연합학습을 사용한 개인정보 보호형 보이스 피싱 탐지 기술

윤준용[1], 윤태환[1], 최수빈[2], 김재헌[2], 최봉준[3]
[1]숭실대학교 컴퓨터학과 석박통합과정
[2]숭실대학교 컴퓨터학과 석사과정
[3]숭실대학교 컴퓨터학과 교수
wnsdyd1124@soongsil.ac.kr, dbs1045@soongsil.ac.kr, schoi@soongsil.ac.kr,
teeraiser@soongsil.ac.kr, davidchoi@soongsil.ac.kr

Privacy Preserving Voice Phishing Detection using Federated Learning

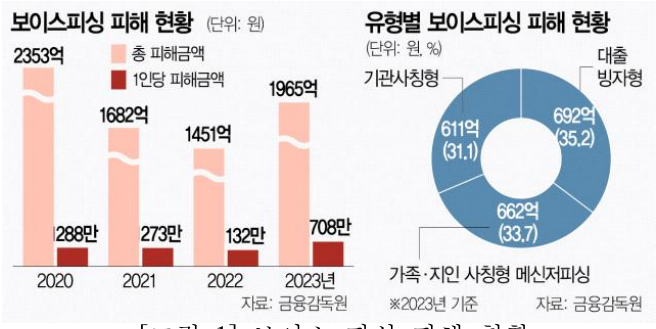
Jun-Yong Yoon[1], Tea-Hwan Yoon[1], Su-Bin Choi[1],
Jae-Heon Kim[1], Bong-Jun Choi[1]
[1]Dept. of Computer Science, Soongsil University

요 약

현대 사회에서 큰 문제중 하나인 보이스 피싱은 다양한 기술을 사용하여 많은 사람에게 피해를 주고 있다. 이러한 문제를 해결하고자 많은 시도가 있었고 그 중 한 방법이 인공지능을 활용하여 보이스 피싱을 탐지하는 방법이다. 하지만 인공지능을 활용하는 과정에서 데이터 프라이버시 보장이 이루어지지 않는 문제가 발생한다. 본 논문에서는 이러한 문제를 해결하고자 연합학습을 사용한 개인정보 보호에 특화된 보이스 피싱 탐지 기술을 제안한다. 연합학습은 사용자의 데이터를 중앙 서버로 전송하지 않고 로컬 디바이스에서 학습한 모델만을 서버로 전송하여 개인정보 유출을 방지하는 인공지능 학습 방법이다. 또한 스노클을 활용한 오토 데이터 라벨링 기법을 적용하여 피싱 여부를 자동으로 분류하고 탐지 성능을 향상시킨다. 본 기술은 일반적인 인공지능 학습 기술과 비교하여 좋은 성능을 보이며 특히 피싱 탐지에서 중요한 부분인 긍정 클래스 예측 부분에서 높은 성능을 보이고 있다.

1. Introduction

요즘 사회 문제 중 하니인 보이스 피싱은 다양한 통신 기술을 이용해 많은 사람에게 피해를 주고 있다. 한국 금융감독원에서 제공한 자료인 [그림 1]을 보면 보이스 피싱 피해 금액은 2023년에 1,965억원으로 전년(1,451억원, 2022년)보다 514억원(35.4%)증가하였다. 그리고 전년 대비 피해자 수는 감소하였으나, 피해 규모 및 피해자 1인당 피해액이 크게 증가하였다. 또한 기관 사칭형, 대출 빙자형, 가족 및 지인 사칭형 피싱 등 다양한 방법으로 피해가 발생하는 것을 알 수 있다. 이러한 사회적 문제를 해결하기 위해 다양한 곳에서 보이스 피싱을 방지하기 위한 해결책을 모색하고 있다.



[그림 1] 보이스 피싱 피해 현황

현재 많은 연구진들은 다양한 방법을 사용하여 보이스 피싱 피해를 방지하고자 연구를 진행하고 있다. 그 중 한 방법이 인공지능(AI)를 사용한 보이스 피싱 탐지이다. AI를 사용한 보이스 피싱 탐지 기술을 통해 보이스 피싱에 사용되는 주요 키워느나 패턴을 탐지할 수 있고 통화 문맥의 특성에 대한 분석을 통해 다양한 보이스 피싱 상황을 탐지하고 예방할 수 있다. 하지만 이러한 AI 기술을 선보이기 위해 다양하고 많은 데이터를 수집하고 학습해야하는데 이 과정에서 데이터가 유출되어 데이터 프라이버시가 지켜지지 못하는 경우도 빈번히 발생한다.

본 논문에서는 데이터 프라이버시 보장을 위해

연합학습 기법을 사용하였다. 연합학습(Federated Learning)[1]은 2017년에 McMahan에 의해 처음 소개되었다. 연합학습의 가장 큰 특징은 데이터 수집을 위해 각 로컬 디바이스에서 데이터를 중앙 서버로 직접 전송하여 서버에서 학습 하는 것이 아닌 각각의 로컬 디바이스에서 데이터를 학습하고 학습된 결과 모델만 중앙 서버로 전송한 뒤 중앙 서버에 수집된 각 로컬 디바이스의 모델을 집계하여 글로벌 모델을 만드는 학습 방식이다. 다음과 같은 방식은 모든 데이터를 서버로 전송하여 네트워크 트래픽과 서버 내 데이터 저장공간 비용이 증가하는 일반적인 인공지능 학습에 비해 통신 비용을 크게 절감함과 동시에 학습 데이터의 외부 유출을 피할 수 있다. 따라서 연합학습은 기존 인공지능 학습과 비교하여 데이터 프라이버시 보장과 통신 효율성 향상이라는 두 가지의 장점을 가지고 있어서 사용자의 데이터 보호가 중요하며 많은 사용자가 있어 통신의 효율성이 중요한 보이스 피싱 탐지에서 중요한 역할을 할 수 있다.

이러한 연합학습의 특징을 적용하여 본 논문에서는 사용자의 개인정보 보호에 특화된 보이스 피싱 탐지 알고리즘을 제안한다. 사용자의 개인정보가 포함된 데이터를 외부로 유출하지 않으면서 피싱 피해를 방지하기 위한 피싱 탐지 모델을 제공한다. 일반적인 중앙 집중식 모델 생성 기술과 달리 사용자의 정보를 외부로 유출하지 않고 온전히 클라이언트 내부에서만 사용하여 프라이버시를 보장함과 동시에 오토 라벨링 기법을 사용하여 자동으로 최신화된 데이터를 수집하고 학습하여 모델에 반영함으로써 피싱 탐지 모델의 성능을 높여 피싱 피해를 효과적으로 방지할 수 있다.

2. Related Works

보이스 피싱에 관련된 몇 가지 논문을 살펴보고자 한다.

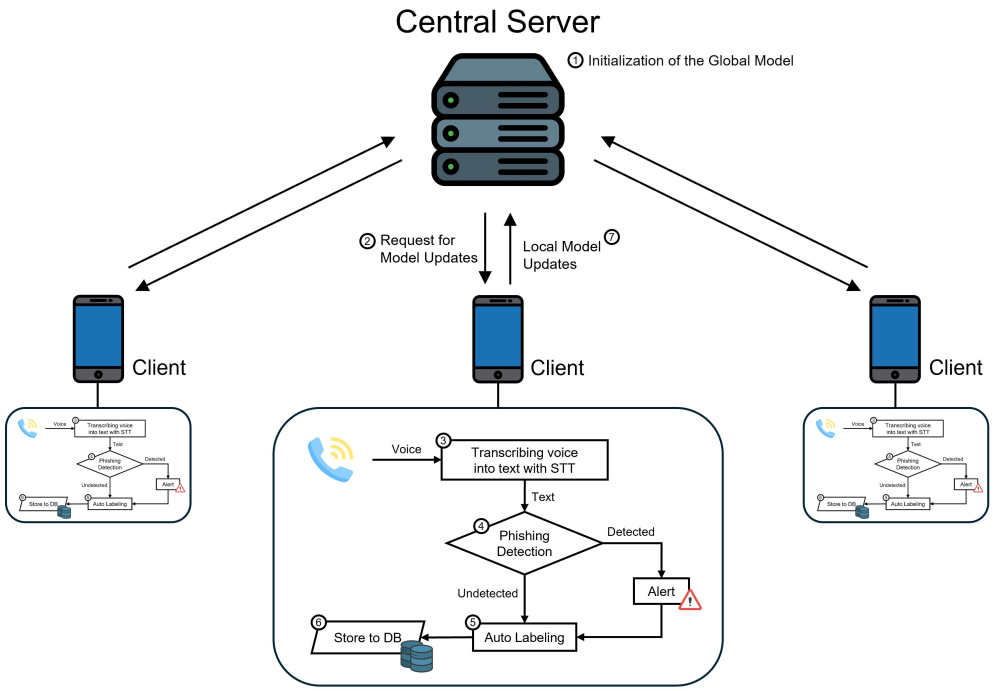
스팸 및 피싱 이메일 필터링에서 기계 학습의 적용 가능성을 다룬 논문이 있다[2]. 스팸과 피싱 이메일은 개인 정보 도용과 같은 심각한 보안 위협을 초래하며, 이를 효과적으로 탐지하기 위한 다양한 기계 학습 기법들이 연구되고 있다. 이 논문에서는 이메일의 본문, 제목, 발신자 주소, URL 등에서 추출한 특징들을 기반으로 스팸 및 피싱 이메일을 분류하는 여러 기계 학습 알고리즘을 비교하고, 높은 정확도를 달성하기 위해 중요한 특징 선택 방법을 설명한다. 나이브 베이즈, 서포트 벡터 머신, 랜덤 포레스트 등의 알고리즘이 사용되었으며, 최종적으로 99%의 정확도를 기록한 기법이 소개됩니다. 논문은 다양한 머신러닝 모델의 학습 및 일반화 가능성을 검토하고, 스팸 및 피싱 이메일 필터링 성능을 향상시키는 방법을 제시한다.

Basnet et al, 에서는 피싱 URL을 탐지하기 위한 머신러닝 기반의 방법론을 제안한다[3]. 저자들은 기존 블랙리스트 기반 방식의 한계를 극복하기 위해 URL의 텍스트적 특성, 키워드, 신뢰도 및 검색 엔진 기반의 피처들을 활용하여 피싱 URL을 분류하는 휴리스틱 접근법을 제시한다. PhishTank와 같은 데이터 소스에서 피싱 URL을 수집하고, Yahoo와 DMOZ에서 정상 URL을 수집한 후, 다양한 머신러닝 알고리즘(SVM, Random Forest 등)을 사용해 분류 성능을 비교한 결과, Random Forest가 가장 우수한 성능을 보였으며, 0.3% 이하의 낮은 에러율을 보여주고 있다. 이 연구는 실질적인 피싱 탐지 도구 개발에 기여할 수 있는 경량화된 접근법을 제시한다.

한국어 음성 피싱을 실시간으로 탐지하기 위한 머신러닝 기반 시스템을 제안하는 논문도 있다[4]. 본 논문에서는 실제 음성 피싱 데이터를 수집하고, 이를 텍스트로 변환한 후 자연어 처리 기술을 적용해 머신러닝 모델을 학습시켰다. 서포트 벡터 머신(SVM), 로지스틱 회귀(LR), 랜덤 포레스트(RF) 등의 모델을 사용한 결과, SVM이 가장 높은 정확도를 기록했고, LR은 가장 빠른 탐지 속도를 보여주고 있다. 이 연구는 복잡한 딥러닝 모델 대신 전통적인 머신러닝 기법 또한 실시간 탐지에 적합하며, 한국어처럼 리소스가 부족한 언어에서도 효과적으로 작동할 수 있음을 입증한다.

3. Proposed Method

본 논문에서는 보이스 피싱에서 문제되는 점 중 하나인 사용자 개인정보 유출 문제를 방지하는 보이스 피싱 탐지 알고리즘을 제안한다.



[그림 2] 사용자의 개인정보 보호에 특화된 보이스 피싱 탐지 알고리즘

본 논문에서는 수집된 데이터의 데이터 라벨링을 위해 스노클(Snorkel)[5]을 사용했다. 스노클(Snorkel)은 머신러닝 모델을 훈련할 때 대량의 데이터 라벨링을 수작업으로 진행하지 않고 자동으로 데이터를 라벨링할 수 있도록 도와준다. 이를 통해 텍스트 분류, 데이터 추출 등의 작업을 더 효율적으로 처리할 수 있도록 도와주며 본 논문에서는 새로운 보이스 피싱 데이터가 수집되었을 때 자동적으로 피싱인지 아닌지 라벨링을 하는 과정에 사용되었다.

[그림 2]는 연합 학습과 오토 라벨링을 사용한 피싱 탐지 시스템의 전체 구조도이다. 먼저 중앙 서버는 글로벌 모델을 초기화하고(①), 클라이언트들을 선택하여 글로벌 모델의 전달과 함께 모델 업데이트를 요청한다(②). 각 클라이언트는 전화를 수신할 경우 음성 인식(STT) 기술을 이용해 텍스트로 변환한다(③). 변환된 텍스트는 피싱 탐지 알고리즘을 통해 피싱 여부를 탐지하며(④), 만약 피싱으로 판단된다면 클라이언트의 사용자에게 알람을 표시한다(⑤). 그 후, 탐지 여부와 관계없이 모든 데이터는 자동으로 라벨링되어 클라이언트의 데이터베이스에 저장된다(⑥). 이와 같이 클라이언트에서 처리된 모델은 중앙 서버로 다시 전송되어(⑦) 글로벌 모델을 업데이트한다.

4. Experiments

실험에 사용된 데이터셋은 한국 보이스피싱 데이터를 모은 KorCCVi(Korean Call Content Vishing) 데이터[6]를 사용한다. KorCCVi는 경찰청에서 제공한 보이스 피싱 예시를 모아서 제작한 데이터셋으로 한국어 보이스 피싱의 특징을 모아 놓은 데이터셋이다. 실험은 일반적인 인공지능 학습과 본 논문이 제안하는 연합학습을 사용한 보이스 피싱 탐지 기술을 비교하는 방향으로 진행된다. 일반적인 인공지능 학습에서는 60 Epochs 동안 베치크기(Batch Size) 64로 바이너리 크로스엔트로피(Binary Crossentropy) 손실함수를 사용하여 진행하였고, 연합학습에서는 총 8개의 클라이언트에 데이터를 균일하게 나누어 학습에 참여시키고 각각 나누어진 데이터에서 배치크기를 64로 2 Epochs 동안 학습을 진행, 이를 30라운드에 걸쳐 학습하는 것으로 일반 인공지능과 최대한 비슷한 환경에서 학습을 진행하였다.

다음 [표 1]은 일반적인 인공지능 학습을 적용한 보이스 피싱 탐지 기술(Central Model)과 본 논문이 제안하는 연합학습을 사용한 보이스 피싱 탐지 기술

Model	Accuracy	Precision	F1-Score	Recall	BCELoss
Central Model	97.35%	78.24%	77.49%	76.82%	0.6156
Fed-Avg	96.60%	80.40%	79.35%	78.50%	0.6187
Fed-Prox	96.57	78.22	77.17	76.31	0.6189

[표 1] 인공지능 학습 기술과 제안된 기술의 보이스 피싱 탐지 성능 비교

(Fed-Avg[1], Fed-Prox[7])의 성능을 비교한 내용이다. [표 1]에 따르면 Central Model은 정확도(97.35%)와 BCE 손실값(0.6156)에서 가장 우수한 성능을 보였지만, Fed-Avg는 정밀도(80.40%), 재현율(78.50%), F1-Score(79.35%)에서 더 나은 성능을 보여준다. Fed-Prox는 모든 지표에서 다른 모델들보다 약간 낮은 성능을 보였으며, 특히 Fed-Avg는 긍정 클래스 예측 성능(정밀도와 재현율의 균형, 해당 모델이 얼마나 잘 피싱을 탐지하는지에 대한 성능)에서 가장 뛰어난 모델로 평가된다. 일반적으로는 정확도가 높은 Central Model이 더 좋은 성능을 보일 수 있지만, 본 논문이 제안하는 연합학습을 사용한 보이스 피싱 탐지 알고리즘이 긍정 클래스 예측 성능이 뛰어난, 즉 피싱을 좀더 잘 찾아내며 데이터 보호까지 보장되기 때문에 충분히 더 좋은 성능을 보이고 있음을 알 수 있다.

5. Conclusion

본 논문은 연합학습을 사용한 데이터 보호 보이스 피싱 탐지 기술을 제안한다. 이 기술은 연합학습 알고리즘을 사용하였기 때문에 사용자의 데이터 프라이버시를 보호함에 있어 우수하다. 이에 맞물려 개인정보가 담겨있는 민감한 데이터셋을 활용하는데 용이하다. 그리고 스노클을 활용한 오토 데이터 라벨링을 통하여 새로운 데이터를 재빠르게 적용할 수 있으며 모델 학습에 활용하여 보다 빠르게 최신 데이터를 적용한 모델을 생성할 수 있다.

5. Acknowledgement

본 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022R1A2C4001270). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2024-RS-2020-0-01602).

참고문헌

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 20-22 Apr, 2017.

[2] Gangavarapu, Tushaar, C. D. Jaidhar, and Bhabesh Chanduka. "Applicability of machine learning in spam and phishing email filtering: review and approaches." Artificial Intelligence Review 53.7 (2020): 5019-5081.

[3] Basnet, Ram B., and Tenzin Doleck. "Towards developing a tool to detect phishing URLs: A machine learning approach." 2015 IEEE International Conference on Computational Intelligence & Communication Technology. IEEE, 2015.

[4] Lee, Minyoung, and Eunil Park. "Real-time Korean voice phishing detection based on machine learning approaches." Journal of Ambient Intelligence and Humanized Computing 14.7 (2023): 8173-8184.

[5] *Programmatically Build Training Data.* n,d. *Snorkel.* https://www.snorkel.org/

[6] Milandu K. M. B. and D.J. Park, "A Real-time Efficient Detection Technique of Voice Phishing with AI," Proceedings of the Korean Information Science Society Conference, pp. 768 - 770, 2021.

[7] Li, Tian, et al. "Federated optimization in heterogeneous networks." Proceedings of Machine learning and systems 2 (2020): 429-450.