30th Humantech Paper Award

NLP 알고리즘을 활용한 AI 보이스피싱 탐지 솔루션

AI voice phishing detection solution using NLP Algorithms

Abstract

본 논문은 디지털 소외계층과 사회적 약자를 고려한 보이스피싱 예방 솔루션을 제안한다. 통화 내용을 AWS Transcribe를 활용한 STT 과정과 NLP 알고리즘 및 PLM을 거쳐, 실시간으로 보이스피싱 위험 여부를 파 악하고 해당 결과를 사용자에게 전달하도록 하는 솔루션으로, PLM으로는 KoBIGBIRD와 DeBERTa 모델 각각을 데이터에 맞게 customize하여 보이스피싱 탐지에 적절하게 파인튜닝했다. 이후, 성능과 인퍼런스를 비교하여 더 좋은 성능을 보인 Ko BIGBIRD 모델로 보이스피싱 탐지를 수행하고, 보다 좋은 결과 전달을 위한 웹앱을 구현하였다.

1. Introduction

과거 불특정 다수를 향한 보이스피싱 사례들이 많았다면, 최근에는 피해자의 개인정보를 기반으로 한, 이른바 '피해자 맞춤형 보이스피싱 시나리오'를 이용한 범죄 사례들이 증가하고 있다. 이런 상황에서 VoIP 및 유선전화 사용자를 포함한 디지털 소외계층과 보이스피싱에 취약한 사회 적 약자를 고려한 서비스가 필요하다. 이에, 본 논문에서는 KoBIGBIRD, DeBERTa 등 우수한 PLM 모델 들을 목적에 맞게 커스텀 및 비교하여 대화의 문맥을 파악해 새로운 피싱 시나리오가 나오더라도, 해당 시나리오에서 피싱 여부를 정확하게 탐지하는 솔루션을 제안한다.

여기서 PLM 모델이란 Pretrained Language Model의 약자로, 누구나 손쉽게 SOTA에 근접한 성능 달성을 가능하게 한다는 장점이 있기에 이번 솔루션에 적용하였다. 여기에, 더 나아가 디지털 소외계층과 사회적 약자들도 해당 서비스를 쉽게 활용할 수 있도록 라즈베리 파이를 활용한 LED 알림과 반응형 웹을 사용해 실사용 시 진행 과정 및 결과물을 구현했다. 다만, 본 솔루션에서는 라즈베리 파이를 노트북으로 대체하여 진행하였다.

2. Data processing

2.1 Data Acquisition

가장 먼저, 금융감독원 홈페이지의 ' 그놈 목소리' 카테고리 속 보이스피싱 음성 데이터 MP3 파일들을 동적 크롤링을 활용하여 한 폴더에 다운로드하였다. 해당 MP3 파일들은 음성 파일이기에 아마존에서 제공하는 AWS Transcribe를 활용해 AWS S3 Bucket에 해당 파일들을 업로드한 후, STT(Speech-To-Text) API를 활용해 text 형식으로 변환하였다. 피싱이 아닌, ' 정상' 데이터는 AI Hub에서 제공하는 민원(콜센터) 질의응답 데이터 중 금융/보험, 이체/출금, 대출 서비스 항목 데이터를 사용하였으며 이 데이터 역시 STT 과정을 거쳤다. 이렇게 STT 과정을 거친 결과, 피싱 데이터는 370 건, 정상 데이터는 9698 건이 수집되었다.

피싱, 정상 관계없이 데이터의 전반적인 양 자체가 매우 부족하고, 두 클래스(피싱-정상) 간 데이터 불균형이 존재하면 과대적합 발생의 가능성 및 위험도가 높아질 수 있다고 판단하여 피싱 데이터에 한정하여 데이터 증강을 진행하였다.

2.2 Data Augmentation

데이터 증강의 경우, 증강하고자 하는 데이터의 특성에 따라 그 기법 및 결과물이 매우 다양하다. 본 솔루션에서는 STT를 거친 Text 형태의 데이터를 증강하고자 하였는데, 텍스트 데이터의 경우 한 단어만 바꿔도 문장의 의미가 달라질 수 있기에 통상적으로 데이터 증강이 어렵다고 알려져 있다. 그럼에도, 아래와 같은 방법을 활용해 증강이 가능하다.

- text를 다른 언어로 번역한 후 원 언어로 번역 (번역 후 회귀)
- 특정 단어를 유의어로 교체
- 임의의 단어를 삽입 혹은 삭제
- 문장 내 두 단어의 위치 교환

위와 같이 4가지 기법을 사용하였고, 결과적으로 증강을 통해, 피싱 데이터 15 ,612 개, 정상 데이터 개수 58 ,187 개로 이뤄진, 총 73 ,799 개의 학습 데이터 셋을 구성할 수 있었다. 각 기법별, 유형별 증강된 데이터 수는 아래 표와 같다.

[도표 1] 데이터 증강 결과

원본 데이터 개수	370	"	"	"
증강 기법	단어 삭제	단어 추가	문장 재구성	번역 후 회귀
수사기관 사칭형	2232	3760	940	188
대출 사기형	1629	5792	905	166
총합	3861	9552	1845	354

3. MODEL (PLM)

위 Introduction 부분에서 언급하였듯, PLM은 Pretrained Language Model의 약자로 누구나 활용할 수 있는 오픈소스 자연어처리 모델을 의미한다. 오픈소스인 만큼 비용을 절감하기에 수월하며, 우리의 데이터를 활용한 customizing 및 fine tuning이 어렵지 않고, 접근성을 높여 본 솔루션의 장점을 극대화할 수 있다고 판단해 PLM 중에서도 한국어 모델에 해당하는 KoBIGBIRD와 DeBERTa를 활용하게 되었다.

BERT 모델의 경우 104개 언어로 구성된 위키피디아(Wikipedia) 데이터셋을 학습시킨 모델을 의미한다. 때문에 광범위한 활용이 가능하다는 장점이 있지만, 모델의 크기가 과도하게 크며, 형태소와 조합하여 만들 수 있는 글자 수가 타 언어에 비해 상당히 많은 한글의 특성상 활용하기는 어렵다.

이를 개선시킨 것이 KorBERT, KrBERT 등의 한국어 모델이나, 본 솔루션에서는 '보이스피싱'의 특성상, 이보다도 더 SOTA 성능을 잘 보이는 모델이 필요하다고 판단하여 KoBIGBIRD 모델을 활용하게 되었다.

3-1.Customized KoBIGBIRD

해당 모델은 R-BERT 모델로부터 영감을 얻어 KoBIGBIRD 모델의 Architecture (내부 구조)를 수정한 모델이다. 본 모델은 기존에 나와 있는 Kr-BERT의 CLS 토큰과 KoBIGBIRD의 CLS 토큰의 임베딩 값을 결합한다. [1] 그다음, 전체 대화 데이터와 키워드 및 형태소만 을 추출한 데이터를 벡터로 분리한다. 이 과정에서 문장의 끝을 나타내는 인덱스를 사용했다. 이는 전 체 대화 데이터를 통해 대화의 문맥을 이해하고, 형 태소 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

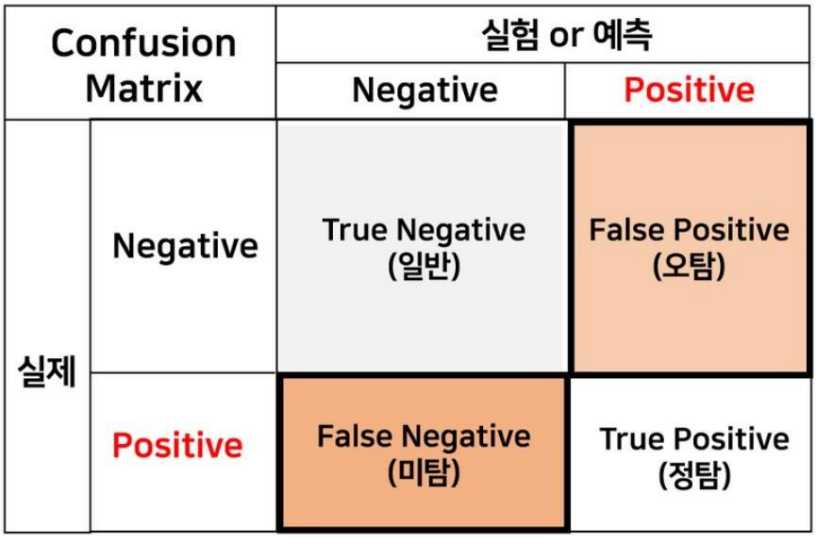
DeBERTa는 "Disentangled" 어텐션 메커니즘을 사용하여 단어 간의 상호작용을 더욱 세분화하고, 이전 단어와의 연관성을 강조하면서도 다음 단어에 대한 정보 또한 보다 효과적으로 포착할 수 있는 모델이다. 이는 효과적인 문맥 파악을 가능하게 해준다는 장점이 있다. 보이스피싱 상황의 경우, 진행되는 말의 앞뒤 문맥을 정확히 파악할 필요가 있는 만큼 해당 특성은 큰 장점으로 작용한다. 보다 구체적으로는, 이 모델은 Enhanced mask decoder BERT 부분을 차용하여 불필요한 정보의 누출을 방지하고 단어의 절대적 위치 정보를 고려할 수 있도록 기능한다. 이로써 각 단어가 문장에서 어떤 역할을 하는지 파악하여 더 정확한 디코딩을 수행할 수 있다. 따라서 DeBERTa를 fine-tuning하여 음성 통화 텍스트 데이터라는 input을 받았을 때, 해당 텍스트 상에서의 문맥을 잘 파악하고 정상 통화와 보이스피싱 통화를 분류하는 작업을 수행하게 된다.

4. Model Comparison & Selection

모델의 평가 지표에는 크게 정확도, 정밀도, F1-Score가 있다. 아래 오분류표에서 정확도란 전체 중 True로 분류된 것들의 비중으로 (True Negative + True Positive) / (True Negative + False Positive + False Negative + True Positive)로 계산된다.

정밀도란 True Positive / (True Positive + False Positive) , 즉 예측 결과가 Positive인 결과 중 실제로 Positive인 것들의 비율을 의미한다. 재현율은 유사하지만, 실제 True인 것들 중에서 모델이 True라고 예측한 것의 비율로 True Positive / True Positive + False Negative로 계산한다. 그러나 정확도를 계산한 상황에서 이를 모두 계산하는 것보다는, 정밀도와 재현율의 조화평균인 F1 Score를 계산하는 것이 보다 효과적일 것이라 판단하여 정확도와 F1 Score를 평가지표로 활용하게 되었다.

[도표 2] 오분류표



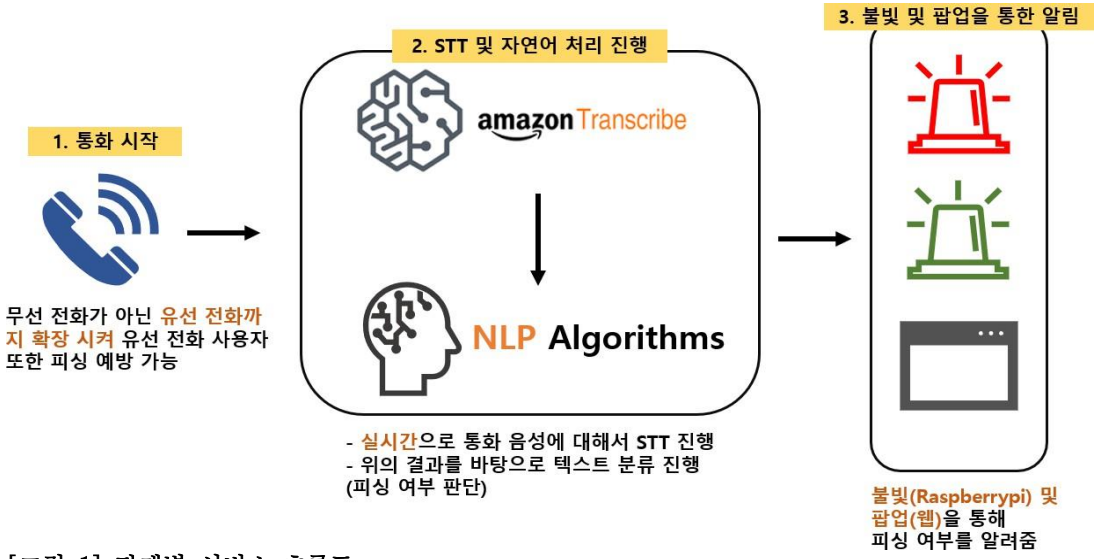
[도표 3] 각 모델별 정확도 및 F1 Score

Model Evaluation Matrix	Accuracy	F1-score
Customized KoBIGBIRD	**0.9997**	**0.9998**
DeBERTa	**0.9988**	**0.9997**

증강한 후, STT를 거친 데이터를 csv 형식으로 변환하여 각 모델들의 input(입력값) 파일로 사용하였다. 그 결과로 나온 성능을 비교한 결과, 정확도와 F1-score 두 지표 모두 KoBIGBIRD가 우수함을 확인할 수 있었다. 또한, 새로운 테스트 데이터 셋 정상 10 개, 피싱 10 개를 통한 모델 테스트 결과는 두 모델이 동일하게 20 개 중 19 개를 옳게 분류하였기 때문에 보이스피싱 탐지를 위한 모델을 KoBIGBIRD로 선정하여 남은 솔루션을 진행하였다.

5. Service Flow Chart

통화를 하면서 실시간으로 피싱 여부를 인지할 수 있어야 금전적인 피해를 보다 효과적으로 방지할 수 있을 것이라고 판단하여, 실시간성을 추가한 서비스의 기획 및 흐름도를 작성하였다. 해당 흐름도는 아래와 같다.

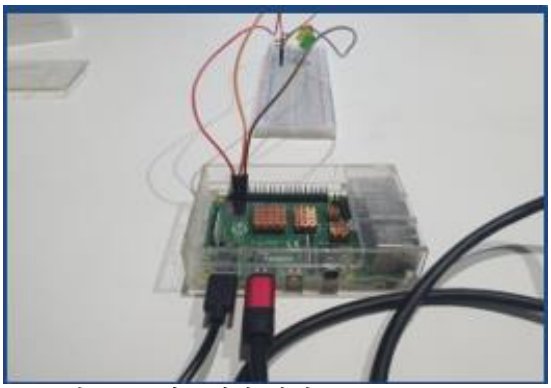


[그림 1] 단계별 서비스 흐름도

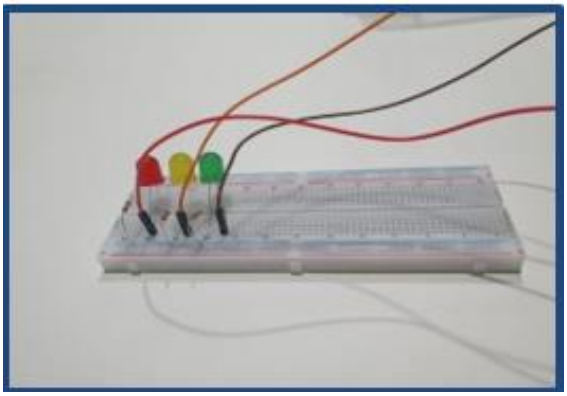
6. Raspberry Pi

라즈베리 파이(Raspberry Pi) 란 초소형, 초저가 PC를 의미한다. 일반 PC에 비해 저렴한 가격이기 때문에 본 솔루션의 접근성을 높이자는 취지에도 부합하며, 연결만 된다면 실시간성을 반영하기에 가장 적합하다고 판단하였다. 때문에, Amazon Transcribe 및 STT API를 활용한 텍스트로의 통화 내역 변환 및 위에서 선정한 KoBIGBIRD 모델을 활용한 피싱 여부 판단을 라즈베리 파이 상에서 진행하였다. 다만, 이번 솔루션에서는 라즈베리 파이는 알림 서비스에 보다 집중하고자 추가적인 대체 Laptop을 라즈베리 파이와 연동시켜 STT 및 NLP 알고리즘 처리는 이 Laptop에서 진행하였다.

먼저, 라즈베리 파이가 알려줄 서비스는 위험도 피싱 여부이다. 모델이 결과값 피싱 여부를 1(피싱) 혹은 0(정상)으로 반환하면, 이를 AWS S3 Bucket에 전달하고, 전달된 결과값을 라즈베리 파이가 읽어온 후, 피싱이면 빨간색 불빛을, 안전하다면 초록색 불빛을 깜박이게끔 함으로써 시각적으로 피싱 여부를 판단할 수 있도록 하였다. 여기에 추가적으로 반응형 웹 결과도 추가하여 활용도를 제고하였다.



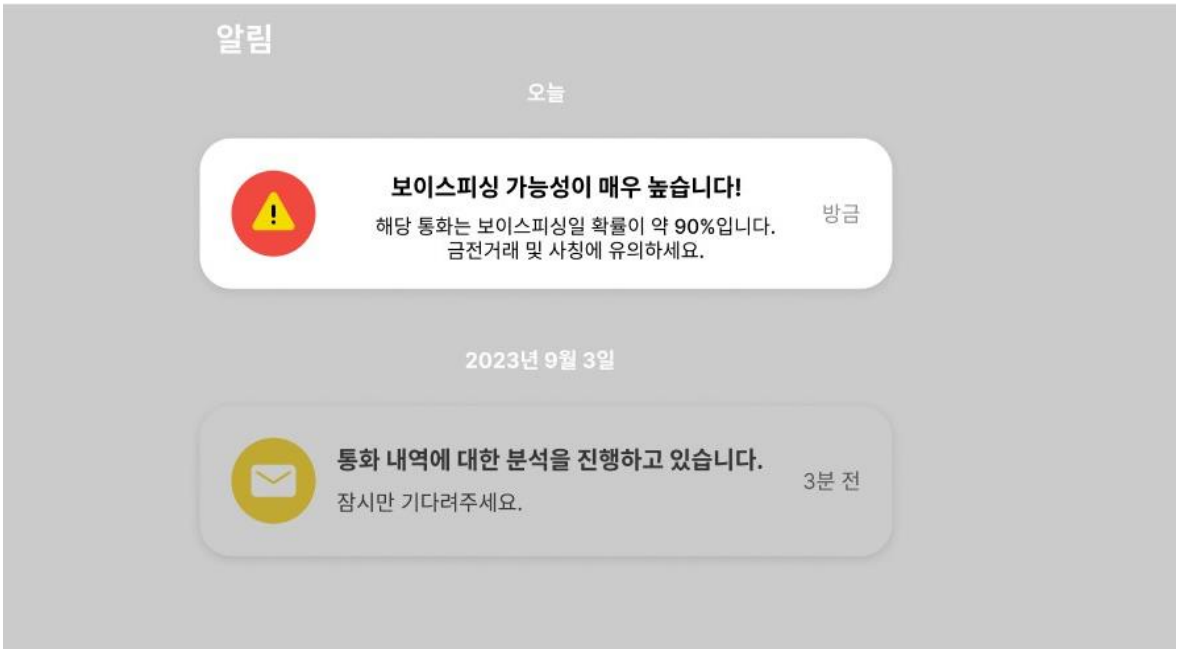
[그림 2-1] 라즈베리 파이 구조도



[그림 2-2] 불빛 알람 및 회로

7. Implementation

공익적인 주제의 솔루션인 만큼, 많은 사람들이 보다 손쉽게 사용할 수 있도록 디자인할 필요가 있다고 판단하여 Figma 툴을 활용해 먼저 디자인을 진행한 후, 이를 javascript 언어를 활용한 HTML 및 CSS 문서로 변환하여 웹앱을 구현하였다. 웹 호스팅 사이트로는 goorm(구름) IDE를 활용하였다.



[그림 3] 알림 웹앱 디자인을 위한 원본 피그마(Figma) 파일

[그림 4] 알림 웹앱 대기 화면

[그림 5] 피싱여부에 따른 웹앱 메인 화면 및 팝업 알림

위 그림(그림 4)는 본 솔루션의 웹앱을 goorm IDE에서 실제로 작동시키며 각 상황별 화면을 캡처해온 화면이다. 특히 보이스피싱일 가능성이 높다고 판단되는 경우, 본 솔루션을 구현하면서 가장 어려웠던 부분이 데이터였기 때문에 클릭하면 곧바로 데이터베이스에 신고 내역을 쌓을 수 있도록 바로 신고가 가능하게끔 링크를 제작하였다. 마지막으로는, 통화량 및 STT 변환 시 텍스트 분량이 적지 않을 수 있음을 고려한 팝업 알림을 구현하였다. 통화량 및 텍스트 분량이 늘어나면 이에 대한 문맥을 탐지하고, 피싱 여부를 분류하는 데는 시간이 더 소요될 수밖에 없고, 자연스럽게 사용자가 기다리는 시간이 늘어나게 된다. 때문에 분류작업이 완료되었을 때 이를 알려줄 수 있는 팝업창이 필요하다고 생각해 추가적으로 구현하였다.

8. Conclusion

본 논문은 디지털 소외계층 및 사회적 약자를 고려한 ' NLP 알고리즘 기반 보이스피싱 탐지 솔루션' 을 제안하였다. 가장 핵심이라고 할 수 있는 보이스피싱 여부 분류를 진행할 모델은 우리가 수집 및 전처리한 데이터를 input으로 넣는 파인튜닝을 통해 KoBIGBIRD 와 DeBERTa 두 모델을 customize하는 과정을 거쳐, 분류 성능을 비교하여 최종 모델을 선정했다. 그 후, 실용성을 위한 웹앱 및 라즈베리 파이를 통해 피싱 후 경고를 주는 솔루션을 기획 후 구현하였다. 새로운 유형의 보이스피싱으로 인한 피해액이 커지면서 이에 대한 방지책이 필수적이지만, 스마트폰이 아닌 기기를 사용하는 경우 구제책 및 예방책이 마땅치 않은 현실이다. 이러한 현실 속에서, 본 논문에서 제안한 자동화된 보이스피싱 예방 및 알림 솔루션이 피해 예방에 실질적인 도움을 줄 수 있기를 기대한다.

덧붙여, 해당 솔루션의 상용화를 통해 피싱 통화내역 데이터가 쌓인다면 지속적인 모델 input의 업데이트를 통해 솔루션의 보완 및 개선이 이루어지기를, 그렇게 더 많은 사람들이 보이스피싱으로부터 안전한 사회를 만들 수 있기를 고대한다.

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 10 명, Big Bird : Transformers for Longer Sequences, Neural Information Processing Systems, Virtual Only, 30 -32 p age

.