ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

작성자 분석 기반의 공격 메일 탐지를 위한 분류 모델☆

A Classification Model for Attack Mail Detection based on the Authorship Analysis

홍 성 삼[1] 신 건 윤[1] 한 명 묵[1]*
Sung-Sam, Hong Gun-Yoon, Shin Myung-Mook, Han

요 약

최근 사이버보안에서 악성코드를 이용한 공격은 메일에 악성코드를 첨부하여 이를 사용자가 실행하도록 유도하여 공격을 수행하는 형태가 늘어나고 있다. 특히 문서형태의 파일을 첨부하여 사용자가 쉽게 실행하게 되어 위험하다. 저자 분석은 NLP(Neutral Language Process) 및 텍스트 마이닝 분야에서 연구되어지고 있는 분야이며, 특정 언어로 이루어진 텍스트 문장, 글, 문서를 분석하여 작성한 저자를 분석하는 방법들은 연구하는 분야이다. 공격 메일의 경우 일정 공격자에 의해 작성되어지기 때문에 메일 내용 및 첨부된 문서 파일을 분석하여 해당 저자를 식별하면 정상메일과 더욱 구별된 특징들을 발견할 수 있으며, 탐지 정확도를 향상시킬 수 있다. 본 논문에서는 기존의 기계학습 기반의 스팸메일 탐지 모델에서 사용되는 특징들과 문서의 저자 분석에 사용되는 특징들로부터 공격메일을 분류 및 탐지를 할 수 있는 feature vector 및 이에 적합한 IADA2(Intelligent Attack mail Detection based on Authorship Analysis)탐지 모델을 제안하였다. 단순히 단어 기반의 특징들로 탐지하던 스팸메일 탐지 모델들을 개선하고, n-gram을 적용하여 단어의 시퀀스 특성을 반영한 특징을 추출하였다. 실험결과, 특징의 조합과 특징선택 기법, 적합한 모델들에 따라 성능이 개선됨을 검증할 수 있었으며, 제안하는 모델의 성능의 우수성과 개선 가능성을 확인할 수 있었다.

☞ 주제어 : 텍스트마이닝, 기계학습, 분류, 작성자분석, 공격자 식별

ABSTRACT

Recently, attackers using malicious code in cyber security have been increased by attaching malicious code to a mail and inducing the user to execute it. Especially, it is dangerous because it is easy to execute by attaching a document type file. The author analysis is a research area that is being studied in NLP (Neutral Language Process) and text mining, and it studies methods of analyzing authors by analyzing text sentences, texts, and documents in a specific language. In case of attack mail, it is created by the attacker. Therefore, by analyzing the contents of the mail and the attached document file and identifying the corresponding author, it is possible to discover more distinctive features from the normal mail and improve the detection accuracy.

In this pager, we proposed IADA2(Intelligent Attack mail Detection based on Authorship Analysis) model for attack mail detection. The feature vector that can classify and detect attack mail from the features used in the existing machine learning based spam detection model and the features used in the author analysis of the document and the IADA2 detection model. We have improved the detection models of attack mails by simply detecting term features and extracted features that reflect the sequence characteristics of words by applying n-grams. Result of experiment show that the proposed method improves performance according to feature combinations, feature selection techniques, and appropriate models.

☞ keyword : Text Mining, Machine Learning, Classification, Authorship Analysis, Attacker Identification

1. 서 론

최근 사이버보안에서 악성코드를 이용한 공격은 메일에 악성코드를 첨부하여 이를 사용자가 실행하도록 유도하여 공격을 수행하는 형태가 늘어나고 있다. 특히 문서형태의 파일을 첨부하여 사용자가 쉽게 실행하도록 유도하고 있다. 업무에서 문서파일(HWP, DOC, PDF)들이 주로 사용되고 있기 때문에 공격 대상이 되는 사용자는 쉽게 문서를 열 수밖에 없으며, 이를 이용한 악성코드는 탐지가 어렵고 실행가능성이 높아 그 위험도가 높아지고 있다. 이를 탐지하기 위한 방법으로 기계학습이 활용되고 있다. 문서 파일에서 추출할 수 있는 특징(feature)들을 이용하여 학습을 통해 과거의 악성 파일, 스팸메일 등을 분류하여 공격을 사전에 탐지하고 분류할 수 있는 연구들

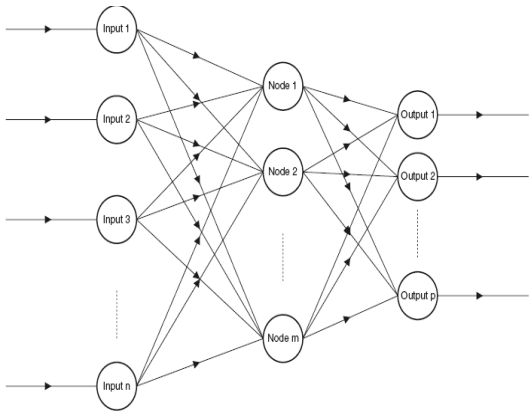
1 Department of Computer Engineering, Gachon University, Seongnam-si, 13120, Korea.
* Corresponding author (mmhan@gachon.ac.rk)
[Received 25 October 2017, Reviewed 26 October 2017, Accepted 1 November 2017]
☆ 이 논문은 2017년도 국방과학연구소의 지원으로 수행되었음 (UD170005ED)

이 수행되고 있다. 주로 메일에서 사용되는 특징들로는 메일의 헤더, EML, 메일의 내용, 단순 텍스트를 텍스트마이닝 방법으로 이용한 특징들이 있으며, 경우에 따라 키보드 캐릭터 셋, 폰트스타일등을 이용하는 경우도 있다. 또한 문서 파일의 경우 문서 포맷이 갖고 있는 고유한 구조를 이용하여 메타 정보로부터 추출한 특징, 파일에 첨부된 URL, 데이터 스트림, OLE 개체 등[1]을 이용하여 특징을 추출하기도 한다. 이런 추출된 특징으로 구성된 feature vector를 이용하여 학습셋으로 모델을 구축하고, 새로운 데이터를 분류를 통해 악성 문서를 탐지한다.

저자 분석은 NLP(Neutral Language Process) 및 텍스트마이닝 분야에서 연구되어지고 있는 분야이며, 특정 언어로 이루어진 텍스트 문장, 글, 문서를 분석하여 작성한 저자를 분석하는 연구 분야이다. 저자 분류는 도메인 전문가에 의한 분류 방법, 규칙기반 분류 방법, 지도 학습(supervised learning)에 의한 학습기반 분류 방법으로 구분할 수 있다. 특히 학습기반 접근 방법은 새로운 메일 및 문서에 대해 작성자의 특성이 잘 표현된 특징들이 필요하며, 이로 인해 문서 데이터에 대한 일반화 및 작성자별로 정확한 프로파일을 얻을 수 있다는 장점이 존재한다[2].

악성 메일의 경우 일정 공격자에 의해 작성되어지기 때문에 메일 내용 및 첨부된 문서 파일을 분석하여 해당 저자를 식별하면 정상메일과 더욱 구별된 특징들을 발견할 수 있으며, 탐지 정확도를 향상시킬 수 있다. 문장이나 어휘의 구성 스타일, 작성자의 버릇에 의해 발생하는 특성, 문장 구조, 비정상적 언어 패턴, 숫자/대문자/특수문자 등의 과도한 사용 등을 특징으로 추출할 수 있다면 악의적인 공격 메일을 탐지하는데, 의미있는 데이터 특징으로 사용될 수 있다[3].

본 논문에서는 기존의 기계학습 기반의 스팸메일 탐지 모델에서 사용되는 특징들과 문서의 저자 분석에 사용되는 특징들로부터 공격메일을 분류 및 탐지를 할 수 있는 feature vector와 이에 적합한 IADA2(Intelligent Attack mail Detection based on Authorship Analysis)탐지 모델을 제안하였다. 단순히 단어 기반 특징들로 탐지하던 스팸메일 탐지 모델들을 개선하고, n-gram을 적용하여 단어의 시퀀스 특성을 반영한 특징을 추출하였다. 또한 문서저자분석에서 적용되는 특징들을 분석하여, 공격 메일 탐지에 필요한 주요 특징들을 추출하였다. 추출된 특징들의 구성을 분석하여, 필요한 특징선택 기법 및 적절한 탐지 모델을 제시하기 위해 실험을 통해 모델들의 성능을 검증하였다. 실험결과, 특징의 조합과 특징선택 기법, 적합한 모델들에 따라 성능이 개선됨을 검증할 수 있었으며, 제안하는



(그림 1) BNPP(3-Layer Back-Propagation Neural Network)

(Figure 1) BNPP((3-Layer Back-Propagation Neural Network)

모델의 성능의 우수성과 개선 가능성을 확인하였다.

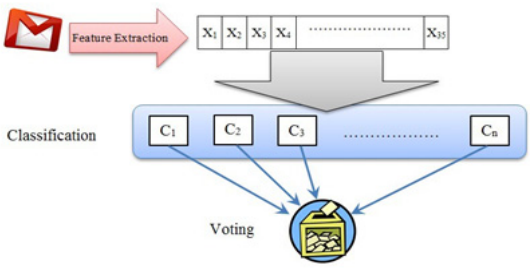
본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 제안하는 모델 및 프로세스에 대해 서술하였다. 4장에서는 성능 평가를 위해 다양한 실험을 수행하여 성능을 검증하였으며, 5장에서 결론을 맺는다.

2. 관련 연구

2.1 기계학습 기반 스팸메일 탐지

기계학습을 이용한 스팸메일 탐지는 기존의 필터링 기반의 스팸메일 탐지 기법들을 개선하고, 자동화된 스팸메일 분류 및 탐지 시스템을 구축하기 위해 연구되어지고 있는 분야이다. 주로 메일의 내용에서 term feature(단어들)을 추출하여, 각 단어 특징의 빈도수(TF : Term Frequency), TF-IDF(Term Frequency-Inverse Document Frequency) 등을 이용하여, 데이터 셋을 구축하고 기계학습 모델에 학습 및 분류를 수행하는 모델들을 제안하고 있다[24]. [4]에서는 스팸메일 탐지를 위해 그림 1에서 나타나듯이 3계층 역전파 신경망(3-Layer BPNN : three-layer Back-Propagation Neural Network)을 이용하였다.

여기서는 CFC(Concentration based Feature Construction) 접근법을 제안하여 'Self'와 'non-Self') gene 라이브러리를 통해 이메일을 표현하기 위해서 2개의 요소의 concentration 특징 벡터를 생성한다. CFC에 의해서 효율적으로 BPNN이 메일을 자동으로 분류하여 스팸과 정상메일을 탐지한다.



(그림 2) body features and readability features 추출 방법

(Figure 2) Extraction Method of body feature and readability features

[5]에서는 이메일을 통한 악성코드 및 웜을 탐지하기 위해 메일에서 추출할 수 있는 특징을 추출하여 베이지안 네트워크와 결정트리를 구성하여, 악성 메일을 탐지하는 방법을 제안하였다. 연구의 목적은 새로운 (보이지 않는) 악성 메일에 대한 정확한 탐지모델을 제안하는 것이다. 탐지 시스템을 구축하기 위해서 베이지안 확률론적 네트워크를 구축하여 제시하였으며, 비교대상으로 결정트리 유도 방법을 사용하였다. 데이터 세트에서 각 이메일에서 프로필을 추출하고, 프로필에서 분류기에서 사용할 특징을 추출했다. 본 연구에서는 메일에 포함된 악성코드의 정적분석 특징을 메일 탐지에 사용하였다.

[6]에서는 그림 2와 같이 스팸메일을 탐지하기 위해 e-mail에서부터 두 가지 유형의 특징을 추출하는데 body features and readability features를 추출한다. 이를 기계학습 기반의 분류기를 이용하여 스팸메일을 탐지한다. 일반적으로 전자 메일 특징은 본문, 제목 또는 헤더 필드의 텍스트에서 추출된다. 이러한 유형의 기능을 컨텐츠 기반 특징이라고 한다. 본 논문에서는 전자 메일 스팸 필터링 (즉, 용어 - 빈도 분석 접근법, 경험적 접근법 및 행동 기반 접근법)을 위한 특징을 구성하기 위한 다양한 접근법을 제안하였다 용어 - 빈도 분석에서 전자 메일의 모든 단어는 특징으로 정의되고 단어 벡터는 전자 메일을 나타내는 데 사용된다. 제안하는 탐지 방법은 다양한 접근법으로 특징을 추출하여, 다수의 분류기로부터 분류결과를 도출하여 이를 투표방식의 앙상블방법을 이용하여 최종 탐지 결과를 도출하였다.

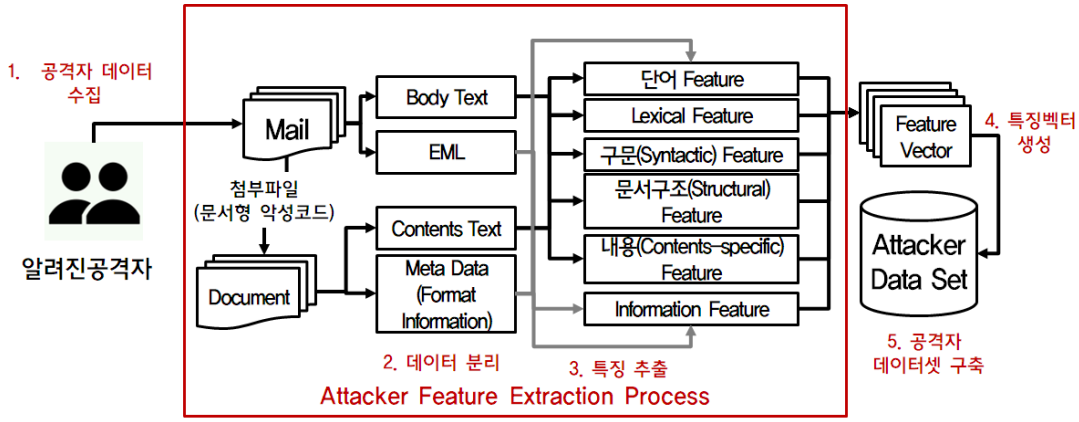
2.2 텍스트 데이터의 작성자 분석을 이용한 탐지

텍스트 데이터로부터 작성자 분석은 텍스트 마이닝 분야에서 계속적으로 연구되어지는 분야이다[21, 22]. 단순한 글들의 저자 식별뿐만 아니라 컴퓨터 분야에서 소스코드, 메일 등의 저자를 분석하여 공격자를 식별하는 연구도 진행되고 있다[3, 7, 8, 9]. 즉, 작성자 분석은 e-maill messages, source code 등의 텍스트 데이터를 이용하여 작성자를 식별하는 방법으로, 이를 통해 작성자에 대한 증거, 특징을 찾는 접근방법이다. 작성자 분석은 일반적으로 3가지 방식으로 구분한다[2].

- Authorship Identification(Attribution)
  - 특정 작성자에 의해 작성된 다양한 코드, 메시지 등을 분석하여 작성자의 특징을 추출
  - 특정 작성자만을 식별할 수 있는 유일한(unique) 특징을 찾는 것이 목표
  - 작성자가 알려지지 않은 코드, 메시지 등과 특징을 비교하여 특정 작성자에 의해 작성되었는지 여부
- Authorship Characterization
  - 특정 작성자 특성(Characterization)을 파악하고, 프로필을 생성하는 방법
  - 작성자의 성별, 교육, 문화, 배경, 작성 스타일 등을 기반으로 프로필 생성
  - 특정 작성자만을 식별할 수 있는 유일한(unique) 특징을 찾음
- Similarity Detection
  - 여러 코드, 메시지를 분석 및 비교하여 모방이 아닌 한사람의 작성자가 작성하였는지 여부 파악
  - 어휘, 기능, 구문 특징, 구조적 특징, 내용 등 다양한 특징을 가지고 특징 추출
  - 프로그램 저작권 등을 파악할 때 사용

3. IADA2 : 공격 메일 탐지 모델

제안하는 IADA2(Intelligent Attack mail Detection based on Authorship Analysis)는 기존의 스팸메일 탐지 모델에서 사용하는 term feature들과 작성자 분석에서 사용하는 특징들, 그리고 문서형 악성코드 분석에서 사용되는 특징들을 분석하여 공격 메일 탐지에 필요한 주요 특징들을 추출하였다. 이를 기반으로 공격 메일 탐지에 적용할 특징 벡터 생성을 위해 추출해야할 특징 유형 셋을 제안하였다. 또한 term feature의 n-gram feature vector의 경우 아주 많은 수의 특징이 발생하여, 고차원 데이터를 생성하기 때문에 특징 선택을 적용하였다. 공격 메일 탐지를 위해 기존의 정의된 각 특징들을 재구성하여 추출하였으며,



(그림 3) 공격 메일 특징 추출 프로세스 구조
(Figure 3) Attack Mail Feature Extraction Process

필요한 특징들로 구분하였다. 특징 유형에 따라 적합한 특징 선택 및 탐지 모델이 필요하므로 특징에 대한 분석을 수행하였다.

3.1 IADA2 특징 추출 및 탐지 프로세스

공격 메일을 탐지하기 위한 IADA2의 특징 추출 프로세스는 그림 3과 같으며, 그 과정은 아래와 같다.

① 알려진 공격자/공격 그룹에서 사용한 메일 및 문서형 악성코드를 수집
② 수집한 데이터를 메일과 문서형 악성코드로 분리하고, 메일은 body text와 eml, 문서형 악성코드는 content text(문서에 포함된 내용들)과 meta data(파일 포맷, 그 외 정보들)로 분리함
③ 분리한 raw data로부터 각 유형별 feature 들을 그림과 같이 추출함
- Body Text, Contents Text : 단어, 어휘, 구문, 문서구조, 내용 Feature
- EML, Meta Data : 단어, Information Feature
④ 각 데이터 포인트별로 추출한 특징들을 feature vector로 구성
⑤ 만들어진 feature vector들로 attacker data set 구축
- 알려진 공격자들의 메일 및 문서형 악성코드로부터 구축하였기 때문에 label을 정의

3.2 공격 메일 탐지를 위한 특징

그림 3의 공격 메일 탐지를 위한 특징들을 유형별로 정리하면 아래와 같다. [3, 9] 논문의 저자 분석 기반의 특징들과 문서형 악성코드 분석에서 사용한 특징, 그리고 기타 악성코드 및 침해사고 분석에서 사용한 메일 특징들을 분석하여, 이를 기반으로 구성된 공격메일 탐지를 위한 특징 벡터를 제안하였다. 각 특징 유형은 아래와 같으며, 본 모델에서는 각각의 방법들에서 공격 메일 탐지에 필요한 특징들을 선택하여 추출하였다. 각 특징들이 갖고 있는 의미가 상이하기 때문에 특징들이 갖는 값들에 대한 일관성 확보 및 일반화를 위해 특징 선택 및 정규화 방법을 적용한다.

3.2.1 단어 특징

일반적으로 문서내 내용, 메일에서는 body text의 내용을 텍스트 마이닝의 tokenize 방법[23]을 이용하여 feature extraction과정을 거쳐서 생성되는 term feature들(예 : 출현한 단어들 - computer, software, love, finance 등의 문서에 사용된 기본 단어들) 이다. 각 문서가 갖고 있는 기본적인 내용이나 의미, 단어별 중요도 등을 분석한다.

3.2.2 어휘적(Lexical) 특징

어휘특징은 문서에서 얻을 수 있는 가장 기본적인 특징들로 본문에 포함된 단어나 문자를 특징으로 하여 빈도수나 엔트로피 등으로 특징에 대한 값을 정하여, 특징 벡터를 생성할 수 있다. 단어별 사용 빈도, 사용되는 단어 길이 등이 특징으로 추출되어서 사용될 수 있으며, 언어마다 단어를 추출하는 방법이 필요하다. 메일 분석에서는

메일의 body text부분을 분석할 때 이 특징들을 사용한다. 총 단어의 수, 단어의 평균길이, 숫자캐릭터의 개수, 탭의 개수 등이 이 특징에 해당한다[3].

3.2.3 구조적 특징

일반적으로 구조적 특징은 글을 작성한 사람의 글쓰기 레이아웃을 구성하는 방식을 나타낸다. 이메일에서도 body 에 대한 글의 구조적 특징을 추출할 수 있다. 문장의 수, 문장의 평균 길이, 숫자의 개수, 들여쓰기, 서명 등 이 구조적 특징에 해당한다. [7]에서는 e-mail의 body에서 특징을 추출하여 일부 구조적 특징을 사용하여 메일을 분류하는 모델을 제안하였다. 대표적 구조적 특징은 아래와 같다.

- Total number of lines
- Total number of sentences
- Total number of paragraphs
- Number of sentences per paragraph
- Number of characters per paragraph
- Number of words per paragraph ...

3.2.4 구문적 특징

기능적 단어(function word), 구두점 및 품사를 포함한 구문 특징은 문장 수준에서 작성자의 작성 스타일 분석할 수 있다. 이를 이용하여 비슷하게 만들어진 메일이나 문서들을 분석하고, 또한 그 저자를 분석해 낼 수 있다. 다른 특징들에 비해 각 도메인과 나라의 언어마다 선택할 수 있는 특징들이 다르게 나타난다[3]. 기능적 단어들은 문장에서 특별한 역할이나 기능을 하는 단어들을 뜻하는데(표 2) 예를 들어 영어에서는 a, by, is, on, nor, at, as 등, 한국어에서는 ~에, ~는, ~도 등의 조사가 대표적인 기능적 단어 또는 character로 각 나라의 언어마다 다르게 구성되어 있다.

3.2.5 콘텐츠적 특징

콘텐츠 즉 내용과 상관없는 특징들이 있는 반면, 문서의 내용에 의존적인 특징들도 있다. contents-specific feature는 특히 온라인 메시지들(메일을 포함)에 중요한 차별화된 특징들이다. 특정 응용 프로그램 도메인에 따라 다르게 나타나기 때문이다. 온라인, 메일 작성자들의 경우 특정 주제와 밀접한 관련이 있는 단어들을 주로 사용할 수 있다. 악성메일을 식별할 때도 공격자 별로 주로 쓰는 문서의 주제나 내용들을 기반으로 공격자를 특정할 수 도 있다. 이러한 이유로 특정 주제와 밀접한 관련이 있는 특수한 단어나 문자가 저자의 신원에 대한 단서를 제공할 수 도 있다. 예를 들어[3], 해적판 소프트웨어를 판매하는 범죄자는 "판매", "소프트웨어", "free" 등의 단어를 사용할 수 있다. 음란물 등의 경우 "섹시한"과 같은 단어를 자주 사용한다. 이러한 콘텐츠적 특징들은 다른 단어들과 구별하여 가중치나 특징셋을 구성하여 식별 및 분류할 때 차별화된 특징으로 사용될 수 있다.

3.2.6 Information Feature

Information Feature는 메일 구조 및 내용 외에서 추출할 수 있는 특징들로 첨부된 문서형 악성코드로부터 추출하는 파일 포맷에 따른 특징과 기타 특징들로 구분할 수 있다. 정보 특징의 경우 특정 문서형 악성코드, 특정 메일, 특정 도메인 등에서만 추출이 가능하므로 모든 상황에서 적용하기에는 어렵지만 특수한 경우 주요 특징으로 활용될 수 있는 특징들이다.

① 파일 포맷에 따른 특징 - 파일의 구조 및 메타데이터에서 추출

파일 포맷에서부터 특징을 추출하기 위해서는 파일 포맷 구조를 이해하여야하며, 파일 포맷 구조가 공개되어 있어야한다. 또한 기존의 공격에 사용되었던 악성 문서들의 공격 형태를 사전지식으로 알고 있으면, 특징을 추출하는데 유용하다.

- PDF 형식

PDF의 악성여부를 식별하기 위해서는 PDF 메타데이터 또는 구조에서부터 특징을 추출할 수 있다[4]. 특정 문자열이나 바이트 시퀀스에 의존하지 않도록 feature를 설계하였다. 악성 pdf 문서의 경우 문서내 임베디드 이미지나 오브젝트에 악성코드를 삽입하여 실행되게 만드는 형태의 악성코드들이 다수 존재하기 때문에 문자열 외 특징을 추출할 필요가 있다. 따라서 메타데이터나 구조, 본문에서 추가적으로 매개변수화를 시킨 feature들을 추출하였다. 예를 들면 특정 메타데이터 필드나 구조내 필드의 문자수, 폰트, 인코딩 방법, 객체의 수, 객체 유형 등이 있다. [10]논문의 경우 PDF로부터 202개의 feature를 추출하였다. count_font, count_javascript, count_js는 /font, /javascript, /js 마커의 인스턴스 수를 표현하는 특징들을

추출하였다. 이러한 메타데이터 및 구조내에서 추출한 특징들과 본문의 내용에서 추출하는 어휘, 구문 등의 언어적 특징을 혼합하여 저자/공격자/악성 문서 식별에 사용할 수 있다.

- DOC 형식

DOC 문서 구조는 디렉토리 형태의 트리 구조를 보이고 있다. [3]에서는 각 doc 파일 구조에 나타난 path들을 특징으로 추출하여 각 문서가 갖고 있는 path 들을 추출하였다. 이 path들이 갖고 있는 의미는 파일의 속성과 동작을 나타낼 수 있으며, 이에 대한 count나 존재 여부 등을 feature로 추출하면 공격 메일 탐지에 의미있는 feature로 사용될 수 있다.

② 기타 정보를 담고 있는 특징

모든 경우에 특징으로는 사용하기 어려울 수 있으나 공격 메일을 분류하고, 공격자를 식별하는데 활용될 수 있는 몇몇 정보로부터 추출할 수 있는 특징들을 설명한다. 실제 [11]에서 이러한 특징들로 분석한 사례들을 소개하고 있으며, 이들에 대한 자동화된 특징 추출 방법이 있다면, 각 특징들의 연관성을 통해 특정 공격자 추론 및 공격 메일 탐지가 가능하다.

- 작성된 컴퓨터의 키보드의 charset 특징

몇몇 메일의 경우 eml 헤더에 charset 정보를 담고 있다. 이 charset은 악성코드나 메일을 작성할 때 사용한 키보드의 charset 정보를 알 수 있는데 대부분은 표준 키보드 레이아웃을 사용하나, 일부의 경우 북한이나 중국 키보드 레이아웃을 사용할 경우 악성코드 및 문서, 메일을 작성한 공격자의 지역이나 위치, 국적을 식별할 수 있다. 실제 공격에 사용된 이메일의 EML에서 charset정보를 얻은 연구 사례[11]도 있다.

- 작성된 문서의 폰트 특징

일부 문서형 악성코드내 작성된 문서 내용을 보면 다른 나라의 언어를 사용하여 작성하였지만 실제 폰트는 특정 나라에서만 사용하는 폰트로 작성된 경우가 있다. 이런 경우 폰트가 작성자나 공격자의 국적이나 위치를 특정할 수 있고, 비정상 메일임을 고려해볼 수 있다. [11]에는 실제 공격 사례로 러시아어로 작성되었지만 폰트는 한국어 폰트인 바탕체, 특히 북한에서 사용하는 KPCheongPong체를 사용하여 비정상 문서임을 식별한 분석 결과가 소개되어 있다.

- 작성자의 동일한 행동 특징

이메일을 작성한 작성자의 동일한 행동 습관을 알 수 있는 특징들을 추출한다면 공격자를 식별하는데 유용하게 활용될 수 있다. 실제 서로 다른 메일들에서 eml 헤더의 ip정보가 같게 나타나 같은 공격자가 작성한 것임을 확인할 수 있었던 공격 사례가 있다. 이 전에 각 메일들은 타겟의 직업 유형(종교인)으로 같았기 때문에 선분별할 수 있었다. 접근법이나 스타일을 알 수 있는 특징이 있다면 공격자를 식별하는데 주요한 특징으로 활용될 수 있다. 피해자의 직업, 직책 등을 feature로 활용할 수 있으며, 각 메일헤더에 있는 ip정보도 특징으로 사용될 수 있다.

3.3 공격 메일 탐지 프로세스

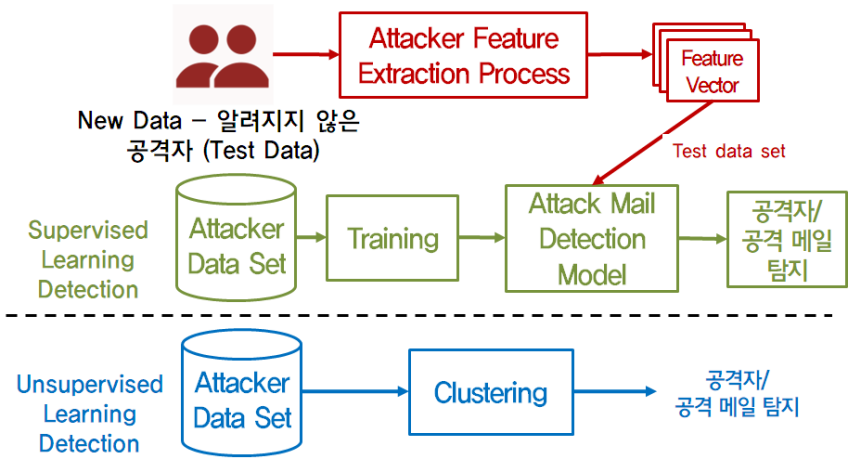
- Supervised Learning Detection
  - 앞서 구축된 Attacker Data Set을 기계학습 알고리즘으로 학습(training)하여 공격자 식별 모델을 구축
  - 새로운 공격자(알려지지 않은 공격자) 데이터 또는 데이터 셋에 대해 Attacker Feature Extraction Process하여 특징을 추출하고, 이로부터 만들어진 feature vector들을 이용하여 공격자 식별 모델에 test data로 적용
  - 공격자 식별 모델로부터 test data에 대한 식별 결과를 얻고, 실험 결과를 분석함
- Unsupervised Learning Detection
  - 앞서 구축된 Attacker Data Set로부터 clustering 알고리즘을 수행
  - 수행한 결과 도출된 cluster들을 분석하여, 유사 공격자/공격 그룹을 식별
  - 식별된 유사 공격자/공격 그룹들을 분석하여, 다양한 알려지지 않은 결과를 도출하고, cluster 결과들을 추후 학습에 사용되는 label로 활용될 수 있음(가능성 평가)

4. 실험 및 결과

4.1 실험 환경 및 방법

실험에 사용된 하드웨어 및 OS 환경은 다음과 같다.

- CPU : Intel Core i7 6700k 4.00Ghz
- RAM : 16GB



(그림 4) 공격 메일 탐지 프로세스
(Figure 4) Attack Mail Detection Process

- OS : Windows 10 64bit

실험을 위해 사용된 tool은 Python을 사용하였으며 버전은 3.61이다[12]. 분류 모델을 사용하기 위해 사용한 패키지는 scikit-learn이며 버전은 0.19.1[13]이다.

4.2 데이터 셋, 특징 벡터, 분류 모델

실험을 위해 사용된 데이터 셋은 CSDM 2010[14]으로 스팸메일과 정상메일 분류 테스트를 할 수 있도록 구성된 데이터 셋이다. 언어는 영어로 구성되어 있다. eml을 포함한 메일의 전체 내용을 포함하고 있으며, 이중 1700개의 메일을 실험에 사용하였다. 정상과 공격 메일의 구성은 정상 1,139개, 공격 562개이다. 실험에 사용한 특징은 term feature, term feature에 대해 PCA(Principle Component Analysis) 특징 선택을 적용한 특징 벡터, 어휘적(lexical) feature 8개이며, 추출 방법은 아래와 같다.

- Term Feature
  - 문서에 대한 전처리과정으로 특수문자, 이메일주소, 숫자를 제거하고, 영어에서 사용하는 stopword(불용어)를 제거
  - 단어 단위로 특징을 추출
  - 추출된 raw term feature로부터 n-gram을 적용하여 3-gram feature를 구성함
  - 총 366,083개의 n-gram feature중 최상위 빈도 1000개의 feature만 사용(빈도수가 적은 feature들일수록 분류 성능을 낮추기 때문에 높은 빈도수 feature들을 사용)
  - 모든 n-gram feature에 대해 PCA를 적용하여, 도출된 주성분 중 8개의 주성분을 사용하여 특징 벡터 생성
- Lexical Feature
  - eml을 포함한 문서 전체에 대해 어휘적 feature를 추출하여 생성함
  - 총 8개의 어휘적 feature를 추출하여 feature vector를 생성 : "total_number_of_words","average_length_of_words", "total_number_of_characters","total_number_of_alphabet", "total_number_of_digit", "total_number_of_uppercase", "total_number_of_whitespace","total_number_of_punctuation"

각각의 특징 벡터를 n-gram feature, PCA feature, 어휘적 feature만 사용한 경우와 각각을 조합하여 사용한 경우에 대해 다양한 분류 모델에 적용하여 공격 메일 탐지성능을 실험하였다.

실험에 적용한 분류 모델은 SVM(Support Vector Machine)[15], DT(Decision Tree)[16], kNN(k Nearest Neighbor)[17], RF(Random Forest)[18], 인공신경망 중 MLP(Multi-Layer Perceptron)[19]를 사용하여 성능을 비교분석하였다.

4.3 성능 측정 방법

본 논문에서는 탐지성능을 측정하기 위해 F1-measure

와 분류 정확도를 사용한다. F1-measure는 데이터 분류, 문서 분류, 분류탐지에서 단순 정확도나 탐지율 등의 성능평가 방법을 개선한 방법이다. TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)으로 precision과 recall값을 구하면 각 값의 비중을 동일하게 하여 조화 평균을 구한다[20]. 높을수록 분류탐지기의 성능이 높다고 평가한다. 본 평가에서는 탐지 성능을 체크하기 위해서 positive 클래스를 attack로 정하고 성능평가를 한다. P는 precision, R은 recall이며 각 식 (1)에 의해 구해지며 F1-measure는 식 (2)과 같다. 먼저 supervised 방법에 대한 결과를 평가하는 방법이다.

- TN: Normal data correctly classified as normal.
- TP: Anomalous data correctly classified as anomalous.
- FP: Normal data classified as anomalous.
- FN: Anomalous data classified as normal.

$$P = \frac{TP}{(TP+FP)} \quad R = \frac{TP}{(TP+FN)} \tag{1}$$

$$F1-measure = \frac{2 \times P \times R}{P+R} \tag{2}$$

4.4 실험결과

4.4.1 단일 특징 벡터별 실험 결과

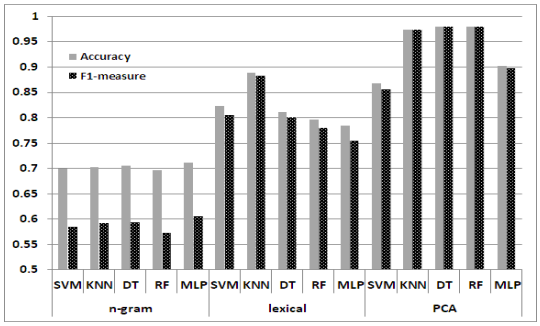
먼저 단일 특징 벡터별로 5개의 분류 모델을 이용하여 공격 메일 탐지를 수행한 결과는 표 1과 같다. 기본적인 term feature로부터 생성한 n-gram(3-gram) term feature 벡터를 이용한 탐지 실험의 경우 MLP 모델이 F1-measure: 0.6056, 정확도: 0.7117로 가장 우수한 성능을 보여주었다. 하지만 모든 모델의 성능을 보면, 평균 F1-measure: 0.5893, 정확도: 0.7034로 다소 낮은 성능치를 보여주었다. 어휘 feature의 경우 kNN 모델이 F1-measure: 0.8830, 정확도: 0.8882로 가장 좋은 성능을 보여주었다. 모든 모델의 성능은 평균 F1-measure: 0.8044, 정확도: 0.8211로 n-gram feature에 비해 좋은 탐지 성능을 보여주었다.

n-gram feature로부터 PCA 특징 선택을 수행한 주성분으로 구성된 특징벡터에 대한 분류 모델 실험에서는 DT와 RF가 각각 F1-measure: 0.9794, 정확도: 0.9794로 우수한 성능을 나타내었다. 모든 모델의 평균 성능은 F1-measure: 0.9371, 정확도: 0.9406으로 아주 우수한 성능을 나타내는 것을 볼 수 있다.

(표 1) feature vector 별 분류 모델 실험 결과
(Table 1) Results of Classification Model Experiments by Feature Vector

Term feature (n-gram) only				
	Precision	Recall	Accuracy	F1-measure
SVM	0.6901	0.6999	0.7000	0.5846
KNN	0.7168	0.7029	0.7029	0.5911
DT	0.7931	0.7058	0.7058	0.5927
RF	0.4858	0.6970	0.6970	0.5726
MLP	**0.7960**	**0.7117**	**0.7117**	**0.6056**
Average	0.6964	0.7035	0.7035	0.5893
Lexical feature only				
	Precision	Recall	Accuracy	F1-measure
SVM	0.8348	0.8235	0.8235	0.8048
KNN	**0.8915**	**0.8882**	**0.8882**	**0.8830**
DT	0.8073	0.8117	0.8117	0.8004
RF	0.7936	0.7970	0.7970	0.7799
MLP	0.8000	0.7852	0.7852	0.7537
Average	0.8254	0.8211	0.8211	0.8044
PCA feature selection (by term feature) only				
	Precision	Recall	Accuracy	F1-measure
SVM	0.8854	0.8676	0.8676	0.8556
KNN	0.9734	0.9735	0.9735	0.9734
DT	**0.9794**	**0.9794**	**0.9794**	**0.9794**
RF	**0.9794**	**0.9794**	**0.9794**	**0.9794**
MLP	0.9105	0.9029	0.9029	0.8977
Average	0.9456	0.9406	0.9406	0.9371

본 실험결과에 대해 전체적으로 분석해볼 때, 각각 단일 특징의 경우 어휘적 특징이 단어 특징보다는 분류 및 탐지 성능이 좋게 나타났다. 이를 통해, 어휘적 특징이 문서 전체의 특성을 좀 더 반영할 수 있다는 것을 알 수 있다. 또한 PCA 특징선택을 통해 만들어진 특징 벡터의 경우 위 두 가지 경우에 비해 상당히 높은 성능을 보여주었는데 이는 특징 선택 기법인 단순한 raw feature를 사용하는 것에 비해 분류 및 탐지 모델의 성능을 높여줄 수 있음을 판단할 수 있다. 보편적으로 kNN과 DT는 모든 데이터셋에서 무난한 성능과 성능 편차가 적게 나타났으나, 각 특징별로 가장 좋은 성능을 나타내는 모델이 다르게 나타난 것을 볼 수 있다. 이는 특징의 유형 및 특성에 따라 적합한 모델이 있음을 알 수 있었으며 따라서 분류 및 탐지 모델 선정에 있어서 특징과 특징 선택 그리고 분류 모델과의 적합성을 고려할 필요가 있음을 알 수 있다.



(그림 5) Feature vector 별 분류 모델 실험 결과(차트)
(Figure 5) Results of Classification Model Experiments by Feature Vector(Chart)

4.4.2 특징 벡터 조합별 실험 결과

• term feature (n-gram) + lexical feature

먼저 n-gram과 어휘 feature를 조합한 특징 벡터를 이용한 실험결과는 표 2와 같다. 가장 우수한 성능을 나타낸 모델은 kNN으로 F1-measure: 0.8887, 정확도: 0.8941의 성능을 보여주었다. kNN 분류 모델의 성능은 나쁘지 않게 나타났지만, 전체 모델의 평균 성능은 평균 F1-measure: 0.7316, 정확도: 0.7823으로 좋지 않은 성능을 나타내고 있다. 다만 앞서 실험한 n-gram feature만 사용한 경우보다는 개선된 성능을 나타내는 것을 볼 수 있다.

• pca feature selection + lexical feature

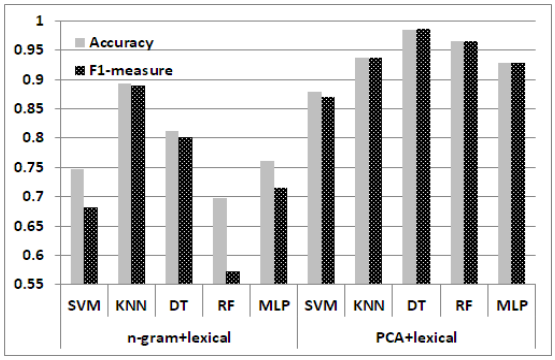
다음으로는 PCA 특징 선택 후 특징 벡터와 어휘적 feature를 조합한 데이터셋의 분류 모델별 실험 결과이다. 성능은 DT가 F1-measure: 0.9853, 정확도: 0.9582로 가장

(표 2) term feature (n-gram) + lexical feature 분류 모델 실험 결과
(Table 2) Results of term feature (n-gram) + lexical feature Classification Model Experiments

	Precision	Recall	Accuracy	F1-measure
SVM	0.7875	0.7470	0.7470	0.6820
KNN	**0.8996**	**0.8941**	**0.8941**	**0.8887**
DT	0.8073	0.8117	0.8117	0.8004
RF	0.6970	0.6970	0.6970	0.5726
MLP	0.7806	0.7617	0.7617	0.7146
Average	0.7944	0.7823	0.7823	0.7317

(표 3) PCA(by term feature) + lexical feature 분류 모델 실험 결과
(Table 3) Results of PCA(by term feature) + lexical feature Classification Model Experiments

	Precision	Recall	Accuracy	F1-measure
SVM	0.8916	0.8794	0.8794	0.8705
KNN	0.9391	0.9382	0.9382	0.9369
DT	**0.9855**	**0.9852**	**0.9852**	**0.9853**
RF	0.9646	0.9647	0.9647	0.9645
MLP	0.9330	0.9294	0.9294	0.9270
Average	0.9428	0.9394	0.9394	0.9368



(그림 6) Feature vector 조합별 분류 모델 실험 결과(차트)
(Figure 6) Results of Classification Model Experiments by Combined Feature Vector(Chart)

높은 성능을 나타냈으며 이는 모든 실험을 종합하여 가장 좋은 성능 수치를 보여준 결과이다. DT 모델의 경우 두 개의 특징을 조합할 경우 성능이 개선되는 것을 볼 수 있었다. 모든 모델의 평균 성능은 F1-measure: 0.9368, 정확도: 0.9393으로 우수한 성능을 나타내었다. 앞서 PCA feature만 활용한 것과 비슷한 결과를 나타내었으나 모든 모델에서, 특징을 조합했을 때 결과가 개선되는 것이 아님을 알 수 있었다. 이는 모델의 특성이나 데이터 셋에 특성에 따라 조합에 따른 결과가 다르게 나타난다는 것을 알 수 있는 결과이다.

결론적으로 특징을 단일 특징을 사용할 때보다는 전체적으로 조합한 특징을 사용했을 때가 보편적으로 우수한 성능을 나타내는 것을 볼 수 있었으며, 가장 우수한 결과는 특징의 조합과 특징 선택, 그리고 적절한 분류 모델을 선택했을 때 가장 좋은 결과를 나타내는 것을 볼 수 있다.

이를 통해 공격 메일 탐지 모델을 구축함에 있어서 주요 특징 추출, 이에 따른 특징 선택 방법, 적절한 기계학습 모델에 대한 연구 및 설계가 우수한 탐지 모델의 성능을 도출하는데 중요한 요소임을 확인할 수 있었다. 또한 제안하는 IADA2에서 2가지 요소의 특징만을 사용하여도 실용가능한 성능을 나타내는 것으로 보아 공격 메일 탐지 모델로의 우수한 성능을 보여준 것을 확인할 수 있었다. 또한 개선 가능성을 확인할 수 있었으며 추후 연구 및 실험을 통해 모든 특징 추출과 특징 선택, 탐지모델을 설계한다면 개선된 성능을 나타낼 것이라 예상할 수 있다.

5. 결 론

본 논문에서는 최근 메일을 통한 문서형 악성코드 공격이 증가함에 따라, 공격 메일을 분류하고 탐지하기 위한 작성자 분석 기반의 IADA2 모델을 제안하였다. 단순한 단어 특징으로부터 기계학습 모델을 적용하는 것이 아닌 작성자를 특정할 수 있는 특징들과 악성코드 분석에서 사용되는 특징을 활용하여 공격 메일 및 공격자를 특정할 수 있는 탐지 모델을 제안하였다. 그 중 일부 특징 및 특징 조합, 특징선택기법을 다양한 기계학습 분류 모델에 적용하여 실제 공격 메일이 탐지되는지 실험을 통해 검증하였다. 실험 결과 제안하는 모델에서 term feature의 n-gram 및 PCA 특징선택을 적용한 feature vecotr와 어휘 feature vector를 조합하여, DT에 적용한 결과가 가장 우수한 성능을 나타내었다. 또한 다양한 실험을 통해 각 특징의 유형이나 특성별로 적합한 특징선택방법, 분류 및 탐지모델이 나타나는 것을 확인할 수 있었다. 또한 제안하는 모델의 모든 특징들을 구성하고 적합한 특징 선택 및 기계학습 모델을 설계한다면 더욱 우수한 성능으로 개선될 수 있는 가능성을 검증할 수 있었다.

추후 연구로는 공격 메일로부터 얻을 수 있는 주요한 특징 벡터와 공격 메일 탐지에 적합한 특징 선택 기법을 연구하여 모델에 적용하고 개선된 모델을 연구하고자 한다. 또한 공격 메일 탐지 및 공격자 식별에 적합한 기계학습 모델에 대한 연구와 이를 기반으로 악성코드, 공격 메일, 침해사고 보고서 등 비정형의 데이터로부터 공격자를 식별하기 위한 특징 추출 방법 및 공격자 식별 모델을 연구하고자 한다.

참고문헌(Reference)

[1] Nir Nissim, Aviad Cohen, and Yuval Elovici, "ALDOCX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology," IEEE Transactions on Information Forensics and Security, vol.12, no.3, pp.631-646, 2017 https://doi.org/10.1109/tifs.2016.2631905

[2] Nathan Rosenblum, Xiaojin Zhu, Barton P. Miller, "Who Wrote This Code? Identifying the Authors of Program Binaries," Proceedings of the 16th European conference on Research in computer security, pp.172-189, 2011 https://doi.org/10.1007/978-3-642-23822-2_10

[3] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," Journal of the Association for Information Science and Technology, vol.57, no.3, pp.378-393, 2006 https://doi.org/10.1002/asi.20316

[4] Ruan, Guangchen, and Ying Tan. "A three-layer back-propagation neural network for spam detection using artificial immune concentration." Soft computing, vol.14, no.2, pp.139-150, 2010 https://doi.org/10.1007/s00500-009-0440-2

[5] Shih, Dong-Her, Hsiu-Sen Chiang, and C. David Yen. "Classification methods in the detection of new malicious emails." Information Sciences, vol.172, no.1, pp.241-261, 2005 https://doi.org/10.1016/j.ins.2004.06.003

[6] Al-Shboul, Bashar Awad, et al. "Voting-based classification for e-mail spam detection." Journal of ICT Research and Applications, vol.10, no.1, pp.26-42, 2016 https://doi.org/10.1016/j.comnet.2008.11.012

[7] De Vel, Olivier. "Mining e-mail authorship." Proceeding of Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000), 2000 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.6277

[8] Alsmadi, Izzat, and Ikdam Alhami. "Clustering and classification of email contents." Journal of King Saud University-Computer and Information Sciences vol.27, no.1, pp.46-57, 2015 https://doi.org/10.1016/j.jksuci.2014.03.014

[9] Ahmed Abbasi and Hsinchun Chen, "Applying Authorship Analysis to Extremist-Group Web Forum Messages," IEEE Intelligent Systems, vol.20, no.5, pp.67-75, 2005 https://doi.org/10.1109/mis.2005.81

[10] Smutz, Charles, and Angelos Stavrou. "Malicious PDF detection using metadata and structural features." Proceedings of the 28th annual computer security applications conference. ACM, 2012 https://doi.org/10.1145/2420950.2420987

[11] Digital Bread Crumbs, Focusing Seven Clues To Identifying Who's Behind Advanced Cyber Attack, FireEye Report, RPT.DB.EN-US.082014, 2014

[12] https://www.python.org/

[13] http://scikit-learn.org/stable/

[14] K. Bache and M. Lichman, "UCI machine learning repository," 2013.

[15] Vapnik, V., The nature of statistical learning theory. Springer-Verlag New York, 2000

[16] Altman, N. S., "An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician, vol.46, no.3, pp.175 – 185, 1992 https://doi.org/10.2307/2685209

[17] Kamiński, B.; Jakubczyk, M.; Szufel, P. "A framework for sensitivity analysis of decision trees". Central European Journal of Operations Research, 2017 https://doi.org/10.4135/9781412971980.n103

[18] Ho, Tin Kam "Random Decision Forests," Proceedings of the 3rd International Conference on Document Analysis and Recognition, pp. 278 – 282, 1995 https://doi.org/10.1109/icdar.1995.598994

[19] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[20] Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," IEEE Communications Surveys & Tutorials, Vol.16, No.1, pp.303-336, 2014 https://doi.org/10.1109/surv.2013.052213.00046

[21] Rocha, Anderson, et al. "Authorship attribution for social media forensics." IEEE Transactions on Information Forensics and Security, Vol.12, No.1, pp.5-33, 2017 https://doi.org/10.1109/tifs.2016.2603960

[22] Alsulami, Bander, et al. "Source Code Authorship Attribution Using Long Short-Term Memory Based Networks." European Symposium on Research in Computer Security, 2017 https://doi.org/10.1007/978-3-319-66402-6_6

[23] Singh, Shashi Pal, et al. "Intelligent Text Mining Model for English Language Using Deep Neural Network." International Conference on Information and Communication Technology for Intelligent Systems, Springer, 2017 https://doi.org/10.1007/978-3-319-63645-0_54

[24] Hong, Sung-Sam, Jong-Hwan Kong, and Myung-Mook Han. "The Adaptive SPAM Mail Detection System using Clustering based on Text Mining." KSII Transactions on Internet and Information Systems (TIIS), vol.8, no.6, pp.2186-2196, 2014 https://doi.org/10.3837/tiis.2014.06.022

◐ 저 자 소 개 ◑

홍 성 삼(Sung-Sam Hong)
2009년 가천대학교 전자거래학과(공학사)
2011년 가천대학교 일반대학원 전자계산학과(공학석사)
2016년 가천대학교 일반대학원 전자계산학과(공학박사)
2016년~현재 가천대학교 컴퓨터공학학과 연구교수
관심분야 : 정보보호, 인공지능, 데이터 마이닝, 데이터 분석, 지능형 시스템
E-mail : sungsamhong0@gachon.ac.kr

신 건 윤(Gun-Yoon Shin)
2017년 가천대학교 인터랙티브 미디어 융합학과 학사
2017년~현재 가천대학교 일반대학원 IT융합공학과 석사과정
관심분야 : 기계 학습, 악성코드 분석, 공격자 식별
E-mail : bobo7754@naver.com

한 명 묵(Myung-Mook Han)
1980년 연세대학교 공학부(공학사)
1987년 뉴욕공과대학교 대학원 컴퓨터공학과(공학석사)
1997년 오사카시립대학교 대학원 정보공학부(이학박사)
1998년~현재 가천대학교 컴퓨터공학과 교수
관심분야 : 정보보호, 알고리즘, 데이터 마이닝
E-mail : mmhan@gachon.ac.kr