NLP 알고리즘을 활용한 A.I 보이스피싱 탐지 솔루션

김태경[1], 박은주[2], 박지원[3] ,한아림[4]
[1]동국대학교 산업시스템공학과 학부생
[2]이화여자대학교 수학과 학부생
[3]숙명여자대학교 통계학과 학부생
[4]서울여자대학교 정보보호학과 학부생
wendy1080@dgu.ac.kr, pej0918@ewhain.net, orilove1@sookmyung.ac.kr, convertme@swu.ac.kr

A.I voice phishing detection solution using NLP Algorithms

Tae-Kyung Kim[1], Eun-Ju Park[2], Ji-Won Park[3], A-Lim Han[4]
[1]Dept. of Industrial Systems Engineering, Dongguk University
[2]Dept. of Mathematics, Ewha Womans University
[3]Dept of Statistics, Sookmyung Women's University
[4]Dept. of Information Security, Seoul Women's University

요 약

본 논문은 디지털 소외계층과 사회적 약자를 고려한 보이스피싱 예방 솔루션을 제안한다. 통화 내용을 AWS Transcribe를 활용한 STT와 NLP 알고리즘을 사용해 실시간으로 보이스피싱 위험도를 파악하고 결과를 사용자에게 전달하도록 한다. NLP 알고리즘은 KoBIGBIRD와 DeBERTa 모델 각각을 커스터마이즈하여 보이스피싱 탐지에 적절하게 파인튜닝 했다. 이후, 성능과 인퍼런스를 비교하여 더 좋은 성능을 보인 KoBIGBIRD 모델로 보이스피싱 탐지를 수행한다.

1. Introduction

과거 불특정 다수를 향한 보이스피싱 사례들이 많았다면, 최근에는 피해자 정보 기반의 맞춤형 보이스피싱 시나리오를 이용한 범죄 사례들이 증가하고 있다. 이런 상황에서 VoIP 및 유선전화 사용자를 포함한 디지털 소외계층과 보이스피싱에 취약한 사회적 약자를 고려한 서비스가 필요하다. 이에, 본 연구에서는 KoBIGBIRD, DeBERTa 등 우수한 PLM 모델들을 테스크에 맞게 커스텀 및 비교하여 대화의 문맥을 파악해 다양한 피싱 시나리오에서도 피싱 여부를 정확하게 탐지하는 솔루션을 제안한다. 더 나아가 디지털 소외계층과 사회적 약자들도 해당 서비스를 쉽게 활용할 수 있도록 라즈베리 파이를 활용한 LED 알림과 반응형 웹으로 결과물을 구현했다.

2. Data processing

2.1 Data Acquisition

금융감독원 홈페이지의 '그놈 목소리' 카테고리 속 피싱 음성 데이터 MP3 파일을 동적 크롤링을 활용하여 수집한 후, 해당 MP3 파일들을 AWS Transcribe를 활용해 text 형식으로 변환하였다. 피싱이 아닌, '정상' 데이터는 AI Hub의 민원(콜센터) 질의응답 데이터 중 금융/보험, 이체/출금, 대출 서비스 항목 데이터를 사용했다. 이때 피싱 데이터는 370개, 정상 데이터는 9698개가 수집되었으나 학습이 어려울 만큼 양이 부족하고, 클래스 간 균형이 맞지 않다고 판단하여 증강을 진행하였다.

2.2 Data Augmentation

크게 단어 삭제 및 추가, 문장 재구성, 번역 후 회귀 등 4가지 기법을 사용하였고, 결과적으로 피싱 데이터 15,612개, 정상 데이터 개수 58,187개로 이뤄진, 총 73,799개의 학습 데이터 셋을 구성했다.

3. Model

3-1. Customized KoBIGBIRD

해당 모델은 R-BERT 모델로부터 영감을 얻어 KoBIGBIRD 모델의 Architecture를 수정한 모델이다. 본 모델은 Kr-BERT의 CLS 토큰과 KoBIGBIRD의 CLS 토큰의 임베딩 값을 결합한다. [1] 그다음, 전체 대화 데이터와 키워드 및 형태소만을 추출한 데이터를 벡터로 분리한다. 이 과정에서 문장의 끝을 나타내는 인덱스를 사용했다. 이는 전체 대화 데이터를 통해 대화의 문맥을 이해하고, 형태소 및 키워드를 통해 대화에서 중요한 부분을 학습하고자 설계된 부분이다. 해당 과정을 걸쳐 Kr-BERT의 언어 이해 능력과 KoBIGBIRD의 장문 처리 능력을 결합함으로써 입력 텍스트의 다양한 특징을 포함하고, 긴 대화 데이터를 다루는 데 있어

KoBIGBIRD의 장점을, [2] 문장 내의 미묘한 의미나 표현을 이해하는 데 Kr-BERT의 장점을 활용할 수 있다.

3-2. DeBERTa

DeBERTa는 "Disentangled" 어텐션 메커니즘을 사용하여 단어 간의 상호 작용을 더욱 세분화하여 이전 단어와의 연관성을 강조하면서도 다음 단어에 대한 정보도 제대로 포착할 수 있도록 한다. 이는 효과적인 문맥 파악을 가능하게 해준다는 장점이 있다. 보이스피싱 상황은 현재 진행되는 말의 문맥을 정확히 파악해야 한다. 해당 모델은 Enhanced mask decoder BERT 부분을 사용하여 불필요한 정보의 누출을 방지하고 단어의 절대적 위치 정보를 고려할 수 있다. 이로써 각 단어가 문장에서 어떤 역할을 하는지 파악하여 더 정확한 디코딩을 수행할 수 있다. 따라서 DeBERTa를 파인튜닝하여 음성 통화 텍스트 데이터 대한 문맥을 잘 파악하고 정상 통화와 보이스피싱 통화를 분류하는 classification을 수행하도록 한다.

4. Model Comparison & Selection

	Accuracy	F1-score
Customized KoBIGBIRD	0.9997	0.9998
DeBERTa	0.9988	0.9997

도표 1

모델 성능 비교 결과, 정확도와 F1-score에서 KoBIGBIRD가 우수함을 확인했다. 또한, 새로운 테스트 데이터 셋 정상 10개, 피싱 10개를 통한 인퍼런스 테스트 결과는 두 모델이 동일하게 20개 중 19개를 옳게 분류했다. 따라서, 보이스피싱 탐지를 위한 모델을 KoBIGBIRD로 선정하여 솔루션을 진행했다.

5. Flow Chart

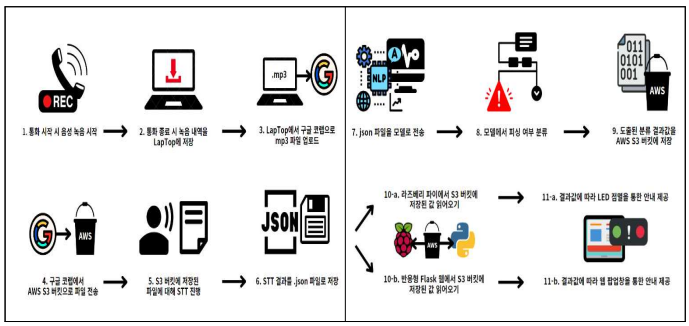

그림 1. 전체 흐름도

해당 모델을 이용한 전체 솔루션 진행 흐름도이다.

6. Implementation

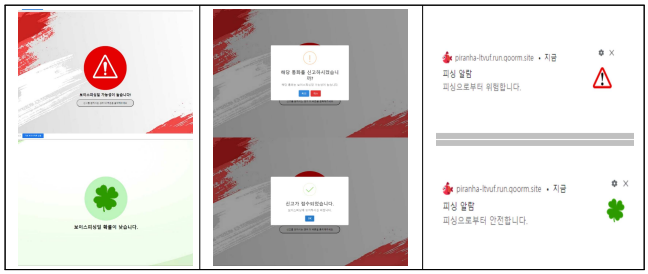

그림2. 웹앱 구현

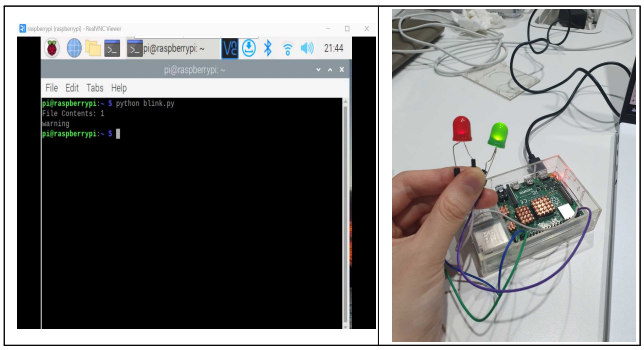

그림 3. 라즈베리파이 결과물 출력 구현

7. Conclusion

본 논문은 디지털 소외계층 및 사회적 약자를 고려한 'NLP 알고리즘 기반 보이스피싱 탐지 솔루션'을 제안하였다. 모델은 KoBIGBIRD 와 DeBERTa를 피싱 탐지 수행에 맞게끔 파인튜닝하였고 성능을 비교하여 최종 모델을 결정했다. 그 후, 실용성을 위해 웹앱 및 라즈베리를 통해 피싱 후 경고를 주는 솔루션을 구현하였다. 보이스피싱으로 인한 피해액이 커지는 현재, 본 논문에서 제안한 자동화된 솔루션이 범죄 예방에 실질적인 도움을 주기를 기대한다.

참고문헌

[1] Shanchan Wu 외 1명 ,Enriching Pre-trained Language Model with Entity Information for Relation Classification, 28th ACM International Conference on Information and Knowledge Management, Beijing China ,2019, 2-3page

[2] Manzil Zaheer 외 10명, Big Bird: Transformers for Longer Sequences, Neural Information Processing Systems, Virtual Only, 30-32 page

※ 본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.