언어모델 전이학습 기반 해외 직접 구매 상품군 분류

오교중○, 최호진, 차원석, 김일구, 우찬균

한국과학기술원(KAIST), 아일리스프런티어, 통계청 조사시스템과
aomaru@kaist.ac.kr, hojinc@kaist.ac.kr, cos2745@aift.kr 19kim@aift.kr, ckwoo@korea.kr

A Method of Classification of Overseas Direct Purchase Product Groups Based on Transfer Learning

Kyo-Joong Oh○, Ho-Jin Choi, Wonseok Cha, Ilgu Kim, Chankyun Woo
KAIST, Ailys Frontier, Statistics Korea

요 약

본 논문에서는 통계청에서 매월 작성되는 온라인쇼핑동향조사를 위해, 언어모델 전이학습 기반 분류모델 학습 방법론을 이용하여, 관세청 제공 전자상거래 수입 목록통관 자료를 처리하기 위해서 해외 직접 구매 상품군 분류 모델을 구축한다. 최근에 텍스트 분류 태스크에서 많이 이용되는 BERT 기반의 언어모델을 이용하며 기존의 색인어 정보 분석 과정이나 사례사전 구축 등의 중간 단계 없이 해외 직접 판매 및 구매 상품군을 94%라는 높은 예측 정확도로 분류가 가능해짐을 알 수 있다.

주제어: 언어모델, 전이학습, 온라인쇼핑동향조사, 상품군분류

1. 서론

본 논문에서는 통계청에서 매월 작성 및 공표되는 온라인쇼핑동향조사의 관세청 제공 전자상거래 수입 목록통관 자료를 처리하기 위해서, 언어모델 전이학습 기반 분류모델 학습 방법론을 이용하여, 해외 직접 구매 상품군 분류 모델을 구축하는 과정을 설명한다.

'온라인쇼핑동향조사' 는 국내외 쇼핑몰 사업체의 취급 상품과 운영 형태에 대하여 상품 및 서비스 거래액 등의 정보를 제공하는 통계청 작성 통계이다. 이 같은 온라인쇼핑 동향조사 자료는 조사 전월 자료를 바탕으로 약 22일 정도의 조사 및 분석 기간을 거쳐 매월 정부 전산망(인터넷), 간행물, 언론(보도자료) 등을 통해 공표하고 있으며, 정부의 정책수립, 기업의 경영계획 수립, 연구소 및 각종 협회 등에 이용되고 있다.

도소매 서비스를 제공하는 전국 사업체(현재 약 1,100개 사업체, 연간거래액 24백만원 이상, 금융업 제외)를 대상으로 조사가 이루어지며, 해외 직접 판매 통계의 경우, 국내의 사업체가 인터넷 상에서 해외로 상품을 판매하는 사업체의 정보가 주 대상이며, 해외 직접 구매 통계는 관세청 수입통관 자료 중 전자상거래로 통관된 목록통관, 간이 및 일반신고 자료를 활용하여 작성된다.

그림 1 관세청 제공 수입 물품 목록통관 자료 발췌

본 논문에서 사용하는 관세청 전자상거래 수입 목록통관 자료는 통계청 해외 직접구매 통계 작성에 이용되는 관세청 수입통관 자료이며, 데이터의 구성은 다음 그림 1과 같다. 데이터는 주로 통관 심사에 이용되는 신고일자, 가격, 발송국가 등의 정보 뿐만 아니라, 통관목록 품명(텍스트, 영문)과 특송허용품목부호(범주형 정보)로 이루어져 있다.

이 관세청 제공 전자상거래 수입 목록통관 자료는 통계청 조사에서 추가적으로 처리하는데, 2자리로만 제공되는 특송허용품목부호(HS코드의 류에 해당) 정보를 통관목록품명을 바탕으로 10자리 HS코드로 추론하고, 추론된 10자리 코드를 다시 통계청 조사에 사용하는 상품군 분류 체계로 연계하여, 최종 상품군 정보를 얻어낸다. 그 결과를 취합하여 상품군 별 거래액 및 증감 등의 통계 자료를 작성하는데, 그 결과는 다음 그림 2의 예시와 같다. 이 자료를 바탕으로 국내 수입 수출에 온라인쇼핑몰 거래액 정보를 종합하여 공표할 수 있다.

기존에는 자료에 포함된 색인어 지식을 기반으로 검색 방식의 코드 추론을 수행했으며, 매월 평균 200만 건 이상의 자료를 처리한다. 기존의 색인어 기반의 방법을 적용하기 위해서는 분류 항목별 색인어 분석 과정이 필요하여 자료 입력 후 익일에 처리가 가능했으며, 대용량의 데이터 처리를 위해서 3일 정도의 처리시간이 필요했다.

본 논문에서는 지속적으로 확보되는 내검(수정) 데이터만을 활용하여 분류 시스템을 구축하기 위해 기계학습 방법론을 적용하고, 사전학습 언어모델에 기반한 전이학습 방법의 분류 모델을 학습하여 분류 정확도를 최적화하고자 한다. 이 같은 방법론은 텍스트 데이터에 범용적으로 적용될 수 있으며, 특히 영문으로 이루어진 자료에도 적용하여 통계청의 효율적인 자료 처리 및 그림 2와 같은 통계 생산업무에 도움이 되고자 한다.

그림 2 상품군 별 온라인쇼핑 거래액(전년동월비) 증감 [출처: 22년 7월 조사 자료]

2. 관련 연구

2.1 품목분류코드(HS code)

HS코드는 국제통일 상품분류체계에 따라 대외 무역거래 상품을 총괄적으로 분류한 품목 분류 코드(분류 체계)를 말하며, 국제적으로 총 6자리로 구성되어 있으며 우리나라는 물품에 따른 관세율 적용을 위해 4자리를 추가(HSK 코드로 정의)하여 사용하고 있다.

이 코드는 국제 "HS협약"에 의해 일반적으로 5년을 주기로 신상품 출현, 최신 기술분야, 수송기기 추가 등의 이유로 개정을 하고 있으며, 우리나라 또한 개정 주기에 맞춰 22년부터 적용하고 있다. 최근 22년 1월 개정으로 HSK상 품목이 12,242개에서 11,293개(신설 341개 삭제 1,290개)로 변경되어 이에 맞춰 기존 시스템도 대응이 필요하다.

HS 코드는 다음 그림 3의 예시와 같이 류, 호, 소호, 기타 코드와 같이 계층형으로 이루어져 있다.

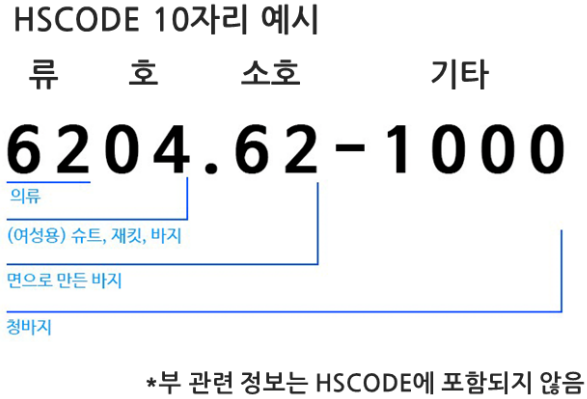

그림 3 HS코드 설명 및 예시 [출처: tradlinx.com]

2.2 색인어 기반 HS코드 분류 시스템

앞서 설명한 바와 같이 현재까지 해외 온라인쇼핑 수출입 통계작성을 위해, 관세청의 전자상거래 수입 목록통관 자료를 상품분류 정보로 처리하는데, 기존에 색인어 정보에 기반한 검색 기반의 코드 추론 방법[1, 2]을 적용하여 통계청에서 구축한 시스템을 이용한다.

관세청 제공 자료의 상품군을 분류하기 위해, 품명에 해당하는 텍스트 정보에서 미리 정의된 색인어 기반의 코드 추론 방법을 이용하여 HS코드로 재분류하며, 이와 연계된 상품군 매칭 결과를 데이터베이스에 적재하는 시스템을 만들었으며, 이를 통해 2016년부터 목록통관 상품군, 지역별 분류, 거래액 산출에 이용하고 있다.

기존에 구축한 분류 시스템은 검색엔진 기반의 텍스트 입력 정보(상품명 파싱 및 형태소 분석)를 이용하는 방법론을 사용한다. 각 상품군 분류 항목에 해당하는 색인어 정보를 추출하여 HS 코드 추론에 활용하였으며, 품명 및 HS코드 매칭 정확도를 약 85% 수준으로 구축하였다. 매월 평균 200만 건 이상의 대용량 텍스트 자료 처리를 위해서 색인어 분석 과정 생략, 내검 자료를 통한 지속적인 정확도 고도화, GPU 장비를 통한 배치처리 업무 효율 제고하기 위해 딥러닝 기반의 기계학습 방법론이 요구되었다.

3. 데이터 분석

3.1 학습 및 검증 데이터 구성

모델 학습에 이용되는 학습 데이터는 그림 4와 같다. 2022년 4월 자료(총 2,219,613 중, 중복데이터 제거 후 661,517 건)에 대해서 학습 및 평가를 진행했으며, 입력에 해당하는 품명 텍스트의 입력 길이는 평균 64개의 토큰 길이를 보임(전체 데이터의 95%에 해당)을 알 수 있었다. 모델 학습 시 오버피팅 및 최적의 모델을 찾기 위해서 검증 셋을 분리(전체 학습 데이터의 10%)하였으며, 편향 정보를 최소화하기 위해서 층화 추출(stratify) 방법을 적용하였다.

입력 항목과 출력항목은 그림 4와 같은데 최종 결과로 제공해야 하는 상품군 분류 정보는 최대 2자리의 코드(최대 24항목)로 이루어진다 [그림 2, 4].

입력	10자리		2자리
상품	HS	년월	상품군
JACKET	6211499000	202204	7
JACKET	6211499000	202204	7
JACKET	6211499000	202204	7
JACKET	6211499000	202204	7
JACKET	6211499000	202204	7
...	...	...	...
CLOTHING JOHN SMEDLEY JUMPERS GENDER 100% VIRG...	6109101000	202204	7
FOOTWEAR CULT SANDALS GENDER SOFT LEATHER RUBB...	6404199000	202204	7
FOOTWEAR ADIDAS ORIGINALS TRAINERS GENDER SOFT...	6404199000	202204	7
CLOTHING MAJE MIDI SKIRTS GENDER 100% VISCOSE ...	6205200000	202204	7
CLOTHING MARNI KIDS DRESSES GENDER 50POLYESTER...	6205200000	202204	7

그림 4 학습 및 검증 데이터 구성 예시

그림 5 온라인쇼핑동향과 해외 판매/구매 상품군 분류 항목 간 연계 및 비교

3.2 상품군 분류 체계

특히 상품 및 서비스 거래액 세부내역 통계자료 작성에 이용되는 분류는 통계청에서 지정한 23개(기타 제외) 상품군 별 온라인상품(모바일 포함) 및 서비스 거래액을 토대로, 9개 국가(대륙) 및 14개(기타 제외) 상품군에 대해 해외 직접 판매 및 구매 상품군을 다시 정의하여 작성된다. 온라인쇼핑 상품군 분류와 해외 직접 판매 및 구매 상품군의 연계 관계는 그림 5와 같다.

22년 4월 수입목록통관 자료를 분석해 보았을 때 [그림 6 참고], 해외 직접 구매 상품군의 상위 5개의 분류 항목이 전체 자료의 89%를 차지하는 매우 편향된 데이터 분포를 보임을 알 수 있었으며, 의류 및 패션 관련 상품(1,320,852건), 화장품(264,446건), 생활/자동차용품(190,220건), 통신기기(131,122건), 기타(81,713건)의 순이었다.

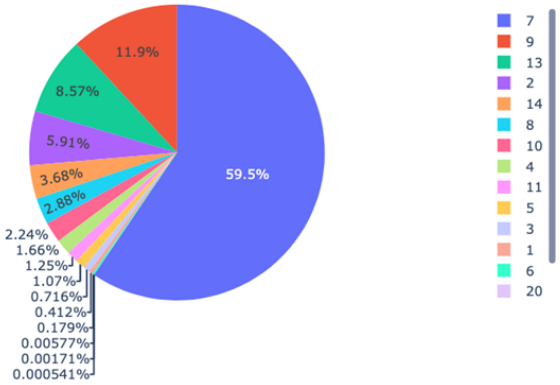

그림 6 수입목록통관 자료 상품군 별 데이터 분포

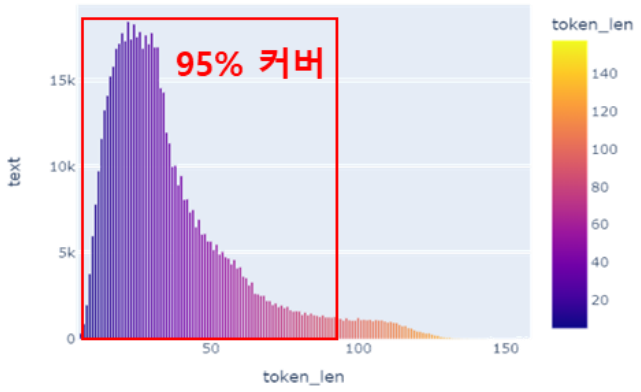

그림 7 입력 텍스트 데이터 길이 분포

4. 분류모델 지도학습 방법 및 결과 분석

4.1 적용 언어모델 및 전처리 과정

본 논문에 사용한 실험 데이터는 관세청 전자상거래 수입 물품목록 통관 데이터로 입력이 주로 영문으로 이루어져 있으며, 따라서 입력 임베딩을 위한 언어모델로 XLM-RoBERTa-large[3]와 DistilBERT[4]를 이용하였다. 이 언어모델은 BERT 모델의 일종으로, 다국어 말뭉치로 학습하여 250,002개의 토큰 어휘로 구성되어 한국어를 포함 영어 텍스트 입력에도 같은 모델로 입력 임베딩을 수행할 수 있다[5]. 본 논문의 학습 및 실험에서는 그림 7과 같이 95%의 품명 입력에 해당하는 정보의 평균 길이 64개로 설정하고 입력 배치를 만들었다.

4.2 학습 및 검증 데이터 구분

학습 시 오버피팅 현상을 방지하고 최적의 학습 결과를 얻기 위해서 편향된 데이터의 분포를 반영하여 학습 데이터 셋을 구축하고자, 데이터를 중복 및 편향되지 않도록 층으로 나눈 다음 각 층에서 표본을 추출하는 층화기법(Stratified sampling)을 사용하여 검증데이터를 추출하고 다음 2단계에 걸쳐 각 모델을 학습하였다.

4.1 HS코드 예측

우선 기존 연구 및 구축 시스템의 결과를 신뢰한다는 가설에 따라 HS코드(10자리, 총 2,654 항목) 추론 결과를 학습하는 모델을 구축해 보았다. 학습 데이터 1,776,182 건, 검증 데이터 443,431 건으로 학습 셋을 구성하였으며, 분류를 위한 파인튜닝 모델로는 언어모델과 함께 구현된 SequenceClassification 모델을 이용하였으며, Nvidia Geforce RTX 3090 환경에서 batch size는 128개, epoch 8번으로 학습을 수행하였다. 학습 결과는 다음 그림 8과 같다. 최적 학습 스텝에서 학습 에러율은 0.27, 검증 에러율은 0.42로 측정되었다.

4.3 상품군 분류

두번째 단계로, HS코드 예측 결과와 해외 직접 구매 상품군 분류 정확도 차이를 비교해 보기 위해 통계청 공표 정보인 상품군 분류 정보(총 16종, 주류, 담배 추가)를 정답으로 이전 학습 과정과 마찬가지로 Sequence-Classification 모델을 이용하여 그림 9와 같이 분류 모델을 학습해 보았다. 품명 텍스트 정보 외에 HS코드에서 류 정보에 해당하는 2자리 범주형 정보를 입력으로 사용하였으며, 16항목의 상품군 정보를 출력으로 사용하여, 중복 데이터를 제거한 학습용 529,215건, 검증용 132,303건 데이터셋으로 구성하였다. 위에서 구축한 같은 환경에서 epoch 5에서 최적 모델이 학습되었다. 학습과 검증 셋 모두에서 에러율 0.17이 나오는 것을 확인하였다.

그림 8과 9의 그래프는 각 분류 모델을 학습한 과정을 보여주고 있으며, X축은 학습 단계(Step)을 말하며, 전체 데이터수×반복 횟수(epoch)를 배치수(128)로 나눈 값이 총 단계(step) 수가 된다. 각 그래프의 Y늑은 각각의 예측 정확도(Accuracy), F-1 점수, 학습 및 검증 과정에서의 에러율을 나타낸다.

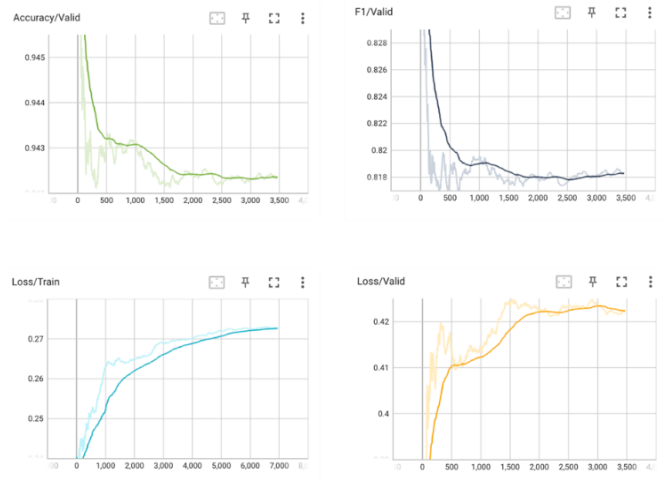

그림 8 HS 코드 예측 모델 학습 결과

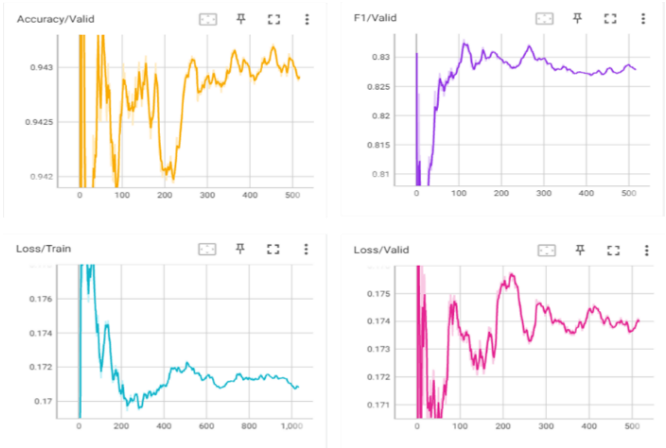

그림 9 상품군 분류 모델 학습 결과

표 1 학습 분류 모델 평가 결과

사용 언어모델	분류 코드	Accuracy (%)	F-1 (weighted)	F-1 (Macro)
XLM-RoBERTa-Large [3]	HSCODE	94.23	0.938	0.829
	상품군	94.48	0.942	0.834
DistilBERT [4]	HSCODE	93.64	0.929	0.816
	상품군	93.88	0.931	0.818

4.4 분류 결과 분석

HS 코드 및 상품군 분류 양쪽 모두 예측 정확도는 94%, F1-score는 83%가 나왔다.

정확도가 높은 항목은 표 2에서 알 수 있듯이 학습 및 검증 데이터의 수가 많은 상품군 7번은 의류-패션, 상품군 11번 음식료품, 상품군 4번 서적 순으로 최소 1,600개 이상의 평가 데이터가 존재했다.

정확도가 낮은 항목은 상품군 12번은 농축수산물, 상품군 4번 소프트웨어, 상품군 7번 음반, 비디오에 해당하는 데이터이며, 평가 셋 기준 평균 1,000개 이하의 항목에 해당하는 항목이었다. 정확도가 떨어지는 항목의 입력 데이터를 분석(4번 소프트웨어)해 보았을 때, 골프채, 보석 상자, 음반 등 다른 항목의 데이터들이 혼재하여 학습 및 평가 데이터에서 발견되었으며, 이 같은 데이터 들은 추후 자료 내검 과정을 통해 정제되어야 할 데이터로 보인다.

표 2 상품군 분류 모델 평가 결과

상품군	Precision	Recall	F1-Score	Support
7 (의류/패션)	0.998	0.998	0.998	75,528
11 (음식료품)	0.996	0.987	0.992	1,604
4 (서적)	0.979	0.999	0.989	2,834
…	…	…	…	…
3 (소프트웨어)	0.653	0.487	0.558	1,448
6 (음반, 비디오, 악기)	0.573	0.592	0.582	475
12 (농축수산물)	0	0	0	6
평균/합계	**0.936**	**0.737**	**0.825**	**661,517**

이 결과로 알 수 있는 결론은 HS코드의 예측 결과와 연계하여 상품군 분류가 이루어진다는 점이었으며, 기존의 색인어 기반의 분류 시스템의 결과도 높은 정확도로 신뢰할 수 있다는 점이었다.

5. 결론

본 논문의 실험에 사용된 데이터는 통계청 서비스업동향과와 조사 시스템과의 협조를 통해 학습 데이터를 확보할 수 있었으며, 최근에 자연어 처리 분야의 텍스트 분류 태스크에서 많이 이용되는 BERT 기반의 언어모델을 이용하여 전이학습 방법으로 분류 모델을 구축하였을 때, 기존의 색인어 정보 분석 과정이나 사례사전 구축 등의 중간 과정 없이 관세청 제공 자료로부터 해외 직접 판매 및 구매 상품군 분류가 가능해짐을 알 수 있었다.

후속 실험으로 다양한 언어모델과 파인튜닝 모델을 이용하여 비교 실험을 진행할 예정이며, 추가적인 자질로 활용 가능성이 있는 입력(예를 들어 수입 국가, 금액 등) 항목 분석, 이전 분류 체계에 따른 데이터 분포의 편향 등을 보완하여 자료 분포가 많은 항목의 자료는 세분화하는 등의 학습 및 평가 데이터 전처리 및 정제 과정이 추가된다면 더 좋은 분류 정확도 성능을 얻을 수 있을 것으로 기대된다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 1711159666, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

참고문헌

[1] 임희석, "예제기반의 학습을 이용한 한국어 표준 산업/직업 자동 코딩 시스템", 한국콘텐츠학회논문지, 제5권, 제4호, pp. 169-179, 2005.

[2] Y. Jung, J. Ryu, S.-H. Myaeng, and D.-C. Han, “A web based automated system for industry and occupation coding,” The 9th International Conference on Web Information Systems Engineering, pp. 443-457, 2008.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, DOI:10.18653/v1/2020.acl-main.747, Jan. 2020.

[4] V. Sanh, el al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019, arXiv:1910.01108, Feb. 2020.

[5] 오교중, 최호진, 안현각, "기계학습 기반 단문에서의 문장 분류 방법을 이용한 한국표준산업분류", 제32회 한글 및 한국어 정보처리 학술발표 논문집, 2020.