韓國IT서비스學會誌
第19卷 第1號
2020年 2月, pp.103-112

Journal of Information Technology Services
https://doi.org/10.9716/KITS.2020.19.1.103

Doc2Vec 모형에 기반한 자기소개서 분류 모형 구축 및 실험*

김영수** · 문현실*** · 김재경****

Self Introduction Essay Classification Using Doc2Vec for Efficient Job Matching*

Young Soo Kim** · Hyun Sil Moon*** · Jae Kyeong Kim****

■ Abstract ■

Job seekers are making various efforts to find a good company and companies attempt to recruit good people. Job search activities through self-introduction essay are nowadays one of the most active processes. Companies spend time and cost to reviewing all of the numerous self-introduction essays of job seekers. Job seekers are also worried about the possibility of acceptance of their self-introduction essays by companies. This research builds a classification model and conducted an experiments to classify self-introduction essays into pass or fail using deep learning and decision tree techniques. Real world data were classified using stratified sampling to alleviate the data imbalance problem between passed self-introduction essays and failed essays. Documents were embedded using Doc2Vec method developed from existing Word2Vec, and they were classified using logistic regression analysis. The decision tree model was chosen as a benchmark model, and K-fold cross-validation was conducted for the performance evaluation. As a result of several experiments, the area under curve (AUC) value of PV-DM results better than that of other models of Doc2Vec, i.e., PV-DBOW and Concatenate. Furthmore PV-DM classifies passed essays as well as failed essays, while PV__DBOW can not classify passed essays even though it classifies well failed essays. In addition, the classification performance of the logistic regression model embedded using the PV-DM model is better than the decision tree-based classification model. The implication of the experimental results is that company can reduce the cost of recruiting good d job seekers. In addition, our suggested model can help job candidates for pre-evaluating their self-introduction essays.

Keyword : Machine Learning, Text Mining, Document Classification, Doc2Vec, Self-Introduction Essay

Submitted : January 15, 2020 1st Revision : February 7, 2020 Accepted : February 10, 2020

* 이 논문 또는 저서는 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2014S1A5B8060940).

** 경희대학교 일반대학원 경영학과 석사과정
*** 경희대학교 AI경영연구센터 박사
**** 경희대학교 경영대학 교수, 교신저자

1. 서 론

글로벌 경제 위기 이후 양질의 일자리 공급이 감소하고, 기업에서는 경력직과 비정규직을 선호하고 있으며, 고학력자 증가로 인해 구직자들은 대기업으로 몰리고 있다(김정수 외, 2016). 기업들은 다양한 채용 방식을 통해 기업이 원하는 역량있는 인재들을 선발하기 위해 많은 비용과 시간을 소모하고 있다(천영민, 2017). 구직자들이 작성하여 기업에 제출하는 자기소개서는 취업의 당락에 직접적인 영향을 끼치는 것으로 인식되고 있다(신정숙, 2015). 데이터 마이닝(Data mining) 및 최근 딥 러닝(Deep Learning)으로 대표되는 기계 학습의 발달로 정형 데이터뿐만 아니라, 이미지와 소리, 텍스트 등의 비정형 데이터를 저장, 관리 및 분석을 하는 것이 용이하게 되었다. 최근 텍스트 데이터를 활용한 연구가 다양한 분야에서 진행하고 있다. 특히 머신 러닝을 이용한 텍스트 마이닝은 텍스트 분류, 감정 분석, 토픽(Topic) 추출하는 단계를 넘어, 딥 러닝을 활용한 텍스트 생성에서 더 나아가 신문 기사 및 소설을 작성하는 단계까지 발전하고 있다. 텍스트 분석에 활용되는 데이터로는 상품 리뷰(review), 뉴스 기사, 소셜 네트워크 서비스(SNS), 백과사전 등의 데이터가 있으며, 다양한 분야에서 텍스트 마이닝을 이용한 연구가 활발히 진행 중이다. 하지만 취업과 관련된 데이터를 이용한 텍스트 마이닝 연구는 다른 분야의 연구에 비해 지금까지 많은 연구가 되고 있지 않다. 따라서 본 연구에서는 취업에 사용하는 자기소개서 데이터를 이용한 텍스트 마이닝 연구를 진행하여, 구직자가 작성한 자기소개서를 기업에서의 서류전형 합격 여부를 미리 예측할 수 있는데 도움을 주고자 한다. 이를 통해 기업 입장에서는 많은 비용과 노력을 기울이지 않고, 구직자들이 작성한 방대한 양의 자기소개서를 사전 필터링(filtering)하는데 도움을 주고자 한다. 이러한 연구 목적을 위하여 본 연구에서는 문서 분류에 우수한 성능을 보이는 Doc2Vec 모델과 로지스틱 회귀 모델(Logistic Regression model)을 결합한 분류 모델을 제시하였으며, 제시한 모델의 정확도를 비교 분석하기 위하여 의사결정 나무를 이용하였다. K-겹 교차분석 방법으로 실제 데이터를 대상으로 실험을 진행하였다. 실제 실험에서는 PV-DM 모델을 이용하여 임베딩(Embedding)한 로지스틱 회귀 모델의 분류 성능이 Doc2Vec의 다른 모델인 PV-DBOW와 Concatenate 모델 보다 특성 곡선 아래 영역(Area Under Curve; AUC) 값이 가장 우수하게 나타났다. 이는 분류 정확도가 가장 좋음을 의미한다. 또한 이 연구에서 제시한 PV-DM모델을 이용하여 임베딩(Embedding)한 로지스틱 회귀모델의 분류 성능이 서류전형 불합격자뿐만 아니라 합격자를 대상으로 한 실험에서도 PV-DBOW와 Concatenate 모델 보다 더 우수하게 나타났다. 본 연구를 통해 구인기업은 좋은 인재를 선발하기 위한 시간과 비용을 절감할 수 있을 것이며, 또한 구직자에게는 본인의 자기소개서를 연구에서 제시한 분류 모델에 대입함으로써 합격 가능성을 미리 추정할 수 있는 보조자료 역할을 할 수 있을 것으로 기대한다.

2. 이론적 배경

2.1 텍스트마이닝

인터넷과 SNS 등 다양한 방면에서 최근 방대한 양의 텍스트 데이터가 나타나고 사라지고 있다. Hearst(1999)와 Sebastiani et al.(2002)는 텍스트 마이닝을 다량의 비정형 데이터를 분석하여 이전에는 찾을 수 없던 새롭고 의미 있는 정보를 추출하는 과정으로 정의했다. 텍스트 마이닝을 통해 분류, 군집화, 연관 관계 분석 뿐만 아니라 자연어 처리, 정보 검색, 토픽 분석, 텍스트 범주화 등을 할 수 있다(Mooney et al., 2015; Stavrianou et al., 2017). 텍스트 마이닝 중 텍스트 분류를 하기 위해서는 텍스트를 정형 데이터로 변환해야 하는데, 이때 텍스트를 벡터(Vector) 형식으로 표현하는 임베딩 방식이 가장 일반적으로 사용된다(정지수 외, 2019). 임베딩 방식을 이용한 방법으로 TF-

IDF, Word2Vec, Doc2Vec 등이 있다(정지수 외, 2019). 김영수 외(2018)의 연구에 따르면 텍스트 처리와 워드 임베딩 모델의 적절한 조합을 찾으면 분류 모델 성능 향상을 도모 할 수 있음을 제시하였다. 최도한 외(2013)는 기술 특허와 관련한 토픽을 분석하여 특정 기술의 동향을 파악하는 연구를 하였고, 백민지 외(2018)는 한국과 일본의 특허 문서에서 사용된 용어 분석을 통해 기술용어의 유사성을 산출하고 이를 기반으로 의미기반 해외 유사 특허를 검색하는 방안을 제시하였다. 송혜지 외(2013)는 국내 경제 관련 학술 논문의 토픽 분석으로 경제 분야의 연구 동향을 분석했다. 박상현 외(2017)는 아마존 사이트의 온라인 후기 데이터를 이용해 토픽모델링과 인공신경망 모델을 사용하여 상품의 평점을 예측하는 모델을 구축하였다. 채민성 외(2012)는 SNS 게시물을 분석하여 SNS 친구에 대한 친밀도를 알 수 있는 시스템을 제안하였다.

2.2 텍스트 분류

텍스트 분류의 전통적인 알고리즘 모델로는 나이브베이즈(NaiveBayes), 서포트 벡터 머신(Support Vector Machine), 랜덤 포레스트(Random Forest)가 있고, 최근에는 신경망의 발달로 CNN(Convolution Neural Network), RNN(Recurrent Neural Network)에 의한 텍스트 및 문서 분류 연구가 진행중에 있다(김나랑 외, 2019). 방대한 정보를 직접 처리하기 힘든 한계점을 극복하기 위해 다양한 분야에서 텍스트 분류를 기술을 적용하고 있다(이재성 외, 2018). 규칙 기반의 기존 텍스트 분류 기술은 기계 학습을 기반으로 이루어지고 있다(Yang and Liu, 1999; Yang, 1999; Sebastiani, 2002; Hong et al., 2014). 문서 분류를 적용한 분야는 민원, 학회, 기사, 감성 분류 등 다양한 분야에서 분류 기술을 적용하고 있다(김나랑 외, 2018; 이수경 외, 2017; 김유영 외, 2016; 김도우 외, 2017). 하지만 자기소개서를 데이터를 이용한 문서 분류 연구는 미미한 실정이다. 전상홍 외(2019)의 연구에서 자기소개서를 분류하였는데 TF-IDF를 이용하여 임베딩하고, 의사결정나무로 분류하였다. 본 연구에서는 자기소개서 데이터를 이용하여 합격, 불합격자를 분류하여 구직 과정에서 소요되는 시간과 비용을 절감하는데 기여하고자 한다.

2.3 Doc2Vec

Doc2Vec은 단어 기반 모델인 Word2Vec에서 확장된 개념이다(Quec et al., 2014). Doc2Vec은 문장, 문단 기반 모델로 Word2Vec에 비해 문서 전체에 대한 고려가 잘 이루어진다. 동일한 문서에 등장한 단어들이 서로 유사성을 갖도록 문서 벡터 값을 조정하며 학습한다(김동성, 2019). 문장 기반 모델인 TF-IDF(Term Frequency-Inversed Document Frequency)를 적용해서 문서 관련 모델에 빈도를 활용하는 점이 Word2Vec과 유사한 특징을 보인다. 하지만 Word2Vec은 고정 크기의 벡터를 생성하지 않는 반면, Doc2Vec은 사용자가 지정하는 고정 크기의 벡터를 생성함으로써 자원 사용량 낭비 및 훈련 시간 증가 등의 비효율적인 낭비를 줄일 수 있다(김도우 외, 2017).

Doc2Vec에는 DM(Distributed Memory) 모델(Model)과 DBOW(Distributed bag Of Words) 모델이 있다. PV-DM은 Word Vector에 문장 벡터(Paragraph Vector)를 추가한 모델이다. 문장 벡터는 메모리(memory)와 유사한 기능을 가져서 단락의 주제나 문맥에서 빠진 내용을 기억하는 역할을 한다. PV-DM은 벡터(Vector)를 입력 데이터로 활용한다. 연구자가 벡터 크기(Vector size)를 설정하고, 해당 크기만큼 문장을 학습하는데, 이때 창(Window)만큼 옆으로 이동하면서 다음 단어를 예측하는 방식으로 학습한다. 학습 내용을 문장 벡터에 저장 후 최종적으로 발생하는 벡터를 해당 문서 전체의 벡터로 정의하여 문서를 분류하는 방식이다. 반면에. PV-DBOW는 단어 벡터(Word Vector)를 따로 생성하지 않고, 문장 벡터만으로 학습하여 단어를 무작위로 예측하는 모델이다. 해당

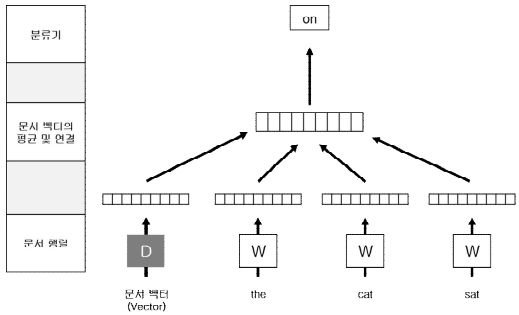
모델은 단어 벡터를 생성하지 않기 때문에 메모리 관리에 용이하다는 특징이 있다(Quec et al., 2014). DM모델과 DBOW 모델을 합친 조합 모델(Concatenated Model)도 있다. 조합 모델의 경우 앞서 설명한 두 모델보다 우수한 분류 성능을 보이기도 하였다(Quec et al., 2014).

Dai et al.(2015)는 영어 위키피디아(wikipedia) 레이디 가가(Lady Gaga) 관련 문서에 Doc2Vec을 적용하여 동시대 여성 팝가수인 리한나(Rihanna)와 비욘세(Beyonce) 같은 인물과의 유사성을 찾아냈다. Jiang 외(2016)는 트립 어드바이저(Trip Advisor)와 옐프(Yelp)의 리뷰데이터로 Word2Vec과 Doc2Vec의 성능을 비교하였는데, 두 모델이 유사한 성능을 나타내어 Doc2Vec이 특징 추출(Feature extraction)과 분류(classification)에 이용될 수 있음을 시사했다.

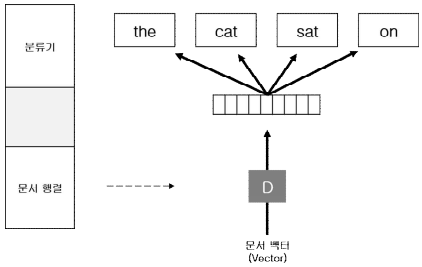
Doc2Vec은 특징 추출과 분류가 동시에 가능하기에 연구에 소요되는 시간을 절감할 수 있다는 장점이 있다(육지희, 2018). 김도우 외(2017)는 한국어 기사를 분류하는 연구를 진행하였는데, word2Vec 모델을 단독으로 사용하였을 때보다 Doc2Vec을 결합한 모델의 분류 성능이 우수한 성능을 보였다. 문서 분류, 감정 분석, 정보 검색에 좋은 성능을 보인다는 Quec et al.(2014)의 연구 결과에 따라서 본 연구에서는 Doc2Vec 모델을 사용했다.

2.4 K겹 교차검증

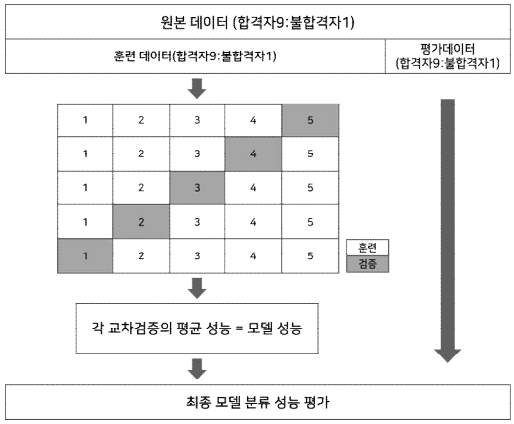
연구 모델을 구축하기 위해 데이터 셋(Data Set)을 학습용 데이터(train data)와 평가용 데이터(test data)로 나누는 작업을 실시한다. 데이터 셋을 분할하는 방법으로 연구자가 임의의 비율로 데이터를 나누거나, 시계열 데이터의 경우 시간을 반영하여 데이터를 분할하기도 한다. 대부분의 경우, 무작위로 분할하여 데이터 편향을 피하는 방법을 선택한다(Galit Shmueli, 2017). 교차 검증은 주어진 자료를 반복적으로 성능을 측정하여 성능 결과를 평균한 것으로 모델을 평가한다. 본 연구에서 사



[그림 1] PV-DM 모델(Model) 구조



[그림 2] PV-DBOW 모델 구조



[그림 3] K겹 교차 검증

용하는 데이터는 합격자와 불합격자 데이터 비율이 1:9로 편향되어 있어, 데이터 편향으로 인해 발생하는 문제점을 교차 검증을 통해 보완하고자 하였다. 하나의 평가 데이터만 사용하는 것이 아닌 서로 다른 평가 데이터를 이용하여 여러 차례 검증하는 작업을 통해 비교적 정확한 검증 값이 나올 것으로 예상하여 교차 검증을 실시하였다.

3. 연구방법

본 연구는 9,226명의 자기소개서 데이터를 활용하여 합격자와 불합격자를 분류하는 실험을 진행하였다.

1단계. 데이터 전처리

원본 데이터(Raw Data)에는 인적성검사 합격, 필기합격, 신체검사합격, 면접 합격 등을 나타내는 '현재 상태' 속성이 총 10단계로 구분되어 있다. 일부 속성의 경우 모델이 충분히 학습하기에 데이터 수가 적어 '현재 상태'를 '1차 면접 합격' 이상의 전형 합격자를 '합격자'로 정의하여 '현재 상태'를 이진화하여 합격 여부를 분류했다.

또한 파이썬(Python)을 사용하여 한국어 자연어 처리를 위해 코엔엘파이(KoNLPy)의 트위터(Twitter) 모듈을 이용하여 명사를 추출했다. 분류 정확도를 높이기 위해 불용어 사전에 '저', '이', '그' 등의 의미 없는 명사 810여 개를 추가하여 데이터 분석에 사용하지 않도록 설정했다. 이와 함께 합격 여부(0,1)를 나타내는 태그(tag)를 저장하는 작업을 함께 진행했다.

2단계. 데이터 표본 추출

전체 데이터 중 합격자와 불합격자의 데이터 비율이 약 9:1의 비율로 나타났다. 이는 불합격자의 데이터 수가 적어 데이터가 희소한 문제가 발생한다. 층화표본추출법을 이용하여 합격자와 불합격자의 비율(9:1)을 유지하면서, 학습데이터와 검증데이터를 7:3 비율로 분리하여 모델 학습을 진행했다. 본 연구에서는 이진화 데이터(합격, 불합격) 분류에 적합한 로지스틱 회귀 분석 모델을 이용하였다. 벤치마크 모델로 의사결정나무 모델을 이용하여 문서를 분류하였는데, 벤치마크 모델 또한 명사만을 추출하였으며, 임베딩 방식은 TF-IDF를 이용하였고, 의사결정나무 모델을 통해 구직자를 합격, 불합격으로 분류하였다.

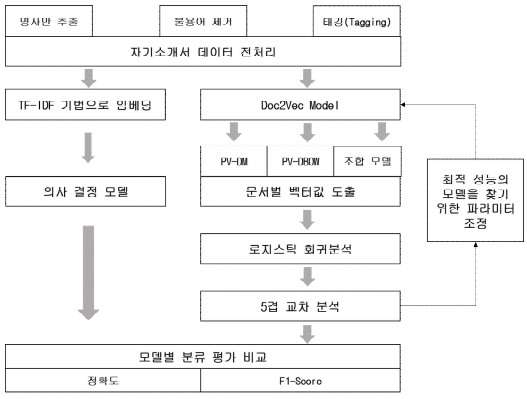
3단계. Doc2Vec 모델의 최적 속성 값 탐색

Doc2Vec 모델의 최적 파라미터를 찾기 위해, 학습 훈련 횟수(epoch), 벡터 크기(vector size), 창(window) 크기 등 파라미터를 다르게 하며 수차례 실험한 결과, PV-DM 모델의 경우 벡터 크기 = 100, 최소 단어 등장 수(min_count) = 6, 창 크기(windows) = 4일 때, 분류 정확도가 가장 높았다. PV-DBOW 모델의 경우 벡터 크기 = 300, 최소 단어 등장 수 = 4일 때, 분류 정확도가 높았다. PV-DM과 PV-DBOW 모델을 조합한 모델의 성능이 우수하다는 기존 연구에 따라(Quec et al., 2014), 위 실험에서 도출한 파라미터의 모델을 합친 조합 모델을 추가로 생성했다. 모델 성능을 평가할 때 K-겹 교차 검증(K-fold cross validation)을 이용하였다. 이때, K 값을 5로 설정하여 5번의 교차분석을 걸쳐 성능의 평균 값을 추출하여 보다 정확한 결과 값을 도출하고자 하였다.

4단계. 분류 성능 평가

본 연구에서 실험한 모델 성능을 정확도(accuracy), F1-score를 이용하여 분류 성능을 평가했다. 정확도(Accuracy)는 전체 샘플(sample) 중 정확하게 예측한 샘플 수의 비율을 의미한다. 정밀도(Precision)는 합격자(또는 불합격자) 클래스에 속한다고 예측한 샘플 중, 실제로 합격자(또는 불합격자) 클래스에 속하는 샘플 수 비율을 의미한다. 재현율(Recall)은 실제 합격자(또는 불합격자) 클래스에 속한 샘플 중 정확하게 예측한 샘플 수 비율을 나타낸다. F1-score는 정밀도와 재현율의 가중조화평균 값을 나타낸다.

본 연구는 [그림 4]에 표현한 작업 흐름도의 절차대로 실험을 진행했다.

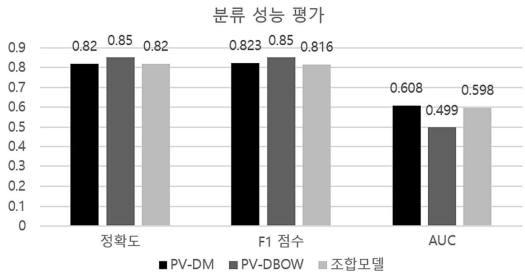


[그림 4] 작업 흐름도

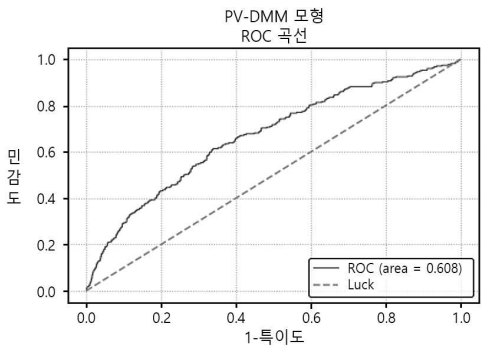
4. 연구결과

Doc2Vec의 모델 중 concatenate > DM > DBOW 모델의 순서로 성능이 좋다는 기존 연구에 따라 3가지 모델을 모두 실험했다(Quec et al., 2014). DM, DBOW 두 모델의 최적 파라미터를 찾은 후, 각각 최적의 모델을 조합한 (Concatenate) 모델을 이용하여 문서 임베딩을 다시 진행하였다. 합격자, 불합격자 분류 작업에는 이진 분류에 가장 우수한 성능을 보이는 로지스틱 회귀분석 모델을 이용하여 진행했다. 3가지 모델에서 실제 불합격자를 불합격자로 분류한 정확도는 평균적으로 약 83%를 기록했다. 그러나 합격자를 올바르게 분류한 정확도는 평균 약 23%로 비교적 낮은 수치를 기록했다. 전체 가중 평균값은 약 83%로 나타나는 것으로 보아 해당 문제는 학습데이터 셋에서 합격자의 데이터 수가 충분하지 않아서 발생하는 편차로 판단했다. 여러 파라미터를 조절하며 얻은 PV-DM과 PV-DBOW 모델, 그리고 조합 모델을 이용하여 각각 실행한 결과 PV-DBOW의 모델의 가중 평균 정확도 값이 가장 높은 결과(0.85)를 보였다. F1 점수 또한 PV-DBOW 모델이 가장 높은 점수(0.85)를 나타냈다. 하지만 합격자를 정확하게 분류하는 모델은 DM모델이 가장 높은 정확도(0.22)를 보였으며, PV-DBOW 모델은 전혀 분류할 수 없는 결과를 보였다. 이는 민감도와 특이도를 통해 표현하는 ROC 그래프의 면적을 나타내는 AUC 값으로 표현되는데, DM 모델의 AUC(Area Under Curve; AUC)값이 0.608로 가장 높게 나타났다. 도출된 AUC 값을 통해 DM 모델이 나이브(Naïve) 규칙 모델과 다른 모델보다 우수한 성능을 보이고 안정된 예측을 할 수 있다는 의미로 해석할 수 있다. 벤치마크 모델인 의사결정나무는 모든 데이터를 불합격으로 분류하였다. 합격과 불합격 데이터의 불균등성 때문에 모델의 정확도는 높지만, 합격자 모두를 올바르게 분류하지 못하면서 유의미한 분류 결과를 보이지 못하였다.

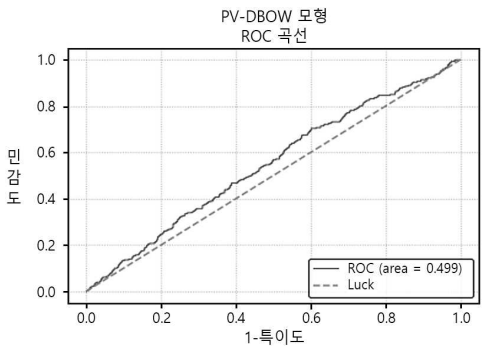
[그림 5]를 통해 각 모델의 분류 성능 평가 값을 정리했다. [그림 6], [그림 7], [그림 8]을 통해 각 모델의 ROC 그래프와 AUC값을 나타냈다.



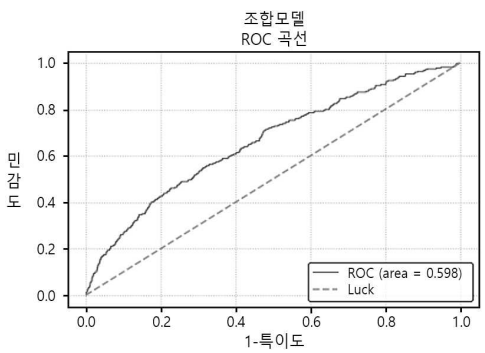
[그림 5] 분류 성능 평가



[그림 6] PV-DM 모델 ROC 그래프



[그림 7] PV-DBOW 모델 ROC 그래프



[그림 8] 조합 모델 ROC 그래프

5. 결론 및 시사점

본 연구에서는 9,226명의 자기소개서 데이터를 이용하여 구직자의 합격 여부를 분류하는 연구를 진행하였다. 연구 방법으로 기존의 Word2Vec에서 발전한 모델인 Doc2Vec 모델을 이용하여 문서를 임베딩하였다. 이진 데이터 분류에 우수한 성능을 보이는 로지스틱 회귀 모델을 이용하여 임베딩한 문서를 합격, 불합격 문서로 이진 분류하였다. 해당 모델의 성능을 비교하기 위해 TF-IDF를 이용하여 임베딩하고, 의사결정나무 모델을 이용한 분류 모델을 벤치마크 모델로 활용하였다. 분류 결과 PV-DBOW 모델의 경우 전체적인 문류 정확도가 조금 높게 나타났으나 합격자에 대한 분류는 전혀 되지 않은 결과를 보였다. 이는 자기소개서 분류에는 적합하지 않음을 의미하며, 기존의 Doc2Vec 분석 결과와도 일치한다.

이번 연구에서 문서 임베딩 방식을 Doc2Vec을 채택하였는데, Doc2Vec은 문서의 태그를 함께 저장함으로써 기존의 Word2Vec이나 TF-IDF 방식보다 비교적 쉽게 문서를 분류할 수 있음을 보였다. 기존 연구에서는 Doc2Vec의 3가지 모델 중 또한 자기소개서 데이터를 활용한 기존 연구가 미미하였는데, 본 연구에서 자기소개서 데이터를 활용함으로써 텍스트 마이닝 기술의 적용 분야를 한 층 더 넓혔다는데 의의를 두었다. 자기소개서를 자동으로 합격자, 불합격자를 분류하는 이번 연구를 기업에 적용하면 인재 채용 프로세스(process)에 소모되는 비용 및 시간을 절감할 수 있을 것으로 보인다. 또한, 취업을 준비하는 구직자에게 본 서비스를 제공하면, 구직 기업에 자기소개서를 제출하기 전에 미리 평가를 받은 후 보완하여 수정할 수 있는 가이드 라인(Guide Line)을 제시할 수 있을 것으로 생각한다. 이를 통해 보다 완성도 높은 자기소개서로 기업에 지원함으로써 구직 활동에 도움이 될 수 있을 것으로 생각한다.

본 연구의 한계점으로는 크게 3가지로 꼽을 수 있다. 첫 번째로 상대적으로 데이터가 부족한 합격자를 올바르게 분류한 비율은 매우 낮게 나타난 점이 한계점이다. 데이터 불균형 문제를 해결하기 위해 층화표본추출법을 이용한 데이터 분할, 교차검증 등을 적용하였다. 적용 전보다 성능이 미미하게 개선되었지만, 연구자의 기대치(95% 정확도)에는 미치지 못한 결과를 보였다. 실효성 있는 모델을 구현하기 위해서는 보다 많은 자기소개서 데이터와 서류 전형에 합격한 자기소개서 자료가 더 필요하다고 판단된다. 두 번째로 합격자를 구분하는 기준이 모호하다는 점을 본 연구의 한계점으로 꼽을 수 있다. 각 기업마다 인재 채용에 대한 상이한 심사 기준과 프로세스를 운영하고 있다. 자기소개서를 서류 전형에서 검토하거나 면접 전형에서 검토하는 등의 자기소개서를 활용하는 시점이 다양할 것으로 생각된다. 하지만 본 연구에서는 서류 전형 이후의 채용 과정을 합격자로 분류함으로써 서류 전형에서 자기소개서를 검토할 것으로 기준을 세우고 연구를 진행했다. 이를 보완하기 위해서 다양한 기업의 실무자의 의견을 검토하여 기업마다 다른 채용 기준과 프로세스를 데이터 전처리 과정에서 적절하게 반영하여 본 연구에서 제안하는 모델에 적용하는 것이 필요하다.

이번 연구에서는 자기소개서의 명사만을 이용하여 문서를 분류하였는데, 추후 연구에서는 다양한 품사를 추가하는 방법도 고려할 수 있을 것으로 보인다. 불용어 사전을 생성하는 과정에서 본 연구보다 면밀히 보완하여 작성한다면 성능 개선을 할 수 있을 것으로 생각한다.

본 연구에서 사용한 자기소개서 데이터를 이용해 토픽 모델링(Topic modeling)을 통한 각 기업별 인재상과 자기소개서 키워드를 매칭(Matching)하거나, 유사한 자기소개서의 특징을 가진 구직자를 특정 기업에 추천해주는 연구로도 발전할 수 있을 것으로 생각한다. 추후에 개선된 모델을 통해 구직자가 자기소개서를 작성할 때 가이드 라인 역할로써 참고 용도로도 사용할 수 있을 것으로 생각한다.

참고문헌

김나랑, 마렌드라 라마디니, “CNN과 Bidirectional LSTM을 활용한 부산시 민원 자동 분류 연구”, *전산회계연구*, 제17권, 제2호, 2019, 81-98.

김도우, 구명완, “Doc2Vec과 Word2Vec을 활용한 Convolutional Neural network 기반 한국어 신문 기사 분류”, *정보과학회논문지*, 제44권, 제7호, 2017, 742-747.

김동성, “Doc2Vec 단어 임베딩 언어 모델을 활용한 텍스트 장르 구분”, *언어와 정보*, 제23권, 제2호, 2019, 23-43.

김영수, 이승우, “문서 분류를 위한 신경망 모델에 적합한 텍스트 전처리와 워드 임베딩의 조합”, *정보과학논문지*, 제45권, 제7호, 2018, 690-700.

김정수, 이석준, “취업준비생 토픽 분석을 통한 취업난 원인의 재탐색”, *경영과 정보연구*, 제35권, 제1호, 2016, 85-116.

박상현, 문현실, 김재경, “토픽 모델링에 기반한 온라인 상품 평점 예측을 위한 온라인 사용 후기 분석”, *한국IT서비스학회지*, 제16권, 제3호, 2017, 113-125.

백민지, 김남규, “Word2Vec 학습을 통한 의미 기반 해외 유사 특허 검색 방안”, *한국IT서비스학회지*, 제17권, 제2호, 2018, 129-142.

송혜지, 박경수, 정혜은, 송 민, “텍스트 마이닝 기법을 활용한 한국의 경제연구 동향 분석”, *한국정보관리학회 학술대회논문집*, 제2013권, 제8호, 2013, 47-50.

신정숙, “취업용 자기소개서 지도방안 연구”, *동남어문논집*, 제1권, 제40호, 2015, 83-113.

육지희, 송 민, “토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구”, *정보관리학회지*, 제35권, 제2호, 2018, 63-88.

이재성, 전승표, 유형선, “한국표준산업분류를 기준으로 한 문서의 자동 분류 모델에 관한 연구”, *지능정보연구*, 제24권, 제3호, 2018, 221-241.

전상홍, 문현실, 김재경, “의사결정나무에 기반한 취업 지원자의 지원결과 분석”, *한국IT서비스학회 학술대회 논문집*, 제2019권, 제2호, 2019, 240-243.

정지수, 지민규, 고명현, 김학동, 임헌영, 이유림, 김원일, “문서 유사도를 통한 관련 문서 분류 시스템 연구”, *방송공학회논문지*, 제24권, 제1호, 2019, 77-86.

채민성, 인관호, 김응모, “텍스트, 오피니언 마이닝을 이용한 SNS 친구 친밀도 분석 시스템”, *한국정보과학회 학술발표논문집*, 제39권, 제2호(C), 2012, 98-100.

천영민, “기업 인재상 분석과 직무역량 기반 채용 확산”, *한국직업자격학회 동계학술대회*, 제12권, 2017, 33-66.

최도한, 김갑조, 박상성, 장동식, “텍스트 마이닝 기반의 특허키워드 정량분석을 이용한 AMOLED 부상기술 예측”, *한국콘텐츠학회 종합학술대회 논문집*, 제2013권, 제5호, 2013, 365-366.

Dai, A.M., C. Olah, and Q.V. Le, “Document embedding with paragraph vectors”, arXiv e-prints, 1507.07998, 2015.

Shmueli, G., P.C. Bruce, and N.R. Pateli, “Data Mining for Business Analytics Concepts, Techniques, and Applications In R”, WILEY, 2017.

Hearst, M.A., “Untangling text data mining, In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics”, *Association for Computational Linguistics*, 1999, 3-10.

Hong, J.S., N. Kim, and S. Lee, “A Methodology for Automatic Multi-Categorization of Single-Categorized Documents,” *Journal of Intelligence and Information Systems*, Vol.20, No.3, 2014, 77-92.

Mooney, R.J. and R. Bunescu, “Mining knowledge from text using information extrac-

tion", *ACM SIGKDD explorations newsletter*, Vol.7, No.1, 2005, 3-10.

Le, Q. and T. Mikolov, "Distributed representation of sentences and documents", *Proceedings of the 31st International Conference on Machine Learning, PMLR*, Vol.32, No.2, 2014, 1188-1196.

Jiang, S., J. Lewris, M. Voltmer, and H. Wang, "Integrating rich document representations for text classification", *Systems and Information Engineering Design Symposium (SIEDS) 2016 IEEE*, 2016, 303-308.

Sebastiani, F., "Machine learning in automated text categorization", *ACM computing surveys(CSUR)*, Vol.34, No.1, 2002, 1-47.

Stavrianou, A., P. Andritsos, and N. Nicoloyannis, "Overview and semantic issues of text mining", *SIGMOD Record*, Vol.36, No.3, 2007, 23-34.

Yang, Y. and X. Liu, "A Re-examination of Text Categorization Methods", *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99)*, 1999, 42-49.

Yang, Y., "An Evaluation of Statistical Approaches to Text Categorization", *Journal of Information Retrieval*, Vol.1, No.1, 1999, 67-88.

About the Authors

김 영 수 (sooperman@khu.ac.kr)

현재 경희대학교 일반대학원 경영학과 빅데이터 경영을 전공으로 석사과정에 재학 중이다. 텍스트 마이닝을 비롯한 데이터 마이닝, 추천 시스템 등에 관심을 갖고 연구하고 있다.

문 현 실 (pahunter@khu.ac.kr)

경희대학교에서 경영학 학사, 동 대학원에서 경영정보시스템(MIS) 전공으로 석사 및 박사학위를 취득하였다. 주요 관심분야로는 빅데이터 분석, 추천 시스템, 텍스트 마이닝, 네트워크 분석 등이다. International Journal of Information Management, Asia Pacific Journal of Information System, 지능정보연구, IT서비스학회지 등에 관련 논문을 게재하였다.

김 재 경 (jaek@khu.ac.kr)

서울대학교에서 산업공학 학사, 한국과학기술원에서 산업공학 전공으로 석사 및 박사학위를 취득하였다. 현재 경희대학교 경영대학 교수로 재직 중이며 경영대학원장을 역임하고 있다. 주요 관심분야로는 Business analytics, 개인화 서비스, IoT(Internet of Things), 딥러닝 등이다. IEEE Transactions on services computing, IEEE Transactions on Systems, Man, and Cybernetics, International Journal of Human-Computer Studies, International Journal of Information Management, Information & Management 등 다수의 학술지에 논문을 게재하였다.