JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 4, pp. 21-29, April 2023
https://doi.org/10.9708/jksci.2023.28.04.021

Improving the Classification of Population and Housing Census with AI: An Industry and Job Code Study

Byung-Il Yun*, Dahye Kim*, Young-Jin Kim*, Medard Edmund Mswahili**, Young-Seob Jeong**

*Reseacher, FS Inc, Daejeon, Korea
*Reseacher, FS Inc, Daejeon, Korea
*Reseacher, FS Inc, Daejeon, Korea
**Student, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea
**Professor, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea

[Abstract]

In this paper, we propose an AI-based system for automatically classifying industry and occupation codes in the population census. The accurate classification of industry and occupation codes is crucial for informing policy decisions, allocating resources, and conducting research. However, this task has traditionally been performed by human coders, which is time-consuming, resource-intensive, and prone to errors. Our system represents a significant improvement over the existing rule-based system used by the statistics agency, which relies on user-entered data for code classification. In this paper, we trained and evaluated several models, and developed an ensemble model that achieved an 86.76% match accuracy in industry and 81.84% in occupation, outperforming the best individual model. Additionally, we propose process improvement work based on the classification probability results of the model. Our proposed method utilizes an ensemble model that combines transfer learning techniques with pre-trained models. In this paper, we demonstrate the potential for AI-based systems to improve the accuracy and efficiency of population census data classification. By automating this process with AI, we can achieve more accurate and consistent results while reducing the workload on agency staff.

▸ **Key words**: Natural language process, Classification, Population survey, Ensemble model, Industrial and Occupation code

• First Author: Byung-Il Yun, Corresponding Author: Young-Seob Jeong
*Byung-Il Yun (fspanda@fscom.kr), FS Inc
*Dahye Kim (ekgp908@gmail.com), FS Inc
*Young-Jin Kim (bada0179@nate.com), FS Inc
**Medard Edmund Mswahili (medardedmund25@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
**Young-Seob Jeong (ysjay@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
• Received: 2023. 02. 14, Revised: 2023. 03. 20, Accepted: 2023. 03. 21.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문에서는 인구 조사에서 산업 및 직업 코드를 자동 분류하기 위한 인공지능 기반 시스템을 제안한다. 산업 및 직업 코드의 정확한 분류는 정책 결정, 자원 할당 및 연구를 위해 매우 중요하지만, 기존의 방식은 사람이 작성한 사례 사전에 의존하는 규칙 기반 방식으로 규칙 생성에 필요한 시간과 자원이 많이 소요되며 오류 발생 가능성이 높다. 우리는 본 논문에서 통계 기관에서 사용하는 기존의 규칙 기반 시스템을 대체하기 위해 사용자가 입력한 데이터를 이용하는 인공지능 기반 시스템을 제안하였다. 이 논문에서는 여러 모델을 학습하고 평가하여 산업에서 86.76%의 일치율, 직업에서 81.84%의 일치율을 달성한 앙상블 모델을 개발하였다. 또한, 분류 확률 결과를 기반으로 프로세스 개선 작업도 제안하였다. 우리가 제안한 방법은 전이 학습 기술을 활용하여 사전 학습된 모델과 결합하는 앙상블 모델을 사용하였으며, 개별 모델과 비교하여 앙상블 모델의 성능이 더 높아짐을 보였다. 본 논문에서는 인공지능 기반 시스템이 인구 조사 데이터 분류의 정확성과 효율성을 향상시키는 잠재력을 보여주며, 인공지능으로 이러한 프로세스를 자동화함으로써 더 정확하고 일관된 결과를 달성하며 기관 직원의 작업 부담을 줄일 수 있다는 점을 보여준다.

▸ **주제어**: 자연어 처리, 분류, 인구 주택 총 조사, 앙상블 모델, 산업 및 직업 코드

I. Introduction

통계청의 인구 주택 총 조사 사업은 우리나라의 모든 사람과 주택을 조사하여 정책 수립 및 평가의 기초자료를 제공하고, 각종 인구 및 가구 대상 조사의 표본 추출 틀을 제공, 대학·연구기관·민간 기업체 등 각종 학술연구와 경영 기초 자료로 활용하기 위해 수행하는 사업이다.

이러한 인구 총 조사에서는 응답자의 설문 결과를 토대로 산업 코드 및 직업 코드를 분류한다. 과거에는 이러한 코딩 작업은 책자를 이용해 사람이 수작업으로 수행하였으나, 현재는 규칙 및 사례 기반 DB를 이용한 자동 코딩 시스템을 이용해 분류를 수행하고 있다.

그러나 이러한 자동 코딩 분류 시스템은 대량의 사례 색인 DB를 구축해야 한다는 문제를 가지고 있으며, 규칙 생성 방식도 사용자의 경험에 의존적이다. 또한 규칙에 해당하지 않은 데이터를 처리하기 위해서는 내용검토 요원이 직접 수작업을 이용해 분류하여야 한다.

이러한 몇 가지 문제점을 해결하기 위해 본 연구에서는 딥러닝 기반 자동 분류 모델을 제안하고자 한다. 딥러닝 기반 자동 분류 모델은 트랜스포머 기반의 사전 언어 모델을 이용해 자연어 처리를 수행하며, 통계청에서 수집한 분류 데이터를 이용해 지도 기반 분류의 학습을 통해 구축된다. 이 과정에서 한국어 사전 언어 모델 중 공개되어 있는 여러 모델들의 성능을 비교 평가하였으며, 다양한 모델을 결합한 앙상블 모델을 설계하고 그 성능 또한 평가하였다. 마지막으로 딥러닝 모델들을 이용해 코드를 분류하며, 얻을 수 있는 분류 확률을 통해 내용 검토 요원의 수동 검토 작업량을 최소화할 수 있는 내용 검토 프로세스에 대해 제안하였다.

본 논문의 기여도를 나열하면 다음과 같다. 첫째, 통계청에서 수집한 대용량의 인구 총 조사 데이터를 이용해 코드를 자동 분류할 수 있는 딥러닝 모델을 제안하였고, 기존 모델보다 더 높은 성능을 보일 수 있음을 보였다. 둘째, 여러 한국어 사전 언어 모델들에 대한 산업/직업 분류 성능을 평가하고 여러 언어 모델을 결합한 앙상블 모델을 설계하고 이를 통해 그 성능이 더 높아질 수 있음을 확인하였다. 마지막으로 이러한 딥러닝 모델을 통해 자동 분류를 하였을 때 얻을 수 있는 분류 확률을 이용해 내용 검토 작업량을 최적화시킬 수 있는 검토 프로세스 개선안에 대해 제안하였다.

본 논문의 구성은 다음과 같다. 먼저 챕터 2에서는 관련 연구 및 본 논문의 이해를 위한 백그라운드 지식에 대해 설명한다. 챕터 3에서는 데이터와 사용한 모델에 대한 설명 및 실험 결과와 그 결과에 대한 분석으로 이루어져있다. 추가로 실험 결과 및 분석을 토대로 실제 업무에서 직업 및 산업 코드 분류에 대한 노동력 소모를 줄이기 위한 개선된 방식을 제안하였다. 마지막으로 챕터 4에서는 결론에 대해 작성하였다. 결론은 논문의 내용을 정리하고, 한계점에 대해 이야기하였다.

II. Preliminaries

1. Related works

미국에서는 ACS(American Community Survey)를 토대로 산업/직업에 관한 데이터를 수집하여, 자동 분류하려는 연구가 진행되고 있다. Thompson et al.[1]은 수집한 데이터를 기반으로 데이터 사전을 구축하여, 이를 통해 코드의 후보를 생성하고, 간단한 모델을 사용하여 최종 코드를 결정한다. 독일에서는 IAB(German Institute for Employment Research)에서 수집한 데이터를 통해 직업 코드를 분류한 연구가 있다. Bethmann et al.[2]은 Naive Bayes와 Bayesian Multinomial을 사용하여 기계학습을 통한 자동 코딩 분류 시스템의 성능과 실현 가능성을 평가했다.

국내에서도 인구 총 조사의 데이터를 이용한 자동 분류 연구가 진행되고 있다. Kang et al.[3]은 과거 조사 결과를 기반으로 인덱스 DB를 구축하고, p-norm 모델을 사용한 코드 자동 부여 시스템을 개발하였다. 문장에서 명사 추출을 위한 형태소 분석기와 단어의 상대적 중요도를 반영한 가중치 계산을 통해 코드를 변환한다. Lim[4]은 형태소 분석기와 품사 태거를 기반으로 가중치를 계산한 색인어 추출기로 색인 DB를 구축하였고, 불필요한 명사를 제거하기 위한 불용어 사전을 사용하였다. kNN(k nearest neighbors) 을 통해 분류 코드의 후보를 생성하고, DVF(discrete value function)와 SF(similarity based function)를 정의하여 최종 분류 코드를 결정한다. Lim[5]은 코드 분류에 적합한 색인어를 추출하기 위해 형태소분석, 명사 추출, 바이그램의 3가지 방법을 비교 평가하였다. 결과적으로 바이그램과 kNN 기반의 색인어 DB를 사용한다. 이 시스템은 잘못 분류된 코드에 대해 자동으로 피드백 하여 학습 데이터의 신뢰성을 재조정하여 개선한다.

최근 국내에서는 색인 DB를 사용하지 않는 딥러닝 기반의 연구들이 등장하고 있다. 통계청에서 실시하는 조사의 결과를 입력받아 Woo and Lim[6]은 CNN과 LSTM을 사용하여 대분류 코드를 분류하였으며, Lim et al.[7]은 사전 학습된 언어 모델인 KoBERT를 사용하여 높은 성능을 보였다. 그러나 기존 연구에서는 특정 분류의 코드만을 대상으로 실험을 진행하거나, 상위 10개 포함의 정확도는 높지만 완전 일치 정확도(Accuracy)는 낮은 한계가 있다.

2. Backgroud

2.1 Pre-Trained Language Model

사전 학습 언어 모델을 이용한 분류 과정은 Fig. 1에 설명되어 있다. 분류가 필요한 문장은 첫 번째로 모델의 토크나이저에 의해 토큰, 혹은 단어의 시퀀스로 변화시킨다. BERT로 대표되는 트랜스포머 모델들은 이러한 단어의 시퀀스를 단어의 자체 의미뿐 아니라 위치 정보를 포함한 임베딩 데이터로 변환시킨다. 이러한 임베딩 데이터는 Attention 알고리즘을 사용하는 Transformer 인코더 모델을 통해 각자의 임베딩 데이터에 영향을 받아 문맥 정보를 포함한 최종 임베딩 토큰으로 출력된다. 이러한 모델은 임베딩 과정에서 [CLS] 토큰 및 [SEP] 토큰을 추가하며, [CLS] 토큰은 문장의 전체 의미를 담고 있는 토큰이다. 따라서 분류 작업에서는 [CLS] 토큰을 이용해 최종 분류 작업을 수행한다. Classification Layer는 이러한 [CLS] 벡

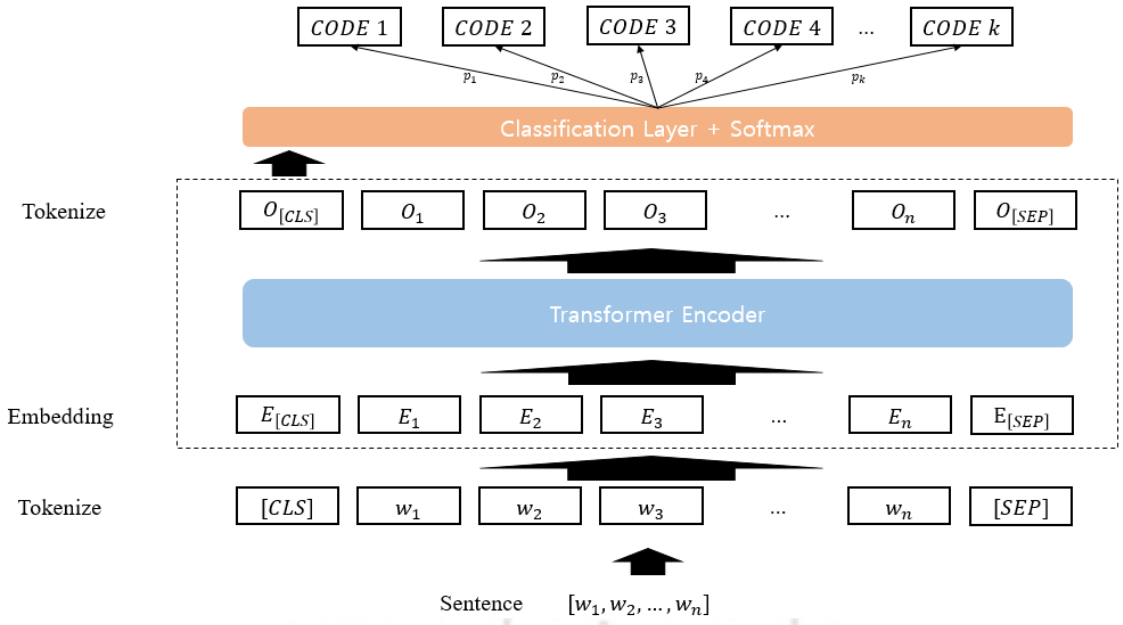

Fig. 1. Classification process using transformer language model

터를 입력받고 클래스 레이블을 예측하는 데 사용할 수 있는 정규화 되지 않은 값을 출력한다. 따라서 입력받는 벡터의 크기는 언어 모델이 출력하는 임베딩 벡터의 크기이며, 출력 크기는 예측하고자 하는 클래스 레이블의 수와 같다. 정규화 되지 않은 값을 확률 분포로 매핑하기 위해 소프트 맥스 활성화 함수를 사용한다. 이 함수를 사용하여 Classification Layer가 각 라벨의 예측 값을 $a=[a_1,a_2,...,a_k]$로 출력했을 때, x 번째 라벨의 확률 값을 표현하는 수식은 (1)과 같다.

$$softmax(x)=[\exp(a_x)/\sum_{i=1}^{k}(\exp(a_i)] \quad (1)$$

여기서 exp는 지수 함수를 의미한다. 각 라벨의 소프트맥스 값의 합은 1이 되며, 이러한 소프트맥스 값은 라벨 예측에 대한 확률 값으로 사용할 수 있다. Fig. 1의 p_x는 x번째 라벨의 소프트맥스 값을 의미한다.

2.2 Ensemble Model

앙상블 모델은 보다 성능이 높은 모델을 만들기 위해 여러 개별 모델을 조합한 형태의 모델을 의미한다. 이 모델의 목표는 여러 모델의 예측을 결합해 예측 성능을 향상시키는 것이다. 이러한 예측을 결합하는 방법은 Bagging이나 Boosting 등의 방법이 있다. 이러한 앙상블 모델을 사용하면 단일 모델과 비교해 얻을 수 있는 몇 가지 이점이 있다. 첫 번째는 단일 모델의 분산을 줄일 수 있어 안정적인 예측이 가능해진다. 둘째, 여러 모델의 장점을 결합하여 모델의 전체적인 정확도를 높이는 데 도움을 줄 수 있다. 셋째, 모델 하나에 대한 의존도를 줄임으로써 견고성을 높이는 데 도움이 될 수 있다. 마지막으로, 여러 모델의 예측을 함께 분석해 패턴과 추세를 식별할 수 있기 때문에 더 해석 가능한 결과를 얻을 수 있다.

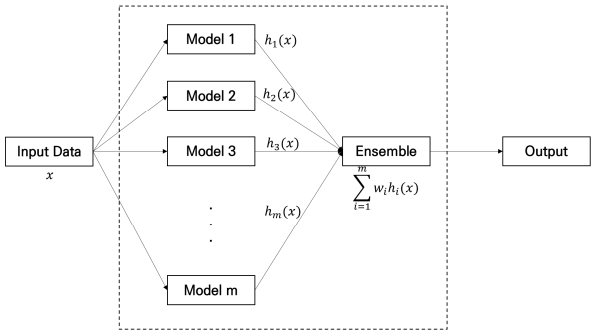

Fig. 2. Ensemble using bagging

사전 언어 모델 기반의 분류 모델의 앙상블은 일반적으로 Bagging 방식으로 이루어진다. Bagging 방식은 각 모델이 동일한 데이터로 학습한 후, 각 모델이 출력한 예측 결과에 가중치를 고려하여 더하는 방식이다. 본 연구에서도 Bagging 방식으로 여러 언어 모델을 학습한 후 최종 예측은 각 모델의 예측치를 합산하여 구하는 방식이다.

Fig. 2는 앙상블 모델의 구현 방식을 설명한 그림이다. 특정 문장 x 에 대한 모델 y의 라벨 예측 값을 $h_y(x)$라고 정의할 때 $h_y(x)$는 라벨 수만큼의 길이를 가진 1차원 벡터 값이다. 총 m개의 모델을 대상으로 앙상블 모델을 구성할 경우 $h_1(x)$ ~ $h_m(x)$ 의 총 m 개의 예측 값이 나오며 각 예측 값에 가중치 $w=[w_1,w_2,\dots,w_m]$를 곱해 최종 예측 값을 얻는다. 이 과정을 설명한 수식은 아래 (2)와 같다.

$$y(x)=\sum_{i=1}^{m} w_i h_i(x) \quad (2)$$

III. Methodology

1. Data & Model

1.1 Data

본 연구에서는 2020년 인구 주택 총 조사의 산업/직업 분류 내용 검토 데이터를 이용하였다. 최종 산업/직업분류 코드는 통계청 내용 검토 시 3개의 단계를 거쳐 확정되는데 분석데이터는 최종 내용 검토 파일을 이용하였다. 총 데이터는 4,099,010개로, 학습데이터와 시험데이터의 비율은 각각 80%, 20%로 하였으며, 학습데이터를 다시 학습데이터와 검증 데이터로 90%:10%의 비율로 분리하였다. 분석 데이터의 레이아웃 및 변수는 아래의 Table 1 과 같다. 산업/직업분류는 세분류(4자리) 수준에서 코딩이 되어 있으며, 연관 정보(성별, 만 나이, 교육 정보, 교육 상태, 종사상 지위, 근로 장소)를 산업/직업 텍스트에 추가로 고려하여 코드 예측 시 성능 개선 여부를 확인하고자 하였다. 여기서 '산업 직업 텍스트'는 산업의 2가지 텍스트(직장, 사업체명, 주된 사업내용)와 직업의 3가지 텍스트(일의 종류, 근무부서, 직책)를 공백을 사이에 두고 모두 연결한 텍스트를 의미한다.

Table 1. Data Layout

Column		Type	Length
ID		NUMBER	8
Industrial	Office Name	VARCHAR2	200
	Main Business	VARCHAR2	200
	Code	VARCHAR2	4
Occupation	Type of Work	VARCHAR2	200
	Department	VARCHAR2	200
	Position	VARCHAR2	200
	Code	VARCHAR2	4
Correlated Variable	Sex	VARCHAR2	1
	Age	NUMBER	4
	Education	VARCHAR2	1
	Education Status	VARCHAR2	1
	Job Status	VARCHAR2	1
	Job Place	VARCHAR2	1

1.2 Pre-Trained Model

최근 한국어 기반 사전 학습 언어 모델은 기업과 개인을 가리지 않고 오픈소스로 활용할 수 있도록 공개된 모델들이 많이 존재한다. 우리는 각 모델들을 이용해 산업 및 직업 분류 태스크 성능을 살펴보았으며, 각 태스크에서 좋은 성능을 보인 5개의 모델을 이용해 앙상블 모델을 구현하였다. 실험에 사용한 모든 모델들은 Table 2에 기록되어 있으며, 각 모델의 서로 다른 버전도 이용하였다.

앙상블 모델은 Fig. 3과 같이 인구 총조사 데이터의 코드 분류 학습 데이터로 각 사전 언어 모델을 분류 모델로 학습한 파인 튜닝 모델들에 특별한 가중치를 주지 않은 Soft Voting 방식을 이용해 구현하였다.

Table 2. List of Pre-Trained Language Model

Type	Model
Organization	KE-T5 (KETI) [8]
	KLUE BERT(KLUE) [9]
	KoBERT (SKT) [10]
	KoGPT (KAKAO) [11]
	KoGPT2 (SKT) [12]
	TUNib Electra (TUNib) [13]
Personal	LASSL BERT [14]
	KcBERT [15]
	KcElectra [16]
	KoBigBird [17]
	KoELECTRA [18]

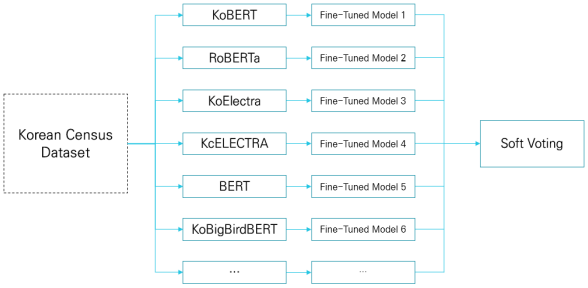

Fig. 3. PLM Ensemble Model

2. Result

Table 3은 개별 모델 중 높은 일치 율을 보이는 모델의 결과 내용이다. 평가 지표는 F1 Score의 클래스 평균값을 사용하였다. 여기서 평균값은 Macro 평균과 Micro 평균, Weight 평균을 모두 사용하였다.

$$f(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} \quad (3)$$

$$\mathrm{Precision}(c) = \frac{\sum_{i=1}^{n} (f(y_i,\hat{y}_i) \cdot f(y_i,c))}{\sum_{i=1}^{n} f(\hat{y}_i,c)} \quad (4)$$

$$Recall(c) = \frac{\sum_{i=1}^{n} (f(y_i,\hat{y}_i) \cdot f(y_i,c))}{\sum_{i=1}^{n} f(y_i,c)} \quad (5)$$

$$F1\,Score(c) = 2 \cdot \frac{\mathrm{Precision}(c) \cdot Recall(c)}{\mathrm{Precision}(c) + Recall(c)} \quad (6)$$

$$Macro\,F1 = \frac{\sum_{i \in c} F1\,Score(i)}{Nmber\,of\,c} \quad (7)$$

Table 3. Result of Classification (Single & Ensemble Model) ; Accuracy = F1 (Micro)

Industrial Code				Job Code			
Model	F1 (Macro)	F1 (Micro)	F1 (Weight)	Model	F1 (Macro)	F1 (Micro)	F1 (Weight)
Ensemble Model	**78.601**	**86.758**	**86.649**	**Ensemble Model**	**72.565**	**81.841**	**81.642**
KcBERT-large	77.388	85.889	85.823	KoBigBirdBERT-base	71.131	81.045	80.386
KoBigBirdBERT-base	76.988	85.830	85.734	BERT-base	71.393	80.956	80.869
KcBERT-base	77.078	85.828	85.734	KcBERT-base	71.042	80.808	80.635
BERT-base	76.729	85.707	85.608	KcBERT-large	71.289	80.511	80.773
KcELECTRA v3	75.918	85.195	85.079	KcELECTRA v3	69.816	80.227	80.008

Table 4. Data Review Result of the Inspector

Type	Data Label		AI Label		Etc		Sum
	Count	Ratio(%)	Count	Ratio(%)	Count	Ratio(%)	
IND	5,880	38.97	9,200	60.97	10	0.07	15,090
JOB	5,822	39.11	9,065	60.89	0	0	14,887

$$Micro\,F1 = \frac{\sum_{i=1}^{n} f(y_i, \hat{y_i})}{n} \quad (8)$$

$$WeightF1 = \frac{\sum_{i \in c} p(i) \cdot \quad F1\,Score(i)}{Nmber\,of\,c} \quad (9)$$

여기서 n은 검증 데이터 셋의 데이터 개수이며, y_i는 i 번째 데이터의 실제 label 값, $\hat{y}_i$는 i 번째 데이터에 대해 모델이 예측한 label 값이다. 또 c는 특정 label 값을 의미하며, $p(i)$는 전체 데이터 중 라벨 i인 데이터가 차지하는 비율을 의미한다. 위와 같은 식으로 라벨별 F1 Score를 구할 수 있으며 라벨별 F1 Score의 평균값도 구할 수 있다. 본 연구에 사용한 데이터는 1개의 데이터가 1개의 라벨 값만을 가지므로 Micro F1 Score는 accuracy와 같은 값을 가지게 된다.

결과를 보면 각 모델들을 결합한 앙상블 모델이 다른 모델들보다 더 좋은 성능을 보이는 것을 확인할 수 있었다. Macro F1 Score 기준으로는 산업 코드 예측에서는 가장 좋은 성능을 보인 개별 모델보다 약 1.213%, 직업 코드 예측에서는 1.434%의 성능 향상을 보였으며 다른 F1 Score 평균도 모두 더 높은 성능을 보였다. 이를 통해 앙상블 모델이 개별 모델을 사용하는 것보다 전반적인 성능 향상이 일어남을 확인할 수 있었다. 개별 모델도 유사한 연구[7]과 비교했을 때 정확도 수치에서 20~30% 이상의 높은 수치가 나타남을 확인할 수 있었다. 이는 모델 선택의 차이로 해석되며, BERT의 한국어 초기 모델인 KoBERT 이후 많은 종류의 데이터로 학습한 공개 모델들이 늘어나며 Task에 적합한 모델들에 차이가 있다고 볼 수 있다. 본 연구에서도 동일한 인풋 데이터를 사용함에도 산업 코드와 직업 코드라는 다른 코드를 예측하는 작업에서 모델별로 성능이 달라짐을 확인할 수 있었다. 또한, 데이터의 양과 질이 향상됨으로 인해 fine-tuning 과정에서 모델의 예측력이 향상되었다고 볼 수 있을 것이다.

3. Analysis

분류 학습 모델에서 테스트 데이터의 품질은 모델 결과 분석에 상당한 영향을 미칠 수 있다. 실제에서 대상 변수의 분포를 정확하게 나타내는 테스트 데이터 세트는 모델의 성능을 잘 평가할 수 있다. 반면, 테스트 데이터 세트의 품질이 낮으면 편향되거나 신뢰할 수 없는 평가 결과를 보이며 모델의 효과에 대한 잘못된 결론을 초래할 수 있다. 본 연구에서 사용한 데이터는 인구 총 조사의 실제 데이터이며, 사람의 내용검토까지 끝낸 데이터로 품질이 높을 것으로 기대되나, 매우 많은 양의 데이터를 서로 다른 사람이 평가하였기 때문에 필연적으로 오류가 발생할 수밖에 없다. 본 연구에서는 이러한 오류가 모델의 성능을 실제보다 높거나 낮게 평가할 수 있음을 확인하기 위해 모델이 분류한 데이터를 다시 내용 검토하였다.

검토 요원은 인구 총 조사 데이터에 대한 내용검토를 진행한 경험이 있는 2명의 인원이며, 총 30일 동안 산업 15,090, 직업 14,887개의 데이터를 검토하였다. 검토 방식은 Table 1 과 같은 데이터를 토대로 라벨 결과 후보를 기존 데이터 라벨과 AI 분류 라벨의 2가지를 제공하였으며, 두 라벨 중 어떤 라벨이 기존 라벨인지 밝히지 않았다. 검토 요원은 주어진 데이터를 토대로 2개의 라벨 중 어떤 라벨이 더 적합한지 판단하며, 두 라벨 모두 적합하지 않은 경우 기타를 선택하도록 하였다.

검토 결과는 Table 4에 나타나 있다. 결과에서 볼 수 있듯 검토 요원은 불일치 데이터에 대해 AI 분류가 더 적합하다고 선택한 비율이 산업 60.97%, 직업 60.89%로 기존 데이터의 라벨을 선택한 비율인 38.97%, 39.11%보다 더 높음을 확인할 수 있다. 이러한 결과는 모델의 성능이 Table 3에 표시된 결과보다 실제로는 더 좋은 분류 성능을 나타낼 수 있음을 의미할 수 있다.

또한 우리가 사용한 데이터가 이미 검토가 끝난 데이터임에도 불구하고 불일치 데이터를 대상으로 다시 검토를 진행하자 기존 라벨의 선택 비중이 낮았다는 건 사람이 진행하는 검토 과정에도 오류가 발생하기 쉽다는 것을 의미한다. 이러한 오류가 발생하는 이유는 검토 요원이 데이터를 검토하는 과정에서 개인 경험에 따른 선택 차이가 발생하며, 또 대규모 데이터를 다루며 피로감으로 인한 오류가 발생할 수 있기 때문이다[19].

다음 챕터에서는 이러한 오류를 해결하기 위해 분류 확률을 이용한 검토 프로세스의 개선 방법에 대해 제안한다.

4. Improving Content Review Process

인구 총 조사와 같이 대규모의 데이터를 검토하기 위해서는 많은 검토 요원이 투입되어야 하며, 시간과 비용의 소모가 크다. 또한 이러한 많은 검토 요원은 동일한 매뉴얼을 이용해 판단한다고 하더라도, 서로 다른 경험과 생각을 갖고 있기 때문에 동일한 데이터에 대해 서로 다른 분류 검토를 진행해 데이터 오류를 발생시킬 수 있다[20].

이를 해결하기 위해 본 연구에서는 분류라벨의 확률을 분류 신뢰도로 정의하고 이를 이용해 검토가 필요한 데이터의 양을 줄일 수 있는 프로세스를 제안한다.

인공지능 기반 코드 분류로 얻을 수 있는 결과는 코드 자체뿐 아니라 코드가 가지고 있는 분류 확률을 얻을 수 있다. 이 분류 확률은 Fig. 1에서 설명한 Classification Layer의 가중치를 소프트맥스 함수를 사용해 확률 분포로 나타낸 값으로, 완벽한 확률 값으로 볼 수는 없지만 더 높은 확률 값을 가질 때 분류 정확성이 높아지는 경향을 보이는 사실은 많은 연구를 통해 밝혀졌다[21,22].

본 연구에서도 이를 확인하기 위해 분류 확률에 따른 데이터의 분류 정확성을 측정하였다. Fig. 4는 분류 확률을 임계값으로 두었을 때, 임계값을 넘는 데이터 비율과 분류 확률이 임계점 이상인 데이터의 일치율을 나타낸 그래프이다. 임계값이 90%일 때 임계값을 넘는 데이터의 비율은 산업 전체의 72.495%, 직업 전체의 61.187%이며, 이때 분류된 데이터가 기존 라벨과 일치할 확률은 산업 98.092%, 직업 98.050%이다. 임계값이 80%일 때 임계값을 넘는 데이터의 비율은 산업 전체의 78.862%, 직업 전체의 69.917%이며, 이때 분류된 데이터가 기존 라벨과 일치할 확률은 산업 96.849%, 직업 96.288%이다.

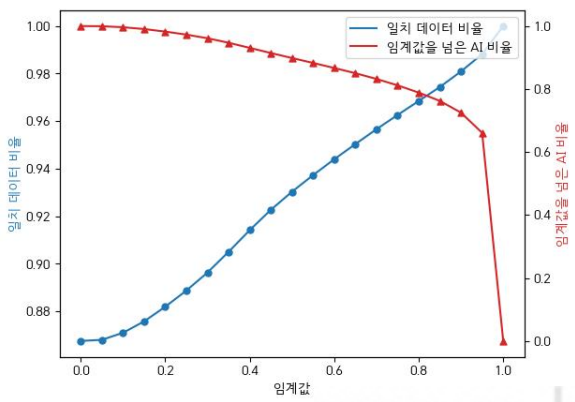

Fig. 4. Threshold-based Data Exceeding and Match Rate

이를 토대로 사용자의 기준에 따라 일치 데이터 비율 및 임계값 초과 데이터 비율을 근거로 임계값을 설정하고 이러한 임계값을 초과한 데이터는 검토 대상에 넣지 않을 수 있다. 만약 임계값을 0.9로 설정한다면 이 값을 넘는 분류 확률을 가진 데이터는 검토 대상이 되지 않으므로, 검토 요원은 산업 27.505%, 직업 38.873%의 데이터만 검토하게 되어 시간, 비용 소모 및 대규모 데이터를 검토하며 발생하는 오류 문제도 줄어들 수 있다. 이때 모델의 코드 분류 성능은 본래 데이터의 98%가 넘는 일치율을 보이므로 상당히 신뢰할 수 있는 데이터 분류기로 활용할 수 있을 것이다. 또 모델의 신뢰성을 높이기 위해 불일치 샘플에 대한 검토 후 재학습을 계속 수행한다면, 모델의 성능 향상을 기대할 수 있을 것이다.

이러한 방식은 기존 자동 코딩 방식으로는 수행할 수 없는 방식이며 검토 요원을 줄여 시간과 비용의 소모를 줄일 수 있을 뿐 아니라 데이터 오류 발생률도 오히려 줄일 수 있는 방안이 될 것이다.

IV. Conclusions

본 연구는 인구 총 조사 데이터를 이용해 데이터의 산업 코드 및 직업 코드를 분류하는 인공지능 모델을 개발하고 그 성능 및 활용 방안에 대해 연구하였다. 인공지능 모델의 개발은 공개되어 있는 한국어 사전 모델을 이용하였다.

기존 유사한 연구에서는 한 가지 모델에 대해서만 성능을 확인하였으나, 본 연구에서는 여러 모델을 비교 분석하여 모델간의 성능을 비교하였으며, 높은 성능을 보여주는 모델들만 결합해 앙상블 모델로 만들어 성능을 확인하였다. 이러한 앙상블 모델의 성능이 단순 개별 모델만 사용하는 것보다 산업과 직업 모든 분야에서 더 높은 정확성을 보여주는 것을 확인하였으며, 기존 연구와 비교하였을 때 매우 높은 성능 향상을 이루었음을 확인하였다.

또한 모델이 분류 결과와 기존 분류 결과를 다시 한 번 비교 검토함으로써 모델의 분류 결과와 테스트 데이터 셋의 결과가 불일치하더라도 항상 모델의 분류 결과가 틀린 것은 아님을 확인하였고, 모델의 성능이 실험 결과보다 더 나을 수 있음을 보였다. 또 대규모 데이터의 내용검토 과정에서 발생하는 시간과 비용의 소모를 줄이기 위해 모델이 출력하는 라벨 확률을 이용한 검토 프로세스를 제안하였다. 이러한 방식을 사용하면 많은 검토 요원을 사용함으로써 발생하는 시간, 비용 소모를 줄일 뿐 아니라 데이터 분류 오류를 줄일 수 있음을 주장하였다.

본 논문의 한계점은 앙상블 모델 결합 방식을 단순한 Soft Voting만을 사용했다는 점이다. 본 논문에서는 여러 가지 모델의 분류 성능을 평가하였는데, 각 클래스에 대해 여러 모델의 성능을 평가하여 클래스 분류 성능에 따라 모델 가중치를 주는 방식을 사용하면 더 나은 성능 분류를 보일 수 있었을 것이다.

향후 연구에서는 이러한 점을 반영하여 분류 결과의 분석을 다양화하고 이를 반영한 앙상블 모델을 설계하고자 한다. 예를 들어 각 업종 별로 분류 성능이 높은 모델들의 가중치를 높이고 낮은 모델들의 가중치를 낮춰, 업종별 성능을 향상시키는 방안을 고려할 수 있다.

ACKNOWLEDGEMENT

This paper is a basic research project supported by the National Research Foundation of Korea with funding from the government (Ministry of Education) in 2020 (No. NRF-2020R1I1A3053015).

REFERENCES

[1] Thompson, Matthew, Michael E. Kornbau, and Julie Vesely. "Creating an automated industry and occupation coding process for the American Community Survey." Unpublished. (accessed October 10, 2016) (2012).

[2] A. Bethmann, M. Schierholz, K. Wenzig and M. Zielonka, “Automatic Coding of Occupations.” In Proceedings of Statistics Canada Symposium, Quebec, Canada, August 29-31, 2014 (accessed October 10, 2016).

[3] Y. K. Kang, “Automatic coding system for industry and occupation classification.” The Korean Association for Survey Research. Fall Conference 2001, pp. 33-45, 2001.

[4] H. S. Lim, “An automated Classification System of Standard Industry and Occupation Codes by Using Information Retrieval Techniques.” The Journal of Korean Association of Computer Education, pp. 51-60, July, 2004.

[5] H. S. Lim, “An Example-based Korean Standard Industrial and Occupational Code Classification”, Journal of the Korea Academia-Industrial cooperation Society, pp. 594-601, July 2006.

[6] C. K. Woo and H. S. Lim, “Comparison of Korean Standard Industrial Classification Automatic Classification Model on Deep Learning”, Proceedings of the Korea Information Processing Society Conference, pp. 516-518, 2020.

[7] J. Lim, H. Moon, C. Lee, C. Woo, and H. Lim, “An Automated Industry and Occupation Coding System using Deep Learning”, Journal of the Korea Convergence Society, pp. 23-30, December 2021.

[8] Kim, San, et al. "A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems." Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.

[9] Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).

[10] SKTBrain, “KoBERT”, https://github.com/SKTBrain/KoBERT

[11] Kim, Ildoo, et al. "Kogpt: Kakaobrain korean (hangul) generative pre-trained transformer." Opgehaal van https://github.com/kakaobrain/kogpt (2021).

[12] SKT-AI, “KoGPT2”, https://github.com/SKT-AI/KoGPT2

[13] Ha, Sangchun et al. "TUNiB-Electra.", https://github.com/tunib-ai/tunib-electra (2021).

[14] Lassl, “lassl/bert-ko-base”, https://huggingface.co/lassl/bert-ko-base

[15] Lee, Junbum. "Kcbert: Korean comments bert." Annual Conference on Human and Language Technology. Human and Language Technology, 2020.

[16] Lee, Junbum. "KcELECTRA: Korean comments ELECTRA." GitHub repository. Opgehaal van https://github.com/Beomi/KcELECTRA (2021).

[17] Jangwon Park, et al. "KoBigBird: Pretrained BigBird Model for Korean." (2021).

[18] Park, Jangwon. "KoELECTRA: Pretrained ELECTRA Model for Korean." . (2020).

[19] Alqershi, Fattoh, et al. "A robust consistency model of crowd workers in text labeling tasks." IEEE Access 8 (2020): 168381-168393.

[20] Draws, Tim, et al. "The effects of crowd worker biases in fact-checking tasks." 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

[21] Le, Quoc V. "A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks." Google Brain 20 (2015): 1-20.

[22] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Authors

Byung-Il Yun received the B.S., M.S. and Ph.D. degrees in Industrial and System Engineering from KAIST, Korea, in 2010, 2012 and 2021, respectively. Dr. Yun joined the FS Inc, Daejeon, Korea, in 2021. He is currently a Senior Research in the Department of AI development, FS Inc. He is interested in natural language processing, time series forecasting and artificial intelligence.

Dahye Kim received the B.S. and M.S. degrees in Bigdata Engineering from SoonChunHyang University, Korea, in 2020 and 2022, respectively. She joined the FS Inc, Daejeon, Korea, in 2022. She is currently a Research in the Department of AI development, FS Inc. She is interested in natural language processing and artificial intelligence.

Mr. Young-Jin Kim founded FS Corporation in Daejeon, South Korea in 2009 and has been managing the company ever since. He and FS Corporation is engaged in the domains of big data and AI business.

Medard Edmund Mswahili received his M. Sc. Eng in Big Data Engineering, ICT convergence department from Soonchunhyang University, Asan, South Korea in 2022. He's currently pursuing his Ph.D. as a research assistant at Data Analysis and Artificial Intelligence Lab in the Computer Engineering department at Chungbuk National University, Cheong-ju city, South Korea. His research interest mainly focuses in Machine & Deep Learning in Pharmaceutical data analysis (drug discovery & development) and natural language processing.

Young-Seob Jeong received the M.S., Ph.D degree in computer science from KAIST, Daejeon, Korea, in 2012 and 2016, respectively. He is a faculty member of the department of computer engineering, Chungbuk National university, Cheong-ju city, Korea. His current research topics include malware detection using deep learning techniques, language models, healthcare system for patients, and pharmaceutics.