LLM을 활용한 보이스피싱 실시간 탐지 및 대응 연구

이정현[1*], 최진우[1*], 강평종[1], 이동호[1], 김진우[2†]

[1,2]광운대학교 (학부생, 교수)

Real-time Voice Phishing Detection and Response Using LLMs

Jung-Hyun Lee[1*], Ji-Nu Choi[1*], Pyung-Jong Kang[1], Dong-Ho Lee[1], Jin-Woo Kim[2†]

[1,2]Kwangwoon University(Undergraduate Student, Professor)

요 약

보이스피싱은 공공기관 사칭, 심리적 압박, 유창한 언어 구사로 탐지가 어렵고 피해가 증가하고 있다. 본 연구는 실시간 통화 내용을 텍스트로 변환하여 보이스피싱 여부를 탐지하고, 유사 사례 및 대응 지침을 제공하는 통합 시스템을 제안한다. 제안 시스템은 KLUE-RoBERTa 기반 보이스피싱 탐지 모델과 LG-EXAONE 및 Ko-SBERT 임베딩 기반 사례 검색 기능으로 구성된다. 실험 결과, 실시간 탐지와 대응 정보 제공을 결합한 구조가 효과적으로 작동함을 확인하였으며, 향후 발화 특성을 반영한 모델 확장을 통해 실용성을 더욱 높일 수 있음을 보여준다.

I. 서론

보이스피싱은 전화, 문자, 메신저 등을 통해 수신자를 속여 금전적 이득을 취하는 범죄 행위로, 디지털 사회의 발전과 함께 점차 고도화되고 지능화되고 있다. 경찰청 통계에 따르면, 국내 보이스피싱 피해는 최근 5년간 연평균 약 2만 건 이상 발생하고 있으며, 2024년 1인당 피해액은 4,100억 원으로 전년(2,366억 원) 대비 1,734억 원(73%) 증가하였다[1]. 최근에는 공공기관을 사칭하거나 개인정보 노출, 범죄 연루와 같은 허위 사실을 제시해 피해자에게 심리적 압박을 가하는 방법을 사용하고 있다. 또한 초기의 피싱과 달리 유창한 한국어를 사용하고 있어 구분이 어려워졌다[2].

기존 보이스피싱 대응 애플리케이션들(예: 후후, 피싱아이즈, 시티즌코난 등)은 통화 상대의 번호가 기존의 의심 이력에 포함되어 있는지를 확인하거나, 통화 내용을 키워드 중심으로 분석하는 방식을 주로 사용하였다. 그러나 이러한 방식은 규칙 기반 탐지 방식에 의존하여, 신고되지 않은 전화번호나 자연어로 위장되어 맥락적 의미 파악이 필요한 경우는 보이스피싱에 효과적으로 대응하기 어렵다.

본 연구에서는 실시간 통화 내용을 텍스트로 변환하고 이를 분석하여 보이스피싱 여부를 판단한 뒤, 유사 사례 및 대응 요령을 함께 제공하는 방식을 제안한다. 이를 위해 사전학습된 한국어 언어모델을 활용하여 통화 세션 단위로 보이스피싱 위험도를 탐지하고, 의심 내용을 탐지 시 사용자에게 유사 사례와 대응 가이드를 제시한다. 이는 단순 탐지를 넘어 사용자가 상황에 맞는 실질적인 대응 판단을 도와준다는 점에서 기존 방식과 차별된다.

본 논문의 주요 기여는 다음과 같다. 첫째, 실시간 통화 환경에 최적화된 분류 모델을 구축하고, 다양한 한국어 사전학습 언어모델의 탐지 성능을 비교하였다. 둘째, 보이스피싱 탐지 이후 실제 사례와 대응 요령을 함께 제공하는 시스템을 구현하였다. 이를 통해 보이스피싱에 대한 조기 대응과 피해 예방을 위한 기술적 가능성을 제시하고자 한다.

II. 관련 연구

보이스피싱 탐지를 위한 기존 연구들은 문서 임베딩, 개체명 인식, 사전학습 언어 모델 등 다양한 자연어 처리 기법을 기반으로 텍스트 분류 및 유사도 분석 접근 방식을 활용해왔다.

Kim 등이 제안한 연구[3]에서는 금융감독원이 제공한 보이스피싱 음성 데이터를 활용하여 음성 인식 후 생성된 텍스트를 대상으로 Doc2Vec 기반

* 공동 1저자
† 교신저자, jinwookim@kw.ac.kr

임베딩과 코사인 유사도 계산을 수행하는 방식을 제안하였다. 실험 결과, Doc2Vec 기반 임베딩 방법은 정확도 0.61, F1-score 0.74를 기록했으며, 음성을 텍스트로 변환한 후 문서 간 유사도를 분석하는 방식이 효과적임을 알 수 있었다.

Boussougou 등이 제안한 연구[4]에서는 KorCCVi 데이터셋을 기반으로 여러 머신러닝 및 딥러닝 모델의 성능을 비교한 결과, 문장 내 장기 의존성과 문맥을 효과적으로 포착할 수 있는 KoBERT 모델이 가장 높은 성능을 기록하였다.

본 연구는 이러한 기반 기술들을 바탕으로 탐지 결과와 유사 사례를 연계함으로써 사용자에게 실시간으로 보이스피싱 대응에 도움이 되는 정보를 제공한다는 점에서 차별성을 가진다.

III. 제안 방법

본 연구에서는 실시간 통화 내용을 텍스트로 변환한 뒤, 해당 문장들을 분석하여 보이스피싱 여부를 판단하고, 유사한 사례와 그에 따른 대응 요령을 제공하는 통합 시스템을 제안한다. Fig. 1은 제안하는 시스템의 전체적인 구조와 처리 과정을 나타낸다.

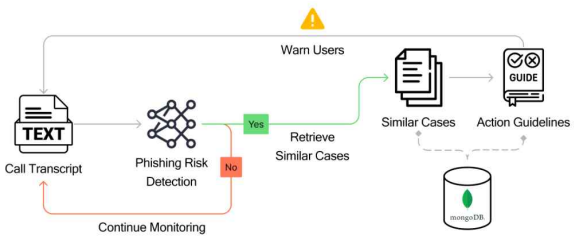

Fig. 1. Phishing Model Architecture

3.1 실시간 통화 세션 단위 탐지 모델

보이스피싱은 일상적인 대화를 모방하며, 명백한 범죄적 표현 없이도 사용자를 속이는 경우가 많다. 특히 실제 상담 대화와 유사한 문맥이 포함되어 단순한 키워드 탐지나 규칙 기반 필터링만으로는 정확한 탐지가 어렵다. 이를 극복하기 위해, 사전학습된 한국어 언어 모델을 활용하여 통화 세션 단위로 실시간 탐지가 가능한 분류 모델을 구축하였다.

사용된 모델은 KoBERT [5], KoELECTRA [6], KLUE-RoBERTa [7]로, 모델별 성능을 비교하기 위해 동일한 데이터셋과 전처리 방식을 적용하여 파인튜닝하였다. 학습 데이터는 총 1,012건으로 구성되며, 금융감독원 보이스피싱 체험관 음성 데이터 506건[8], AIHub 민원(콜센터) 질의응답 텍스트 데이터 중 금융/상담과 관련된 데이터 253건, 그리고 AIHub 일상대화 주제별 텍스트 데이터 253건[9]을 결합하여 구성하였다.

각 대화 세션은 '보이스피싱' 또는 '정상'으로 이진 라벨링되었으며, 통화 세션 전체 텍스트를 하나의 입력으로 처리하여 각 모델의 입력 형식에 맞게 전처리하였다.

3.2 유사 사례 기반 대응 정보 제공 기능

탐지 모델이 특정 대화를 보이스피싱으로 분류한 경우, 해당 대화와 유사한 실제 사례를 검색하여 사용자에게 대응 정보를 제공한다. 금융감독원이 공개한 실제 보이스피싱 사례 데이터를 기반으로 총 82건의 피해 사례 유형을 수집하여 비교를 위한 데이터베이스를 구축하였다. 각 사례는 피해 사례의 핵심을 요약한 문장과 이에 맞는 행동 요령으로 구성되며, GPT-4o를 활용해 생성 및 정리되었다.

생성된 사례 요약 문장은 Ko-SBERT [10]를 이용해 768차원 임베딩 벡터로 변환하였으며, 세 가지 정보 모두 MongoDB에 저장하였다. 실제 서비스 운용 시, 수신된 통화 내용 전체를 LG-EXAONE [11] 모델에 입력하여 요약을 수행한다. 생성된 요약문은 임베딩 벡터로 변환한 후 데이터베이스에 저장된 사례 임베딩들과의 코사인 유사도를 계산하여 가장 유사한 사례를 검색한다. 유사도 점수가 사전에 정의된 임계값 이상일 경우에만 해당 사례의 대응 정보를 사용자에게 제공함으로써, 부정확하거나 불필요한 정보 제공을 방지한다.

IV. 평가

앞서 제안한 실시간 보이스피싱 탐지 및 대응 시스템의 핵심 과정에 대해 실험을 수행하고, 모델별 성능을 정량적으로 비교하였다. 실험은 두 가지 측면에서 진행되었다.

4.1 보이스피싱 탐지 모델 성능 평가

보이스피싱 탐지 성능을 평가하기 위해 KoBERT,

KoELECTRA, KLUE-RoBERTa 세 모델을 동일한 분류 구조로 파인튜닝하여 성능을 비교하였다. 전체 1,012건의 데이터를 8:2 비율로 학습 및 검증 세트로 나누었고, 배치 크기 32, 학습률 2e-5, 에폭 수 5로 설정하였다. 각 모델은 통화 세션 단위 텍스트를 입력으로 받아 '보이스피싱' 여부를 판단하며, Accuracy, F1-Score, 추론 시간을 주요 평가 지표로 사용하였다. 실험은 Google Colab의 NVIDIA T4 GPU 환경에서 수행되었다.

Table. 1. Voice Phishing Detection Model Performace

Model	Accuracy	F1-Score	Time(s)
KoBERT	0.9954	0.9954	0.0533
KoELECTRA	0.9954	0.9954	0.0606
KLUE-RoBERTa	1.0000	1.0000	0.0518

세가지 모델 모두 비슷한 정확도, F1-Score, 추론 시간을 기록하였으며, 그 중 KLUE-RoBERTa 모델이 약간 높은 성능을 보여 실시간 탐지에 가장 적합한 모델로 판단되었다.

4.2 유사 사례 검색 및 대응 정보 제공 실험

유사 사례 검색 기능의 실효성을 평가하기 위해, 실제 통화 시나리오를 요약 및 임베딩 하여사례 데이터베이스와의 유사도를 계산하였다. 생성된 요약문은 다음과 같다: "서울중앙지검 특수부가 2021년 사건 조사 중 개인정보 유출로 인해 김창호 관련 성매매 알선 및 불법 자금 은닉 사건에 대한 소환장을 발부하려는 상황입니다."

이 문장은 "검찰·공공기관 사칭으로 CD기 유인 후 피해자 모르게 송금 유도" 사례와 가장 높은 유사도를 보였으며, 유사도 점수는 0.4258이었다. 해당 사례에 기반하여, "현금지급기 유도는 보이스피싱으로 간주하고, 계좌 조작 요구는 즉시 거절" 등의 대응 지침이 제시되었다. 이 결과를 통해, LG-EXAONE 기반 요약 및 Ko-SBERT 임베딩에 기반한 유사 사례 검색이 실제 서비스 운용에 적합한 성능과 정확도를 제공함을 확인하였다.

V. 결론

본 연구는 실시간 통화 내용을 기반으로 보이스피싱 여부를 탐지하고, 유사 사례 및 대응 정보를 제공하는 통합 시스템을 제안하고 구현하였다. 통화 세션 단위 보이스피싱 분류에는 KLUE-RoBERTa 모델이 적합하였으며, 대응 정보 제공은 LG-EXAONE 기반 요약 및 Ko-SBERT 임베딩 기반 유사 사례 검색을 통해 실시간 환경에서도 실효성을 확인하였다. 제안된 시스템은 단순 탐지를 넘어 실질적인 대응 지침까지 제공함으로써 사용자 판단을 지원하는 실시간 대응 도구로서의 가능성을 제시한다.

다만, 실시간 통화 환경은 잡음, 발화 단절 등 특수성을 가지므로 추가적인 학습 기법이 요구된다. 또한, 실제 서비스 환경에서의 안정성 검증과 다양한 발화 유형 및 복합 시나리오 대응 범위 확대가 향후 과제로 남는다.

[참고문헌]

[1] Money Today, "[Exclusive] 41 million KRW stolen per voice phishing call... One victim lost 3 billion KRW," http://news.mt.co.kr/mtview.php?no=2025022109373997591, February 2025. Accessed: May 2, 2025.

[2] Woori Investment Securities, "Voice Phishing Prevention," https://fundsupermarket.wooriib.com/fmc/FMC790000/main.do, 2025. Accessed: May 2, 2025.

[3] J.-W. Kim, G.-W. Hong, and H. Chang, "Voice recognition and document classification-based data analysis for voice phishing detection," Human-centric Computing and Information Sciences, vol. 11, pp. 1-13, January 2021.

[4] M. K. M. Boussougou and D.-J. Park, "Exploiting Korean language model to improve korean voice phishing detection," KIPS Transactions on Software and Data Engineering, vol. 11, pp. 437-446, October 2022.

[5] SKTBrain, "SKTbrain/kobert: Korean bert pre-trained cased (kobert)." https://github.com/SKTBrain/KoBERT, 2025. Accessed: May 2, 2025.

[6] J. Park, "Koelectra: Pretrained electra model for Korean." https://github.com/monologg/KoELECTRA, 2020. Accessed: May 2, 2025.

[7] S. Park, J. Moon, et al., "Klue: Korean language understanding evaluation," 2021. Accessed: May 2, 2025.

[8] Financial Supervisory Service, "Financial Supervisory Service Main Page," https://www.fss.or.kr, Accessed: May 2, 2025.

[9] National Information society Agency, "AIHub," https://www.aihub.or.kr, Accessed: May 2, 2025.

[10] J. Ham, Y. J. Choe, et al., "KorNLI and korSTS: New benchmark datasets for korean natural language understanding," arXiv preprint arXiv:2004.03289, 2020.

[11] LG AI Research, S. An, et al., "EXAONE 3.5: Series of Large Language Models for Real-world Use Cases," arXiv preprint arXiv:2412.04862, 2024. [Online]. Available: https://arxiv.org/abs/2412.04862