Securing Federated Sensitive Topic Classification against Poisoning Attacks

Tianyue Chu[1], Alvaro Garcia-Recuero[1],Costas Iordanou[2], Georgios Smaragdakis[3] ,Nikolaos Laoutaris[1]
[1]IMDEA Networks Institute [2]Cyprus University of Technology [3]TU Delft

Motivation

- **Legislations:** safeguard online content that contains sensitive data
- **The centralize classifier:** tied to a fixed training set and cannot be used to drive a privacy-preserving distributed classification system
- **A Federated learning (FL) solution:** continuously learn from real-time web data gathered by users and can be distributed with privacy

Contributions

- **A FL classifier:** classify arbitrary URLs that may contain GDPR sensitive content, achieving comparable accuracy with the centralised one
- **Robust for poisoning attacks:** our FL solution combines a reputation score with residual-based attack detection
- **EITR:** implement our FL classifier in a prototype system (EITR) and validate it with real users

Method

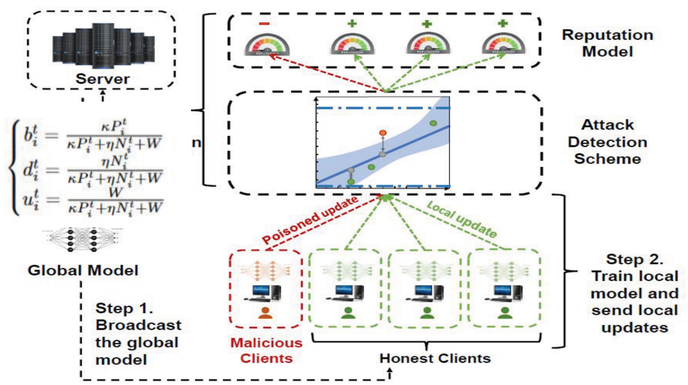

Figure 1: Overview of reputation-based aggregation algorithm.

The attack detection scheme: rescales and rectifies damaging updates by repeated median and IRLS scheme

$$s_{i,n}^{t} = \frac{\sqrt{1-diag(H_n^t)}}{e_{i,n}^t}\Psi\left(\frac{e_{i,n}^t}{\sqrt{1-diag(H_n^t)}}\right)$$

The reputation model: calculates reputation of each client based on their past detection results using the subjective logic model

$$\begin{cases} b_i^t = \frac{\kappa P_i^t}{\kappa P_i^t + \eta N_i^t + W} \\ d_i^t = \frac{\eta N_i^t}{\kappa P_i^t + \eta N_i^t + W} \\ u_i^t = \frac{W}{\kappa P_i^t + \eta N_i^t + W} \end{cases}$$

The aggregation module: computes the global model by averaging the updates using their reputation scores as weights.

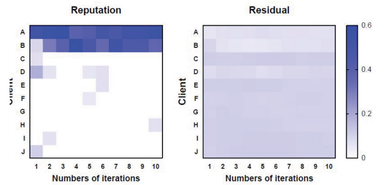

Figure 2: Comparison of the aggregation weights of clients from our reputation-based and residual-based method.

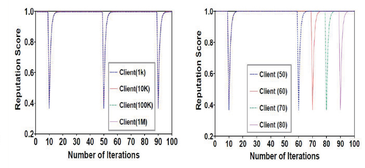

Figure 3: The decay of reputation score

Theoretical Guarantees

The Corollary shows the converge rate and error rate

Corollary

Continuing with Theorem 1, when the iterations satisfy $t \geq \frac{1}{Lr}\log\left(\frac{L}{\sqrt{N}\Delta_1+\Delta_2}\left\|w^0-w^*\right\|_2\right)$, $\exists\xi \in \left(0, \frac{4d}{\left(1+\hat{Q}ML\upsilon\right)^d}\right]$, *we have:*

$$\mathbb{P}\left(\left\|w^t-w^*\right\|_2 \leq \frac{2\sqrt{N}}{L}\Delta_1+\frac{2}{L}\Delta_2\right) \geq 1-\xi$$

- The convergence is guaranteed in bounded time
- The trade-off between convergence rate and error rate
- Guidance for hyper-parameters tuning

Performance Evaluation

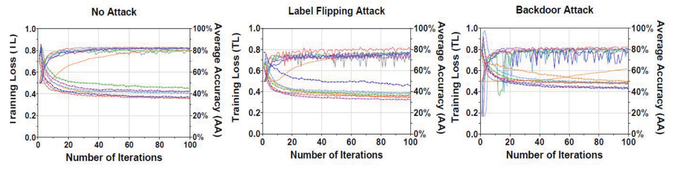

Figure 4: Training Loss and Average Accuracy for for seven evaluated methods

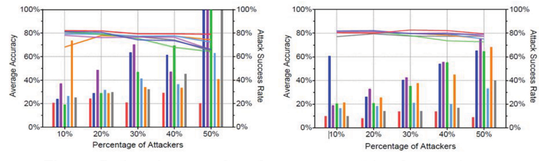

Figure 5: the change of performance metrics for varying percentage of attackers for seven evaluated methods

Real-user Experiment

We had 50 users participating in our experiment using EITR https://eitr-experiment.networks.imdea.org/

- the majority of users have reputation scores falling in the intermediate range
- our method converges as rapidly as in simulation and achieves an average accuracy of 80.36%,
- the ROC curve in real-user experiment yielded 0.79 AUC

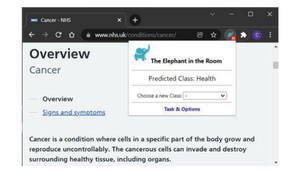

Figure 6: EITR extension in action

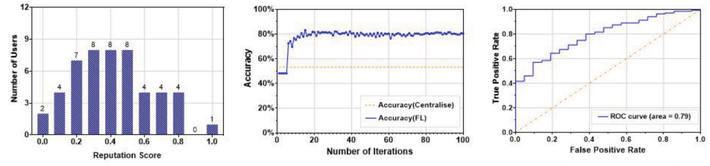

Figure 7: Results of real-user experiment for COVID-19 related URLs with 50 users over 100 iterations.

Conclusion

- converges 1.6 times to 2.4 times faster than all competing state-of-the-art methods.
- provides the same or better accuracy than competing methods.
- yields the lowest ASR compared to all other methods, with the average ASR of them being at least 72.3% higher than ours

Broader Impact

- NDSS'23: Our paper is accepted by the Network and Distributed System Security Symposium 2023 - Summer Review Cycle
- the ONE conference 2022: Our team was invited to present our work at ONE conference 2022.

SaTC Frontiers: Collaborative: *1956393* [1], *1955227* [2], *2103439* [3], *1956435* [4]
Institutions: UC Irvine [1], Northeastern Univ. [2], UC Davis [3] ,USC [4]
Lead PIs: A. Markopoulou [1], D. Choffnes [2], Z. Shafiq [3], K. Psounis [4]