졸업과제 착수 보고서

멀티모달 기반 스팸 필터링 플랫폼 개발

지도교수	최윤호
팀명	멀티모발
이름	**201824534** 윤상호
	201824636 이강우
	202055651 조재홍

1. 과제 배경 및 목표

1.1 과제 배경

전자우편(e-mail)은 영리 목적의 광고성 정보인 스팸메일과 정상적인 내용을 포함하는 햄메일으로 구분할 수 있다. 스팸메일이란 전자우편이나 휴대폰 등 정보통신서비스를 이용하는 이용자의 단말기로 본인이 원하지 않음에도 일방적으로 이용자에게 전송되는 영리 목적의 광고성 정보를 말한다. 그러나 최근에는 일반적으로 스팸메일을 상업적인 목적의 전자우편만으로 한정하지 않고, 본인이 원하지 않는 전자우편 또한 스팸메일이라고 하며, UCE(Unsolicited commercial e-mail) 또는 UBE(Unsolicited bulk e-mail)라고 부르기도 한다.

스팸메일의 큰 특징 중 하나가 대량성(bulk)으로 프로그램을 통해 매우 손쉽게 불특정 다수에게 무수히 많은 양의 스팸메일을 반복적으로 전송할 수 있다는 점이다. 이러한 특징때문에 스팸메일을 받는 이용자의 불편을 유발, 필요한 정보 수신을 방해 및 메일 서버의 과부하 초래 등 많은 문제점을 야기한다.

이러한 스팸메일을 분류하는 기법은 규칙기반 스팸메일 필터링과 기계학습 기반 스팸메일 필터링으로 나누어진다.

규칙기반 스팸메일 필터링은 사용자가 정의한 문자열 집합에 따라서 스팸메일을 탐지한다. 하지만 스패머들에 의해 광고의 내용이 자주 변하기 때문에 키워드나 규칙을 이용하는 것으로는 스팸메일을 높은 확률로 탐지하기 어렵다.

기계학습 기반 스팸메일 필터링은 최종 이용자가 수신 전자우편의 특징을 분석하여 수작업으로 규칙을 만드는 대신, 소프트웨어가 수신 전자우편의 스팸메일 여부를 판별하여 필터링 규칙을 자동으로 생성하는 장점이 있다.

하지만 이러한 기계학습 모델들은 단일 데이터 형태, 즉 하나의 모달리티만을 학습하여 task를 수행하기 때문에 가끔 이질적인 결과를 출력하는 한계점을 보인다. 따라서 한계점을 극복하기 위해 다양한 모달리티를 학습에 이용하여 보다 사람처럼 학습하고 결과를

추론하는 방법을 멀티모달 러닝이라 부르며 이를 활용하여 스팸메일을 필터링하는 머신러닝 플랫폼을 개발하고자 한다.

1.2 과제 목표

본 졸업과제는 멀티모달 딥러닝/머신러닝 기반 스팸 필터링 플랫폼 개발을 목표로한다.

- 수집 텍스트 스팸 데이터, 생성 텍스트 스팸 데이터, 수집 이미지 스팸 데이터를 모달리티로 활용한다.
- 해당 모달리티들을 기반으로 스팸 필터링 모델을 학습시키고 학습된 모델로 스팸메일 여부를 판별한다.
- 시각화 인터페이스를 통해 결과(필터링 된 스팸메일 수, 필터링 된 이유 등 스팸 분류에 유의미한 정보)를 확인한다.

2. 요구 조건 분석

2.1 멀티모달 활용

수집 텍스트 스팸 데이터, 생성 텍스트 스팸 데이터, 수집 이미지 스팸 데이터, 총 세종류의 데이터 셋을 스팸 필터링 모델의 학습과정에 사용한다.

각 데이터 셋의 수집 과정은 다음과 같다.

- 수집 텍스트 스팸 데이터: Kaggle 데이터 저장소에서 공개 데이터를 다운로드하여 사용한다. 스패머들은 일반적으로 스팸 필터링을 우회하기 위해 새로운 기법들을 지속적으로 개발하기 때문에, 하나의 데이터 셋을 사용하는 것은 부적절하다.
  따라서 데이터 수집시 데이터가 업로드된 시점을 다양화하여 여러 데이터셋을 수집한다.
  또한, 한국어로 이루어진 텍스트 스팸데이터는 포털사이트 스팸메일함에서 크롤링을 통해 수집한다.
- 생성 텍스트 스팸 데이터: 생성을 위해 생성형 AI서비스인 ‘Chat GPT’, ‘Google Bard’, ‘MS Bing’을 사용한다.

이러한 생성형 AI서비스는 스팸메일 생성과 같은 불법적인 행위를 제공하지 않기 때문에, 스팸메일의 키워드가 되는 단어가 일정 횟수 이상 반복되도록하는 규칙을 통해 영리 목적의 광고성 메일로 분류되어지는 스팸 데이터를 생성한다.

- 수집 이미지 스팸 데이터: Kaggle 데이터 저장소에서 스팸 이미지 데이터를 다운로드하여 사용한다. 공개 데이터저장소에 이미지 스팸 데이터셋이 많지 않기 때문에, beautifulsoup4 패키지를 활용하여 웹에서 이미지 크롤링을 통해 데이터를 추가적으로 확보한다.

2.2 데이터 전처리

2.2.1 텍스트 데이터 전처리

- 영문 메일인 경우 알파벳과 숫자를 이외의 특수문자들은 모두 제거하여 하나의 문자열을 생성한다.
- 한글 메일인 경우 한글과 숫자 이외의 특수문자들은 모두 제거하여 하나의 문자열을 생성한다.
- NLTK, KoNLP 패키지를 사용하여 문자열을 형태소(morpheme)라는 최소 의미 단위로 분리한다.
- WordPunctTokenizer, spaCy, kss, kkma 등 토크나이저 패키지를 사용하여 문장을 토큰 시퀀스로 분절하는 토큰화를 수행한다.
- 토큰화된 데이터는 Keras에서 제공하는 도구인 embedding()을 이용하여 임베딩 벡터로 만든다.

2.2.2 이미지 데이터 전처리

- Open CV 패키지를 사용하여 이미지의 색상 특성을 Grayscale로 변환한다.
- Grayscale로 변환한 이미지를 이진화(Binary) 이미지로 변환한다.
- 위의 두 과정은 이미지를 흑백화하여, 텍스트 추출을 용이하게 한다.
- Google Vision api를 사용하여 이미지에서 텍스트 만을 추출한다.
- 이후 과정은 텍스트 데이터 전처리와 동일하다.

2.3 스팸 필터링 모델 선택

- 입력한 메일 스크립트가 스팸 메일인지, 정상 메일인지 판별해주는 모듈.
- 스팸 필터링은 입력이 스팸인지 정상인지 분류하는, 이진 분류 모델이다.
- 스팸 메일 분류 머신 러닝 기술로는 Clustering, Naive Bayes Classifiers(이하 NBC), Neural Network(이하 NN), Firefly Algorithm(이하 FA), Rough Set Classifiers(이하 RSC), Support Vector machine(이하 SVM), Decision Tree(이하 DT), Ensemble(이하 앙상블), Random Forests(이하 RF), Deep Learning(이하 DL) 등이 있다.
- 스팸 필터링 모델로 가장 많이 사용되는 NBC는 스팸 내용에 특정 단어들이 많을수록 스팸으로 분류될 확률이 높다는 점을 이용하는 방식인데, 해당 단어들을 스패머가 경험을 통해 알게되면, 특정 단어를 회피하여 스팸 메일을 작성할 수 있다는 단점이 있어 선택하지 않았다.
- 본 졸업과제에서는 학습 데이터가 많을수록 정확도가 높아지는, 인공 신경망을 활용한 스팸 필터링 분석 모델이 해당 주제에 적합하다고 생각하여 선택하였다.

2.4 Graphic User Interface

2.4.1 PyScript

위의 과정에서 전처리가 완료된 데이터를 이용해 학습모델이 학습한 후에, 실사용에서 모듈의 수행 능력을 검증하기 위한 출력 데이터를 받아 웹페이지에 시각화하기 위해 PyScript를 이용한다.

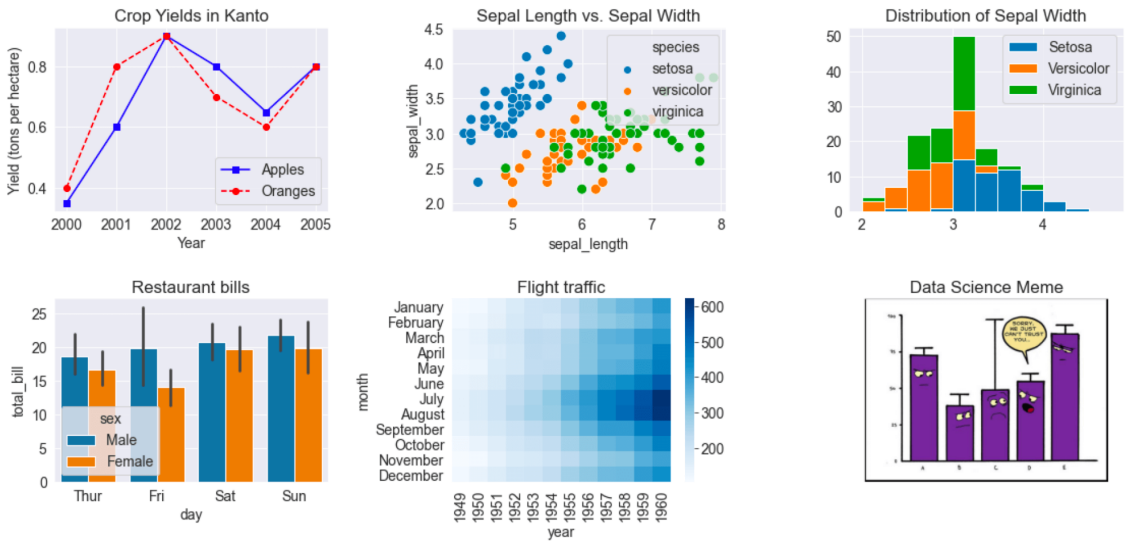
- HTML과 Javascript, Python의 라이브러리들은 그대로 사용하면서 파이썬으로 모든 웹 개발을 할 수 있다.
- 웹브라우저에서 열기만 하면 바로 코드를 실행할 수 있다.
- 데모 프레임워크 없이도 웹에서 시연할 수 있다.

[1]

[그림 1] PyScript

2.4.2 Matplotlib

- PyScript에서 생성한 element에 Matplotlib과 각종 Event Handler를 이용해 스팸 필터링 전과 후의 비교를 여러 그래프로 시각화한다.
- 기존의 스팸 필터링 모듈에 비해 개선점이 있다면 기존 모듈과의 차이점을 그래프로 시각화한다.
- 어떤 단어가 스팸으로 선정 되었는지 이유 (홍보성, 가입성 등등)



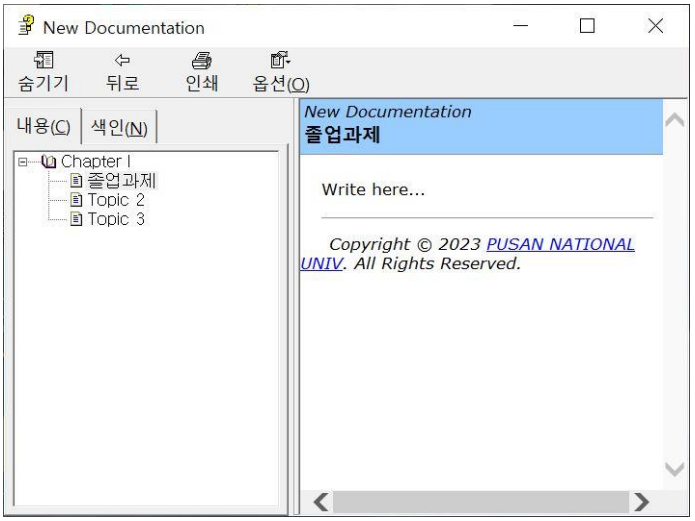
[2]

[그림 2] Matplotlib

[1] https://velog.io/@taekkim/PyScript-%EB%9E%80
[2] https://jovian.com/aakashns/python-matplotlib-data-visualization

2.4.3 CHM editor

- 프로그램 실행 매뉴얼
- 구현 과정 및 실행 사진
- 테스트 케이스
- 개발과정 기록



[그림 **3] CHM editor**

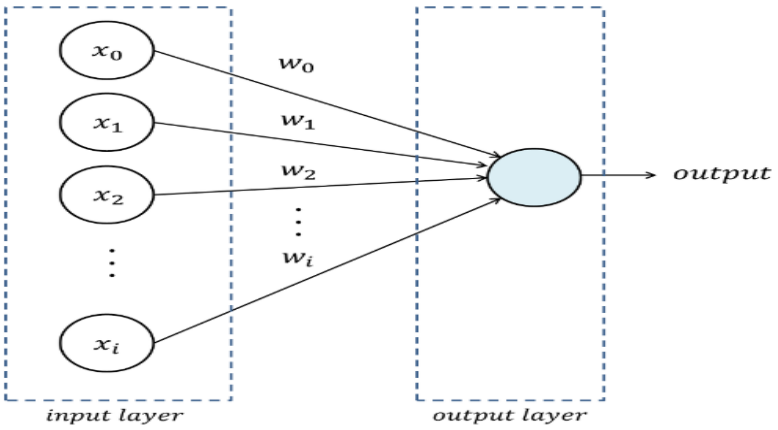
3. 연구 방향

3.1 인공신경망을 활용한 스팸 필터링 모듈 개발

3.1.1 인공신경망 개요

인공신경망은 인간의 뇌 신경망 뉴런의 구조를 본 떠서 만들어진 기계학습 알고리즘이다. 스팸메시지 혹은 정상메시지라는 결과를 알고 있는 분류된 메시지들을 통해 인공신경망을 학습시켜, 각 메시지들에 가중치를 통해 스팸 메시지를 효과적으로 탐지할 수 있다.

인공신경망은 앞으로 새로운 유형의 스팸메시지들을 대처하기위해, 새로운 데이터를 통해 학습할 수 있으므로 본 졸업과제에 적합하다고 판단했다.



[그림 **4]** 단층 퍼셉트론 예시

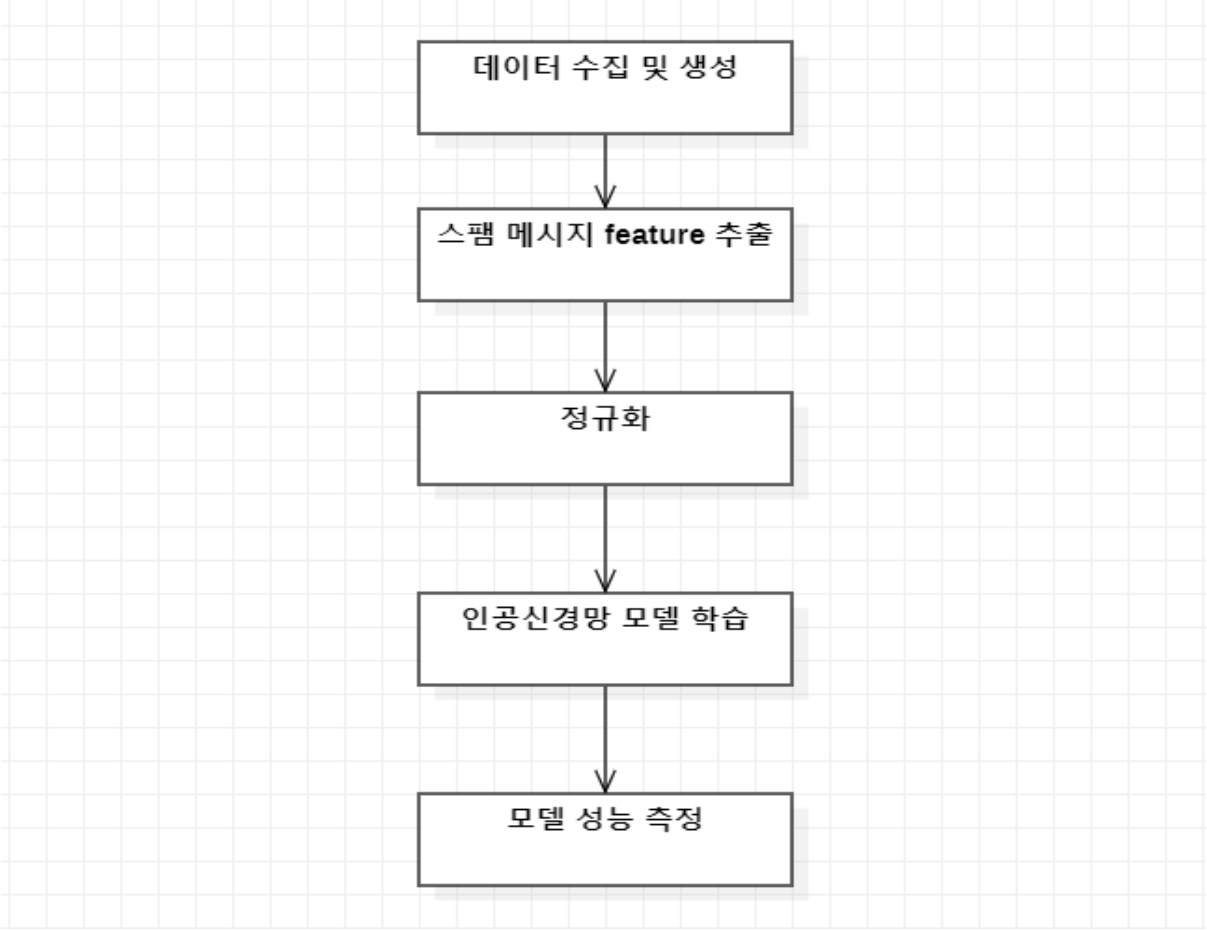
인공 신경망은 신경세포를 모델로 한 뉴런을 노드라고 지칭하고, 이 노드들을 결합하여 하나의 네트워크를 만드는 방식이다.

입력 계층의 입력으로는 보통 원본의 특징을 기반으로 여러개의 입력으로 분류하여 각 입력마다 서로 다른 가중치(weight)를 가지게된다.

예를 들어, 스팸 메시지의 경우 스팸 메시지의 특징을 기반으로 스팸 키워드 빈도수, 홍보성 문구의 빈도수, URL 빈도수 등과같이 스팸으로 분류하기위한 특징들을 분석해서 feature를 추출하여 입력계층의 입력으로 설정할 수 있다.

여러개의 입력 값들은 각각의 가중치가 부여되고 모든 입력의 가중치의 합을 시그모이드 함수 등과 같은 활성화함수에 적용하여 최종 결과값을 얻는다.

3.1.2 인공신경망 모델 학습 과정



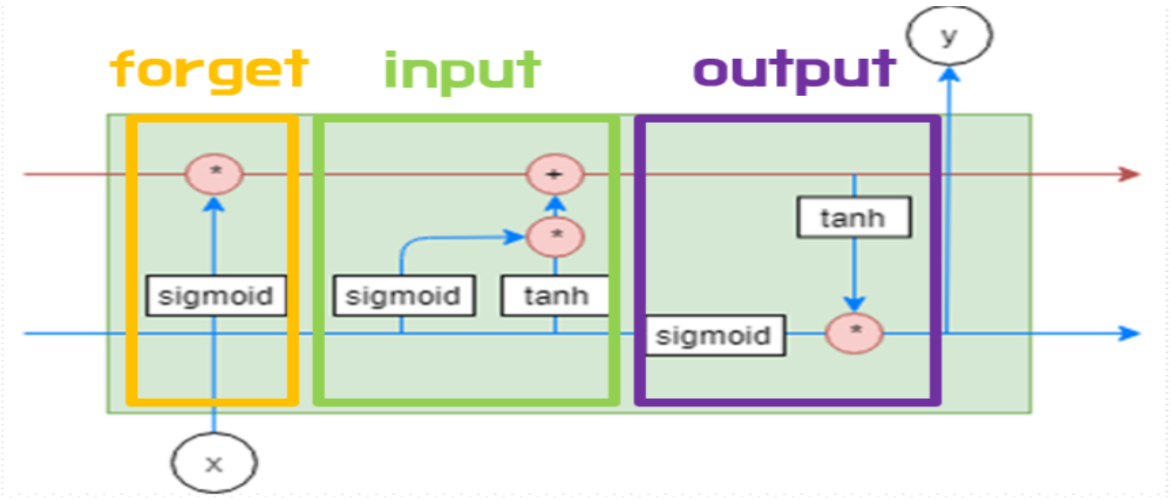
[그림 **5]** 인공신경망 학습 과정 예시

신경망의 입력은 가중치를 이용하므로 0과 1사이의 값을 부여받게 되며, feature에 따라 학습 데이터를 0과 1사이의 값으로 만들어주는 것을 정규화라고 한다.
예를 들어, feature를 키워드 “가입” 의 빈도수라고 할 때, 개수가 0~3개인 것은 0으로, 4~6개인것은 0.1로 50개 이상인것은 1로 mapping 하는 것이 추출한 feature를 이용한 정규화 과정이다.
인공신경망 모델의 성능은 feature 값과 정규화범위, 입력 계층에서 입력 변수들의 개수 등에 따라 달라지게 되는데, 본 실험에서는 성능 변화 요인들을 경험적으로 바꿔가며 가장 성능이 좋은 모델을 찾을 예정이다.
모델의 성능을 측정하기 위한 지표로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), FP-Rate(FPR, 거짓 참 비율) 등의 지표들을 이용할 예정이다.
추가적으로 단일 퍼셉트론의 한계인 XOR과 같은 복잡한 논리게이트를 구현할 수 없다는 점등을 고려할 때, 다중 퍼셉트론을 고려하여 성능을 높이는 점도 고려해볼 생각이다.

3.1.3 인공신경망 분류 및 적합한 모델 선정**.**

인공신경망(Artificial Neural Network, 이하 ANN)은 사용되는 알고리즘에따라 여러 분류 방식과 모델이 사용되고 있다. ANN의 종류로는 Deep Feedforward Network(이하 DFN), Recurrent Neural Network(RNN), Long Short Term Memory(LSTM), Autoencoder(이하 AE), Variational Autoencoder(이하 VAE), Convolutional Neural Network(CNN), Deep Residual Network(DRN), Generative Adversarial Network(GAN), Graph Neural Network(GNN), Spiking Neural Network(SNN), Graph Convolutional Network(GCN) 등이 있다.
많은 인공신경망 종류가 있지만, 디지털콘텐츠학회논문지에 탑재된 ‘인공신경망을 적용한 악성 댓글 분류 모델들의 성능 비교’ 논문을 참고하여 LSTM을 적용한 모델이 정확도가 대부분의 결과에서 높은 순위를 차지하는 것을 토대로 LSTM을 선정하였다.

3.1.4 LSTM(Long Short Term Memory)

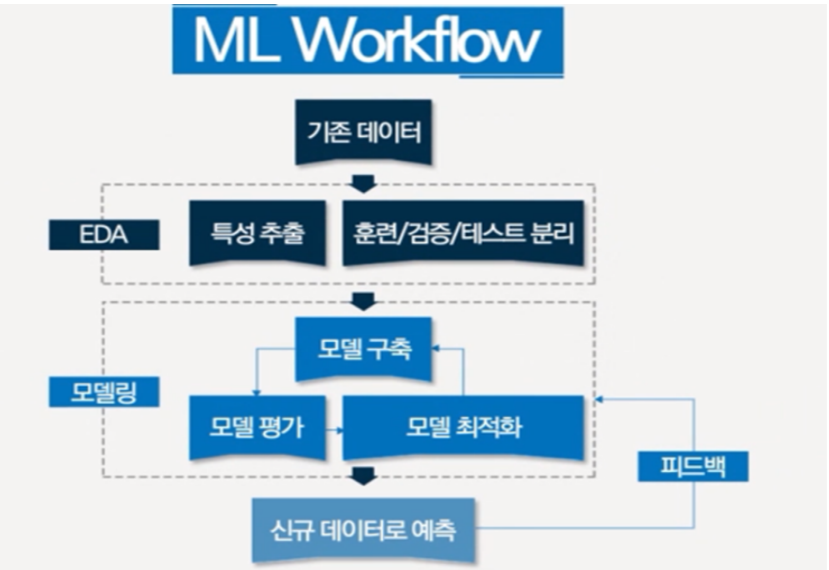


[그림 6] LSTM (Long Short Term Memory)

LSTM은 RNN에서의 히든 계층을 input, output, forget 3개의 게이트로 구성하는 메모리 블럭으로 대체한 구조이다. 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장기/단기 기억을 가능하도록 설계한 신경망 구조이다.

즉, LSTM의 신경망은 3개의 게이트 조절을 통해 이전 신경망 정보가 현재 신경망에 끼치는 영향을 조절할 수 있으며, 현재 입력과 연관된 정보를 추가할 수도 있고 출력에 끼치는 영향 수준을 정할 수도 있다.

3.2 머신러닝 워크플로우



[그림 7] 머신러닝 워크플로우

머신 러닝의 전반적인 작업 과정을 서술한다.

- 먼저, 머신 러닝에서 활용할 데이터의 수집이 필요하다. 본 졸업과제에서는 e-mail 형식의 텍스트 데이터와 이미지 데이터를 수집하고자 한다.
  이미지 데이터는 텍스트로 변환하여, 머신 러닝 모델에 입력으로 들어갈 수 있게끔 해주어야 한다. 또한, 텍스트 데이터들은 스팸으로 분류되기 위한 특성(feature)들을 추출하거나, 필요한 데이터 형식으로 정제하여야 한다. 이를 데이터 전처리 과정이라고 한다.
- 전처리가 완료된 데이터를 모델에 적용시켜 모델을 학습시킨 뒤 신규 데이터를 생성하여 해당 모델을 테스트하고 성능을 평가한다.
- 본 졸업과제에서는 기존의 스팸 분류 모델을 사용하고, 직접 생성한 데이터와 기존에 있는 데이터를 사용하여 모델을 테스트할 것이다.
- 이러한 과정을 통해, 스팸 필터링 모델을 평가하고 필터링 된 항목, 필터링이 된 이유, 필터링 정확도 등을 벤치마크 형식으로 시각화할 것이다.

4. 제약 사항 및 대책

4.1 제약 사항

- 기존의 머신 러닝 기반 필터링 모델들은 학습데이터의 한계로 인해, 새로운 데이터 유형에 대한 필터링이 부정확할 수 있다.
- 이미지 형식으로 수신되는 메일의 경우 스팸 필터링의 정확도가 낮아질 수 있다.
- 이미지 데이터를 텍스트로 변환할 때, 완벽하게 변환되지 않을 수 있다.
- 기존의 chatgpt와 같은 생성 툴들을 이용하여 스팸 메일 데이터를 생성할 때, 해당 툴의 정책에 의해 반려될 수 있다.

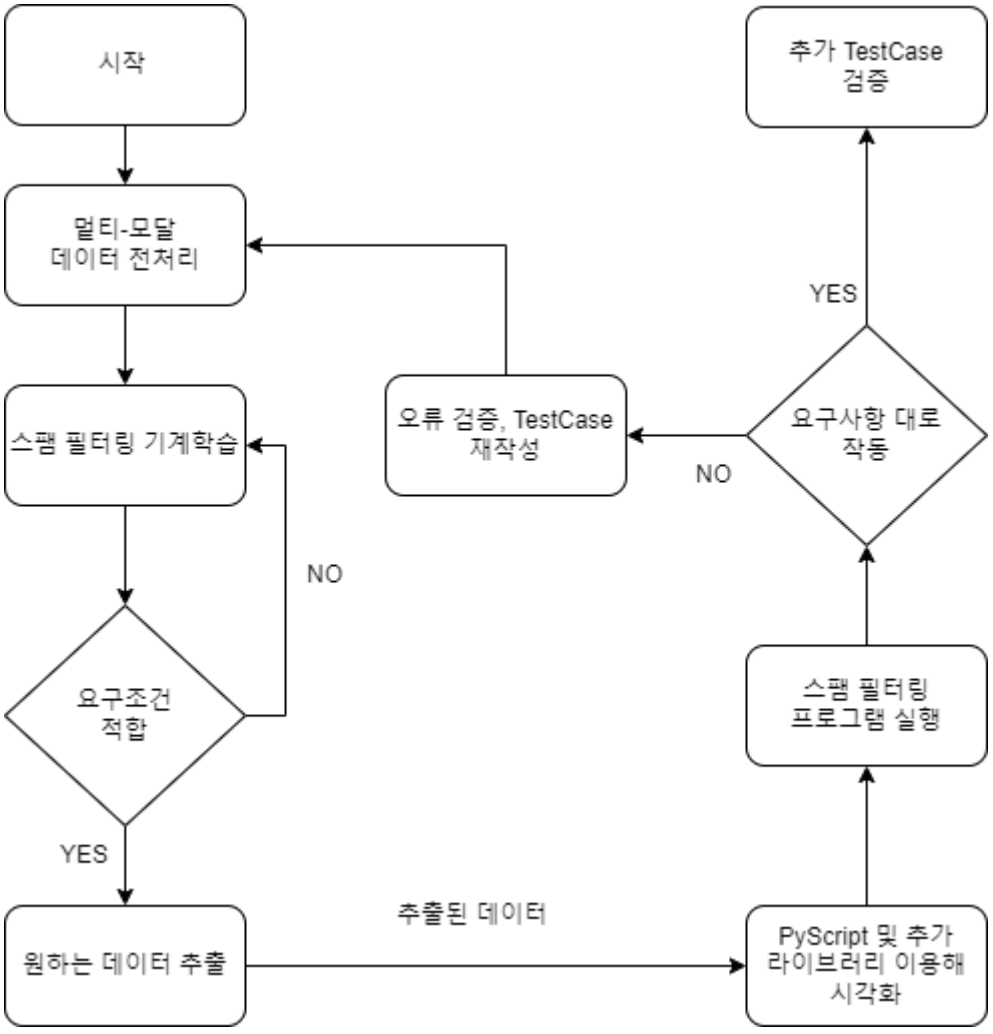
4.2 대책

- 새로운 데이터를 직접 생성하는것을 병행하여 기존의 모델을 학습시킨다.
- 본 과제에서는 이미지 형태의 메일을 텍스트 기반으로 변환하여 진행한다.

- 복잡한 이미지 데이터의 변환이 힘들 경우, 직접 만든 이미지 데이터를 변환해서 테스트를 진행한다.
- 기존 툴을 사용해서 스팸 메일을 생성할 때, 직접적으로 스팸이라는 단어를 사용하지 않고 우회할 수 있는 키워드를 선택하여 생성할 수 있도록 한다.

5. 설계

5.1 시스템 구성도



[그림 **8]** 시스템 구성도

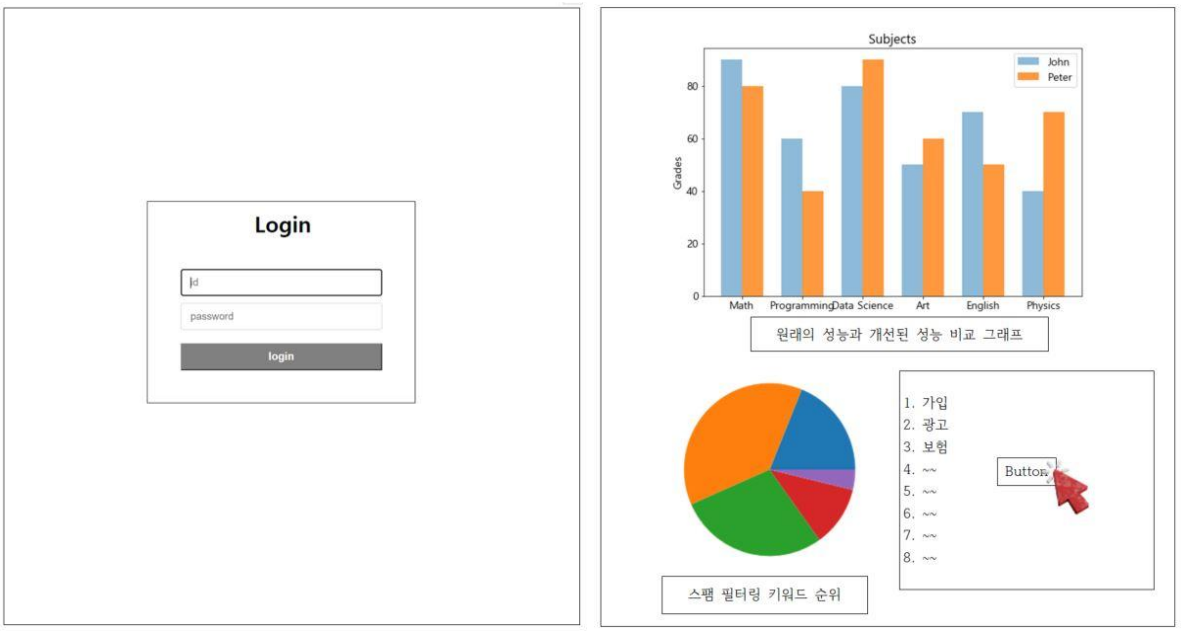
5.2 개발 환경

- **Python GUI**
    - **Microsoft Visual Studio Code 1.77.3**
- **Manual editor**
    - **CHM Editor 3.1.0**
- **Machine-learning**
    - **Anaconda Jupyter notebook 3.9.5**
- **Data-Processing**
    - **Anaconda Jupyter notebook 3.9.5**

5.3 사용 기술

- 머신러닝: sklearn, tensorflow 등
- 데이터 전처리 : Pandas, Numpy, OpenCV, Google Vision API (https://cloud.google.com/vision?hl=ko)
- 데이터 시각화: PyScript, Visual Studio Code, Python Library (Matplotlib), Chm editor

5.4 결과물 예시



[그림 **9]** 졸업과제 결과물 **(**예시**)**

- 아이디와 비밀번호로 로그인 후 멀티모달 스팸 데이터 필터링 실행
- 원래 모델의 성능과 개선된 모델의 성능을 바(bar) 그래프로 시각화
- 3제일 가중치가 높은 키워드들 내림차순으로 상위n개 까지 파이(pie)그래프로 시각화
- 옆의 버튼을 클릭하면 그래프 외 스팸 데이터 스크롤박스 형태로 끝까지 출력
- 모바일 기기로 접속 시 모바일화면에 맞추어 출력

6. 개발 일정 및 역할 분담

6.1 개발 일정

주요일정	6月	7月	8月	9月	10月
각각 파트 개발 착수					
추가 개발 착수 및 모듈 통합					
중간보고서 및 중간평가표 제출					
개발 마무리 및 최종 테스팅					
프로젝트 디버깅					
최종보고서, 최종평가표 제출					
졸업과제 발표 심사					
결과물 업로드					

[표 1] 개발 일정

6.2 역할 분담

윤상호

1. 스팸 필터링 모듈 개발
    - Jupyter notebook(파이썬 기반), 구글 colab 으로 sklearn 또는 tensorflow 등의 라이브러리를 활용하여 개발.
    - LSTM 기반 스팸 필터링 모델 개발 및 학습.
    - 적합한 히든 계층 개수와 입력 노드 개수를 경험적으로 선정.
    - 적합한 입력 feature를 고려.
2. 스팸 필터링 모듈 성능 분석
    - 해당 모델의 성능을 잘 알려진 성능 평가 지표(FP Rate, Accuracy, Precision, ReCall 등)를 이용하여 수치화.
    - 기존의 데이터만으로 측정한 성능과 직접 생성한 데이터를 활용하였을 때 측정한 성능을 비교.
3. 데이터 셋 관리
    - 학습 결과와 입력, 히든 계층의 조절과정을 저장.
    - 학습 결과나 검증 결과는 CSV 파일 형태로 저장.

조재홍

1. 이미지 스팸 데이터 수집 및 변환
    - Kaggle 데이터 저장소와 웹 이미지 크롤링을 통해 이미지 스팸 데이터 수집.
    - 수집된 이미지 데이터를 전처리과정을 통해 텍스트 데이터로 변형.
2. 텍스트 스팸 데이터 생성
    - 생성형 AI 서비스인 Chat GPT, Google Bard, Ms Bing을 활용하여 특정 단어가 일정 횟수 이상 반복되도록 하는 규칙을 통해 스팸 데이터를 생성.

3. 텍스트 스팸 데이터 전처리
    - 수집, 생성 과정을 통해 확보한 데이터를 인공신경망 모델에 적합한 데이터 형태로 변형.

이강우

1. GUI 시각화 도구 개발
    - VisualStudioCode를 이용해 학습이 완료된 데이터 셋을 받아 웹 상에 구현.
2. Python 라이브러리로 시각화
    - Pyscript를 이용해 HTML에 파이썬 언어 활용.
    - Matplotlib 라이브러리를 이용해 그래프로 시각화.
3. 개발과정 기록, 테스트 케이스 작성
    - CHM Editor를 이용하여 개발과정을 기록.
    - 졸업과제가 어떤 기능을 하고 어떻게 동작하는 지 설명.

공통

1. 기존 학습 데이터 수집
    - Kaggle / 네이버, 구글 플랫폼 계정 메일 / 크롤링 등을 활용하여 스팸 메일 데이터 셋 수집.
2. 테스트 및 디버깅
3. 보고서 작성 및 발표
    - 착수보고서, 중간보고서, 최종보고서를 Google Docs를 통해 공동으로 작성.