에브리타임 강의평 감성분석

Everytime Course Evaluation Sentimental Analysis

CUAI 5기 NLP 15팀
김민기(경영학부), 이강민(산업보안학과), 오창준(AI학과)

[요약] 본 연구는 강의평에 대한 감성분석을 목표로 진행되었다. 학습데이터로는 감성분석에 적합한 AI-hub 감성대화 말뭉치와 네이버 영화 리뷰를 사용하였다. 학습 모델로는 LSTM, Bi-LSTM, GPT2, KoBERT를 사용하였고, 4개의 모델 모두 학습 결과에서 큰 차이를 보이지는 않았다. 따라서 학습된 모든 모델에 강의평을 넣고 그 결과를 이용해 강의 점수를 산출하였다.

1. 서 론

매 학기가 시작되기 전 대학생들은 보다 좋은 강의를 수강하기 위해 흔히 과거에 작성된 강의평가를 참고한다. 그중 학생들이 가장 많이 이용하는 '에브리타임' 강의평가에서는 크게 별점과 텍스트 리뷰로 각 강의에 대한 평가가 기록되어 있다. 별점은 텍스트 리뷰에 비해 간단한 평가 지표로서 빠른 판단을 돕지만, 정보의 양이 현저히 부족하다는 단점이 존재한다. 따라서 단순히 별점만을 가지고 강의를 평가하는 것은 불가능하며 텍스트 리뷰를 참고하는 것이 필수적이다.

텍스트 리뷰는 별점에 비해 많은 양의 정보를 포함하고 있지만, 그 리뷰들을 모두 읽은 뒤 해당 강의의 전반적인 평가를 한 번에 판단하는 것은 쉽지 않다. 따라서 본 연구에서는 텍스트의 긍정 또는 부정 정도를 수치화하여 많은 양의 정보가 담긴 텍스트 리뷰를 빠르게 판단할 수 있는 모델을 제작하고자 한다.

2. 본 론

1) 데이터 선정 및 레이블링

감성분석을 위한 모델의 학습데이터로는 AI-hub의 감성대화 말뭉치[1]와 네이버 영화 리뷰[2] 데이터를 선정하였다.

AI-hub 데이터는 응답자의 연령, 성별, 감정_대분류, 감정_소분류, 상황 키워드로 구성된 6개의 특성(feature)을 가지고 있다. 그 중 감정_대분류를 기준으로 긍정과 부정 레이블링을 진행하였다.

감정_대분류는 기쁨, 불안, 당황, 슬픔, 분노, 상처 6종류의 감정들로 구성되어 있다. 기쁨을 긍정, 그 외의 감정들을 부정으로 분류한 결과, 데이터 불균형 문제가 발생하였다.

데이터 불균형 I

데이터	부정(0)	긍정(1)
AI-hub	68298	13460

데이터 불균형은 분류 문제의 성능을 저하시키는 요인으로 작용할 수 있다.[3] 따라서, 이와 같은 문제를 해결하기 위해 네이버 영화 리뷰 데이터(이하 NSMC)를 사용하였다. NSMC는 리뷰 문장과 해당 문장의 레이블로만 구성되어 있어 따로 레이블링은 진행하지 않았다. 결과적으로 AI-hub 데이터와 NSMC를 합쳐 학습데이터를 생성하는 것으로 데이터 불균형 문제를 해결하였다.

데이터 불균형 II

데이터	부정(0)	긍정(1)

AI-hub + NSMC	181760	111142

2) 데이터 전처리

AI-hub 데이터는 하나의 응답 데이터가 여러 문장으로 되어있는 경우가 많았다. 따라서 데이터에 2개 이상의 문장이 포함되어 있는 경우, nltk의 문장 토큰화 기능을 사용하여 문장을 분리하였다. 분리한 문장의 경우 본래의 데이터와 같은 레이블을 할당하였으며, 3개 이상의 문장으로 구성된 데이터는 레이블링 오류를 고려하여 학습에서 제외하였다.

NSMC의 경우 인터넷 리뷰의 특성상 'ㅋㅋㅋ', 'ㅠㅠㅠ'와 같은 반복적인 자음, 모음이 자주 포함되었다. 따라서 soynlp의 정규화 기능을 사용하여 자·모음의 반복으로 이루어진 단어들을 하나의 단어로 만들어주었다.

이렇게 각각 전처리된 AI-hub 데이터와 NSMC를 합친 뒤 KoNLPy의 Okt를 통해 형태소 분석을 진행하였다. 한국어의 특성상 중요한 의미가 들어있지 않은 접미사, 조사, 어미를 불용어로 처리하였고, 등장 빈도가 2회 이하인 단어들은 어휘 집합에 포함하지 않았다. 패딩 길이는 데이터의 95%를 포함하는 길이인 25로 설정하였다.

3) 사용한 딥러닝 모델

감성분석에는 일반 머신러닝 모델보다는 자연어 처리에 특화된 딥러닝 모델이 더 좋은 성능을 낼 것이라고 예상하여 총 4가지 딥러닝 모델을 사용하였다.

LSTM[4] 모델은 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장·단기 기억을 가능하게 설계한 신경망 구조이다. 또한 Bi-LSTM[5] 모델은 LSTM 구조에 역방향으로 정보를 처리하는 은닉층을 추가하여 성능을 보다 개선한 구조이다. LSTM과 Bi-LSTM은 시계열 데이터에 적합한 모델이지만, 자연어 처리에서도 좋은 성능을 보인다고 하여 선택하였다.

GPT2[6] 모델은 트랜스포머[7]의 디코더를 기반으로 하며, 2018년에 발표한 GPT1보다 레이어는 12개에서 48개로, 가중치는 117만 개에서 1,542만 개로 증가한 모델이다. 모델의 입력값도 BPE(Byte Pair Encoding)라는 방식을 사용해 높은 성능을 보이는 등 여러모로 기존의 GPT1보다 더 좋은 성능을 낼 것으로 기대되어 모델로 선정하였다.

BERT[8] 모델은 트랜스포머의 인코더를 사전 훈련시킨 모델이다. 양방향의 사전훈련을 통해 문맥의 오른쪽과 왼쪽의 맥락을 통합하고, 다운스트림 태스크에 대한 미세조정(fine-tuning)을 효과적으로 수행한다. 그럼에도 불구하고 Multilingual BERT는 영어 이외의 언어에서는 성능이 저하되며 모델의 파라미터 수가 많아 학습이 느리다는 단점을 가진다. 따라서 SKT에서는 BERT_base_multilingual_cased의 한국어 성능 한계를 극복하고자 5백만 개의 문장과 5천 4백만 개의 단어로 사전훈련된 KoBERT를 개발하였다.[9]

4) 딥러닝 모델 학습 결과

LSTM에서 출력된 혼동 행렬(confusion matrix)은 다음과 같다.

혼동 행렬 I

Confusion Matrix		*Predicted Label*	
		0	*1*
True Label	*0*	41447	3961
	1	5605	22143

이를 통해 계산한 정확도(accuracy), 정밀도(precision), 재현율(recall), 그리고 f1-score(이하 세부 지표)는 다음과 같다.

세부 지표 I

Accuracy		0.87
Precision	*부정(0)*	0.91
	긍정(1)	0.80
Recall	*부정(0)*	0.88
	긍정(1)	0.85
F1-score	*부정(0)*	0.90
	긍정(1)	0.82

Bi-LSTM에서 출력된 혼동 행렬은 다음과 같다.

혼동 행렬 II

Confusion	*Predicted Label*

Matrix		*0*	*1*
True Label	*0*	40924	4484
	1	5273	22475

이를 통해 계산한 세부지표는 아래와 같다.

세부 지표 Ⅱ

Accuracy		0.87
Precision	*부정(0)*	0.90
	긍정(1)	0.81
Recall	*부정(0)*	0.89
	긍정(1)	0.83
F1-score	*부정(0)*	0.89
	긍정(1)	0.82

GPT2에서의 혼동 행렬은 다음과 같다.

혼동 행렬 Ⅲ

Confusion Matrix		*Predicted Label*	
		0	*1*
True Label	*0*	41833	3651
	1	5137	22656

이를 통해 계산한 세부지표는 아래와 같다.

세부 지표 Ⅲ

Accuracy		0.88
Precision	*부정(0)*	0.92
	긍정(1)	0.82
Recall	*부정(0)*	0.89
	긍정(1)	0.86
F1-score	*부정(0)*	0.90
	긍정(1)	0.84

KoBERT에서의 혼동 행렬은 다음과 같다.

혼동 행렬 Ⅳ

Confusion Matrix		*Predicted Label*	
		0	*1*
True Label	*0*	43666	3799
	1	4182	24222

이를 통해 계산한 세부지표는 아래와 같다.

세부 지표 Ⅳ

Accuracy		0.89
Precision	*부정(0)*	0.91
	긍정(1)	0.86
Recall	*부정(0)*	0.92
	긍정(1)	0.85
F1-score	*부정(0)*	0.91
	긍정(1)	0.85

모든 성능 지표에서 긍정과 부정 데이터에 대한 유의미한 성능 차이를 확인할 수 있다.

3. 결 론

1) 강의평 적용 결과

특정 조건에 맞는 강의의 텍스트 리뷰를 가져오고 학습 데이터와 동일한 방식으로 전처리하였다. 그 후 각 학습된 모델의 단어 집합과 토크나이저를 사용해 최종 입력값을 만들었다. 이 입력값과 학습된 모델을 이용해 예측을 진행하였고, 그중 대표적인 두 텍스트 데이터를 모델에 넣은 결과를 보겠다.

다음은 긍정으로 보이는 실제 텍스트 리뷰인 '교수님이 친절하시고 학생에게 배려를 많이 줍니다'를 모델에 예측한 결과이다.

텍스트 리뷰 예측 Ⅰ

모델	예측
LSTM	긍정(1)
Bi-LSTM	긍정(1)
GPT2	긍정(1)
KoBERT	긍정(1)

이를 통해 모든 모델이 올바른 예측을 하였음을 알 수 있다.

다음은 또 다른 긍정으로 보이는 실제 텍스트 리뷰인 '수업 나긋나긋하게 해주시고 어렵지 않았어요'를 모델에 예측한 결과이다.

텍스트 리뷰 예측 Ⅱ

모델	예측
LSTM	부정(0)
Bi-LSTM	부정(0)
GPT2	긍정(1)
KoBERT	긍정(1)

이를 통해 LSTM 모델과 Bi_LSTM 모델은 잘못된 예측을 하였음을 알 수 있다.

2) 강의 점수 산출

특정 강의의 모든 데이터를 전처리한 뒤 학습된 모델에 넣어 결과가 각 리뷰마다 부정(0) 또는 긍정(1)이 나오게 만들었다. 그 후 전체 샘플 중 긍정 샘플의 비율을 백분율로 환산하여 점수를 산출하였다.

다음은 특정 강의의 강의 점수를 4가지 모델을 통해 산출한 결과이다.

강의 점수

모델	강의 점수
LSTM	66.7점
Bi-LSTM	67.9점
GPT2	73.5점
KoBERT	73.9점

4가지 모델을 통해 같은 강의에 대해 서로 다른 강의 점수가 산출된 것을 볼 수 있다.

3) 한계점 및 해결 방안

아래와 같이 전처리 과정에서 여러 개의 문장이 제대로 나눠지지 않은 경우에는 예측이 제대로 이루어지지 않았다.

「교수님 너무 좋으세요 수업도 쉬웠어요 조 모임은 힘들었고요ㅠㅋㅋ」

예측 결과 Ⅰ

모델	예측
LSTM	부정(0)
Bi-LSTM	부정(0)
GPT2	긍정(1)
KoBERT	긍정(1)

이를 해결하기 위해서는 성능이 더 좋은 문장 토크나이저를 찾거나 만들어야 한다.

또한 아래와 같이 사람이 봐도 애매한 문장을 긍정 또는 부정으로만 판단하다 보니 강의 점수에 악영향을 주었다.

「최종 개인과제에서 두 장 분량으로 피드백 받았습니다」

예측 결과 Ⅱ

모델	예측
LSTM	긍정(1)
Bi-LSTM	긍정(1)
GPT2	긍정(1)
KoBERT	긍정(1)

이를 해결하기 위해서는 모델의 예측 결과에 긍정도 부정도 아닌 중립 영역을 추가해야 한다. 우선 긍정과 부정의 각 확률로 결과가 나오도록 만든 후, 서로 간의 확률 차이가 크지 않으면 중립 문장이라고 판단하여 강의 점수에 반영하지 않는 쪽으로 설계해야 한다.

마지막으로 모델의 예측 과정에 대한 설명이 불가능하다는 한계가 존재한다. 문장에서 어떤 단어 혹은 구절이 예측에 영향을 미쳤는지 확인할 수 있다면 학생들에게 보다 상세한 정보를 제공할 수 있을 것이다. 이를 위해서는 단어 임베딩 벡터 간의 유사도를 수치화하여 점수 산출에 반영하는 것을 생각해볼 수 있다.

참고 문헌

[1] Ai-hub 데이터,

https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=86
[2] Huggingface datasets NSMC, https://huggingface.co/datasets/nsmc
[3] 강필성 외, '데이터 불균형 문제에서의 SVM 앙상블 기법의 적용', 한국정보과학회 가을 학술발표논문집 Vol.31, No.2, 2004
[4] Sepp Horchreiter & Jurgen Schmidhuber, "LONG SHORT-TERM MEMORY", NEURAL COMPUTATION, 1997
[5]Alex Graves & Jurgen Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM Networks", International Joint Conference on Neural Networks, 2005
[6] Alec Radford, et al. "Language Models are Unsupervised Multitask Learners", OpenAIblog, 2019
[7] Ashish Vaswani, et al. "Attention Is All You Need.", Google AI Language, 2017
[8] Jacob Devlin, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Google AI Language, 2018
[9] https://github.com/google-research/bert/blob/master/multilingual.md