한국진로창업경영학회지 제6권 4호 pp. 161~179 http://dx.doi.org/10.48206/kceba.2022.6.4.161
2022년 11월

딥러닝기반 텍스트 분석을 통한 직업분류시스템 구축에 관한 연구*

장지연*, 심지환**, 정준호***, 전병유****

< 목 차 >

< 요 약 >

본 연구의 목적은 온라인구인공고 텍스트 데이터를 활용하여 해당 일자리의 직종을 판별하는 분류 모델을 생성해 평가하는 것이다. 워크넷 온라인구인공고(OJPs) 텍스트 자료에 딥러닝 기계학습 기법을 적용하여 자동으로 직업을 판별하는 것이다. 텍스트 자동 분류를 위한 기계학습 기법이 규칙기반 모델에서 인공신경망 모델로 전환하는 연구 흐름을 반영하고, 대규모의 온라인구인공고 자료와 텍스트의 문맥적 의미를 잘 다룰 수 있다는 점을 고려하여, 인공신경망의 최신 모델인 Bi-LSTM과 KoBERT 모델을 적용하였다. 1999-2021년 간의 워크넷 구인공고 데이터 800만 개에 모델을 적용한 결과, 0.62-0.82 정도의 매칭 정확도를 달성했다. 특히, 직무 기술(job description)이 특수하고 정확한 전문직에서 높은 정확도를 달성했다.

키워드 : 기업가정신, 기업가정신교육, 창업의지, 성별 조절효과

---

* 이 논문은 한국고용정보원 연구용역보고서 “온라인 구인-구직 텍스트 데이터 분석을 통한 직업 분류와 기능(skills) 수요 변화 연구”를 추가 분석을 통해 수정·보완하였습니다.

* 제1저자, 한국노동연구원 선임연구위원, jchang@kli.re.kr
** 공동저자, 국민대학교 데이터사이언스 학과, 박사과정, sim2080@gmail.com
*** 공동저자, 강원대학교 부동산학과, 교수, jhj33@kangwon.ac.kr
**** 교신저자, 한신대학교 사회혁신경영대학원 교수, bycheon@hs.ac.kr

• 논문투고일 : 2022-10-25 • 수정일 : 202-11-17 • 게재확정일 : 2022-11-24

Ⅰ. 서 론

온라인 구인-구직 과정에서 얻어지는 노동시장 관련 데이터가 노동시장의 변화를 파악하고 전망하는 새로운 원천으로 노동시장의 분석, 연구, 정책에 활용되기 시작하고 있다. 특히, 온라인구인공고(Online Job Postings, 이하 OJPs)는 고용주의 노동 수요를 실시간으로 더 세밀하게 파악할 수 있게 한다. 모든 구인공고가 온라인에 올라오는 것은 아니지만 점점 더 많은 일자리가 온라인을 통해서 매칭되고 있으며, 노동시장 관련 지표의 실시간 생산이 가능해지고 있다. OJPs의 텍스트 데이터를 활용하여 직업을 분류하는 것은 노동시장 정보를 개선하여 많은 이점을 제공한다. 구직자와 고용주가 서로를 더 쉽게 찾을 수 있게 하여 일자리 매칭을 개선하며, 노동시장 분석의 기초 정보로 활용되어 고용주와 정책 담당자에게 더 유용한 통찰력과 시사점을 제공할 수 있다. 기술 변화에 따른 과업, 숙련, 역량 등과 같은 직업 특성들의 변화를 빠르게 포착하고 이를 분류하고 표준화하려는 정책담당자와 연구자들의 손을 덜어주기도 한다. 직업 관련 문서들(구인공고나 이력서 등)을 자동으로 분류하는 것은 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 줄일 수 있다. 특히, 기계학습(Machine Learning, 이하 ML)을 활용하는 텍스트 분석(Text Analytics)은 전문가들의 도메인 지식에 의존하지 않으면서도 대량의 정보를 다룰 수 있는 자동 분류 체계 구축을 가능하게 한다.

본 연구는 워크넷(Work-Net)의 OJPs 텍스트 데이터를 분석하여, 기존의 직업분류체계(한국고용정보원의 ‘한국고용직업분류’ KECO)로 자동분류하는 시스템을 구축하는 것을 목적으로 한다. 방법론적으로는 가장 최신의 텍스트 분석 기법인 딥러닝(deep learning) 모델을 적용한다. 최근 ML 생태계에서는 텍스트 분석 방법론은 규칙 기반 확률 모형에서 인공신경망과 딥러닝 모델로 발전하고 있다. 본 연구에서 딥러닝 모델을 적용한 것은 OJP 텍스트 데이터에 대한 자연어 처리 능력이 뛰어나기 때문이었다.

본 논문의 구성은 2장에서 OJPs를 활용한 직업분류에 관한 선행 연구를 정리하고, 3장에서는 연구 방법과 활용한 데이터 등을 소개한다. 4장에서 분석 결과를 보여주고, 5장에서 요약과 결론, 시사점을 제시한다.

Ⅱ. OJPs를 활용한 직업분류에 관한 선행 연구

노동시장에서의 비정형 텍스트 데이터에서 의미 있는 정보를 추출하는 것은 이력서를 직무와 일치시키기 위해 이력서 관리를 자동화하려는 전자 모집 프로세스(Electronic Recruit Process)를 지원하는 데 주로 사용되었다. 그러나, 최근 들어, 구인공고(Job Postings or Job

Vacancies)가 고용주의 노동수요를 일자리 수준에서 잘 반영하기 때문에 노동시장 수요 분석에 활용되고 있다.

OJPs를 활용하는 기존의 연구들은 대부분 민간기업에서 만들어준 직업분류체계를 활용하는 방식이었다. 예를 들어, Acemoglu et al.(2020), Deming & Noray(2020), Burke et al.(2020), Das et al.(2020), Hershbein & Kahn(2018) 등은 민간기업인 BGT(Burning Glass Technologies)가 제공하는 OJPs와 연계된 직업코드를 활용하여 직업 내 숙련이나 직업별 과업 비율(task-share)의 동태적 변동을 분석하였다. Calanca et al.(2019)의 연구도 영국의 온라인 일자리매칭 사이트인 Adzunda(https://www.adzuna.com/)로부터 OJPs 데이터와 29개 직업분류로 연계하는 코드를 제공받아서 숙련을 파악하는 연구를 진행하였다. 그런데, BGT나 Adzunda와 같은 민간기업들은 OJPs의 데이터를 활용하여 어떻게 직업을 분류하고 연계하였는지에 대해서는 일반에게는 공개하지 않고 있다(Gnehm & Clematide, 2020).

그러나, OJPs의 직업명(job titles)이나 직무기술(job descriptions)과 같은 텍스트 형태의 데이터를 ML 방법으로 분석하여, 기존 직업분류체계에 매칭시키는 연구들이 제출되기 시작하고 있다. 관련 연구들을 정리한 것이 <표 1>이다.

Turrell et al.(2019), Hoang et al. (2018), Colombo, Mercorio & Mezzanzanica(2019) 등은 자국의 온라인 일자리 매칭사이트의 OJPs 데이터를 활용하고, 명시적 규칙(explicit-rules) 모델*, Linear SVC, LDA, Perceptron 등과 같은 알고리즘 등을 사용하여 구인공고 데이터를 표준직업분류시스템으로 분류하는 알고리즘을 개발하였다.

개인 연구 목적이 아니라 정부나 기관 차원에서 OJPs를 기존 직업분류체계에 연계한 대표적 사례가 이태리 the University of Milano-Bicocca 연구팀이 Cedefop European Agency의 지원을 받아, 개발한 WoLMIS(a labor market intelligence system for classifying web job vacancies)이다. 이는 EU 차원의 온라인 구직 사이트에서 OJPs를 수집하고 국제적 직업분류체계 ISCO(International Standard Classification of Occupations)에 따라 OJPs를 분류하는 시스템이다. WoLMIS의 내용과 개발 과정에 대한 자세한 설명은 Boselli et al.(2018a, 2018b, 2017)와 Marrara et al.(2017)에 잘 제시되어 있다. EU 이외의 국가들에 대해서도 유사한 연구들이 이루어지고 있다. Xu et al.(2017)는 중국에서 직업 분류를 위한 대규모 OJPs 말뭉치 JCTC(Job posting Corpus for Text Classification)를 구축하는 계획의 첫 번째 작업으로, OJPs를 기존의 중국의 직업분류체계인 CGCO(People's Republic of China Grand Classification of Occupations)에 연계하였다. 여기에 CNN, LSTM, GRU(gated recurrent unit) 등을 벤치마킹 모델로 사용하였다.

최근 들어 딥러닝 기반의 알고리즘이 OJPs의 직업분류에 좋은 성능을 보여주는 연구들이

* 규칙기반 접근 방식은 특정 도메인의 용어들을 모델링한 분류(taxonomy)를 활용하여 텍스트 분류 규칙을 정의한 후에 온라인 구인공고에서 사용된 관련 용어를 식별하고, 이를 기존 직업분류체계에 맵핑하는 방식이다.

제출되고 있다. Gnehm & Clematide(2020), Tamburri, Van Den Heuvel & Garriga(2020), Choi , Kim & Lee(2020) 등은 딥러닝 기반 모델인 LSTM, BERT 모델 등을 활용하여 직업 분류의 성능을 크게 개선하고 있다. Gnehm & Clematide(2020)은 BERT 모델을 사용하여 문맥을 고려하는 임베딩(contextualized embedding)의 이점과 이러한 목적을 위한 다중 작업 모델(multi-task models)의 잠재력을 경험적으로 발견하였다. 즉, 문맥을 고려하는 딥러닝 기반 모델인 BERT-CRF를 활용한 결과 정확도가 91%에 달하였고*, 하나의 모델에서 직업(34 클래스), 산업(11 클래스), 관리기능(2 클래스)을 동시에 분류하는 결합분류(joint classification) 방법(a multi-tasking BERT text classifier)을 활용하여 좋은 성능을 낼 수 있음을 보여주었다. Tamburri, Van Den Heuvel & Garriga(2020)도 BERT 모델을 사용하여 80% 이상의 정확도로 매칭시킨 결과를 보여주었고, Choi, Kim & Lee(2020)도 Bi-LSTM(Long Short Term Memory)과 Bi_LSTM attention 모델을 사용하여 인터넷 구직사이트인 Careerbuilder의 OJPs를 특정 직무 역할로 분류하였다.

한국에서 기존의 직업 연구에서는 전문가 판단이나 설문조사에 기초했었다(오현주・김미경 2020; 오성욱・김태희, 2020). 최근, ML 기법을 활용하여 직업을 자동으로 분류하는 시스템에 대한 연구들이 제출되고 있다. 다만, OJPs를 활용하기보다는 주로 통계청 서베이 조사의 텍스트 자료들을 활용하는 방식이었다. 이들 연구는 사회과학보다는 컴퓨터과학 쪽의 연구들이 대부분이며, 자동화된 직업분류체계 자체에 관심이 있다기보다 ML 알고리즘의 성능 평가가 연구목적인 경우가 많았다. 다만, 임정우 외(2021)의 연구는 딥러닝 기반의 LSTM, BERT 모델을 적용했다는 점에서 관심을 가져볼 만하다. 임정우 외(2021)는 통계청의 인구주택총조사와 사업체기초조사에서 조사대상자들이 응답한 텍스트 데이터**에 Bi-LSTM과 BERT 모델을 적용하였다. 우찬균・임희석(2020)도 통계청의 조사 자료를 활용하여 딥러닝 기반으로 자동산업분류기를 만들었다. 기존에 코드가 부여된 자료를 가지고 자동화된 산업분류 알고리즘들을 생성한 것인데, CNN, LSTM, CNN-LSTM 등의 성능을 평가한 결과, CNN-LSTM이 좋은 성능을 보인 것으로 평가하였다.

---

* 문맥을 고려한 임베딩을 위해 토큰 수준 시퀀스 라벨링(token level sequence labeling)으로 구인공고에 대한 텍스트 영역화(text zoning)를 하여 구인공고를 구조화하는 작업을 하였다.

** 임정우 외(2021)의 연구에서는 인구주택총조사 데이터에서는 각 개인이 종사하고 있는 일과 재직중인 사업체, 사업체에서의 부서 및 직위, 해당 사업체의 사업내용 등의 데이터와 수작업으로 레이블링된 산업코드와 직업코드 등을 활용하였고, 사업체기초조사에서는 사업체의 이름, 사업내용, 주요 품목 등과 수작업으로 분류된 산업-직업 코드 등을 사용하였다.

<표 1> OJPs를 활용한 기존 연구 정리

구분	연구자	OJPs	분석 기법	분석 내용
민간 데이터 활용	Turrell et al.(2019)	Reed.co.uk	TF-IDF	직업 숙련수요 예측
	Hoang et al. (2018)	CareerBuilder	SVM	다중 레이블 분류
	Xu et al.(2017)	ChinaHR, Zhaopin, 51job	CNN, LSTM, GRU	JCTC 구축
	Colombo, Mercorio & Mezzanzanica(2019)	wollybi.com	SVM	직업 숙련 파악
기관 차원 연구	Boselli et al. (2018a, 2018b, 2017)	EU 온라인구직사이트	다양한 분석기법 적용	WoLMIS 구축
딥러닝 기반 연구	Gnehm & Clematide(2020)	Swiss Job Market Monitor, SJMM	BERT	직업, 산업, 관리기능별분류
	Tamburri, Van Den Heuvel & Garriga(2020)	네델란드, 벨기에 OJPs	BERT	직업분류시스템
	Choi, Kim & Lee(2020)	CareerBuilder	Bi-LSTM	직무역할분류
국내 연구	임정우 외(2021)	인구주택총조사 사업체기초조사	Bi_LSTM, BERT	자동직업분류
	우찬균·임희석(2020)	지역별고용조사 전국사업체조사	CNN, LSTM, CNN-LSTM	자동산업분류기

딥러닝 기반의 모델은 단어 간의 숨겨진 관계와 의미를 포착하는 데 효과적이어서, 다양한 자연어 처리 작업에서 유망한 결과들을 보여주고 있다. 본 연구에서도 자동화된 직업분류 시스템을 구축하는 연구들이 빠르게 발전하는 흐름을 고려하여, 딥러닝 기반의 모델인 Bi-LSTM과 KoBERT 모델을 고용노동부의 Work-Net 구인공고 자료에 적용해보고 그 성능을 평가해보고자 한다. 딥러닝 기반 모델은 데이터 규모가 매우 커야 하며, 좋은 결과를 얻기 위해서는 모델 패러미터의 미세조정 방법을 적용해야 한다. 본 연구도 워크넷 구인공고라는 대규모 데이터를 사용할 수 있기 때문에 딥러닝 적용하는 데 무리가 없을 것으로 판단된다.

Ⅲ. 연구 방법

본 연구는 온라인구인공고 데이터에 포함되어 있는 직업 관련 텍스트 데이터를 활용하여 온라인구인공고의 직업을 파악하고 분류하는 방법을 개발하기 위해, 고용노동부가 운영하는 공공고용서비스(Public Employment Service) 구인-구직 사이트인 워크넷(Work-Net)에 게시된 구인공고를 분석 자료로 사용한다. 워크넷의 구인공고에는 채용하고자 하는 일자리의

이름(직업명, job title)이 채용 제목에, 그리고 담당하게 될 직무에 대한 설명(직무기술, job description)이 담당 업무에 텍스트 형태로 포함되어 있다. 본 연구에서는 기본적으로 직업명과 직무기술 텍스트 자료를 가지고 직종을 식별하고자 한다. 구인공고에는 요구되는 경력이나 학력에 대한 부가적인 정보도 포함되어 있어서 이러한 변수들도 활용하였다. 딥러닝 기법을 적용하기 위해서는 풍부한 자료가 많을수록 좋겠으나, 컴퓨터의 처리능력이나 분석에 소요되는 시간 등 현실적인 문제를 고려하여 9개 연도 자료로 제한하였다. 1999-2001년, 2009-2011년, 2019-2021년도의 9개 년도 자료를 활용하였다. 매년 약 백만 건 내외의 자료로 총 8백 만여 건의 구인공고 데이터이다. 분석에 사용된 자료에서 직무설명 텍스트가 전혀 없거나 직종코드가 부여되어있지 않은 일부 사례들은 제외하였다.

워크넷의 온라인구인공고 데이터에는 이미 한국고용직업분류(KECO) 체계에 따른 직종코드가 명시되어 있다. 워크넷에서는 자체적으로 가지고 있는 워크넷취업알선직업분류코드에 따라, 구인처의 구인공고 게시자가 직업분류코드를 일차적으로 입력하고 고용센터의 직업상담사들이 사후적으로 체크하고 있다. 본 연구에서는 이 정보를 타겟(target)으로 하고 구인공고 제목(직업명)과 직무기술(job description) 등 두 가지 텍스트 데이터를 특성(Features)*으로 사용하여(<표 2> 예시 참조) 지도학습 방식의 딥러닝을 수행한다. 일단 직종분류에 적합한 딥러닝 모델이 개발되면 이를 직종분류기로 활용하여 직종코드가 제공되지 않는 여타 구인공고에 직종코드를 부여할 수 있다. 본 연구에서는 한국고용직업분류(KECO) 소분류(3-digit)를 기준으로 분석하였다. 1999-2021년 간 전체 구인데이터에는 136개의 직종이 포함되어 있다. 본 연구의 과제는 각각의 구인공고 일자리를 136개의 직종 하위범주에 자동으로 매칭하는 자동직업분류시스템을 구축하는 것이다.

워크넷 구인공고에는 직무설명 이외에 자격증과 전산능력 요구사항을 표기하는 항목이 따로 있다. 이 정보도 분석에 포함시키는 것이 직종을 예측하는 데 도움이 될 것이므로 직무설명에 추가하였다. 예컨대 위 2번 사례의 경우, '지게차운전기능사' 자격증을 요구한다고 되어 있으므로 실제 분석에 포함된 텍스트는 '지게차운전기능사. 벼를 매입하여~ 도정해서~ 쌀로 포장하는 작업을 합니다' 가 된다. 이러한 텍스트 데이터는 일차적으로 불필요한 구두점과 기호들을 제거하고, 불용어(stop word) 제거, 오해의 소지가 있는 단어 및 숫자 제거, 특수문자 대체 등 데이터 전처리를 실행해야 한다. 이러한 토큰화 작업에는 한국어 정보처리를 위한 파이썬 패키지인 코엔엘파이(KoNLPy)에서 제공하는 한국어 형태소 분석기 'Okt' 를 사용하였다. 전처리 된 텍스트 형태는 <표 3>에 제시되어 있다.

---

* ML에서 target은 목표인데, 레이블(label)이라고도 한다. feature는 학습 및 예측을 할 데이터의 특성 변수들(특징이나 항목들)을 의미한다. feature 대신에 attribute라는 용어가 사용되기도 한다. 지도학습(Supervised Learning) 기반의 텍스트 분류는 데이터의 feature와 target(label) 사이의 관계를 모델링하는 것으로 볼 수 있다. 비지도학습(Unupervised Learning)은 target 없이 Features만 가지고 클러스터링하는 것으로 볼 수 있다.

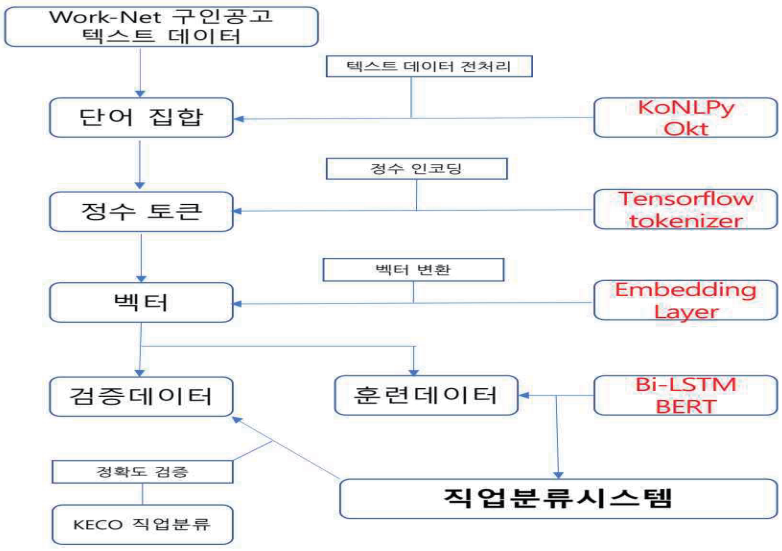
〈표 2〉 구인공고 제목과 직무설명 예시

연번	채용 제목	직무 기술
1	플라스틱 사출 관리직 채용	플라스틱 자동차부품 사출 생산 관리원(경력자 우선채용)
2	□□ 정미소 단순 생산직	벼를 매입하여~ 도정해서~ 쌀로 포장하는 작업을 합니다
3	0000000(회사명) 마케팅 본부 디지털 마케팅 경력 사원 채용	온라인 광고 기획 및 운영. 공식 웹사이트 및 온라인몰 사용자 행동 분석. 고객 데이터 관리, 분석 기반 디지털 캠페인 기획 및 실행

〈표 3〉 전처리 후 모델에 투입될 텍스트 형태

연번	채용 제목	직무 기술
1	플라스틱 사출 관리직 채용	플라스틱 자동차 부품 사출 생산 관리원 경력 우선 채용
2	(회사명) 정미소 단순 생산직	벼 매입 도정 쌀 포장 작업 지게차 톤 미만 자격증 필수
3	(회사명) 마케팅 본부 디지털 마케팅 경력 사원 채용	온라인 광고 기획 및 운영. 공식 웹사이트 및 온라인몰 사용자 행동 분석. 고객 데이터 관리, 분석 기반 디지털 캠페인 기획 및 실행

〈그림 1〉 자동직업분류시스템 생성 프로세스



이렇게 각 문장을 명사집합으로 구분한 뒤 구글 텐서플로(tensorflow)가 제공하는

Tokenizer를 이용하여 정수로 인코딩하였다. 이 토큰들은 본 연구에서 사용하게 될 모델인 Bi-LSTM과 BERT에 넣기 전에 신경망의 하나인 임베딩 레이어(embedding layer)에 투입한다. 임베딩셀은 입력된 문장을 분석하고 단어들 사이의 관계를 고려하여, 각 단어를 벡터로 변환시킨다. 본 연구에서는 100개의 숫자로 된 벡터로 임베딩하였다. 벡터로 변화된 데이터를 훈련데이터(training data) 80%와 검증데이터(test data) 20%로 랜덤하게 분할하고, 훈련데이터에 Bi-LSTM과 BERT 기법을 적용하여 구인공고에 직업분류코드를 매칭하는 알고리즘(직업자동분류시스템)을 생성한다. 이후 검증데이터를 가지고 이 알고리즘이 직업을 제대로 분류하였는지 성능을 평가한다.

Ⅳ. 연구 결과

본 연구에서는 Bi-LSTM 모델을 전체 데이터에 우선 적용하여, 모델의 정확도를 평가해보고, 일부 데이터를 직종별로 동일한 비율로 추출하여 Bi-LSTM과 KoBERT 모델의 성능을 평가해보고자 한다. 우선 Bi-LSTM 모델을 4가지 형태의 데이터에 적용해보았다. 전체 데이터의 80%를 훈련데이터(train data), 나머지 20%를 검증데이터(test data)로 사용하였다.

첫 번째 모델은 단순히 직무기술 텍스트만을 가지고 그 일자리(구인공고)의 직종을 분류하는 것이고, 두 번째 모델은 직무설명과 함께 구인공고의 제목(직업명)도 텍스트 형태로 투입하여 직종분류에 활용하는 모델이다. 세 번째 모델은 두 가지 텍스트 변수와 함께 학력과 경력과 같은 메타데이터를 투입하는 모델이다. 마지막으로 네 번째 모델은 위의 세 모델과는 달리 워크넷에 탑재되어 있는 '직업사전' 에 등장하는 어휘만을 사용하도록 사전적으로 제한하여 구인공고 데이터에 LSTM을 적용하는 모델이다.

Model-1 : 직무 기술
Model-2 : 직무 기술 + 구인공고제목(직업명)
Model-3 : 직무 기술 + 구인공고제목(직업명) + 교육년수(수치값), 경력년수(수치값)
Model-4 : Model3 + KECO 직업사전으로 사전 학습 후 구인공고에 적용하는 모델

우선, 위의 네 가지 모델의 성능을 정확도(accuracy)로 비교하여 살펴본 것이 <표 4>, <표 5>이다. * 정확도는 전체 사용된 총 케이스 수에 대한 올바른 예측의 비율로, 전체 KECO 136개 전체에 대해서 정의된 요약 값이다. 직무설명 부분의 텍스트만을 분석에 사용한 모델1에 비해서 구인공고 제목을 추가로 투입한 모델2의 성능이 훨씬 더 좋은 것을 확인할 수 있다. 또한, 모델2와 모델3의 성능은 대체로 비슷한 것으로 나타났다.

* 전체 데이터의 약 1/100 분량으로 샘플데이터를 만들어 적용해 본 결과이다. 전체 데이터에 대해서도 모델을 적용해 보았으나 컴퓨터의 성능 한계로 모델4는 돌아가지 않았다.

학력과 경력 요구사항을 메타데이터로 추가하는 것의 효과는 크지 않은 것으로 판단된다. 또한, '직업사전' 에 등장하는 정제된 단어를 활용하여 사전적으로 훈련시킨 모델4 역시 모델2에 비해 더 훌륭하다고 할 수 없었다. 샘플데이터(8만)에 비해 데이터 규모가 약 100배 큰 전체 데이터(800만)를 훈련에 투입하였을 때 성능이 더 좋아져, 대략 81.7% 수준에서 정확한 직종분류코드를 얻을 수 있었다.

<표 4> Bi_directional LSTM의 정확도 비교 (표본 데이터)

구분	훈련 데이터		검증 데이터	
	loss	accuracy	val_loss	val_accuracy
모델1	0.8780	0.7657	1.8282	0.6402
모델2	0.3001	0.9141	1.5307	0.7153
모델3	0.2683	0.9237	1.5415	0.7124
모델4	0.9687	0.7268	1.1863	0.6884

주: 2021년 데이터의 5% 표본인 약 8만 개의 데이터를 사용한 결과임.

<표 5> Bi_directional LSTM 성능비교 (전체데이터 이용)

구분	훈련 데이터		검증 데이터	
	loss	accuracy	val_loss	val_accuracy
모델1	1.2266	0.6932	1.3601	0.6713
모델2	0.5561	0.8443	0.6858	0.8169
모델3	0.7283	0.8015	0.7631	0.7936

주: 약 800만 개의 전체 데이터를 이용한 결과임. 데이터 사이즈가 커서 모델4는 기존 컴퓨터가 감당하기 어려워 시도하지 않았음.

따라서, 본 연구에서 적용할 Bi-LSTM과 KoBert 모델의 성능을 평가하기 위해, 직무기술과 구인공고제목(직업명)만을 활용하고, 워크넷 구인공고 자료가 전체 노동시장의 직종별 취업자수 분포와는 비례하지 않는다는 점을 고려하여, 직종별로 3,000개씩만 표본을 선정하여 모델에 투입하는 방법을 채택하였다. 모델의 성능을 다음과 같이 정의되는 precision, recall, f1-score, accuracy 지표*로 평가하였다.

* precision(정밀도)란 모델이 True라고 분류한 것 중에서 실제로 True인 것의 비율을 의미하며, recall(재현율)이란 실제 True인 것 중에서 모델이 True로 예측한 것의 비율, f1-score는 precision과 recall의 조화평균이다. f1-score는 데이터의 레이블이 불균형구조일 때 모델의 성능을 정확하게 평가할 수 있으며, 성능을 하나의 숫자로 표현할 수 있고, 조화평균은 산술평

모델1(bi-directional LSTM(1))은 구인공고 제목 텍스트와 직무기술명 텍스트를 결합하여 하나의 자료로 인풋으로 사용한 것이며, 모델2(bi-directional LSTM(2))는 구인공고 제목 텍스트와 직무설명 텍스트를 분리하여 각각 인풋으로 사용한 모델이다. 모델3은 모델2에 사용된 데이터에 트랜스포머 모델(Transformer model)을 적용한 것이며, 모델4는 같은 데이터에 KoBERT 모델을 적용한 것이다.

다음 <표 6>는 최근 3년간(2019-2021년) 자료만을 가지고 평가한 것이다. 직종별로 3천 개씩 추출하여 표본을 구성했고 최종데이터 수는 약 40만 개이며, 일부 직종은 3천 개에 미달하는 사례도 존재한다. 직종별로 3천 개의 표본을 추출하여 적용했기 때문에 가중치를 적용한 경우와 그렇지 않은 경우에 큰 차이가 나타나지는 않았다.

우선 Bi-LSTM(2) 모델의 경우 정확도가 0.62로 나타나 앞에서 전체 데이터를 활용해서 분석한 0.82에 비해서는 낮은 편이다. 데이터의 수가 클수록 정확도는 높아지는 것으로 볼 수 있다. 한편 Bi-LSTM 모델들보다는 KoBERT 모델이 상대적으로 더 높은 정확도를 보여준다. KoBERT 모델의 정확도는 약 0.75 수준을 나타낸다. 자동화된 직업분류기 적용에서도 최근 개발된 딥러닝기반의 BERT 모델이 좋은 성능을 보여주는 것으로 나타났다.

<표 6> Bi_directional LSTM과 Transformer 성능 평가 지표

구분	세부사항	precision	recall	f1-score	accuracy
Bi_LSTM(1)	macro avg.	0.55	0.51	0.52	0.58
	weighted avg.	0.60	0.58	0.58	
Bi_LSTM(2)	macro avg.	0.62	0.59	0.60	0.62
	weighted avg.	0.63	0.62	0.62	
Transformer	macro avg.	0.67	0.62	0.63	0.66
	weighted avg.	0.66	0.66	0.66	
KoBERT	macro avg.	0.74	0.73	0.73	0.75
	weighted avg.	0.75	0.75	0.75	

주: 2019-2021년 데이터에서 직종별로 3,000개의 표본을 추출하여 활용하였다. 최종데이터 수는 약 40만개이며, 일부 직종은 3,000개에 미달하는 사례도 존재한다.

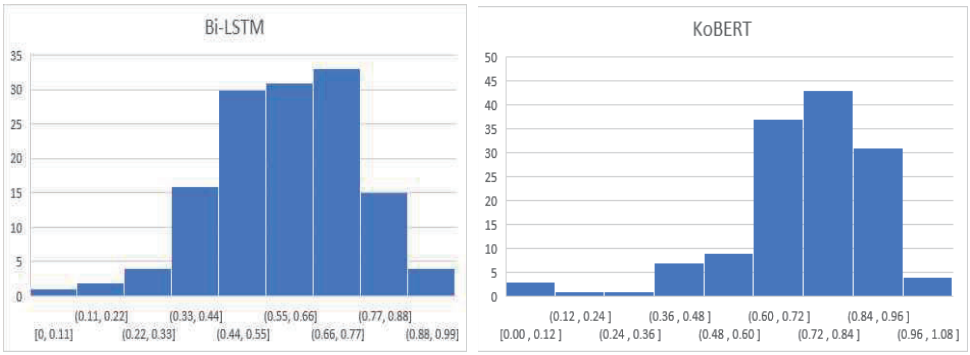
균에 비해서 커다란 비중의 클래스가 초래하는 편향(bias)을 줄일 수 있다. 정확도(accuracy) true를 true라고 예측한 것뿐만 아니라 false를 false라고 예측한 것까지 포함하는 개념이다.

<그림 2> 에서 f1-score 점수의 직종별 분포를 볼 경우, Bi-LSTM의 경우 0.66-0.77이 33% 정도로 나타나고 있으며, Ko_BERT의 경우 0.72-0.84가 42% 정도로 나타나고 있다. f1-score가 0.7 이상인 비율이 Bi-LSTM의 경우 50% 미만이지만, KoBERT의 경우 70%가 넘는 것으로 나오고 있다. KoBERT 모델의 성능이 많은 직종에서 더 좋다는 점을 확인할 수 있다.

한편, <표 7>은 대분류별로 매칭의 성능을 보여주고 있다. f1-score 기준으로 보면, 보건의료직종들이 0.95로 매우 높은 점수를 보여주고 있고, 예술디자인방송스포츠 직종이 높은 점수를 나타낸 반면, 경영·사무·금융·보험직, 교육·법률·사회복지・경찰·소방직 및 군인 등이 낮은 점수를 보이고 있다. 이렇게 직종별로 매칭 성능에서 상당한 차이가 있다는 것을 알 수 있다.

직종별로 자세한 성능 지표는 <부표 1>을 참조할 수 있는데, 영양사, 의사, 약사, 돌봄서비스종사자 등 전문직의 경우 상대적으로 높은 점수를 나타내는 반면, 표본수가 작은 케이스를 제외하면, 단순종사자, 관리자, 사무원 등의 경우 점수가 낮은 것으로 나타나고 있다. 다만, 법률사무원이나 금융보험사무원 등 전문성이 높은 직종의 경우 점수가 높은 것으로 나타나고 있다. 전문직의 경우, 구인공고에서 직무기술이나 직업명 등이 더 자세하게 서술되고 있기 때문인 것으로 판단된다.

<그림 2> 직종별 성능지표의 분포(히스토그램)



<표 7> 직종별 분류 성능 평가 지표(KoBERT 모델 적용

대분류 직종	평균값			표준편차			직종수
	precision	recall	f1_score	precision	recall	f1_score	
전체	0.74	0.73	0.73	0.16	0.20	0.18	136
경영·사무·금융·보험직	0.62	0.59	0.60	0.16	0.23	0.21	18
연구직 및 공학 기술직	0.71	0.69	0.70	0.10	0.18	0.14	19
교육·법률·사회복지·경찰·소방직 및 군인	0.77	0.65	0.69	0.26	0.26	0.24	12
보건의료직	0.95	0.95	0.95	0.03	0.05	0.04	7
예술·디자인·방송·스포츠	0.85	0.87	0.86	0.07	0.07	0.07	8
미용·여행·숙박·음식·경비·청소직	0.79	0.78	0.79	0.25	0.25	0.25	13
영업·판매·운전·운송직	0.73	0.77	0.75	0.11	0.11	0.11	11
건설·채굴직	0.75	0.72	0.73	0.09	0.11	0.10	6
설치·정비·생산직	0.73	0.73	0.73	0.12	0.15	0.14	37
농림어업직	0.75	0.74	0.75	0.12	0.22	0.17	5

Ⅴ. 결론

본 연구는 온라인구인공고(OJPs) 텍스트 자료에 최신의 딥러닝 ML 방법을 적용하여 자동으로 직업을 분류하는 모델을 생성하고 그 모델의 성능을 평가했다. 온라인노동시장 정보는 구인-구직 매칭 알고리즘 개발, 숙련 파악 및 새로운 노동통계 지표의 생성 등 여러 가지 방식으로 활용될 수 있지만, 본 연구에서는 온라인구인공고를 기존 직업분류체계에 자동으로 분류하는 알고리즘을 생성하고 평가하는 데 집중하였다.

텍스트 분석을 위한 방법론으로는 인공신경망 기반의 딥러닝 기법을 적용하였다. 자동으로 문서의 텍스트를 분류하는 방법은 최근 ML을 활용하여 빠르게 발전하고 있다. 규칙 기반의 확률 모형에 근거한 방법에서 인공신경망을 활용하는 모델로 빠르게 전환하고 있고, 인공신경망 모델 자체도 빠르게 진화하고 있다. 본 연구에서는 이러한 연구의 흐름을 반영하고 온라인구인공고의 텍스트 데이터가 가지는 특성을 반영하여 대규모 자료와 텍스트의 문맥적 의미를 잘 다룰 수 있는 인공신경망의 최신 모델인 Bi-LSTM과 KoBERT 모델을 적용했다.

Bi-LSTM 모델을 적용해본 결과, 800만 개의 대규모 데이터를 활용할 경우 약 0.82 정도의 정확도(accuracy)를 보여주는 것으로 나타났다. 실무에서 활용하기에는 만족스럽지는 않지만, 워크넷 구인공고 데이터가 아직은 직무에 대한 충분한 기술(description)이 이루어지지 않고 있다는 점을 고려하면, 그리고 높은 정확도를 보여준 여타 연구들이 특정 업종이나 직무에 한정해서 분석한 연구라는 점에서, 정확도가 그리 낮은 것은 아니라고 판단된다.

워크넷 데이터가 직종별로 구인공고 수에서 편차가 크다는 점, 그리고 대규모 데이터에 딥러닝을 처리하기에는 컴퓨터 처리 용량이 제한적이라는 점 등을 고려하여, 직종별로 3천개씩 표본 규모를 줄여서 두 가지 모델을 적용한 결과, Bi-LSTM은 0.62, KoBERT는 0.75 정도의 정확도를 나타냈다. 다만, 직종별로 볼 경우, f1-score가 보건의료직종들은 0.95로 높은 점수를 나타내고 있고, 단순종사자, 관리자, 사무원 등 낮은 점수를, 전문직은 높은 점수를 나타내고 있다. 이는 전문직일수록 직무기술이 직업명이 분명하게 구분되는 특수한 키워드를 많이 포함하고 있기 때문인 것으로 판단된다.

본 연구는 이러한 시스템 구축을 위해 국내에서 거의 처음으로 시도되는 연구라는 의미를 가진다. 워크넷 구인공고 텍스트 데이터를 활용하여 직업을 자동으로 분류하는 시스템은 그동안 활용되지 못한 노동시장 정보를 적극적으로 활용할 수 있도록 한다. 고용주와 구직자에게 더 많은 맞춤형 정보를 제공할 수 있는 기반을 제공하고, 직업상담사의 업무를 효율화하고 지원하는 데 기여할 수 있고, 정책담당자들에게 더 통찰력있는 유용한 정보를 제공할 수 있다.

물론, 온라인노동시장정보가 통계편향과 선택편향의 가능성이 있고, OJPs를 활용한 자동화된 분류가 사용자(고용주)의 필요를 단기적으로만 과도하게 대표할 수도 있어 단편적이고 대표성이 없는 그림만을 제공할 수 있다는 비판도 있다. 워크넷 데이터도 저숙련 일자리의 비중이 크다는 편향을 가지고 있어 전체 구인 수요를 반영하는 완전한 직업분류체계를 생성하는 것이 현재로서는 쉽지 않다. 더 타당한 형태의 자동화된 분류시스템을 구축하려면, OJPs의 데이터들을 더 정밀하게 처리하고, 더 다양한 모델들을 적용하여 성능을 높여야 한다.

자동분류시스템의 정확도가 아직은 높지 않지만, 기술은 항상 현실과의 상호작용하면서 발전하기 마련이다. 기술의 발전이 현실을 변화시킬 수도 있고, 현실이 기술 발전을 가속화하기도 한다. 한국의 노동시장이 좀 더 직무형으로 바뀌면서 구인공고 텍스트가 더 구체적으로 기술될 수 있으면 온라인구인공고자료에 기초한 완전한 자동화된 직업분류시스템이 구축될 수 있을 것이다. 노동시장 채용 관행이 직무 중심으로 변하고, 직업 관련 한국어 말뭉치와 직업 온톨로지 등이 더 정교하게 구축되어 데이터의 전처리가 더 정밀해지고, 모델을 다양하게 파인튜닝한다면 모델의 성능이 더 높아질 수 있을 것으로 기대한다.

<부표 - 1> 직종별 분류 성능 평가 지표(KoBERT 모델 적용)

keco_code	keco_title	평가지표			support
		f1_score	precision	recall	
305	영양사	0.99	0.99	0.99	600
301	의사, 한의사 및 치과의사	0.97	0.97	0.96	110
303	약사 및 한약사	0.97	0.98	0.97	94
550	돌봄 서비스 종사자	0.97	0.95	0.99	600
306	의료기사·치료사·재활사	0.96	0.95	0.97	600
412	기자 및 언론 전문가	0.96	0.95	0.97	315
302	수의사	0.94	0.94	0.94	18
304	간호사	0.94	0.92	0.96	600
511	미용 서비스원	0.94	0.94	0.94	600
512	결혼·장례 등 예식 서비스원	0.94	0.95	0.93	248
222	법률 사무원	0.93	0.93	0.94	600
32	금융·보험 사무원	0.92	0.91	0.93	600
561	청소·방역 및 가사 서비스원	0.92	0.90	0.93	600
213	유치원 교사	0.91	0.92	0.91	553
413	학예사·사서·기록물관리사	0.91	0.90	0.92	424
823	단조원 및 주조원	0.91	0.90	0.93	600
871	제과·제빵원 및 떡제조원	0.91	0.89	0.93	600
902	낙농·사육 종사자	0.91	0.87	0.96	600
904	어업 종사자	0.91	0.88	0.94	600
157	식품공학 기술자 및 시험원	0.90	0.87	0.94	600
411	작가·통번역가	0.90	0.92	0.89	284
812	운송장비 정비원	0.90	0.91	0.89	600
521	여행 서비스원	0.89	0.90	0.88	425
611	부동산 컨설턴트 및 중개인	0.89	0.85	0.93	193
307	보건·의료 종사자	0.88	0.91	0.85	600
416	연극·영화·방송 전문가	0.88	0.85	0.91	600
623	물품이동장비조작원(크레인·호이스트	0.88	0.86	0.90	600
881	인쇄기계·사진현상기 조작원	0.88	0.87	0.89	600
523	숙박시설 서비스원	0.87	0.85	0.90	520
613	텔레마케터	0.86	0.87	0.85	600
704	건설·채굴 기계 운전원	0.86	0.87	0.85	600
825	도장원 및 도금원	0.86	0.84	0.87	600
231	사회복지사 및 상담사	0.85	0.87	0.83	600
862	패턴사, 재단사 및 재봉사	0.85	0.82	0.89	600
883	가구·목제품 제조·수리원	0.85	0.83	0.88	600
414	창작·공연 전문가(작가, 연극 제외)	0.84	0.84	0.84	237
420	스포츠·레크리에이션 종사자	0.84	0.84	0.85	600
542	경비원	0.84	0.83	0.84	600
616	매장 계산원 및 매표원	0.84	0.83	0.84	600
864	제화원, 기타 섬유·의복 기계조작원	0.84	0.84	0.83	387
133	소프트웨어 개발자	0.83	0.82	0.83	600
155	에너지·환경공학 기술자 및 시험원	0.83	0.83	0.82	600
13	전문서비스 관리자	0.82	0.83	0.80	600
158	소방·방재·산업안전·비파괴 기술자	0.82	0.80	0.84	600
531	주방장 및 조리사	0.82	0.81	0.84	600

541	경호·보안 종사자	0.82	0.83	0.81	600
703	배관공	0.82	0.80	0.83	600
853	환경관련 장치 조작원	0.82	0.81	0.83	600
110	인문·사회과학 연구원	0.81	0.81	0.81	258
842	방송·통신장비 설치·정비원	0.81	0.78	0.85	600
861	섬유 제조·가공 기계 조작원	0.81	0.82	0.81	600
154	화학공학 기술자 및 시험원	0.80	0.77	0.83	600
532	식당 서비스원	0.80	0.83	0.76	600
822	판금원 및 제관원	0.80	0.80	0.80	600
824	용접원	0.80	0.82	0.79	600
851	석유·화학물 가공장치 조작원	0.80	0.79	0.81	600
882	목재·펄프·종이 생산기계 조작원	0.80	0.79	0.80	600
884	공예원 및 귀금속세공원	0.80	0.80	0.81	93
27	회계·경리 사무원	0.79	0.78	0.79	600
135	정보보안 전문가	0.79	0.74	0.84	126
156	섬유공학 기술자 및 시험원	0.79	0.80	0.78	105
705	기타 건설 기능원(채굴포함)	0.79	0.80	0.78	302
122	생명과학 연구원 및 시험원	0.78	0.76	0.80	475
214	문리·기술·예능 강사	0.78	0.77	0.80	600
232	보육교사 및 기타 사회복지 종사자	0.78	0.77	0.79	600
415	디자이너	0.78	0.77	0.79	600
417	문화·예술 기획자 및 매니저	0.77	0.76	0.77	187
23	회계·세무·감정 전문가	0.76	0.79	0.73	121
621	항공기·선박·철도 조종사 및 관제사	0.76	0.71	0.81	88
622	자동차 운전원	0.76	0.82	0.70	600
815	자동조립라인·산업용로봇 조작원	0.76	0.74	0.78	454
872	식품 가공 기능원	0.76	0.76	0.77	600
901	작물재배 종사자	0.76	0.74	0.78	600
31	금융·보험 전문가	0.75	0.72	0.78	82
22	경영·인사 전문가	0.74	0.72	0.75	333
841	정보통신기기 설치·수리원	0.74	0.77	0.71	600
221	법률 전문가	0.73	0.83	0.66	58
233	성직자 및 기타 종교 종사자	0.73	0.80	0.67	6
215	장학관 및 기타 교육 종사자	0.72	0.73	0.72	600
814	냉·난방 설비 조작원	0.72	0.72	0.72	600
524	오락시설 서비스원	0.71	0.69	0.72	141
817	운송장비 조립원	0.71	0.69	0.74	600
140	건축·토목공학 기술자 및 시험원	0.70	0.65	0.75	600
612	영업원 및 상품중개인	0.70	0.64	0.77	600
624	택배원 및 기타 운송 종사자	0.70	0.64	0.76	600
813	금형원 및 공작기계 조작원	0.70	0.66	0.75	600
24	광고·조사·상품기획·행사기획 전문가	0.69	0.67	0.71	600
25	정부·공공 행정 사무원	0.69	0.65	0.73	600
159	제도사 및 기타 인쇄·목재 등 공학 기	0.69	0.68	0.71	600
212	학교 교사	0.69	0.71	0.67	404
562	검침·주차관리 및 기타 서비스 단순 종	0.69	0.74	0.66	600
863	의복 제조원 및 수선원	0.69	0.72	0.66	178
873	식품 가공 기계 조작원	0.69	0.67	0.72	600
826	비금속제품 생산기계 조작원	0.68	0.71	0.65	600

831	전기공	0.68	0.70	0.67	600
152	금속·재료공학 기술자 및 시험원	0.67	0.71	0.63	240
833	발전·배전 장치 조작원	0.67	0.67	0.68	176
131	컴퓨터하드웨어·통신공학 기술자	0.66	0.66	0.66	297
614	소규모 상점 경영 및 일선 관리 종사자	0.66	0.63	0.68	600
701	건설구조 기능원	0.66	0.68	0.64	600
832	전기·전자 기기 설치·수리원	0.66	0.67	0.65	600
852	고무·플라스틱 및 화학제품 생산기계	0.66	0.65	0.67	600
885	악기·간판 및 기타 제조 종사자	0.66	0.66	0.66	600
11	의회의원·고위공무원 및 기업 고위임원	0.64	0.62	0.66	32
151	기계·로봇공학 기술자 및 시험원	0.64	0.61	0.67	600
706	건설·채굴 단순 종사자	0.64	0.69	0.59	600
834	전기·전자 설비 조작원	0.64	0.65	0.62	600
617	판촉 및 기타 판매 단순 종사자	0.63	0.65	0.61	600
702	건축마감 기능원	0.63	0.63	0.63	600
816	기계 조립원(운송장비 제외)	0.63	0.63	0.63	600
14	미용·여행·숙박·음식·경비·청소관	0.62	0.64	0.59	600
211	대학 교수 및 강사	0.62	0.89	0.47	34
811	기계장비 설치·정비원 (운송장비 제외)	0.62	0.62	0.62	600
12	행정·경영·금융·보험 관리자	0.61	0.62	0.59	600
836	전기·전자 부품·제품 조립원	0.61	0.59	0.62	600
905	제조 단순 종사자	0.60	0.68	0.53	600
28	무역·운송·생산·품질 사무원	0.59	0.56	0.62	600
134	데이터·네트워크 및 시스템 운영 전문	0.59	0.58	0.60	600
153	전기·전자공학 기술자 및 시험원	0.57	0.56	0.59	600
615	판매 종사자	0.57	0.56	0.59	600
903	임업 종사자	0.55	0.60	0.50	117
132	컴퓨터시스템 전문가	0.53	0.59	0.48	222
240	경찰관, 소방관 및 교도관	0.53	1.00	0.36	11
835	전기·전자 부품·제품 생산기계 조작원	0.53	0.57	0.49	600
29	안내·고객상담·통계·비서·사무보조	0.48	0.48	0.47	600
821	금속관련 기계·설비 조작원	0.48	0.52	0.45	600
16	건설·채굴·제조·생산 관리자	0.47	0.50	0.46	600
15	영업·판매·운송 관리자	0.44	0.50	0.40	600
121	자연과학 연구원 및 시험원	0.43	0.57	0.34	103
26	경영지원 사무원	0.41	0.47	0.37	600
136	통신·방송송출 장비 기사	0.38	0.67	0.26	23
33	금융·보험 영업원	0.31	0.55	0.21	28
890	제조 단순 종사자	0.21	0.28	0.17	600
21	정부·공공행정 전문가	0.04	0.20	0.02	48
250	군인	0.00	0.00	0.00	2
522	항공기·선박·열차 객실승무원	0.00	0.00	0.00	9

참고문헌

(1) 국내 문헌

오현주・김미경(2020). 취업취약계층의 취업장애요인 유형 분류에 관한 연구: 취업성공패키지 참여자를 중심으로. **한국진로창업경영학회지**, 4(1), 71-100.

오성욱・임태희(2020). 대학생의 취업준비행동이 졸업 후 첫 직업에 대한 만족요인들 간의 인과관계. **한국진로창업경영학회지**, 4(1), 23-50.

우찬균・임희석(2020). 딥러닝 기반 한국 표준 산업분류 자동분류 모델 비교, *한국정보처리학회 학술대회논문집*, 27(1), 516-518.

임정우・문현석・이찬희・우찬균・임희석(2021). 딥러닝 기법을 활용한 산업/직업 자동코딩 시스템, **한국융합학회논문지**, 12(4), 23-30.

(2) 국외 문헌

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P.(2020). **AI and jobs: Evidence from online vacancies (No. w28257).** National Bureau of Economic Research.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., & Viviani, M.(2018a). WoLMIS: a labor market intelligence system for classifying web job vacancies. **Journal of Intelligent Information Systems,** 51(3), 477-502.

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M.(2018b). Classifying online job advertisements through machine learning. **Future Generation Computer Systems,** 86, 319-328.

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M.(2017). *Using machine learning for labour market intelligence*. In Joint European Conference on Machine Learning and Knowledge Discovery in Database, Italy: Springer, Cham.

Burke, M. A., Sasser, A., Sadighi, S., Sederberg, R. B., & Taska, B.(2020). *No longer qualified? Changes in the supply and demand for skills within occupations.* Working Papers, Boston: Federal Reserve Bank of Boston.

Calanca, F., Sayfullina, L., Minkus, L., Wagner, C., & Malmi, E.(2019). Responsible team players wanted: an analysis of soft skill requirements in job advertisements. **EPJ Data Science,** 8(1), 1-20.

Choi, I. H., Kim, Y. S., & Lee, C. K.(2020). *A Study of the Classification of IT Jobs Using LSTM and LIME*. In The 9th International Conference on Smart Media and

Applications, Jeju:Association for Computing Machinery.

Colombo, E., Mercorio, F., & Mezzanzanica, M.(2019). AI meets labor market: Exploring the link between automation and skills. **Information Economics and Policy**, 47, 27-37.

Das, S., Steffen, S., Reddy, P., Brynjolfsson, E., & Fleming, M.(2020). *Forecasting Task-Shares and Characterizing Occupational Change across Industry Sectors*. In Harvard CRCS Workshop on AI for Social Good, Cambridge: CRCS.

Deming, D. J., & Noray, K. (2020). Earnings dynamics, changing job skills, and STEM careers. **The Quarterly Journal of Economics**, 135(4), 1965-2005.

Gnehm, A. S., & Clematide, S.(2020). *Text zoning and classification for job advertisements in German, French and English*. In Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science. Online: Association for Computational Linguistics.

Hershbein, B., & Kahn, L. B.(2018). Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. **American Economic Review**, 108(7), 1737-72.

Hoang, P., Mahoney, T., Javed, F., & McNair, M. (2018). Large-scale occupational skills normalization for online recruitment. **AI Magazine**, 39(1), 5-14.

Marrara, S., Pasi, G., Viviani, M., Cesarini, M., Mercorio, F., Mezzanzanica, M., & Pappagallo, M.(2017). *A language modelling approach for discovering novel labour market occupations from the web*. In Proceedings of the International Conference on Web Intelligence, Leipzig: Association for Computing Machiner.

Tamburri, D. A., Van Den Heuvel, W. J., & Garriga, M.(2020). *DataOps for societal intelligence: A data pipeline for labor market skills extraction and matching.* In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science(IRI), Las Vegas: IEEE.

Turrell, A., Speigner, B. J., Djumalieva, J., Copple, D., & Thurgood, J.(2019). Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings. **National Bureau of Economic Research**, WP No. w25837.

Xu, H., Gu, C., Zhou, H., Kou, S., & Zhang, J.(2017). *JCTC: A Large Job posting Corpus for Text Classification.* arXiv preprint arXiv:1705.06123.

<ABSTRACT>

A study of the establishment of a job classification system with deep learning-based online job posting text analysis

Chang, Jiyeun*, Sim, Jiwhan**, Jeong, Jun Ho***, Cheon, Byung You****

The purpose of this study is to create a classification model that can identify the type of job by using online job posting text data and evaluate the performance of the model. By applying the latest deep learning machine learning method to Work-Net online job postings(OJPs) text data, it is to automatically determine the occupat ional code of the OJPs. Considering the research trends shifting from a rule-based model to an artificial neural network model, and the merit of handling large-scale online job posting materials and the contextual meaning of text well, the latest models of artificial neural networks, Bi-LSTM and KoBERT models, were applied. As a result of applying the model to 8 million text data of employment insurance Work-Net job posting data from 1999 to 2001, matching accuracy of 0.62 to 0.82 was achieved. The result is not very high performance, but it is generally judged to be a model that can determine the occupation. In particular, high accuracy was achieved in professions where job descriptions were specific and precise. Although it is not yet perfect for practical use, it is expected that the performance of the automatic occupational classification system will improve in the future when recruitment practices into the job-type labor market change and more precise data pre-processing and model applications are made.

Keywords : Deep Learning, Machine Learning, LSTM, BERT, Online Job Posting, job classification

---

* First Author, Korea Labor Institute, Senior Researcher, jchang@kli.re.kr

** Co-Author, Kookmin University, Dept. of Data Science, Ph.d. Course, sim2080@gmail.com

*** Co-Author, Kangwon University, Dept. of Estate, Professor, Professor, jhj33@kangwon.ac.kr

**** Communication Author, Hanshin University, Graduate College of Social Innovarion Business, Professor, bycheon@hs.ac.kr