Open STT API와 머신러닝을 이용한 AI 보이스피싱 예방 솔루션

모시은[1], 양혜인[1], 조은비[2], 윤종호[3]
[1]을지대학교 의료it학과
[2]금오공과대학교 전자통신과
[3]코즈비즈 시스템즈
ahtlse0917@naver.com, helen6662@naver.com, reenee21@kumoh.ac.kr, hagis@naver.com

AI voice phishing prevention solution using Open STT API and machine learning

Shi-eun Mo[1], Hye-in Yang[1], Eun-bi Cho[2], Jong-Ho Yoon[3]
[1]Dept. of Medical IT, Eul-ji University
[2]Dept. of Electronic Communication, Kumoh National Institute of Technology
[3]COZBIZ Systems

요 약

본 논문은 보이스피싱에 취약한 VoIP와 일반 유선전화 상의 보안을 위해 유선전화의 대화내용을 Google STT API 및 텍스트 자연어 처리를 통해 실시간으로 보이스피싱 위험도를 알 수 있는 모델을 제안했다. 보이스피싱 데이터를 Data Augmentation와 BERT 모델을 활용해 보이스피싱을 예방하는 솔루션을 구상했다.

1. 서론

최근 코로나19로 인해 정부의 긴급 재난지원금과 같은 공공기관을 사칭한 보이스피싱 수법이 다양하고 더욱 교묘한 방법으로 진화하고 있다. 또한 비대면 사회의 등장으로 금융사기, 사칭피해 사례가 증가하고 있다. 이와 관련해 보이스피싱 피해를 사전에 예방하기 위해 적극적으로 대응책을 마련해야 한다.

보이스피싱 발생건수는 경제활동 위축 및 코로나19 발생 이후 줄어들었지만 1건당 피해금액은 늘어났다. 2020년은 7000천억 원, 2021년 상반기는 4천 3백억원으로 전체 피해액이 1조 원 대에 이를 것이라고 예상되었다.[1]

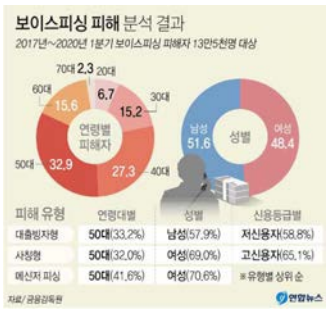
구분	발생건수	피해금액	검거건수	검거인원
'17년	24,259	2,470	19,618	25,473
'18년	34,132	4,040	29,952	37,624
'19년	37,667	6,398	39,278	48,713
'20년	31,681	7,000	34,051	39,324
'20년 상반기	16,050	3,298	17,257	20,192
'21년 상반기	17,814	4,351	13,331	12,421

출처 : 경찰청

(그림 1) 경찰청 집계 보이스피싱 연도별 발생 및 검거 현황(단위: 건, 억 원, 명)

스마트폰 분야에서는 이미 AI로 통화내용을 분석해 전화금융사기(보이스피싱)를 잡아내는 어플리케이션이 .출시되어 현재 사용되고 있으나 VoIP(인터넷 전화) 또는 일반 유선전화에는 보이스피싱을 예방하기 위한 하드웨어와 소프트웨어가 없는 실정이다. 보이스피싱은 스마트폰, 음성 이메일, 유선전화 등을 통해 이루어지는데, 이중 VoIP가 사용되었을 때 추적이 어렵다. 심지어 일반 유선전화를 사용하는 대상은 대부분 어르신이고 연세로 인해 순간 판단력이 낮아져 보이스피싱에 쉽게 피해를 입을 수 있다.

금융감독원에 의하면 2017년부터 2020년 1분기까지 보이스피싱을 당한 피해자 13만5천명 중 피해자를 연령별로 살펴보면 50대(32.9%)가 가장 많았다. 40대(27.3%), 60대(15.6%)가 뒤를 이었고 신용등급이 낮을수록 대출 빙자형 피해에 취약했다.[2]



(그림 2) 보이스피싱 피해 분석 결과

이러한 보이스피싱 취약계층을 위한 저렴한 비용으로 대화내용을 실시간으로 분석해 보이스피싱 여부

를 가려내기 위한 시스템의 개발이 필요하다고 판단되어 Open STT API와 머신러닝을 이용한 AI 보이스피싱 예방 솔루션을 구상해 보았다.

2. 관련 연구

2.1 Google STT

Google STT는 사람의 음성 인터페이스를 통해 텍스트 데이터(문자)를 추출해내는 것이다. 이는 크게 음성 데이터와 언어 데이터로부터 인식 네트워크 모델을 생성하는 오프라인 학습단계와 사용자가 발성한 음성인식하는 온라인 탐색 단계로 구분된다. STT엔진은 음성 및 언어 데이터에 대한 사전지식을 사용해 음성신호로부터 문자정보를 출력하게 된다. 이때 STT알고리즘의 해석이라는 차원을 디코더(Decoder)라고도 부른다.

디코딩 단계에서는 학습 단계 결과인 음향 모델(Acoustic Model), 언어 모델(Language Model)과 발음 사전(Pronunciation Lexicon)을 이용하여 입력된 특징 벡터를 모델과 비교, 스코어링(Scoring)하여 단어 열을 최종 결정 짓는다.

2.2 EDA (Exploratory Data Analysis, 탐색적 데이터 분석)

EDA는 벨연구소의 미국인 수학자 '존 튜키'가 개발한 데이터 분석 방법론에 대한 개념으로 데이터를 분석하고 결과를 내는 과정에 있어서 지속적으로 해당 데이터에 대한 '탐색과 이해'를 기본으로 가져야 한다는 것을 의미한다. 이는 오늘날 데이터 감지 프로세스에서 지속적으로 널리 사용되고 있는 방법으로 가설을 세우고 그 가설을 검정하는 기존의 통계학에서의 문제점을 주어진 자료만으로 충분한 데이터가 될 수 있도록 하는 과정으로 기존에 보유한 데이터를 단어 단위로 교체, 삽입, 위치변경, 삭제해 데이터 양의 증량으로 데이터 모델학습 정확도 향상에 효과적이다. EDA의 주요 목적은 어떠한 가정을 하기 전에 데이터를 살펴볼 수 있도록 지원하는 것으로 오류 식별 및 데이터 내의 패턴을 파악하고 아웃라이어 또는 이례적 이벤트를 감지해 변수들 간의 관계파악을 돕는다. 이는 주로 Python과 R을 개발언어로 사용하고 특히 Python과 EDA를 함께 사용하면 데이터 세트의 누락된 값을 식별하고 머신러닝에서 누락된 값을 처리하는 방법을 결정할 수 있도록 하기 때문에 매우 중요하다. [3]

2.3 BERT

BERT는 2018년 구글이 공개한 사전 훈련된 모델로 등장과 함께 수많은 NLP 태스크에서 최고 성능을 보여주었다. BERT는 레이블이 없는 데이터로 사전 훈련을 시켜 레이블이 있는 다른 작업에서 추가 훈련을 하면서 하이퍼파라미터를 재조정하여 이 모델을 사용하면 높은 성능을 얻는 기존의 사례들을 참고했다. 이를 통해 BERT는 높은 성능을 얻을 수 있었다. 다른 작업에 대해서 파라미터 재조정을 위한 추가 훈련 과정을 파인 튜닝(Fine-tuning)이라고 한다.[4]

BERT는 3가지 특징을 가진다. 첫 번째, 문맥의 의존한 단어 벡터 표현을 만들 수 있다. 두 번째, 파인 튜닝이 자연어 처리 작업으로 가능해졌다. 세 번째, Attention을 활용한 설명과 시각화가 쉽다. [5]

3. 제안 모델

본 논문에서는 딥러닝을 이용한 보이스피싱 여부를 판별하는 모델을 제안한다.

(그림 3) 보이스피싱 여부 판별 흐름도

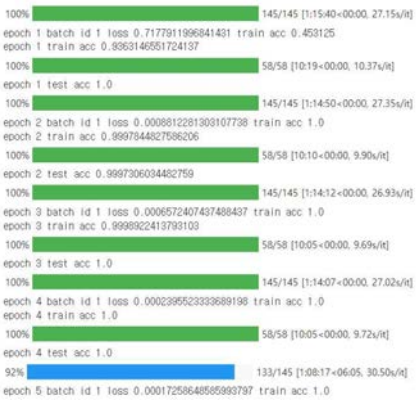
그림 3은 전체적인 흐름을 도식화 한 것이다. 먼저 보이스피싱 과정에서는 빈번히 나타나는 특정 단어가 존재한다. 주로 받을 수 있는 혜택이 존재한다며 개인정보 유출을 유도하거나, 공공기관의 일원을 사칭하며 금융사기에 연루되었다는 내용의 통화를 진행한다. 이러한 요소를 통하여 보이스피싱 여부에 대하여 판별한 결과값에 가중치를 부여한다.

우선 일정시간 통화가 진행되면, 진행 동안에 STT로 보이스피싱 조직원의 음성 데이터를 수집한다. 수집된 데이터를 보이스피싱 시나리오를 이용하여 학습된 딥러닝 모델의 입력값으로 사용하여 해당 통화가 보이스피싱인지에 대한 여부를 판별하여 사용자에게 알린다. 그 후 STT API를 이용하여 VoIP STT 녹취 저장 및 분석 모듈을 이용해 분석함으로 최신 보이스피싱 데이터셋을 보이스피싱 탐지 시스템에 제공한다.

3.1 보이스피싱 탐지시스템

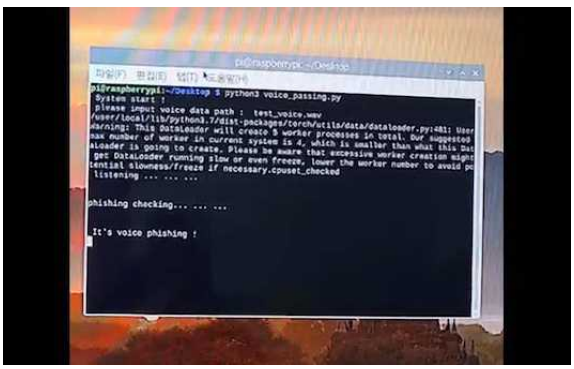
보이스피싱 탐지 시스템은 학습된 인공지능 모델이다. 학습을 위해 수집한 보이스피싱 데이터의 Data Augmentation으로 EDA를 사용한다. 머신러닝 학습을 위해 BERT 모델을 사용해 정확도 높은 결과를 기대한다. 본 모델은 한국어 특화 모델인 KoBERT를 사용하여 단어를 임베딩하고, Softmax 함수를 활용한 Binary Text classification으로 보이스피싱임을 판별한다. 이를 라즈베리파이 및 젯슨나노를 통해 LED 경고등을 띄워 보이스피싱 여부를 알려준다.

3.2 구현



(그림 4) 보이스피싱 탐지 모델 학습

위 그림은 제안한 모델을 기반으로 보이스피싱 탐지 모델을 학습시키는 과정이다. 학습데이터로는 경찰청에 제공되어 있는 보이스피싱 사례들과 공개되어 있는 상담전화데이터를 사용했다. 전체 데이터셋에서 7:3으로 학습데이터와 테스트데이터를 나누어서 학습을 진행하였다. 학습 과정에서 각 에폭마다 test set을 활용하여 탐지 결과를 보여주는데, 3에폭부터 train set과 test set에 대해서 1.0의 정확도를 보여준다. 이것은 탐지 모델이 보이스피싱을 제대로 판별한다는 의미이다.



(그림 5) 보이스피싱 제품 프로토 타입

그림 5는 제안 시스템의 프로토타입이다. 실제로 보이스피싱 음성이 들어갔을 때, “it‘s voice phishing”을 화면에 표시하며, 보이스피싱을 판별해주는 것을 볼 수 있다.

4. 결론

본 논문은 보이스피싱에 취약한 VoIP와 일반 유선전화 상의 보안을 위해 유선전화의 대화내용을 Google STT API 및 텍스트 자연어 처리를 통해 실시간으로 보이스피싱 위험도를 알 수 있는 모델을 제안한다. 본 모델은 학습할 데이터 양을 Data Augmentation을 이용해 늘리고, 2018년 공개되어 다양한 NLP 태스크에서 좋은 성능을 보여준 BERT를 사용해 학습의 정확도를 높이려고 한다. 보이스피싱으로 인한 피해금액이 커지는 현재, 본 논문에서 제안한 모델이 범죄 예방에 도움이 되기를 기대하며 추후 VoIP와 유선전화 Sniffing과 연결해 시스템을 구현할 계획이다.

- 본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다 -

참고문헌

[1] 송석배, 보이스피싱 범죄, 2021년 피해 다시 증가 추세 !, 열린뉴스., 2021,11,14, http://www.yeollinnews.co.kr/news/articleView.html?idxno=21593

[2] 이재윤, [그래픽] 보이스피싱 피해 분석 결과, 연합뉴스, 2020.08.10., https://www.yna.co.kr/view/GYH20200810002500044,2020.08.10

[3] IBM, 탐색형 데이터 분석, https://www.ibm.com/kr-ko/cloud/learn/exploratory-data-analysis, 2022.09.15.

[4] 딥러닝을 이용한 자연어 처리 입문, 02) 버트(Bidirectional Encoder Representations from Transformers, BERT), https://wikidocs.net/115055, 2022.09.15.

[5] 오가와 유타로, 만들면서 배우는 파이토치 딥러닝, 서울, 한빛미디어, 2021.

[6] 이승아, “보이스피싱에 대한 텍스트언어학적 연구.” 텍스트언어학 45 (2018): 177-195.