이진분류 논문 소개

한글 텍스트 감정 이진 분류 모델 생성을 위한 미세 조정과 전이학습에 관한 연구

A Study on Fine-Tuning and Transfer Learning to Construct Binary Sentiment Classification Model in Korean Text

한국산업정보학회논문지 = Journal of the Korea Industrial Information Systems Research, v.28 no.5, 2023년, pp.15 - 30

김종수 ((주)골드브릿지 기업부설연구소)

발표자 : 서진형

요약

- 트랜스포머(Transformer) 구조를 기초로 하는 생성 모델이 크게 유행 ex)ChatGPT
- BERT 다국어 생성 모델을 미세 조정한뒤 한국 학습 데이터셋을 사용하여 전이학습을 진행
- 영화 리뷰 댓글을 긍정 또는 부정으로 분류하는 모델 생성

관련연구(모델간 비교)

1. 텍스트 감정 이진 분류 Colab 기본모델
2. RNN(Bidirectional) 기반 영문 텍스트 분류모델
3. BERT 기반 영문 텍스트 분류 모델

Colab 기본 모델

Fig. 2 Graph of learning and validation accuracy(left), training and validation loss(right)

- 정확도 0.8678
- 손실 0.3357
- F1 점수 0.8626

Bi-RNN 기반 모델

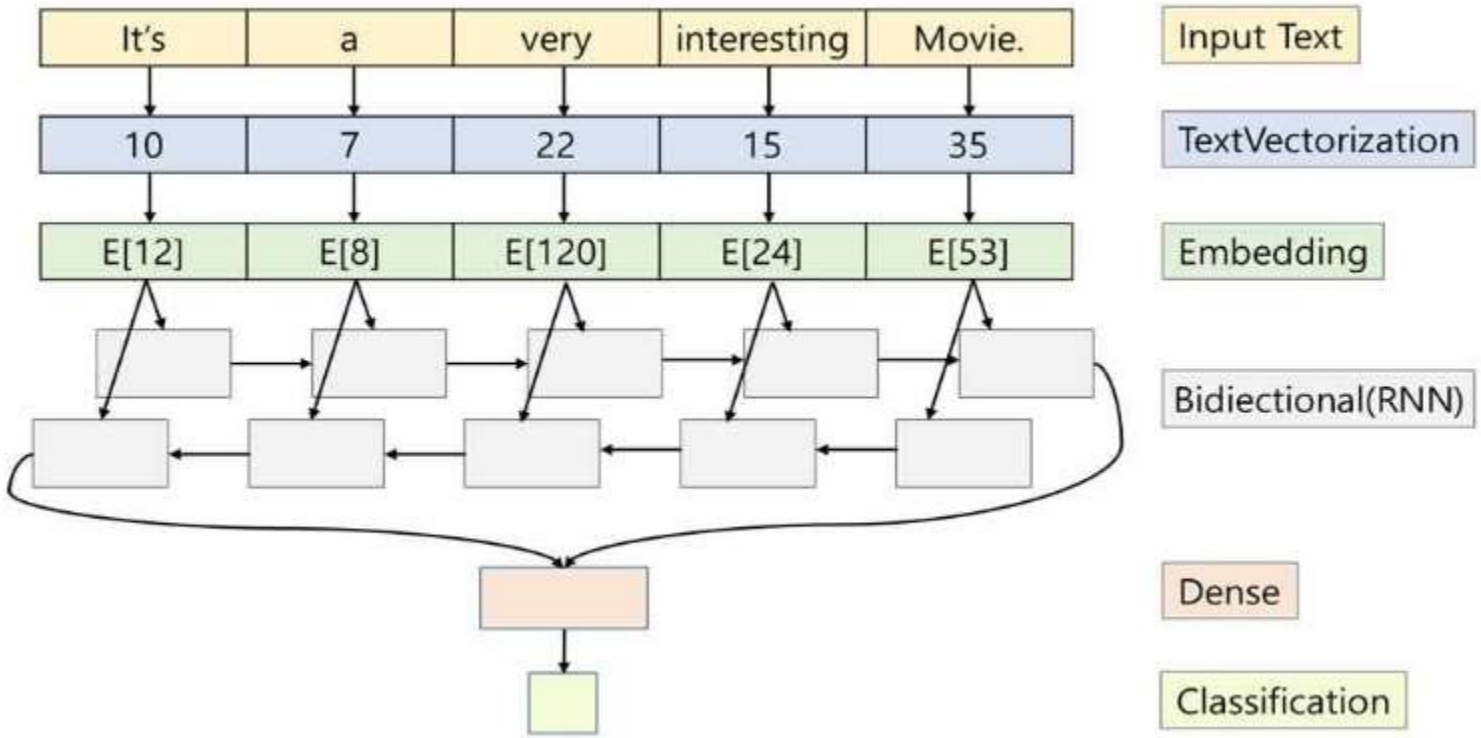


Fig. 3 RNN(recurrent neural network) model for text binary classification

- 정확도 0.8754
- 손실 0.2955
- F1 점수 0.8665

BERT 기반 영문 텍스트 분류 모델

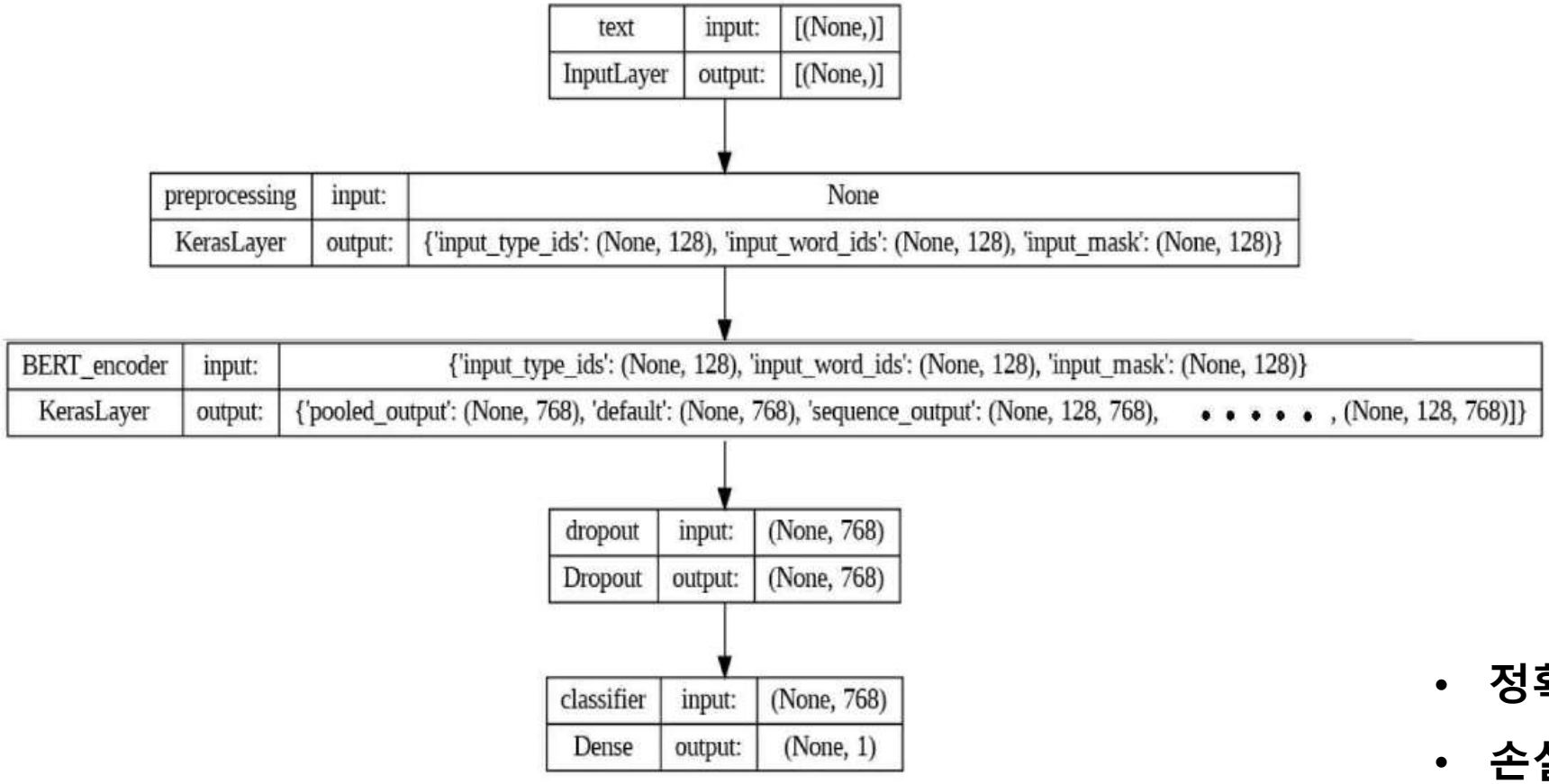


Fig. 4 Structure of a pre-trained BERT-based model for an English text binary classification

- **정확도 0.9805**
- **손실 0.0672**
- **F1 점수 0.9792**

BERT-Bases 모델 미세조정

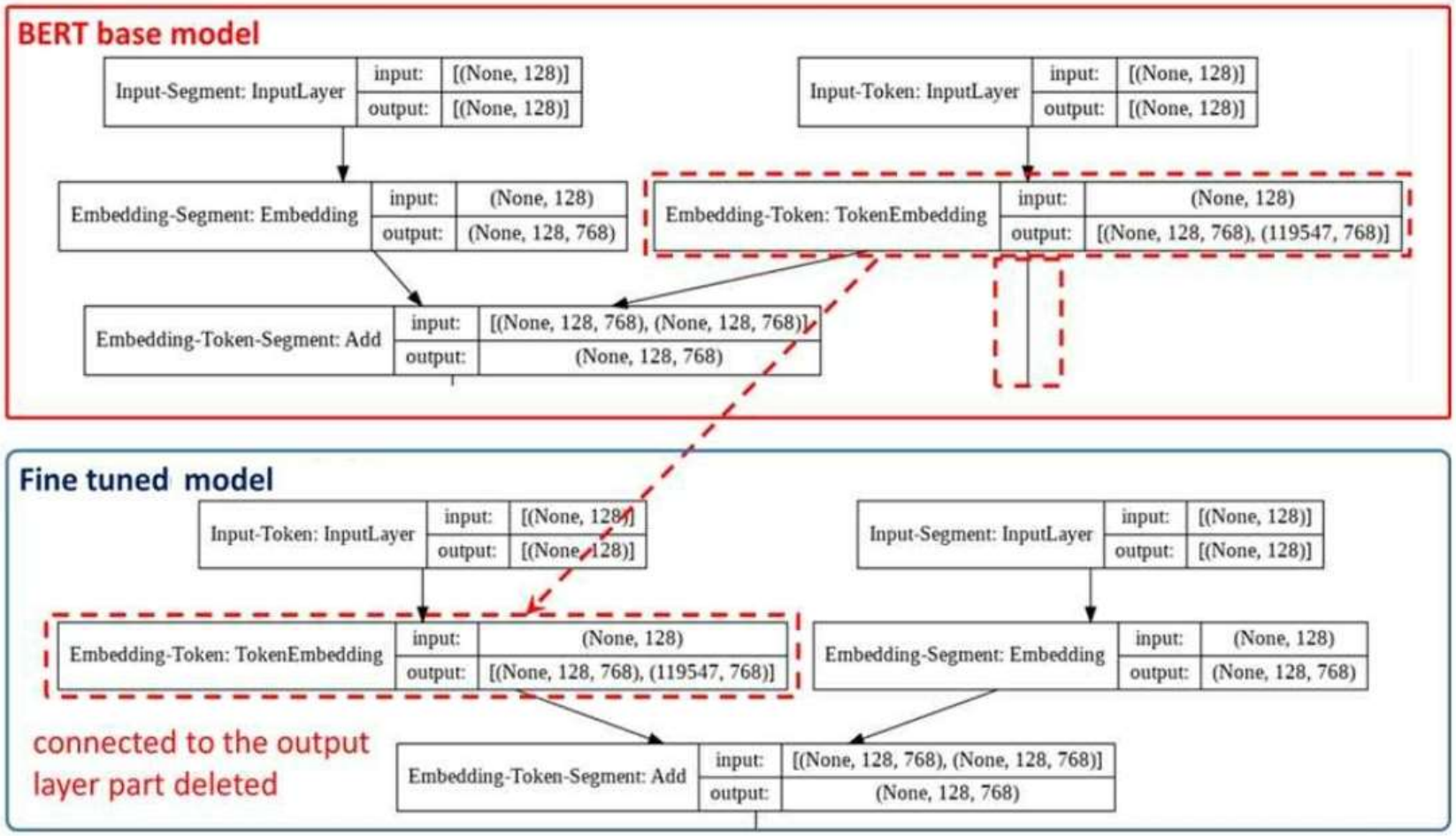


Fig. 8 Comparison of model input layers before(top) and after(bottom) refinement

BERT-Bases 모델 미세조정

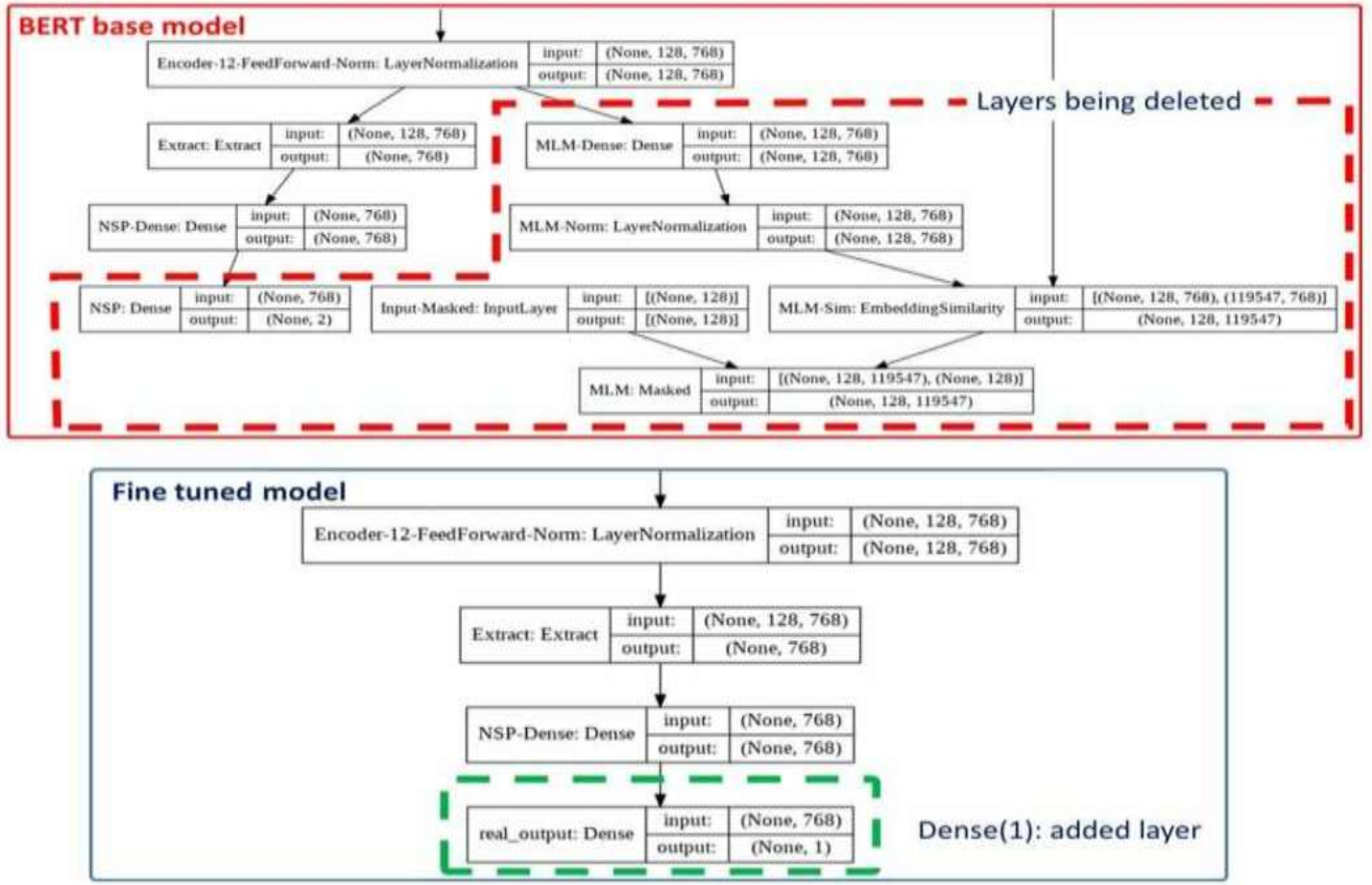


Fig. 9 Comparison of model output layers before(top) and after(bottom) refinemen

한글 문장 데이터셋 토큰화

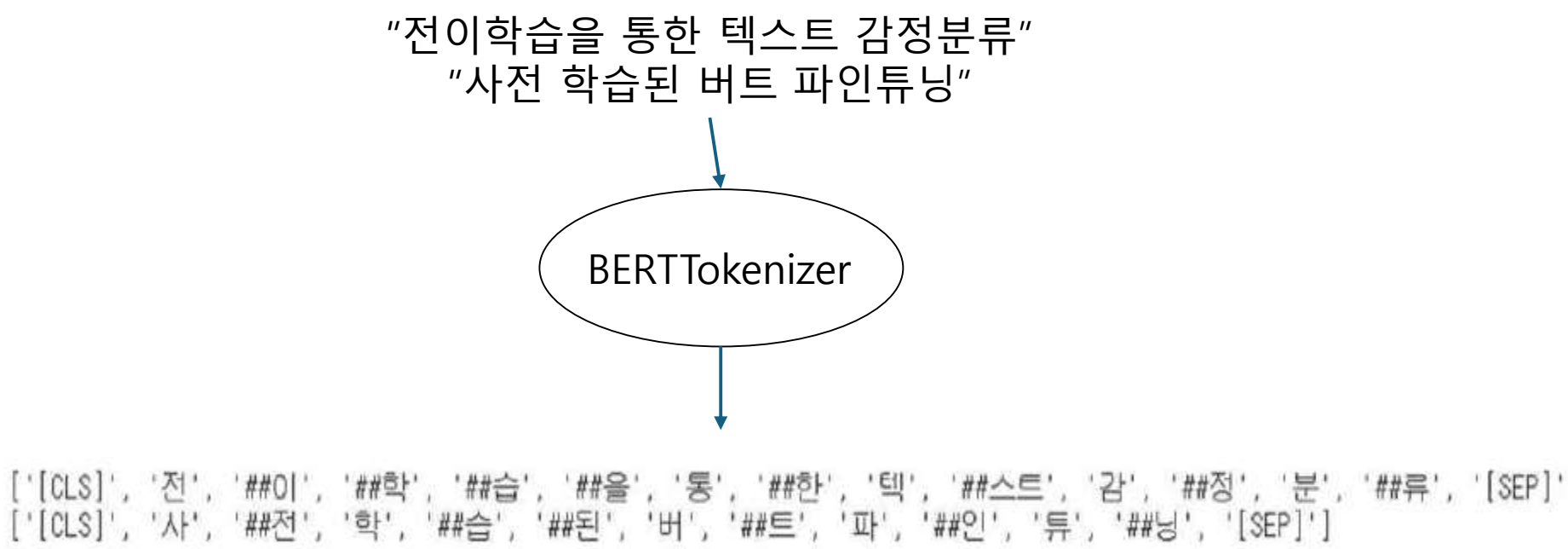


Fig. 11 An example of token generation

성능 평가

Category.	Dataset		loss	Acc.	Val loss	Val Acc.	F1 Score
	Train	Test					
Basic Mode	IMDB(English)		0.1407	0.9536	0.3008	0.8818	0.86
RNN(1 layer)			0.2968	0.8678	0.3194	0.8562	0.85
RNN(2 layer)			0.2955	0.8754	0.3187	0.8484	0.84
BERT			**0.0672**	**0.9805**	**0.7291**	**0.8604**	**0.86**
BERT	NEWS Review		0.0929	0.9656	0.2488	0.9124	0.91
Proposed BERT model (dataset=nsmc)	10000	5000	0.1177	0.9582	0.5237	0.8258	**0.81**
	20000	10000	0.0669	0.9497	0.6543	0.8430	**0.84**
	30000	15000	0.1107	0.9604	0.4917	0.8467	**0.85**
	40000	20000	0.1423	0.9477	0.4109	0.8576	**0.86**
	50000	**25000**	**0.1202**	**0.9562**	**0.4135**	**0.8621**	**0.86**

Table 4 Performance comparison according to the amount of learning data

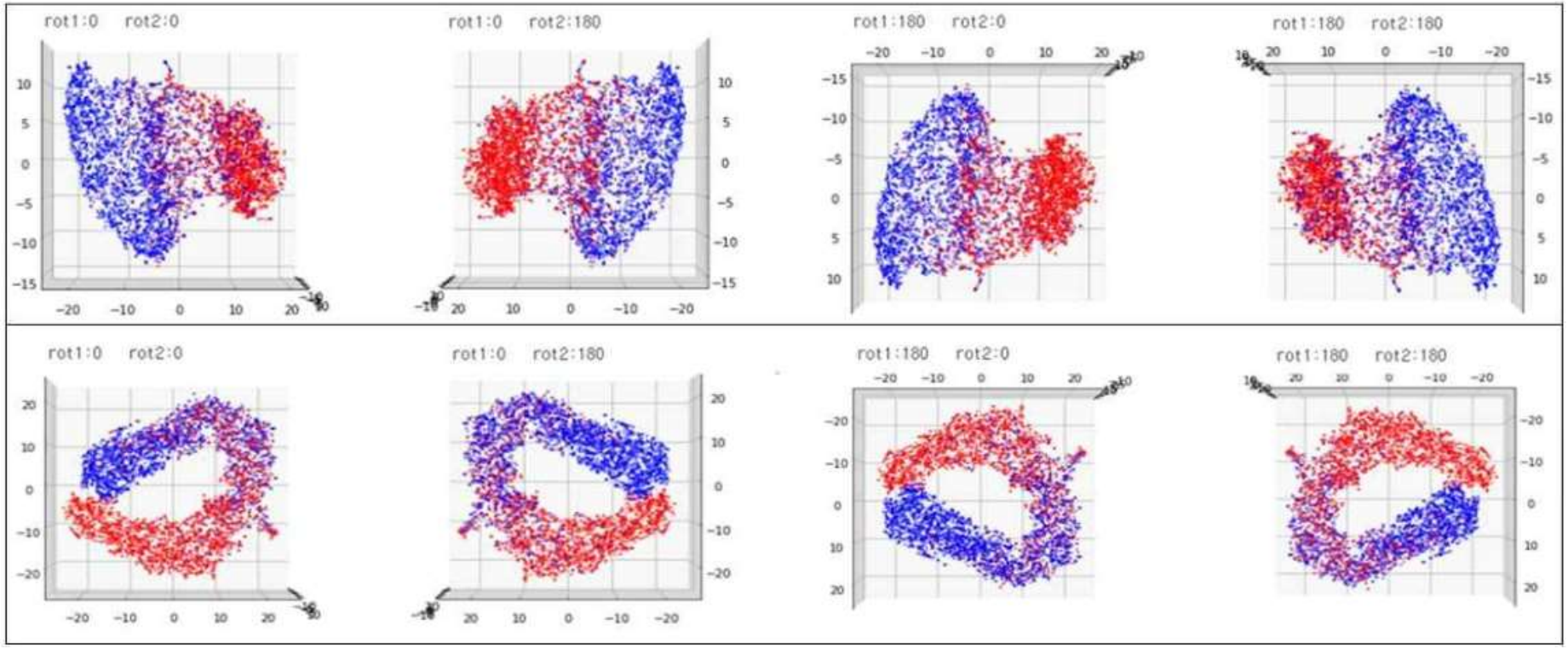
다른 데이터셋 학습 (기사 댓글)

document	label
나이 들어서 걸리면 치사 율도 높을껀데 전부 자기 들 자초 일이 의료 진 죽어나는거지	0
축하 합니다	1
저 것 들 한 곳 몰아 넣고 지 들 끼리 감염 되서 뭘 뜬 고함 다 디지길	0
몇 일 전 부터 우한 중국 입국 시키더니 우한 폐렴 확 진자 급증 이 거 아무래도 ...	0
왜 또 교회 몰려서 이런 난리 피우는 건지 집단 장소 가지말라고 그렇게 당부 메세지...	0
주님 만나러 일찍 가니 좋겠네요	1
뎨러 리스트 들	0
저승 길 친구 들 많이 있어 좋겠다 적당히 쫌해	0
미친 개 제 일 교회 확 진자 다 가라 하늘나라	0
빨리 죽 병상 확보 하자	0

Table 5 Dataset of news article comments

- **정확도 0.9656**
- **손실 0.0929**
- **F1 점수(부정) 0.95**
- **F1 점수(긍정) 0.31**

결과 분포 시각화


rot1:0 rot2:0
rot1:0 rot2:180
rot1:180 rot2:0
rot1:180 rot2:180
rot1:0 rot2:0
rot1:0 rot2:180
rot1:180 rot2:0
rot1:180 rot2:180

결론

- LLM 분야에서 추론을 위한 AI 프로세스 연구가 진행되고 있다.
- BERT 모델이 성능이 우수
- 본 논문에서 소개한 방법을 사용한다면 한국어 바탕 모델을 신속하게 테스트 가능
- 추후 발전 가능성이 무궁무진함