딥러닝 기반 스마트폰 피싱 공격 대응 방법

이재경[O], 서진범[*], 조영복(교신저자)[*]
[O]대전대학교 정보보안학과,
[*]대전대학교 정보보안학과
e-mail: forever408916@gmail.com[O], ybcho@dju.ac.kr[*]

A Deep Learning-Based Smartphone Phishing Attacks Countermeasures

Jae-Kyung Lee[O], Jin-Beom Seo[*], Young-Bok Cho(Corresponding Author)[*]
[O]Dept. of Information Security, Daejeon University,
[*]Dept. of Information Security, Daejeon University

● 요 약 ●

스마트폰 사용자가 늘어남에 따라 갖춰줘야 할 보안성이 취약하여, 다양한 바이러스 및 악성코드 위험에 노출되어 있다. 안드로이드는 운영체제 중 가장 많이 사용되는 운영체제로, 개방성이 높으며 수많은 악성 앱 및 바이러스가 마켓에 존재하여 위험에 쉽게 노출된다. 2년 넘게 이어진 코로나 바이러스(Covid-19)으로 인해 꾸준히 위험도가 높아진 피싱공격(Phshing attack)은 현재 최고의 스마트폰 보안 위협 Top10에 위치한다. 본 논문에서는 딥러닝 기반 자연어처리 기술을 통해 피싱 공격 대응 방법 제안 및 실험 결과를 도출하고, 또한 향후 제안 방법을 보완하여 피싱 공격 및 다양한 모바일 보안 위협에 대응할 수 있는 앱을 설계할 것이다.

키워드: 피싱 공격(Phishing attack), KoNLPy(Korean NLP in Python), 딥러닝(Deep Learning), 자연어 처리(Natural Language Processing), NLTK(Natural Laguage Toolkit)

I. Introduction

4차 산업혁명이 발전됨에 따라, 일상생활에서 사람들의 편의성이 점차 증대되고 있다. 그에 따른 금융 서비스 및 회사 업무 같은 일상생활에서 스마트폰이 널리 사용되고 있으며 그에 따른 보안 위협도 증가하고 있다. 특히 운영체제 중 안드로이드 운영체제를 대상으로 하는 악성코드가 꾸준하게 증가하고 있다[1]. 안드로이드는 운영체제 중 가장 높은 점유율을 차지하고 있으며, 또한 안드로이드 앱 마켓의 개방성이 높아서 악성 앱 배포가 쉽다. 2021년 기준 모바일 보안 위협 Top 10이 공개되었고, 그 대상으로 취약한 암호 보안, 악성 앱, 데이터 유출, 피싱 공격 등이 있다. 피싱 공격은 다른 보안 위협과 달리, 불과 몇 년 전부터 가파른 상승세로 보안 위협에 지목되었다[2]. 본 논문에서는 딥러닝 기반 자연어 처리 기술을 이용하여, 형태소 분류를 통한 특정 단어를 찾아내는 기술을 통한 모바일 피싱 공격 대응 방안을 제안한다.

II. Related works

2.1 모바일 보안 위험 분석

본 연구에서는 악성 앱 탐지 방법을 고안하기 위해 정적 분석, API 호출 정보 분석, 유사도 분석을 통해 방법을 제안하였다. 정상 앱 6320개, 악성 앱 6320개의 데이터 셋을 통해 탐지 성능 평가 및 실행 시간 성능 평가를 한 결과, 성능 평가는 약 98%, 실행 시간 성능 평가는 약 98.05%의 정확도를 보였다[3].

2.2 자연어 처리

NLTK(Natural Language Toolkit)는 자연어처리 토크나이징 라이브러리 중 가장 기본적이고 먼저 만들어진 토크나이징 라이브러리로, 자연어 처리를 위한 파이썬 패키지이다. NLTK는 Classfication Toeknization, Stemming, Pos tagging, parsing 기능을 제공한다[5]. KoNLPy(Korean NLP in Python)는 파이썬을 기반으로 한국어 대상 자연어 정보처리가 가능한 패키지를 말한다. KoNLPy의 형태소 분석기는 총 5가지로, kkma, Komoran, Mecab, Okt, Hannanum 형태소 분석기로 이루어져 있다.

III. The Proposed Scheme

3.1 데이터 셋

본 논문의 데이터 셋은 국내를 기준으로, 피싱 메일에서 가장 빈도수가 높은 단어를 활용하였다. 데이터셋은 “대출”, “주식” “광고” “결제”가 포함된 가짜 스팸 메일로 만들어 활용하였으며, 약 50만 글자 정도를 활용하여 속도 측정에 용이하도록 하였다.

3.2 제안 방법

자연어 처리 모델은 국내 피싱 메일을 기준으로, 한국어 처리에 특화되어 있는 KoNLPy 토크나이징 라이브러리를 사용하였다. KoNLPy의 형태소 분석기 5가지 중 kkma, twitter(Okt), komoran 형태소 분석기를 대상으로 각각 형태소 분석을 하였다. 형태소 분석 결과, 각각 명사, 형용사, 조사, 부사 등 각각 나누어서 분류되어 있는 것을 확인할 수 있으며, 또한 Okt 형태소 분석기, kkma 형태소 분석기, Komoran 형태소 분석기는 각각 형태소를 분류하는 기준이 다른 점도 확인할 수 있다.

3.3 속도 측정

속도 측정은 형태소 분석기(kkma, Okt, Komoran)을 대상으로 각각 약 50만개로 이루어진 데이터셋을 활용해 실행 속도를 측정하였다. 형태소 분석기를 각각 속도 측정한 결과, kkma 형태소 분석기는 0.019 sec Okt 형태소 분석기는 0.006 sec, Komoran 형태소 분석기는 0.008 sec의 결과가 나왔다.

IV. Conclusions

본 논문에서는 딥러닝 기반 자연어 처리를 활용한 피싱 공격 대응 방안을 제안하였다. 한국어 처리에 특화된 KoNLPy 토크나이징 라이브러리 모델을 사용하였으며, KoNLPy 라이브러리에 포함된 형태소 분석기 5개 중, Okt, Kkma, Komoran 형태소 분석기를 활용하여 실험을 진행하였다. 데이터셋은 국내 피싱 메일 기준으로 가장 빈번한 단어를 첨가하여 약 50만 글자가 포함된 텍스트 파일로 구성되었다. 형태소 분석을 하였을 때, Okt 형태소 분석기, Kkma 형태소 분석기, Komoran 형태소 분석기 중 Kkma 형태소 분석기와 Komoran 형태소 분석기는 2~3개 정도 형태소 분석에 문제가 있었으며, Okt 형태소 분석기는 가장 정확하게 형태소 분석을 한 것을 확인하였다. 각 형태소 분석기의 속도 측정은 1초를 기준으로 소숫점 셋째 자리까지 나타내었다. 속도 측정 결과, Kkma 형태소 분석기는 0.019secs, Okt 형태소 분석기는 0.006secs, Komoran 형태소 분석기는 0.008secs로 측정되었으며, Okt 형태소 분석기가 가장 빠른 시간으로 측정된 것을 확인하였다. 실험을 통해 형태소 분석과 실행 속도 측정 모두 Okt 형태소 분석기가 가장 성능이 좋은 것을 확인하였다. 데이터셋은 변동성이 있기 때문에, 향후 모든 데이터셋에 대해 형태소 분석기 성능 비교 연구가 지속되어야 하며, 국내의 모든 피싱 메일을 확실하게 대처할 수 있는 연구가 지속되어야 할 것이다.

REFERENCES

[1] Jinhyun Yu, In Hyuk Seo, Seungjoo Kim "Study on DNN Based Android Malware Detection Method for Mobile Environment" KIPS Tr. Comp. and Comm. Sys. Vol 6, No. 3 pp.159-168 2020

[2] Joo-hun Boo, Kyung-ho Lee "Adavanced Feature Selection Method on Android Malware Detection by Machine Learning" Journal of The Korea Institute of Information Security & Cryptology, Vol. 30, No. 3, Jun. 2020.

[3] Choi Heesik, Cho Yanghyun, “Analysis of Security Threats from Increased Usage of Mobile App Services” *The korea Institute of Digital Industry Information Society,* Vol. 14, No.1 Mar. 2018

[4] JaeKyung-Lee, YoungBok-Cho “Analysis of the Korean Tokenizing Library Module” T*he Korea Information and Communication Society* Vol 25, No.1 pp.78-80 May. 2021

[5] Google Cloud by Natural-Language-Processing def https://cloud.google.com/learn/what-is-natural-language-processing?hl=ko