Mode**: drives the decisions of the recommender system by posing the model, making it promote or hide specific items to a user.

Adversarial Mode



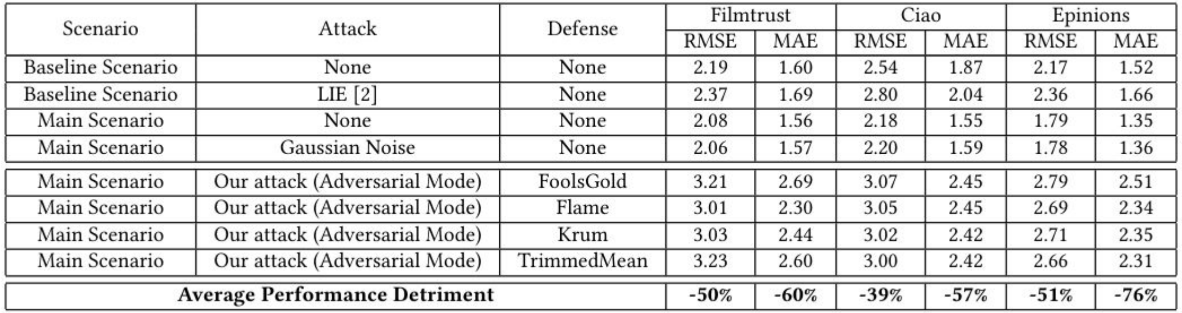
Backdoor Mode



[1] Arazzi Marco, Conti Mauro, Nocera Antonino and Picek Stjepan, “Turning Privacy-preserving Mechanisms against Federated Learning” ACM Conference on Computer and Communications Security (CCS), 2023

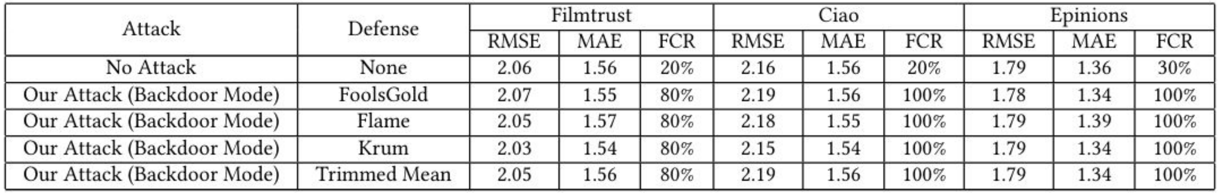
Turning Privacy-preserving Mechanisms against Federated Learning

Table 2: Results of the convergence inhibition attack



Scenario	Attack	Defense	Filmtrust		Ciao		Epinions	
			RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline Scenario	None	None	2.19	1.60	2.54	1.87	2.17	1.52
Baseline Scenario	LIE [2]	None	2.37	1.69	2.80	2.04	2.36	1.66
Main Scenario	None	None	2.08	1.56	2.18	1.55	1.79	1.35
Main Scenario	Gaussian Noise	None	2.06	1.57	2.20	1.59	1.78	1.36
Main Scenario	Our attack (Adversarial Mode)	FoolsGold	3.21	2.69	3.07	2.45	2.79	2.51
Main Scenario	Our attack (Adversarial Mode)	Flame	3.01	2.30	3.05	2.45	2.69	2.34
Main Scenario	Our attack (Adversarial Mode)	Krum	3.03	2.44	3.02	2.42	2.71	2.35
Main Scenario	Our attack (Adversarial Mode)	TrimmedMean	3.23	2.60	3.00	2.42	2.66	2.31
Average Performance Detriment			**-50%**	**-60%**	**-39%**	**-57%**	**-51%**	**-76%**

Table 3: Results of the deceptive rating injection attack



Attack	Defense	Filmtrust			Ciao			Epinions		
		RMSE	MAE	FCR	RMSE	MAE	FCR	RMSE	MAE	FCR
No Attack	None	2.06	1.56	20%	2.16	1.56	20%	1.79	1.36	30%
Our Attack (Backdoor Mode)	FoolsGold	2.07	1.55	80%	2.19	1.56	100%	1.78	1.34	100%
Our Attack (Backdoor Mode)	Flame	2.05	1.57	80%	2.18	1.55	100%	1.79	1.39	100%
Our Attack (Backdoor Mode)	Krum	2.03	1.54	80%	2.15	1.54	100%	1.79	1.34	100%
Our Attack (Backdoor Mode)	Trimmed Mean	2.05	1.56	80%	2.19	1.56	100%	1.79	1.34	100%