딥러닝을 이용한
비정상 문자 조합으로 구성된 스팸 문자 탐지 기법

김가현 [1], 유헌창 [2]
[1] 고려대학교 컴퓨터정보통신대학원 석사과정
[2] 고려대학교 정보대학 컴퓨터학과 교수

kimgahyun@korea.ac.kr, yuhc@korea.ac.kr

A Technique to Detect Spam SMS with Composed of Abnormal Character Composition Using Deep Learning

Ka-Hyeon Kim[1], Heonchang Yu[2]
[1]Dept. of Software Security, Graduate School of Computer & Information Tech., Korea University
[2]Dept. of Computer Science and Engineering, Korea University

요 약

대량 문자서비스를 통한 스팸 문자가 계속 증가하면서 이로 인해 도박, 불법대출 등의 광고성 스팸 문자에 의한 피해가 지속되고 있다. 이러한 문제점을 해결하기 위해 다양한 방법들이 연구되어 왔지만 기존의 방법들은 주로 사전 정의된 키워드나 자주 나오는 단어의 출현 빈도수를 기반으로 스팸 문자를 검출한다. 이는 광고성 문자들이 시스템에서 자동으로 필터링 되는 것을 회피하기 위해 비정상 문자를 조합하여 스팸 문자의 주요 키워드를 의도적으로 변형해 표현하는 경우에는 탐지가 어렵다는 한계가 있다. 따라서, 본 논문에서는 이러한 문제점을 해결하기 위해 딥러닝 기반 객체 탐지 및 OCR 기술을 활용하여 스팸 문자에 사용된 변형된 문자열을 정상 문자열로 복원하고, 변환된 정상 문자열을 문장 수준 이해를 기반으로 하는 자연어 처리 모델을 이용해 스팸 문자 콘텐츠를 분류하는 방법을 제안한다. 그리고 기존 스팸 필터링 시스템에 가장 많이 사용되는 키워드 기반 필터링, 나이브 베이즈를 적용한 방식과의 비교를 통해 성능 향상이 이루어짐을 확인하였다.

1. 서론

대량 문자서비스를 통한 스팸 문자의 증가[1]로 인해 사용자들은 도박, 불법대출 등의 광고성 스팸 문자로부터 지속적인 피해를 입고 있다. 스팸 문자로 인한 문제는 사용자의 편의성과 개인 정보 보호에 대한 중요한 고려사항이 되었으며, 이에 효과적인 대응 방식이 필요하다.

기존 스팸 문자 필터링 시스템은 주로 기존 신고된 스팸 발신 번호를 차단하는 방식으로, 발신 번호 차단 방식은 전화번호 변작기를 통해 매번 다르게 조작된 발신번호로 문자가 보내지는 최근의 방식[2]에 적합하지 않다. 이에 상보적인 기법으로 문자 메시지의 내용을 분석하여 스팸 메시지의 여부를 확인하는 콘텐츠 분석 기법이 있다. 그러나 지금의 콘텐츠 분석 기법은 문자 메시지 내 사전 정의된 키워드를 찾는 수준에 지나지 않아 아직 그 성능이 미흡하다. 이러한 한계점을 이용해 광고 문자들은 주요 키워드를 특

수문자 등으로 변형 및 왜곡하여 필터링을 회피하고 있다.

그 외에도 일반적으로 스팸 필터링에 사용되는 머신러닝 알고리즘으로 나이브 베이즈(Naive Bayes) 모델이 있다. 하지만 자주 나타나는 단어의 출현 빈도수를 기반으로 스팸 콘텐츠를 분류하기 때문에 변형된 문자열에 대해서는 분류 성능이 좋지 않다. 따라서 더욱 정확성이 높고, 폭넓은 수준으로 콘텐츠를 분석할 수 있는 기법에 대한 필요성이 대두되고 있다.

본 연구에서는 딥러닝 기반의 이미지 처리와 자연어 처리 기법을 적용하여 스팸 콘텐츠가 포함된 문자 메시지를 탐지하는 기법을 제안한다. 제안하는 기법의 평가를 위해 실제 스팸 문자 데이터를 이용하여 기존 스팸 문자 필터링 기법들과의 비교 실험을 수행하였고, 성능 향상을 확인하였다.

2. 관련 연구

2.1 키워드 기반 문자 콘텐츠 필터링

가장 일반적인 방법으로 '카지노', 'Casino', '대리운전' 등의 스팸 문자에 많이 등장하는 키워드를 지정하여 필터링한다. 하지만 의미가 유사하더라도 정의되지 않은 새로운 단어는 탐지하지 못하며, 지정한 철자 그대로 사용된 단어만 필터링이 가능하다는 한계점이 있다. 광고성 문자들은 이와 같은 점을 이용하여 특수문자 등을 사용해 주요 키워드를 의도적으로 변형 및 왜곡시켜 자동 스팸 필터링을 회피한다. 예를 들면 'Casino'를 'ℂÅsIΠΟ', '€A $ IΠO' 로 변형하는 등 비정상 문자 조합으로, 시스템은 탐지하지 못하지만 사람은 알아볼 수 있는 형태의 단어를 만들어 낸다. 하지만 일일이 변형된 형태를 포함하여 모든 키워드를 필터링 지정하는 것은 불가능하며, 또한 이러한 방식은 전체 문자 내용과 상관없이 지정된 키워드가 포함되는 경우 일반 문자임에도 불구하고 무조건적으로 스팸 문자로 분류되는 false-positive 오류(false alarm) 문제가 있다.

2.2 나이브 베이즈 분류기

나이브 베이즈 분류기는 조건부 확률을 계산하는 베이즈 정리를 이용하여 스팸 문자를 분류하는 머신러닝 기반 방법이다. 콘텐츠에 포함된 단어들을 각각 하나의 특성(feature)로 간주하여 각 단어가 스팸 문자에서 나타날 확률과 정상 문자에서 나타날 확률을 각각 계산하여 해당 문자 메시지가 스팸 문자일 확률을 계산한다.

이 기법은 빠르고 쉽게 구현이 가능하며 확장성이 좋다고 알려져 전통적으로 많은 스팸 분류 시스템에 사용되고 있다[3]. 하지만 단어들의 출현 빈도수에 따라 확률이 결정되기 때문에 광고 스팸문자들과 같이 의도적으로 단어를 변형해서 시스템이 탐지하지 못하게 하는 경우에는 적합하지 않다.

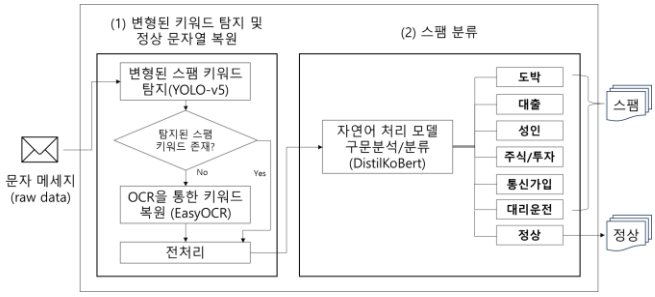
3. 딥러닝 기반의 변형된 스팸 키워드가 포함된 스팸 문자 필터링 시스템

딥러닝 기반의 객체 탐지 모델 및 OCR 기술로써 비정상 문자 조합으로 구성된 변형된 키워드를 정상 문자열로 복원하고, 복원된 문자열을 문장 수준 이해를 기반으로 하는 자연어 처리 모델을 적용하여 스팸 콘텐츠를 분류한다.

객체 탐지 모델은 YOLO-v5[4], OCR 기술로는 EasyOCR[5] 모듈을 이용했으며, 언어 모델은 DistilKoBERT[6]로, 경량화 된 한국어 BERT 모델을 사용하였다.

3.1 전체 설계

본 논문에서 제안하는 딥러닝 기반의 스팸 문자 필터링 시스템은 (그림 1)과 같이 설계되었다. 주요 단계는 크게 두 가지로, (1)변형된 키워드 탐지 및 정상 문자열로의 복원과 복원된 문자열을 (2)스팸/정상 분류로 이루어진다.



(그림 1) 딥러닝 기반 스팸 필터링 시스템 구성도

클래스명	개수(개)
정상	5138

(1) 변형된 키워드 탐지 및 정상문자열로의 복원

문자 메시지 데이터를 이미지로 인식하여 변형된 키워드를 탐지하고, 해당 이미지를 텍스트로 변환하는 OCR 기술로써 변형된 문자열을 정상 문자열로 복원시키는 방법을 사용했다. 그리고 특수문자 제거, 외국어 단어를 한글로 변환, 자동 띄어쓰기 모듈을 사용하여 전처리함으로써 복원율을 높였다.

(2) 스팸 분류

사전 학습된 자연어 모델인 Bert 기반의 모델을 fine-tuning 하여 스팸을 분류했다. 스팸 문자의 경우 내용 유형이 다양하여 각 클래스가 그 특성을 더 잘 이해할 수 있도록 스팸 문자의 카테고리를 나누어 다중분류로 구현했다. 스팸문자는 한국인터넷진흥원의 분류 기준을 참고하여 '주식/투자, 도박, 성인, 대출, 통신가입, 대리운전'으로 나누었고, '정상'까지 7 개 클래스로 정의되었다. 이 중 결과가 '정상' 클래스로 분류되는 경우를 정상, 그 외 스팸 유형 클래스로 분류되는 경우를 모두 스팸으로 분류한다.

3.2 데이터셋

스팸 문자 데이터는 스마트 치안 빅데이터 플랫폼에 공개된 스팸문자 데이터[7]와 한국인터넷진흥원에서 제공하는 '휴대전화 간편내역 수집 내역'[8]을 샘플링하여 사용하였다.

정상 문자 데이터는 개인의 사생활 정보가 포함될 수 있어 공개된 데이터가 없다. 따라서 본인의 휴대폰 문자 메시지 데이터를 직접 수집, 샘플링하여 사용하였다.

본 연구에서 학습과 검증에 사용된 데이터의 개수는 클래스별로 <표 1>과 같다.

<표 1> 스팸문자와 정상 무자 분류 클래스별 사용 데이터 개수

클래스명	개수(개)
주식/투자	4240
도박	2504
성인	581
대출	445
대리운전	356
통신가입	278
합계	8404

3.3 변형된 문자열 복원

변형된 문자열의 유형은 다양하며, 각각에 대응할 수 있는 탐지와 복원 방식이 다르다.

(1) 특수문자로 이루어져 텍스트로는 검출되지 않지만 사람이 인식 가능한 경우

1) 변형 문자열 예시(Casino)

ℂÅsIΠƗO, ℃CASINO, ℃CASlN★, €A $ IΠO, ℂasIJƗO, ¢ A $ INO,

2) 탐지 및 복원 방식

문자열을 이미지로 간주하여, 해당 키워드를 하나의 객체로 학습하고, 검출한다. OCR 기술을 통해 정상 문자열로 복원할 수도 있다.

(2) 세로형으로 작성된 경우

1) 변형 문자열 예시

(그림 2) 세로형 변형 문자열 포함 문자 예시

(그림 2)는 사람이 봤을 때엔 구분이 가능하지만,

시스템에서 순차적으로 텍스트를 읽어 들였을 때엔 키워드 탐지가 불가능하다.

2) 탐지 및 복원 방식

문자열을 이미지로 간주하여, 해당 키워드를 하나의 객체로서 학습하고, 검출한다. 그리고 해당 객체에 대한 텍스트 정보를 데이터에 추가한다.

(3) 문자열 사이 특수문자를 넣은 경우

1) 변형 문자열 예시

카.지~노, 바@카!라, 슬_롯, 대 리운 전

2) 탐지 및 복원 방식

특수문자를 제거하여 복원한다.

3.4 YOLO 를 이용한 변형된 키워드 탐지 및 정상 문자열 복원

YOLO 는 객체 탐지와 인식 분야에서 많이 사용되는 모델 중 하나이다. 본 연구에서는 컴퓨터 텍스트로는 검출되지 않지만 사람에게는 이미지 텍스트로 인식되는 비정상 문자 조합으로 구성된 스팸 키워드를 검출하기 위해 YOLO-v5 모델을 fine-tuning 하여 사용했다. 문자 메시지에서 스팸 키워드를 이미지로 탐지하도록 하는 방식이다. 스팸 키워드가 라벨링된 이미지를 학습했으며, 하이퍼파라미터를 조정했다. YOLO 는 모델 탐지의 정확도가 높고, 연산속도가 빠르다는 장점이 있어 실시간성이 중요한 스팸 문자 필터링 시스템에도 적합하다.

학습에 사용된 스팸 키워드는 주로 변형되는 키워드인 '카지노', '슬롯', '바카라', '토토'에 대해 각각 한글형(가로형, 세로형)/영어형(가로형, 세로형) 형태로 16 개 클래스를 정의했다. 이미지를 탐지하고, 인식하는 모델의 특성상, 폰트 종류에 따라 같은 문자열을 사용하더라도 다른 이미지로 받아들이고, 그에 따라 성능이 다를 수 있어 모든 문자 데이터에 동일한 폰트를 적용한 후 이미지로 변환하는 전처리 과정을 거쳐 학습 데이터를 만들었다. 검증 데이터 또한 같은 방식을 사용하였다.

(그림 3) 실제 학습 및 탐지된 문자 데이터 예시

<표 2> Yolo-v5 를 fine-tuning 한 변형된 스팸 키워드 탐지 모델 성능

Precision	Recall	mAP50	mAP50-95
0.92	0.971	0.959	0.722

학습된 모델의 성능은 Precision, Recall 이 모두 0.9 이상으로 좋게 평가되었으며<표 2>, 테스트된 예시는 (그림 3)과 같다. 특히 '세로형 변형 문자열'의 경우 해당 과정에서만 탐지가 가능하기 때문에 효과적이다.

탐지 후 탐지된 영역의 위치를 얻어와 해당 부분을 정상 문구열로 대체하여 복원한다.

3.5 EasyOCR 을 이용한 변형된 문자열 정상 문자로 복원

OCR 기술을 이용하여 스팸 키워드를 이미지로 인식하고, 텍스트로 변환하는 기법이다. 딥러닝 기반의 OCR 모델인 EasyOCR 을 이용하였다. 3.4 의 이미지 기반 객체 탐지 모델에 비해 비교적 성능이 떨어지는 경향이 있기 때문에 3.4 의 과정을 보완하는 용도로, 3.4 에서 탐지된 키워드가 없는 경우에만 적용하여 시스템의 연산량을 줄였다.

3.6 문자열 전처리

(1) 외국어 단어를 한글로 변환

외국어 단어를 한글로 자동으로 변환해주는 모듈[9]을 사용하여 외국어로 표기된 부분을 한글로 변환하여 정상 문자열과 가장 유사한 형태로 복원한다.

- (변환 전):카.Zl. NO → (변환 후):카.즈ㅣ.노

(2) 특수문자 제거, URL 마스킹

분류 성능에 영향을 미치지 않도록 정규화 식을 사

용하여 특수문자는 제거하고, URL 은 <URL>의 형태로 변환한다.

4. 자연어 처리 기반 스팸 분류 모델

3 장에서 복원된 정상 문자열을 단어가 아닌 문장 수준 이해를 기반으로 하는 자연어 처리 모델을 이용해 최종적으로 스팸 문자 콘텐츠를 분류한다. 본 연구에서는 한국어 이해에 대표적으로 사용되는 KoBERT 를 경량화한 모델인 DistilKoBERT 를 이용하여 연산량이 작은 분류 모델을 설계함으로써 전체 연산 오버헤드를 줄이고자 하였다.

또한 기존 연구들에서 일반적으로 '스팸/정상'으로 이진 분류를 하는 것과 달리 스팸 메시지를 내용 유형별로 나누어 다중 분류를 함으로써 언어 모델이 좀 더 쉽게 각 유형을 구성하는 문장들의 특성을 이해할 수 있도록 설계했다. 이진 분류에 비해 비교적 적은 데이터로도 각 유형의 특성을 효과적으로 학습할 수 있다. 분류 기준은 **3.1** 의 **(2)**에서 확인할 수 있다.

5. 성능평가

동일한 학습데이터와 검증데이터를 사용해 기존 스팸 콘텐츠 분류에 사용되는 방식 두 가지와 비정상 문자열의 정상 복원 기법 없이 자연어 처리만을 통한 분류 방식, 그리고 본 연구에서 제안하는 기법을 모두 각각 적용하여 성능을 비교했다.

키워드 기반 필터링 방식의 경우, 각 유형별로 자주 나오는 명사 키워드들을 TF-IDF 방식으로 뽑아내 적용했다.

<표 3>에서 본 연구에서 제안한 방식인 ④가 기존 방식인 ①, ②와 비교하여 높은 성능을 보여주는 것을 확인할 수 있다. 또한 ③과도 성능 차이가 존재하는 것으로 보아 '비정상 문자열을 포함한 스팸 콘텐츠를 정상 문자열로 복원' 하는 과정이 스팸 필터링의 효과를 높이는 데에 크게 영향을 주는 것을 알 수 있다.

<표 3> 성능 평가

방식	Precision	Recall	F1-score
① 키워드 기반 필터링	0.80	0.34	0.48
② 나이브 베이즈 분류기	0.74	0.68	0.71
③ 비정상 스팸 키워드 복원 없이 자연어 처리 분류	0.83	0.76	0.79
④ 비정상 스팸 키워드 정상 복원 + 자연어처리	0.91	0.95	0.93

6. 결론

스팸 콘텐츠는 시스템에 필터링 되지 않기 위해 계속해서 진화하며, 비정상 스팸 키워드의 변형 형태도 다양해지고 있다. 본 연구는 이와 같은 상황에서도 스팸 탐지 분류의 정확성을 높여 보다 효과적인 프로세스를 구축하는 데에 기여할 수 있을 것이다. 또한, 제안 기법은 다른 콘텐츠 이상 탐지 관련 방안으로도 확장하여 사용할 수 있을 것으로 기대된다.

참고문헌

[1] 방송통신위원회, 한국인터넷진흥원, "2022 년 하반기 스팸 유통 현황", 2023.3.

[2] 오형주기자, "070→010 바꿔치기로 피싱 시도, 이미 작년의 두 배 넘었다", 한국경제, 2022.10.01., https://www.hankyung.com/article/202210019497i.

[3] Vangelis Metsis, "Spam Filtering with Naive Bayes-Which Naive Bayes?," CEAS, 2006.

[4] Glenn Jocher, ultralytics/ YOLOv5, 2020.10., GitHub, https://github.com/ultralytics/yolov5.

[5] A. Randika, N. Ray, X. Xiao, & A. Latimer, "Unknown-box Approximation to Improve Optical Character Recognition Performance", arXiv preprint arXiv:2105.07983. (2021). DOI: https://doi.org/10.48550/arXiv.2105.07983.

[6] Park, Jangwon, DistilKoBERT: Distillation of KoBERT, 2019, GitHub, https://github.com/monologg/DistilKoBERT.

[7] 스마트치안 빅데이터 플랫폼, "스팸문자", https://www.bigdata-policing.kr/product/view?product_id=PRDT_395.

[8] 통신 빅데이터 플랫폼, "휴대전화 간편신고 수집내역", https://www.bigdata-telecom.kr/invoke/SOKBP2603/?goodsCode=KIS00000000000000023.

[9] sublee, Hangulize-외래어 자동 한글 변환 모듈,

Github, https://github.com/sublee/hangulize.