KoBERT 기반 VoIP Voice Phishing 탐지 솔루션

조윤지 [1], 이경윤 [2], 이윤서 [3], 정재희 [4], 박세진 [5], 윤종호 [6]
[1] 서울과학기술대학교 전자 IT 미디어공학과 학부생
[2] 서울시립대학교 통계학과 학부생
[3] 덕성여자대학교 컴퓨터공학과 학부생
[4] 한국외국어대학교 전자물리학과 학부생
[5] 한양대학교 도시공학과 학부생
[6] 코즈비즈 시스템즈 대표

a01056187820@gmail.com, lky3685@gmail.com, dorothy4171@gmail.com,
jaeheeyajaehee@gmail.com, vivabsj@hanyang.ac.kr, hagis007@naver.com

The Solution for VoIP Voice Phishing Detection Based on KoBERT Model

Yun-Ji Cho [1], Kyeong-Yoon Lee [2], Yun-Seo Lee [3], Jae-Hee Jeong [4], Se-Jin Park [5], Jong-Ho Yoon [6]
[1] Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology
[2] Dept. of Statistics, University of Seoul
[3] Dept. of Computer Engineering, Duksung Women's University
[4] Dept. of Electronic Physics, Hankuk University of Foreign Studies
[5] Dept. of Urban Planning & Engineering, Hanyang University
[6] COZBIZ Systems

요 약

본 논문은 보이스피싱 취약 계층을 위해 통화 내용을 신속하게 처리하여 실시간으로 범죄 여부를 판별하는 VoIP 에 특화된 시스템을 제안하였다. 실제 보이스 피싱 통화 유형을 학습한 탐지 모델을 개발하여 API 로 배포하였다. 또한 보이스피싱 위험도가 일정 수준에 도달할 경우 사용자에게 보이스피싱 가능성을 경고하는 장치를 제작하였다. 본 연구는 보이스피싱을 사전에 탐지함으로써 개인정보의 유출 및 금융 피해를 예방하고 정보 보안을 실천하는 데 기여할 것으로 기대된다.

1. 서론

VoIP 의 사용량이 증가하며 이에 따른 보안 위협도가 증가하고 있다. 다양한 플랫폼에서 보이스피싱 사례가 지속적으로 보고되고 있으며, 개인 정보의 유출 및 금융 피해를 초래하고 있다.

최근 자연어 처리 분야에서는 BERT, GPT 와 같은 대형 언어모델이 높은 성능을 보이고 있으며, SKTBrain 에서 제공하는 KoBERT 는 한국어 특화 모델로서 국내 보이스피싱 문제의 해결을 위한 연구에 활용되고 있다.

BERT 는 transformer 의 encoder 부분을 여러 층으로 쌓은 구조이며, 주어진 입력에 대한 문맥 정보를 Bidirectional 하게 확인할 수 있다. BERT 는 BPE 알고리즘을 사용하여 WordPiece-Tokenizing 을 사용한다.

언어 모델을 학습할 때 masked Language Modeling (masked LM) 구조로 다음 문장예측을 함께 학습하게 된다.

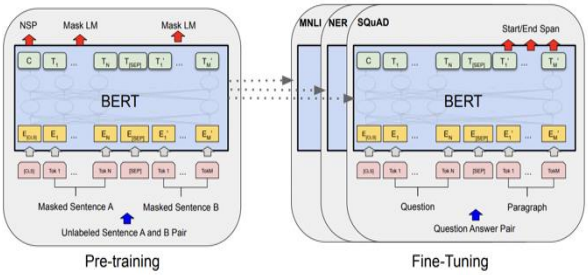


그림 1. BERT 모델 구조도[2]

본 연구에서는 문맥을 인식하는 Encoder-Decoder 모델의 특성을 살려 Dialogue-Dataset 으로 모델을 학습시켜 정확도를 향상시켰다. 그리고 GPT-3.5 모델, nlpaug 라이브러리를 사용하여 데이터를 증강하여 Data Imbalance 를 해결하였다. 또한 리소스가 제한된 임베디드 환경에서 모델 API 를 사용하여 실행시간을 단축하였다.

2. 제안 모델

2.1 데이터셋 수집

본 연구에서는 금융감독원 보이스피싱 통화 녹음 데이터[3]와 AI Hub의 주요 영역별 회의 음성인식 데이터[4]를 통해 모델을 학습시켰다.

2.2 데이터 전처리

회의 음성인식 데이터와 보이스피싱 데이터는 257313:200 으로 Data Imbalance 가 발생하였다. 따라서 nlpaug 라이브러리와 GPT 를 사용하여 Upsampling 을 진행하고 정상 데이터를 랜덤으로 추출하여 Downsampling을 수행하였다. 최종적으로는 각각 12000, 8000개의 데이터를 사용하였다. 문장의 최대 길이가 BERT의 최대 input 인 512 token을 넘지 않도록 500 이하로 처리하였고 특수문자, 숫자, 단위를 제거하였다. '그러면', '해서' 등 143개의 의미가 없는 단어를 불용어로 간주하여 제거하였으며 이후 '피해자:', '사기범:' 으로 토큰화 하여 대화형 데이터셋으로 바꾸었다.

2.3 모델 학습

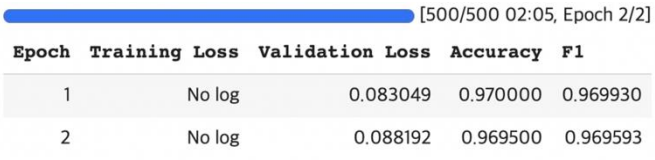
Hugging Face에 업로드 된 모델 중 skt/kobert-base-v1 으로 BertForSequenceClassification 을 수행하였고 모델 하이퍼파라미터는 다음의 표 1 과 같다. 데이터는 각각 8:1:1 비율로 Train Set, Validation Set, Test Set에 배정하였다.

Max_len	64
Batch_size	64
Warmup_ratio	0.1
Num_epochs	2
Max_grad_norm	1
Log_interval	200
Learning_rate	5e-5

표 1. 모델 하이퍼파라미터

2.4 모델 분석

학습 완료된 모델의 정확도는 그림 2 와 같으며 Accuracy와 F1 score가 0.96 정도로 높은 성능을 보임을 확인할 수 있다.


[500/500 02:05, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	No log	0.083049	0.970000	0.969930
2	No log	0.088192	0.969500	0.969593

그림 2. KoBERT 학습 결과

3. 보이스 피싱 탐지 시스템

모시은 외(2021)가 제안한 보이스피싱 예방 솔루션[1]의 시스템 구성도를 참조하여 라즈베리 파이 기반 시스템을 설계하였다. 본 시스템은 서비스 이용자의 통화 음성을 sniffing하고 Clova Speech API를 이용하여 텍스트로 변환하였다. 2항에서 제안한 모델로 보이스피싱 여부를 판별하고 LED를 이용하여 판별 여부를 확인할 수 있게 하였다.

[1]과 달리 Dialogue-Dataset을 구성하기 위해 한국어 화자 분리를 지원하는 API로 교체했으며, 로컬에서 계산 량을 줄여 판별 시간을 단축하기 위해 Hugging Face에 배포한 모델 API를 사용하였다.

4. 결론

본 연구에서는 보이스피싱에 취약한 계층을 보호하기 위해, 통화 내용을 실시간으로 분석하여 범죄 여부를 신속하게 판별하는 VoIP 기반 시스템을 제시하였다. KoBERT를 기반으로 한국어 보이스피싱 판별에 특화된 모델을 만들었으며, 최종적으로는 보이스피싱 탐지 임베디드 시스템을 구축하였다.

본 연구에서 개발한 시스템은 보이스피싱의 IP 접속을 사전에 차단하거나 딥보이스와 같은 음성 변조 기술의 부정 사용을 방지하는 시스템으로 응용할 수 있다.

- 본 프로젝트는 과학기술정보통신부
정보통신창의인재양성사업의 지원을 통해 수행한
ICT 멘토링 프로젝트 결과물입니다.-

참고문헌

[1] 모시은, 양혜인, 조은비, 윤종호.
Open STT API 와 머신러닝을 이용한 AI 보이스피싱 예방 솔루션. 한국정보처리학회 학술대회논문집, 29(2), 1013-1015. 2022.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[3] 금융감독원, 보이스피싱 체험관, 2021, https://www.fss.or.kr/fss/bbs/B0000203/list.do?menuNo=200686

[4] AI Hub, 주요 영역별 회의 음성인식 데이터, 2022, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=464