2023 전기 졸업과제 착수보고서

멀티모달 기반 스팸 필터링 플랫폼

스팸도시락

202055550 박혜경

202055578 이승현

202055608 천영채

목차

1. 과제 배경

스팸은 요청하지 않은 메시지(이메일, 문자 또는 전화)를 대량으로 불특정 다수에게 전송하는 것을 의미한다. 개인정보 침해와 스팸은 인터넷 기술의 발전으로 더욱 빈번해졌다.

한국인터넷진흥원에 따르면 2020 년에 조사한 결과, 약 800 만 건의 스팸 문자가 발송된 것으로 나타났다. 특히, 정부의 재난 지원금이나 손실 보상을 지원한 2020~2021 년에는 금융회사를 사칭해 대출을 해준다는 스팸이 많아졌다.

단순히 수신자가 원하지 않는 정보를 퍼뜨리는 목적이었던 과거와는 달리, 최근에는 개인정보를 탈취하고 데이터를 삭제하며 금전을 요구하는 악의적인 스팸이 더 많아졌다. 이러한 스팸은 수신자의 개인정보를 탈취하고 데이터를 삭제하며, 금전을 요구하는 악성 행위가 대부분이다. 또한 수신자들은 불편함을 느끼게 되어 시간을 낭비할 뿐만 아니라, 바이러스에 감염될 가능성도 있다.

기존의 스팸 필터링 시스템은 이미 신고된 스팸 데이터베이스와 비교해 동일 문자나 이미지 등을 필터링하는 방식이다. 그러나 이 방식은 발신자가 문구를 교모하게 바꾸거나 특수문자를 사용하는 경우에는 필터링이 어렵다. 반면, 딥러닝을 기반으로 한 스팸 필터링은 AI 가 반복된 학습을 통해 단어들의 패턴을 발견해 정확도를 높일 수 있어 더욱 정확한 스팸 필터링이 가능하다.

2. 과제 목표

2.1

다양한 형식의 스팸 필터링 기능//현재 SNS 및 보안이 취약한 웹페이지의 회원가입 등으로 개인정보가 많이 퍼져 있는 상태이다. 이에 많은 사람들이 개인 이메일, 휴대폰 번호 등을 통해 스팸 메시지를 받고 있으며 이는 개인에게 불필요한 내용을 전달하는 것뿐만 아니라 보이스피싱 등 사기나 범죄의 피해자가 될 수도 있는 큰 문제이다. 이를 예방하기 위한 적절한 대응책이 필요한 시점이다.

효율적으로 스팸 메시지를 막기 위해서는 텍스트뿐만 아니라 이미지, 음성 등 다양한 형태의 데이터로 존재하는 스팸 메시지를 구별할 수 있어야 한다. 데이터의 유형별로 존재할 수 있는 형태는 다음과 같다.

텍스트	이메일, 문자 메시지, 채팅
이미지	첨부파일, SNS 의 프로필 이미지
음성	음성 메시지, 통화

성공적인 스팸 필터링을 위해서 기계학습을 통해 여러 스팸 데이터를 학습시키고, 새로운 데이터가 주어졌을 때 스팸 여부를 정확하게 판단할 수 있도록 해야 한다. 또한 사용자가 설정한 키워드 및 규칙에 대해서 스팸으로 분류하도록 규칙 설정이 가능해야 하도록 구현해야 한다. 딥러닝과 머신러닝을 기반으로 위와 같은 조건들을 만족하는 효과적인 스팸 필터링 모델을 개발하도록 한다.

2.2

서비스 제공 웹페이지를 통해 스팸 필터링 결과를 확인할 수 있도록 한다. 필터링 된 메시지의 개수와 필터링 된 이유로 원본 데이터와 판단 근거 등을 제시해주는 웹을 개발한다. 추가적으로 스팸으로 분류된 메시지에 대해 사용자가 피드백을 제공할 수 있도록 한다. 스팸이 아닌데 스팸으로 분류된 것에 대해 피드백을 받고 스팸 필터링의 정확도를 개선하도록 한다.

3. 요구 조건 분석

본 졸업과제는 텍스트 데이터, 이미지 데이터, 음성 데이터 스팸에 대한 필터링 플랫폼을 개발하는 것을 목표로 한다. 스팸 필터링 모델을 통해 나온 결과들은 스팸으로 인한 사용자의 피해를 줄여줄 수 있을 것이다.

- 데이터 수집
  - 텍스트 데이터
    - 온라인 상에 존재하는 스팸 데이터(Enron dataset, SMS spamdataset)를 수집한다.
  - 이미지 데이터
    - 온라인 상에 존재하는 Image Spam Dataset 을 수집한다.
    - 피싱 사이트에 존재하는 이미지를 수집한다.
  - 음성 데이터
    - 피싱 전화 사기 수법 패턴에 대해 수집한다.
- 수집한 데이터의 스팸 여부 분석
  - 텍스트 데이터
    - Python 을 이용하여 수집한 데이터를 전처리, 분석하여 모델에 넣을 수 있게 가공한다.
    - 가공한 데이터를 모델에 넣어 스팸 여부를 계산한다.
  - 이미지 데이터
    - Python 을 이용해 이미지의 텍스트를 추출하여 텍스트 토큰으로 만든다.
    - 만들어진 토큰을 연결하여 스팸 필터링 모델에 넣을 수 있는 데이터로 가공하여 모델에 넣어 스팸 여부를 계산한다.
  - 음성 데이터
    - Python 과 음성인식 API 를 활용하여 화자를 분리하여 대화 스크립트 형식으로 추출한다.
    - 추출된 문장을 모델에 넣을 수 있는 데이터로 가공하여 모델에 넣고 스팸 여부를 계산한다.
- 서비스 제공 형식
  - 웹 형식으로 제공하여 메일이나 이미지를 받았을 경우 데이터를 넣어 확인할 수 있게 한다.
  - 음성 서비스의 경우 파일을 넣는 것과 실시간 탐지 기능을 구분하여 제공하며, 모바일 웹을 통해 통화중에도 사용할 수 있게 하여 실시간 통화의 스팸 여부도 계산할 수 있도록 한다.

4. 현실적 제약 사항 및 대책

4.1 제약사항

- 기존 필터링 시스템을 우회하기 위한 지능적인 스팸 메시지가 증가했다. (ex) 대출 → ㄷ H★ ㅊㅜㄹ)
- 음성 스팸 메시지의 경우 기존 자료가 많지 않다.
- 음성 스팸 메시지의 경우 스팸 음성 감지와 동시에 전화를 차단하기에 한계가 있다.

4.2 대책

- 딥러닝을 기반으로 한 반복된 학습과 빅데이터 분석기술을 통해 교묘하게 단어를 바꾸어 보내는 메시지도 탐지할 수 있게 한다.
- 직접 음성 스팸 데이터를 모은다. 기존 데이터와 직접 모은 데이터를 같이 활용할 예정이다.
- 녹음된 음성을 첨부하는 형식과, 실시간으로 음성을 들려주면 스팸 여부를 알려주는 기능을 추가할 예정이다.

5. 설계 문서

5.1 개발 환경

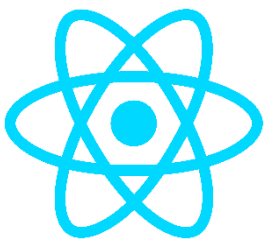
파이토치

스팸 필터링을 위해 스팸데이터를 학습시켜야 하므로 머신러닝을 위해 파이토치를 사용한다.

파이썬

스팸 필터링 결과의 통계 분석 및 시각화를 위해 파이썬을 사용한다.

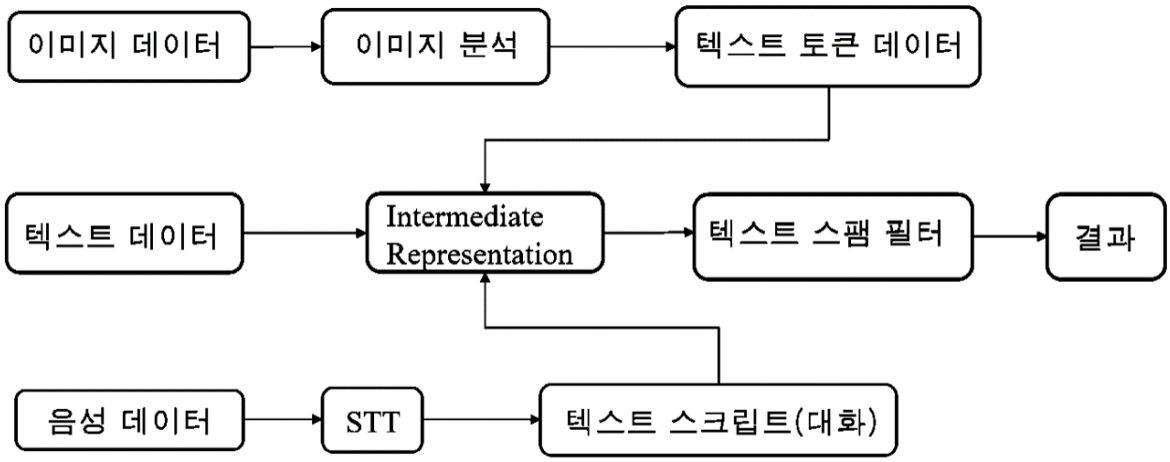
리액트 & 스프링



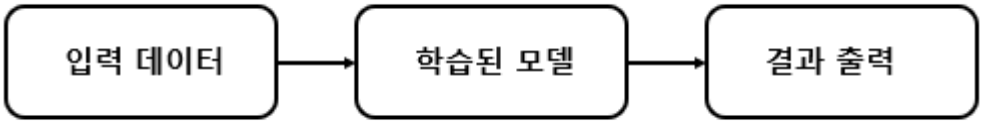
스팸 필터링 현황 및 결과를 사용자에게 제공할 웹을 구성하기 위해 리액트와 스프링을 사용한다.

5.2 프로세스

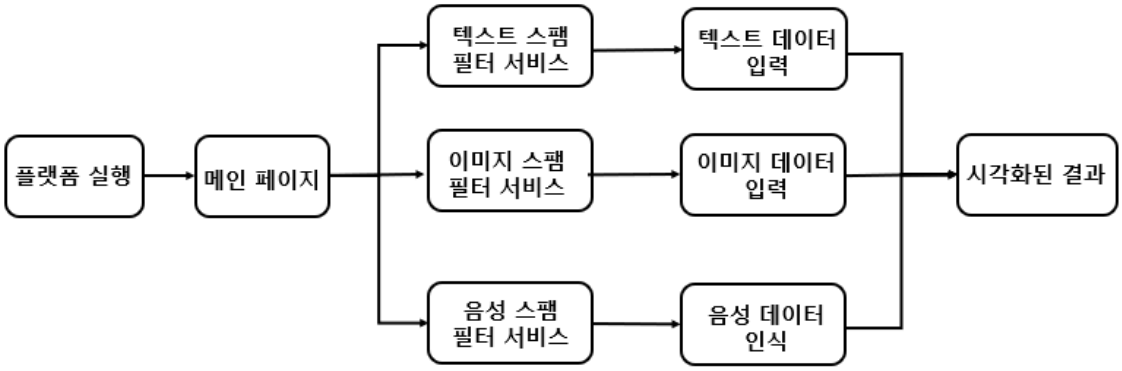
1. 학습 모델 구상 : 이미지, 텍스트, 음성에서의 텍스트 형식이 다르다고 판단하여 각각 다르게 데이터를 추출하고 이를 스팸 필터에 넣을 수 있는 IR(Intermediate Representation)으로 가공한다.



2. Server

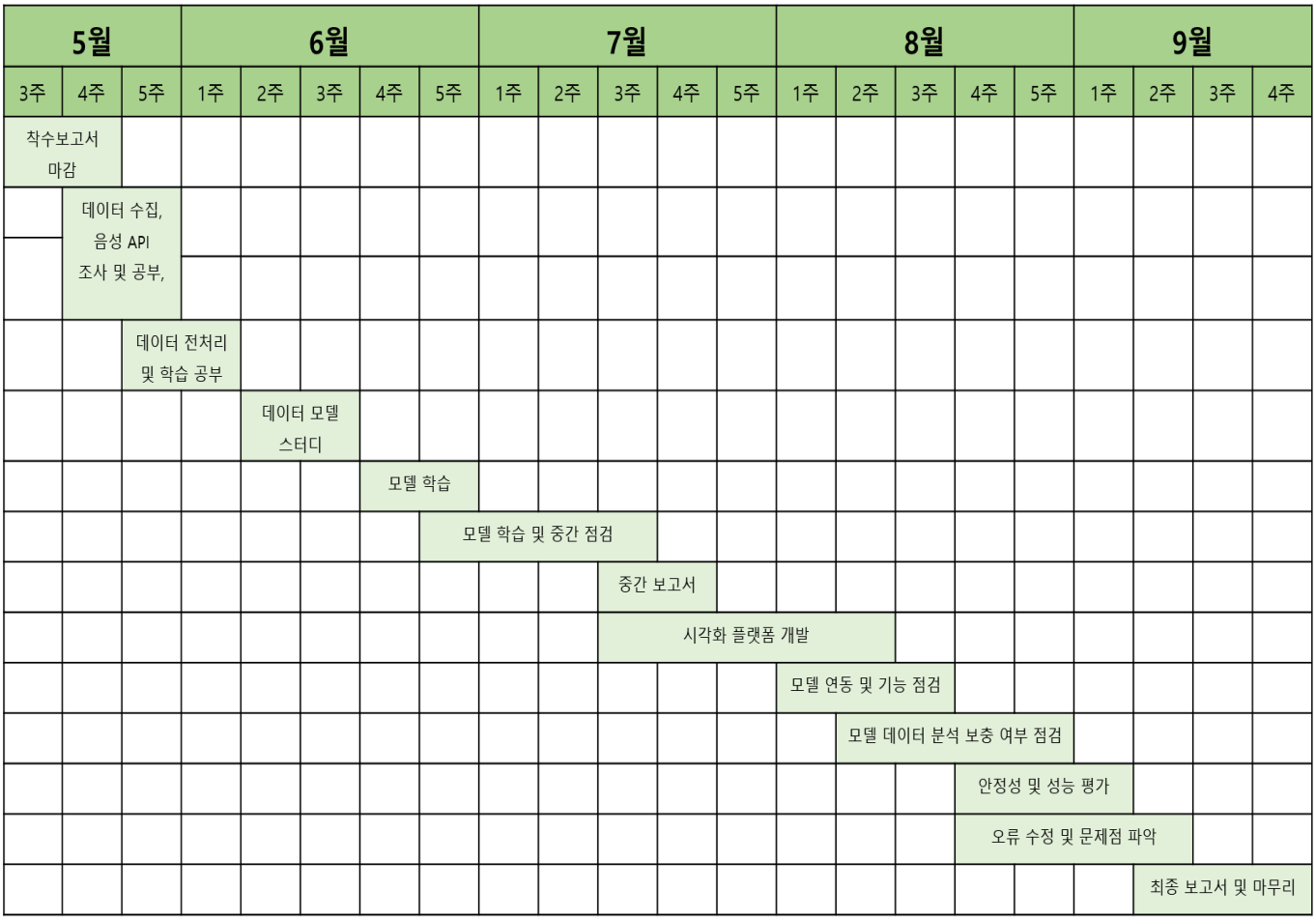


3. Client



6. 개발 일정 및 역할 분담

6.1 개발 일정



6.2 역할 분담

이름	역할 분담
박혜경	- 텍스트 데이터 수집 - 데이터 전처리 - 모델 학습
이승현	- 이미지 데이터 수집 - 모델 학습 - 시각화 플랫폼 개발
천영채	- 음성 데이터 수집 - 모델 연동 - 모델 학습