반도체디스플레이기술학회지 제21권 제2호(2022년 6월)
Journal of the Semiconductor & Display Technology, Vol. 21, No. 2. June 2022.

인턴십 지원자를 위한 기계학습기반 취업예측 모델 개발

김현수*†·김선호*·김도현*
*† 명지대학교 산업경영공학과

Development of the Machine Learning-based Employment Prediction Model for Internship Applicants

Hyun Soo Kim*†, Sunho Kim* and Do Hyun Kim*
*† Department of Industrial and Management Engineering, Myongji University

ABSTRACT

The employment prediction model proposed in this paper uses 16 independent variables, including self-introductions of M University students who applied for IPP and work-study internship, and 3 dependent variable data such as large companies, mid-sized companies, and unemployment. The employment prediction model for large companies was developed using Random Forest and Word2Vec with the result of F1_Weighted 82.4%. The employment prediction model for medium-sized companies and above was developed using Logistic Regression and Word2Vec with the result of F1_Weighted 73.24%. These two models can be actively used in predicting employment in large and medium-sized companies for M University students in the future.

Key Words : IPP, work-study Internship, machine learning, employment prediction model, large company, mid-sized company, Logistic Regression, Random Forest, Word2Vec, F1_Weighted

1. 서 론

청년 실업률을 줄이고자 고용노동부는 2015년부터 4년제 대학생들의 장기현장실습 및 일학습병행을 통해 중소, 중견기업들의 신입채용의 안정적인 확보와 학생들의 직무능력 역량을 조기에 달성할 수 있는 사업을 해오고 있다[1]. 지금까지 청년층의 취업에 관련된 연구들은 평균학점, 인턴십 등 독립변수 요인들과 취업여부 예측 관련 연구들이 대부분이고 자기소개서를 포함한 취업예측모델 개발 연구는 부족한 편이다. 선행연구를 보면 자기소개서가 없는 H 대학교의 학생들 데이터로 취업에 영향을 주는 요인들을 가지고 로지스틱 회귀모형으로 기업에 합격할 확률을 예측했다[2]. 머신러닝의 랜덤 포레스트 기법을 사용하여 자기소개서 포함 없이 영어점수, 취업동아리 등 독립변수 요인만으로 대졸자의 취업 여부를 예측했다. 또한 대졸자 직업이동경로조사(GOMS : Graduate Occpational Mobility Survey) 데이터를 이용하여 의사결정나무, 랜덤 포레스트, 인공신경망 등 기계학습 기반으로 대졸자의 취업 여부를 예측해 취업 여부에 영향을 미치는 중요 변수를 도출해내어 대졸자들의 고용 관련 정책 수립 및 다양한 취업 지원 프로그램을 분석을 했지만 기업규모 합격에 따른 취업예측 모델 개발은 연구되지 않았다[4]. 따라서 본 논문에서는 M대학교 인턴십에 지원한 학생들의 데이터(16개 독립변수, 3개의 종속변수)와 선행논문에서 사용하지 않은 자기소개서 텍스트파일을 벡터화해서 로지스틱 회귀모형, 랜덤 포레스트, Light GBM 모델과 임베딩 기법 Word2Vec, BERT 등 기계학습 기반으로 대기업합격 및 중견기업이상 합격 취업예측모델을 제안한다.

†E-mail: hskcpu@mju.ac.kr

2. 취업관련 모델

이장에서는 취업예측 모델에 사용된 기계학습 관련 모델들을 간략하게 소개한다.

2.1 로지스틱 회귀모형

로지스틱 회귀모형은 분포의 가정이 일반 회귀분석에 비해 완화되어 있고, 독립변수의 분포에 대한 어떠한 가정도 필요하지 않기 때문에 분석이 용이하다. 취업예측모형에서 로지스틱 회귀모형에 대한 5가지 적합도 검정도 활용되었다[2]. 본 논문에서는 로지스틱 회귀모형을 취업예측모델 개발에 사용하였다.

2.2 랜덤 포레스트

랜덤 포레스트는 각각의 트리가 독립적으로 표본 추출된 임의 벡터 값에 따라 달라지는 동시에, 모든 나무는 동일한 분포를 갖는 나무 예측 변수의 조합으로 구성되어 있고 각각의 나무는 입력 벡터를 분류하는데 적합하며[5] R 함수를 사용해서 정확도, 민감도, 특이도를 검증하여 대졸자의 취업여부를 예측할 수 있다[4]. 랜덤 포레스트는 일반화 및 성능이 우수하며 파라미터 조정이 쉽고 데이터 스케일 변환이 불필요하다. 또한 오버피팅에 강하며 분류 예측에 안정하다. 랜덤 포레스트 모델은 선행연구에서 취업예측에 많이 사용하고 있으며 본 논문에서도 취업예측모델 개발에 적용하였다.

2.3 Light GBM(Gradient Boosting Machine)

Light GBM은 학습시간이 적고 빠른 속도와 범주기능의 자동 변환과 최적 분할이 장점이다. Light GBM 모형을 본 연구의 취업 예측모델 개발을 위해서 비교 검토하였다.

2.4 Word2Vec

구글 연구팀이 발표한 Word2Vec은 가장 널리 쓰이고 있는 단어 임베딩 모델이다. Word2Vec 기법은 「Efficient Estimation of Word Representations in Vector Space(Mikolov et al., 2013a)」와 「Distributed Representations of Words and Phrase and their Compositionality(Mikolov et al., 2013b)」 논문으로 나누어 발표되었고 네거티브 샘플링 학습 최적화 기법을 제안한 내용이 두 논문의 핵심 골자다[6]. Word2Vec 학습을 통한 한국과 일본의 실제 특허 문서에 대해서도 사용하였으며[7] 한국어 신문기사 문장의 분류에 있어 성능이 입증된 Word2Vec을 활용했다[8]. 본 논문에서도 자기소개서를 텍스트 파일을 벡터화로 변환시켜 Word2Vec을 취업 예측모델 개발 시 사용하였다.

2.5 BERT

BERT(Bidirectional Encoder Representations from Transformers)는 언어 이해를 위해 텍스트 파일의 양방향 표현을 훈련하도록 설계된 임베딩 기법으로 사전 학습되어 있는 모델이다[9]. 본 논문에서는 Word2Vec보다 벡터화 차원을 3배 이상 증가시킨 BERT도 취업 예측모델 개발 시 Word2Vec과 비교하기 위해 사용하였다.

2.6 샘플링 방법

모델개발 시 사용하는 샘플링 방법은 언더 샘플링, 오버 샘플링, SMOTE(Synthetic Minority Over sampling Technique), ADASYN(Adaptive Synthetic Sampling) 등 4가지 방법을 적용했다.

3. 취업예측모델 설계

3.1 개발모델 설계 과정

1단계는 독립변수 및 종속변수 조사이다. 2단계는 데이터수집 정제에서 설문조사 및 상담과 데이터 수집, 데이터 정제로 이루어지고 3단계는 데이터를 활용한 모델 개발에서 학습데이터 선정, 테스트 데이터 선정, 변수 검증 및 모델 검증으로 진행되며 4단계는 모델개발 순서로 설계되었다.

3.2. 개발 모델 선정

본 논문에서는 로지스틱 회귀 모형, 랜덤 포레스트, Light GBM, Word2Vec, BERT를 선정하여 대기업합격 취업예측 모델 개발과 중견기업이상합격 취업 예측 개발 모델로 선정했다.

3.3. 개발 모델 변수 선정

개발모델 변수 설정은 많은 선행 논문에서 연구되었다. 취업예측에 있어 H대학 사례에서는 취업동아리 등 14개 독립변수와 취업여부를 종속변수로 설정했다[2]. 또한 평균학점, 재학 중 일자리 경험 등 7개 독립변수와 종속변수를 대기업 취업률로 분석하였다[10]. 본 논문에서는 성별, 전공계열, 전공일치, 입학유형, 교환학생 경험,공학인증,비교과활동, IPP(Industry Professional Practice: 장기현장실습), 일학습병행, 인턴십 경험 무, 전공활동, 자격증&수상, 해외여행, 평균학점, 영어점수, 자기소개서를 포함한 16개 독립변수와 종속변수는 대기업 합격, 중견기업이상 합격, 미 취업으로 3개이다. 특히 선행 논문에서 독립변수로 사용하지 않은 자기소개서를 포함시켜 기계학습에 기반된 Word2Vec(200차원), BERT(768차원) 임베딩 기법을 적용하여 자기소개서 텍스트파일 내용을 벡터화로 변환시키고

독립 변수를 증가시켜 정확도를 향상시킬 수 있었다.

4. 데이터 수집 및 정제

4.1 조사

설문조사는 전화인터뷰와 한국 사회과학 데이터센터(KSDC : KOREAN SOCIAL SCIENCE DATA CENTER)의 DB(Data Base)를 사용하여 온라인 설문지를 작성, 배포, 결과확인, 자료 분석을 하였다. KSDC DB는 한국의 대표적인 학술 DB로서 주요 설문조사 및 통계 데이터를 수집 및 표본화 하여 제공하는 통합 DB이다.

4.2 상담

학생상담은 2017~2021년 장기현장실습 및 일학습병행(2018~2021년)에 지원한 학생들을 상담 진행하였다. 상담 내용은 전공과목이수, 필수 및 일반 교양과목 이수, 평균 학점, 어학, 공학인증 유무, 원하는 직무 및 회사, 본인 거주지, 수상, 자격증, 해외연수, 전공활동, 성격 등이다.

4.3 데이터 수집

장기현장실습 데이터수집은 2017년~2021년에 지원한 학생 1284명의 데이터다. 이중에 장기현장실습으로 확보된 데이터는 219명이다. 또한 장기현장실습에 지원했지만 인턴십을 하지 않았거나 장기현장실습을 진행했지만 미취업으로 확보된 데이터는 135명이다. 일학습병행 데이터 수집은 2018년~2021년까지 지원한 학생이 249명 데이터다. 이중에 일학습병행 인터십에서 확보된 데이터는 154명이다. 이들 데이터 확보는 대면상담, 전화상담, 전화인터뷰, 설문조사를 통해서 확보하였다.

4.4 데이터 정제

데이터 정제(Data Cleansing)는 결측 값을 채우거나 이상값을 제거하는 것으로 진행하였으며 장기현장실습 219명, 일학습병행 154명, 미 취업 135명 등 총 508명의 데이터를 확보하였다. 장기현장실습과 일학습병행 데이터 정제(Data Cleansing)는 16개의 독립변수를 기준으로 데이터를 정리하였다. 독립변수는 성별, 전공계열, 입학유형, 평균학점, 전공 일치, 교환학생 경험, 공학인증, 영어점수, 비교과활동, 장기현장실습, 일학습병행, 인턴십 경험 무, 전공활동, 자격증&수상, 해외여행, 자기소개서 등 총 16개 항목이며 종속변수는 대기업, 중견기업, 미 취업 등 총 3개이다. 데이터 정제는 첫째 장기현장실습 사이트에 등록한 이력서, 자기소개서 및 상담 내용을 중심으로 2017년 ~ 2021년에 지원한 학생들의 데이터를 정리하였다. 둘째 학생들과 이력서 및 자기소개서를 동일하게 지원하였는지 확인한 후 정리하였다. 셋째 데이터 정제는 정확한 데이터 확보를 위한 것이며 지원한 학생들의 이력서, 자기소개서, 전화인터뷰와 설문조사에 참여하고 장기현장실습 및 일학습병행을 경험하거나 경험하지 못한 졸업생들에 대한 데이터 비교 및 일치성 확인 후 정리하였다. Table 1은 적합성 테스트 시 독립변수와 종속변수 변환 데이터를 나타낸 것이다.

Table 1. Independent variable and dependent variable transformation data

Variable	Item		
Independent Variable	Gender	Male	1
		Female	0
	Major	Humanities	1
		Nature	0
	Match Major	Yes/No	1/0
	Admission Type	Regular	1
		Transfer	0
	Exchange Student Experience	Yes/No	1/0
	Engineering Certification	Yes/No	1/0
	Extracurricular Activities	Yes/No	1/0
	IPP	Yes/No	1/0
	Work-Study Internship	Yes/No	1/0
	No Internship Experience	Yes/No	1/0
	Major Activities	Yes/No	1/0
	Certifications & Awards	Yes/No	1/0
	Overseas Travel	Yes/No	1/0
	Grades	Graduation Credits, Standardization(0~1)	
	English Score	TOEIC, Standardization (0~1)	
	Self Introduction	Text File, Vector Dimension	
Dependent Variable	Large Company	1	1
	Mid-Sized Company	0	1
	Unemployed	0	0

5. 데이터를 활용한 모델 개발

5.1 학습데이터 선정

학습데이터 선정은 2017년 ~ 2021년 장기현장실습 및 일학습병행에 지원한 학생들 중에서 장기현장실습은 219명, 일학습병행 154명, 인턴십 무 135명 등 총 508명 중에서 70%356명을 데이터를 학습데이터로 선정하였다.

5.2 테스트 데이터 선정

테스트 데이터 선정은 2017년 ~ 2021년 장기현장실습 219명, 일학습병행 154명, 인턴십 무 135명 등 총 508명 중에서 30%인 152명 데이터를 테스트 데이터로 선정하였다.

5.3 적합성 테스트 진행

적합성 테스트 진행은 성별, 전공계열, 평균학점, 전공일치, 입학유형, 교환학생, 공학인증, 영어점수, 비교과활동, 장기현장실습, 일학습병행, 인턴십 경험 무, 전공활동, 자격증&수상, 해외여행 등 15개의 독립변수와 자기소개서 독립변수를 텍스트파일 처리하여 Word2Vec 200차원, BERT 768차원을 구성한다. 사전학습모델(이미 만들어진 모델 사용)을 적용하여 학습데이터 70%, 테스트데이터 30%로 적합성 평가를 진행하였다. 모델은 로지스틱회귀모형, 랜덤 포레스트, Light GBM과 BERT와 Word2Vec을 적용 및 비교 평가하여 정확도가 높고 최적한 조합으로 구성된 대기업합격, 중견기업이상합격 취업 예측 모델을 개발하는 것으로 진행했다.

5.4 로지스틱 회귀모형 적합성 테스트 결과

중견기업이상 합격 예측 모델에서는 자기소개서가 포함되지 않은 SMOTE에서 F1_Weighted는 70.52%, 자기소개서가 포함된 언더 샘플링 Word2Vec에서 F1_Weighted는 73.24%로 2.72% 증가하는 제일 좋은 값을 얻었다.

Table 2. F1_Weighted of medium-sized company or higher

Model	Sample Method	Text Treatment Method	F1_ Weighted
Logistic Regression	Under Sample	Word2Vec	73.24%

따라서 중견기업이상 합격 예측 모델에서 로지스틱 회귀모형으로 언더 샘플링 자기소개서가 포함된 Word2Vec으로 모델을 개발하는 결과를 확보했다. 위에 Table.2는 중견기업이상 합격 취업예측 모델 개발에 대한 F1_Weighted 결과이다.

5.5 랜덤 포레스트 적합성 테스트 결과

대기업 합격 예측 모델에서는 자기소개서가 포함되지 않은 원데이터에서 F1_Weighted는 77.14%, 자기소개서가 포함된 ADASYN에서 F1_Weighted는 82.48%로 5.34% 증가했다. 따라서 대기업 합격 예측 모델에서 랜덤 포레스트 ADASYN 샘플링 자기소개서가 포함된 Word2Vec으로 모델을 개발하는 결과를 확보했다. 아래는 Table.3은 대기업 합격 취업예측 모델 개발에 대한 F1_Weighted 결과이다.

Table 3. F1_Weighted of passing large company

Model	Sampling Method	Text Treatment Method	F1_ Weighted
Random Forest	ADASYN	Word2Vec	82.34%

5.6 Light GBM 적합성 테스트 결과

대기업 합격 예측 모델에서는 자기소개서가 포함한 ADASYN BERT에서 F1_Weighted는 81.08%, 중견기업이상 합격 예측 모델에서는 오버 샘플링 자기소개서가 포함되지 않은 F1_Weighted 74.72%를 얻었다.

5.7 변수 검증

기계학습 기반으로 취업 예측 모델을 개발했기 때문에 모델에서 독립변수들이 얼마나 중요하게 영향을 미치는지 변수들을 검증하였다. 모델개발에 영향을 미치는 것을 검증하기 위해 Shap을 사용하여 feature importance(기능 중요도)를 나타냈다. 아래 Fig. 1에서 보면 대기업 합격 취업 예측 모델 개발에서 독립변수가 영향을 미치는 5위 이내 순위는 일학습병행, 평균학점, 장기현장실습, 성별, 영어점수로 나타났다. 대기업 취업을 목표로 하는 학생들은 5위 이내 속한 5개 독립변수를 보면 일학습병행, 장기현장실습 인턴십을 반드시 해야 하고 대기업에서 요구하는 평균학점 및 영어점수가 높아야 한다. 또한 Fig. 2에서 보면 중견기업이상 취업을 목표로 하는 학생들은 5위 이내 속한 5개 독립변수를 보면 영어점수, 성별, 전공계열, 장기현장실습, 평균학점으로 나타났다. 이는 대기업 합격 취업 예측모델과 비교해보면 일학습병행 독립변수를 제외하고 4개의 독립변수는 순위는 다르지만 같게 나타났다. 중견기업이상 합격 취업 예측모델도 대기업합격 취업 예측모델에 영향을 미치는 정도가 80%가 일치하는 경향을 알 수 있었다. 중견기업이상 합격률을 높이려면 장기현장실습 인턴십을 반드시 해야 된다고 생각한다.

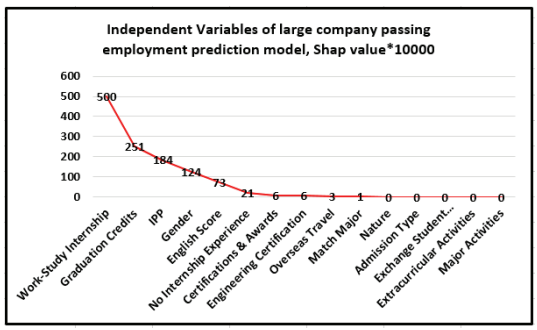

Fig. 1. Shap value of large company passing model.

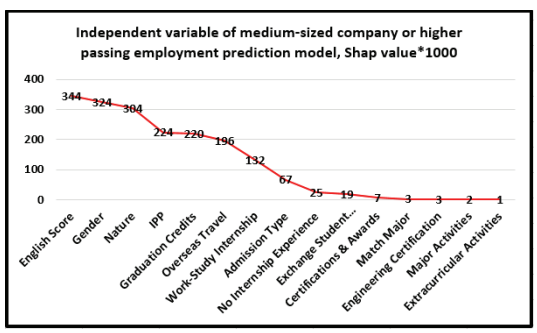

Fig. 2. Shap value of medium-sized company or higher passing model.

5.8 대기업합격 취업예측 모델 개발

장기현장실습 및 일학습병행 인턴십을 진행한 508명의 학생 데이터와 취업예측 성능평가를 위해 사용한 학생 44명의 모델 검증 데이터를 가지고 랜덤 포레스트, ADASYN, Word2Vec(200차원)으로 대기업 합격 취업예측 모델을 개발했다. Fig. 3에서 보면 취업예측모델 개발 시에는 F1_Weighted는 82.48%, 모델검증 예측성능평가에서는 70.69%를 나타내어 대기업합격 취업예측 모델로 사용할 수 있다고 판단된다.

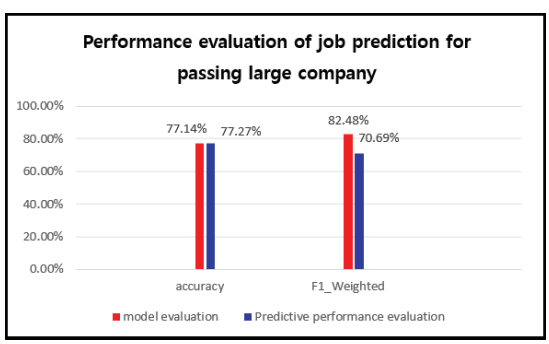

Fig. 3. Performance evaluation of job prediction for passing large company

5.9 중견기업이상합격 취업예측 모델 개발

장기현장실습 및 일학습병행 인턴십을 진행한 508명의 학생 데이터와 취업예측 성능평가를 위해 사용한 학생 44명의 모델 검증 데이터를 가지고 로지스틱 회귀모형, 언더 샘플링, Word2Vec(200차원)으로 중견기업이상합격 취업예측 모델을 개발했다. Fig.4에서 알 수 있듯이 취업예측모델 개발 시에는 F_1 Weighted는 73.24%를 나타내었고 모델검증 예측성능 평가에서는 74.55%를 나타내어 중견기업이상합격 취업예측 모델로 사용할 수 있다고 판단된다.

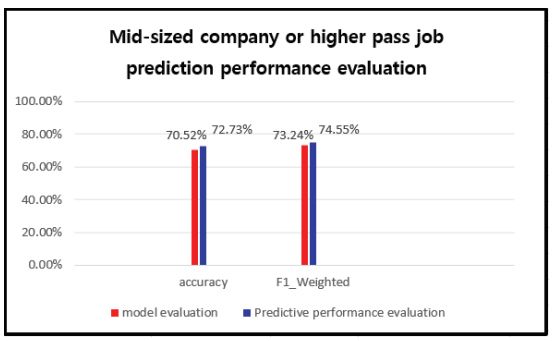

Fig. 4. Mid-sized company or higher pass job prediction performance evaluation.

6. 결 론

M대학교 인턴십 지원자들을 통해 독립변수로 자기소개서를 포함해서 Word2Vec, BERT를 사용한 기계학습 기반으로 대기업 합격 및 중견기업이상 합격 취업예측 모델 개발을 개발하였다. 자기소개서를 포함해서 대기업합격 및 중견기업이상 취업예측모델 개발 시 자기소개서를 포함하지 않을 때보다 정확도가 5.34%, 2.72% 향상되었다. 이번 취업예측 모델 개발에서 독립변수 영향도 Shap 값을 보면 장기현장실습, 일학습병행 등 인턴십 경력이 있어야 취업을 할 수 있고 인턴십 경력이 없다면 본인이 원하는 대기업, 중견 기업 직무에 취업할 수 없다는 것을 더욱 확인할 수 있었다. 또한 M대학교 학생들에게 기계학습 기반의 대기업과 중견기업에 컨설팅 및 취업시킬 수 있는 취업예측 모델 시스템을 마련했으며 이번 모델 개발로 대학과 기업에 신입 취업 및 인력 양성에 있어 서로 협업관계가 잘 진행될 것으로 생각된다. 그러나 이번 논문에는 개발 모델의 한계가 있다. 독립변수를 좀 더 세분화하여 봉사활동, 동아리활동, 경진대회참가 등과 면접 스킬을 적용할 수 있었지만 그렇게 하지 못했고 인턴십 및 신입 채용 결정에 중요한 마지막 면접에 대한 항목 등 관련 면접 내용과 평가 점수도 적용하지 못했다. 앞으로

기업들과 면접 항목 개발과 면접결과 데이터베이스를 구축해 나갈 예정이다. 본 논문에서 추후 보완할 사항은 비교과활동 등 독립변수를 좀 더 세분화 및 최적화하여 모델의 정확도를 향상시켜 나갈 예정이다. 또한 자기소개서 임베딩 기법 최적화와 인턴십 및 신입 채용 결정에 중요한 기업 면접에 대한 항목 등 면접 내용과 평가 점수를 확보해 모델 정확도를 더욱 증가시켜 나갈 것이다. 마지막으로 취업예측 개발모델을 산업별, 직무별로 구축해서 M대학 학생들에게 제공하여 인턴십 및 신입채용 합격률을 더욱 높여 나갈 예정이다.

감사의 글

본 논문이 나오기까지 많은 도움을 준 최기정, 권정을, 박희준학생에게 깊은 고마움을 전합니다.

참고문헌

1. Hyunsoo, K., Sunho, K., Sangjin, H., Minseok, C., and Youngsoo, S., “A Study on the Improvement of Employment Competency through Corporate Field Experience”, J. of The Korean Society of Semiconductor & Display Technology, pp. 78, 2019.
2. Seonyoung, E., “Derivation of a 4-year college student employment prediction model based on university employment support: Focusing on the case of H University”, Hanyang Graduate School, pp. 1-78, 2017.
3. Pilseon, C., and Insik, M., “Employment Prediction Model for College Graduates Using Machine Learning Technique”, Study on Vocational Competency Development, Vol. 21(1), pp. 31-54, 2018.
4. Donghoon, L., and Tae-hyung, K., “A Study on the Prediction Model for Job Seekers for College Graduation Using Machine Learning Technique”, Korea Information System Research, Vol. 29, pp. 287-306, 2020.
5. Breiman, L., “Random Forests, Prediction Games and Algorithms”, pp. 1-33, 1999.
6. Kichang, L., “Korean Embedding,” Acorn Publishing Co., Ltd., 121p, 2020.
7. Minji, B., and Namgyu, K., “Meaning-based search for similar overseas patents through Word2Vec learning”, The Journal of the Korean IT Service Society, Vol. 17, pp. 129-142, 2018.
8. Dowoo, K., and Myunghwan, K., “Classification of Korean Newspaper Articles Based on Convolutional Neutral Network Using Doc2Vec and Word2Vec”, Vol. 44, pp. 742-747, 2017.
9. Delvin, J., Chang, M, W., Lee, K., and Toutanpva, K., “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810. 04805, 2018.
10. Gigon, N., Ji-ho Y., and Sikyun, L., “The Effect of University Activities on Labor Market Performance”, Economic Development Study, Vol. 16, pp. 143-172, 2010.
11. Wonseok, L., and Hyunhee, K., “Interpretable convolutional neural network model for yield prediction in semiconductor fabrication”, Journal of the Korean Society for Data Information and Information Science, Vol. 31, pp. 691 - 720, 2020.

접수일: 2022년 6월 16일, 심사일: 2022년 6월 20일, 게재확정일: 2022년 6월 23일