한국빅데이터학회지
제4권 제2호, 2019, pp. 23-34

소셜데이터 분석 및 인공지능 알고리즘 기반 범죄 수사 기법 연구

Artificial Intelligence Algorithms, Model-Based Social Data Collection and Content Exploration

안동욱[1]·임춘성[2†]

㈜ 미소정보기술[1], 연세대학교 산업공학과[2]

요 약

최근 디지털 플랫폼을 활용한 민생 위협 범죄는 `15년 약 14만여 건, '16년 약 15만여 건 등 사이버범죄 지속 증가 추이이며 전통적인 수사기법을 통한 온라인 범죄 대응에 한계가 있다고 판단되고 있다. 현행 수기 온라인 검색 및 인지 수사 방식만으로는 빠르게 변화하는 민생 위협 범죄에 능동적으로 대처 할 수 없으며, 소셜 미디어 특성상 불특정 다수에게 게시되는 콘텐츠로 이루어 졌다는 점에서 더욱 어려움을 겪고 있다. 본 연구는 민생 침해 범죄가 발생하는 온라인 미디어의 특성을 고려한 콘텐츠 웹 수집 방식 중 사이트 중심의 수집과 Open API를 통한 방식을 제시한다. 또한 불법콘텐츠의 특성상 신속히 게시되고 삭제되며 신조어, 변조어 등이 다양하고 빠르게 생성되기 때문에 수작업 등록을 통한 사전 기반 형태소 분석으로는 빠른 인지가 어려운 상황이다. 이를 해소 하고자 온라인에서 벌어지는 민생 침해 범죄를 게시하는 불법 콘텐츠를 빠르게 인지하고 대응하기 위한 데이터 전처리인 WPM(Word Piece Model)을 통하여 기존의 사전 기반의 형태소 분석에서 토크나이징 방식을 제시한다. 데이터의 분석은 불법 콘텐츠의 수사를 위한 지도학습 기반의 분류 알고리즘 모델을 활용, 투표 기반(Voting) 앙상블 메소드를 통하여 최적의 정확도를 검증하고 있다.
본 연구에서는 민생경제를 침해하는 범죄를 사전에 인지하기 위하여 불법 다단계에 대한 사례를 중심으로 분류 알고리즘 모델을 활용하고, 소셜 데이터의 수집과 콘텐츠 수사에 대하여 효과적으로 대응하기 위한 실증 연구를 제시하고 있다.

■ 중심어 : AI, 범죄 수사, 데이터수집, 딥러닝

Abstract

Recently, the crime that utilizes the digital platform is continuously increasing. About 140,000 cases occurred in 2015 and about 150,000 cases occurred in 2016. Therefore, it is considered that there is a limit handling those online crimes by old-fashioned investigation techniques. Investigators' manual online search and cognitive investigation methods those are broadly used today are not enough to proactively cope with rapid changing civil crimes. In addition, the characteristics of the content that is posted to unspecified users of social media makes investigations more difficult. This study suggests the site-based collection and the Open API among the content web collection methods considering the characteristics of the online media where the infringement crimes occur. Since illegal content is published and deleted quickly, and new words and alterations are generated quickly and variously, it is difficult to recognize them quickly by dictionary-based

2019년 11월 22일 접수; 2019년 12월 10일 수정본 접수; 2019년 12월 31일 게재 확정.
† 교신저자 (leem@yonsei.ac.kr)

morphological analysis registered manually. In order to solve this problem, we propose a tokenizing method in the existing dictionary-based morphological analysis through WPM (Word Piece Model), which is a data preprocessing method for quick recognizing and responding to illegal contents posting online infringement crimes. In the analysis of data, the optimal precision is verified through the Vote-based ensemble method by utilizing a classification learning model based on supervised learning for the investigation of illegal contents. This study utilizes a sorting algorithm model centering on illegal multilevel business cases to proactively recognize crimes invading the public economy, and presents an empirical study to effectively deal with social data collection and content investigation.

■ Keyword : AI, Content Investigation, Web Crawlers, Deep Running

Ⅰ. 서 론

최근 경제・사회의 양극화 심화로 인하여 중산층 비율을 가구 소득 순을 기준하여 중위 소득의 75~200% 수준인 가구를 중산층으로 잡았을 때 1980년대 중반 64%이던 OECD기준 중산층 비율이 점차 내려가 2010년대 중반에는 61%까지 붕괴되었으며 뿐만 아니라 대출사기, 보이스 피싱 등 서민경제 침해 범죄는 지속적으로 서민 경제에 어려움을 심화시키고 있다. 또한 2018. 12월말 금융거래에서 소외자(신용등급 7등급 이하) 수는 388만 명 정도 되고 있으며 청년들의 실업률(15～29세)은 실업률 전체 비율 중에서 9%대 후반을 보이면서 점차 문제가 발생 되고 있다. 이러한 상황에 민생 경제 침해 범죄가 증가하고 있으며, 민생 범죄의 증가는 또 다른 민생경제를 더욱 악화시키고 있다. 특히 서민과 중산층의 생활 자체를 위협 할 정도이며 붕괴시킬 수 있다. 또한 디지털 플랫폼을 활용한 민생위협 범죄도 증가 추세에 있다. 2015년 약 14만여 건, 2016년 약 15만여 건 등 사이버범죄 지속 증가 추이를 보이고 있으며 취업과 생활이 어려운 서민을 대상으로 하는 불법 대부 및 다단계의 유혹이 상존 하고 있다. 이를 자세히 살펴보면 취업 미끼·고수익 보장 등으로 유인하여 청년층에 피해가 확산되는 등 민생 침해 경제범죄는 나날이 발전 해가고 있다고 할 수 있다. 하지만 전통적인 수사기법을 통한 온라인 범죄 대응에는 한계가 있다. 현행 수기 온라인 검색 및 인지 수사 방식만으로는 빠르게 변화하는 민생 위협 범죄에 능동적으로 대처하기 어려운 것이 현실이며 온라인 미디어의 특성상 빅데이터를 다루는 콘텐츠 수사에서 직접 눈으로 확인해야 하는 수작업 방식은 상당히 비효율적이다.

근래에 경제 범죄에 대한 일반적이고 거시적인 연구는 수행된 적이 있으나, 일반 국민들의 생활과 경제에 직접 침투하는 온라인 민생 범죄의 사전 차단 및 인지 방안에 대한 연구는 수행된 바 없다. 이에 본 연구는 지속적으로 변하는 불법 콘텐츠를 자동 분석 및 분류하여 민생 위협 요소 경감에 기여하고자 한다. 불법 콘텐츠 수사 시 단순 반복 업무에는 인공지능 기술을 활용하여 효율적 개선이 가능하다. 또한 소셜미디어 불법콘텐츠의 특성을 고려한 수집 방법을 연구하고, 불법 콘텐츠에 대한 자연어 처리, 학습, 분석, 분류를 자동화하여 수사 기법 및 역량 향상에 기여하고 실증분석 결과에 따라 불법 콘텐츠 수사의 가치 창출과 경쟁력향상에 기여하고자 한다.

Ⅱ. 이론적 배경 및 관련 연구

2.1 소셜 데이터 수집

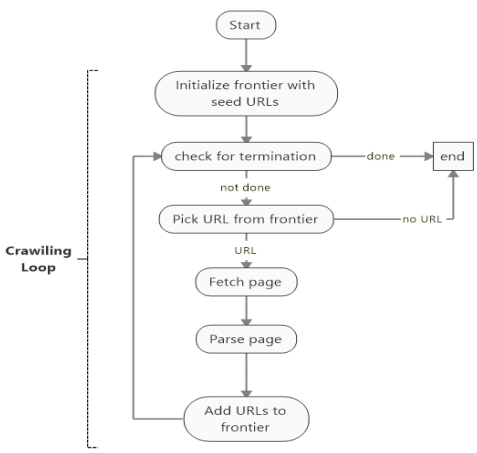
검색엔진의 기초가 되는 웹크롤러는 인터넷상

에 등록되는 문서를 추적하여 필요한 정보를 수집하는 기술이다. 이는 구글, 네이버와 같은 인터넷 검색 사업자 외에도 e-commerce, 상품 리뷰, 브랜드 광고등 대부분 인터넷 산업에 적용되는 핵심 기술이다.

예로는 네이버 카페, 블로그, 다음, 트위터, 페이스북, 인스타그램 등에서 데이터를 추출하고 분석하여 마케팅이나 브랜드 효과, 제품의 설계 등에 반영할 수 있다.

인터넷의 웹문서는 방대한 양의 다양한 자료구조를 가진 빅데이터라는 점에서 자료의 생성과 소멸, 전파 등이 역동적이고 물리적 경계가 낮은 것이 특징이다.

소셜 데이터를 수집하는 웹크롤러는 웹문서가 저장되어 있는 서버를 순환하면서 각 사이트에 있는 대량의 정보를 수집하는 프로그램이다. 이는 개별 사용자가 홈페이지의 각 링크를 설정하여 정보를 획득하는 수작업 기반의 반복 작업을 자동화하는 것으로 사이트의 주소, 패턴, 콘텐츠 형태를 파악한 후 SEED URL을 접속하여 변형된 패턴에 대응 또는 확장하여 정보를 수집하게 된다. <그림 1>은 큐(que)를 사용한 스케줄러 및 멀티쓰레드로 고성능 웹크롤링 방법에 적용하는 것을 보여준다. 그림에서 start는 방문해야 하는 URL (혹은 URL 패턴)들의 seed url 리스트이며, 일반적으로 FIFO(first input first output) 큐로 구성된다. 하나의 URL에 대한 페이지를 가져오면 그 안에 또 다른 웹링크나 패턴의 변경 등이 포함될 수 있으며, 웹크롤러는 링크 페이지도 함께 수집 되어야 한다. 또한 필요한 경우 웹링크나 게시판의 수집 깊이를 제한하여 중복성 데이터를 방지해야 한다.



〈그림 1〉 웹크롤링의 기본 구성

2.2 자연어 처리 모델 연구

2.2.1 자연어

NLP(Natural Language Processing)란 컴퓨터를 이용하여 사람 언어의 이해, 생성 및 분석을 다루는 인공지능 기술을 뜻한다.(한국정보통신기술협회 IT용어사전, 2015).

여기서 자연어(natural language)란 사회가 형성되면서부터 자연스럽게 발전되어 생겨난 의사소통을 행하기 위한 수단 중 하나로 인류가 사용하는 언어를 의미 한다. 컴퓨터 프로그래밍에서 자연어를 위한 특별하게 개발되어진 언어가 인공어(artificial language) 또는 프로그래밍 언어(programming language)로 정의할 수 있다.

근래 자연어 처리는 음성(STT), 콘텐츠의 요약, 언어 간의 번역, 인터넷 문서 또는 글의 감성분석, 텍스트 분석 작업 (스팸 메일 분류, 뉴스 기사, 카테고리 구축), 단순 질의응답 챗봇과 같은 곳에서 사용되고 있으며 근래 딥러닝이 결합되면서 제4차 산업혁명의 중요 기술로 떠오르고 있다. 자연어 처리는 기계에게 인간의 언어를 이해시킨다는 점에서 필수적인 중요한 연구 분야이며 한국어의 특징을 고려한 연구가 계속되어야 하는 것이 특히나 중요한 연구 요소라고 할 수 있다.

2.2.2 형태소 분석

형태소 분석은 일정한 의미를 가진 작은 말의 단위로 문장 내에서 품사에 맞추어 분리하는 과

정을 의미한다. 형태소는 분리 과정 중 가장 하위 단위로 뜻이 없어지는 말의 최소 단위이며 다양한 언어별 의미로 나타내어질 수 있다. 자연어 처리에서 형태소 분석은 문장 어절에서 볼 수 있는 품사를 기준으로 출력하는 것을 의미한다.

이러한 한글 형태소 분석의 세부과정은 먼저 특수문자와 숫자 등을 제거하고 단어를 추출하는 전처리 과정, 품사 기준의 형태소를 분리하는 과정, 접미사 분리 과정, 동사와 형용사를 분리해내는 용언 분석 과정, 명사, 대명사, 수사를 분리해내는 체언 분석 과정을 거치며, 이렇게 분리된 형태소를 기반으로 복합어 추정, 조사 생략, 준말처리 등을 통해 사전에 등록된 단어로 문장이 분석되게 된다. 또한 특수한 단어나 어휘를 위하여 사용자 사전이 등록 되어 최종의 문장에서 분리된 단어로 처리하게 된다[1].

2.2.3 WPM(Word Piece Model)

기존의 자연어 처리에서는 형태소 분석, 품사 분석, 문장 의미 분석 등으로 진행되나, WPM은 음성 검색 시스템 구축을 위한 방법으로 시작하여 언어 또는 비즈니스 용어에 대한 사전에 구축된 지식 없이도 혼잡도(perplexity)를 최소화 시켜 어휘를 생성하는 방법이라 할 수 있다[2].

WPM은 국제 발음기호(the International Phonetic Alphabet, PA)기반 Set으로 유닛을 코드로 변경시킨 후, 통계를 활용한 기법을 사용하여 도출 빈도수에 따라 조합을 해서 신규 유닛을 생성한다. WPM의 장점은 독립적인 언어이며, 통계적인 방식을 사용하므로 특정 비즈니스 또는 도메인에 대하여 아직 의미가 정확히 정의 되지 않은 언어에도 적용할 수 있다.

2.2.4 앙상블 알고리즘

일반적으로 기존의 분류 모델의 정확도가 떨어지는 분류와 예측 시 낮은 성능을 해결하고자 앙상블 알고리즘을 사용한다. 앙상블의 경우 지도학습 영역에서 연구가 시작된 것은 1970년 대 부터이며, 많은 연구와 발전이 진행되었던 때는 1990년대에 진행 되었다. 앙상블 알고리즘은 지도학습 만이 아닌 비지도 학습, 클러스터링에서도 품질 및 성능의 향상에 적용되어 발전 해왔다[3].

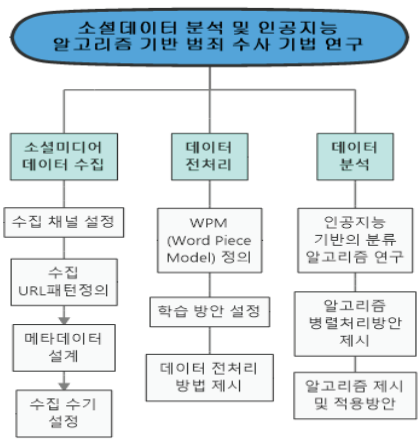
앙상블 알고리즘은 여러 가지의 예측을 하고자 하는 성능이 약한 분류(weak classifiers)모델들에서 생성 및 구성 하는 단계로 부터 시작한다. 우선 학습데이터(training data)에서 랜덤한 샘플 데이터를 통하여 랜덤 서브 데이터 셋(sub data-set)이 추출 및 구성 된다. 이후 추출된 랜덤 서브 데이터 셋에 대하여는 상호 독립적인 다수의 약한 분류 모델들이 생성된다. 생성되는 약한 분류 모델들은 예측 면에서 정확도가 0.5에 근접한 모델들로 되며 개별모델의 예측력이 랜덤 예측보다 더 우수한 수준이 지만 앙상블 알고리즘 내에서는 생성된 다수의 약한 분류 모델들은 알고리즘의 통합 과정을 진행하고 상호 결합하여 더 우수한 예측 성능을 가지는 강한분류 모델로서 진행하게 된다[3].

Ⅲ. 연구 설계 및 연구 방법

본 연구에서는 소셜 데이터 분석 및 인공지능 알고리즘 기반 범죄 수사 기법 연구를 위하여 소셜 미디어의 수집, 데이터 전처리, 데이터 분석방법을 사용하였고, 인공지능 기반의 연구 모형을 만들었다. 상위 계층 모형은 데이터 수집, 전처리, 분석 등 3개의 요인들로 설정하였다.

이에 따른 세부 모형은 소셜 미디어 수집을 위해 콘텐츠 특성을 고려한 채널 선정을 해야 하며, 이 채널들의 url패턴을 인지해야 한다. 이는 전문 수집이 필요한 Scraper 타입과 Open API 타입으로 진행하게 된다. 수집 채널의 패턴을 파악하게 되면 메타데이터 설계를 통해 주기적인 수집을 고려해야 하며 Scraper가 수집하는 사이트의 게시판이나 패턴

url을 인식하게 된다. 또한 Open API의 경우 반복적인 키워드를 호출하여 중복을 제거 하게 된다. 데이터의 전처리는 분석을 위한 1차적인 형태소 분석을 진행하나 일반적인 사전을 활용한 형태소 분석이 아닌 WPM 모델을 사용함으로 새로운 신조어나 변조어에 대한 자동 인지를 강화 하고자 한다. 데이터 전처리 후 분석 시 이러한 형태의 지도학습 데이터를 통해 분류알고리즘의 정확도를 기준하여 분석 방법의 모형을 제시하고자 한다.



〈그림 2〉 소셜데이터 분석 및 인공지능 알고리즘 기반 범죄 수사 기법 연구 모형

3.1 콘텐츠 수사를 위한 데이터 수집 및 처리

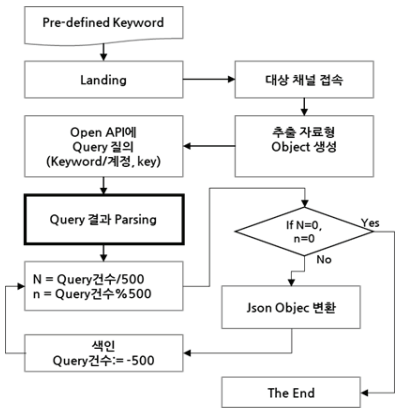
인터넷을 통한 정보 발생은 필요로 하는 정보의 유무가 아닌 어떠한 정보를 사용할지에 대한 선별이나 분류가 더 중요한 문제가 된다. 이를 위하여 검색 엔진 및 서비스가 발전하였고 검색 서비스가 원활하고 정확하게 정보를 제공하기 위해 정보 수집 시스템의 발전도 같이 진행되었다. 사전에 정의한 URL 수집 리스트 정보를 참조하여 해당 사이트의 패턴을 바탕으로 수집하는 방식이 일반적이며 이러한 수집 시 유의할 점은 해당되는 url의 하부 패턴의 범위 설정이다.

불법콘텐츠의 특성상 해당 주소 웹 페이지를 기반으로 관련 링크들을 추출하고 정리한 후 다음 관련 대상 페이지를 호출하게 되며 다음과 같은 문제점이 발생한다. 첫째, 수집 시 많은 광고들로 인한 서버 부하가 발생할 수 있다. 둘째, 반복적 호출 및 검색 시 변조어에 의한 키워드 검색이 안될 수 있다. 셋째, 수사를 피하기 위해 텍스트로 작성된 것이 아닌 이미지를 통한 글을 작성할 수 있다.

따라서 본 연구는 소셜 데이터 분석 및 인공지능 알고리즘 기반 범죄 수사를 하고자 불법 콘텐츠의 등록부터 홍보까지의 특성을 고려한 수집 기술을 연구하기 위한 최적의 방법론을 제시하고자 한다. 불법 콘텐츠의 특성상 일반적인 홍보 보다는 폐쇄성이 존재하는 곳에 일반적 광고를 통해 회원가입 및 소개발표를 하는 것이 특징이다. 본 장에서는 불법콘텐츠가 일반적 광고를 통해 배포되는 사이트를 선정하고 이의 특징을 가진 키워드 검색 후 제공 받는 Open API방식의 수집과 회원가입을 해야 활동할 수 있는 폐쇄성이 존재하는 사이트를 기준으로 Scraper를 통해 특성을 고려한 수집 방법을 고려하게 된다. 불법콘텐츠의 특성상 이미지를 활용한 게시물이 발생하지만 본 연구의 경우 이미지 수집은 고려하지 않으며 텍스트에 준하여 연구하고자 한다.

일반적으로 국내에서 가장 많이 사용하는 포털 사이트인 네이버와 다음, 밴드, 카카오스토리를 기준으로 수집대상을 선정하고 이중 실제 사례가 있을 수 있는 대상을 수집하여 분석을 하게 되며 이에 맞는 최적의 수집 방식과 수집의 주기를 도출한다.

<그림 3>에 따른 진행 절차 중 사전 정의된 키워드를 기준으로 대상 채널 접속, Open API기준 질의를 통해 결과를 받게 된다. 표 1 기준으로 불법콘텐츠에 대한 키워드 수집과 중복 체크, 캐쉬 등의 컬럼을 가진 테이블을 설계하여 진행하였다. 또한 Open API의 경우 인증을 위한 Open API 키를 사용할 수 있게 설계되어 있으며 만약의 키

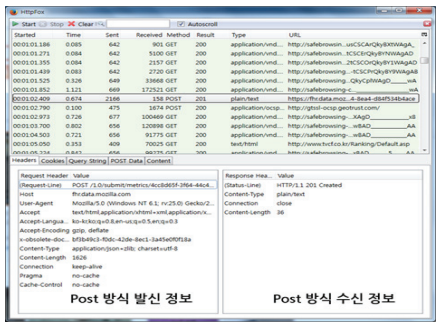


〈그림 3〉 OPEN API 수집 구성

할당의 제한 문제를 방지하기 위해 추가 키를 입력될 수 있게 고려하였다.

3.2.2 Scraper 수집

Scraping은 기존의 Seeding방식인 nutch기반의 수집기에서 수집하기 어렵고 Open API를 제공하지 않는 구조의 사이트를 대상으로 데이터를 수집하고자 할 때 활용하는 방법으로 사이트의 구조와 Logic을 분석하여 맞춤형 Scraper를 개발하여 사용한다. 이는 nutch기반의 수집기에서 판단하기 어려운 문제를 판단하게 되며 다음과 같이 적용한다. 첫째 Data 수집을 위해 자동 Log-In이 필수인 경우,



〈그림 4〉 해당 Site의 동작 Logic을 분석

둘째 JS, AJAX등 동적 웹 기술이 적용된 경우 셋째 Post방식으로 URL을 호출해야 하는 경우 넷째 수집영역이 iframe으로 Nesting된 경우를 고려하여 결정하며 불법 콘텐츠의 특징인 폐쇄성 사이트의 대부분은 이러한 특징을 보이고 있다. 또한 Scraper 개발 시 쓰이는 기술요소는 다음과 같다. 첫째 HTTPPostURL : Post Parameter전달이 필요 한 경우 둘째 HTMLUnit : JS, AJAX등 동적 Paging처리가 필요 한 경우 셋째 Web Driver : Page를 Rendering 해야 할 경우이다

이러한 동작을 감지하기 위해서는 <그림 4>와 같이 브라우져를 통해 간단히 확인할 수 있으며 구조적 판단을 사전에 하고 결정되어야 한다.

Open API의 기능축소/사용제한으로 수집이 어

〈표 1〉 OPEN API 수집 적재 테이블

Table Name	Description
I_keyword	개별 키워드 관리 테이블, Group단위로 runtime에서 호출
I_abuse	불용어 설정. 제목/본문/URL에 불용 Keyword, Pattern 적용
I_context_preg	Naver, daum Portal Crawler의 Parsing Pattern 설정
I_config	Crawling 서버 배정 및 키워드 그룹 배정
I_authentication	Open API 인증키 관리 테이블
I_cache_blog	수집 URL Cache 테이블 최근 6시간 데이터 유지 Daily 1회 전체 삭제
I_cache_café	
I_result_runtime	Naver, Daum Portal Crawler 수행 로그

려워진 Daum Blog, Cafe의 수집은 Open API와 Scraping 형식을 혼합하여 설계한다. 이러한 방식은 다음과 같은 사항을 고려해야 한다.

1. Naver Open API와 동일한 Keyword Set 사용
2. Daum Portal의 경우 Open API가 아닌 검색 서비스를 활용
3. 수집에 필요한 설정 정보를 DB로 관리하여 관리 / 변경 용이
4. 최근 수집시간/수집문서 수를 활용한 키워드 선별 알고리즘 적용
5. Thread를 활용하여 수집 효율 향상 필요
6. 모바일 검색 페이지를 이용하여 속도 및 Parsing 기능 향상

3.3 데이터 전처리

언어 모델(Language Model), 기계 번역(Neural Machine Translation) 등을 사용한 머신 러닝을 이용한 자연어 처리의 목표는 기계가 사람 이상의 성능을 내는 것을 기대하는 것이다. 그런데 아무리 많은 단어들을 기계에게 학습시켜도, 일반적으로 사람이 생성하는 신조어, 단축어 등은 기계에게 모든 단어를 사전에 알려줄 수는 없으며 기계에게 더 많은 단어를 알려주려고 하면, 그만큼 계산 비용도 늘어난다는 부담이 있다.

기계가 훈련 단계에서 학습한 단어들의 집합을 단어 집합(vocabulary)이라고 하며 기계가 암기한 단어들의 리스트라고 정의 할 수 있다. 그리고 테스트 단계에서 기계가 미처 배우지 못한 모르는 단어가 등장한다면 이 단어들을 OOV(Out-Of-Vocabulary)라고 한다. 단어 집합에 없는 단어라는 의미로 이를 UNK(Unknown Word)로 표현하기도 한다. 결국 기계가 모르는 단어로 인해 문제를 풀지 못하는 상황은 OOV 문제이다. 내부 단어 분리(Subword Segmentation)는 기계가 아직 배운 적이 없는 단어더라도 대처할 수 있도록 도와주는 기법으로 이제는 기계 번역 등에서 주요 전처리로 사용되고 있다. 이를 활용하는 것이 WPM으로 WPM은 음성 검색 시스템 구축을 위한 방법으로 많이 사용되고 있으며 자연어 처리에 대한 지식 없어도 혼잡도를 이용하여 단어나 어휘를 생성하는 자동화 방법이다.

본 연구는 불법콘텐츠의 특성상 변조어 및 자주 단어를 교체하여 기존의 자연어 처리(Natural Language Processing)로 진행하기에 많은 량의 사전 어휘관리가 필요한 문제를 해결하기 위하여 WPM을 사용하며 이를 기반 하여 학습을 시키는 연구방법을 선택하였다. 이는 불법 콘텐츠에 대응하기 위한 방안을 제시하고자 한다.

3.3.1 BPE알고리즘

기존의 자연어 처리(Natural Language Processing)는 형태소 분석 엔진에 따라 사전을 사용하게 되고 필요에 의해 사용자 사전을 만들게 된다. 이는 많은 어휘를 추가해야하는 문제가 발생하기 때문에 기본적인 BPE알고리즘 기반의 WPM을 제시하고자 한다. BPE(Byte pair encoding) 알고리즘은 1994년에 제안된 데이터 압축 알고리즘이다. 이후에 자연어 처리의 단어 분리 알고리즘으로 응용되고 있다. 이 방법을 적용하면 문장 구성 시 인식하는 단위를 '단어'로 기준하여 기존의 사전 대비 경량의 사전을 이용하게 된다.

BPE 알고리즘은 단어의 분리(word segmentation) 알고리즘이다. 기존에 있던 단어를 분리한다는 의미인데, 어떤 의미인지는 뒤에서 나올 BPE 알고리즘의 최종 결과를 보면 이해할 수 있다. BPE 알고리즘을 요약하면, 글자(charcter) 단위에서 점차적으로 단어 집합(vocabulary)을 만들어 내는 Bottom up 방식의 접근을 사용한다. 우선 훈련 데이터에 있는 단어들을 모든 글자(Chracters) 또는 유니코드(Unicode) 단위로 단어 집합(vocabulary)를 만들고, 가장 많이 등장하는 유니그램을 하나의 유니그램으로 통합한다.

```
vocab = {'b e s t </w>' : 15,
         'l o w e r </w>' : 14,
         'b e s t e s t </w>': 16,
         'g o o d e s t </w>': 13
         }
```

```
max_num = 15

for i in range(max_num):
    pairs = get_stats(vocab)
    fit = max(pairs, key=pairs.get)
    vocab = merge_vocab(fit, vocab)
    print(fit)
```

```
('e', 's')
('es', 't')
('est', '</w>')
('b', 'est')
('best', 'est</w>')
('b', 'est</w>')
('l', 'o')
('lo', 'w')
('low', 'e')
('lowe', 'r')
('lower', '</w>')
('g', 'o')
('go', 'o')
('goo', 'd')
('good', 'est</w>')
```

〈그림 5〉 Byte Pair Encoding 기법의 Python 동작 예

Byte Pair Encoding은 언어 단위 처리 체계에서 Subword를 기본으로 하는 Segmentation 기법으로, 특정 언어에 종속적인 문법적, 의미적 규칙이 필요하지 않고 오로지 데이터에만 의존하는 학습 방법이다. BPE 방법은 다음과 같은 알고리즘을 통해 진행된다.

1. 먼저 단어를 캐릭터 단위로 다 분할을 한 뒤, 전체 코퍼스에서 등장한 횟수를 기록한다.
2. 등장한 횟수가 가장 많은 공통의 캐릭터 bi-gram을 시작으로 병합하며, 더 이상 공통의 캐릭터 bigram이 존재하지 않을 때까지 BPE 기법을 적용하여 Subword 사전을 저장시켜 나간다.
3. Test time에는 새로 입력된 문장을 마찬가지로 캐릭터 단위로 다 분할을 한 뒤, 사전에 저장되어 있는 Subword를 적용시켜 입력 문장을 Subword 기반으로 분해한다.

<그림 6>는 Byte Pair Encoding 기법이 Python에서 어떻게 동작하는지 예시를 나타낸다. 각각을 먼저 캐릭터 단위로 다 분할을 한다. 그 다음 등장한 횟수가 가장 많은 공통의 캐릭터 bigram을 합쳐 나가기 시작하는데, 맨 처음 등장 횟수가

```
vocab = {'b e s t </w>' : 15,
         'l o w e r </w>' : 14,
         'b e s t e s t </w>': 16,
         'g o o d e s t </w>': 13
         }
```

```
max_num = 15

for i in range(max_num):
    pairs = get_stats(vocab)
    fit = max(pairs, key=pairs.get)
    vocab = merge_vocab(fit, vocab)
    print(fit)
```

```
('e', 's')
('es', 't')
('est', '</w>')
('b', 'est')
('best', 'est</w>')
('b', 'est</w>')
('l', 'o')
('lo', 'w')
('low', 'e')
('lowe', 'r')
('lower', '</w>')
('g', 'o')
('go', 'o')
('goo', 'd')
('good', 'est</w>')
```

〈그림 6〉 Byte Pair Encoding 기법의 Python 동작 예

많은 bigram은('s', 't')라서 이 둘을 먼저 합쳤고, 그 다음은 ('st', '$'), ('e', 'st$'), ('l', 'o') 순으로 합쳐 나간다.

3.3.2 Sentencepiece

SentencePiece는 주로 많이 사용되는 분야가 인공 신경망(Neural Network) 기반의 텍스트 생성에 주로 사용되는 비지도 학습 개념의 텍스트 토큰화 및 해독기이다. SentencePiece는 문장에서 직접 학습으로 확장하여 서브 워드 단위의 BPE 및 유니 그램 언어 모델을 구현하며 이를 사용하면 언어 별 사전 및 사후 처리에 의존하지 않는 순수한 end-to-end 분석을 할 수 있다.

SentencePiece는 sub-word units를 다시 구현 한 것으로 인공 신경망(Neural Network)에서 어휘 문제를 해소하는 것에 효과적인 방법이다. SentencePiece는 BPE 및 유니 그램 언어 모델의 두 가지 세그먼트화 알고리즘을 지원하며 다른 차이점은 구현에 대한 부분으로 다음과 같다.

첫째 고유 토큰의 수는 사전 정의를 하게 된다.

일반적으로 고정 어휘로 작동하고 대량의 어휘를 가정하는 대부분의 비지도학습의 단어 분할 알고리즘과 달리 SentencePiece는 최종 어휘 크기가 고정되도록 (예 : 8k, 16k 또는 32k) 분할 모델을 학습한다.

SentencePiece는 훈련을 위한 최종 단어 크기를 지정하는데, 이는 병합 할 수 있는 연산의 수를 사용하는 subword-nmt와 다르다. 병합 조작의 수는 BPE 특정 매개 변수이며 유니 그램, 단어 및 문자를 포함한 다른 세그먼트에는 적용되지 않는다.

둘째 원시 문장 학습으로 이전의 subword-unit 구현에서는 입력 문장이 사전 토큰화 된 것으로 가정하면 사전에 언어에 의존하여 토크 나이저를 실행해야 하므로 사전 처리가 복잡해진다. SentencePiece의 구현은 원시 문장으로부터 모델을 훈련시키기 때문에 단어 사이에 명확한 공백이 없는 중국어 및 일본어의 토크 나이저를 훈련시키는 데 유용하게 사용된다.

서브 워드 정규화(Subword regularization)는 모델의 정확성과 견고성을 향상시키는데 도움이 되는 온더 플라이 서브 워드 샘플링으로 학습 데이터를 사실상 보강하는 간단한 정규화 방법이며 하위 단어 정규화를 활성화하려면 SentencePiece 라이브러리를 호출하여 (C ++ / Python)를 통합하며 표준 오프라인 데이터 준비와는 다른 각 매개 변수 업데이트에 대해 하나의 세그먼트를 샘플링 하게 된다. 그림 7은 Python 라이브러리의 예로 'New York'은 각 SampleEncode 호출마다 다르게 세그먼트화를 하는 과정을 볼 수 있다.

```
>>> import sentencepiece as spm
>>> s = spm.SentencePieceProcessor()
>>> s.Load('spm.model')
>>> for n in range(5):
...     s.SampleEncodeAsPieces('New York', -1, 0.1)
...
['_', 'N', 'e', 'w', '_York']
['_', 'New', '_York']
['_', 'New', '_Y', 'o', 'r', 'k']
['_', 'New', '_York']
['_', 'New', '_York']
```

〈그림 7〉 Sentencepiece python 예시

3.4 데이터 분석

본 연구에서 제안하는 분석 모델은 기본적으로 분류 모델로써 체계적으로 구성된 데이터 형태의 불법성 콘텐츠를 분리 하였다. WPM을 사용하여 데이터의 학습셋을 기반으로 학습 및 분류를 통하여 정확도 측정 및 검증을 진행하고자 한다. 또한 불법 콘텐츠라는 학습데이터를 기준으로 지도학습을 통한 분류를 해야 한다.

따라서 지도학습에 많이 사용되는 NaiveBayes, RandomForest, ExtraTrees, AdaBoost, GradientBoost 등의 알고리즘을 선정하여 연구를 진행하였다. 각각의 조건 및 변수는 많은 연구과정을 통하여 설정을 조절하여 최적의 값을 찾는 것으로 방법을 수행한다. 수행 시 고려해야 하는 사항은 추가적인 학습 데이터를 확보함으로 분류의 정확도를 향상 시킬 수 있다.

앙상블 메소드는 학습기로부터 얻어진 예측들을 조합하여 voting한 예측을 만드는 메소드이다. 앙상블메소드는 Soft Voting으로 간단히 구현 가능하고 Soft Voting이 Hard Voting보다 합리적인 추론에 도움 되는 경우를 확인 할 수 있다. 이러한 앙상블 메소드는 학습 단계에서 여러 개의 머신러닝 알고리즘 모델에 대하여 학습시킨 후 모델들을 활용하여 신규 데이터에 대해 개별 모델의 예측값을 가지고 다수결 투표를 통해 최종 단계를 예측하는 방식을 말한다. 이러한 분류기는 보통 직접 투표(hard voting) 분류기라고 한다. Soft voting은 Hard Voting과 다르게 확률 기반의 사고를 반영하여 불법콘텐츠를 판단하기 위한 보다 합리적인 추론이 가능하다. 또한 분류기가 다수인 경우에도 특별한 정책 없이 사용 가능하다는 장점이 있다. 일반적으로 앙상블 메소드는 여러 가지의 알고리즘을 이용한

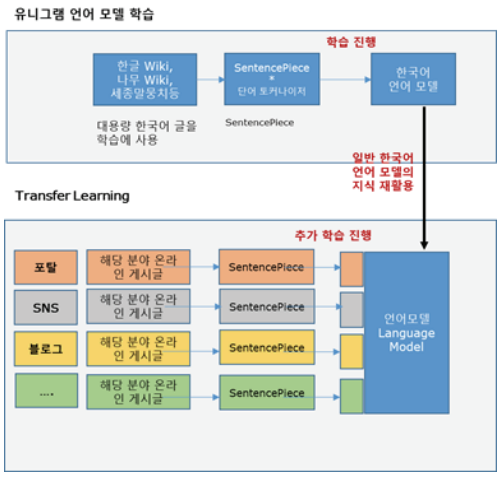
예측 모형을 이용하는 것이 바람직하다. 이는 앙상블 메소드에 적용된 모델 간의 연관성이 적을수록 앙상블 메소드의 효과가 효율적이기 때문이다.

본 논문에서는 불법 콘튼츠의 특성을 감안한 지도학습으로 모형을 만들게 되고 이에 따른 지도학습 분류 알고리즘을 선정하여 앙상블 메소드로 값을 도출하고자 한다.

Ⅳ. 인공지능 알고리즘 관련 콘텐츠 수사 연구

본 논문은 다양한 분야에 포괄적으로 적용가능하나 정확도 판정 및 학습을 위하여 불법 다단계라는 분야를 선정하여 실증을 진행하였다. 불법다단계에 대한 분야를 머신러닝 기술을 활용하여 수집된 콘텐츠의 불법성을 판단·분류(classification)하는 알고리즘을 도출하였으며 소셜데이터의 수집 및 전처리를 통하여 보다 정확한 민생위협 콘텐츠로 판단되는 게시물에 포함된 업체명 등을 추출, DB화 할 수 있는 알고리즘 또한 도출하고자 한다. 이러한 객체인식의 알고리즘은 등록업체인지 불법 다단계업체인지에 대한 판단이 필요하기 때문이며 이를 기반으로 콘텐츠 수사를 진행하게 된다. 이는 수집된 콘텐츠 중 불법행위가 의심되는 게시글을 자동으로 판단할 수 있는 기계학습 알고리즘으로 수사 시 필요한 유의도를 기준으로 추천되어 실제 확인 및 수사를 진행하였다. 불법 다단계가 의심스러운 사이트나 게시글의 소셜데이터 수집 후 판단하기 위해서는 학습단계가 필요하며 학습의 경우 언어모델(Language Model)을 사용하였다.

민생범죄의 불법다단계에 대한 데이터 처리를 위한 학습데이터의 요소는 작성자, 작성자그룹, 콘텐츠를 기초로 학습데이터를 진행하며 sentencepiece 토크나이져를 활용하여 텍스트를 <그림 8>과 같이 유니그램 언어모델 기반으로 형태소로 분리시키며 형태소는 위키피디아, 수집데이



〈그림 8〉 학습 단계 개념도

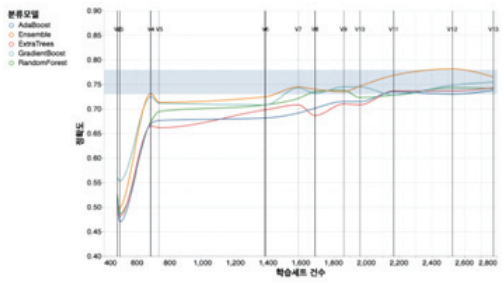
터, 세종말뭉치 등으로 코퍼스기반의 학습을 진행하였다. 또한 불법성이 있는 다단계 확정글에 대한 전문가 태깅 데이터를 12차수 준비하여 각각의 학습단계의 정확도를 향상시켰다.

소셜데이터 수집 및 이미지의 처리를 통한 데이터를 학습 진행 후 앙상블 모델 및 수사유의도 값(Ensemble model and Predictions)을 산출하기 위한 절차는 <그림 9>와 같은 개념도를 통하여 연구하였다.

〈그림 9〉 분류 알고리즘 선정 개념도

앙상블 모델로는 그림 9와 같은 알고리즘을 앙상블메소드를 사용하여 정확도를 산출하였다.

본 연구는 <표 2>와 같이 12차에 걸친 언어모델 학습을 통해 Gradient Boost 등 다양한 알고리즘 적용으로 앙상블 메소드 기준의 정확도 82%값을 제시하였다. 기존 불법콘텐츠를 일일이 검색하여



〈그림 10〉 모델별 분류 정확도

차수	Vaildation Acurracy	F1-score	비고
1 차수	0.50	0.66	
2 차수	0.51	0.67	
3 차수	0.70	0.75	
4 차수	0.74	0.77	
5 차수	0.74	0.78	
6 차수	0.74	0.77	
7 차수	0.75	0.77	
8 차수	0.78	0.79	
9 차수	0.77	0.79	
10 차수	0.78	0.80	
11 차수	0.81	0.82	
12 차수	0.82	0.82	

〈표 2〉 차수별 정확도

확인하는 수작업 방식에서 인공지능 알고리즘을 통해 분류 정확도 높이고 자동 수사 또는 사전인지를 위한 모니터링을 한다면 수사의 효율성을 향상하고 민생 범죄 예방에 기여할 수 있다는 것을 확인 하였다.

Ⅴ. 결 론

본 연구는 인터넷과 소셜 미디어의 발전으로 인해 민생경제에 대한 범죄가 소셜 미디어를 통해 전파되고 모집되며 광고되는 현상을 포착하여 사전 인지 정보화를 연구하였다. 본 논문의 사례인 불법 다단계 콘텐츠의 소셜 미디어 수집 및 전처리 방안을 제시하였으며 이를 분석한 후 분류 알고리즘을 통한 정확도를 산출하였다. 이는 불특정 다수에게 전파되는 불법성 콘텐츠를 수작업으로 찾아서 확인하는 방식을 크게 개선 할 수 있다. 수집 자동화와 알고리즘을 활용해서 방대한 데이터를 효율적으로 다룰 수 있으며 유의성에 대한 82% 정확도를 제시하였다. 이를 활용하여 불법 콘텐츠를 수사 한다면 기존보다 효율적인 수사가 가능할 것으로 판단되며 많은 불법 콘텐츠에 대한 사전 인지 및 불법 콘텐츠의 확산 방지에 기여할 수 있다.

본 논문은 실증 사례로 불법 다단계에 대하여 다루다 보니 많은 불법콘텐츠들이 있는 분야에 대하여 확장을 하지 못하였다. 즉 불법대출, 의약품판매, 방문 판매 등 지속적으로 발생되는 민생침해 범죄에 대하여 추가적인 연구가 필요한 상황이며 분류 알고리즘의 정확도만으로 검증하다 보니 재현율에 대한 초점이 연구되지 못하여 추가적인 연구가 필요하다.

참 고 문 헌

[1] 강정배, “자연어 처리 기술을 활용한 문제행동 유형 분석 연구”, 대구대학원 박사논문, 2012.

[2] Mike Schuster and Kaisuke Nakajima, “JAPANESE AND KOREAN VOICE SEARCH”, Google Inc, USA, 2012.

[3] Rokach, L., 2010, “Ensemble-based classifiers.”, *Artificial Intelligence Review*, vol. 33(1-2), pp.1-39.

[4] Polikar, R. (2006). “Ensemble based systems in decision making”. *IEEE Circuits and Systems Magazine*, 6 (3): 21-45. doi:10.1109/MCAS.2006.1688199.

[5] Rokach, L. (2010). “Ensemble-based classifiers”. *Artificial Intelligence Review*, 33(1-2): 1-39. doi: 10.1007/s10462-009-9124-7.

[6] 이재환, 김보성, 허광호, 고영중, 서정연, *Subword*

유닛을 이용한 영어-한국어, 2009.

[7] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

[8] Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90-94).

[9] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[10] 데이비드 M. 비즐리, *파이썬 완벽 가이드*, 2012.

[11] 황승구, *빅데이터 플랫폼 전략*, 2013.

[12] 김경수, *웹 크롤링 수집주기의 동적 설계 및 구현*, 2011.

[13] 장문수, 정준영, "URL 패턴 스크립트를 이용한 효율적인 웹문서 수집방안", 퍼지 및 지능시스템학회 논문지, 제17권, 제6호, pp.849-854, 2007.

[14] C. Bertoli, V. Vrescenzi, and P. Merialdo, "Crawling Programs for Wraller-based Applications", In Proc. IEEE Intl. *Conference on Information Reuse and Integration (IRI '08)*, pp.160-165, 2008.

[15] M. L. Vidal, A. S. da Silva, E. S. de Moura, and J. M. B. Cavalcanti, "Go GetIt!: a tool for generating structure-driven web crawlers", *InProc. 15th international conference on World Wide Web*, pp.1011-1012, 2006.

저 자 소 개

안 동 욱(Dong-Uk An)

·2016년 : 연세대학교 공학대학원 공학경영전공
·2019년 : 연세대 융합기술경영 공학 박사 과정
·제2회 코리아빅데이터 어워드 최우수기술부문 중소기업청장상 수상
·제12회 신성장 경영대상 우수상 수상
·제3회 코리아빅데이터 어워드 대상 미래창조과학부장관상 수상
·2017 대한민국디지털경영혁신대상 중소벤처기업부장관상 수상
·2019 스마트시티 헬스케어부문 국토교통부장관상
·현재 : ㈜ 미소정보기술 대표이사
·관심분야 : 빅데이터, 데이터 마이닝, 딥러닝, AI, 자연어처리

임 춘 성(Choon Seong Leem)

·1985년 : 서울대학교 산업공학과 (학사)
·1987년 : 서울대학교 산업공학과 (석사)
·1992년 : Univ. of California at Berkeley (박사)
·1993년~1995년 : 미국 Rutgers University 산업공학과 조교수
·현재 : 연세대학교 산업공학과 교수
·관심분야 : 비즈니스 모델(BM) 개발, 신기술 융합서비스 모델 개발, 산업경쟁력 평가개발