https://doi.org/10.7840/kics.2025.50.10.1603

개인정보 탐지 기반 Self-Destructing 메신저

소 예 나◆, 이 선 우°

Self-Destructing Messenger Based on PII Detection

Ye-na So◆, Sun-woo Lee°

요 약

본 연구에서는 개인정보 유출 방지를 위해 보안성을 강화한 Self-Destructing 메신저 애플리케이션을 제안한다. 제안된 애플리케이션은 개인정보의 민감도를 4등급으로 분류하고, 실시간 개인정보 탐지 기능을 통해 유출 위험을 사전에 방지한다. 개인정보의 등급에 따라 적절한 보호 조치가 활성화되며, 사용자는 열람 조건(시간, 열람 횟수)을 설정하여 메시지 접근 및 열람을 제어할 수 있다. 이를 통해 민감한 정보가 설정한 조건에 따라 자동 삭제되도록 하여 유출 가능성을 최소화한다. 또한, 본 애플리케이션은 AES 암호화와 Android KeyStore를 적용하여 메시지를 안전하게 저장하며, 삭제 후에도 복구가 어렵도록 보안을 강화한다. 실험 결과, 개인정보의 등급별 탐지 정확도는 96.25%에 달했으며 대표적인 포렌식 도구를 활용한 복구 시도에서도 메시지 복구가 불가능함을 확인하였다.

키워드 : 일회용 메신저, 개인정보 탐지, 메신저 프레임워크
Key Words : Slef-Destructing Messenger, PII Detection, Messenger Framework

ABSTRACT

This study proposes a security-enhanced self-destructing messenger application to prevent personal information leakage. The proposed application classifies personal information into four sensitivity levels and utilizes real-time personal information detection to preemptively mitigate leakage risks. Depending on the sensitivity level, appropriate protection measures are automatically activated. Additionally, users can set access conditions (e.g., time limits and view counts) to control message access and viewing. This mechanism ensures that sensitive information is automatically deleted according to predefined conditions, minimizing the risk of exposure. Furthermore, the application employs AES encryption and Android KeyStore to securely store messages and enhance security by making message recovery difficult after deletion. Experimental results show that the detection accuracy for different sensitivity levels reached 96.25% and attempts to recover deleted messages using leading forensic tools proved unsuccessful.

※ 이 성과(논문)는 정부(교육부)의 지원을 받아 수행된 연구임 (2024년 부처 협업형 인재양성사업[정보보안 분야], No. 2024개인정보보호-002)
※ 이 성과(논문)는 서울여자대학교 교내연구비의 지원을 받아 수행된 연구임 (2025-0014)
◆ First Author : Seoul Women's University, yenas0@swu.ac.kr, 학생회원
° Corresponding Author : Seoul Women's University, sun.lee@swu.ac.kr, 정회원
논문번호 : 202503-063-D-RE, Received March 21, 2025; Revised May 2, 2025; Accepted May 18, 2025

Ⅰ. 서 론

개인 식별 정보(Personal Identifiable Information; PII)는 특정 개인을 직접 또는 간접적으로 식별할 수 있는 정보를 의미한다. 예시로는 이름, 생년월일, 여권 번호, 이메일 주소, 전화번호, IP 주소, 금융 정보 등이 포함된다[1]. PII가 유출될 경우 신원 도용, 금융 사기 등 각종 범죄에 악용될 가능성이 높으며, 이는 개인뿐만 아니라 사회 전체의 보안 위협이 될 수 있다. [2]는 이러한 문제를 해결하기 위해 구조화되지 않은 대규모 텍스트 데이터(예: 이메일)에서 PII를 자동으로 탐지하고 분류하는 기법을 제안하였다[2]. 해당 연구에서는 개인 이메일의 본문 길이가 짧을수록 PII 유출 빈도가 증가한다는 점을 밝혀내었으며, 이는 메신저와 같은 짧은 메시지 기반 플랫폼에서도 유사한 개인정보 유출 위험이 존재할 가능성을 시사한다.

이러한 문제를 해결하기 위해 일부 메신저 서비스는 'Self-Destructing' 기능을 도입하여 일정 시간이 지나면 메시지가 자동으로 삭제되도록 설계하였다. 해당 기능은 사용자 프라이버시 보호를 위한 효과적인 방법 중 하나지만, 많은 사용자가 이를 보안 기능이 아닌 단순 부가 기능으로 인식하는 경향이 있다[3]. 또한, 현재의 Self-Destructing 메신저는 단순히 일정 시간이 지나면 메시지를 삭제하는 방식을 따르고 있어 전송 중 평문 노출, 서버 내 데이터 잔존, 삭제된 메시지의 복구 가능성 등의 보안 취약점을 내포하고 있다[4]. 즉, 기존 Self-Destructing 메신저는 단순한 자동 삭제 기능만 제공할 뿐 근본적인 보안 문제를 해결하지 못하는 한계가 있다.

본 연구에서는 기존 Self-Destructing 메신저의 보안 취약점을 개선하고, 보다 강력한 개인정보 보호 기능을 갖춘 새로운 Self-Destructing 메신저 애플리케이션을 제안한다. 이를 위해 Support Vector Machine(SVM) 모델을 활용하여 메시지 내 개인정보를 실시간으로 탐지하고, 정보의 민감도에 따라 4단계로 분류하는 자동 보호 시스템을 도입하였다. 또한, 사용자가 메시지의 열람 가능 시간 및 횟수를 직접 설정하거나 자동으로 설정되도록 설계함으로써 보안성을 강화하였다. 기존의 Self-Destructing 메신저가 단순한 자동 삭제 기능만 제공하는 것과 달리, 본 연구에서는 개인 정보 보호 수준을 사용자가 직접 제어할 수 있도록 하여 유출 가능성을 최소화하는 점에서 차별성을 갖는다.

아울러, AES 암호화 및 Android KeyStore를 적용하여 메시지를 안전하게 저장하고, 삭제된 메시지가 포렌식 도구를 활용한 복구 시도에서도 재구성되지 않도록 보안을 강화하였다. 기존 Self-Destructing 메신저가 일정 시간이 지나면 자동으로 메시지를 삭제하는 데에만 초점을 맞춘 반면, 본 연구는 메시지의 저장부터 삭제 이후의 복구 가능성까지 고려한 보안 설계를 적용하여 보다 종합적인 개인정보 보호를 제공한다.

이를 통해, 본 연구는 단순한 자동 삭제 기능을 넘어 실시간 PII 탐지, PII 민감도 기반 차등 보안 적용, 암호화 및 포렌식 복구 방지 기법을 결합한 Self-Destructing 메신저의 새로운 보안 프레임워크를 제안한다.

Ⅱ. 공격자 모델

본 연구는 사용자의 메시지 데이터를 안전하게 보호하기 위해 다양한 유형의 공격자를 가정한다. 공격자 모델은 사용자의 실수나 일상적 행위로 인한 일반적인 위협부터, 전문적인 기술을 이용한 고도화된 공격까지 폭넓은 시나리오를 포함한다. 이에 따라, 본 논문에서는 다음과 같은 네 가지 대표적인 공격자 유형을 정의한다.

2.1 사용자의 실수 또는 수신자에 의한 유출

- 공격자 A (발신자 유출): 발신자가 실수로 개인정보가 포함된 메시지를 캡처하거나 외부로 전송함으로써, 의도치 않게 민감한 정보를 유출할 수 있다. 이는 사용자의 부주의로 인해 타인에게 개인정보가 전달되는 대표적인 사례에 해당한다.
- 공격자 B (수신자 유출): 메시지를 수신한 사용자가, 타인의 개인정보가 포함된 메시지를 캡처한 뒤 이를 외부로 공유하거나 저장할 수 있다. 이와 같은 방식은 실행이 간단하고 흔하게 발생할 수 있어 메시지 유출 위험이 매우 높다.

2.2 기기 접근을 통한 메시지 노출

- 공격자 C (물리적 접근자): 공격자가 사용자 스마트폰에 물리적으로 접근할 수 있는 상황에서, 기기의 잠금이 해제된 상태라면 메시지 앱을 열람하여 민감한 개인정보를 직접 확인할 수 있다. 이는 기기를 도난당하거나 방치한 경우 현실적으로 발생 가능한 위협이다.

2.3 삭제된 데이터의 복구 시도

- 공격자 D (포렌식 기반 고급 공격자): 공격자가 기기를 확보한 후, 포렌식 도구를 활용해 삭제된 메시지를 복원하려는 시도를 할 수 있다. 이는 전문적인 복구 소프트웨어나 하드웨어를 통해 데이터를 추출하는 방식으로, 고급 기술을 사용하는 공격자 시나리오에 해당한다.

Ⅲ. 설계 및 구현

3.1 SVM 기반 민감 정보 탐지 기능

본 연구에서는 Support Vector Machine(SVM) 모델을 활용하여 메시지 내 민감한 개인정보를 실시간으로 탐지하고, 탐지된 경우 Self-Destructing 기능을 자동으로 활성화하는 방식을 제안한다. SVM모델[5]은 데이터를 고차원 공간으로 매핑하여 최적의 분류 경계(초평면)를 설정함으로써 민감 정보와 일반 정보를 효과적으로 구분할 수 있다. 특히 고차원 특성을 가지는 데이터에서도 우수한 일반화 성능을 보여, 민감 정보 탐지에 적합하다.

선행 연구 [6]에서는 TF-IDF 벡터화를 기반으로 다양한 분류 모델을 적용해 PII 탐지 성능을 평가하였으며, 이 중 LSTM, SVM, RF(Random Forest) 모델이 상대적으로 높은 정확도를 보였다. 본 연구에서는 이 세 가지 모델을 기반으로 비교 실험을 수행하였다. 실험 결과, LSTM은 level 2 데이터에 대한 오분류율이 높아 전체 정확도는 78.25%에 그쳤으며, RF는 전체 정확도는 97%로 SVM과 유사했지만 level 2 및 level 3 데이터를 낮은 민감도 등급으로 잘못 분류하는 비율이 높았다. 반면, SVM은 전체 정확도 뿐 아니라 고민감도 데이터의 과소 분류를 가장 잘 억제하는 특성을 보여, 민감 정보 보호가 중요한 본 연구의 목표에 가장 적합하다고 판단되어 최종 분류 모델로 채택하였다.

학습 데이터는 SQLite 기반 데이터베이스에서 추출한 실제 메시지 데이터를 기반으로 구축되었으며, 각 메시지 항목의 열을 문자열로 변환 후 하나의 필드로 결합하여 텍스트 입력 데이터를 구성하였다. 이후 Scikit-learn의 TfidfVectorizer를 이용해 1,000차원의 희소 벡터로 변환하였으며, 해당 벡터화 과정에 L2 정규화가 포함되어 있어 별도의 정규화 처리는 생략하였다. 데이터의 불균형 문제는 SMOTE(Synthetic Minority Over-sampling Technique) 기법을 통해 보완하였다. 분류 모델로는 선형 커널(linear kernel)을 사용하는 SVM을 적용하였으며, 하이퍼파라미터로는 정규화 계수 C(0.01, 0.1)와 gamma(0.01)를 설정하여 GridSearchCV 기반의 3-fold 교차검증을 통해 최적 파라미터를 탐색하였다. 최종적으로 도출된 모델은 전체 학습 데이터로 재학습되었으며, 테스트셋을 통해 성능을 검증하였다.

마지막으로, 본 연구에서는 개인정보 보호를 효과적으로 구현하기 위해 민감도에 따른 등급별 분류 체계를 적용하였다. 기존 연구들 역시 정보의 중요도에 따라 등급을 구분하는 방식이 효과적임을 강조하고 있으며[7], 특히 금융 정보나 건강 정보와 같은 고위험 데이터에 대해서는 보다 강력한 보호 조치의 필요성이 제기된 바 있다[8]. 본 연구는 이러한 선행 연구를 바탕으로, 개인정보가 개인에게 미칠 수 있는 부정적 영향을 기준으로 4단계로 분류하였다(표 1).

표 1. 개인정보 민감도별 분류
Table 1. Personal Data Sensitivity Classification

0 level	Non-personal data	
1 level	- Non-threatening personal information - Messages can be sent without warning	Email address, Religion, Name, Place of birth, Date of birth, Weight
2 level	- Identifiable contact informatoin - Confirm with sender before sending	Address, Phone number, Service number, Employee number
3 level	- Risk of economic loss or criminal misuse if leaked - Confirm with sender before sending - Notify sender if recipient captures or risks leaking	SSN, Driver's license number, Vehicle registration number, Credit card number, IP address, Bank account number, Passport number, Password, Medical history

3.2 AES 암호화 기반 메시지 전송

본 연구에서 제안하는 메신저 애플리케이션은 AES(Advanced Encryption Standard) 알고리즘[9]을 사용하여 메시지 데이터를 안전하게 암호화한다. AES는 비대칭키 암호화보다 연산이 간결하여 실시간 통신 환경에서도 성능 저하 없이 빠르고 안정적인 암·복호화가 가능하다[10]. 이러한 특성으로 인해 카카오톡, WhatsApp, Facebook Messenger 등 주요 메신저 애플리케이션에서도 AES 암호화가 널리 활용되고 있다[11,12]. 본 연구에서는 이러한 AES의 빠른 연산 속도와 효율성을 활용하여, 메시지 전송 과정에서 암호화를 적용함으로써 보안성을 강화하고 데이터 유출 및 위·변조 위험을 최소화한다.

3.3 화면 캡처 방지 및 탐지

본 연구에서는 메시지가 화면 캡처를 통해 외부로 유출되는 것을 방지하기 위해 캡처 제한 기능을 설계하였다. 이를 구현하기 위해 안드로이드 애플리케이션의 Window 객체에 FLAG_SECURE 플래그를 설정하여 화면 캡처 및 녹화 기능을 비활성화하거나 이미지 저장 경로를 추적하여 screenshot으로 저장될 경우 알림을 보내도록 하였다. [그림 1]은 캡처 방지 구조를 나타내며, 수신자가 Self-Destructing 메신저를 캡처할 경우

메시지의 PII를 분석하여 [표 1]에 따라 분류된 민감도 레벨을 확인한다. 확인 결과 0 level~ 2 level 일 경우 메시지를 보낸 발신자에게 수신자가 메시지를 캡처했다는 알림을 보낸다. 3 level의 경우에는 민감한 정보를 유출할 가능성이 높으므로 캡처를 완전히 차단하도록 하였다.

또한, 발신자가 무심코 스크린샷을 촬영하는 경우에도 캡처된 이미지를 탐지하여 경고 메시지를 제공함으로써, 사용자의 실수로 인한 정보 유출 위험을 줄이는 데 기여한다. [그림 2]은 스크린샷 캡처 시 경고 팝업이 표시되는 과정을 나타낸다. 이러한 설계는 발·수신자 양측에서의 캡처 시도를 효과적으로 방지하여 민감 정보의 보안을 더욱 강화하는 역할을 하며 이러한 설계는 공격자 A 와 공격자 B에 대한 대응이 가능하다.

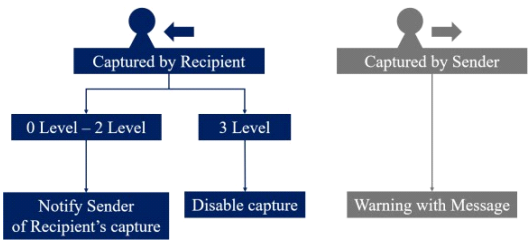

그림 1. 캡처 방지 파이프라인
Fig. 1. Anti-Capture Pipeline

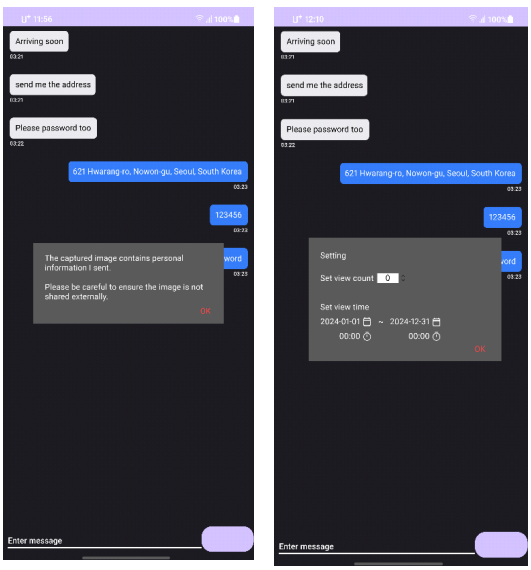

그림 2. 캡처 알림 화면(좌), 메시지 열람 조건 설정 화면(우)
Fig. 2. Capture Notification (Left), Set Message Viewing Conditions (Right)

3.4 메시지 열람 제한 및 자동 삭제 설정

본 연구에서는 사용자가 보낸 메시지를 길게 누르면 [그림 2]와 같이 열람 조건을 추가할 수 있는 사용자 인터페이스(UI)가 제공된다. 발신자는 메시지 전송 시 열람 가능 시간과 횟수를 지정할 수 있으며, 이를 통해 수신자는 설정된 시간 동안 또는 횟수만큼만 메시지를 열람할 수 있도록 제한된다. 이러한 조건은 수신자가 발신자의 의도를 벗어나 무단으로 메시지를 열람하거나 민감한 정보를 유출할 가능성을 최소화하며, 설정된 조건이 충족되면 메시지는 자동으로 삭제되도록 설계되었다. 이 과정은 발신자가 직접 수동으로 설정할 수도 있으며, 민감 정보의 등급에 따라 자동으로 적용되도록 구현되었다. 특히, 본 기능은 기기에 직접 접근할 수 있는 공격자 C가 메시지를 유출하려는 시도를 방지하는 역할을 한다. 메시지에 열람 조건이 적용되면 공격자는 설정된 시간과 횟수를 초과하여 메시지를 확인할 수 없으므로, 민감 정보의 유출 가능성이 크게 낮아진다. 이는 사용자 편의성을 유지하면서도 개인정보 유출을 예방하는 실질적인 보호 수단으로 작용한다.

이러한 애플리케이션 전체 구조는 [그림 3]에서 확인할 수 있다. 본 연구에서 제안하는 애플리케이션은 메시지를 보내면 보낸 메시지에서 PII의 존재 유무를 확인하여 [표 1]에 따라 민감도를 구분한다. 사용자(발신자)가 수동으로 Self- Destructing 기능의 옵션을 설정한 경우에는 해당 조건에 맞추어 메시지가 전송되지만 옵션을 설정하지 않았을 경우 민감한 데이터가 보호되지 않고 전송될 가능성이 있어 PII가 포함된 데이터의 경우 [표 1]의 조치에 맞추어 시스템 상에서 자동적으로 옵션을 설정하도록 한다.

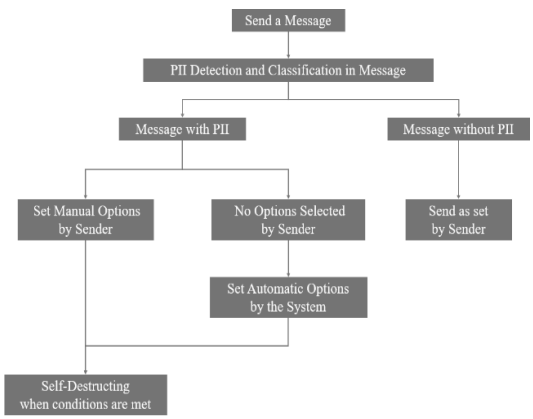

그림 3. 애플리케이션 파이프라인
Fig. 3. Application Pipeline

Ⅳ. 실험 및 평가

4.1 실험 환경

본 실험은 삼성 갤럭시 S20(모델명 SM-G980, 저장 용량 256GB) 기기에서 수행되었으며, Android 12 운영체제에서 개발 및 테스트를 진행하였다. 메신저 애플리케이션은 Android Studio Hedgehog(2023.1.1) 버전을 이용하여 구현하였다. 기기의 주요 하드웨어 사양은 다음과 같다. 프로세서는 Samsung Exynos 990(S5E9830) 칩셋으로, 2개의 Exynos M5 (2.73GHz), 2개의 ARM Cortex-A76(2.50GHz), 4개의 ARM Cortex-A55(2.00GHz)로 구성된 8코어 CPU를 탑재하고 있으며, GPU는 ARM Mali-G77 MP11(850MHz), 메모리는 8GB LPDDR5 SDRAM(5,500 MT/s)이다. 또한, 삼성 1세대 듀얼 NPU(최대 933MHz DSP 포함)를 탑재하고 있다.

성능 측정의 정확도와 재현 가능성을 확보하기 위해, 테스트 전 디바이스에 내장된 시스템 최적화 기능(Device Care)을 사용하여 저장 공간 정리, 불필요한 파일 삭제, 비정상적인 배터리 소모 앱 종료, 백그라운드 프로세스 정리 등을 수행하였다. Android 시스템 설정에서 배터리 최적화 기능은 해제하였으며, 실험은 충전기 미연결 상태의 배터리 전원 환경에서 수행되었다. 테스트의 일관성을 확보하기 위해 매 실험 전 기기 초기화를 통해 동일한 조건을 구성하였고, 성능 측정값(예: 메시지 분석 지연 시간)은 동일 환경에서 여러 차례 반복 측정하여 평균값을 산출하였다.

4.2 개인 정보 탐지 결과

본 실험은 [표 1]에서 정의한 20개 개인정보 항목에 대해 각 500개씩, 총 10,000개의 메시지 데이터를 수집하여 수행되었다. 모든 입력 데이터는 TF-IDF 기반으로[6] 벡터화된 후 SVM 모델에 적용되었다. [그림 4]는 실제 등급과 예측 등급 간의 관계를 나타낸 혼동 행렬을 보여준다. 이를 기반으로 정확도, 정밀도, 재현율, F1-score 등의 성능 지표를 산출할 수 있다[13].

SVM 모델의 전체 정확도는 96.25%로 나타났으며, 이는 개인정보 탐지 및 분류에 있어 높은 신뢰성을 확보하였음을 의미한다. 특히, 위험 등급이 높은 2단계와 3단계 항목에 대한 오분류율이 매우 낮게 유지되었다. 2단계 데이터의 FNR(False Negative Rate)과 FPR(False Positive Rate)은 각각 0.01이며, 3단계는 FNR이 0.00, FPR이 0.03으로 측정되었다. 이는 고위험 개인정보가 잘못 탐지되거나 낮은 등급으로 분류되는 비율이 거의 없음을 보여주며, 보안적으로 민감한 정보를

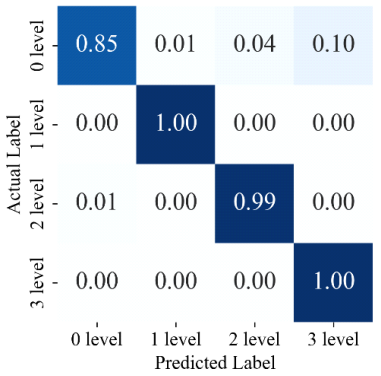

그림 4. 혼동 행렬 기반 개인정보 탐지 결과
Fig. 4. Confusion matrix-based PII Detection Results

정확히 판별할 수 있는 탐지 모델의 특성을 반영한다.

반면, 비민감 정보(0단계)의 경우 FNR이 0.14로 비교적 높은 편이었으며, 이는 개인정보가 포함되지 않은 메시지가 민감 정보로 잘못 탐지되는 경향이 있음을 시사한다. 그러나 해당 클래스의 FPR은 0.00으로, 민감 데이터를 비민감으로 간주하는 위험은 거의 발생하지 않았다. 이러한 결과는 본 시스템이 오탐보다는 과탐을 허용하는 보수적인 설계 방식을 따르고 있음을 의미하며, 보안 중심의 응용 환경에서는 긍정적으로 해석될 수 있다.

결과적으로, 제안된 SVM 기반 탐지 시스템은 전체적으로 높은 정확도와 함께, 민감 정보를 비민감 정보로 잘못 분류하는 오류를 최소화함으로써 설계 목표였던 정보 과소 탐지 방지를 효과적으로 달성하였음을 보여준다.

4.3 개인 정보 탐지 시간

개인정보 탐지 모델의 실시간 응답 성능을 평가하기 위해, SVM 모델이 하나의 입력 데이터를 특정 민감도 등급으로 분류하는 데 소요되는 시간을 측정하였다. 100개의 테스트 데이터를 대상으로 평균 탐지 시간을 산출한 결과, 개인정보 등급을 예측하는 데 평균 0.2325초가 소요되었다([그림 5]). 이 결과는 사용자가 메시지를 입력한 직후 민감 정보 포함 여부를 판단할 수 있을 만큼 실시간에 가까운 응답 성능을 제공함을 의미하며, 메신저 환경에서도 개인정보 보호 기능을 사용자 경험을 해치지 않고 적용할 수 있음을 보여준다.

추가적으로, 대화량 증가에 따른 처리 성능 저하가능성을 분석하기 위해 입력 메시지 수를 100개, 500개, 2,000개, 10,000개로 단계적으로 확장하여 예측 시간을 측정하였다. 그 결과, 처리 시간은 각각 0.233초, 0.997초, 4.136초, 17.810초로 측정되었으며, 입력 수 대비

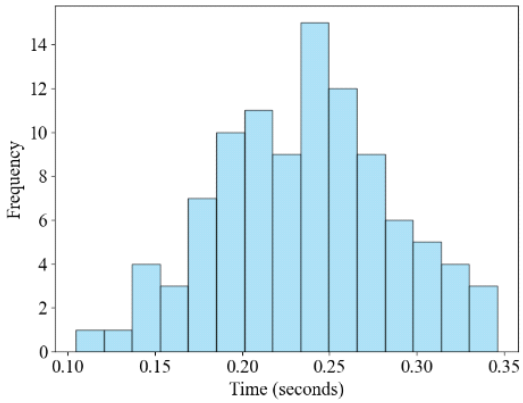

그림 5. 개인정보 탐지 시간
Fig. 5. PII Detection time

처리 시간은 선형보다 느린 증가율을 보였다. 메시지 수를 5배 증가시킨 500개 입력 시 처리 시간은 약 4.29배, 20배 증가한 2,000개 입력 시 약 17.79배, 100배 증가한 10,000개 입력 시 약 76.6배 증가하였다.

이는 SVM 기반 탐지 모델이 다량의 메시지를 연속적으로 처리하는 상황에서도 sub-linear한 처리 특성을 보이며, 시스템 확장성 측면에서 안정적인 성능을 유지할 수 있음을 시사한다. 이러한 결과는 본 메신저 시스템이 대규모 대화 흐름 속에서도 실시간에 가까운 개인정보 탐지가 가능하다는 점에서, 실제 환경 적용에 적합한 설계임을 보여준다.

4.4 삭제된 메시지의 복구 가능성 평가 및 보안 요구사항 도출

삭제된 메시지의 복구 가능성을 평가하기 위해 포렌식 도구를 활용하여 기기 내 저장 데이터를 분석하였다. 이를 위해, Odin[14]을 사용하여 테스트 기기의 루트 권한을 획득한 후, FTK Imager[15]를 이용하여 데이터베이스(.db 파일)를 추출하였다. 분석 결과, 삭제되지 않은 메시지는 평문으로 저장 되어있어 직접적인 확인이 가능하였으나([그림 6]), Self-Destructing 기능이 적용된 메시지는 암호화된 상태로 저장된 후 삭제되었으며,

sender_id	receiver_id	username	message_content
Filter	Filter	Filter	Filter
1	2	Bob	Test
1	2	Bob	Hello?
2	1	Alice	Hello
2	1	Alice	Social security number Test
1	2	Bob	000101-1111111
1	2	Bob	testtest

그림 6. 삭제되지 않은 메시지 확인
Fig. 6. Undeleted Message

sender_id	receiver_id	username	message_content
Filter	Filter	Filter	Filter
1	2	Bob	Test
1	2	Bob	Hello?
2	1	Alice	Hello
2	1	Alice	Social security number Test
1	2	Bob	BLOB
1	2	Bob	testtest

그림 7. 삭제된 메시지 확인
Fig. 7. Deleted Message

데이터베이스에서 복구할 수 없음을 확인하였다([그림 7]).

본 연구에서는 암호화 키의 보안을 강화하기 위해 Android KeyStore를 활용하였다. Android KeyStore 시스템은 암호화 키를 기기 내 보안 컨테이너에 저장하여 외부에서 직접 추출하는 것이 어렵도록 설계되어 있다. 또한, 하드웨어 기반 보안 모듈을 적용하여 키 관리의 안전성을 더욱 강화하였다[16]. 이러한 설계는 루트 권한을 이용한 복구 시도에도 불구하고 개인정보 탈취가 어렵도록 보호할 수 있으며, 공격자 D에 대한 효과적인 대응이 가능함을 보여준다.

이러한 실험 결과를 바탕으로, 삭제된 메시지의 복구 가능성을 최소화하고 보안성을 강화하기 위해 몇 가지 기능적 요구사항을 도출하였다. 먼저, 메시지 저장 방식의 보안성을 강화해야 하며, 삭제된 메시지는 완전히 제거될 수 있도록 설계되어야 한다. 이를 위해 단순한 논리 삭제(logical deletion)가 아니라, 복구가 불가능한 물리적 삭제(secure deletion) 방식을 적용해야 한다. 또한, 암호화된 저장 방식을 적용해야 하며, 삭제 전 메시지가 저장될 때 평문 저장 방식이 아닌 AES-256과 같은 강력한 암호화 기법을 활용해야 한다. 마지막으로, 파일 시스템 및 캐시 데이터 처리를 철저히 수행해야 한다. 포렌식 분석을 통한 삭제된 메시지의 복구를 방지하기 위해, OS 레벨에서 생성되는 캐시 파일 및 백업 데이터까지 삭제하는 보안 정책이 필요하다.

이러한 기능적 요구사항을 반영함으로써, 삭제된 메시지의 복구 가능성을 최소화하고 보안성을 강화할 수 있다.

V. 관련 연구

5.1 PII detection method

[1]은 대규모 언어 모델(LLM)에서 PII 및 민감정보 보호를 위한 적응형 프레임워크를 연구하였다. 기존 규칙 기반 PII 탐지 시스템의 한계를 극복하기 위해 문맥

기반 PII 식별 기법 및 동적 규제 준수 시스템을 개발하였으며, GDPR, CCPA 등의 규정에 맞춘 맞춤형 마스킹 및 익명화 기법을 적용하였다[1]. [2]는 비정형 데이터에서 PII가 다양한 형태로 표현될 수 있다는 점에 주목하여, 기존의 구조화된 데이터 중심 연구의 한계를 극복하기 위한 탐지 기법을 제안하였다. 이를 위해 공백 및 특수 문자 처리를 포함한 데이터 전처리 기법을 적용하고, 이메일 헤더와 본문을 분리하여 분석하는 방식을 도입하였다. 또한, 기존 모델과의 분류 성능을 비교하였다[2]. [6]은 10,000개 이상의 텍스트 데이터셋을 활용하여 문자열 토큰화, 형태소 분석, 정규 표현식을 기반으로 PII 라벨링을 수행하고, TF-IDF 벡터화를 이용하여 주요 특징을 추출하는 방법을 제안하였다[6]. 연구 결과, LSTM(Long Short-Term Memory), SVM(Support Vector Machine), RF(Random Forest) 모델이 높은 탐지 성능을 보여주었다. [17]은 2013년부터 2018년까지 발생한 대규모 데이터 유출 사고 4건을 분석하며, PII 자동 탐지 및 보호 시스템의 필요성을 제시하였다. 이를 위해 비정형 데이터에서 PII를 탐지하는 딥러닝 및 머신러닝 모델을 제안하였으며, 실험 결과 SVM 모델이 가장 우수한 성능을 보여주었다.

PII Detection에 대한 연구는 대부분 PII가 어떤 특정 정보를 담는지 혹은 PII 포함 여부를 가르는 반면 본 연구는 PII탐지를 레벨별로 수행하여 높은 정확도로 구별할 수 있도록 하였다. 또한, 탐지 및 분류에 소요되는 시간을 평가하여 Self-Destructing 메신저 환경에서의 적용 가능성을 검증하였다는 점에서 기존 연구와의 차별성을 가진다.

5.2 Capture prevention method

[18]은 사용자가 민감한 메시지를 보낼 때 수신자의 화면 캡처를 방지하는 기능이 필요함을 강조하였다. 해당 연구에서는 Self-Destructing 기능이 존재하더라도 스크린샷을 통해 데이터가 유출될 가능성이 있으며, 이를 예방하기 위해 캡처 방지 기능을 구현하였다. 또한, 운영체제 수준에서 개인 및 기관의 중요한 정보 유출을 방지하기 위한 화면 캡처 방지 모듈을 제안한 연구도 진행되었다.

[19]는 API후킹 기술을 활용하여 화면 캡처 방지 모듈을 개발하고, 이를 통해 보안을 강화하는 방법을 제안하였다. 해당 연구는 개별 애플리케이션에서 동작하는 스크린샷 방지 기능이 아니라, Windows 운영체제 전체에 적용되는 캡처 방지 기능을 구현하여 운영체제 차원의 보안 방법을 제안하였다. 반면, 본 연구에서는 운영체제가 아닌 애플리케이션 수준에서 보안을 적용하였으며, 캡처 방지 시나리오를 발신자와 수신자의 개인정보 유출 가능성에 따라 구분하여 보다 적절한 보안 조치를 적용할 수 있도록 설계하였다.

5.3 Self-Destructing method

기존 Self-Destructing 관련 연구들은 데이터를 안전하게 삭제하는 암호학적 접근에 중점을 두어 왔다. 예를 들어, [20]은 AES 대칭 암호화와 OpenPGP 기반 RSA 비대칭 암호화를 결합하여 보안을 강화하였으며, Ephemeral Public Key Manager를 통해 오프라인 환경에서도 암호화 키를 관리할 수 있도록 설계하였다. 또한, 암호화 및 키 생성, 해시 계산 등의 성능을 시간 단위로 평가하여 모바일 환경에서의 실용성을 입증하였다. 네트워크 관점에서는 분산 P2P 네트워크를 활용해 일정 시간이 지난 후 데이터 접근을 차단하는 시스템이 제안되었으며[21], Self-Destructing 기술은 모바일 환경뿐 아니라 이메일 등 다양한 응용 분야로도 확장되어 왔다[22]. [23]은 다중 키 관리 기법을 기반으로 메시지를 일회용 키로 암호화하고, 설정된 조건(예: 시간, 빈도, 위치 등)이 충족되면 해당 키를 삭제하여 메시지 열람을 차단하는 Self-Destructing 메시징 시스템을 제안하였다. 이 연구에서는 스테가노그래피와 위치 기반 제어 기능을 통합하여 보안을 강화하였고, 암호화 기법으로는 Shamir의 비밀 공유 알고리즘을 적용하였다.

최근에는 상용 메신저 앱의 Self-Destructing 기능에 대한 포렌식 분석 연구도 이루어지고 있다. [24]는 WhatsApp, Snapchat, Telegram의 자폭 메시지 기능이 실제로 완전한 삭제를 보장하는지를 디지털 포렌식 도구(Cellebrite, XRY, AXIOM)를 활용하여 분석하였다. 해당 연구는 Android와 iOS 환경 모두에서 삭제된 메시지가 여전히 로컬 데이터베이스나 캐시 영역에 남아 있을 수 있으며, 복구 가능성은 기기 상태나 도구에 따라 달라진다는 점을 보여주었다. 예를 들어, WhatsApp의 경우 메시지가 만료된 이후에도 일부 복구 도구에서는 데이터를 복원할 수 있었으며, Snapchat은 메시지를 서버에 저장하지 않아 분석 가능성이 제한되었고, Telegram은 iOS에서는 복구가 불가능한 반면 Android에서는 일부 데이터의 복구가 가능하였다. 다만, 해당 연구에서는 각 앱의 암호화 방식이나 키 저장 구조에 대한 기술은 제한적이었으며, Signal이나 Wickr와 같은 강력한 보안 기반 메시징 앱에 대해서는 후속 연구가 필요하다고 언급되었다.

[25]는 Signal, Wickr, Threema와 같은 고보안 인스턴트 메신저를 대상으로 암호화된 데이터의 복호화 가능성을 분석하고, 법적 증거 확보를 위한 방법론을 제안

표 2. Self-Destructing 기존 연구와 본 연구의 비교
Table 2. Comparison of Previous Self-Destructing Studies

	Encryption Method	Capture Prevention	PII Detection
[23]	Shamir's secret sharing	X	X
[20]	AES	X	X
[21]	AES	X	X
[18]	E2EE	O	X
Our method	AES	O	O

하였다. 특히 Signal과 Wickr는 Self-Destructing 메시지를 기본 기능으로 제공하며, 프라이버시 보호를 위한 핵심 수단으로 분석되었다. Signal은 Android Keystore를 사용하여 메시지 DB 키를 보호하며, 루팅된 기기에서도 이 키를 직접 획득할 수 없어, 연구팀은 Signal 앱의 키 관리 구조를 정적·동적 분석한 후 이를 모방한 테스트 애플리케이션을 제작하여 복호화를 시도하였다. 그 결과, Android Keystore 기반 암호화는 단순 루팅만으로는 우회가 불가능하다는 점을 입증하였다. Wickr는 AES-256-GCM 기반 이중 암호화를 사용하며, 사용자 비밀번호 없이도 내부 자동 로그인 키를 통해 복호화 경로를 확보하는 방식으로 보안을 유지하고 있음을 보여주었다.

한편, WhatsApp이나 Signal과 같은 상용 메신저 앱 중에서는 PII를 탐지하여 보호 기능과 연계하는 시스템은 존재하지 않으며, 기존 Self-Destructing 연구들 역시 안전한 삭제 기능에만 초점을 맞춘 경우가 대부분이다. 본 연구는 이러한 기존 한계를 보완하여, 머신러닝 기반 PII 탐지, 민감도 기반 분류, 조건 기반 자동 삭제, 캡처 방지 기능을 통합함으로써 실질적인 개인정보 보호 기능을 갖춘 Self-Destructing 메시지 시스템을 제안한다. 기존 연구와 본 연구의 차별점은 [표 2]에 정리되어 있다.

Ⅵ. 결 론

본 연구에서는 빈번한 개인 식별 정보(PII) 유출 문제를 해결하기 위해 보안성이 강화된 Self- Destructing 메신저 앱을 제안하였다. 기존 Self- Destructing 기능이 단순한 메시지 일시적 공유에 초점을 맞춘 것과 달리, 본 연구는 PII 탐지 기술과 조건 기반 자동 삭제 기능을 결합하여 민감 정보의 실시간 보호와 삭제 후 복구 방지를 목표로 하였다.

실제 기기에서 구현된 앱을 이용한 실험 결과, 개인정보 등급별 탐지 정확도는 96.25%로 높은 성능을 기록하였으며, 포렌식 도구를 활용한 복구 시도에서도 삭제된 메시지가 복구되지 않음을 확인하였다. 이를 통해, 본 연구는 기존 Self- Destructing 시스템의 한계를 보완하고, 모바일 환경에서 실질적인 개인정보 보호를 구현하는 새로운 방향을 제시하였다.

References

[1] S. Asthana, et al., *Adaptive PII mitigation framework for large language models*(2025), Retrieved Mar. 20, 2025, from https://arxiv.org/abs/2501.12465. (https://doi.org/10.48550/arXiv.2501.12465)

[2] P. Kulkarni and N. K. Cauvery, "Personally identifiable information (PII) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique," *IJACSA*, vol. 12, no. 9, pp. 605-612, Sep. 2021. (https://doi.org/10.14569/IJACSA.2021.0120957)

[3] F. Roesner, B. T. Gill, and T. Kohno, "Sex, lies, or kittens? Investigating the use of Snapchat's self-destructing messages," in *Proc. Int. Conf. Financial Cryptography and Data Security*, pp. 64-76, Springer, 2014. (https://doi.org/10.1007/978-3-662-45472-5_5)

[4] L. Onwuzurike and E. De Cristofaro, "Experimental analysis of popular smartphone apps offering anonymity, ephemerality, and end-to-end encryption," *arXiv preprint arXiv: 1510.04083*, 2015. (https://doi.org/10.48550/arXiv.1510.04083)

[5] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Wkshp. COLT '92*, pp. 144-152, Pittsburgh, PA, USA, Jul. 1992. (https://doi.org/10.1145/130385.130401)

[6] M. Mitra and S. Roy, "Identification and processing of PII data, applying deep learning models with improved accuracy and efficiency," *J. Data Acquisit. and Process*, vol. 33, no. 6, p. 1337, Dec. 2018.

[7] R. Turn, "Classification of personal

information for privacy protection purposes," in *Proc. National Computer Conf. and Exposition*, pp. 301-307, New York, USA, Jun. 1976. (https://doi.org/10.1145/1499799.1499846)

[8] R. Belen-Saglam, J. R. C. Nurse, and D. Hodges, "An investigation into the sensitivity of personal information and implications for disclosure: A UK perspective," *Frontiers in Comput. Sci.*, vol. 4, art. 908245, 2022. (https://doi.org/10.3389/fcomp.2022.908245)

[9] V. Rijmen and J. Daemen, "Advanced encryption standard," in *Proc. Federal Inf. Process. Standards Publications, National Inst. Standards and Technol.*, vol. 19, no. 22, pp. 1-8, 2001. (https://doi.org/10.6028/NIST.FIPS.197-upd1)

[10] S. Lenz, Evaluation of the messaging layer security protocol: A performance and usability study, *Linköping University, Master's thesis*, 2020.

[11] V. Bhuse, "Review of end-to-end encryption for social media," in *Proc. ICCWS 2023*, vol. 18, no. 1, pp. 35-37, Feb. 2023. (https://doi.org/10.34190/iccws.18.1.1017)

[12] M. Jo and N. S. Chang, "Study on the data decryption and artifacts analysis of KakaoTalk in windows environment," *J. KIISC*, vol. 33, no. 1, pp. 51-61, 2023. (https://doi.org/10.13089/JKIISC.2023.33.1.51)

[13] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection," in *MAICS*, vol. 710, pp. 120-127, 2011.

[14] S.-T. Sun, A. Cuadros, and K. Beznosov, "Android rooting: Methods, detection, and evasion," in *Proc. 5th ACM CCS Wkshp. Secur. and Privacy in Smartphones and Mobile Devices*, pp. 3-14, Denver, USA, Oct. 2015. (https://dl.acm.org/doi/10.1145/2808117.2808126)

[15] J. Stüttgen and M. Cohen, "Anti-forensic resilient memory acquisition," *Digital Investigation*, vol. 10, pp. S105-S115, 2013. (https://doi.org/10.1016/j.diin.2013.06.012)

[16] M. Sabt and J. Traoré, "Breaking into the KeyStore: A practical forgery attack against android keyStore," in *Comput. Security - ESORICS 2016*, pp. 531-548, Heraklion, Greece, Sep. 2016. (https://doi.org/10.1007/978-3-319-45741-3_27)

[17] A. K. Makhija, "Deep learning application – Identifying PII (Personally Identifiable Information) to protect," *J. Accounting, Finance, Economics, and Social Sci.*, vol. 5, no. 2, pp. 10-16, Dec. 2020. (https://doi.org/10.62458/jafess.160224.5(2)10-16)

[18] M. E. Y. M. Al Suwaidi, S. B. Sidek, and S. A. Al-Shami, "A conceptual framework of fintech laws and regulations on the risk management of financial institutions in UAE," *Math. Statistician and Eng. Appl.*, vol. 71, no. 3, pp. 1-7, May 2022.

[19] J. H. Lee, "Implementation of anti-screen capture modules for privacy protection," *J. KIICE*, vol. 18, no. 1, pp. 91-96, Jan. 2014. (https://doi.org/10.6109/jkiice.2014.18.1.91)

[20] T.-Y. Tung, L. Lin, and D. T. Lee, "Pandora messaging: An enhanced self-message-destructing secure instant messaging architecture for mobile devices," in *Proc. 26th Int. Conf. Advanced Inf. Netw. and Appl. Wkshps. (WAINA 2012)*, pp. 720-725, Fukuoka, Japan, Mar. 2012. (https://doi.org/10.1109/WAINA.2012.112)

[21] R. Geambasu, T. Kohno, A. A. Levy, and H. M. Levy, "Vanish: Increasing data privacy with self-destructing data," in *Proc. 18th USENIX Security Symp.*, pp. 299-316, Montreal, Canada, Aug. 2009. (https://www.usenix.org/legacy/events/sec09/tech/full_papers/geambasu.pdf)

[22] J. Clark, P. C. van Oorschot, S. Ruoti, K. Seamons, and D. Zappala, "SoK: Securing email—a stakeholder-based analysis," in *Proc. 25th Int. Conf. Financial Cryptography and Data Security* (FC 2021), pp. 360-390, Berlin, Heidelberg: Springer Berlin Heidelberg, Mar.

2021.
(https://doi.org/10.48550/arXiv.1804.07706)

[23] A. Holkar, P. Powar, P. Mhaske, and S. Tak, "A self-destructing secure messaging system using multi key management scheme," *Int. J. Innovations Eng. Res. Technol.*, vol. 2, no. 2, pp. 1-8, Feb. 2015.
(https://repo.ijiert.org/index.php/ijiert/article/view/293)

[24] H. Heath, Á. MacDermott, and A. Akinbi, "Forensic analysis of ephemeral messaging applications: Disappearing messages or evidential data?," *Forensic Sci. Int.: Digital Invest.*, vol. 46, p. 301585, May 2023.
(https://doi.org/10.1016/j.fsidi.2023.301585)

[25] J. Son, Y. W. Kim, D. B. Oh, and K. Kim, "Forensic analysis of instant messengers: Decrypt signal, wickr, and threema," *Forensic Sci. Int.: Digital Invest.*, vol. 40, p. 301347, Mar. 2022.
(https://doi.org/10.1016/j.fsidi.2021.301347)

소 예 나 (Ye-na So)

2023년 3월~현재 : 서울여자대학교 정보보호학과 학사과정
<관심분야> AI for security, Privacy Enhancing Technologies(PET)
[ORCID:0009-0004-1825-4620]

이 선 우 (Sun-woo Lee)

2015년 2월 : 서강대학교 수학과 학사
2022년 2월 : 고려대학교 정보보호대학원 박사
2022년 3월~9월 : 고려대학교 정보보호대학원 연구교수
2022년 10월~2024년 2월 : Samsung Research Staff Engineer
2024년 3월~현재 : 서울여자대학교 정보보호학부 조교수
<관심분야> Usable security, IoT security, Authentication Side-channel attack, Privacy leakage attack
[ORCID:0000-0001-5216-0266]