경영정보학연구
제25권 제3호, 2023년 8월 http://dx.doi.org/10.14329/isr.2023.25.3.027

머신러닝 기반의 기업 리뷰 다중 분류: 부분 문법 적용을 중심으로

Multi-Label Classification for Corporate Review Text: A Local Grammar Approach

백 혜 연 (HyeYeon Baek) 서강대학교 경영대학 인사조직전략 박사과정
장 영 균 (Young Kyun Chang) 서강대학교 경영대학 인사조직전략 교수, 교신저자

요 약

최근 많은 분야에서 기계학습에 대한 연구가 활발히 진행되고 있는데, 상당수의 연구들이 학습 모델의 성능을 개선하는 최신 방법론을 제시하고 있다. 본 연구에서는 방법론의 개발 못지않게 기계학습에 투입되는 훈련용 데이터의 '품질'을 개선하는 것 역시 중요하다는 점에 착안하여, 코퍼스 분석에서 자주 사용되는 '부분 문법' 처리 프로세스를 통해 훈련 데이터의 품질을 향상시키는 방법을 제시한다. 우리나라 100대 기업에 근무하는 재직자들이 채용플랫폼에 게시하는 방대한 양의 비정형 기업 리뷰 텍스트 데이터를 수집하고, 데이터 품질을 부분 문법 프로세스로 개선한 후, 부분 문법이 적용된 분류 모델이 적용되지 않은 모델보다 분류 성능이 우수함을 확인하였다. 분류 카테고리는 직원 몰입의 5가지 요인으로 상정하였는데, 국내 직장인들이 기업 리뷰가 각 유형별로 빈도에 차이가 있는지를 분석하였다. 추가로 리뷰 양상이 코로나 팬데믹 전후로 어떠한 변화가 있었는지도 분석하였다. 본 연구를 통해 국내 직장인들의 생생한 일터 경험들을 자동적으로 식별하고 분류하여, 이직을 포함한 주요한 조직문화 현상의 행태와 유발 원인 등을 유추해 볼 수 있는 근거를 제공한다.

키워드 : 다중 레이블 분류, 기계 학습, 부분 문법, 기업 리뷰, 직원 경험

Ⅰ. 서 론

사상 초유의 글로벌 팬데믹으로 인해 평생직장의 개념이 희미해지고, 일과 삶의 균형에 대한 직장인들의 열망이 더해지면서 이직이 증가하고 있다. 미국의 경우 팬데믹 기간 동안 이직률이 역사적 고점을 형성하고 있는데(BBC News, 2021), 소위 '대퇴사 시대'(The Great Resignation)라고 불리며 직장인들의 잦은 이직 현상이 공론화되고 있다(Ducci, 2021). 또한 이직을 실제 실행에 옮기지 못하는 상황에서조차 재직은 유지하되 자신의 업무에 최소한의 수준으로 관여하는 소위 '조용한 사직(Quiet Quitting)'이라 불리는 특이 현상도 목격되고 있다(Harter, 2022). 미국보다는 덜 하지만 우리나라도 IT 기술인력들과 MZ 세대를 중심으로

† 본 연구는 2022년도 서강대학교 교내연구비 지원을 받아 수행되었음(과제번호 : 202212019.01).

높은 이직률을 보이고 있다(EBS News, 2022). 이러한 현상들은 기업의 인력관리나 조직운영 측면에서 명백한 부담 요인이다. 실제로 많은 조직에서 구성원들의 몰입도(engagement)와 재직 의지를 높이기 위한 다양한 노력을 기울이고 있다.

이러한 노력의 일환으로 경영진들은 직원들의 다양한 일터 경험을 인식하고 개선하기 위해 사내 소통에 힘쓰고 있다. 그러나 국내 기업의 구성원들은 사내 소통 채널에서 자신의 솔직한 일터 경험을 공유하는 것을 꺼리거나 침묵으로 일관하는 경향이 강하다. 상명하복의 위계적 업무 형태가 지배적이고, 자신의 의견이 공유되면 불이익이 발생할 수 있다는 불안감 때문이다. 직원들의 실제 경험 세계는 사내에서 온전히 수렴되지 못하는 반면, 익명이 보장되는 외부 온라인 기업 리뷰 플랫폼(예: 블라인드, 잡플래닛 등)에서는 솔직한 일터 경험들이 무수히 게시된다. 즉, 기업 리뷰 플랫폼이 직원들의 실제 일터 경험에 관한 방대한 양의 데이터를 확보하고 있는 셈이다.

기업 리뷰 플랫폼에 게시된 직원들의 일터 경험들은 기계학습의 중요한 자원이 될 수 있다. 비정형 텍스트 형태를 띠고 있고, 그 양이 방대한 빅데이터이므로 기계학습에 사용이 가능하기 때문이다. 그럼에도 불구하고, 그간 인사조직과 경영관리연구에서는 기계학습 측면에서의 기업 리뷰 데이터가 거의 활용되지 못했다. 현행 기계학습에 자주 사용되는 데이터들은 직원들의 일터 경험과는 거리가 먼 데이터(예: 네이버 영화 리뷰, 온라인 쇼핑 리뷰 등)이므로 직원들의 행동을 연구하는 인사조직 분야의 연구에 직접적으로 사용하기에는 한계가 있다.

본 연구의 목적은 매일같이 쏟아지는 방대한 양의 기업 리뷰 데이터를 자동 처리하여 조직관리상의 중요한 인사이트를 쉽고 빠르게 도출함으로써 조직 관리자들의 의사결정의 질과 업무 생산성을 높이는 데 있다. 본 연구에서는 '잡플래닛'이라는 기업 리뷰 플랫폼상에 존재하는 비정형 텍스트 데이터를 활용하되, 부분 문법(local grammar) 처리 프로세스(Gross, 1997; 백혜연 등, 2021)를 적용하여 현존하는 기계학습 모델보다 개선된 성능의 모델을 구축하였다. 부분 문법을 적용한 이유는 기계 학습용 데이터 구축 단계에서 학습에 사용되는 데이터의 정확한 식별을 통해 향상된 분류 결과를 얻게 해주는 고품질의 학습 데이터 생성이 가능하기 때문이다. 예를 들어, 서술어까지 고려하는 부분 문법을 적용하지 않고 가장 흔한 방식인 명사 단위로 텍스트를 식별할 경우, '파리'라는 단어학습 시 프랑스의 수도 파리(Paris)와 곤충류인 파리(fly)가 구분 없이 동일한 의미를 가진 어휘로 식별되어 학습용 데이터 품질이 저하된다. 그러나 부분 문법은 '파리'라는 단어를 식별할 때, '파리가 날았다'와 '파리에 도착했다' 같이 서술어까지 고려하여 곤충으로서의 파리와 도시로서의 파리를 구분하여 식별할 수 있다.

부분 문법을 적용하여 구축된 학습 모델을 통해 우리나라 100대 기업에 근무하는 재직자들이 표출하고 있는 일터 경험들은 무엇이고, 각각의 경험이 직원의 몰입에 영향을 주는 요인으로 자동 분류하기 위해 몰입 요인을 5개로 유형화하였다. 그리고 각 유형 별로 경험 빈도에 차이가 있는지를 분석하였다. 추가적으로 일터 경험 양상이 2019년에 발발한 코로나 팬데믹 전후로 어떠한 변화가 있었는지도 분석하였다. 요약하면, 본 연구를 통해 우리나라 직장인들의 생생한 직원 경험을 자동적으로 식별하고 분류(유형화)하여, 몰입이나 이직과 같은 주요한 조직문화 현상의 행태와 유발 원인 등을 유추해 볼 수 있는 근거를 제공하고자 한다.

Ⅱ. 이론적 배경

본 연구에서는 방대한 직장인들의 기업 리뷰를 자동으로 분류하는 기계학습 모델을 구축하되, 학습에 부분 문법이 적용된 데이터를 사용하고자 한다. 본격적인 분석에 앞서 우선 부분 문법의 방법론적 의미와 원리를 소개하고 그 필요성에 대해

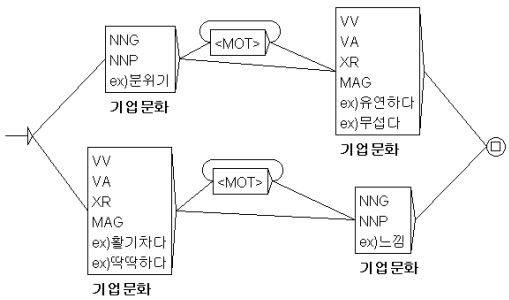
살펴본다. 다음으로, 현행 기계학습에 활용되는 대표적인 분류 방법론 중 본 연구에서 채택하고 있는 다중 레이블 분류에 대해 소개한다.

2.1 부분 문법(Local Grammar)

부분 문법은 프랑스의 전산언어학자인 모리스 그로스(Maurice Gross)가 제안한 유한 문법(finite grammar) 모델로 일종의 어휘적 문법(lexicon grammar)이라고 할 수 있다(Gross, 1997). 전혀 다른 형태이지만 같은 의미를 갖는 표현들이나 중의적인 어휘들 중에서 연구자가 찾는 의미의 표현들만을 정확히 추출할 수 있는 장점이 있다. 예를 들어, '미국 대통령'은 'The president of America', 'The president of United States', 'The president of USA'와 같이 다양한 표현으로 서술될 수 있다. 이 경우 부분 문법을 사용하지 않고, 단순히 'The president of America'라는 표현만 사용할 경우, 사실상 동일한 의미를 갖는 나머지 유사 표현들은 추출 단계에서 탈락하게 된다. 그러나 부분 문법을 활용하면 특정 표현의 패턴 확인이 가능하며 문맥추출(concordance)을 통해 실제 텍스트에 발현된 구체적이고 사실적인 문장들을 보다 정확하게 확인할 수 있다.

부분 문법은 부분 문법 그래프(LGG, local grammar graph)를 이용해 분석할 수 있는데, LGG는 왼쪽에서 오른쪽으로 진행되는 방향성 그래프(directed acyclic graph) 형태를 띤다. LGG 각각은 하나의 유한 오토마타(finite automata)를 표현하며 <그림 1>과 같이 하나의 시작 상태에서 하나의 최종 상태로 이어진다. 언어 사전에서 확인이 가능한 유의어, 유사어가 개별 단어 수준이라면, 부분 문법은 구문 차원으로 생성되어 특정한 분야의 의미를 가지는 구문 추출이 가능할 뿐만 아니라, 특정 분야의 코퍼스 내에서 문장 수준으로도 확장이 가능하다(백혜연 등, 2021).

<그림 1>은 기업문화를 표현하는 두 가지 방식의 예를 보여주고 있다. 상단 노드는 우리 기업은 '분위기(NNG)가 유연하다(VA)'와 같은 LGG를 표현한 이미지이고, 하단 노드는 우리 기업은 '활기찬(VA) 느낌(NNG)이다'와 같은 LGG를 표현한 이미지이다(약어: 명사는 일반명사(NNG), 고유명사(NNP)로, 서술어는 동사(VV), 형용사(VA), 어근(XR), 부사(MAG)로 표현되어 있음).



〈그림 1〉 '기업문화' 부분 문법 그래프(LGG)

부분 문법 방법론은 다양한 연구 분야에서 활용되어 왔다. 남지순(2008)은 동일한 의미지만 다른 형태를 갖는 서술어들을 검색어로 인식하도록 LGG를 구축하여 보다 정교한 정보검색 결과를 얻었다. 김명관, 이영우(2009)는 주식정보 등락을 표현하는 LGG를, 백혜연, 박용석(2019)은 직원 불만의 발화 요인을 분석하기 위해 LGG를 구축하여 활용하였다.

한편 최성용, 남지순(2018)은 형태소 분석기의 성능을 비교하여 미분석어까지 포함하는 보다 정교한 텍스트 분석을 가능하게 하였다. 백혜연 등(2021)은 언론 기사 텍스트를 LGG로 분석하여 프로야구 감독의 리더십 스타일을 유형화하였으며, 팀 성과와의 관계를 분석하였다.

김학준(2022)은 토픽 모델링 결과를 제시하면서 오직 명사만을 사용하면 해당 어휘들만으로는 정확히 전체 의미가 어떤 것인지 파악이 어렵다는 점을 보고하고 있다. 관련하여 대부분 명사 위주의 분석이 주를 이루는 환경에서 변수에 다양한 품사를 사용할 경우 모델 성능이 향상된다는 결과를 제시한 연구도 존재한다(정세민 등, 2021).

상기 연구들은 본 연구에서 주장하는 부분 문법 처리 프로세스의 유용성을 간접적으로 뒷받침한다고 볼 수 있다. 본 연구에서는 최초로 부분 문법 처리 프로세스를 기계학습에 직접 적용하였다. 구체적으로, 텍스트 분류를 위한 기계학습에 사용될 데이터를 2가지 데이터셋으로 구분하였다. 첫 번째 데이터셋은 일종의 실험용 데이터로서, 명사류(일반명사, 고유명사)와 서술어(동사, 형용사, 어근, 부사)의 조합이 사용된 LGG를 통해 구축한 데이터셋이고, 두 번째 데이터셋은 일종의 대조용 데이터로서, 선행 연구들에서 가장 빈번하게 사용하는 명사류만으로 구축된 데이터셋이다. 본 연구는 이 두 데이터셋을 각각 사용한 자동 분류 결과를 비교하여 부분 문법의 우수성을 확인하고자 한다.

2.2 다중 레이블 분류

데이터 분류는 기계학습 중 대표적인 지도학습(supervised learning) 방법 중 하나다. 이중 다중클래스 분류(multi-class classification)는 각각의 데이터를 여러 레이블 중의 하나로 분류하는 것이고, 다중 레이블 분류(multi-label classification)는 1개의 데이터가 1개 이상의 레이블을 한 번에 가지게 되는 경우의 분류이다. 예를 들면, 뉴스 기사를 정치, 사회, 문화 등의 카테고리 중 하나로 분류하는 것이 다중 클래스 분류이고, 영화를 스릴러, 액션, 로맨스, 코미디 등의 장르 중 하나 이상의 장르(예: 로맨스+코미디, 코미디+액션, 스릴러)로 분류하는 것이 다중 레이블 분류라고 할 수 있다.

데이터 자동 분류 시 다중 클래스 분류로 해결되는 경우도 있으나, 실생활에서 마주하게 되는 많은 경험 사례들은 다중 레이블 분류로 해결해야 하는 경우가 많다. 특히 텍스트로 표현된 경우, 문장 단위로 분석을 하더라도 한 개의 문장에 여러 의미 요소가 포함되므로 한가지의 카테고리로 분류하는 것은 결코 쉬운 일이 아니다. 아무리 짧은 단문이라도 하나의 문장에 여러 의미가 포함되어 있을 수 있다. 다음 예시들은 본 연구의 기계학습 데이터에 포함된 실제 기업 리뷰 텍스트 사례들이다. 가령, 아래 문장들은 사내 분위기나 오래된 조직문화(조직문화로 분류)에 대한 설명과 연봉(보상으로 분류)에 대한 설명이 혼합되어 있어서, 해당 문장이 조직문화 카테고리나 보상 카테고리 중 하나로만 분류된다면 데이터 식별의 정확도가 떨어지게 된다.

"사내 분위기가 좋으며 워라밸이 잘 지켜지는 편이지만, 연봉은 높지 않은 편"

"대기업의 높은 연봉. 오래된 조직문화에 따른 답답함"

위의 실제 리뷰 사례만 보더라도 한 개의 문장을 한가지의 의미로 레이블링하여 한 개의 카테고리로만 분류하는 것이 얼마나 부정확한 식별인지 짐작할 수 있다. 심지어 이진 분류에 해당하는 감성분석을 시도하려고 해도 위의 문장들은 모두 긍정/부정 의미를 각각 다른 분야별로 가지고 있어서 '중립'으로 분류하기도 어렵다.

다중 레이블 분류의 방법에는 크게 문제변환 기법(Problem Transformation Method, 이하 PT)과 알고리즘 적용 기법(Algorithm Adaptation Method) 2가지가 있는데, 각 기법들의 세부적인 방법론 차이의 결과를 비교한 연구들이 다수 존재한다(Read, 2008; Tsoumakas and Katakis, 2007; Tsoumakas *et al.*, 2009). 또한 여러 가지 문제변환 기법을 적용하여 특징을 추출하여 결과를 비교한 연구(Spolaôr *et al.*, 2013)와 기존 임베딩 기법들과 비교하여 다중 레이블 분류를 위한 새로운 임베딩 기법을 제안한 연구(박선호, 2013) 등이 존재한다.

김무성, 김남규(2021)의 연구에서는 다중 레이블 분류의 정확도를 높이기 위해 스킵 연결 오토인코더 기반 레이블 임베딩 방법론을 구축하였다. 다중 레이블 분류는 카테고리의 개수가 증가할수록 예측의 난이도가 상승하는 문제가 발생하는데, 이를 해결하기 위한 레이블 임베딩 기법(다수의 레이블을 압축한 후 압축된 레이블을 예측하고,

예측된 압축 레이블을 원래 레이블로 복원하는 방식)을 스킵 연결이라는 심층 신경망 분야의 아이디어를 적용하여 고도화하기도 하였다.

다중 레이블 분류에 관한 선행 연구들은 분류 성능을 개선하기 위한 다양한 방법론적 고도화를 시도하였는데, 본 연구와 같이 훈련용 데이터의 품질 자체를 고도화하는 시도는 없었다는 점에서 본 연구의 차별화된 공헌을 확인할 수 있다.

Ⅲ. 연구 방법

3.1 텍스트 자동 수집

잡플래닛(www.jobplanet.co.kr)에 등록된 3년(2018년 1월 ~ 2020년 12월) 간의 기업별 리뷰 텍스트를 파이썬 크롤링을 통해 자동수집하였다. 수집 대상 기업은 CEO Score에서 선정한 100대 기업 중 3년 연속 100대 기업으로 선정된 89개 기업이다. 3년 동안 89개 기업을 대상으로 등록된 리뷰는 총 35,731건이다. 각 리뷰는 텍스트 형태로 이루어진 요약(summary), 장점, 단점과 수치 형태의 별점(총점, 항목별 개별 별점) 및 '이 기업을 추천합니다', '이 리뷰가 도움이 됩니다' 등의 항목으로 이루어져 있다. 본 연구에서는 직장인들의 일터 경험을 가장 풍성하게 전반적으로 기술하고 있는 요약(summary) 텍스트만을 데이터 생성에 사용하였다.

3.2 한국어 벡터화

텍스트를 기계학습에 적용하기 위해서는 텍스트 데이터가 숫자 형태로 변형되어야 하는데, 벡터화(vectorization)가 그중 한 가지 방법이다. 다양한 텍스트 벡터화 방법이 있으나 한국어에서는 영어 위주의 벡터화 패키지를 그대로 사용하기 어렵다. 여러 가지 한국어 텍스트 벡터화 방법 중 본 연구에서는 TF-IDF를 기계학습에 사용하였다. TF-IDF(Term Frequency - Inverse Document Frequency, 단어 빈도-역문서 빈도)는 특정 문서에 출현한 어휘 빈도와 해당 어휘가 출현한 문서 수의 역수를 곱한 것(출현 어휘 빈도를 어휘가 나타난 문서의 수로 나눈 것)이다. 어휘가 출현한 문서가 많을수록 이 값은 작아진다(박상언 등, 2022). TF-IDF의 기본적인 전제는, 어떤 문서에만 고유하게 출현한 어휘가 해당 문서의 특성을 잘 나타내주는 것으로 보며, 모든 문서에 공통적으로 나타나는 어휘는 의미를 구별하는데 중요하지 않다고 간주한다. 이를 실현하기 위해 우선 사이킷 런(www.scikit-learn.org/)의 CountVectorizer를 통해 정해진 숫자와 특성을 가진 어휘들을 벡터화시켜서 TDM(Term document matrix)을 생성한 후 이 결과를 TF-IDF로 변환하여 기계학습에 사용하게 된다. 그런데 벡터화 과정에서 어떠한 특성의 어휘들을 얼마나 사용할지를 정하는 것이 학습 모델의 결과와 성능을 좌우할 수도 있다. 대부분의 연구가 영어 데이터를 활용하거나 이미 만들어진 한국어 영화 리뷰 데이터를 활용하여 감성분석 정도에 사용하는 상황에서 품사가 처리된 한국어를 사용하여 벡터화 결과를 다양하게 적용한 연구는 매우 드물다. 다만, CountVectorizer 생성 시 한국어 품사별 성능 차이를 제안하고, 오직 명사만을 사용하는 것이 아니라 분석 대상이나 목적에 따라 한국어 품사를 달리 설정하여 기계학습에 사용할 필요가 있다고 언급한 연구가 있다(박상언 등, 2022). 본 연구에서는 기존에 수행되어 왔던 '명사' 위주의 결과와 부분 문법이 활용된 결과(구체적으로 품사 전체를 사용했을 때, 일부 품사 묶음을 사용했을 때, 단일 품사를 사용했을 때의 결과들)를 비교한다.

3.3 한국어 품사 분류 및 훈련용 데이터 생성

텍스트를 분석하기 위해 한국어 형태소 분석 지원 패키지인 코엔엘파이 'Konlpy'(konlpy.org)의 Komoran을 사용하여 품사 분석을 진행하였다. Konlpy에서 제공되는 다양한 형태소 분석기들 중 Okt(Twitter)의 경우 명사 위주로는 최신 어휘들이

많이 등록되어 있다고 하나, 서술어의 원형을 제공하지 않은 채 모든 활용형을 각각 품사로 태깅한다는 문제점이 있다. Mecab의 경우에는 어휘들을 과도하게 작은 단위의 형태소로 분석하여 원래의 의미를 파악하기 어려운 사례(ex. '비추천'을 비+추천으로 분석)가 다수 확인되었다.

본 연구에서는 형태소 분석 결과 확인 시 복합어를 과도하게 분리하지 않고 서술어의 원형 확인이 쉽다는 장점을 높이 사서 Komoran을 선택하였다. 다만, 추후 데이터의 크기가 매우 증가할 경우, 수행시간 등의 문제로 다른 형태소 분석기가 더 나은 선택이 될 수도 있다.

기계학습에 사용된 훈련 데이터는 잡플래닛의 리뷰 텍스트이다. 잡플래닛 리뷰의 경우 기타 온라인 커뮤니티의 게시물이나 댓글들보다 띄어쓰기가 상대적으로 잘 지켜지는 편이고 최종 리뷰 등록 전에 잡플래닛 측의 자체 검토가 이루어진 후 승인이 되므로 욕설이나 불법적인 표현은 포함되지 않는다고 봐도 무방하다. 특정 품사만을 활용할 예정이기에 일괄적인 기호 제거, 불용어 선정, 1음절 어휘 삭제 등의 전처리 작업은 따로 진행하지 않았다. 띄어쓰기 처리도 오히려 과도한 띄어쓰기로 원래 어휘가 대체될 가능성이 있어 사용하지 않았다. 오직 오류를 발생시킨 '줄바꿈' 기호만을 일괄 삭제하였다.

35,731개 리뷰를 품사 분석한 결과, 일반명사(NNG) 5,241가지(type), 고유명사(NNP) 4,404가지, 동사(VV) 1,014가지, 형용사(VA) 254가지, 어근(XR) 368가지, 부사(MAG) 558가지로 확인되었다. 이 중 명사류(일반명사와 고유명사를 한 번에 조회한 후 중복되는 것은 제거함)는 8,885가지다. 명사류가 다른 품사에 비해 등장하는 빈도가 많기 때문에, 명사류는 빈도 25 이상, 동사, 형용사, 어근, 부사류는 빈도 5 이상의 어휘들을 선정하여 자동 분류하였다. 모든 품사들의 빈도를 동일하게 설정하여 확인하려고 했으나, 명사류의 종류(type)가 불균형적으로 많아서 불가피하게 다른 품사들과 다른 빈도로 설정하게 되었다.

3.4 분류 카테고리

본 연구에서는 개별 리뷰 텍스트들을 5개의 카테고리(분야)로 분류하였다. 본 연구에서 사용한 5개의 카테고리는 인사조직 분야의 주요 연구 주제 중 하나인 직원 몰입(work engagement)에 영향을 끼치는 5개의 요인들이다. 직원 몰입은 직원들이 업무를 수행함에 있어서 자신의 신체적, 인지적, 정서적 에너지를 쏟아내는 상태를 말한다(Kahn, 1990). 업무에 몰입하는 직원은 높은 성과를 달성하고 조직에 도움이 되는 다양한 행동을 하기 때문에(Demerouti and Cropanzano, 2010), 학계와 실무계에서는 직원 몰입을 유발하는 요인에 대해 관심을 가져왔다. 대표적으로 글로벌 리서치 기업인 갤럽(Gallup)은 100여 개국 이상의 기업에 종사하는 직원 700만 명 이상의 직원 몰입에 관한 조사를 실시해왔고, 해당 조사를 매년 수행하고 있다(Bakker and Leiter, 2010).

최근 WSA(Workforce Science Associations)는 직원 몰입에 관한 대규모 분석을 실시했다. 미국 인사관리협회(SHRM)가 공식 주관하는 글로벌 컨퍼런스에 초청된 WSA는 지난 40여 년간 전 세계 약 2,300만 명의 직원을 대상으로 직원 몰입과 관련이 있는 핵심 요인을 발표하였는데, 개인의 성장, 회사의 미래, 조직문화(조직에 대한 신뢰와 소통을 통합하여), 리더에 대한 신뢰, 인정과 보상이 그것이다(Erickson and Erinckson, 2021).

본 연구에서는 상기 5개 카테고리를 직원들의 일터 경험을 식별할 수 있는 중요한 식별 기준으로 사용하였다. 다양한 일터 경험들은 궁극적으로 직원들이 업무에 몰입하는 데 도움이 되거나 방해가 되는 경험들일 가능성이 높기 때문에, 직원 몰입 요인을 구분한 5개의 카테고리가 직원 경험을 인식할 수 있는 중요한 기준이라고 판단하였다.

방대한 일터 경험들을 5개의 카테고리 분야에 걸쳐 각각 고유하게 인식하는 것이 최초의 분류 목적이었으나, 여러 분야에서 인식 가능한 어휘들도 존재하여 복수 분야로 라벨링하는 방식을 택했다.

〈표 1〉 품사별로 추출된 어휘 예시

카테고리	NNG+NNP	VV	VA	XR	MAG
개인의 성장	경험, 성장, 자부심	배우다, 버티다, 해보다	어렵다, 쉽다, 벅차다	철저, 뿌듯	스스로
회사의 미래	비전, 타이틀, 매출	망하다, 살아나다	없다	유명	현재, 매년
조직문화	환경, 느낌, 수평, 수직	느끼다, 참다, 돕다	활기차다, 낡다, 무섭다	다양, 딱딱, 유연	모두, 점점, 여전히
리더신뢰	관리, 평가, 차별, 존중	따르다, 시키다, 잡다, 듣다	없다	꼼꼼	철저히
인정 및 보상	연봉, 복지, 대우, 보상	받다, 벌다, 모으다	많다, 적다, 아쉽다	미흡, 넉넉, 취약	확실히, 충분히

<표 1>은 5개의 품사(명사류, 동사, 형용사, 어근, 부사)별로 위에서 언급한 5개 분야에 맞게 라벨링된 결과의 일부를 예시하고 있다.

5개 카테고리별로 추출된 품사별 어휘들은 단순한 형태의 부분 문법 구축(Gross, 1997)에 사용되었다. 전술하였듯이, 부분 문법을 활용하지 않고 단일 품사(주로 명사류)만을 사용하는 경우 텍스트 식별에 심각한 오류가 발생할 수 있기 때문이다. 예를 들어, “연봉이 낮다”는 문장은 “월급이 짜다”는 문장과 겹치는 단어가 하나도 없음에도 의미가 동일하며, “업무에 대한 보상이 적어서 아쉽다”라는 문장과도 매우 유사한 의미를 갖는다. 명사인 “연봉”만 사용할 경우 위의 세 문장 중 첫 번째 문장만 인식하는 오류가 발생한다. 하지만 명사와 서술어의 조합을 고려하는 부분 문법 프로세스를 통하면 이러한 유사 문장들까지 ‘보상’에 대해 언급한 문장임을 정확하게 인식할 수 있게 되는 것이다.

그런데 명사의 품사 분류(NNG와 NNP의 구분)가 쉽지 않기 때문에, 우선 NNG, NNP 태그와 무관하게 라벨링 된 명사들에서 중복을 제거하여 품사 태그 없이 분류별 의미를 가지는 명사들만 포함된 리뷰들을 35,731건에서 5개 분야로 1차로 추출하였다. 그 다음 각 분류별로 의미를 가지는 서술어(동사, 형용사, 어간)나 부사가 포함된 리뷰들을 다시 추출하여 <표 2>와 같은 결과를 얻게 되었다. 자동 수집된 35,731건의 리뷰 중 부분 문법이 적용된 총 22,397건의 리뷰가 5개 분야에 대한 다중 레이블 분류를 위한 학습과 검증에 사용될 학습 데이터셋으로 추출되었다.

명사류와 서술어가 조합된 부분 문법의 우수성을 비교하기 위해, 명사들만이 포함된 데이터셋(엄밀한 의미의 부분 문법이 적용되지 않은 데이터, 개인의 성장 22,521건, 회사의 미래 13,805건, 조직문화 16,888건, 리더 신뢰 16,937건, 인정 및 보상 16,347건, 총 86,490건)도 구축하였다.

<표 3>은 부분 문법을 사용하여 구축된 데이터셋에서 추출한 상대적으로 길이가 짧은 2개의 원문을 분야별로 제시하고 있다. 실제 경험을 나타내는 표현들이다 보니 긍정적인 표현과 부정적인 표현이 혼재되어 있음을 확인할 수 있다. 또한, 의도하지 않은 듯한 오타, 일부 띄어쓰기가 무시되는 현상이 존재하기도 한다.

〈표 2〉 품사를 통해 구축된 데이터셋 현황

구분	2018년	2019년	2020년	합계
개인의 성장	1,087	1,473	1,506	4,066
회사의 미래	664	812	922	2,398
조직문화	1,640	2,192	2,206	6,038
리더신뢰	702	894	889	2,485
인정 및 보상	1,998	2,756	2,656	7,410
합계	6,091	8,127	8,179	22,397

〈표 3〉 분야별 리뷰 원문 예시

구 분	리뷰 원문
개인의 성장	1. 오래 다니긴 나쁘진 않지만 자기 개발을 하기 어려운 회사 2. 커리어를 쌓기도 좋고 기술을 배우기도 좋습니다.
회사의 미래	1. 과거에 좋은 회사, 조선업이 살아나면서 부활하기를 희망 2. 국내최고의 회사이나 패스트팔로워 전략의 한계에 부딪혔다
조직문화	1. 안정적인 직장이며 근무 문화가 빠르게 바뀌고 있는 중임. 2. 겉과 속이 다른 회사. 기업문화 바꾸지 않으면 힘든 회사.
리더신뢰	1. 글로벌 프로세스를 경험할 수 있다는 점이 메리트. 단점은 책임질 리더가 없다는 것. 2. 과다한 팀장. 과도한 인원 이동. 마케팅에 대한 이해없이 쪼기만 하는 사람들로 가득함.
인정 및 보상	1. 워라밸만 좋음. 직원들 인건비 대비 일이 너무 많은 회사 2. 받는 연봉의 비해 대우가 매우 열악하고 환경 또한 열악하다

<표 2>에 나타나듯이, 직원들의 기업 리뷰 빈도를 살펴보면, 인정 및 보상에 대한 리뷰가 3개년도 모두 가장 높은 빈도를 보였고, 다음으로 조직문화와 개인의 성장에 대한 리뷰가 많았다. 회사의 미래와 리더 신뢰에 대한 리뷰는 비교적 적은 편이었다. 즉, 직원들은 자신과 직접적으로 관련된 경험(인정 및 보상, 개인의 성장, 조직문화)을 회사와 관련된 간접적인 경험(회사의 미래와 리더 신뢰)보다 높은 관심도를 가지고 리뷰하고 있음을 알 수 있다.

Ⅳ. 연구 결과

4.1 다중 레이블 분류 결과

본 연구에서는 문제변환(PT) 방법론 중 다수의 레이블에 대한 분류를 각각의 이진 분류로 변환하는 BR(binary relevance) 방식을 사용하였다. BR 방식은 상대적으로 파이썬 코드 구현에 용이할 뿐만 아니라, 여러 레이블로 분류될 수 있는 데이터들을 대상으로 '각각의 레이블에 해당 여부'만을 우선 판단한다는 점에서 복잡한 문제가 이진 분류로 단순하게 치환되므로 상대적으로 명쾌한 답을 얻게 되는 장점이 있다. Keras의 sequential model에서 activation에 sigmoid를 적용하여 5개 카테고리에 대한 이진 분류를 각각 수행한 후 최종 결과가 정확도와 손실률로 측정된다(김기현, 2019). CountVectorizer 생성 후 결과를 TF-IDF로 변환하여 다중 레이블 분류를 위한 학습에 사용했다.

Keras에서 sigmoid를 적용하면서 일괄적으로 적용한 수치는 아래와 같다. 가능한 동일한 조건하에서 결과를 확인하기 위해 훈련 데이터 75%, 테스트(test) 데이터 25%, TF-IDF 최대 특성(max feature) 2,000, batch_size 400, epochs 20, validation_split 0.2, optimizer 'rmsprop'로 일괄 설정하였다. 5개의 sigmoid를 한 번에 적용하면서 layer 1, dense 1이 적용되었다. 초반에 epoch를 100으로 설정했다가 과적합(overfitting)이 다소 이르게 발견되어 일괄 20으로 축소하여 결과를 비교하였다.

부분 문법이 적용된 분류와 적용되지 않은 분류의 성능 비교를 위한 분석을 실시하였다. 체계적인 비교를 위해 전체 품사를 적용한 경우, 특정 품사를 6가지, 5가지, 3가지, 2가지 적용한 경우, 명사만 적용한 경우, 동사만 적용한 경우의 각기 다른 결과 8개를 도출하였다.

부분 문법을 적용하지 않은(즉, 서술어 없이 명사류만을 사용함) 데이터로 학습 데이터를 구축하고 다중 레이블 분류한 결과는 <표 4>와 같다. 한편, 부분 문법을 적용한 데이터로 학습 데이터를 구축하고 다중 레이블 분류한 결과는 <표 5>와 같다.

분석 결과, 부분 문법이 적용되지 않은 다중 레이블 분류보다 부분 문법이 적용된 다중 레이블 분류가 전반적으로 모든 구분 기준에서 개선된 결과를 보여주었다.

예를 들어, 명사류만 사용한 NNG의 경우, 부분 문법이 적용된 다중 레이블 분류의 정확도는 0.5707이고, 부분 문법이 적용되지 않은 다중 레이블 분류의 정확도는 0.4376이다. 명사+서술어 혼합 NNG+NNP+VV+VA+XR+MAG의 경우, 부분 문법이 적용된 다중 레이블 분류의 정확도는 0.4370이고, 부분 문법이 적용되지 않은 다중 레이블 분류의 정확도는 0.4137이다. 따라서, 부분 문법이 적용된 학습 데이터가 기계학습 모델의 성능을 개선하고 있음을 알 수 있다.

기존 명사 위주의 연구들이 <표 4>의 NNG 혹은 NNG+NNP의 결과를 나타냈다면, 부분 문법을 통해 최소 <표 5>의 NNG 혹은 NNG+NNP의 결과를 얻는다. 학습 데이터를 변경한 것만으로 10% 이상의 정확도가 향상되는 것이다.

〈표 4〉 다중 레이블 분류 결과(명사류만으로 구축된 86,498건)-Sigmoid

구 분	Accuracy	Loss
품사전체(tagger.pos)	0.4112	2.2190
NNG+NNP+VV+VA+XR+MAG	0.4137	2.1979
NNG+NNP+VV+VA+XR	0.4058	2.1908
NNG+NNP+VV	0.4149	2.2053
NNG+NNP(tagger.nouns)	0.4163	2.2094
NNG+VV	0.4405	2.3396
NNG	0.4376	2.3358
VV	0.4844	2.5004

〈표 5〉 다중 레이블 분류 결과(부분 문법으로 구축된 22,397건) -Sigmoid

구 분	Accuracy	Loss
품사전체(tagger.pos)	0.4265	1.6965
NNG+NNP+VV+VA+XR+MAG	0.4370	1.7234
NNG+NNP+VV+VA+XR	0.4374	1.7862
NNG+NNP+VV	0.4979	1.9773
NNG+NNP(tagger.nouns)	0.5408	2.1986
NNG+VV	0.5001	1.9983
NNG	0.5707	2.2609
VV	0.5342	1.9772

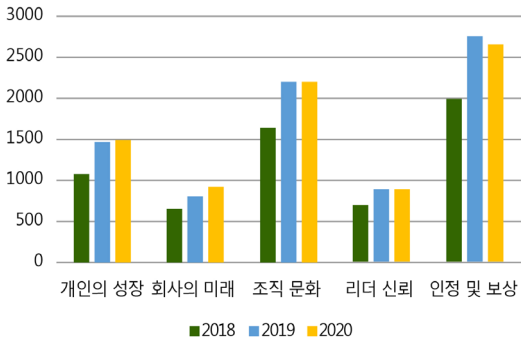
다만, 해석상의 주의가 필요한 분석 결과도 존재한다. 부분 문법이 적용되건 적용되지 않건, 오직 한 개의 품사만을 사용하여 분류한 경우가(예: NNG, VV) 여러 개의 품사를 혼합하여 분류한 경우보다(예: NNG+NNP+VV+VA+XR+MAG) 높은 정확도를 보인다는 점이다. 그 이유는 한 개의 어휘만 일치하면 분류되는 조건보다 여러 개의 어휘가 모두 정확하게 일치해야 분류되는 조건이 정확도를 유지하면서 분류해야 하는 까다로움을 가지고 있기 때문이다. 그러나, 한편으로 정확도는 더 높을 수 있으나 손실(loss, Binary Crossentropy)역시 더 높았기 때문에 한 개의 품사만을 사용하여 분류하는 것이 더 바람직하다고 주장하기에는 무리가 있다.

4.2 추가 분석: COVID-19 이전 이후 비교

앞선 분석에서는 지난 3년간의 직장인들의 리뷰 데이터를 연도 구분없이 사용하였다. 그러나 직원들의 일터 경험은 사회적, 경제적 상황 등의 외부 변화에 따라 패턴 상의 변화가 발생할 수 있다. 따라서, 2019년 말에 발생한 코로나 바이러스 팬데믹(COVID-19)이라는 초대형 이벤트는 직장인들의 일터 경험에 어떤 식으로든 영향을 미쳤을 것이므로, 팬데믹 이전과 이후의 직원 경험 트렌드가 어떻게 변화했는지를 추가적으로 분석해 보았다.

코로나 이전과 이후를 통틀어 빈번하게 리뷰된 경험 카테고리는 인정 및 보상 > 조직 문화 > 개인의 성장> 리더 신뢰 > 회사의 미래 순서였다. 이는 지난 3개년도에 걸쳐 직원들이 가장 많이 경험했거나 관심이 집중된 영역이 '인정 및 보상'임을 알 수 있다. 그러나 변화 추이를 자세히 들여다보면, '인정 및 보상'에 대한 언급은 2019년에 비해 2020년에 다소 감소하는 추세를 보이는 반면, '회사의 미래'에 대한 언급은 증가하고 있음을 알 수 있다. 코로나로 인해 회사의 존망이 위협받는 상황에서 개인의 인정과 보상에 대한 관심 보다는

회사의 미래에 대한 불확실성과 염려가 더 컸기 때문으로 추측된다.



〈그림 2〉 COVID전후 5개 분야 리뷰 빈도 비교

Ⅴ. 결 론

최근 많은 분야에서 기계학습에 대한 연구가 활발히 진행되고 있다. 상당수의 연구들이 학습 모델의 성능을 개선하는 최신 방법론을 소개하고 있는데, 본 연구에서는 방법론의 개발 못지않게 기계학습에 투입되는 훈련용 데이터의 '품질'을 개선하는 것 역시 중요하다는 점을 강조하였다. 국내 최초로 코퍼스(corpus) 분석에서 자주 사용되는 부분 문법 처리 프로세스를 통해 훈련 데이터의 품질을 높일 수 있음을 제안하였고, 모델 성능 비교를 통해 유효한 접근법임을 입증하였다.

구체적으로, 기업 리뷰 플랫폼상에 존재하는 다양한 직장인들의 리뷰 텍스트를 활용하여 직원 몰입 분류에 관한 기계학습 모델을 구축하고, 부분문법 처리 프로세스를 가미하여 개선된 성능의 모델을 구축하였다. 부분 문법을 활용하지 않고 기존의 학습 모델과 같이 단일 품사(주로 명사류)만으로 훈련을 시키는 경우 텍스트 식별의 정확성에 심각한 오류가 발생할 수 있었다는 점에서 본 연구의 기여는 의미심장하다.

이러한 방법론적 공헌뿐 아니라 기계학습을 통해 도출된 분석 결과는 중요한 실무적 시사점을 제공하고 있다.

첫째, 본 연구에서는 직원들의 기업 리뷰의 '빈도'를 분석하여 의미 있는 시사점을 도출하였다. 분석 결과, 국내 직장인들은 인정과 보상 경험에 대해 사람들과 가장 많이 공유한다는 사실이다. 직원들이 가장 관심을 가지고 중요하게 여기는 주제이기 때문일 수도 있지만, 회사 내부에서는 민감한 주제인 인정과 보상에 대해 공론화하기가 어려워 익명이 보장된 외부 플랫폼에 여과 없이 공유하려는 현상의 발로일 수도 있다. 그 이유에 대해서는 보다 정밀한 분석이 필요하지만, 현상적으로는 직원들은 인정과 보상에 대해 가장 많은 경험을 세상과 공유하고 있다는 사실이다.

경영진들은 직원들이 게시한 기업 리뷰를 면밀히 분석하여 그들이 인정과 보상에 관해 어떤 pain point를 경험하는지 검토하고 제도를 개선할 필요가 있다. 구직자나 재직자들은 잡플래닛과 같은 채용플랫폼의 기업 리뷰를 입사 지원이나 이직과 같은 중요한 의사결정의 참고자료로 활용하고 있다는 점에서 pain point 개선이 더욱 중요하다고 할 수 있다.

둘째, 전반적으로 직원들은 회사와 관련된 2차적 경험(회사의 미래와 리더 신뢰)보다 자신과 직접적으로 관련된 1차적 경험(인정 및 보상, 개인의 성장, 조직문화)에 대해 더 많은 관심을 가지고 소통하고 있음을 파악하였다. 그러나 코로나 팬데믹 발발 전후의 변화를 동태적으로 살펴보면 흥미로운 패턴이 발견되었는데, 인정 및 보상에 대한 언급은 코로나 사태 전보다 후에 다소 감소하는 추세를 보이는 반면, 회사의 미래에 대한 언급은 증가하고 있었다. 코로나로 인해 회사의 존망이 위협받는 상황에서 개인의 인정과 보상보다는 회사의 미래에 대한 불확실성과 염려가 더 컸기 때문으로 추측된다. 따라서 기업은 직원들이 관심 갖는 일터 경험의 영역을 보다 동태적으로 이해할 필요가 있고, 조직관리의 효과성을 높이기 위해 시의성 높은 관리 방향을 설정할 필요가 있음을 시사한다. 예를 들어, 코로나 시기와 같은 불확실성이 증가하는 시기에는 최고경영진이 회사의 미

래에 대한 소통을 확대하고, 회사의 미래에 대한 확신을 구성원들에게 심어줄 필요가 있다. 즉, 본 연구의 결과는 코로나 사태와 같이 기업에 영향을 주는 중요한 이벤트가 발생했을 때 이벤트 전후로 직원들의 일터 경험이 어떻게 변화하는지를 기계학습 모델을 통해 신속히 파악하고 효과적인 대응 근거로 활용할 수 있음을 보여준다.

셋째, 인력 및 조직관리 상의 중요한 시사점을 얻기 위해 본 연구에서는 외부 기업 리뷰 플랫폼의 데이터를 활용하여 직원들의 "진솔한" 경험세계를 파악했다는 점이다. 국내 기업의 구성원들은 상명하복의 위계적 조직문화 속에서 자신의 솔직한 일터 경험을 사내에 공유하기를 기피하는 경향이 강하다. 그러다 보니 직원들은 익명이 보장되는 외부 온라인 기업 리뷰 플랫폼에서 솔직한 일터 경험들을 공유하고 있다. 본 연구에서는 이점에 착안하여 플랫폼 상에 게시된 기업 리뷰가 조직 현장의 목소리를 보다 정확히 대변한다고 가정하고 기계학습의 훈련 자료로 활용하였다.

현재 많은 기업에서 직원들의 경험세계를 이해하기 위해 정량화된 설문이나 표준화된 진단 도구를 활용하고 있다. 그러나 이러한 도구들은 진단 문항이 사전에 정의되어 있어서 주어진 측정의 범위 내에서만 이해하게 되는 한계가 있다. 엄밀히 말하자면 직원들의 경험 세계를 있는 그대로 포착하기보다는 직원들의 주관적 '견해'를 수렴하는 것에 가깝다. 본 연구에서는 있는 그대로의 직원 경험을 분석하여 사안의 실제성(actuality)을 최대한 포착하였다.

넷째, 자동화 프로세스를 통해 인력과 조직관리상의 중요한 시사점을 제공하였다는 점이다. 기업 리뷰 데이터는 온라인의 여러 공간에서 매일같이 쏟아지는데, 이러한 방대한 양의 데이터를 자동으로 처리하여 중요한 인사이트를 쉽고 빠르게 도출함으로써 관리자들의 의사결정의 질을 높이고, 인사관리 업무의 생산성과 비용 효율성을 높일 수 있다.

본 기계학습 모델을 기반으로 자동화를 고도화하여 RPA(Robotic Process Automation) 프로그램이나 다양한 사무용 인공지능 프로그램(예: 챗봇) 등과 연계한다면 그러한 장점이 극대화될 수 있다. 가령, 인사담당자는 본 연구에서 개발한 직원 경험 자동 분류 모델을 활용하여, 자사 직원들의 불만 경험이 어떤 카테고리에 속하는지를 신속히 파악할 수 있을뿐더러, 업계 전반 혹은 경쟁 기업군의 직원 경험 상황도 손쉽게 파악할 수 있다. 즉, 직원들의 불만 경험 양태가 자사의 고유한 현상인지, 경쟁사도 유사한 현상을 겪고 있는지, 또는 산업계 전반의 추세인지 등을 매우 손쉽게 파악할 수 있다. 이를 통해, 큰 비용 없이 조직관리 의사결정의 질과 속도에 도움을 얻을 수 있다.

마지막으로, 본 연구를 통해 기업 리뷰 플랫폼 기업들의 사업 역시 고도화될 수 있다. 본 연구에서 사용한 '잡플래닛'은 가장 활발한 기업 리뷰가 공유되는 플랫폼 기업 중 하나로서, '알리'라는 서비스를 제공한다. '알리'는 기업의 리뷰에서 언급되는 위험 키워드(폭행, 성범죄, 비리 행위 등)를 감지해 대상 기업에게 인사이트를 제공하여 기업의 리스크 관리에 도움을 준다. 하지만 해당 서비스는 주로 명사 기반의 키워드를 활용하기 때문에 본 연구에서 소개한 기계학습 모델을 적용한다면, 리스크 센싱의 정확성을 높일 수 있고 이는 플랫폼 기업의 사업 고도화에 도움을 줄 수 있다.

이러한 기여점에도 불구하고 본 연구는 몇 가지 한계점을 가지고 있다. 본 연구의 가장 큰 한계점은 부분 문법을 사용하여 생성된 데이터들을 일일이 검수하지 못한 채로 학습이 진행되었다는 점이다. 이러한 한계를 개선하는 가장 확실한 방법은 생성된 데이터들의 정답 여부를 수동으로 재확인하는 것이다. 이러한 방식이 사용될 경우 다수의 데이터 라벨러들이 작업하여 데이터를 구축하는 기존의 방식과 다를 바가 없어 보이나, 최초의 라벨링 단계가 사라지고 검수만 하면 된다는 점, 레이블링 대상이 되지 않는 데이터들을 미리 분리하고, 필터링이 된 유의미한 데이터들을 빠르게 얻게 된다는 점에서 의의가 있다.

또 다른 한계점은 직원들의 일터 경험을 직원 몰입의 원인에 해당하는 5개의 카테고리로만 분류했다는 점이다. 직원들의 경험을 분류하는 기준은 매우 다양한데 '몰입'이라는 한 가지 기준으로만 분류 모델을 수립한 것은 단일 주제 연구의 한계임이 분명하다. 향후에는 직원 몰입 이외에 다른 분류 기준을 사용했을 때에도 부분 문법을 적용한 분류 모델의 성능이 여전히 우수한지 검증할 필요가 있다.

요약하면, 본 연구는 기업 리뷰 데이터를 기계학습에 활용하여 다양한 가치를 창출할 수 있음을 제시하였다. 본 연구가 기계학습 연구와 실무에 중요한 도전과 공헌을 제공하였기를 기대한다.

참 고 문 헌

[1] 김기현, *(김기현의) 자연어 처리 딥러닝 캠프: 파이토치 편*, 한빛미디어, 서울, 2019.

[2] 김명관, 이영우, "웹 문서 정보추출과 자연어 처리를 통한 온톨로지 자동구축에 관한 연구", *한국인터넷방송통신학회 논문지*, 제9권, 제3호, 2009, pp. 61-67.

[3] 김무성, 김남규, "다중 레이블 분류의 정확도 향상을 위한 스킵 연결 오토인코더 기반 레이블 임베딩 방법론", *지능정보연구*, 제27권, 제3호, 2021, pp. 175-197.

[4] 김학준, *보통 일베들의 시대 : '혐오의 자유'는 어디서 시작되는가*, 오월의봄, 파주, 2022.

[5] 남지순, "자연언어 검색 질의문 인식을 위한 유한 그래프 문법의 구축", *언어과학*, 제15권, 제1호, 2008, pp. 39-69.

[6] 박상언, 강주영, 정석찬, 파이썬 텍스트 마이닝 완벽 가이드 : 자연어 처리 기초부터 딥러닝 기반 BERT 모델까지, 위키북스, 파주, 2022.

[7] 박선호, *Embeddings for Multi-class and Multi-label Learning* (박사학위논문), 포항공과대학교 일반대학원, 2013.

[8] 백혜연, 박용석, "기업 리뷰 웹 사이트 텍스트 분석을 통한 직원 불만 표현 추출과 불만 원인 도출 및 해소 방안", *한국정보통신학회논문지*, 제23권 제4호, 2019, pp. 357-364.

[9] 백혜연, 장영균, 양동훈, "공유리더십의 정성적 측정 도구에 관한 연구: 국내 프로야구 감독의 언론 기사 코퍼스 분석을 중심으로", *리더십 연구*, 제12권, 제1호, 2021, pp. 135-162.

[10] 정세민, 이세영, 안유나, 김보경, "품사에 따른 영화 리뷰 감성분석 연구", *KIIT Conference 2021.11*, 2021, pp. 651-654.

[11] 최성용, 남지순, "소셜 미디어 텍스트의 의존명사 내포 비정규토큰의 부분문법그래프 패턴 사전 구축", *한국사전학*, 제32호, 2018, pp. 114-150.

[12] Bakker, A. B. and M. P. Leiter, "Where to go from here: Integration and future research on work engagement", in A. B. Bakker and M. P. Leiter (Eds.), *Work engagement: A handbook of essential theory and research*, Psychology Press, New York, 2010, pp. 181-196.

[13] BBC News, "US job openings hit record high, with more Americans quitting", 2021, Available at https://www.bbc.com/news/business-58170391.

[14] Delip, R., Brian, M. (박해선 역), *파이토치로 배우는 자연어 처리*, 한빛미디어, 서울, 2021.

[15] Demerouti, E. and R. Cropanzano, "From thought to action: Employee work engagement and job performance", in A. B. Bakker, and M. P. Leiter (Eds.), *Work Engagement: A Handbook of Essential Theory and Research*, Psychology Press, New York, 2010, pp. 147-163

[16] Ducci J., "The Great Resignation: Why Employees Are Quitting in Droves", Forbes, 2021, Available at https://www.forbes.com/sites/jackieducci/2021/06/23/the-great-resignation-why-employees-are-quitting-in-droves/?sh=22f8200a2eb9.

[17] EBS News, "<뉴스브릿지> 대퇴사 시대…회사 떠나는 MZ세대", 2022, Available athttps://ne

ws.ebs.co.kr/ebsnews/allView/60247678/N.

[18] Erickson, K. and B. Erickson, “The new voice of our employees: What you need to know from employees aroudn the world”, *Annual Virtual Conference of Society for Human Resource Management(SHRM)*, 2021, Concurrent Session

[19] Gross M., “The construction of local grammars”, in E. Roche and Y. Schabès(eds.) *Finite-State Language Processing*, MIT Press, 1997, pp. 329-354.

[20] Harter J., “Is Quiet Quitting Real?”, GALLUP, 2022, Available at https://www.gallup.com/workplace/398306/quiet-quitting-real.aspx.

[21] Jens, A., R. Sidharth, and W. Christian (심상진 역), *파이썬 라이브러리를 활용한 텍스트 분석: 텍스트에서 통찰을 이끌어내는 98가지 자연어 처리 전략*, 한빛미디어, 서울, 2022

[22] Kahn, W. A., “Psychological conditions of personal engagement and disengagement at work”, *The Academy of Management Journal*, Vol.33, No.4, 1990, pp. 692-724.

[23] Keras, “The sequential model”, 2020, Available at https://keras.io/guides/sequential_model/.

[24] Komoran, “품사표(PoS Table)”, 2019, Available at https://komorandocs.readthedocs.io/ko/latest/firststep/postypes.html.

[25] Read, J., “A pruned problem transformation method for multi-label classification”, *New Zealand Computer Science Research Student Conference*, 2008, pp. 143-150.

[26] Scikit learn, “1.12 Multiclass and multioutput algorithms”, 2023, Available at https://scikit-learn.org/stable/modules/multiclass.html.

[27] Sohom, G. and G. Dwight, (김창엽, 최민환 역), *예제로 배우는 자연어 처리 기초 : NLP 알고리즘, 텍스트 분류와 요약, 감성 분석*, 에이콘, 서울, 2020.

[28] Spolaôr, N., E. A. Cherman, M. C. Monard, and H. D. Lee, “A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach”, *Electronic Notes in Theoretical Computer Science*, Vol 292, 2013, pp. 135-151.

[29] Tsoumakas, G. and I. Katakis, “Multi-label classification: An overview”, *International Journal of Data Warehousing and Mining 3*, 2007, pp. 1-13.

[30] Tsoumakas, G., I. Katakis, and I. Vlahavas, “Mining multi-label data”, in Maimon, O., Rokach, L. (eds.), *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 2009, pp. 667-685.

Information Systems Review
Volume 25 Number 3
August 2023

Multi-Label Classification for Corporate Review Text: A Local Grammar Approach

HyeYeon Baek* · Young Kyun Chang**

Abstract

Unlike the previous works focusing on the state-of-the-art methodologies to improve the performance of machine learning models, this study improves the 'quality' of training data used in machine learning. We propose a method to enhance the quality of training data through the processing of 'local grammar,' frequently used in corpus analysis. We collected a vast amount of unstructured corporate review text data posted by employees working in the top 100 companies in Korea. After improving the data quality using the local grammar process, we confirmed that the classification model with local grammar outperformed the model without it in terms of classification performance. We defined five factors of work engagement as classification categories, and analyzed how the pattern of reviews changed before and after the COVID-19 pandemic. Through this study, we provide evidence that shows the value of the local grammar-based automatic identification and classification of employee experiences, and offer some clues for significant organizational cultural phenomena.

Keywords: ***Multi-label classification, Machine learning, Local grammar, Corporate Review, Employee Experience***

* Ph.D Student, Sogang Business School, Sogang University
** Corresponding Author, Professor, Sogang Business School, Sogang University

◐ 저 자 소 개 ◑

백 혜 연 (bequette@sogang.ac.kr)

한국외국어대학교에서 전산언어학 석사수료 후 세종사이버대학교대학원에서 정보보호학 석사학위를 취득하였으며 현재 서강대학교 경영학과 박사과정 재학 중이다. 주요 관심 분야는 텍스트를 활용한 기계학습, 정보 추출, 기업 윤리, 직원경험 분석이다.

장 영 균 (changy@sogang.ac.kr)

미국 위스콘신대학교 경영학부 교수를 역임하였고, 현재 서강대학교 경영대학 인사조직전략 계열 교수로 재직 중이다. Journal of Applied Psychology, Journal of Management, Journal of Business Research, Journal of Business Ethics 등의 저널에 다수의 논문을 게재하였으며, 현재 Business Ethics, CSR/ESG, Stakeholder Activism, Employee Experience Design 등을 연구하고 있다.

논문접수일 : 2023년 03월 07일 **게재확정일 : 2023년 05월 15일**
1차 수정일 : 2023년 05월 02일