2025년 새싹 해커톤(SeSAC Hackathon) AI 서비스 기획서

팀명	깜자와 꽁주
팀 구성원 성명	오소민,이나경

1. AI 서비스 명칭

피싱 세이프 AI

2. 활용 인공지능 학습용 데이터

	활용 데이터명	분야	출처
1	한국어 spam/ham 이메일 데이터	한국어 이메일 spam/ham이메일 분류	자체 수집 수동정리
2	SNS/문자 기반 한국어 스팸 메시지 데이터	SNS스팸,문자 사기 메시지	GitHub 공개 한국어 스팸 메시지 GitHub – Ez-Sy01, "KOR_phishing_Detect-Dataset" (Public Dataset)

3. 핵심내용

우리 서비스는 "사용자의 스팸 대응 능력을 높이기 위한 'AI 기반 스팸 예방 교육 플랫폼"이다. 이메일·문자·SNS메시지 등 실제 환경에서 자주 등장하는 스팸 사례를 기반으로, 사용자가 직접 스팸 여부를 판단하고 즉시 피드백을 받을 수 있는 체험형 보안 교육 서비스를 제공한다. 이 서비스는 기존의 단순 스팸 필터와 다르게 사용자 행동 데이터를 중심으로 한 능력향상형 구조를 채택한다. 사용자의 선택 기록·정오답 패턴을 분석하여 AI가 개인 맞춤형 위험도 점수·약점유형·보완 가이드를 자동 생성해준다. 또한 자체 수집한 한국어 스팸/햄 이메일 및 SNS 스팸 데이터를 기반으로 구축한 AI 모델이 메시지를 분석해 위험도(리스크 스코어)를 제공하며, 사용자는 실제처럼 메시지를 확인해보며 학습할 수 있다.

사용자 교육·경각심 제고/사례 기반 학습을 동시에 달성하는 스팸 대응 교육형 보안 서비스의 MVP를 목표를 한다.

4. 제안배경 및 목적

최근 이메일·메시지·SNS플랫폼을 기반으로 한 스팸·피싱 공격이 지속적으로 증가하고 있다. 특히 한국어 기반 스미싱·문자 사기·계정 탈취형 피싱은 사용자의 심리를 이용하는 방식으로 발전하고 있어, 기존의 단순 필터링 기술만으로는 모든 피해를 예방하기 어렵다. 문제는 많은 사용자가 스팸의 위험성을 알고 있음에도 불구하고 실제 메시지에서 어떤 부분이 의심해야 할 요소인지 스스로 판단하는 능력이 부족하다는 점이다. 이로인해 링크클릭, 개인정보 입력 등 사용자 행동 기반 피해가 반복적으로 발생한다.

따라서 우리는 사용자 스스로 스팸을 판별하는 능력을 향상시키는 교육형 보안 서비스의 필요성을 느꼈다. 기존 보안 솔루션이 기술적 차단에 초점을 맞춘 반면, 본 서비스는 "사용자 교육+AI 분석"을 결합해 실제 피해 가능성을 근본적으로 감소시키는데 목적이 있다.

본 프로젝트의 목적은 다음과 같다.

1. **현실 기반 스팸·햄 메시지를 활용한 실습형 교육 제공**

   사용자가 직접 메시지를 보고 스팸 여부를 판단하며, 실전 감지 능력을 체득할 수 있도록 한다.

2. **AI 기반 위험도 분석으로 정확한 피드백 제공**

   단순히 정답만 제시하는 것이 아니라, AI가 메시지의 패턴을 분석하여

   "어떤 요소가 위험한지", "왜 스팸일 가능성이 높은지" 를 직관적으로 설명한다.

3. **개인 맞춤형 취약점 파악을 통한 보안 인식 강화를 목표**

   사용자의 오답 유형을 분석하여 취약한 스팸 유형을 파악하고

   취약 영역에 대한 추가 학습·훈련을 제공한다.

4. **실제 피해 감소에 기여할 수 있는 '사용자 중심' 보안 서비스 개발**

   기술적 필터링을 넘어, 사용자의 보안 의식과 대응 능력을 높여

   스팸·피싱 피해를 근본적으로 줄이는 데 목적이 있다.

5. 세부내용

활용 데이터

본 서비스는 한국어 스팸 메시지 환경에 맞춘 모델을 구축하기 위해

팀이 직접 수집한 다양한 형태의 스팸·햄 데이터를 활용한다.

먼저, 실제 사용자가 받는 한국어 이메일을 기반으로 공지메일, 일반 광고메일, 스팸메일 등 다양한 유형의 메시지를 직접 수집하였다.

또한 GitHub에서 공개된 한국어 스미싱·문자 스팸 데이터셋을 내려받아 문자·SNS 기반 공격 사례도 함께 포함시켰다.

이후 이메일 데이터와 문자 데이터를 통합하여 하나의 학습용 데이터셋으로 재구성했으며, 메시지 내 긴급 문구, 금전 요구 표현, 계정 정지 경고, 링크 포함 여부 등

실제 공격자가 사용하는 패턴이 그대로 포함되어 있어 모델 학습과 사용자 교육 모두 활용할 수 있다.

활용 AI 모델

본 서비스는 머신러닝 기반 텍스트 분류 모델을 중심으로 구현되었다.

가장 성능이 우수했던 모델은 **SVM(Support Vector Machine)**이며, TF-IDF 벡터화를 적용해 학습한 결과 약 **93% 정확도**를 달성하였다.

SVM 모델의 decision score는 Flask 서버에서 0~100 위험도 점수로 변환되어 사용자에게 제공된다.

또한 나이브 베이즈(Naive Bayes) 모델도 비교 모델로 함께 사용하였고, 예측 속도가 빠르기 때문에 퀴즈형 학습에서 즉각적인 피드백을 제공하는 용도로 적용할 수 있다.

서비스 아이디어 개요

본 서비스는 단순히 스팸 여부를 자동으로 분류하는 기술을 넘어서, 사용자가 직접 메시지를 보고 "스팸인지 햄인지" 판단해보는 과정을 통해 실제 상황에서 스스로 위험 메시지를 구별할 수 있는 능력을 키워주는 교육형 보안 서비스를 목표로 한다.

사용자가 문제를 풀 듯 메시지를 판별하면 AI가 해당 메시지에서 어떤 부분이 위험한 요소인지 분석해 설명해주고 위험도를 숫자로 제공하여 사용자가 이해하기 쉽도록 구성하였다.

즉, 자동 차단 중심의 기존 스팸 필터와 달리 사용자 참여+AI 분석이 결합된 형태로, 실제 보안 인식 향상에 초점을 둔 서비스다.

적용 기술

서비스는 자연어 처리 기반 텍스트 분석 기술을 중심으로 구성된다.

메사지는 불용어 제거, 전처리, TF-IDF 백터화 과정을 거쳐 SVM 및 나이브 베이즈 모델이 입력된다.

AI 모델이 산출한 결과는 Flask 서버를 통해 웹 화면으로 전달되며, 스팸 여부 이외에도 메시지의 위험도 점수와 위험 키워드 (예:긴급,계정정지,URL 링크)를 함께 제공한다.

이과정에서 Flask는 웹 서버 역할을 담당하고, 모델은 메시지를 분석하는 AI 역할을 수행하여 각자의 기능이 명확하게 분리된 구조로 설계된다.

서비스 방법

사용자는 웹 페이지를 통해 실제 스팸·햄 메시지와 비슷한 형식의 텍스트를 확인한다. 메시지를 읽은 뒤, 스스로 '스팸 / 햄'을 선택하는 방식으로 서비스를 이용하게 된다. 선택이 이루어지면, Flask 서버는 해당 메시지를 AI 모델(SVM)에 전달하여 스팸 여부를 판단하고, 모델이 계산한 결과를 토대로 0~100 사이의 위험도 점수를 생성해 사용자에게 보여준다. 이와 함께 메시지 안에서 위험성이 높은 문구나 패턴을 간단한 설명 형태로 제공하여 사용자가 왜 이 메시지가 위험한지 이해할 수 있도록 돕는다. 초기 MVP 단계에서는 사용자의 선택 결과와 AI 판단 결과를 간단한 형태로 기록해두어, 사용자가 어떤 유형의 메시지를 어려워했는지 기본적인 경향을 파악할 수 있게 한다. 이를 기반으로 사용자가 자주 헷갈리는 유형의 메시지를 다시 연습할 수 있도록 안내하는 기본적인 재학습 기능을 제공한다.

서비스의 창의성 및 구현 가능성

이 서비스의 차별점은 기존의 스팸 필터처럼 "AI가 자동으로 판단하는 기능"뿐 아니라, 사용자 스스로 판단해보고 학습할 수 있도록 만든 교육형 구조에 있다. 실제 메시지 기반의 실습형 경험과 AI 설명·위험도 분석이 결합되어 있기 때문에,보안인식을 자연스럽게 강화할 수 있다. 또한 머신러닝 모델(SVM)과 Flask 웹 구조만으로도 MVP 수준의 구현이 가능하여 해커톤 기간 내에 충분히 개발 가능한 서비스 형태를 가진다.

UI/UX 이미지 시각화

본 서비스의 UI/UX는 사용자가 스팸 여부를 손쉽게 확인하고, OX 형태의 학습 기능을 자연스럽게 사용할 수 있도록 단순하고 직관적인 흐름으로 설계하였다.

주요 화면 구성은 다음 두 가지 기능을 중심으로 이루어진다.

1) 메시지 분석 화면(UI 설명)

서비스의 핵심 기능은 사용자가 직접 메시지를 입력하면, AI가 스팸 여부와 위험도를

분석해주는 기능이다. 이를 위해 메시지 입력 중심의 단순한 UI 구조를 사용한다.

- 화면 상단에 텍스트 입력 박스를 배치하여

  이메일·문자 내용을 그대로 붙여넣을 수 있도록 한다.

- 입력 후 "분석하기" 버튼을 누르면

  AI가 메시지를 분석하여 결과 페이지로 이동한다.

분석 결과 화면(UI 설명)

AI 분석이 완료되면, 사용자는 다음 정보를 한눈에 확인할 수 있다.

- **스팸 / 햄 분류 결과**

  메시지가 스팸인지 정상 메시지인지 명확하게 표시한다.

- **위험도 점수(0~100%)**

  스팸 가능성을 직관적으로 이해할 수 있도록

  숫자 또는 간단한 게이지 형태로 위험도를 제공한다.

- **위험 요소 설명**

  메시지 안에서 AI가 감지한 핵심 위험 문구(예: 긴급성, 금전 요구, 링크 포함 등)를 간단한 텍스트 형태로 요약하여 보여준다.

전체 화면은 복잡한 그래픽 요소 없이 "결과를 빠르게 확인할 수 있는 구조"를 목표로 설계된다.

2) OX 학습 페이지(UI 설명)

본 서비스의 두 번째 주요 기능은 사용자가 랜덤으로 제공되는 메시지를 보고

직접 스팸 여부를 판단해보는 OX 학습 기능이다.

- 메시지는 카드 형태의 간단한 박스로 표시되며

  짧은 문장부터 실제 스팸 사례까지 다양하게 구성된다.

- 사용자는 O(스팸이라고 판단) / X(정상이라고 판단)

  두 버튼 중 하나를 선택해 자신의 판단을 기록한다.

OX 선택 후 피드백 화면

사용자의 선택 후에는 즉시 피드백이 제공된다.

- 사용자의 선택이 정답인지 오답인지 명확히 표시하며
- 해당 메시지가 왜 스팸 또는 정상인지에 대한 AI 분석 근거를 짧게 설명한다.
  (예: 계정 정지 문구 포함, 링크 패턴 이상 등)
- 필요 시, 간단한 위험도 점수도 함께 제공해 학습 효과를 높인다.

6. 기대효과

본 서비스는 스팸 메시지로 인한 피싱, 개인정보 유출, 계정 탈취 등 실제 피해를 줄이는 데 직접적인 도움을 줄 수 있다. 기존 스팸 차단 서비스가 자동 필터링에만 초점을 맞추는 것과 달리, 본 서비스는 사용자가 스스로 메시지를 판단해보는 경험을 제공하여 보안 인식을 자연스럽게 높인다.

첫째, 사용자들은 실제 스팸 사례를 기반으로 메시지를 직접 분석해보면서 스팸 유형과 위험징후를 빠르게 학습할 수 있다. 단순한 정보 제공을 넘어 스스로 선택하고 결과를 확인하는 방식은 교육 효과를 극대화한다.

둘째, AI가 제공하는 위험도 점수와 간단한 위험 요소 설명을 통해 사용자는 메시지에서 어떤 부분을 의심해야 하는지 명확히 이해할 수 있다. 이는 기존 필터링 시스템이 제공하지 못하는 '이해 기반 보안'이라는 가치를 제공한다.

셋째, 메시지 분석 기능과 OX 학습 기능을 결합함으로써 사용자는 반복적인 실습을 통해 실제 상황에서 사기 메시지를 구별할 수 있는 능력을 향상시킬 수 있다. 이는 스미싱, 계정 도용형 피싱 등 일상에서 자주 발생하는 보안 사고를 예방하는 데 실질적인 효과를 가져올 것으로 기대된다.

넷째, 서비스 구조가 단순하고 확장성이 높아, 향후 메시지 유형 추가, 사용자 맞춤형 학습 기능 강화 등 지속적인 발전 가능성이 크다. 초기 MVP 단계에서 구축한 모델과 UI 기반으로 다양한 교육형 보안 도구로 확장할 수 있다.

종합적으로 본 서비스는 개인의 보안 인식을 높이고 실제 피해 발생 가능성을 감소시키며, 사회 전반의 보안 수준 향상에 기여할 수 있을 것으로 기대된다.

※ 상세 설명을 위해 도표, 스케치 등 별도파일 추가 가능

※ 제출한 기획서는 온라인 예선 심사 전 구체화하여 깃허브(GitHub)에 필수로 게시