LLM을 활용한 보이스피싱 실시간 탐지 및 대응 연구

이정현[1*], 최진우[1*], 강평종[1], 이동호[1], 김진우[2]
1,2광운대학교 소프트웨어학부 (학부생, 교수)

I. 연구의 배경 및 목적

- 최근 보이스피싱 범죄는 공공기관 사칭, 개인정보 유출 협박 등 고도화된 수법과 유창한 언어 사용으로 탐지가 점점 어려워지고 있다.
- 기존에는 신고된 번호 조회나 키워드 중심 분석에 의존했지만, 자연어 문맥에 기반한 위장 대화는 이러한 방식으로는 효과적으로 대응하기 어렵다.
- LLM의 발전으로 문맥 이해와 상황 인식이 가능한 자연어 처리 모델들이 실시간 분석에 활용될 수 있게 되었다.

본 연구에서는 LLM을 활용하여 통화 내용을 분석하고, **보이스피싱 여부 판단과 함께 유사 사례 및 대응 정보를 제공**하는 통합 시스템을 제안한다.

단순 탐지를 넘어 실제 사용자 대응을 돕는 실용적인 접근을 통해, 발전하는 보이스피싱 위협에 대응할 수 있는 가능성을 제시한다.

II. 제안 방법

- 본 시스템은 통화 중 음성 데이터를 실시간으로 텍스트로 변환하고, LLM 기반 한국어 분류 모델을 활용하여 보이스피싱 여부를 탐지한 뒤, 탐지 시점에 유사 사례와 그에 따른 대응 지침을 제공한다.
- 이는 단순 경고 수준을 넘어 사용자의 실질적인 행동 판단을 돕는 것을 목표로 하며, 실시간 대응이 가능하도록 전체 프로세스를 자동화한 구조로 설계되었다.

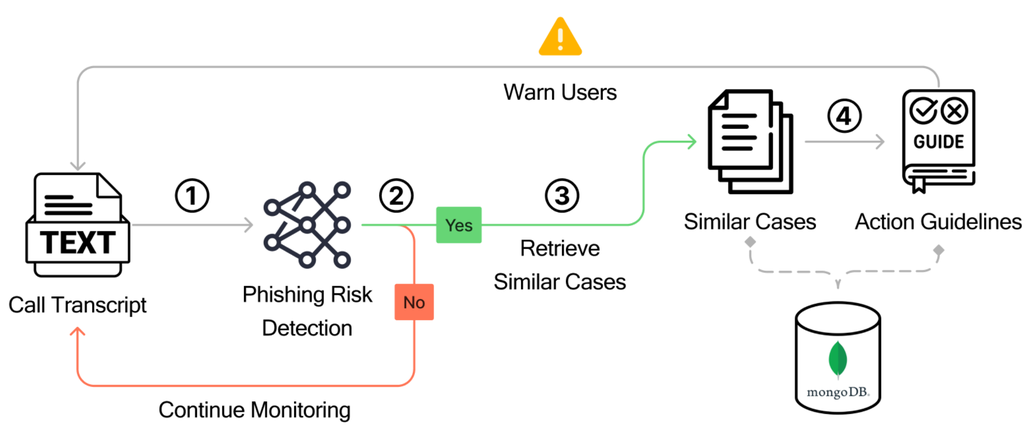

Fig. 1. Phishing Model Architecture

① Call Transcript

- 통화 중인 음성 데이터를 실시간으로 STT 모델을 통해 텍스트로 변환한다.
- 변환된 텍스트는 이후 보이스피싱 탐지 모델의 입력값으로 사용된다.

② Phishing Risk Detection

- 변환된 통화 텍스트는 한국어 언어모델 기반의 분류기에 입력된다.
- 보이스피싱 여부를 판단하기 위해 KoBERT, KoELECTRA, KLUE-RoBERTa 세 모델을 파인튜닝하여 탐지 성능을 실험하였다.
- 실험 결과, KLUE-RoBERTa가 가장 우수한 성능을 보여 최종 탐지 모델로 시스템에 적용되었다.
- 보이스피싱이 아닌 것으로 판단된 경우, 시스템은 통화 종료 시점까지 지속적으로 모니터링을 수행한다.

③ Retrieve Similar Cases

- 보이스피싱으로 탐지된 경우, 의심되는 통화 내용을 요약한 뒤 유사한 실제 피해 사례를 검색한다.
- 요약은 LG-EXAONE 모델을 통해 수행되며, 요약된 문장은 Ko-SBERT로 임베딩되어 MongoDB에 저장된 피해 사례들과 코사인 유사도로 비교된다.
- 유사도가 사전 정의한 임계값을 초과하면 가장 유사한 사례 한 건을 추출하여 사용자에게 제공한다.

④ Action Guidelines

- 검색된 유사 사례에는 해당 상황에 대응하기 위한 행동 지침이 포함되어 있다.
- 단순 경고를 넘어 상황에 맞춘 실질적 판단을 도와주는 핵심 기능이다.

III. 실험 결과

보이스피싱 탐지 모델 성능 평가

- **모델**: KoBERT, KoELECTRA, KLUE-RoBERTa
- **데이터셋 구성**: 총 1,012건
  - 금융감독원 보이스피싱 체험관 데이터 506건
  - AIHub 민원(콜센터) 금융/상담 질의응답 텍스트 253건
  - AIHub 일상대화 주제별 텍스트 데이터 253건
- **환경**: Google Colab (NVIDIA T4 GPU)
- **설정**: 배치 크기 32, 학습률 2e-5, 에폭 수 5
- **지표**: Accuracy, F1-Score, 추론 시간 기준 성능 비교
- **실험결과**

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

Table. 1. Voice Phishing Detection Model Performance

- **결과**: 세 모델 모두 높은 정확도를 기록했으며, 그 중 KLUE-RoBERTa[1]가 가장 안정적이고 빠른 탐지 성능을 보여 최종 모델로 선정되었다.

유사 사례 검색 및 대응 정보 제공 실험

- **요약 모델**: LG-EXAONE[2]
- **임베딩 모델**: Ko-SBERT[3] (768차원 벡터 변환)
- **사례 데이터베이스**: 금융감독원이 공개한 실제 보이스피싱 사례 82건
- **검색 방식**: 요약된 통화 내용을 Ko-SBERT로 임베딩한 후, 사례 DB와 코사인 유사도 계산
- **평가 목적**: 실제 통화 내용을 기반으로 유사 사례를 얼마나 정확히 찾아내고, 사용자에게 실질적인 대응 가이드를 제공할 수 있는지 확인

유사 사례 매칭 결과 예시

입력 요약문 예시

"서울중앙지검 특수부가 2021년 사건 조사 중 개인정보 유출로 인해 김창호 관련 성매매 알선 및 불법 자금 은닉 사건에 대한 소환장을 발부하려는 상황입니다."

매칭된 유사 사례와 대응 방법

[유사 사례]
검찰, 공공기관 사칭으로 CD기 유인 후 피해자 모르게 송금 유도

[대응 방법]
- 검찰, 경찰, 금융감독원 등은 전화로 계좌 송금을 요구하거나 출석 요구서를 발부하지 않음
- CD기로 이동을 유도하는 경우, 즉시 통화를 종료하고 해당 기관에 직접 사실 여부를 확인할 것①②③④

IV. 결론 및 향후 과제

결론

- 본 연구에서는 KLUE-RoBERTa 기반 분류 모델로 통화 내용을 으로 분석하여 보이스피싱 여부를 탐지하고, LG-EXAONE 요약 및 Ko-SBERT 임베딩을 통해 유사 사례와 대응 지침을 제공하는 통합 시스템을 구현하였다.
- 제안된 시스템은 단순 탐지를 넘어 실제 상황에 맞는 대응까지 지원함으로써, 보이스피싱 피해 예방을 위한 실질적인 실시간 대응 도구로서의 가능성을 보였다.

향후 과제

- 실시간 통화 환경은 잡음, 발화 단절 등 다양한 특수성이 존재하므로, 이를 보완할 수 있는 전처리 및 학습 기법의 추가 적용이 필요하다.
- 보이스피싱 시나리오가 점차 복잡해짐에 따라, 다양한 유형을 포괄할 수 있는 학습 데이터의 확장도 요구된다.

참고문헌

[1] S. Park, J. Moon, et al., "Klue: Korean language understanding evaluation," 2021. Accessed: May 2, 2025.
[2] 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
[3] J. Ham, Y. J. Choe, et al., "KorNLI and korSTS: New benchmark datasets for korean natural language understanding," arXiv preprint arXiv:2004.03289, 2020.