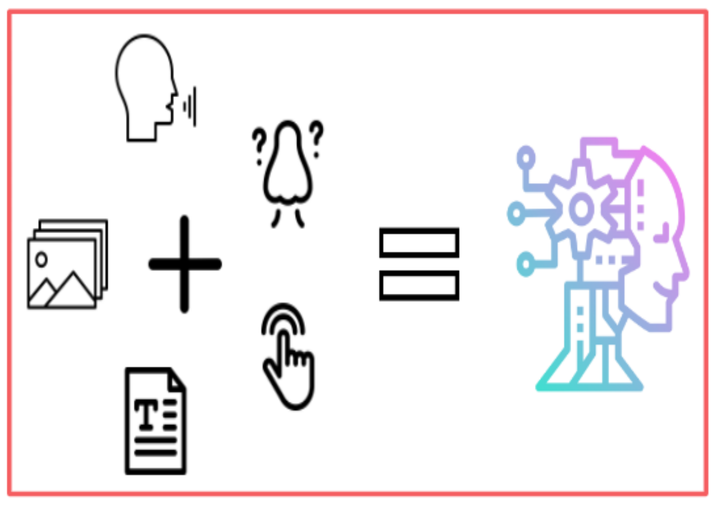
49 멀티모달 기반 스팸 필터링 플랫폼 개발

소속 정보컴퓨터공학부 분과 D 팀명 멀티모발 참여학생 이강우, 윤상호, 조재홍 지도교수 최윤호

과제 소개 및 목표

과제 소개 - 멀티모달이란?

- **모달리티** 라는 것은 데이터의 형태, 유형을 의미한다. 예를 들어, 언어라는 것은 **시각적으로 글**로 표현될 수 있고, **청각적으로 소리**로 표현될 수 있다. 또한 **점자는 촉각의 형태**로 언어를 표현한 것이다. 이처럼 **시각, 촉각, 청각 등을 하나의 모달리티라고 할 수 있다.**



- 앞서 소개한 기사에 따르면 **요즘은 대부분의 텍스트 필터링을 우회하기 위해서 이미지 형식으로 스팸메일을 보내는 경우가 많아지고 있다.**

- 본 졸업과제에서는 멀티모달을 활용하여 **수집한 텍스트 메일**뿐만 아니라, **이미지**와 **Chat-GPT로 생성한 텍스트 메일**을 각각의 모달리티로 보고 이 3가지 모달을 잘 처리할 수 있는 스팸 필터링 모델을 개발하였다.

결론은, 텍스트와 이미지가 섞여 들어와도 잘 처리할 수 있다!

- 최근 텍스트 스팸은 대부분의 필터링 시스템에서 잘 처리되고 있다.

- BUT, 최근에는 텍스트 필터링을 우회하기 위해서 이미지 형식으로 보낸다거나, 이미지와 텍스트를 같이 동봉하여 전송해서 기존의 필터링 시스템을 우회하는 스패머들이 많이 등장하고 있다.


'문자' 걸러내니 '이미지'·'카톡'으로...스팸의 진화

머니투데이 | 이해인 기자

VIEW 6,779 | 2018.07.01 11:00

휴대전화 스팸 중 약 30%가 이미지 형식...필터링 피하기 위한 변종 스팸 활개

신종 스팸 메시지가 새로운 문제로 떠오르고 있다. 문자 메시지형 스팸 메시지에 대해 필터링과 차단을 강화하자 필터망을 피해가기 위해 이미지 스팸, 카카오톡 스팸 등 변종을 생산해 내면서다. 관련 기관들은 이미지 인식을 통한 스팸 필터링 시스템을 개발하는 등 변종 스팸 막기에 나서고 있다.

1일 한국인터넷진흥원(KISA)에 따르면 최근 신고되는 휴대전화 스팸 중 약 30% 가량이 이미지 형식으로 집계됐다. 이미지 형식 스팸이란 글자 없이 이미지만을 첨부한 스팸 메시지다. 특히 불법 도박과 성인 관련 문자 스팸 메시지가 이미지 형식을 많이 취하고 있다.

손등에 기미, 검버섯 '이것'으로 싹~ 진작 할 걸!

月 o만원대 "암보험" 인기! 암진단금 '1억'+생...

회전근개파열 비수술치료—당일회복!

과제 목표

- 멀티모달리티를 활용하여 멀티미디어 형식으로 전송되는 스팸메일 필터링을 목표로 한다.

- 각 모달리티에 대한 피처를 직접 분석 및 선정해 새로운 유형의 스팸메일에 유연하게 대처할 수 있다.

- 필터링 결과를 눈으로 확인할 수 있도록 시각화 인터페이스 (스팸메일 여부, 스팸 분류 이유 등 유의미한 정보)를 구현한다.

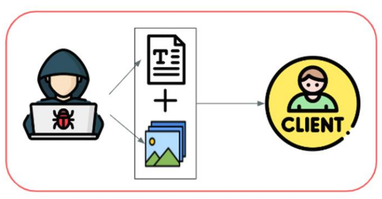
과제 내용

동작 과정

동작 과정 - 이미지 + 텍스트 입력

<주어진 상황>

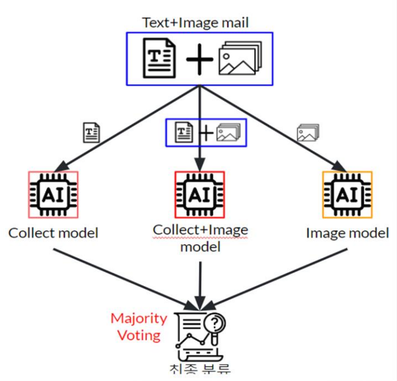
- 어떤 스패머가 악의적인 의도를 가지고 스팸 필터링을 우회하려고 한다.
- 실제로 네이버와 구글에 정상적인 텍스트와 스팸 이미지를 같이 동봉해서 전송했을 경우 자동 필터링이 되지 않는 것을 확인했다.



- 본 졸업과제에서 사용하는 5가지 모델(분류기)

동작 과정 - 이미지 + 텍스트 분류

- 이미지와 텍스트가 동봉되었으므로 서로 분리해주어야 한다.
- **이미지 / 텍스트 / 이미지 + 텍스트**의 3종류로 구분된다.
- **기존의 모델은 이미지+텍스트 유형의 메일을 잘 처리하지 못했는데** 이를 극복하기 위해 멀티모달을 잘 처리하는 분류기를 제작하였다.
- **이미지 / 텍스트**는 **각각 단일 모달리티** 이므로 단일 입력모델로 들어가서 결과가 나온다.
- **이미지 + 텍스트**는 **다중 모달리티**이므로 멀티모달 분류기로 들어가서 결과가 나온다.
- **3개의 결과를 과반수로 종합**하여 최종 결과를 도출해내는 방식이다.

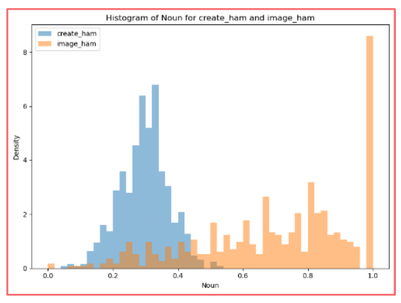


분석 및 결과

분석 - 피처 분석 방법

피처들의 각 모달리티간의 차별점 여부를 시각적으로 확인하기 위해 히스토그램을 그려 분포를 확인

- **각 모달리티의 스팸메일 간의 피처 분포도**를 보면 답이 나오지 않을까? 라는 생각을 했고 실제로 **히스토그램**으로 그려서 살펴보았을 때, 두 개의 그래프가 **겹치는 면적이 50% 미만인 것**을 선택하면 **두 모달리티간의 차이를 명확**하게 표현할 수 있다라고 판단하였다.

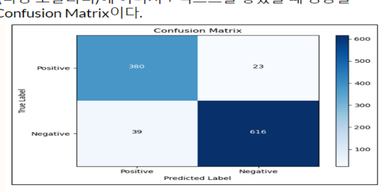


	생성+이미지(spam)	생성+이미지(ham)
special_ratio	0.527458788	0.493304871
number_ratio	**0.320865888**	**0.355738076**
url_count	0.447033034	0.518343747
upper_ratio	**0.1480129**	**0.130628516**
blank_ratio	0.772814351	0.306008086
crlf_ratio	**0.400054735**	**0.284124516**
Noun	0.579741564	0.179713874
Pronoun	**0.316468297**	**0.172116825**
Verb	0.706450695	0.335565043
Adjective	**0.46197787**	**0.388980463**
Adverb	0.644793187	0.305567022
avg_word_sentences	0.494451361	0.549831774
avg_char_sentences	0.471555175	0.523792021
avg_word_paragraphs	0.111520134	0.534824281
avg_char_paragraphs	0.106599339	0.556213068

- 각 feature별 겹치는 비율을 나타낸 표

결과 - 최종 성능(이미지)

아래 그래프는 이미지 단일 모델과, 이미지 다중 모델(다중 모달리티)에 이미지 + 텍스트를 넣었을 때 성능을 나타낸 그래프와 최종 완성된 모델의 성능을 나타낸 Confusion Matrix이다.



- Accuracy (정확도): **0.9414**
- Recall (재현율): 0.9069
- Precision (정밀도): 0.9429
- Specificity (특이도): 0.9640
- F1 Score: 0.9246

WEB, DATABASE

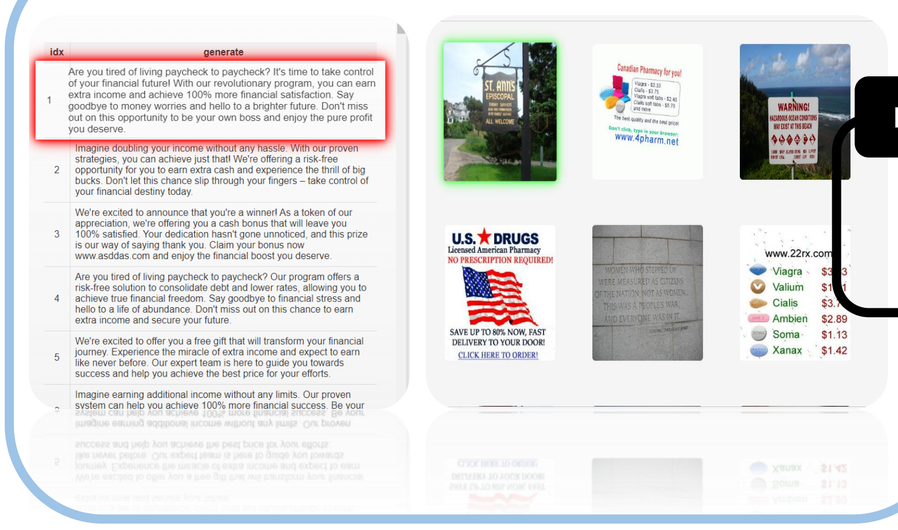
HTML JS CSS

php

phpMyAdmin

과제 결과

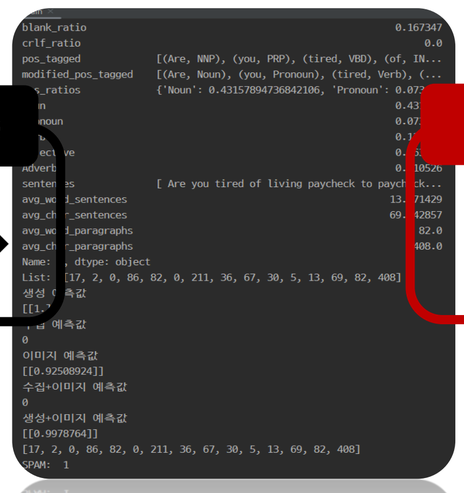
멀티모달 메일



idx	generate
1	Are you tired of living paycheck to paycheck? It's time to take control of your financial future! With our revolutionary program, you can earn extra income and achieve 100% more financial satisfaction. Say goodbye to money worries and hello to a brighter future. Don't miss out on this opportunity to be your own boss and enjoy the pure profit you deserve.
2	Imagine doubling your income without any hassle. With our proven strategies, you can achieve just that! We're offering a risk-free opportunity for you to earn extra cash and experience the thrill of big bucks. Don't let this chance slip through your fingers – take control of your financial destiny today.
3	We're excited to announce that you're a winner! As a token of our appreciation, we're offering you a cash bonus that will leave you 100% satisfied. Your dedication hasn't gone unnoticed, and this prize is our way of saying thank you. Claim your bonus now www.asddas.com and enjoy the financial boost you deserve.
4	Are you tired of living paycheck to paycheck? Our program offers a risk-free solution to consolidate debt and lower rates, allowing you to achieve true financial freedom. Say goodbye to financial stress and hello to a life of abundance. Don't miss out on this chance to earn extra income and secure your future.
5	We're excited to offer you a free gift that will transform your financial journey. Experience the miracle of extra income and expect to earn like never before. Our expert team is here to guide you towards success and help you achieve the best price for your efforts.
6	Imagine earning additional income without any limits. Our proven system can help you achieve 100% more financial success. Be your

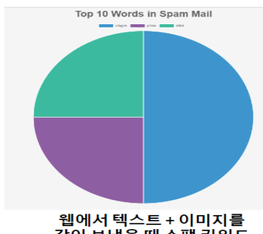
메시지 전송

메시지 분석



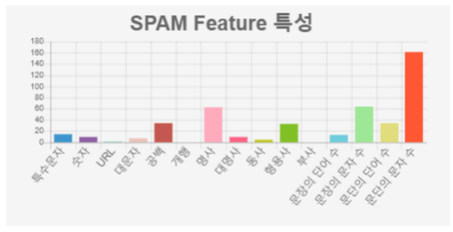
스팸 탐지

분석 결과



웹에서 텍스트 + 이미지를 같이 보냈을 때 스팸 키워드 파이차트

모델 분류 결과에 의해 스팸이라서 사이렌이 울림



입력한 메일의 분석 결과를 본 졸업과제에서 분석한 피처에 빈도수 그래프 나타남

분석결과

부산대학교