대화패턴 분석을 이용한 메신저 피싱 개인화 예방시스템

박준희O 박상근
경희대학교 소프트웨어융합학과
pjh1313@khu.ac.kr, sk.park@khu.ac.kr

Messenger Phishing Personalized Prevention System Using Conversation Pattern Analysis

Junhee ParkO Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

2022년 보이스 피싱 발생 건수의 89%를 메신저 피싱이 차지하였고, 지인을 사칭하며 금전적인 도움을 요구하는 유형의 메신저 피싱 피해액은 최근 2년간 554억 원 증가하였다. 본 연구에서는 피싱범이 타인을 사칭해서 발생하는 메신저 피싱을 예방하기 위해 기존의 메신저 피싱 예방 방법과는 달리 개개인에게 맞춤화된 예방 방법을 제안한다. 존댓말/반말 분류 모델과 지칭 유사 단어 사전을 활용하여 메신저 피싱 개인화 탐지 모델을 개발하였고, 메신저 피싱 시나리오의 사용자 스터디를 수행하여 본 모델의 효과성을 입증하였다. 본 모델을 적용한 메신저 웹 서비스 프로토타입을 통해 해당 모델의 활용 가능성을 검증함으로써 본 연구 결과가 지인을 사칭한 메신저 피싱 범죄 예방에 기여할 수 있을 것으로 기대한다.

1. 서 론

피싱(phishing)은 프라이빗 데이터(private data)와 피싱(fishing)의 합성어로, 온라인에서 믿을만한 사람 또는 기관을 사칭하면서 특정 대상인의 정보를 부정하게 얻으려는 행위를 말한다. 전화로 타인을 속여 금전을 요구하는 보이스 피싱(Voice phishing)과 타인의 메신저 아이디를 도용하여 금전을 요구하는 메신저 피싱(Messenger phishing)이 대표적인 예이다.

피싱에 의한 피해액도 상당하다. 2018년부터 2022년까지 최근 5년간 보이스 피싱으로 입은 피해 금액이 1조 7천억 원에 이른다고 한다. 유형별로는 대출 빙자 보이스 피싱 피해액이 9,998억 원(60.1%)으로 과반을 차지했고, 기관 사칭이 3,799억 원(22.8%), 메신저 피싱(지인 사칭)이 2,849억 원(17.1%)으로 그 뒤를 이었다. 지인을 사칭한 메신저 피싱 피해액은 상대적으로 적은 수치이지만, 2020년까지만 해도 373억 원 수준에 머물렀던 피해액이 2022년에는 927억 원으로 급증하였다. 건수 기준으로는 2022년 전체 보이스 피싱 발생 건수의 89%(25,534건)를 메신저 피싱이 차지하며 보이스 피싱 피해 10건 중 9건이 메신저 피싱으로 인한 피해였다 [1].

대출 빙자나 기관 사칭과 같은 피싱은 패턴이 비슷하므로 사기라는 것을 인지하기 상대적으로 쉽지만, 가족이나 지인을 사칭하여 돈을 요구하는 메신저 피싱은 일상적인 말투와 소재로 접근하기 때문에 가족에게 정말 급한 일이 생긴 것을 아닐지 걱정되어 당황하게 되고, 이에 따라 사기라는 것을 빠르게 알아채지 못하는 경우가 많다.

피싱 범죄에서는 대부분 피해액을 돌려받지 못하고 끝나는 경우가 많으므로 금융위원회에서 안전 안내 문자를 발송하는 등의 메신저 피싱 예방에 노력을 기울이고 있으나, 개인의 상황을 고려하여 발생하는 메신저 피싱을 예방하는 데는 한계가 있다 [2].

본 연구에서는 메신저 피싱 상황을 빠르게 알아챌 수 있도록 존댓말/반말 분류 모델과 지칭 관련 단어 유사어 사전을 활용하여 상대방의 대화 패턴이 평소의 패턴과 유사한지 분석해 주는 대화 패턴 분석 모델을 제안한다. 제안한 모델을 활용해 메신저 피싱 시나리오의 사용자 스터디를 수행하여 본 모델이 메신저 피싱을 효과적으로 예방할 수 있음을 확인했다. 그리고 해당 모델을 적용한 메신저 웹 서비스 프로토타입을 개발해 모델의 활용 가능성을 검증하였다.

2. 관련 연구

IT 기술을 활용해 보이스 피싱 및 메신저 피싱을 예방하기 위한 다양한 연구가 수행되었다. 김상국 et al. [3]은 메신저 피싱을 예방하기 위해 URL 문자열을 그래프로 구축했다. 그리고 그래프 추론 알고리즘과 그래프 임베딩 기법을 통해 URL이 피싱을 위한 URL인지 높은 정확도로 예측할 수 있음을 확인하였다.

양지훈 et al. [4]는 딥러닝 기반의 KoBERT를 이용하여 실제 보이스 피싱 통화 음성과 일반 통화 음성을 학습한 탐지 모델을 개발했으며, 채수열 et al. [5]은 469개의 메신저 피싱 대화 내용을 분석해 피싱 범죄로 의

심되는 대화를 탐지하는 연구를 수행했다. 하지만 기존의 연구들은 피싱의 패턴을 분석하고, 이를 기준으로 새로운 입력이 주어졌을 때 피싱인지 아닌지 판단하기 때문에 기존의 피싱 패턴과 다른 새로운 피싱 방법을 사용하면 탐지하기 어렵다는 한계가 존재한다.

본 연구에서는 메신저 피싱에서 대화 상대의 기존 대화 패턴과 현재 대화 패턴을 비교하는 방법으로 대화 상대의 사칭범 가능성을 유추하는 새로운 접근 방법을 고안하였다. 이를 위해, 존댓말/반말 사용과 유사 단어 사전을 활용해서 개인화된 대화 패턴 기반의 메신저 피싱 개인화 예방 시스템을 개발하고 그 효과를 검증한다. 본 연구의 결과물은 기존의 메신저 피싱 예방 방법과는 달리 개개인에게 맞춤화될 수 있으므로, 획일화된 경고 메시지에 비해 보다 큰 경고 효과를 거둘 수 있어 메신저 피싱 범죄 피해 예방에 도움 될 수 있을 것으로 기대된다.

3. 메신저 피싱 개인화 탐지 모델

사람들은 평소 말할 때나 메신저를 통해 채팅할 때, 각기 다른 자신만의 어투를 사용한다. 상대방의 어투가 평소 어투와 다르다면, 다른 사람이 상대방을 사칭하고 있을 가능성이 크다. 본 연구에서는 먼저 어투를 구분할 수 있는 다양한 요소 중 텍스트 정보에서 추출할 수 있는 존댓말/반말 분류와 지칭 유사 단어 사전을 활용해서 개인화 대화 패턴을 분석하는 모델을 개발했다.

3.1. 존댓말/반말 분류 모델

먼저, 존댓말과 반말을 분류하기 위해 스마일게이트 말투 데이터 세트와 AI 허브의 감성 대화 말뭉치를 데이터셋으로 사용한 한국어 존댓말/반말 분류기 모델[6]을 활용했다. 해당 모델은 사전 학습 모델로 KcBERT를 사용한 딥러닝 모델로, 형태소분석기에 의존한 기존의 존댓말/반말 분류보다 높은 성능을 보여준다. 본 모델을 활용하면 사용자가 대화를 입력할 때마다 해당 텍스트의 반말과 존댓말을 분류해 주며 해당 분류의 정확도까지 확인할 수 있다. 이를 활용하면, 존댓말을 쓰던 사람이 갑자기 존댓말 확률이 현저히 낮은 텍스트를 입력하거나, 반말을 쓰던 사람이 존댓말 확률이 높은 텍스트를 사용할 때 이를 빠르고 정확하게 감지할 수 있다.

3.2. 지칭 단어 유사어 사전 생성

표 1. 지칭 관련 유사 단어 예

아버지	어머니	언니
아빠	엄마	언닝
아빵	엄망	온니
아부지	엄니	옹니
아부징	어무니	시스타
	어무닝	씨스타

사람을 지칭할 때 사용하는 지칭 단어를 선정하고, 해당 지칭 단어와 동일한 의미를 갖는 유사한 단어를 찾아 직접 지칭 단어 유사어 사전을 만들었다 [표 1]. 이를 활용하면, 대화 상대가 평소에 사용하던 지칭 단어와 다른 단어를 사용하는 것을 알아채고 사칭범이 상대를 대신해서 대화하고 있을 가능성이 있다고 판단할 수 있다. 예를 들어, “아버지”를 지칭하는 단어는 “아빠”, “아빵”, “아부지”, “아부징” 등이 있다. 나를 평소에 아버지라고 부르던 사람이 갑자기 이와 동일한 의미의 아빠, 아부지 등의 지칭 단어를 사용하면 이러한 이상 패턴을 감지하고 피싱을 의심할 수 있다.

4. 사용자 스터디

4.1. 사용자 스터디 방법

저자 중 한 명이 평소에 많은 대화를 나누는 4명의 참여자(아버지, 어머니, 언니, 지인)를 선정하고, 메신저 피싱범을 가장하여 작성한 시나리오([표 2] 참고)를 기반으로 카카오톡을 통해 참여자들에게 메시지를 보냈다. 참여자가 피싱 시나리오임을 눈치챌 때까지 대화를 진행했으며, 실험이 종료된 후 피싱 시나리오 실험이었음을 밝히고 메신저 피싱 경험에 대한 인터뷰를 진행했다. 모든 참여자는 실제로 계좌번호에 돈을 입금하기 전에 피싱을 눈치채고 실험자에게 전화했다.

4.2. 사용자 스터디 결과

스터디가 종료된 후 각 참여자에게 어떠한 요소로부터 메신저 피싱을 의심하였는지 인터뷰를 진행했다. 참여자 3명은 평소와 다른 존댓말/반말 사용을 보고 피싱임을 의심하기 시작했고, 참여자 2명은 과거에 사용하지 않았던 지칭 단어 사용(예: 평소에 아부지라고 했으나 이번에 아빠라고 부름)을 통해 피싱을 의심했다고 응답했다. 이를 통해, 본 논문에서 메신저 피싱 탐지를 위해 존댓말/반말 분류 및 지칭 단어 유사어 사전이 메신저 피싱 탐지에 유효할 수 있음을 확인했다.

참여자와 한 실험자의 최근 대화 기록을 가져와 분석한 결과, 참여자는 실험자에게 항상 존댓말을 사용했으며, 지칭 단어 유사어 사전에 저장된 단어 중, “아부지”라는 단어만 사용했다는 것을 확인했다. 그리고 메신저 피싱 개인화 탐지 모델을 활용해서, 해당 참여자와 진행한 메신저 피싱 시나리오 사용자 스터디에서의 대화를 분석했다 [표 2].

표 2. 사용자 스터디: 아버지 대상 메신저 피싱 시나리오

화자	대화 내용	반말 확률	아버지 유사어
피싱범	아빠	100.0%	Y
참여자	왜ㅡㅡ	-	-
피싱범	뭐해?	100.0%	-
참여자	일한다ㅡ	-	-
피싱범	졸업사진 촬영비 입금해야 되는데	100.0%	-
피싱범	핸드폰이 꺼져서 못하고 있어	100.0%	-
참여자	언제까지 보내야 되는데?	-	-
피싱범	지금 빨리 보내야 돼	100.0%	-
피싱범	아빠가 먼저 입금해주면 안될까?	100.0%	Y
참여자	알았다ㅡ	-	-
피싱범	미안해	100.0%	-
참여자	카톡은 뭘로 하는거냐ㅡ?	-	-
피싱범	피씨방와서 급하게 연락하는 거야	100.0%	-
피싱범	김ㅇㅇ XX은행 12-**-345 15만 5천원이야 폰 켜지면 바로 줄게	100.0%	-

5. 메신저 피싱 탐지를 위한 웹 서비스 프로토타입

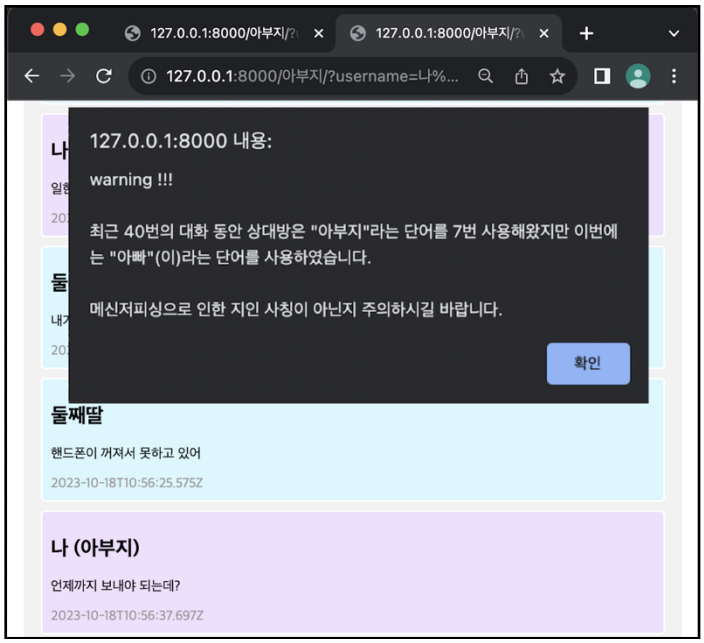

그림 1. 메신저 피싱 탐지를 위한 웹 서비스 프로토타입

위 사용자 스터디를 기반으로, 해당 모델을 적용한 실제 텍스트 메신저에서 피싱을 감지하고 경고할 수 있는 웹 서비스 프로토타입을 개발했다 [그림 1]. 해당 프로토타입 개발을 위해서 Python과 Django 웹 프레임워크를 사용했다.

해당 웹 서비스 프로토타입은 대화를 주고받을 때마다 상대방 메시지의 존댓말/반말 분류 결과 및 해당 분류의 정확도와 지칭 단어 유사어 사전에 포함된 단어 사용 여부를 SQLite3 데이터베이스에 저장한다. 이를 통해, 최근의 대화 패턴과 현재 대화 패턴을 비교하면서 존댓말/반말 사용에 차이가 있거나 평소와 다른 지칭 단어를 사용하는 경우를 빠르게 감지할 수 있다. 단순 피싱 경고가 아니라, [그림 1]과 같이 개인화된 피싱 감지 근거와 함께 피싱 감지 경고 팝업을 띄울 수 있기 때문에 보다 강한 경각심을 전달함으로써 피싱 피해를 막을 수 있다.

6. 결론 및 제언

본 연구에서는 대화 패턴 분석을 이용한 메신저 피싱 개인화 예방 시스템을 제안하였다. 존댓말/반말 분류 모델과 지칭 단어 유사어 사전을 활용하여 어투를 구분하고, 평소 대화의 어투 패턴이 달라지면 메신저 피싱을 감지할 수 있다. 사용자 스터디를 통해 존댓말/반말 분류와 지칭 단어 유사어 사전이 실제 피싱 탐지에 유용하게 활용될 수 있음을 확인하였으며, 이를 기반으로 메신저 피싱 탐지를 위한 웹 서비스 프로토타입을 개발하여 본 논문에서 제안한 개인화 된 피싱 감지의 활용 가능성을 확인했다. 본 연구에서는 실제 메신저에 해당 모델을 적용할 수 없어서 프로토타입으로 그 가능성을 확인하는 데 그쳤으나, 향후 텔레그램이나 슬랙(Slack) 등 API가 잘 제공되고 있는 메신저 서비스와 연동하면 온라인 메신저에서의 피싱을 효과적으로 막을 수 있을 것으로 기대된다.

참 고 문 헌

[1] 김두환, “"메신저 피싱 급증"...5년간 보이스 피싱 피해액 1조7천억원”, 청년일보, 2023.02.21.
[2] 본청 사이버안전 사이버수사과, “가족 또는 지인 사칭해 개인정보와 돈을 요구하는 메신저 피싱 근절. 위해 관계기관 힘 모아”, 경찰청 보도자료, 2020.06.24.
[3] 김태리·홍지원·박노성·김상욱 “그래프 기반의 피싱. URL 탐지" 한국정보과학회, 26(3), 156-160, 2020.
[4] 양지훈·이충훈·김성백, “KoBERT 기반의 통화내용. 분석을 통한 보이스 피싱 예방 서비스 개발 및 활용”, 한국정보과학회, 29(5), 205-213, 2023.
[5] 배지효·채수열·송명준 “메신저 피싱 대화 분석을 통한 탐지 방안", 한국통신학회, 157, 537-538, 2019.
[6] https://github.com/jongmin-oh/korean-formal-classifier