멀티 모달 기반 텍스트-이미지 스팸 탐지 모델

박혜경, 우타리예바 아쎔, 신진명, 김형건, 김보람, 정해원, 최윤호

부산대학교

hyegp16@gmail.com, assemakx@pusan.ac.kr, sinryang@icloud.com, qddd2000@gmail.com, csrcerp@pusan.ac.kr, speedhaewon@gmail.com, yhchoi@pusan.ac.kr

Multi-Modal-BasedText-ImageSpamDetectionModel

Park Hye-Gyoung, Utaliyeva Assem, Shin Jin-Myeong, Kim Hyeong-Geon, Kim Bo-Ram, Jeong Hae-Won, Choi Yoon-Ho

Pusan National University.

요 약

본 논문은 기존의 텍스트 위주의 스팸 탐지 필터에 존재하는 텍스트가 아닌 데이터를 통해 스팸 필터를 간단하게 우회할 수 있는 문제점에 대한 방안으로 텍스트뿐만 아니라 이미지 데이터도 고려한 멀티 모달 기반 텍스트-이미지 스팸 탐지 모델을 제안한다. 제안하는 모델은 텍스트 뿐 아니라 이미지에 대한 내용 분석을 종합하여 스팸 탐지를 수행하며, 기존 텍스트 기반 스팸 탐지 모델 방식을 회피하기 위해 시도하는 이미지로 우회하는 스팸 방식에 대해 효과적인 대비책을 제시한다.

Ⅰ. 서 론

스팸은 대량으로 특정되지 않은 다수에게 전송되는 요청하지 않은 메시지(이메일, 문자, 또는 전화)를 의미한다. 인터넷 사용이 증가할수록 스팸은 증가하고 있으며 이로 인한 개인정보 침해 및 사기 문제가 악화되고 있다. 그러나 기존에 존재하는 스팸 필터는 문자 메시지, 이메일과 같은 텍스트 데이터에 집중되어 있기 때문에 이미지 스팸이나 보이스 피싱과 같은 다른 데이터 유형으로 우회하는 경우 스팸이 잘 탐지되지 않는 한계가 존재한다. 따라서, 스팸 메시지를 효율적으로 차단하기 위해서는 텍스트뿐만 아니라 이미지, 음성 등 다양한 데이터 형태의 스팸 메시지를 식별하고 데이터 유형에 적합한 방식을 통한 분석이 요구된다.

이러한 문제를 해결하기 위해, 본 논문에서는 멀티 모달 기반 텍스트-이미지 스팸 탐지 모델을 제안한다. 제안하는 모델은 이미지 데이터의 특성을 스팸 여부 판단에 반영할 수 있도록 고안되었다. 제안하는 모델은 이미지에 대한 캡션 생성 및 이미지 내부 텍스트 추출을 진행하여 이미지에 반영된 정보를 분석하고, 분석한 정보를 이메일의 텍스트 데이터와 결합하여 스팸 필터 모델에 적용함으로써 이미지 데이터를 스팸 판단에 함께 고려한다.

본 논문의 구성은 다음과 같다. 2장에서 실험에 활용된 데이터의 특징 및 모델을 설명한 뒤 3장에서 실험을 통해 제안한 모델의 우수성을 기술한다. 이후 4장에서 논문의 결론을 맺는다.

Ⅱ. 본 론

본 장에서는 학습에 사용한 텍스트 데이터 및 이미지 데이터, 전처리 기법과 실험에 활용한 모델들에 대해 기술한다.

2.1 텍스트 데이터 셋

본 논문에서는 기존 스팸 탐지 연구에서 보편적으로 활용된 Enron Spam Dataset, SMS SpamCollectiont, SpamAssasin Dataset을 활용하여 총 39,610개의 텍스트 데이터를 학습에 사용하였다.

Enron Spam Dataset은 V. Metsis, I. Androutsopoulos 및 G. Paliouras가 수집한 데이터로 30,494개의 이메일로 구성되어 있다. SMS SpamCollection은 휴대 전화 스팸 조사를 위해 수집된 SMS 레이블 메시지의 공개 데이터셋으로 5574개의 SMS 메시지로 구성되어 있다. SpamAssassin Dataset은 Apache SpamAssassin Project에서 공개 제공한 데이터셋으로 6047개의 이메일로 구성되어 있다.

2.2 이미지 데이터 셋

이미지 데이터의 경우 스팸 이미지의 특징을 학습하는 방식 대신 이미지에 대한 의미적 분석을 캡션으로 생성할 수 있도록 학습시키는 것을 목표로 하였다. 이미지의 의미적 캡션 생성을 위한 모델의 학습에는 Flickr 8K Dataset을 사용하였다. Flickr 8K Dataset은 미국의 기업 야후의 온라인 사진 공유 커뮤니티 사이트인 Flickr에서 수집된 데이터로, 5개의 서로 다른 캡션(이미지에 대한 설명)과 쌍을 이루는 8,000개의 이미지로 구성되어 있다.

2.3 전처리

이미지 데이터의 경우 이미지 캡션[1]과 이미지에서 추출한 텍스트를 결합하여 이미지를 분석 결과를 텍스트 형식으로 변환한다. 텍스트 데이터는 단어 또는 문장으로 나누는 토큰화를 진행한 뒤 의미가 없는 단어나 너무 빈번하게 나타나는 단어(예:"the," "and," "is")를 제거한다. 이후 텍스트를 동일한 기본 형태인 어근으로 변환, 특수 문자 및 노이즈 제거, 대소문자 통일, 줄임말 확장, 동의어 통합과 같은 텍스트 정규화 작업을 수행하여 일관성 있는 데이터로 가공한다. 가공된 데이터를 본 논문에서는 CountVectorizer를 이용해 벡터화를 수행하였다.

전처리를 마친 데이터는 멀티 모달 기반 텍스트-이미지 스팸 탐지 모델에 적용할 수 있는 Intermediate Representation이 된다.

2.4 모델

2.4.1 이미지 캡션 생성 모델

VGG16 모델[2]을 사용하여 이미지의 특징을 추출하고 해당 데이터를

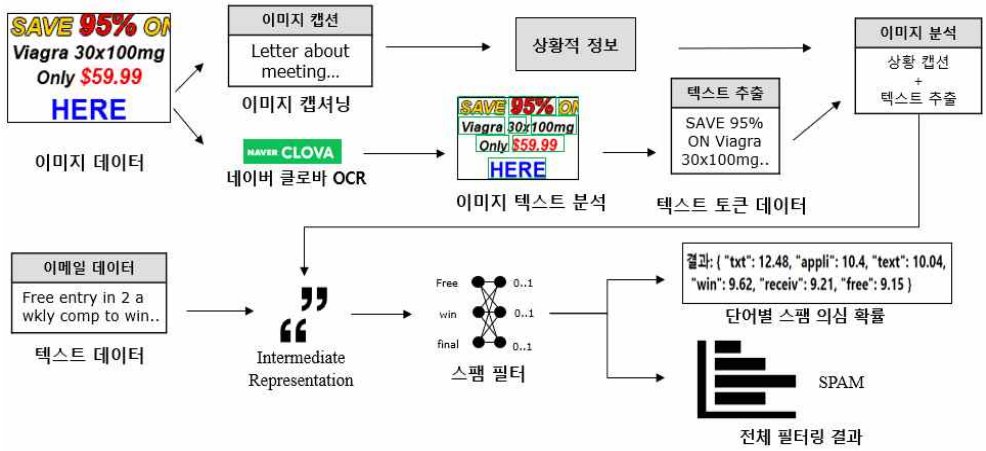

그림 1 멀티 모달 기반 텍스트-이미지 스팸 탐지 모델 구조도

인코더에 입력으로 사용하였다. 디코더로는 LSTM을 사용하여 이미지의 특징을 바탕으로 캡션을 생성하였다.

2.4.2 멀티 모달 기반 텍스트-이미지 스팸 탐지 모델

본 논문에서는 스팸 메일을 판단할 때 텍스트와 이미지를 함께 고려하는 모델을 제안한다. 해당 방식을 사용하면 기존의 텍스트 필터를 이미지를 통해 우회하는 방식에 대해서도 방어할 수 있으며 텍스트와 이미지로 분산되었던 정보를 함께 고려함으로써 스팸 판단 여부에 대한 보다 설명 가능한 결과를 도출해낼 수 있다.

제안하는 모델은 텍스트와 이미지로 구성된 이메일에서, 이미지 캡션을 통한 이미지 내용 정보와 네이버 클로바 OCR을 활용해 추출한 이미지 내부 텍스트를 결합하여 이미지 분석을 진행한다. 이후 이메일에 있는 텍스트와 이미지의 분석 결과를 결합하여 Intermediate Representation으로 형식을 변환하고 스팸 필터를 통해 스팸 여부 판단을 진행한다. 스팸 필터로는 선형 및 비선형 데이터에 대한 효과적인 분류를 제공하며 고차원 데이터에 유연하게 작동하는 SVM(Support Vector Machine)[3]을 사용하였다. 스팸 판단 결과는 XAI를 활용하여 설명 가능한 결과를 함께 도출한다.

Ⅲ. 실험 결과

그림 2 스팸 이미지

스팸이 아닌 이메일

I've updated the project documents. Your feedback is crucial—please review at your earliest convenience. Let me know if any further adjustments are needed.

그림 3 스팸이 아닌 이메일

실험은 Window 11, 16GB RAM 환경, NVIDIA GeForce 3060 GPU 환경에서 진행되었으며 실험을 위한 코드는 Python v3.11.4, tensorflow v2.10.x을 활용하여 구현되었다. 스팸 텍스트 데이터의 학습 데이터와 실험 데이터는 8:2 비율로 구성되어있으며 각 데이터는 동일한 전처리를 진행하였다.

기존의 텍스트 데이터로만 판단하는 스팸 필터의 경우 그림 2(Dredzed의 논문[4]에서 공개된 이미지 스팸 데이터)와 그림 3이 같은 이메일에 존재하는 경우 스팸이 아닌 확률 : 스팸일 확률을 0.967: 0.032로 판단하여 스팸이 아닌 메일로 분류한다. 그러나 본 논문이 제안하는 모델의 경우 그림 2와 그림3을 함께 분석하여 스팸 판단에 활용하므로 스팸이 아닌 확률 : 스팸일 확률을 0.149 : 0.850로 스팸으로 판단하였다.

Ⅳ. 결론

본 논문에서는 텍스트와 이미지를 모두 고려하는 멀티 모달 기반 텍스트-이미지 스팸 탐지 모델을 제안한다. 기존의 텍스트 데이터만을 분석하는 방식의 경우 텍스트가 아닌 데이터를 이메일에 추가하여 텍스트 필터를 우회하는 방식으로 스팸 필터를 우회할 수 있었다. 본 논문에서 제안하는 모델은 이메일의 텍스트와 이미지를 함께 분석하며, 이미지 자체의 캡션을 통해 이미지 내용 분석을 진행하여 기존에 고려되지 않았던 이미지에 대한 정보를 스팸 필터에 활용해 스팸 필터의 활용성을 높였다. 또한 해당 분석은 설명 가능한 결과와 결합하여 제공할 수 있어 향후 시각적 활용 측면에 기여할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. RS-2023-00217689) 및 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원(2022R1I1A3055233)을 받아 수행되었음.

참 고 문 헌

[1]Borneel Bikash Phukan, Amiya Ranjan Panda, "An Efficient Technique for Image Captioning using Deep Neural Network". Under review by an internationally recognized Scopus indexed journal 2020

[2]Deep Learning Bible 2. Classification - 한글[도서] : Understanding Vgg-16 Vgg-19, Available : https://wikidocs.net/164796

[3]민도식, 송무희, 손기준, 이상조, "SVM 분류 알고리즘을 이용한 스팸메일 필터링(SPam-mail Filtering Using SVM Classifier)", 한국정보과학회 03 봄 학술발표논문집(B) 2003 Apr.,2003년,pp.552-554

[4]Mark Dredze, Reuven Gevaryahu, Ari Elias-Bachrach. Learning Fast Classifiers for Image Spam. In proceedings of the Conference on Email and Anti-Spam (CEAS), 2007

[5]Yidong Chai, Yonghang Zhou, Weifeng Li, "An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence", IEEE 2022