Journal of KIIT. Vol. 24, No. 2, pp. 1-15, Feb. 28, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.2.1

합성 데이터와 XAI를 활용한 온디바이스 텍스트 기반 보이스피싱 탐지 시스템

이규혁*[1], 박수진*[2], 김건우**

On-Device Text-based Voice Phishing Detection System Exploiting Synthetic Data and XAI

Gyu-Hyeok Lee*[1], Su-Jin Park*[2], and Gun-Woo Kim**

본 논문은 교육부와 경상남도의 재원으로 지원을 받아 수행된 경상남도 지역혁신중심 대학지원체계(RISE) 연구결과 및 2025년도 산업통상자원부 및 한국산업기술기획평가원(KEIT)의 연구비 지원(RS-2025-02633048)에 의한 연구결과임

요 약

보이스피싱 수법은 시간이 지나면서 고도화되고 있으나, 데이터 희소성과 모델의 블랙박스 한계로 인해 사용자가 모델의 탐지 결과를 신뢰하기 어렵게 한다. 이러한 문제를 해결하기 위해 합성 데이터 증강과 설명 가능 인공지능(XAI)을 결합한 온디바이스 텍스트 기반 보이스피싱 탐지 시스템을 제안한다. 대형 언어 모델(LLM)로 신종 수법 패턴의 합성 데이터를 생성하였으며, 탐지 모델에 Attention 메커니즘을 적용하여 주요 토큰을 추출하고 경량 LLM 기반의 설명 모델과 결합하여 분류 근거를 생성한다. 실험 결과, 탐지 모델인 LSTM-Attention과 KoBERT는 30초 이상 길이의 통화에서 99.5% 이상의 F1-score를 보였으며, 설명 모델인 Gemma 2 2B는 BLEU-4 28.69, ROUGE-1/2/L 각각 62.3, 44.41, 54.76의 성능을 달성하였다. 모델의 정확도와 지연시간 및 모델 크기 간에는 상충 관계가 존재하여 시스템 제약에 맞는 모델 선정이 중요함을 시사한다.

Abstract

Voice phishing scams are becoming more sophisticated, while data sparsity and black-box models hinder user trust in automated detection. we propose an on-device, text-based detection system that combines LLM-based synthetic data augmentation and explainable AI, where an LLM generates synthetic examples of new scam patterns, an LSTM-Attention detector highlights salient tokens, and a lightweight LLM explainer produces rationales, and in experiments LSTM-Attention and KoBERT achieved F1-scores above 99.5% on calls longer than 30 seconds while the Gemma 2 2B explainer reached BLEU-4 28.69 and ROUGE-1/2/L 62.3/44.41/54.76, highlighting trade-offs between accuracy, latency, and model size for on-device deployment.

Keywords
voice phishing, XAI, synthetic data generation, NLP

* 경상국립대학교 컴퓨터공학과 학사과정
- ORCID[1]: https://orcid.org/0009-0004-3679-4540
- ORCID[2]: https://orcid.org/0009-0008-2573-3901
** 경상국립대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-5643-4797

· Received: Sep. 24, 2025, Revised: Nov. 18, 2025, Accepted: Nov. 21, 2025
· Corresponding Author: Gun-Woo Kim
Dept. of ComputerScience and Engineering, College of IT Engineering, Gyeongsang National University, JinJu, Korea
Tel.: +82-55-772-3323, Email: gunwoo.kim@gnu.ac.kr

Ⅰ. 서 론

보이스피싱(전기통신금융사기)은 전화 등 전기통신수단을 이용하여 타인을 속이거나 협박하여 자금을 송금·이체하도록 하거나 개인정보를 탈취하는 범죄를 의미한다. 최근 몇 년 사이 보이스피싱의 발생 건수는 감소했지만 피해 규모는 오히려 커지는 양상을 보인다. 한국 경찰청 통계에 따르면 2019 발생 건수는 37,667건이었으나 2024년에는 18,902건으로 줄었음에도, 2024년 피해액은 8,545억으로 역대 최고를 기록했다[1]. 또한 보이스피싱 범죄 수법은 금융기관 사칭, 검찰 사칭과 같은 통화상 직접적인 송금 유도가 많았으나, 최근에는 개인정보 탈취, 악성 앱 및 링크 클릭 유도와 같은 간접·우회형 수법이 나타나고 있다.

이러한 보이스피싱 수법 변화에 대응하기 위해 탐지 모델 역시 신종 수법의 패턴을 학습할 필요가 있다. 그러나 개인정보 보호 문제로 인해 신종 수법에 대한 데이터는 희소하여 데이터 확보가 어렵다는 문제가 있다. 한편, 기존 AI 기반 보이스피싱 탐지 모델과 기존 통신사의 보이스피싱 탐지 시스템 모두 높은 성능을 보고하고 있으나[2][3]. 내부 의사결정 과정을 명확히 설명하기 어려운 블랙박스 한계로 인해 탐지 결과에 대한 해석이 어렵고, 이는 탐지 결과에 대한 사용자 신뢰를 저하할 수 있다.

본 연구는 이러한 한계를 극복하기 위해 텍스트 분석 기반의 온디바이스 XAI(On-device eXplainable Artificial Intelligence) 보이스피싱 탐지 시스템을 제안한다. 그 과정에서 LLM(Large Language Models) 기반의 합성 데이터 생성 방식을 통해 학습 데이터셋을 확장하여 신종 수법의 데이터 희소성을 완화하였으며, 경량 LLM을 통해 탐지 결과에 대한 근거를 사용자에게 제시함으로써 결과의 해석 가능성과 시스템에 대한 사용자 신뢰를 제고한다. 아울러, 양자화 전후 모델 크기와 모델 지연시간을 비교 및 분석하여 온디바이스 환경에서의 실행 가능성을 분석하고자 한다. 실험 결과, 제안한 시스템은 신종 수법을 포함한 보이스피싱을 정확하게 탐지하고 수행할 수 있는 기반을 마련하였으며, 경량 설명 모델을 통한 탐지 근거를 제공할 수 있음을 확인하였다. 또한 정확도와 운영상 고려 대상(지연 시간, 모델 크기) 간의 상충 관계를 비교 및 분석하여 기기 제약을 고려한 온디바이스 보이스피싱 탐지 앱의 적용 가능성을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서 활용된 기술과 관련 연구를 소개한다. 3장에서는 시스템의 설계와 구성 요소, 합성 데이터 기반 데이터셋 구축 및 전처리, 적용된 탐지 모델과 설명 모델의 아키텍처에 관해서 설명한다. 4장에서는 모델 실험과 성능을 평가하고, 시스템 실행 결과를 논의한다. 마지막으로 5장에서는 결론과 함께 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

2.1 텍스트 기반 보이스피싱 탐지 모델

보이스피싱 문제를 해결하기 위하여 자동음성인식(Automatic speech recognition) 전사 텍스트를 입력으로 머신러닝·딥러닝 모델을 사용한 다양한 연구들이 존재한다. M. K. M. Boussougou et al.[4]은 한국어 보이스피싱 데이터셋인 KorCCVi를 구축하고, 텍스트 기반 이진 분류에서 전통적 머신러닝 및 딥러닝 모델 모두가 높은 성능을 보였다. S. Yu et al.[5]은 NER(Named Entity Recognition)과 문장 단위 N-gram 특징을 결합하여 여러 머신러닝 모델의 탐지 성능을 향상할 수 있음을 보여주었다.

앞서 소개한 두 연구는 보이스피싱 탐지 작업에 대규모 딥러닝 모델이 필수적이지 않음을 시사하며, 특히 전통적 머신러닝은 파라미터 규모와 연산량이 적어 온디바이스 환경에서의 적용 부담이 낮다는 점을 뒷받침한다.

2.2 합성 데이터

합성 데이터는 실제 데이터의 통계적 특성과 구조를 반영하여 알고리즘이나 생성 모델을 통해 인공적으로 생성된 데이터를 의미한다. 공개된 데이터가 제한적인 보이스피싱 탐지와 같은 도메인에서 희소성·불균형 문제를 완화하기 위한 대안으로 주

목받고 있다. E. Igba et al.[6]은 생성 모델을 활용한 금융 사기 데이터 증강을 논의하는 한편, 합성 데이터의 편향·품질 문제를 최소화할 필요성을 강조하였다. J. Y. Sim et al.[7]은 보이스피싱 데이터 확보를 위해 가상 녹취록 생성을 시도하였으나 실제 데이터와의 도메인 괴리가 있어, 기존 사례의 공통패턴을 기반으로 수작업 생성하였다. Z. Shen et al.[8]은 LLM 기반 실시간 보이스피싱 탐지에서 실제 데이터와 합성 데이터의 성능을 비교한 결과, 합성 데이터에서 거짓 양성이 증가하여 정밀도가 하락함을 보고하였으며, 그 원인으로 특정 의심 키워드가 포함된 경우, 보이스피싱으로 분류되는 키워드 기반 분류 편향에 기인한다고 분석하였다.

이러한 선행 연구는 합성 데이터가 데이터의 희소성·불균형 문제를 완화하여 탐지 성능 향상에 기여할 수 있음을 보여주며, 동시에 양질의 데이터를 얻기 위해서는 패턴 기반 생성과 사후 정제 과정이 필수임을 시사한다.

2.3 XAI 모델

XAI는 인공지능 모델의 의사결정 과정을 인간이 이해할 수 있도록 제시하는 방법론이다[9]. 피싱 탐지 분야에서 XAI 기법의 적용 사례가 보고되고 있다. P. R. G. Hernandes et al.[10]은 URL 기반 피싱 탐지 연구에서 LIME(Local Interpretable Model-agnostic Explanations)[11]과 EBM(Explainable Boosting Machine)[12]을 적용하여 블랙박스 모델의 한계를 보완할 수 있음을 보고하였다. 또한 S. R. Alotaibi et al.[13]은 IoT-클라우드 환경에서 LIME으로 의사결정 근거를 제시하며, 사용자 신뢰가 중요한 애플리케이션에서 투명성의 가치를 강조하였다. M. A. Uddin et al.[14]은 LIME과 Transformer-Interpret를 결합하여 각 토큰의 기여도를 시각화함으로써, 블랙박스 한계를 완화하고 사용자 신뢰를 높일 수 있음을 보고하였다.

이러한 선행 연구는 실시간으로 금전적 피해가 발생할 수 있는 보이스피싱 탐지에서, 사용자에게 명확한 탐지 근거를 제시하여 시스템의 신뢰를 확보하는 것이 중요함을 시사한다.

2.4 온디바이스 AI

온디바이스 AI는 데이터가 별도의 서버를 거치지 않고 기기 내에서 로컬로 AI 기능을 수행할 수 있는 기술이다. 따라서 네트워크 독립적이므로 통화 음성과 같은 민감한 데이터가 외부로 유출되지 않는 장점이 존재한다. 이러한 장점들로 인해 보이스피싱 탐지 연구에서 민감 통화 데이터를 단말 내에서 처리하며 실시간 경고를 가능하게 하는 접근으로 주목받고 있다. C. Lee et al.[15]은 콘텐츠를 서버로 전송하지 않고 단말 내부에서 런타임 권한 요청을 실시간 추적하여 보이스피싱 악성 앱을 조기 탐지하는 온디바이스 접근법을 제시하였다. 한편, S. Park et al.[16]은 통화 중 온디바이스 ASR(Automatic Speech Recognition)과 텍스트 분류를 결합한 프로토타입을 제시하였으며, 사용자 관점에서는 온디바이스 AI임에도 일부 사용자에게서 개인정보 보호를 우려하는 점이 관찰되어 사용자의 이해를 높이는 것이 중요함을 보고하였다. 또한 백그라운드를 통한 AI 추론의 오버헤드가 존재함을 나타냈다.

이러한 선행 연구는 온디바이스 AI는 보이스피싱 탐지와 같은 연구 맥락에서 개인정보보호 장점을 제공함과 동시에, 사용자 신뢰를 높이는 설명 제공과 효율적인 모델 설계를 위한 모델 경량화의 필요성이 병행되어야 함을 시사한다.

III. 시스템 설계 및 구현

3.1 시스템 설계 및 구성

본 절에서는 제안된 실시간 보이스피싱 탐지 시스템의 전체적인 시스템 구조에 관해서 설명한다. 그림 1은 본 연구에서 제안하는 시스템 아키텍처를 보여준다. 해당 시스템은 클라이언트-서버 구조를 기반으로 하며 실시간으로 통화 탐지, 음성 녹음 및 추출, 탐지·XAI 모델 호출, 탐지 결과 인터페이스 등을 제공하고 있다. 특히 온디바이스 AI를 적용하여 사용자의 개인정보 유출을 최소화하고 있으며, 탐지 결과를 서버에 추가 저장하는 동의 유무에 따라 보이스피싱 관련 통계 정보 등을 제공한다.

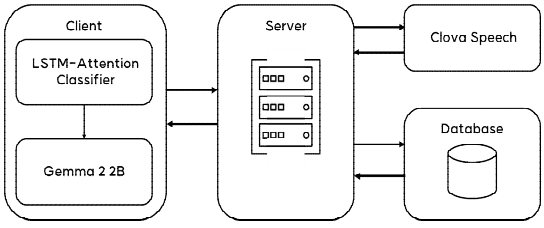

그림 1. 시스템 아키텍처
Fig. 1. System architecture

그림 2는 통화 수발신부터 탐지 결과 출력까지의 과정을 나타낸다. 먼저 클라이언트 애플리케이션에서 사용자의 연락처에 등록되지 않은 전화번호로 수발신한 경우 통화 녹음을 시작하며 녹음된 음성에서 텍스트를 추출한다. 이후 텍스트 추출을 위해 시스템 내에서 Clova Speech API를 사용한다. 추출한 텍스트는 전처리한 후 탐지 모델과 XAI 모델의 입력으로 사용되며, XAI 모델은 탐지 결과와 텍스트 내 핵심 키워드를 바탕으로 설명을 생성한다. 이를 통해 시스템에서 탐지한 모델의 결과와 XAI 모델의 결과를 최종 사용자에게 전달한다.

서버는 크게 두 가지 주요 기능을 수행한다. 첫째, 애플리케이션 사용 중에 발생한 사용자 정보와 보이스피싱 탐지 결과를 데이터베이스에 저장하며 이를 바탕으로 보이스피싱 통계 정보를 생성한다. 둘째, 저장된 사용자 정보를 바탕으로 사용자 인증 및 보안 기능을 제공한다.

먼저 첫 번째 기능에 관해 설명해 보자면, 탐지 결과는 사용자 동의가 있는 경우에만 서버에 저장되도록 설계하였다. 또한 동의한 사용자에게만 보이스피싱 통계 정보를 추가로 제공한다. 이러한 설계는 개인정보 수집과 활용을 최소화하면서, 사용자의 데이터 제공에 따른 부가적인 서비스를 제공함으로써 개인정보 보호와 시스템 운영 간의 균형을 고려하였다.

다음으로 두 번째 기능을 구체적으로 살펴보면, 사용자 인증 및 보안을 위해 JWT(JSON Web Token), HMAC(Hash-based Message Authentication Code), 그리고 HTTPS(Hypertext Transfer Protocol Secure)를 결합하여 구현하였다. 먼저, 클라이언트는 모든 API 요청에 JWT를 포함하며, 서버 측에서는 이를 검증하여 인증을 수행한다. JWT 기반 인증 방식은 경량화된 인증 구조로 설계되어 있지만, 토큰 탈취 시 보안에 취약하다는 단점이 존재한다. 따라서 본 시스템은 JWT의 Refresh Token을 서버 측 DB에 저장하여 보안을 강화하였다. 이와 별도로 비로그인 상태에서도 보이스피싱 탐지 서비스를 제공할 수 있도록 서비스 가입된 사용자 디바이스 식별자와 대칭키 기반의 HMAC 인증 방식을 적용하였다. 위 JWT와 HMAC 두 방식을 통해 비인가 사용자의 API 접근을 효과적으로 제한하면서, 보이스피싱 탐지 서비스는 로그인 없이 지속적으로 이용할 수 있도록 설계하였다. 마지막으로 앞서 언급한 인증 방식들은 MITM(Man-in-the-Middle Attack, 중간자 공격)에 취약하므로, 클라이언트-서버 간 모든 통신에 HTTPS를 도입하여 네트워크 보안을 강화하였다. 이러한 다층 보안 체계는 시스템 전반의 안정성과 신뢰성을 확보한다.

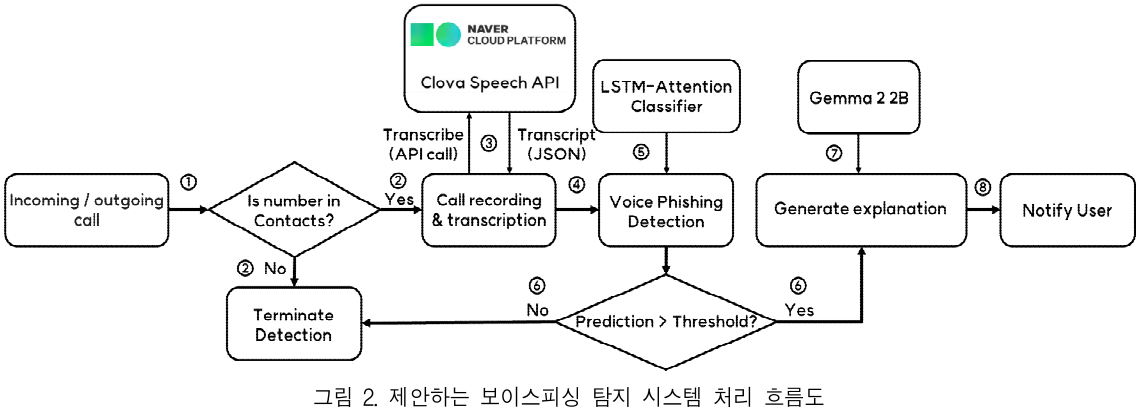

그림 2. 제안하는 보이스피싱 탐지 시스템 처리 흐름도
Fig. 2. Processing flow of the proposed voice phishing detection system

3.2 데이터셋 구축 및 합성 데이터 생성

보이스피싱 탐지 모델에 사용한 전체 데이터셋의 상세 구성은 다음 표 1과 같다. 한국 금융감독원(FSS)에서 제공하는 데이터셋에는 보이스피싱의 수법 중 금융기관 사칭 및 검찰 사칭에 대한 보이스피싱 음성 데이터가 포함되어 있다 [17]. 아울러 KorCCVi v2 데이터셋의 보이스피싱 데이터와 정상 데이터는 각각 FSS와 국립국어원(NIKL)에서 수집한 데이터로 구성되어있다 [18]. 두 데이터셋의 보이스피싱 항목은 둘 다 같은 FSS 출처지만 서로 다른 STT(Speech-To-Text)를 사용하여 같은 통화 음성이라도 전사된 텍스트가 다르다. 따라서 중복 검사를 위해 Sentence-Transformers[19]로 문장 임베딩을 계산하고, 코사인 유사도 0.70 이상인 쌍을 후보로 자동 선별하였다. 이후 전수 수작업 대조를 통해 실제 중복 여부를 확인하여 332개를 제외하였다. 그 결과 KorCCVi v2의 보이스피싱 데이터에서는 360개만 사용하였다.

표 1. 전체 데이터셋
Table 1. Data sets

Source	Vishing	Not vishing	Total
FSS[17]	410	0	410
KorCCVi v2[18]	360	2232	2592
Synthetic data	2000	0	2000
Total	2770	2232	5002

데이터 수집 과정에서 실제 보이스피싱 데이터는 개인정보 및 보안상의 이유로 공개가 제한적이기 때문에 데이터양이 부족했으며, 특히 신종 수법에 대한 보이스피싱 대화 데이터는 거의 확보되지 않아 탐지 모델의 학습에 한계가 있었다. 이러한 한계를 보완하고 모델의 일반화 성능을 향상시키기 위해 신종 수법 기반의 합성 데이터를 추가적으로 생성하였다. 그림 3은 합성 데이터 생성 과정을 나타낸다. 합성 데이터를 생성하기 위해 실제 보이스피싱 사례에 관한 기사, 신고 사례 등 다양한 온라인 자료를 분석하였으며 그 결과, 신종 수법은 크게 3가지(개인정보 탈취형, ARS 기반 사칭형, 악성 앱 클릭 및 유도)로 나타났다. 각 유형에 대해 보안 강화 사칭, 서류 누락 및 보안 유도형 사칭, 통신사 및 앱 보안 사칭, 택배 및 쇼핑몰 사칭 등 세부 시나리오를 구성한 뒤, LLM인 GPT-4.0을 이용하여 총 2,000개를 생성하였다.

3.3 평균 음절 개수 기반 데이터 세분화

효율적인 보이스피싱 탐지를 위해서는 적절한 분석 시간 구간 설정이 필요하다. 너무 짧은 구간은 보이스피싱 발화의 특징을 충분히 담지 못할 수 있고, 반대로 과도하게 긴 구간은 모델이 분석하기 전에 보이스피싱 피해가 발생하거나 사용자에게 긴 통화 시간을 요구한다.

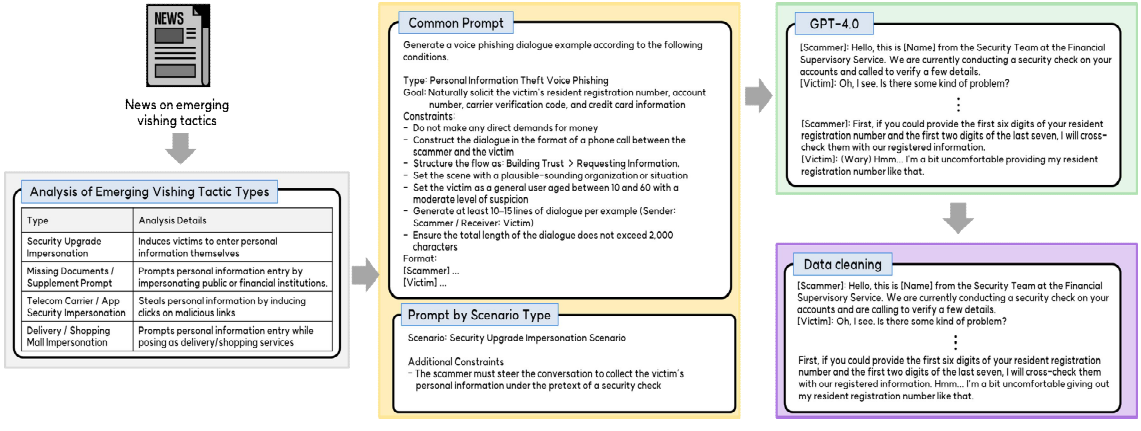

그림 3. 합성 데이터 생성 과정 파이프라인(예시)
Fig. 3. Process of synthetic data generation pipeline(Example)

본 연구에서는 보이스피싱 탐지를 위한 적절한 통화 구간을 선정하기 위해 통화 시간대별 평균 음절 개수를 사용하였다. 평균 음절 개수는 금융감독원의 보이스피싱 음성 데이터를 활용하여 10초 단위로 음성을 자른 뒤 음절 개수를 카운트하였다. 통화 시간대별 평균 음절 개수는 그림 4를 통해 확인할 수 있다.

이렇게 카운트한 통화 시간대별 평균 음절 개수를 기준으로 수집한 데이터셋을 세분화하였다. 그리고 각 구간 단위로 보이스피싱 탐지 성능을 분석하여 탐지 성능과 응답 반환 시간 사이에서 최적의 구간을 도출하고자 하였다.

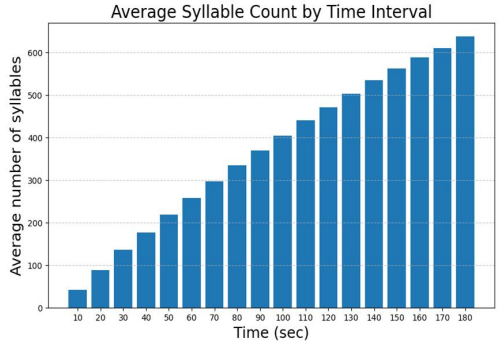

그림 4. 통화 시간대별 평균 음절 개수
Fig. 4. Average syllable count by time interval

3.4 데이터 전처리

보이스피싱 탐지 모델을 학습하기에 앞서, 모델의 탐지 성능 향상을 위해 세분화한 보이스피싱 데이터와 합성 데이터에 대한 전처리 과정을 수행하였다. 보이스피싱과 정상 데이터에서 비식별화를 위해 삽입된 영문자(예: OOO)와 특수 기호를 제거하고 학습 데이터셋에 과적합하는 것을 방지하기 위해 보이스피싱 및 합성 데이터 특유의 인사말과 인물명을 삭제 처리하였다. 특히 합성 데이터 생성 과정에서 반복적으로 생성된 특정 업체명을 본문에서 지움으로써 텍스트 내 불필요한 잡음을 줄이고, 탐지 모델이 의미 있는 패턴 학습에 집중할 수 있도록 하였다.

3.5 탐지 모델 및 설명 모델 설계

그림 5는 시스템에 탑재된 보이스피싱 탐지 모델의 아키텍처를 나타낸다. 그림의 T는 유효 토큰 길이를 의미하며, 훈련 데이터 토큰 수 분포의 95% 분위수를 입력 길이 상한으로 설정하였다. 해당 모델의 입력으로 들어오는 임베딩 벡터를 만들기 위하여 한국어 FastText 사전 학습 모델을 사용하여 변환하였다. 변환한 임베딩 벡터를 바탕으로 보이스피싱 분류를 위해 LSTM[20]을 백본으로 사용하였으며 Attention Layer[21]를 추가로 도입하여 모델이 중요하게 판단한 단어를 추출하도록 하였다. 추출된 단어는 추후 설명 모델의 입력으로 활용되어 탐지 결과에 대한 해석을 제공함으로써 탐지 모델의 해석 가능성(Interpretability)을 확보하였다.

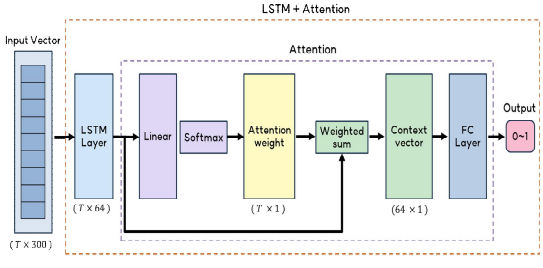

그림 5. 보이스피싱 탐지 모델 아키텍처
Fig. 5. Model architecture for detecting voice phishing

그림 6은 설명 모델의 입력 구성과 생성 절차를 보여준다. 설명 모델은 LLM 모델 기반으로 통화 전사, 탐지 결과 그리고 Attention Layer를 통해 추출한 핵심 토큰(최대 5개)을 입력으로 받아 탐지 결과에 대한 설명을 생성한다. 자연스럽고 일관성 있는 설명을 생성하기 위해 여러 경량 모델을 선정하여 LoRA[22] 기반 미세조정(Fine-tuning)을 수행하였다. 이때 Qwen2.5 7B Instruct[23]를 teacher 모델로 사용하여 기준 설명을 생성하고, 이를 학습 지도 신호 및 평가 기준으로 활용하였다. 최종적으로 Gemma 2 2B를 설명 모델로 선정하였으며 양자화(8bit)하여 시스템에 적용하였다. LoRA는 베이스 가중치를 고정하고 저랭크(Low-rank) 어댑터만 학습하므로 메모리·연산 비용을 크게 절감할 수 있어 온디바이스 환경에 적합하다.

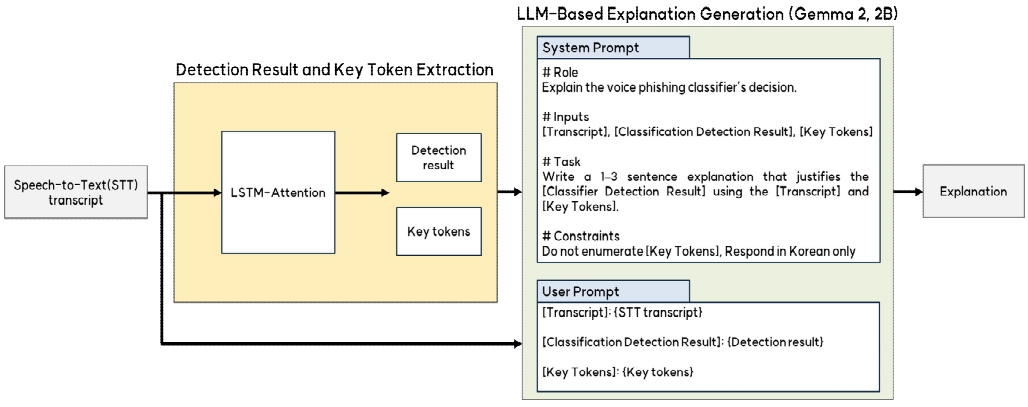

그림 6. LLM 기반 설명 생성 파이프라인(입력 구성 및 절차)
Fig. 6. LLM-based explanation generation pipeline(Inputs and procedure)

Ⅳ. 모델 실험 및 평가

4.1 실험 모델 설정

본 연구에서는 신종 보이스피싱 수법에도 강건한 분류 모델을 만들기 위해 기존 연구에 사용된 모델들을 참고하여 트리 기반 앙상블 모델인 CatBoost[24], LGBM[25], 한국어 사전학습 언어모델 KoBERT[26], 그리고 시계열 특화 모델인 LSTM을 사용하였다. 선정한 후보 모델에는 모두 Attention Layer를 추가하였다. CatBoost와 LGBM의 경우 FastText 임베딩 시퀀스에 선형 Attention 풀링 적용하여 가중합된 문장 벡터를 입력 피처로 사용하는 방식으로 하였다. 후보 모델에 대해 비교 실험을 진행하여 모델들의 재현율(Recall)과 정밀도(Precision)를 기준으로 최적의 분류 모델을 도출하고자 한다. 설명 모델은 온디바이스 환경을 고려하여 경량 모델인 TinyLlama 1.1B Chat v1.0[27], Gemma 2 2B[28], Qwen2.5 1.5B Instruct[29]를 실험 모델로 선정하였다. 설명 모델의 평가지표는 BLEU-4[30], ROUGE-1/2 F1, ROUGE-L F1[31]을 종합적으로 평가하여 최적의 설명 모델을 선정하고자 한다.

4.2 보이스피싱 탐지 모델 평가

탐지 모델의 성능 평가는 층화 5-겹 교차검증(stratified 5-fold)으로 수행하였다. 각 폴드에서 1,001개를 테스트로 두고, 나머지 4,001개는 학습 3,200개 검증 801개로 분할하였다. 학습·검증·테스트 구성은 폴드마다 달라지며, 모든 분할에서 클래스 비율을 유지하였다. 모델의 최종 성능은 각 폴드 테스트 지표의 평균을 성능 지표로 사용했다. 표 2는 보이스피싱 탐지 모델로 선정된 LGBM, CatBoost, LSTM, KoBERT을 전사 길이에 따른 성능을 수치(Recall, Precision, F1-score)로 제시한다. 그림 7은 동일한 설정에서 모델별 추세를 시각화하고, 표에 없는 ROC-AUC까지 함께 보여준다.

모든 모델이 높은 성능을 나타냈지만, 머신러닝 모델 계열인 CatBoost와 LGBM은 특정 구간 이후로 낮게 수렴하는 경향을 보인다. 따라서 추가적인 통화 내용이 들어오더라도 성능이 개선되지 않음을 나타내고 있다. LSTM과 KoBERT는 30초와 60초 구간의 통화 내용으로 대부분 잘 분류할 수 있음을 보여주고 있으며, 이는 통화 시간이 길지 않더라도 보이스피싱의 핵심 패턴을 충분히 감지할 수 있음을 시사한다.

표 3은 각 폴드의 테스트 샘플을 합친 동일 표본에서 LSTM과 KoBERT를 McNemar 검정[32]을 사용하였으며 p-value는 양측 mid-p[33]로 계산하였다. 30초 구간과 60초 구간 모두에서 KoBERT만 정답인 경우(c)가 LSTM만 정답인 경우(b)보다 많았으며, 유의수준 5%에서 통계적으로 유의함을 보였다. 이는 30·60초의 짧은 통화 길이에서도 KoBERT가 LSTM 대비 일관되게 더 정확함을 시사한다.

표 2. 전사 길이(초)에 따른 탐지 모델별 성능 비교
Table 2. Performance comparison of detection models by transcript duration (s)

Duration	LGBM			CatBoost			LSTM			KoBERT		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
10 s	0.957	0.971	0.964	0.955	0.975	0.964	0.983	0.984	0.984	0.996	0.997	0.996
20 s	0.973	0.971	0.972	0.97	0.978	0.974	0.992	0.996	0.994	0.999	0.999	0.999
30 s	0.976	0.976	0.976	0.978	0.977	0.978	0.995	0.996	0.996	1.0	1.0	1.0
60 s	0.984	0.987	0.985	0.979	0.991	0.985	0.999	0.998	0.998	0.999	1.0	1.0
120 s	0.982	0.986	0.984	0.982	0.987	0.985	0.998	1.0	0.999	1.0	1.0	1.0
180 s	0.984	0.99	0.987	0.986	0.985	0.986	0.998	0.999	0.998	1.0	1.0	1.0

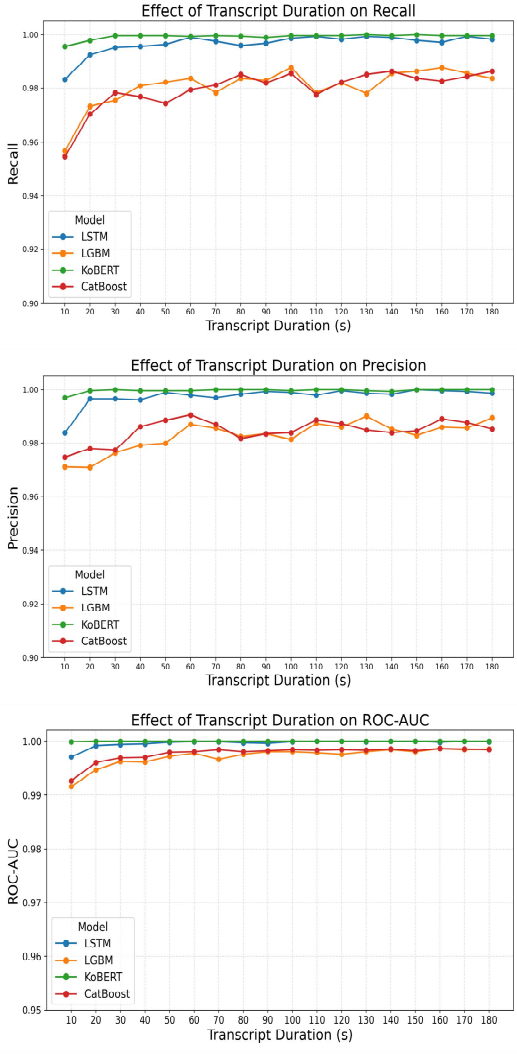

그림 7. 전사 길이(초)에 따른 탐지 모델의 재현율(위), 정밀도(가운데), ROC-AUC(아래)
Fig. 7. Detection performance by transcript duration (s): Recall (top), Precision (middle), and ROC-AUC (bottom)

표 3. 30초 및 60초 구간에서 LSTM - KoBERT McNemar 검정 결과(동일 표본, 양측 mid-p)
Table 3. McNemar's test(paired, two-sided mid-p) for LSTM vs. KoBERT at 30 s and 60 s

Duration	Sample size	LSTM-only correct (b)	KoBERT-only correct (c)	McNemar mid-p
30 s	5002	5	23	5.46×10^{-4}
60 s	5002	0	13	1.22×10^{-4}

4.3 설명 모델 평가

설명 모델의 성능 평가는 Qwen2.5 7B Instruct Teacher 모델로 사용해 생성한 기준 설명 대비 후보 설명 모델 출력 결과의 텍스트 유사도 기반으로 측정하였다. 평가 과정에서 토크나이저는 KiwiPiePy 한국어 토크나이저를 사용했으며, 기준 설명을 이용하여 LoRA 기반 미세조정을 진행한 후 설명 모델 평가를 수행하였다. 표 4는 해당 지표 기반의 평가 결과를 요약한다.

표 4. 설명 모델 텍스트 유사도 평가 결과
Table 4. Results of textual similarity evaluation for the explanation model

Model	BLEU-4	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
TinyLlama 1.1B Chat v1.0	21.3	53.55	40.12	48.37
Gemma 2 2B	28.69	62.3	44.41	54.76
Qwen2.5 1.5B Instruct	24.01	58.01	39.15	49.79

지표로 사용된 BLEU-4, ROUGE-N F1(N=1, 2), ROUGE-L F1 수식을 표현하면 다음과 같다.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases} \quad (1)$$

$$BLEU = BP \times exp(\sum_{n=1}^{N} w_n \log p_n)$$

$$ROUGE_N\, recall = \frac{\sum_{S \in R} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in R} \sum_{gram_n \in S} Count(gram_n)}$$

$$ROUGE_N\, precision = \frac{\sum_{S \in R} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in G} \sum_{gram_n \in S} Count(gram_n)} \quad (2)$$

$$ROUGE\text{-}L_{recall} = \frac{\Sigma_j\ LCS(X_j,\ Y_j)}{\Sigma_j\ |X_j|} \quad (3)$$

$$ROUGE\text{-}L_{precision} = \frac{\Sigma_j\ LCS(X_j,\ Y_j)}{\Sigma_j\ |Y_j|}$$

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision} \quad (4)$$

BLEU-4는 1~4-그램 정밀도의 기하평균과 과도한 축약(생성한 설명이 기준 설명보다 짧음)에 대한 페널티를 반영한다. 따라서 점수가 높다는 것은 생성한 후보 설명이 기준 설명과 n-그램 수준에서 많이 겹치고, 지나치게 짧지 않음을 시사한다. ROUGE-1/2 F1은 각각 기준 설명과의 1-그램, 2-그램 정밀도와 재현율의 조화평균을 계산한 값이다. 값이 높을수록 핵심 단어를 충분히 포함하면서 불필요한 단어가 적음을 의미한다. ROUGE-L F1은 최장 공통 부분 수열(LCS)에 기반해 순서를 유지한 겹침 정도를 측정하므로, 전개 순서의 유사성을 시사한다.

실험 결과, Gemma 2 2B가 모든 지표에서 가장 높은 점수를 기록하였으며, 이는 해당 모델이 Teacher 모델의 설명과 가장 유사한 설명을 생성했음을 의미한다.

4.4 모델 복잡도 및 계산 복잡도

후보 탐지 모델과 설명 모델이 온디바이스에서 실행 가능한지 파악하기 위하여 모델의 파라미터 수(Params)와 모델의 용량(Model size), 추론 연산량(FLOPs)을 평가하고자 한다. 표 5와 6은 각각 탐지 모델과 설명 모델에서 모델 규모 및 추론 연산량 수치를 제시한다. 탐지 모델의 성능은 각 폴드의 모델 결과를 평균하였으며, 통화 길이에 따른 모델의 연산량을 비교하기 위해 30초와 60초에서의 추론 연산량 성능을 측정하였다. 설명 모델은 양자화 전·후를 비교하기 위해 GGUF 포맷의 4-bit 양자화 가중치(Q4_K_M)를 적용하고. 프롬프트 토큰 수와 생성 토큰 수에 따른 연산량도 함께 분석했다.

탐지 모델의 추론 연산량은 LGBM, CatBoost, LSTM, KoBERT 순으로 적으며, 30초에서 60초로 입력 길이를 늘리면 평균 약 1.8배 증가하였다. 설명 모델의 추론 연산량은 TinyLlama 1.1B Chat v1.0 모델이 Gemma 2 2B 대비 약 60.7%, Qwen2.5 1.5B Instruct 대비 약 32.5% 적어 가장 낮았다. 또한 탐지 모델과 설명 모델에 양자화를 적용 후 모델 크기는 기존 대비 대략 67~70% 감소하였다.

표 5. 30초 및 60초 구간에서 탐지 모델별 모델 규모 및 추론 연산량 비교
Table 5. Comparison of model scale and inference workload of detection models at 30 s and 60 s

Duration	Model (variant)	Model size (MB)	FLOPs	Params (M)
30 s	CatBoost (N/A)	2.116	6.08×10^4	0.135
	LGBM (N/A)	1.416	5.80×10^4	0.016
	LSTM (FP32)	0.362	1.07×10^7	0.094
	LSTM (INT8)	0.119		
	KoBERT (FP32)	351.737	1.91×10^{10}	92.188
	KoBERT (INT8)	107.566		
60 s	CatBoost (N/A)	2.258	1.11×10^5	0.144
	LGBM (N/A)	1.032	1.09×10^5	0.018
	LSTM (FP32)	0.362	2.00×10^7	0.094
	LSTM (INT8)	0.119		
	KoBERT (FP32)	351.737	3.58×10^{10}	92.188
	KoBERT (INT8)	107.566		

표 6. 설명 모델별 모델 규모 및 추론 연산량 비교
Table 6. Comparison of model scale and inference workload of explanation models

Model (variant)	Model size (MB)	Prompt tokens	Generated tokens	FLOPs
TinyLlama 1.1B Chat v1.0 (FP16)	2099.05	128	128	5.36×10^{11}
			256	8.08×10^{11}
		512	128	1.36×10^{12}
			256	1.64×10^{12}
TinyLlama 1.1B Chat v1.0 (GGUF Q4_K_M)	636.88	128	128	5.36×10^{11}
			256	8.08×10^{11}
		512	128	1.36×10^{12}
			256	1.64×10^{12}
Gemma 2 2B (FP16)	4992.69	128	128	1.37×10^{12}
			256	2.06×10^{12}
		512	128	3.45×10^{12}
			256	4.16×10^{12}
Gemma 2 2B (GGUF Q4_K_M)	1629.43	128	128	1.37×10^{12}
			256	2.06×10^{12}
		512	128	3.45×10^{12}
			256	4.16×10^{12}
Qwen2.5 1.5B Instruct (FP16)	2950.35	128	128	7.96×10^{11}
			256	1.20×10^{12}
		512	128	2.01×10^{12}
			256	2.42×10^{12}
Qwen2.5 1.5B Instruct (GGUF Q4_K_M)	940.37	128	128	7.96×10^{11}
			256	1.20×10^{12}
		512	128	2.01×10^{12}
			256	2.42×10^{12}

그림 8과 9는 각각 탐지 모델과 설명 모델의 FLOPs를 기준으로 온디바이스 기기의 초당 조 단위의 연산(TOPS)를 이용하여 환산했을 때의 지연시간을 나타낸다. 탐지 모델의 경우 KoBERT를 제외한 나머지는 대부분은 1 TOPS 미만에서도 사실상 무시 가능한 수준의 낮은 지연시간을 나타내고 있다. 반면 KoBERT의 경우 추론 연산량이 크기 때문에 상대적으로 매우 높은 것을 볼 수 있다. 설명 모델의 경우 파라미터 수가 작은 TinyLlama 1.1B도 1 TOPS에서는 지연시간이 약 1,000ms 정도로 탐지 모델보다 매우 크며, 설명 성능이 가장 좋은 Gemma 2 2B는 지연시간이 가장 길며 1 TOPS에서는 수천 ms로 실시간 동작은 힘듦을 나타낸다. 따라서 정확도(탐지/설명), 지연시간 간의 상충 관계가 있기에 지원하고자 하는 기기의 성능 범위, 지연시간 상한선, 보이스피싱 범죄 예방 효과에 맞춰 적절한 모델을 선택할 필요가 있다.

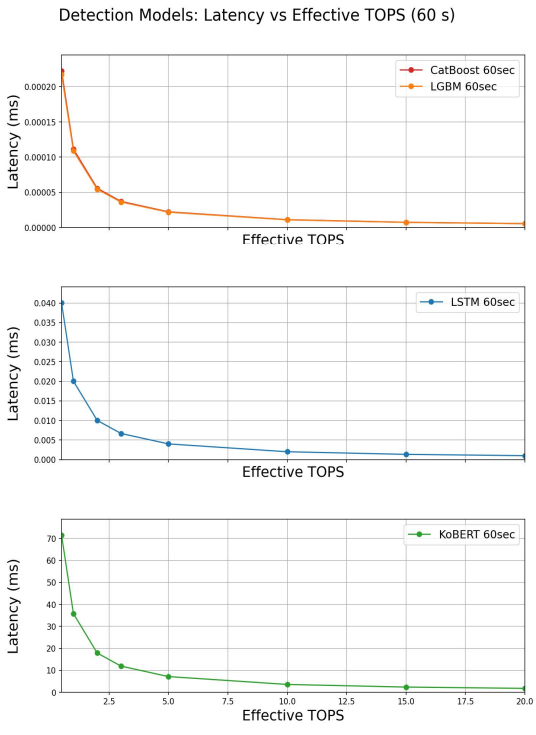

그림 8. 60초 구간 탐지 모델에서 디바이스 유효 TOPS에 따른 탐지 모델 지연시간
Fig. 8. Detection models: latency vs. device effective TOPS at 60 s

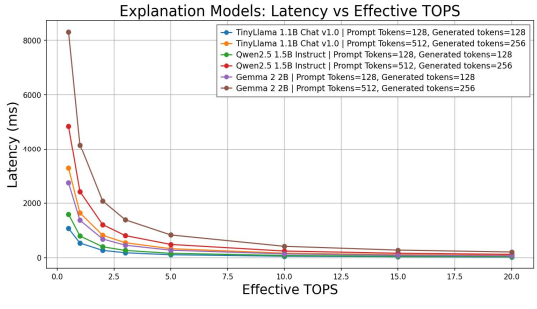

그림 9. 디바이스 유효 TOPS에 따른 설명 모델 지연시간 (프롬프트/생성 토큰 수 조합별)
Fig. 9. Explanation models: latency versus device effective TOPS (by prompt and generated token count combinations)

4.5 오분류 사례 및 원인 분석

표 7은 LSTM 기반 탐지 모델이 60초 통화 내용에서 오탐지 또는 미탐지한 사례를 나타낸다. 여기서 Label 항목의 0과 1 값은 각각 정상과 보이스피

싱을 의미하며 Score는 모델이 예측한 점수이다. Key Tokens와 Weight는 탐지 모델이 해당 통화 원문에서 추출한 주요한 토큰과 Attention 가중치를 의미한다. 표의 첫 번째 행은 보이스피싱을 거의 정상으로 분류하였으며, 두 번째 행은 정상을 보이스피싱으로 예측하였다.

모델이 오분류한 원인은 크게 두 가지로 분석하였다. 첫째, 정상 데이터와 보이스피싱 데이터 간 대화 맥락 차이에서 비롯된 토큰 분포 차이이다. 그림 10은 60초 통화에서 각 라벨의 상위 30개 토큰과 상대적 차이를 동시에 보여준다. 정상 데이터에서는 '그/거/뭐/이제/그래서'와 같은 담화표지가 많이 나타나는 반면, 보이스피싱에서는'고객님/확인/본인/주민등록번호'등 신원확인 중심 어휘가 두드러진다. 이로 인해 모델이 특정 키워드에 과의존하여 경계 사례에서 오분류가 발생할 수 있다. 둘째, 통화 음성을 텍스트로 변환하는 STT 품질 저하에 따른 입력 왜곡이다. 띄어쓰기·철자 인식 오류 등으로 인해 핵심 토큰을 왜곡하여 추가적인 오분류를 유발한다.

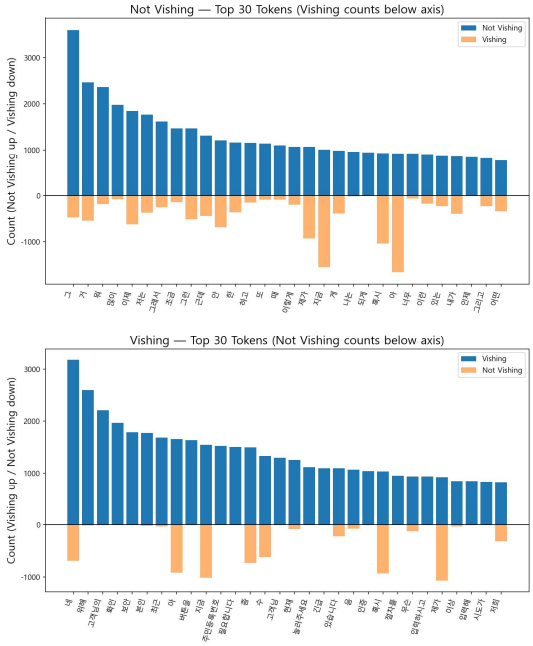

그림 10. 60초 통화에서 라벨별 상위 30개 토큰 분포: 정상(위), 보이스피싱(아래)
Fig. 10. Top 30 tokens on 60 s transcripts: Not Vishing (top), Vishing (bottom)

표 7. 60초 전사에서 LSTM 기반 탐지 모델의 오분류 사례
Table 7. Misclassification examples of the LSTM-based detection model on 60 s transcripts

Label	Score	Transcript (60sec)	Key tokens	Weight
1	0.0062	한 번도 못 들어 본 사람이에요 아니면 한번 정도는 들어봤었던 이름인 거 같아요 한 번도 못 들어본 사람이 맞아요 이제 저희가 체포를 했는데 생략 여쭤본 거고 저 왜냐하면 저희가 최근에 주범으로 하는 금융범죄 사기단들을 7명 검거를 했어요 이 사람들 검거를 하고 사무실 압수수색을 좀 진행을 했는데 사무실에서 압수된 것들이 약 900여 개가 넘는 대포통장들하고 복제된 신용카드 보안카드 그 다음에 위조된 신분증 이런 거 수천장 같이 압수를 했고 지금 압수된 물품들 분류하는 과정에서 씨 명의로 된 은행 계좌를 두 점 저희가 발견을 했고요 이 두 개의 계	같이	0.01282
			넘는	0.01275
			했고	0.01261
			거	0.01258
			좀	0.01256
0	0.9587	허위 봉사에 대해서 어떻게 생각하세요 당연히 하면 안 되겠죠 어떻게 그 시간을 허위로 이렇게 할 수 있으 있는지 이해가 안 됩니다 얼마 되지 않은 시간들을 무엇 때문에 못하는 건지 왜 그래야만 되는지 실로 이해가 안 되네요 그리고 봉사 활동이 왜 필요한 거죠 허위 봉사 같은 경우는 주로 혜택을 많이 받는 연예인이나 국회의원 자녀분들이 있습니다 봉사 활동이 필요한 이유는 대학교 입학이나 취업할 때 유리한 가산점을 얻기 때문입니다 그러나 정작 시간을 투자를 해서 진정한 봉사 활동을 하는 경우는 점수를 그다지 많이 받지 않고 허위 봉사로 인한 허위 점수는 굉장히 고득점으로 알고 있습니다 그래서 이 허위 봉사에 대한 문화와 적폐는 없어	허위	0.01609
			허위	0.01592
			허위	0.01531
			받지	0.01476
			인한	0.01466

4.6 서비스 실행 결과

그림 11은 앱 실행 시 로그인 화면과 약관 동의 화면을 보여준다. 사용자는 사용자 등록 및 통화 안드로이드 접근성 설정 동의 후, 비로그인 상태에서도 보이스피싱 탐지 서비스를 이용할 수 있다.

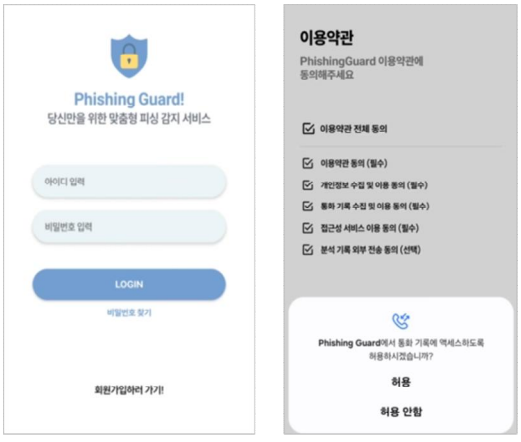

그림 11. 로그인 화면(좌)과 약관 동의 화면(우)
Fig. 11. Login screen(left) and Terms-and-Conditions consent screen (right)

그림 12는 보이스피싱 탐지 실행 화면과 탐지 결과 화면을 보여준다. 연락처에 없는 통화를 수발신하면 통화 녹음과 함께 보이스피싱 탐지를 수행하며, 탐지 시 결과와 근거를 사용자에게 제공한다.

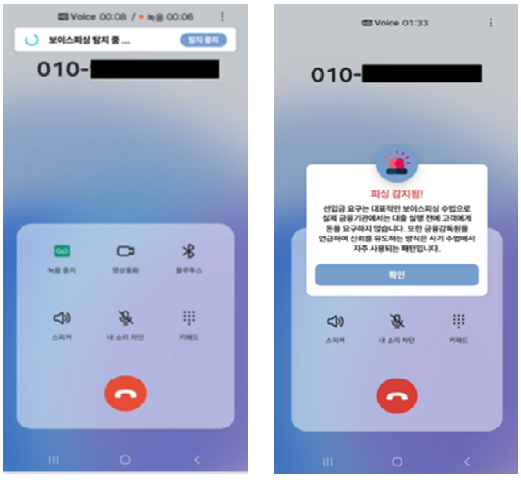

그림 12. 탐지 실행 화면(좌)과 탐지 결과 화면(우)
Fig. 12. Detection-in-progress screen (left) and detection result screen (right)

그림 13은 저장된 탐지 결과를 통해 사용자에게 탐지 기록을 나타내는 기능을 보여준다. 통화 종료 후, 보이스피싱 탐지 유무를 해시태그를 통해 구분할 수 있으며, 보이스피싱의 경우 설명 모델이 생성한 결과를 포함하고 있다.

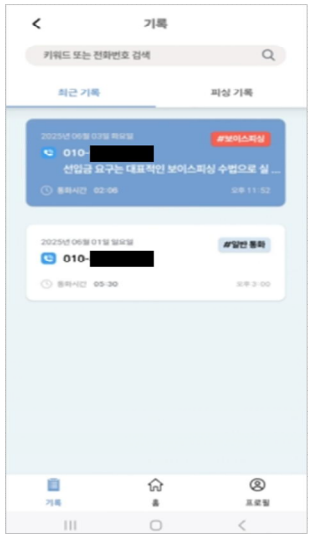

그림 13. 탐지 기록 제공 화면
Fig. 13. Detection history screen

V. 결론 및 향후 과제

보이스피싱 피해 예방 및 모델의 블랙박스 문제를 해결하기 위해 텍스트 분석 기반 온디바이스 XAI 보이스피싱 탐지 시스템을 제안하였다. 제안한 시스템은 사용자가 통화 중에 실시간으로 보이스피싱을 탐지하면서 사용자에게 탐지 이유를 자연어로 설명하도록 하였다. 또한 온디바이스 환경에서 모델 추론을 진행하여 개인정보 보호 및 탐지 기록 관리 기능을 제공함으로써 사용자 신뢰와 사용 편의성을 높였다.

제안한 시스템은 위와 같이 여러 사용자 친화적인 기능을 제공하지만, 한계점도 존재한다. 첫째, 편향된 데이터셋이다. 신종 수법에 강건한 모델을 만들기 위해 합성 데이터를 생성·정제하여 결합했으나 정상 데이터와 보이스피싱 데이터 간의 도메인 차이로 인해 특정 키워드가 포함하지 않는 경우는 오분류가 발생하며, 설명 모델의 결과에도 영향을

주게 된다. 둘째, 완전하지 않은 온디바이스 환경이다. 현재 통화 음성으로부터 텍스트를 추출하기 위해 외부 API를 사용하고 있으며, 이는 온디바이스의 핵심 장점 중 하나인 네트워크 무송신을 통한 정보 유출 최소화를 온전히 보장하지 못하게 한다. 마지막은 온디바이스 환경에서의 성능 파악이다. 온디바이스에서의 모델 실행 가능성을 분석하기 위해 탐지 모델과 설명 모델의 추론 연산량을 기준으로 지연시간을 추정하였으며, 실제 환경에서는 메모리 병목현상, 백그라운드 실행 등 다양한 원인으로 인해 실험 결과와 다를 수 있다.

이러한 한계점을 극복하기 위해 향후 과제는 다음과 같다. 정상 데이터와 보이스피싱 데이터와 도메인 차이를 해소하기 위해 탐지 모델 추론 과정에 사용된 주요 키워드를 제거·완화한 데이터를 생성하거나, 도메인 차이가 나지 않도록 경계(hard negative) 샘플을 추가하여 모델의 강건성을 검증하고자 한다. 또한 온디바이스 Speech-to-Text AI를 적용하여 개인정보유출을 최소화함으로써 사용자 신뢰도를 제고하고자 한다. 마지막으로 실제 온디바이스 환경에서의 성능 비교를 위해 모델을 GPU 가속이 가능한 프레임워크에 맞게 변환하고 최적화한 뒤 온디바이스 환경에서 실험을 진행할 계획이다. 이를 통해 확실한 모델 선정 기준이 마련할 수 있을 것이며, 실시간 통화 환경에서의 사용성을 높여 사용자 경험을 개선할 수 있을 것으로 기대한다.

References

[1] Voice phishing status by the National Police Agency, https://www.data.go.kr/data/15063815/fileData.do. [accessed: Oct. 28, 2025]

[2] M. K. M. Boussougou and D. J. Park, "Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection", Mathematics, Vol. 11, No. 14, Art. 3217, Jul. 2023. https://doi.org/10.3390/math11143217.

[3] AI voice-phishing detection performance of Korean telecom operators (KT/SKT/LGU+), https://www.mk.co.kr/news/it/11380453. [accessed: Oct. 28, 2025]

[4] M. K. M. Boussougou and D. J. Park, "Exploiting Korean Language Model to Improve Korean Voice Phishing Detection", The Transactions of the Korea Information Processing Society, Vol. 11, No. 10, pp. 437-446, Oct. 2022. https://doi.org/10.3745/KTSDE.2022.11.10.437.

[5] S. Yu, Y. Kwon, M. Kim, and K. Lee, "Korean Voice Phishing Detection Applying NER With Key Tags and Sentence-Level N-Gram", IEEE Access, Vol. 12, pp. 52951-52962, Apr. 2024. https://doi.org/10.1109/ACCESS.2024.3387027.

[6] E. Igba, H. S. Olarinoye, V. E. Nwakaego, D. B. Sehemba, Y. S. Oluhaiyero, and N. Okika, "Synthetic Data Generation Using Generative AI to Combat Identity Fraud and Enhance Global Financial Cybersecurity Frameworks", International Journal of Scientific Research and Modern Technology, Vol. 4, No. 2, pp. 1-19, Feb. 2025. https://doi.org/10.5281/zenodo.14928919.

[7] J. Y. Sim and S. H. Kim, "Detecting Voice Phishing with Precision: Fine-Tuning Small Language Models", arXiv:2506.06180. Jun. 2025. https://doi.org/10.48550/arXiv.2506.06180.

[8] Z. Shen, S. Yan, Y. Zhang, X. Luo, G. Ngai, and E. Y. Fu, "It Warned Me Just at the Right Moment: Exploring LLM-based Real-time Detection of Phone Scams", Proc. CHI EA '25 (Extended Abstracts of the CHI Conf. on Human Factors in Computing Systems), Yokohama, Japan, Art. 18, Apr.-May 2025. https://doi.org/10.1145/3706599.3720263.

[9] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. D. Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence", Information Fusion, Vol. 99, Art. 101805, Nov. 2023. https://doi.org/10.1016/j.inffus.2023.101805.

[10] P. R. G. Hernandes, C. P. Floret, K. F. C. D. Almeida, V. C. da Silva, J. P. Papa, and K. A. P. da Costa, "Phishing Detection Using URL-based XAI Techniques", Proc. 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, pp. 1-6, Dec. 2021. https://doi.org/10.1109/SSCI50451.2021.9659981.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, pp. 1135-1144, Aug. 2016. https://doi.org/10.1145/2939672.2939778.

[12] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A Unified Framework for Machine Learning Interpretability", arXiv:1909.09223, Sep. 2019. https://doi.org/10.48550/arXiv.1909.09223.

[13] S. R. Alotaibi, H. K. Alkahtani, M. Aljebreen, A. Alshuhail, M. K. Saeed, S. A. Ebad, W. S. Almukadi, and M. Alotaibi, "Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems", Alexandria Engineering Journal, Vol. 110, pp. 490-505, Jan. 2025. https://doi.org/10.1016/j.aej.2024.09.115.

[14] M. A. Uddin and I. H. Sarker, "An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach", arXiv:2402.13871, Feb. 2024. https://doi.org/10.48550/arXiv.2402.13871.

[15] C. Lee, B. Kim, and H. Kim, "The silence of the phishers: Early-stage voice phishing detection with runtime permission requests", Computers & Security, Vol. 152, Art. 104364, May 2025. https://doi.org/10.1016/j.cose.2025.104364.

[16] S. Park, H. Yoon, J. Kim, H. Kim, and S.-J. Lee, "I know my data doesn't leave my phone, but still feel like being wiretapped: Understanding (Mis)Perceptions of On-Device AI Vishing Detection Apps", Proc. CHI EA '25 (Extended Abstracts of the CHI Conf. on Human Factors in Computing Systems), Yokohama, Japan, Art. 15, Apr.-May 2025. https://doi.org/10.1145/3706599.3719784.

[17] Financial Supervisory Service (FSS), https://www.fss.or.kr/fss/bbs/B0000207/list.do. [accessed: Dec. 26, 2024]

[18] M. K. M. Boussougou, C. K. Hong, S. Hong, and D. J. Park, "Towards Privacy-Preserving Korean Voice Phishing Detection: A Federated Learning Approach with RoBERTa", Proc. Korea Computer Congress (KCC), Jeju, Republic of Korea, pp. 626-628, Jun. 2023.

[19] Sentence-Transformers, https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. [accessed: Nov. 11, 2025]

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", Proc. 31st Conf. on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, pp. 5998-6008, Dec. 2017.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models", Proc. International Conf. on Learning Representations (ICLR), Oline, Apr. 2022.

[23] Qwen2.5 7B Instruct, https://huggingface.co/Qwen/Qwen2.5-7B-Instruct. [accessed: Nov. 11, 2025]

[24] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features", Proc. 32nd Conf. on Neural Information Processing Systems (NeurIPS), Montréal, Canada, pp. 6639-6649, Dec. 2018.

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A

Highly Efficient Gradient Boosting Decision Tree", Proc. 31st Conf. on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, pp. 3149-3157, Dec. 2017.

[26] Korean BERT pre-trained cased (KoBERT), https://github.com/SKTBrain/KoBERT. [accessed: Sep. 03, 2025]

[27] TinyLlama-1.1B-Chat-v1.0, https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0. [accessed: Sep. 03, 2025]

[28] Gemma 2 2B, https://huggingface.co/google/gemma-2-2b. [accessed: Sep. 03, 2025]

[29] Qwen2.5-1.5B-Instruct, https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct. [accessed: Sep. 03, 2025]

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, pp. 311-318, Jul. 2002. https://doi.org/10.3115/1073083.1073135.

[31] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Proc. Text Summarization Branches Out (ACL Workshop), Barcelona, Spain, pp. 74-81, Jul. 2004.

[32] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages", Psychometrika, Vol. 12, No. 2, pp. 153-157, Jun. 1947. https://doi.org/10.1007/BF02295996.

[33] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional", BMC Medical Research Methodology, Vol. 13, Art. 91, Jul. 2013. https://doi.org/10.1186/1471-2288-13-91.

저자소개

이 규 혁 (Gyu-Hyeok Lee)

2020년 2월 ~ 현재 : 경상국립대학교 컴퓨터공학과 학사과정

관심분야 : 인공지능, 데이터 분석, 자연어 처리

박 수 진 (Su-Jin Park)

2022년 2월 ~ 현재 : 경상국립대학교 컴퓨터공학과 학사과정

관심분야 : 인공지능, 멀티모달, Deepfake, Meme Detection

김 건 우 (Gun-Woo Kim)

2006년 12월 : 호주뉴캐슬대학교 컴퓨터공학과(공학사)

2007년 9월 : 호주뉴캐슬대학교 정보공학과(공학석사)

2017년 8월 : 한양대학교 컴퓨터공학과(공학박사)

2021년 9월 ~ 현재 : 경상국립대학교 컴퓨터공학과 부교수

관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝