한국빅데이터학회지
제7권 제2호, 2022, pp. 1-9 https://doi.org/10.36498/kbigdt.2022.7.2.1

국가별 행정체계 특성을 반영한 인공지능 활용 해외 주소데이터 품질검증 기법

Overseas Address Data Quality Verification Technique using Artificial Intelligence Reflecting the Characteristics of Administrative System

김진실 · 이경희 · 조완섭†

충북대학교 대학원 빅데이터학과, 충북대학교 경영정보학과

요 약

글로벌 시대에 들어서면서 수입식품 안전관리에 대한 중요성이 증가하고 있다. 해외 식품업체 주소정보는 수입식품 안전관리를 위한 핵심 정보로써 식품위해 발생시 신속한 대처와 사후관리를 위해 반드시 검증되어야 한다. 그러나 각국의 주소체계가 다른 관계로 하나의 검증시스템이 모든 국가의 주소를 검증할 수는 없다. 또한, 주소검증은 사용하는 분야에 따라 검정목적이 상이할 수 있다. 본 논문에서는 주어진 해외 식품업체 주소로부터 해당 국가의 행정구역 레벨로 분류하는 문제를 다룬다. 수입식품 안전관리를 정확하고 효율적으로 하기 위하여 수입식품제조업체 주소를 해당 국가의 행정구역 수준으로 정확하게 매칭하는 것이 필요하다. 수입식품이 생산·제조되는 위치와 식품제조에 영향을 줄 수 있는 환경정보, 재난재해 정보를 결합함으로써 선제적 수입식품 안전관리가 가능하다. 그러나, 일부 국가에서는 주소를 표기할 때 행정구역 레벨명을 생략하여 작성하고 있으며, 동일한 지명이 여러 행정구역 레벨에서 중복되는 경우가 있어 주소로부터 행정구역 레벨을 정확히 분류하는 일은 쉽지 않다. 본 연구에서는 이러한 경우에 적합한 딥러닝 기반 행정구역 레벨 분류 모델을 제안하고, 실제 해외 식품회사 주소 데이터에 대하여 검증한다. 구체적으로 다중 레이블 분류 모델에서 멱집합(Label Powerset)을 이용해 훈련하는 방식을 사용한다. 제안된 기법의 검증을 위해 식약처에 등록된 에콰도르 및 베트남에 있는 해외 제조업소 주소에 대하여 정확도를 검증하였으며, 기존의 분류 모델보다 정확도가 각각 28.1% 및 13% 정도 향상되었다.

■ 중심어 : 다중 레이블 분류, 텍스트 분류, 딥러닝, RNN, LSTM

Abstract

In the global era, the importance of imported food safety management is increasing. Address information of overseas food companies is key information for imported food safety management, and must be verified for prompt response and follow-up management in the event of a food risk. However, because each country's address system is different, one verification system cannot verify the addresses of all countries. Also, the purpose of address verification may be different depending on the field used. In this paper, we deal with the problem of classifying a given overseas food business address into the administrative district level of the country. This is because, in the event of harm to imported food, it is necessary to find the administrative district level from the address of the relevant company, and based on this trace the food distribution route or take measures to ban imports. However, in some countries the administrative district level name is omitted from the address, and the same place name is used repeatedly in several administrative district levels, so it is not easy to accurately classify the administrative district level from the address. In this study

2022년 10월 12일 접수; 2022년 12월 07일 수정본 접수; 2022년 12월 14일 게재 확정.

† 교신저자 (wscho@cbnu.ac.kr)

we propose a deep learning-based administrative district level classification model suitable for this case, and verify the actual address data of overseas food companies. Specifically, a method of training using a label powerset in a multi-label classification model is used. To verify the proposed method, the accuracy was verified for the addresses of overseas manufacturing companies in Ecuador and Vietnam registered with the Ministry of Food and Drug Safety, and the accuracy was improved by 28.1% and 13%, respectively, compared to the existing classification model.

■ Keyword : Multi-label Classification, Text Classification, Deep Learning, RNN, LSTM

Ⅰ. 서 론

우리나라는 190여 개 국가, 9만여 개 업체로부터 식품을 수입하고 있으며 해마다 그 양이 증가하고 있다[1]. 수입식품의 안전을 관리하기 위해서는 이들 해외제조업소의 정확한 데이터를 확보하는 것이 중요한 과제이다. 해외 제조업소 주변 및 그 국가에서 발생한 자연재해, 방사능 오염, 환경오염, 원자재 부족 등의 이슈는 우리의 식품안전과 경제에 직・간접적인 영향을 주기 때문이다.

해외식품제조업소에 대한 가장 중요한 정보는 주소 데이터이다. 구체적으로 영업소 명칭, 소재지(주소), 제품명, 제조국 또는 생산국, 제조업소, 작업장 등에 관한 정보로 이루어진 해외 제조업소 주소정보이다. 이는 수입식품 유통 이력 추적을 위한 핵심 정보이며, 수입식품에 안전 문제가 발생한 경우 신속한 대처와 사후관리를 위해서 중요하다. 수입식품의 해외제조업소 주소는 수입신고 시 검증하여 저장 관리해야 하지만 다양한 국가별로 다른 주소체계를 갖고 있어 자동화된 검증에 한계가 있다.

주소검증은 다양한 목적과 방법으로 수행될 수 있지만 본 연구는 수입식품 안전관리 업무에 한정하여 유용한 검증방식을 제안한다. 식품안전 업무에서는 해외제조업소의 주소지를 기반으로 여러 가지 관리업무를 수행하므로 주소의 정확성이 중요하다. 특히, 식품 위해 발생시 해외제조업소 주소로부터 해당 국가의 행정지역명과 GPS 정보 등을 인식하고, 데이터 기반으로 통관검사 항목 보완 또는 검사율 증감 등의 수입식품안전 관리 행정업무를 보완하거나 조정하는 것이 필요하다.

이를 위해 주어진 주소 문자열로부터 주소 구성요소들을 분리한 후, 각 구성요소를 그 나라의 행정구역 레벨로 분류하는 과정(:행정구역 레벨 분류)이 필요하다. 행정구역 체계는 우리나라의 경우 “국가-광역시도-시군구-세부지역” 형식으로 구성되며, 국가별로 차이가 있을 수 있다. 본 연구에서는 행정구역 레벨 분류를 위해 각국의 행정구역 체계를 수집하여 데이터베이스로 구축하였으며(:행정구역 데이터셋), 이를 이용하여 주소를 행정구역 레벨로 분류한다.

Saravit 등(2022)의 연구[2]에서는 중국 등 일부 국가에 대하여 행정구역 데이터셋을 사용하여 분류모형을 생성하고, 이를 사용하여 주소를 행정구역 레벨로 분류하였다. 주소 문자열을 받아서 지오코딩을 거쳐 주소 구성요소들로 분리한 후, 각 주소구성 요소를 해당 국가의 행정구역 레벨로 분류하는 딥러닝 (LSTM) 기반의 모형을 제안하였다. 제안된 기법을 중국 등 일부 국가에 적용한 결과 90% 이상의 정확도를 가지는 것으로 발표하였다.

양광[3]은 중국 소재 해외식품업체 주소를 다양한 방식으로 검증하고, 검증된 주소를 기반으로 식품안전에 필요한 여러 가지 부수적인 정보

(기업정보, 환경정보 등)를 수집하여 연계하는 방안을 제시하였다. 특히, 주소검증을 할 때 영어로 된 주소 문자열뿐 아니라 중국어로 번역한 주소, 영문 및 중국어 회사명, POI 등을 사용하여 더욱 정교하게 검증하여 중국회사의 주소 품질을 개선하였다.

그러나, 기존 분류 모형[2,3]의 문제점은 일부 국가의 경우 (에콰도르, 베트남 등) 주소를 표기할 때 행정구역 단위명(시, 도 등)을 사용하지 않거나 혹은 여러 행정구역 수준에서 동일한 지명을 중복 사용하고 있어 검증의 정확도가 낮아지는 문제가 있다. 표 1은 에콰도르의 해외제조업소 주소의 예시로 동일한 지명(Santo Domingo)이 여러 행정구역 수준에서 중복하여 사용됨을 알 수 있다. 본 연구는 이러한 주소체계를 가지는 국가에 대하여 높은 정확도를 가지는 분류기법을 제안한다.

〈표 1〉 동일 지명이 여러 행정구역 수준에서 중복 사용되는 예시(에콰도르)

Country	Level 1	Level 2	Level 3
Ecuardor	Guayas	Guayaquil	Guayaquil
Ecuardor	Santo Domingo	Santo Domingo	Santo Domingo
Ecuador	Manabí	Manta	Manta
Ecuador	Guayas	Balao	Balao

제안된 기법에서는 지역명과 그에 대한 행정구역 레벨로 구성된 행정구역 데이터셋을 학습 데이터로 사용한다. 학습 데이터는 기존의 주소 데이터와 웹크롤링을 통해 확보한 주소체계 정보를 통합하여 수작업으로 구축한 데이터셋이다. 학습 데이터셋의 클래스 불균형 문제를 고려하고, 자연어 분류에 쉬운 딥러닝 알고리즘인 LSTM을 이용한 다중 레이블 분류 예측 모형을 만들어 주소검증 프로그램을 개발하였다.

제안된 모델의 성능을 평가하기 위해 식약처에 등록된 실제 해외제조업소 주소에 대하여 기존의 Saravit 등(2022)의 분류 모델[2,3]과 비교하는 방식으로 검증을 수행하였다. 에콰도르의 경우 제안된 기법의 분류 예측 정확도는 기존 알고리즘보다 28.1%, 베트남의 경우 기존방식보다 13% 정도 정확도가 개선되었다.

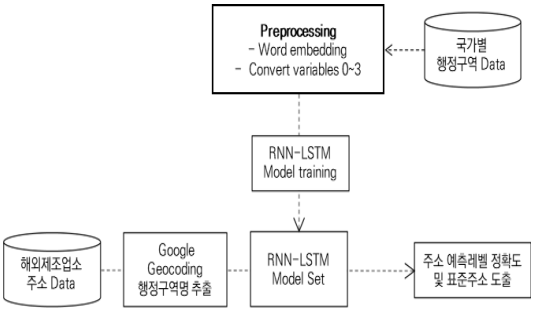
본 논문의 구성은 다음과 같다. 제2장에서 이론적 배경 및 관련 연구를 소개하고, 제3장에서는 연구 방법에 관해 설명한다. 제4장에서는 연구결과를 소개하고 마지막으로 제5장에서 결론 및 한계점을 제시한다.

Ⅱ. 관련연구

본 장에서는 기존의 주소검증 기법들을 살펴본 후, 다중레이블 분류기법에 관해 소개한다.

2.1 주소검증 기법

현대사회에서 주소는 거주지 개념을 넘어 물류, 우편, 전자상거래, 위치기반산업 등 산업 전반과 연결되는 기본요소로서의 역할과 함께 그 활용 범위가 점차 확대되고 있다. 국제 사회는 산업 전반에 걸친 유통체계 비용 절감 등을 위해 주소를 국제표준으로 제정하고 있다. 최근 들어 표준 제정의 범위를 주소의 품질・교환 및 지도 등으로 확대 중에 있으며, IoT 시대를 맞아 사물 주소표준화 등으로 확장해 나가고 있다.



〈그림 1〉 기존 딥러닝을 이용한 주소검증 프로세스

Saravit 등(2022)의 연구[2,3]에서는 딥러닝 기법을 활용하는 주소 정합성 검증 프로그램을 개발하였으며, 검증 프로세스는 <그림 1>과 같다.

먼저 국가별 행정구역 체계 데이터를 학습데이터셋으로 하여 주소 요소로부터 행정구역 레벨을 예측하는 RNN-LSTM 모델을 생성한다. 이후 실제 주소검증 단계에서는 입력된 해외 제조업소 주소 Data를 구글 지오코딩(Google Geocoding)하여 행정구역명을 추출하고, 생성된 RNN-LSTM 모델에 적용하여 예측 행정구역의 정확도 및 표준 주소를 도출한다. 기존 연구에서는 RNN-LSTM 모델에 다중 클래스 분류(Multi-Class Classification)를 구현하여 추출된 각 주소구성 요소에 대하여 총 4개의 클래스('Country', 'Level 1', 'Level 2', 'Level 3') 중에 하나로 분류하였다.

Christen 등(2005)의 연구[4]에서는 호주 국가 우편 주소 지침과 호주 주소 데이터베이스(Geocoded National Address File, G-NAF)를 기반으로 세부 주소 태그를 구축하고, 확률적 은닉 마르코프 모델(HMM)을 사용하여 주소 정리 및 표준화를 위한 자동화된 접근 방식을 제시했다.

Abid 등(2018)의 연구[5]에서는 비표준화된 주소 문제를 해결하고 모든 개체명 인식(Named Entity Recognition, NER) 문제에 적용할 수 있도록 지원하는 딥러닝 기법 Deepparse를 구축했다. 구축된 기법은 비표준화된 주소 데이터에 대해 클래스 혼합 문제와 개체명 인식 문제의 해결 방안을 제시하였다.

민경현 등(2019)의 연구[6]에서는 지오태깅이 되지 않은 텍스트 데이터에서 행정구역이나 기관명, 도서관, 영화관 등이 들어가는 확장된 개념의 장소 정보를 탐지하는 방법을 제안하였다. 뉴스, 기사, 블로그, 소셜미디어 등에서 추출되는 비정형 텍스트 데이터를 가지고 라벨링, 단어 임베딩, 어텐션 기반의 딥러닝 모델을 사용해서 이진 분류기를 만들고 장소 정보의 포함 여부를 예측하였다.

2.2 다중 레이블 분류(Multi-label classification)

다중 레이블 분류는 하나의 데이터를 여러 개의 레이블로 분류하는 기법이다[7-9]. 주소에 포함된 하나의 지명이 두 개 이상의 행정구역 레벨에서 사용된다면 다중 레이블 분류로 모델링하는 것이 적합하다.

다중 레이블 분류 문제를 푸는 방법으로 이진 관련성(Binary Relevance)과 레이블 멱집합(Label Powerset)을 사용하는 기법이 있다. 이진 관련성은 다중 레이블 학습 문제를 참(True)/거짓(False)의 이진 분류 문제로 바꾸어 처리하는 방법으로 분리된 레이블 간에 상호 보완적인 관계가 있을 때 이를 고려하지 못한다는 단점이 있다. 레이블 멱집합 방식은 학습 자료에 나타나는 다중레이블들을 새로운 단일 레이블로 정의하여 다중레이블 분류를 단일 레이블 분류로 변환하여 해결하는 방법이다.

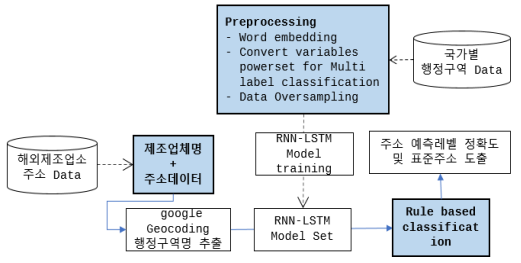
Ⅲ. 연구방법

본 장에서는 주소분류 프로세스를 살펴보고, 단계별로 제안된 기법의 상세 내용을 소개한 후, 정확도 검증결과를 설명한다.

3.1 주소분류 프로세스

본 연구가 제안하는 프로세스는 그림 2와 같다. 가장 큰 특징은 전처리 단계에서 레이블 멱집합 방법을 사용한다는 점과 레이블을 멱집합으로 변형하는 과정에서 발생하는 클래스 불균형 문제를 오버 샘플링으로 해결한다는 점이다. 오버샘플링을 통해 분류기가 모든 지명에 대하여 반드시 한 번씩 훈련할 수 있도록 하였다. 그리고 실제 해외 제조업소 주소 데이터를 사용하는 단계에서 구글 지오코딩 시 좀 더 자세한 값을 반환받기 위해 주소 데이터에 제조 업소명을 함께 입력

한다. RNN-LSTM 모델에서 예측하는 것은 레이블 멱집합의 원소이기 때문에 최종적으로 행정구역 레벨로의 분류를 위해 규칙(Rule)을 기반으로 하는 분류 단계를 추가했다.



〈그림 2〉 본 연구의 주소검증 프로세스

3.2 행정구역 데이터셋

각 국가의 행정구역 데이터셋은 알려진 주소 데이터베이스 및 웹 크롤링을 통해 수작업으로 구성한 것이다, 표 2는 에콰도르의 행정구역 데이터셋 일부를 보여주고 있으며, 학습 데이터는 전체 행정구역 데이터셋을 사용한다.

〈표 2〉 에콰도르 행정구역 데이터셋
(전체 1,332행 중 일부)

iso	country	level1	level1_nm	level2	level2_nm	level3
EC	Ecuador	Morona-	provincia	Santiago	Cantons	Santiago
EC	Ecuador	Morona-	provincia	Santiago	Cantons	Santiago De Mendez
EC	Ecuador	Morona-	provincia	Santiago	Cantons	Tanyuza
EC	Ecuador	Morona-	provincia	Sucua	Cantons	Asuncion
EC	Ecuador	Morona-	provincia	Sucua	Cantons	Huambi
EC	Ecuador	Morona-	provincia	Sucua	Cantons	Logrono
EC	Ecuador	Morona-	provincia	Sucua	Cantons	Santa Marianita De Jesus
EC	Ecuador	Morona-	provincia	Sucua	Cantons	Sucua
EC	Ecuador	Morona-	provincia	Sucua	Cantons	Yaupi
EC	Ecuador	Morona-	provincia	Taisha	Cantons	Huasaga (Cab. En Wampu
EC	Ecuador	Morona-	provincia	Taisha	Cantons	Macuma
EC	Ecuador	Morona-	provincia	Taisha	Cantons	Taisha

표 2에서 보는 바와 같이 에콰도르의 경우 동일한 지명이 여러 행정구역 레벨에서 나타나고 있다. 예를 들어, 표 2의 첫행에서 Santiago는 level 2과 level 3에서 중복하여 나타나고 있으며, 마지막 행에서 Taisha도 마찬가지이다.

베트남도 유사하게 행정구역 데이터셋을 구축하였으며(10,994행), 이를 학습용 데이터로 사용하여 분류 모델을 생성한다.

3.3 연구 내용 및 방법

3.3.1 입·출력 텍스트 데이터 전처리 (Preprocessing)

기존의 알고리즘[2,3]은 출력 데이터로 'Country', 'Level 1', 'Level 2', 'Level 3'의 4개의 클래스 중에 하나로 행정구역 레벨을 예측하였다(다중 클래스 분류). 그러나 일부 국가처럼 동일 지명이 여러 행정구역 레벨에 나타나는 경우 다중 레이블 분류기법을 적용하는 것이 바람직하다. 이를 위해 각 지명에 대하여 출력 데이터를 'Country', 'Level 1', 'Level 2', 'Level 3'으로 원-핫 인코딩하고 이를 조합하여 멱집합을 구성한다.

표 3은 멱집합을 생성하는 과정을 보여주고 있다. Ecuador는 Country, Level3에서 나타나므로 1001 로 인코딩하였다.

〈표 3〉 에콰도르 주소지명에 대한 멱집합 구성예

행정구역명	Country	Level1	Level2	Level3	Power set
Ecuardor	O	X	X	O	1001
Los Ríos	X	O	X	X	0100
Rumiðahui	X	X	O	X	0010
alluriquin	X	X	X	O	0001
pastaza	X	O	O	X	0110
Guayaquil	X	X	O	O	0011
loja	X	O	O	O	0111

구글 지오코딩 결과의 성능을 높이기 위하여, 국가의 특징을 반영한 전처리 과정도 추가하였다. 특히, 스페인 언어 국가의 주소에 사용되는 축약어 리스트를 수집하고 축약어를 모두 완전한

명칭(Full name)으로 변경하였다.

3.3.2 분류 모델 생성

학습데이터로 표 2의 행정구역 데이터셋 전체를 사용한다. 일반적으로 학습 데이터는 7:3 정도로 분할하여 훈련용/검증용 데이터셋을 만든 후, 훈련 데이터셋을 사용하여 분류모형을 생성하며, 검증용 데이터셋을 사용하여 모델의 성능을 평가한다. 그러나 여기서는 표 2의 행정구역 데이터셋이 에콰도르의 행정구역 체계에 대한 전체 집합임을 감안하여 전체를 훈련 데이터셋으로 사용하고, 모델의 성능 평가는 실제 식약처 해외업체 주소 데이터셋을 사용하여 수행한다.

행정구역 분류 모델은 RNN-LSTM 알고리즘을 사용하였다[7-9]. 분류 모델은 RNN 시퀀스에 임베딩 층과 두 개의 LSTM 층 그리고 두 개의 드롭아웃 층과 출력층으로 구성된다. 정수로 인코딩한 입력 데이터는 임베딩 층을 거쳐 글자 단위의 임베딩을 얻고 드롭아웃 층에서는 신경망의 과적합을 방지하기 위해 신경망의 유닛을 무작위로 삭제하여 훈련한다[10]. 드롭아웃 층이 없다면 유닛에 대하여 너무 과도하게 상호작용하기 때문에 과적합의 위험이 있다. 출력층에서는 멱집합 7개의 클래스 중 하나로 출력하고 활성 함수는 소프트맥스 함수를 사용한다. 손실 함수는 카테고리 분류에 사용되는 크로스 엔트로피 함수를 사용하였다.

분류모형은 주어진 주소의 각 요소에 대하여 7개의 멱집합 중 하나로 분류한다. 실제 활용하기 위해서는 각 멱집합을 해당하는 행정구역 레벨(들)로 환산하는 과정이 수반되어야 한다.

Ⅳ. 실험 결과 및 해석

여기서는 주소를 행정구역 레벨로 분류하는 모델의 실험결과를 기존 연구와 비교하여 설명한다. 실험은 수입업자가 식약처에 등록한 에콰도르와 베트남 소재 실제 해외식품제조업소 주소를 대상으로 기존 모델과 제안한 모델의 분류 정확도를 비교한다.

4.1 분류 모델 실험 결과

실험에 사용하는 데이터셋은 2021년 6월 기준으로 식약처에 등록된 에콰도르 및 베트남의 해외식품제조업소 주소 데이터이다. 에콰도르의 경우 300개의 해외식품제조업소가 등록되었으며, 베트남의 경우 3,029개의 해외식품제조업소가 등록되어 있다. 표 4는 식약처에 등록된 에콰도르 해외식품제조업소 주소의 일부를 보여주고 있으며, 베트남도 동일한 형식으로 구성되어 있다.

〈표 4〉 에콰도르 해외제소업소 주소 데이터셋

해외제조업소코드	해외제조업소명	국가	주소	등록시작일	등록종료일
EC000001074	TULICORP.S.A	EC 에콰도르	AV. CARLOS JULIO AROSEMENA KM2 GUAYAQUIL	2017-01-02	2019-01-01
EC000001049	MANYTER C.LTDA	EC 에콰도르	VIA MOCACHE ATRAS DE PICHINGUE	2016-08-04	2018-08-03
EC000001070	CALDERA EXPORT S.A.	EC 에콰도르	RECINTO MARISCAL SUCRE, GUAYAS, ECUADOR	2016-12-20	2018-12-19

4.2 실제 해외업체 주소데이터를 사용한 평가

식품의약품안전처에 등록된 수입식품 해외식품제조업소 주소를 대상으로 제안된 모델의 정확도를 분석한다. 제안하는 주소검증 모델은 동일한 지명이 여러 행정구역 수준에 등장하는 국가인 에콰도르와 베트남를 위한 검증모델이다. 이런 이유로 에콰도르와 베트남 식품업체를 대상으로 주소데이터셋을 확보하고, 이로부터 각 주소요소를 추출한 후, 각 요소에 대한 행정구역 레벨을 분류하고 정확도를 평가한다.

4.2.1 에콰도르의 경우

먼저 표 2와 같은 에콰도르 행정구역 데이터셋(1,332행)으로 분류모형을 만들고, 식약처에 등록된 에콰도르 식품업체 300개의 주소에 대하여 모형을 통해 분류하였다. 제안한 분류 모델을 Saravit 등이 개발한 모형[2,3]과 비교 검증하여 표 5와 같은 결과를 얻었다.

〈표 5〉 에콰도르에 대한 성능 비교

에콰도르 모델 1		에콰도르 모델 2	
정확도	COUNT	정확도	COUNT
50%미만	100	50%미만	53
50%이상 60%미만	15	50%이상 60%미만	10
60%이상 70%미만	170	60%이상 70%미만	0
70%이상 80%미만	15	70%이상 80%미만	27
80%이상 90%미만	0	80%이상 90%미만	0
90%이상 100%미만	0	90%이상 100%미만	196
100%	0	100%	14
합계	300	합계	300

- 국가명을 정확히 분류하면 정확도 10%
- 레벨1을 정확히 분류하면 정확도 30% 추가
- 레벨2를 정확히 분류하면 정확도 30% 추가
- 레벨3을 정확히 분류하면 정확도 30% 추가

여기서 정확도는 다음과 같이 정의한다. 즉, 국가명, 레벨1～레벨3을 모두 정확히 맞추면 정확도가 100%이며, 국가명만 정확히 분류하고 나머지는 모두 틀린 경우에 정확도는 10%로 정해진다. 이러한 정의는 현업의 요구에 맞추어 다르게 정해질 수 있다.

검증에서 주어진 주소의 각 구성요소가 정확히 해당 행정구역 레벨로 분류되었는지의 여부는 표 2의 행정구역 데이터셋을 사용하여 수작업으로 검증하여 판단한다.

기존 모델의 경우 정확도가 50% 미만인 경우가 100건으로 나타났으며(33%), 제안된 모델에서는 53개로 줄어(17%) 분류정확도가 개선되었다. 정확도가 50% 미만이라 함은 국가명과 3개 레벨 중에서 하나 정도를 올바르게 분류한 것을 의미한다. 또한, 기존 모델의 경우 정확도가 90% 이상인 경우가 0건이었지만(0%) 제안된 모델에서는 196건으로 증가하여(65%) 분류가 정확하게 이루어짐을 알 수 있다. 정확도가 90% 이상이라 함은 전체 행정구역 레벨을 정확히 분류하였음을 의미한다. 마지막으로 두 모델의 정확도를 산술평균으로 구하면 기존 모델은 51.6% 이고, 제안된 모델은 79.7% 이므로 약 28.1% 정도 개선되었다.

표 5에서 에콰도르 모델 1을 개선한 모델 2는 에콰도르 주소 300개 중에서 210개의 주소를 90% 이상의 정확도로 분류할 수 있고, 285개를 70% 미만의 정확도로 분류되었던 주소를 63개로 크게 성능 향상되었음을 알 수 있다. 모델 2의 60%이상 70%미만 구간과 80%이상 90%미만 구간에 해당되는 주소의 개수가 0인 것은 모델 1 대비 모델 2가 성능이 향상된 것을 나타낸다.

모델 2에서 정확도 50% 미만으로 분류된 53개의 해외제조업소 주소는 수작으로 검토한 결과 신고업자가 등록한 주소에 에콰도르의 주소 항목이 누락되는 등 오류가 많이 포함되어 있었다.

4.2.2 베트남의 경우

식약처에 등록된 베트남 소재 해외식품제조업소 수는 3,029개이며, 이들에 대하여 표 2와 같은 행정구역 데이터셋(10,994행)을 학습데이터로 사용하여 분류모형을 생성하고 검증하였다. 표 6은 검정결과를 비교한 표이다.

〈표 6〉 베트남에 대한 기존 모델과 개선 모델 비교

베트남 모델 1		베트남 모델 2	
정확도	COUNT	정확도	COUNT
50%미만	712	50%미만	192
50%이상 60%미만	664	50%이상 60%미만	23
60%이상 70%미만	610	60%이상 70%미만	625
70%이상 80%미만	1014	70%이상 80%미만	2081
80%이상 90%미만	8	80%이상 90%미만	11
90%이상 100%미만	21	90%이상 100%미만	48
100%	0	100%	49
합계	3029	합계	3029

표 6에서 기존 알고리즘의 경우 평균 예측 정확도가 50% 미만으로 분류된 경우는 3,029건 중 712건이지만(24%), 제안된 기법은 192건으로(6.33%) 나타나 정확도가 개선됨을 알 수 있었다.

또한, 정확도가 50~60%인 경우도 기존방식에서는 664건이지만(22%), 제안된 기법에서는 23건으로 0.7%로 나타나 개선되었다. 그리고 80% 이상의 평균 예측 정확도를 보이는 주소의 개수는 기존 방식에서는 29건이나 (1%) 제안된 방식에서는 108건으로 증가하였으므로(3.5%) 3배 이상 개선되었다고 볼 수 있다. 산술평균을 구하면 기존 방식의 경우 정확도가 평균 57% 이지만 제안된 기법은 70%로 13% 정도 개선되었다.

Ⅴ. 결론 및 제언

본 논문에서는 주소에 행정구역 단위명을 사용하지 않거나 혹은 하나의 지명이 여러 행정구역 레벨에 나타나는 특이한 주소체계를 가진 국가에 대하여 딥러닝 기반 다중 레이블 분류 예측 모형을 통해 행정구역 레벨 분류의 정확도를 개선하였다. 학습 데이터로는 수작업으로 구축한 각국의 행정구역 데이터셋을 사용하였다. 하나의 지명이 주소체계에서 여러 행정구역 레벨에 나타나므로 다중 레이블 분류 모형을 사용한 것이다.

제안된 기법의 유효성을 검증하기 위해 에콰도르와 베트남 소재 해외식품제조업소의 주소를 대상으로 분류의 정확도를 평가하였다. 제안된 분류 모델의 정확도를 평가한 결과, 기존 분류 모델보다 13% (베트남의 경우)에서 28.1% (에콰도르의 경우)까지 향상되었음을 확인했다.

사 사

본 연구는 2022년도 식품의약품안전처의 연구개발비 (21163MFDS517-1)로 수행되었으며 이에 감사드립니다.

참 고 문 헌

[1] 식약처, https://www.mfds.go.kr/index.do

[2] Soeng, Saravit, Jin-Hyun Bae, Kyung-Hee Lee, and Wan-Sup Cho, “Deep Learning Based Improvement in Overseas Manufacturer Address Quality Using Administrative District Data”, Applied Sciences 12, no. 21: 11129, 2022, https://doi.org/10.3390/app122111129

[3] 양광, 수입식품 안전을 위한 해외기업 정보검증 도구 설계 및 구현, 충북대학교 석사학위논문, 2022.

[4] Peter Christen and Daniel Belacic, “Automated Probabilistic Address Standardisation and Verification”, In Proc. 4th Australasian Data Mining Conference – AusDM05, 2005.

[5] N. Abid, A. ul Hasan and F. Shafait, “DeepParse: A Trainable Postal Address Parser,” 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8, 2018, doi: 10.1109/DICTA.2018.8615844.

[6] 민경현, 송재영, 유기윤, 김지영, “단어 임베딩과 어텐션 기반의 딥러닝 모델을 활용한 장소정보 탐지 기법”. 대한공간정보학회지, 제27권, 5호, pp. 33-39, 2019.

[7] Szymański, P., & Kajdanowicz, T. “scikit-multi-learn: A scikit-based Python environment for performing multi-label classification”. Journal of Machine Learning Research, 20, pp.1-22, 2019.

[8] Zhang, M. L., Li, Y. K., Liu, X. Y., & Geng, X. “Binary Relevance for Multi-Label Learning: An Overview”. Frontiers of Computer Science, 12(2), 191-202, 2018.

[9] Luaces, O., Díez, J., Barranquero, J., del Coz, J. J., & Bahamonde, A., “Binary relevance efficacy for multilabel classification”. Progress in Artificial Intelligence, 1(4), 303-313. 2012.

[10] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. “Dropout: a simple way to prevent neural networks from over-fitting”. The journal of machine learning research, 15(1), pp. 1929-1958., 2014.

저 자 소 개

김 진 실(Jin-Sil Kim)
·2017년 : 청주대학교 광고홍보학과 (학사)
·2020년~현재 :충북대학교 빅데이터협동과정 석사
·관심분야 : 빅데이터, 머신러닝

이 경 희(Kyung-Hee Lee)
·2004년 : 충북대 컴퓨터과학과 (박사)
·2016년~2020년 : 충북대 빅데이터학과 교수
·2020년~2021년 : ㈜힐링소프트
·2022년 현재 : 충북대 경영정보학과 교수
·관심분야 : 빅데이터, 알고리즘

조 완 섭(Wan-Sup Cho)
·1987년: KAIST 전산학과 (박사)
·1996년~현재: 충북대학교 교수
·관심분야: 빅데이터, 블록체인, 빅데이터거버넌스