『2025 제3회 KISIA 정보보호 개발자 해커톤』

개발기획서

팀명	유청 분리 두 번 했어요
프로젝트명	PII-tering
프로젝트 소개	
팀 소개 및 팀원별 역할	서울여자대학교 정보보호학과 학생으로 구성된 팀이고, 기업 유출 사고가 잦아짐에 따라 개인정보보호에 관심을 갖게 되었고 이와 관련된 보안 솔루션 개발에 어쩌구저쩌구.. 정지윤: PM, 프롬프트 엔지니어링, UI/UX 육은서: Ollama 연동 및 RAG, 데이터셋 구축 이시온: API 구성 및 백엔드, docs 작업

<작성 안내사항>
- 모든 기재 내용에는 허위 사실이 없어야하며, 필요시 증빙 내용을 제출해야 함
- 글자는 12포인트로 작성하며 분량의 제한은 없음
- 서류심사시 개발기획서를 기반으로 평가
- 팀명은 참가신청서에 작성한 팀명과 동일해야 함

① 추진 배경 및 필요성

기획 의도의 목적이 명확한가, 기존 서비스와 비교하여 독창적인 특징이 있는가? 등 작성

현대 기업 환경에서는 문서, 이메일, 이미지 등 다양한 포맷의 데이터가 실시간으로 공유되고 있으며, 이 과정에서 의도치 않은 개인정보 유출 및 내부 보안 정책 위반 사고가 빈번히 발생하고 있다. 특히 비정형 문서나 다국어 혼용 자료의 경우 기존 보안 시스템으로는 효과적인 감지가 어려워, 조직의 정보 유출 취약점으로 작용하고 있다.

기존 DLP(Data Loss Prevention) 솔루션은 주로 정규표현식 기반의 정적 탐지 방식을 사용하며, 문맥을 이해하지 못한 채 단순 패턴 일치 여부만으로 판단하기 때문에 오탐과 누락이 자주 발생한다. 또한 한국어에 대한 최적화가 부족하고, 스캔된 이미지 문서나 PDF 등 비정형 데이터에 대한 처리가 제한적이다. 내부 보안 정책 판단 역시 단순한 룰 매칭에 의존하기 때문에, 복잡한 조건이나 유연한 해석이 필요한 상황에서는 효과적인 대응이 어렵다.

본 프로젝트는 이러한 한계를 극복하기 위해, Microsoft Presidio 기반의 고도화된 PII 탐지 기능에 한국어 특화 NER 모델 및 정규표현식을 결합하여 높은 탐지 정밀도를 확보하고자 한다. 여기에 LLM 기반의 정책 판단 시스템을 적용함으로써, 문맥과 보안 정책을 실제로 이해하고 해석하는 능동적인 보안 결정을 가능하게 한다. 특히 조직 내에 정의된 보안 정책 문서를 벡터화한 뒤 RAG 구조를 통해 LLM에 연동함으로써, 조직의 개별 기준을 반영한 정책 위반 여부 판단이 가능하다. 또한 이미지 문서에 대한 OCR 처리, FastAPI 기반의 구조 설계 등을 통해 실무 환경에 쉽게 통합될 수 있도록 구성하였다.

이러한 지능형 보안 분석 시스템은 민감정보 보호와 내부 정보 유출 방지 측면에서 기존 솔루션보다 높은 정확성, 유연성, 실용성을 갖추고 있으며, 실질적인 보안 수준 향상을 이끌 수 있는 효과적인 대안이 될 수 있다.

② 주요 기능 및 개발환경

서비스 기능 설명

문서, 이메일, 이미지 등 다양한 형식의 내부 문서를 업로드하면, 해당 데이터 내에 포함된 개인정보 및 내부 정책 위반 여부를 자동으로 분석하고 차단 여부를 판단한다. Presidio 기반 PII 탐지, LLM 기반 정책 위반 판별, 벡터 검색 기반 RAG 파이프라인 등을 활용하여 정확도 높은 보안 위반 판단과 실시간 대응 기능을 제공한다.

문서/이미지 업로드 및 사전 처리

- 사용자는 PDF, Word, 이미지(JPG, PNG 등) 형태의 문서를 업로드할 수 있다.
- 업로드된 문서는 형식 검증 및 전처리를 거쳐 시스템 내에서 처리된다.
- 이미지 문서는 OCR 엔진을 통해 텍스트로 변환되며, 한글 및 영문 혼용 문서에 대해서도 정확도 높은 텍스트 추출이 가능하다.

민감정보(PII) 탐지 기능

- Microsoft의 Presidio 프레임워크를 기반으로, 주민등록번호, 계좌번호, 전화번호, 이메일 등 주요 개인정보를 탐지한다.
- 한글 이름, 한국 전화번호, 주민등록번호 등은 별도의 정규표현식 패턴으로 고도화하여 탐지 정확도를 높였다.
- 탐지된 민감정보는 문서 내 위치, 타입 등과 함께 JSON 형태로 구조화된다.
- 마스킹 처리를 위한 정보도 함께 제공된다.

정책 기반 판단 시스템 (LLM + RAG)

- 정책 위반 여부를 판단하기 위해, 보안 가이드라인 문서를 사전 벡터화하여 벡터 DB에 저장한다.
- 문서 내 탐지된 민감정보와 메타데이터(발신자 직책, 수신자 도메인 등)를 기반으로 LLM에게 정책 판별을 요청한다.
- RAG 구조를 통해, 정확하고 일관된 판단을 유도한다.
- 민감정보 포함 여부, 발신자 권한, 수신자 도메인 등을 종합하여 해당 내

용의 전송 허용/차단 여부를 판단한다.

API 기반 통합 운영

- 서비스는 FastAPI 기반으로 구축되어 있으며, 외부 시스템에서 API를 통해 연동 가능하다.

주요 기술 구성

PII 탐지 엔진 (Presidio 기반)

- Microsoft Presidio 활용
    - 한국어 지원을 위한 NER 엔진 추가
    - 사내 전용 엔티티(ID, 메일 등) 탐지를 위한 정규표현식 커스터마이징
- OCR 연동
    - CLOVA OCR 등으로 이미지(PNG, JPG, PDF 등)에서 텍스트 추출
    - 텍스트 후처리를 통한 정제 및 오류 수정
- 출력 포맷
    - 민감정보 탐지 결과를 JSON 형태로 반환

보안 정책 위반 분석 (RAG + LLM)

- 문서 벡터화
    - 내부 보안 문서, 메일 정책 등 임베딩 처리
    - Chroma 기반의 벡터 DB 구축
- RAG 파이프라인 구성
    - 탐지 결과 + 정책 context → LLM 입력
    - 위반 여부를 LLM 판단
- LLM 연동
    - Ollama 기반 로컬 LLM (Mistral) 호출

- FastAPI에서 LLM 연동

- 프롬프트 설계 및 튜닝
  - 정책 판별 전용 프롬프트 엔지니어링
  - 다양한 Role, Domain, Content 등에 대한 조합 테스트

통합 시스템 구성

- API 기반 설계
  - 모든 기능을 API로 제공 → 외부 연동 가능
  - 추후 사내 시스템/메일 서버와 쉽게 통합 가능

구현서비스의 기능이 명확하게 정의되었는가?

기능을 구현하기 위해 서술한 개발환경이 적절한가? 등 작성 ex)개발 플랫폼, 언어, API 등 작성

③ 활용방안(공공성) 및 기대효과

실제 사회에 도움이 될 수 있는가? 기존 서비스의 문제점을 제시하고 개선사항을 도출하였는가? 등 작성

기존 DLP 솔루션은 주로 정규표현식 기반의 패턴 탐지나 고정된 룰 세트를 활용하기 때문에, 문맥에 따른 유연한 판단이 어렵고 비정형 데이터에 대한 대응력도 제한적이다. 특히 기관마다 상이한 내부 보안 정책을 반영하기 어려우며, 복잡한 조건이 결합된 정책은 수작업 검토에 의존하는 경우가 많아 보안 사고를 사전에 차단하기 어렵다.

본 시스템은 이러한 문제점을 해결하고자, 다양한 조직에서 이미 정의하고 운용 중인 내부 보안 정책을 기반으로, 실제 문서 내 민감정보와 전송 상황을 정밀하게 분석하고 정책 위반 여부를 자동으로 판단할 수 있도록 설계되었다. 예를 들어, "팀장급 이상만 외부 메일 전송 가능", "주민등록번호 포함은 내부망 문서로 한정" 등과 같은 조직 특유의 규칙들을 LLM이 직접 해석하고, 민감정보의 종류, 발신자 권한, 수신자 도메인 등의 조건과 결합하여 전송 허용 여부를 동적으로

판단할 수 있다. 이는 정형 규칙이 아닌 실제 업무 맥락을 고려한 판단이 요구되는 현장에 특히 적합하다.

공공기관에서는 민원 응답, 결재 문서 등에서 개인정보 유출을 사전에 차단하는 데 활용될 수 있으며, 의료기관이나 교육기관 등 민감정보를 일상적으로 다루는 환경에서도 동일한 효과를 기대할 수 있다. 민간 기업의 경우, 이메일, 협업 도구, 클라우드 저장소 등 다양한 채널에서 문서가 외부로 전송되는 상황에서 자동 점검 기능을 통해 보안 사고를 방지하고, 규제 준수 및 감사 대응 능력을 강화할 수 있다.

또한 이미지 기반 문서까지 지원하는 비정형 데이터 대응력, 기관별 정책에 맞춘 유연한 구조, REST API 기반의 확장성을 바탕으로, 기존 솔루션으로는 대응이 어려웠던 영역을 실질적으로 보완할 수 있다. 이를 통해 조직은 민감정보 보호 수준을 높이는 동시에, 보안 대응의 자동화와 효율화를 실현할 수 있다.