한국컴퓨터정보학회 동계학술대회 논문집 제31권 제1호 (2023. 1)

앙상블 기반의 위조 탐지 알고리즘

타히예프 일킨[O], 조영복[*]
[O]대전대학교 정보보안학과,
[*]대전대학교 정보보안학과
e-mail: taghiyev@edu.dju.ac.kr[O]

Ensemble-based Counterfeit Detection Algorithm

Ilkin Taghiyev[O], Youngbok-Cho[*]
[O]Dept. of Information Security, Daejeon University,
[*]Dept. of Information Security, Daejeon University

● 요 약 ●

본 연구에서는 인터넷 상에서 발생되는 부정행위를 탐지할수 있는 신뢰 모델을 생성하고 개인의 프라이버시를 보장할수 있는 모델을 제시하였다. 인터넷 상에 게시판에 올려진 부정해위를 탐지하기 위해 앙상블 접근 방식 기반의 분류 모델을 제시하고 자동화된 도구를 제안하였다. 본 연구는 데이터에 대한 탐색적 데이터 분석을 수행하고 얻은 통찰력을 사용해 자연어처리 가반 텍스트를 기반으로 앙상블 기반의 위조 탐지 알고리즘을 제안하였다. 제안 알고리즘의 정확도는 99%로 자연어 처리에 높은 탐지율을 보였다.

키워드: 앙상블 접근법, 자연어 처리, 머신러닝, 프라이버시보호

I. Introduction

우리는 전쟁과 코로나19 범유행으로 인해 모든 대륙의 경제에 타격을 입혀 전례 없는 시대에 살고 있다. 실업률이 매일 증가하고 있다. 많은 회사들이 구직자들이 쉽고 시기적절하게 접근할 수 있도록 그들의 빈자리를 온라인에 게시하는 것을 선호하기 때문에 사기꾼들은 이와 같은 상황을 이용하는 것을 좋아한다. 대부분의 사기꾼들은 사이버 범죄에 연루되기 위한 개인정보를 얻거나 그들이 사기를 치려는 사람으로부터 돈을 빼앗기 위해 이것을 한다[1]. 이를 위해 앙상블 기반 머신 러닝 접근 방식은 위조 게시물을 탐지하기 위한 몇 가지 분류 알고리듬을 사용한다. 이 시나리오에서, 분류 기법은 위조 구인 게시물을 더 넓은 구인 게시물 풀과 구별하여 사용자에게 알린다.

온라인 사기 탐지 영역에서 위조 뉴스 식별 및 전자 메일 스팸 탐지가 많은 관심을 받았다. 온라인 사기 탐지 영역에서 위조 뉴스 식별 및 전자 메일 스팸 탐지가 많은 관심을 받았다.

1.1. 위조 뉴스 신원 확인

위조 뉴스 탐지에 대한 주요 연구는 위조 뉴스가 어떻게 작성되는지, 위조뉴스가 어떻게 퍼지는지, 사용자가 위조 뉴스와 어떻게 상호작용하는지 등 세 가지 관점에 달려 있다. 뉴스 콘텐츠 및 소셜 컨텍스트와 관련된 기능을 추출하고 위조 뉴스를 인식하기 위해 머신러닝 모델을 적용된다[2].

1.2 전자 메일 스팸 탐지

많은 연구가 데이터 마이닝을 수행해왔고 지식 검색 및 텍스트 처리 기술을 사용하여 피싱 전자 메일을 찾을 수 있는 자능형 분류 모델을 제공하고 있다[5].

1.3 앙상블 방식

앙상블 방법은 단일 모델을 사용하는 대신 여러 모델을 결합하여 모델에서 결과의 정확도를 향상시키는 것을 목표로 하는 기술이다. 결합된 모델을 사용하면 결과의 정확도가 크게 향상된다. 가장 인기 있는 앙상블 기법은 부스팅, 백킹이다. 앙상블 방법은 편향과 분산을 줄여 모델 정확도를 향상시키는 회귀 및 분류에 이상적이다[3]

II. The Proposed Method

1. 데이터 세트세부 정보

본 연구에 사용된 데이터 세트는Kaggle에서 자유롭게 이용할 수 있는EMSCAD이다. 이Kaggle 데이터 세트는17,880개의 채용

게시 정보 레코드가 포함되어 있다. 데이터 세트은 문자열, HTML 조각, 이진 및 공칭과 같은 다양한 데이터 유형을 결합한다. 데이터 세트는 노이즈가 많기 때문에 기계 학습 모델이나 분류기에 맞추기 전에 먼저 이 데이터를 사전 처리하여 예측을 준비해야 한다[4].

2. 데이터 전처리.

전처리 기법에는 결측값 제거, 중지 단어 제거, 관련 없는 속성 제거, 여백 제거 등이 있다. 그러면 피처 벡터를 얻기 위해 범주형 인코딩을 위한 데이터 세트가 준비된다. 이러한 형상 벡터는 여러 분류자에 맞게 조정된다. 여기서는 예측을 하기 전에NLP(자연 언어 처리)를 사용하여 텍스트 데이터를 숫자 데이터 표현으로 변환했다[4].

3. 특징 추출

특징 추출은 대규모 원자재 세트를 더 작은 그룹으로 나누고 축소하는 차원 축소 프로세스의 단계이다. 이러한 대규모 데이터 집합의 가장 중요한 특징은 변수 수가 많다는 것이다. 이러한 변수를 처리하려면 많은 컴퓨팅 성능이 필요하다. 따라서 변수를 선택하고 특징으로 결합함으로써 특징 추출은 이러한 대규모 데이터 세트에서 최상의 특징를 추출하고 데이터 양을 효과적으로 줄이는 데 도움이 된다[6].

4. 분류기를 구현하기

이 프레임워크에서 분류자는 적절한 매개변수를 사용하여 준비된다. 기본 하이퍼 매개변수는 이러한 모델의 성능을 향상시키기에 충분하지 않을 수 있다. 이러한 매개 변수를 조정하면 이 모델의 신뢰성이 높아지며, 이는 구직자로부터 위조 구인광고를 식별하고 격리하기 위한 최적화된 모델로 간주될 수 있다. 특징을 추출하고 최소화한 후 앙상블 분류를 적용했다.

5. 평가 매개 변수

분류 예측을 수행할 때TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)의 네 가지 유형의 결과가 있다.

정확도(Accuracy)는 우리 모형에 의해 만들어진 정확한 예측의 백분율을 측정하는 것으로 정확도= TP+TN/TP+FP+FN+TN으로 표현한다. 또한 F1 점수(F1 Score)는 모델 정밀도와 리콜의 조화 평균으로 F1 점수= 2*(Recall * Precision) / (Recall + Precision)으로 표현한다.

III. Experiments and Results

위에서 언급된 앙상블 방식는 위조 게시물과 합법적인 게시물을 모두 포함하는 주어진 데이터 세트에서 위조 게시물을 탐지하도록 훈련되고 테스트된다. 다음 표1은 백깅 기반 예측 기법에 대한 결과를 제공한다. 표2는 부스팅 기반 예측 기법에 대한 결과를 제공한다.

Table 1. Performance comparison chart for Bagging-based prediction

Performance Measure Metric	Decision Tree Classifier	Random Forest Classifier
Accuracy	98.31%	99%
F1-Score	76.5%	78%
Best Parameters	'max_depth'=2	'n_estimators':4

Table. 2 Performance comparison chart for Boosting-based prediction

Performance Measure Metric	AdaBoost	GBM
Accuracy	97.42%	97.69%
F1-Score	49.5%	50%
Best Parameters	n_estimators':50	'n_estimators':400

고용 사기 탐지는 구직자들이 회사로부터 합법적인 제안만 받도록 안내할 것이다. 본 논문에서는 고용 사기 탐지를 해결하기 위해 앙상블 접근법 기반 기계 학습 알고리듬을 대응책으로 제안한다. 실험 결과는 Bagging 기반Random Forest Classifier가 다른 분류 도구보다 성능이 우수하다는 것을 나타낸다. 제안된 접근 방식은 기존 방법보다 훨씬 높은99%의 정확도를 달성했다.

REFERENCES

[1] Alghamdi, Bandar, and Fahad Alharby. "An intelligent model for online recruitment fraud detection." Journal of Information Security 10, no. 03 (2019): 155.

[2] Ahmed, Alim Al Ayub, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. "Detecting fake news using machine learning: A systematic literature review." arXiv preprint arXiv:2102.04458 (2021).