기업 비지니스 분석을 위한 한국표준산업코드 앙상블 분류

오교중○, 최호진, 김진원, 차원석, 김일구

한국과학기술원(KAIST), 아일리스프런티어
aomaru@kaist.ac.kr, hojinc@kaist.ac.kr, jw0831@aift.kr, cos2745@aift.kr, 19kim@aift.kr

A Ensemble Classification Method of Korean Standard Industry Code for Corporate Business Analysis

Kyo-Joong Oh○, Ho-Jin Choi, Jinwon Kim, Wonseok Cha, Ilgu Kim
KAIST, Ailys Frontier

요 약

본 논문에서는 기업 비즈니스 분석을 위해 한국표준산업분류에 근거하여 국내 사업체의 산업군을 분류하는 앙상블 분류 모델 구축 방법론을 제시한다. 기업 평가 및 보고서 자동화 시스템 구축을 위해 기업의 재무제표 정보, 기업등록부와 같은 신고 정보, 사업체 조사 정보에 포함된 텍스트 정보를 이용하여, 각 기업이 속해 있는 산업군 정보를 분석해야 하며, 이를 통해 동일한 산업군에 속해 있는 다른 기업에 대한 현황 파악 및 비교 등 비즈니스 정보를 분석할 수 있다.

주제어: 텍스트분류, 앙상블모델, 산업군분류, 사업체분석

1. 서론

최근 기업에 대한 신용평가, 투자분석 등의 목적을 위해 로봇 프로세스 자동화의 일환으로 기업 보고서 자동화 시스템이 구축되고 있다. 이 같은 기업 평가 및 보고서 자동화 시스템 구축을 위해서는 기업의 재무제표 정보, 기업등록부와 같은 신고 정보, 사업체 조사 정보에 포함된 텍스트 정보를 이용하여, 해당 기업이 속해 있는 산업군 정보를 분석해야 하며, 이를 통해 동일한 산업군에 속해 있는 다른 기업에 대한 현황 파악 및 비교 등 비즈니스 정보를 분석할 수 있다.

이 같은 분석에서는 기업의 업종을 구분하기 위해서 일반적으로 한국표준산업분류를 기준으로 활용한다. 이 표준분류는 법정 분류체계로써, 국제표준(유엔 통계처 UNSD)에 의거하여 한국의 실정에 맞게 수정 보완되어, 최대 5자리, 대(21 항목)-중(77 항목)-소(232 항목)-세(495 항목)-세세분류(1195 항목)로 이루어진 계층형 분류 코드 정보를 포함한다. 7년을 주기로 개정되며, 현재 제10차 개정분류를 고시(2017년)하고 있다.

이 같은 산업군 분류 정보는 과세, 규제, 정책 지원 등의 업무에 각종 법적 기준으로 적용된다. 그러나 민원 등의 이유로 정부 기관 및 지방자치단체에서 각 분류 업무에 활용할 수 있도록 분류 체계에 대한 해설 및 사례 정보만 제공하고 있을 뿐, 일관성 있는 분류 정보를 제공하는 서비스나 시스템이 구축되지 않고 있는 실정이다.

본 논문에서는 기계학습 기반의 텍스트 분류 방법을 적용하여, 2021년 경제총조사 샘플링 자료 (2022년 인공지능 기반 통계데이터 활용 경진대회 제공)를 이용하여, 여러 산업군 분류 모델을 구축하고, 이를 앙상블로 이용했을 때 어떤 결과가 나오는지 비교 및 분석한다.

2. 관련 연구

2.1 전국사업체조사와 한국표준산업분류

통계청에서는 94년 이후 매년 우리나라 사업체에 대한 구조를 파악하기 위해 전수 통계조사 자료로 전국사업체 조사를 실시 중에 있다. 이 조사는 국내에서 산업을 수행하고 있는 모든 사업체에 약 659만여 개에 대해 전수조사를 수행하며, 가구 내에서 산업활동을 하는 사업체도 포함된다. 이 자료는 국가 및 지방자치 단체의 정책 수립을 위한 기초자료, 각종 경제통계 조사의 명부로 활용된다.

조사 항목은 주로 조직형태 사업의 종류, 제공하는 상품 및 서비스 등에 해당하는 인터뷰 정보이며, 그 외에도 종사자 수, 매출액 등 10여 종의 데이터를 수집한다. 이중에서 일부 조사 자료는 각 기업에서 직접 제공 및 신고하는 행정 자료도 함께 활용한다.

분류 체계는 대(21개), 중(77개), 소(232개), 세(495개), 세세분류(1,196)까지 총 5단계의 계층형 구조를 통해 분류를 하고 있으며, 통계청 인구총조사, 경제총조사, 지역별고용조사의 경우 세분류(4자리)까지, 전국사업체 조사의 경우 세세분류(5자리)까지 통계조사 별 목적과 범위에 맞춰 적용하여 통계자료 공표에 이용하고 있다.

2.2 기계학습 기반 한국표준산업분류

선행연구 [1]에서는 인구총조사와 지역별고용조사와 같은 가구조사 자료를 이용하여 직장 정보, 업무, 직급, 부서명 등 사용자의 다양한 조사 입력 자료를 한국표준산업/직업분류에 맞춰 분류 코드 정보를 제공해주는 모델을 구축하였다.

기존의 사례사전과 색인어 기반의 분류시스템[2, 3]과 별도로 지도학습 기반의 텍스트 분류 모델을 구축하였으며, 세분류(495 항목) 기준 0.89의 분류 정확도(F-1점수)를 보이는 분류 결과를 얻을 수 있었다.

3. 사용 학습 및 평가 데이터

3.1 학습 및 검증 데이터 구성

학습 데이터는 100만개, 평가 데이터는 10만개가 제공되었으며, 평가 데이터는 통계청에서 공개하지 않고, 시스템에 예측 결과를 올리면 예측 정확도와 F-1 점수를 익일에 제공하였다. 소분류(3자리, 495항목) 분류 결과가 제공되었으며, 각 데이터셋에 대하여 분포를 분석한 결과는 다음 그림 1과 같다.

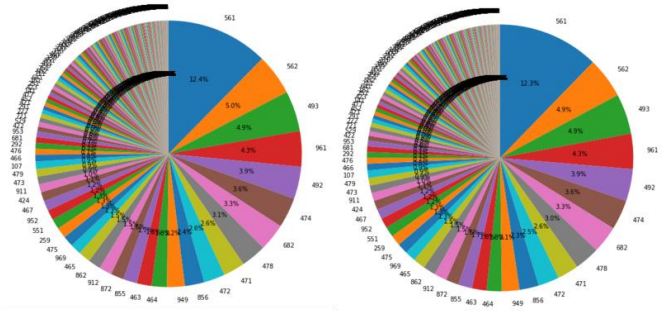

그림 1 학습 및 검증 데이터 분포 비교

학습셋과 검증셋의 분류 항목 별 분류 결과 분포(비율)는 일치했으며 단지 개수만 다를 뿐 비율은 거의 일치하며 전체 데이터의 분포를 정규 분포로 잘 따라 간다는 것을 확인하였다.

3.2 데이터 정제 및 전처리

통계청에서 제공된 학습 데이터는 다음 그림 2와 같다. 입력 항목(변수)는 총 3종류의 항목 정보(목적, 방법, 판매 및 서비스 상품)로 모두 자연어 텍스트 형식이다. 불필요한 특수 문자와 오탈자들이 일부 포함 것을 확인하였으며, 누락 데이터와 중복 데이터는 제거하고, 단일 문자 포함, 정규 표현식 기반의 패턴 정보를 키워드로 치환하는 전처리 과정을 수행하였다. 이 항목 정보들을 띄어쓰기로 구분하여 하나의 입력 문장 형태로 구성하였다, 결과 코드는 3자리 코드를 사용하였으며, 앞의 2자리가 대분류 알파벳 분류 코드에 대응된다.

	AI_id	digit_1	digit_2	digit_3	text_obj	text_mthd	text_deal	X	Y
0	id_0000001	S	95	952	카센터에서	자동차부분정비	타이어오일교환	카센터에서 자동차부분정비 타이어오일교환	952
1	id_0000002	G	47	472	상점내에서	일반인을 대상으로	재소,과일판매	상점내에서 일반인을 대상으로 재소,과일판매	472
2	id_0000003	G	46	467	절단하여사업체에도매	공업용고무등가지고	합성고무도매	절단하여사업체에도매 공업용고무등가지고 합성고무도매	467
3	id_0000004	G	47	475	영업점에서	일반소비자에게	열쇠장금장치	영업점에서 일반소비자에게 열쇠장금장치	475
4	id_0000005	Q	87	872	어린이집	보호자의 위탁을 받아	취학전아동보육	어린이집 보호자의 위탁을 받아 취학전아동보육	872
5	id_0000006	C	29	291	칠	절삭,용접	카프라배관자재	칠 절삭,용접 카프라배관자재	291
6	id_0000007	I	56	561	음식점에서	접객시설을 갖추고	참치회(일본식)	음식점에서 접객시설을 갖추고 참치회(일본식)	561

그림 2 학습 및 검증 데이터 발췌

최종적으로 10만개의 학습 데이터에서 중복을 제거하고 702,685개의 고유한 학습 데이터를 얻을 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 적용하여 학습데이터 632,418개 검증 데이터 70,269개를 분리하였다.

4. 분류 모델 구축 방법 및 결과 분석

4.1 사전학습 언어모델 별 파인튜닝 모델 구축

본 논문의 실험에서는 사전학습 언어모델로 XLM-RoBERTA-Large[4], KoElectra[5], DistilBERT[6]를 사용해 보았으며, 워드임베딩 기반의 지도학습 분류 모델로 FastText[7]를 이용해 보았다.

분류를 위한 파인튜밍 모델로는 각 언어모델과 함께 구현된 SequenceClassification 모델을 이용하였으며, 각 사전학습 언어모델을 이용한 분류 모델 파인튜닝 학습 결과는 다음 표 1과 같다.

표 1 모델 학습(Training) 결과 (KoElectra LM 적용)

	Precision	Recall	F-1 Score	Support
Accuracy	0.98			100,000
Macro Avg.	0.96	0.95	0.95	100,000
Weigted Avg.	0.98	0.98	0.98	100,000

참고로 계층 구조인 한국표준산업분류의 특징을 반영하여, 하위 분류 시 상위 항목의 분류 결과를 추가로 활용하는 파인튜닝 모델(HIBERT)도 적용하였으며, 그 결과 평가 결과 비교는 다음 표 2와 같다.

표 2 구축 분류 모델 평가 결과

LM	Accuracy (%)	F-1 Score (0-1, weighted)
XLM-RoBERTA-Large[4]	88.95	0.739
KoElectra[5]	89.18	0.769
DistilBERT[6]	86.79	0.650
FastText[7]	72.65	0.727
Ensemble (Hard-voting)	**90.80**	**0.793**
[비교] Ensemble (HiBERT + RoBERTa)	91.24	0.819

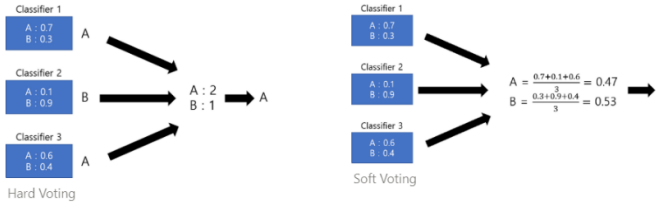

그림 3 Hard/Soft Voting 결과 비교 설명

4.2 앙상블 모델 구축 방법

앙상블 모델을 구축하기 위해서, 앞서 학습한 각각의 파인튜닝 분류 모델을 투표 알고리즘을 통해 최종 분류 결과를 추출해 보았다. 그림 3과 같이 세 개의 언어모델(RoBERTA-Large, KoElectra, DistilBERT)로 학습한 파인튜닝 모델에서 각 모델 별 Top-2 예측 결과를 이용하여, 두 개 이상의 모델에서 같은 결과를 이용하는 Hard Voting 방식으로 정답을 예측하였다.

그 결과 최종적으로 표 2와 같이 평가셋 기준 예측정확도(Accuracy) 90.8%, F-1점수 0.793의 분류 정확도 결과를 얻을 수 있었다.

추가로 토크나이징 결과를 개선하기 위해 띄어쓰기 정제까지 반영하였을 때에는 최종적으로 약 1% 이상의 정확도, 0.03의 F-1 Score 가 개선되는 결과를 얻을 수 있었다.

4.3 분류 결과 분석

같은 평가셋에 대해서 KoElectra, XLM-Roberta-Large 기반 모델들이 서로 다르게 예측한 데이터를 확인하였으며, 전체 10만개의 평가셋 데이터 중에서 9,743 개가 서로 다르게 예측했음을 확인하였다. 이 같은 양상은 같은 데이터로 학습한 기존 연구의 모델과 본 연구의 모델들의 전체적인 예측 정확도인 89% 확률과 유사한 양상을 띄고 있다는 점을 확인할 수 있었다.

5. 결론

최근 딥러닝에 기반한 자연어처리 기술의 발전과 오픈소스의 공개로 여러 도메인 분야에서 한국어 텍스트 처리 기술이 실무에 적용되고 있으며, 기존에는 어려웠던 시스템이나 서비스적인 구현에 대한 제약은 많이 해소되고 있다. 그러나 특수한 업무나 목적에 최적화된 기술적이며 공학적인 연구는 계속되고 있다.

본 논문에서는 공개된 여러 사전학습 언어모델을 적용하여 분류 모델을 파인튜닝 학습 해보면서 성능을 비교하고, 여러 모델의 결과를 앙상블로 적용함으로써 그 결과가 어떻게 나오는지를 확인하였다.

추후 연구로는 지역별고용조사, 전국사업체조사, 인구총조사 등 통계청의 대용량 조사 데이터에 본 연구의 방법론을 적용하고 최적의 성능을 분석해 봄으로써, 일반적으로 적용할 수 있는 통계자료 처리 및 분류 구축 방법론을 수립해 볼 예정이다.

감사의 글

이 논문은 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임. (No. 1415179613, 전문개인투자자 맞춤형 투자 정보 제공을 위한 실시간 금융 텍스트 심층 이해 및 투자 정보 지원 서비스 개발)

참고문헌

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

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

[3] 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.

[4] 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.

[5] JangGwon Park. KoELECTRA: Pretrained ELECTRA Model for Korean, GitHub repository, https://github.com/monologg/KoELECTRA, 2020.

[6] 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.

[7] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of Tricks for Efficient Text Classification, arXiv preprint arXiv:1607.01759, 2016.