대화 데이터의 감정 분석의 한계를 보완한 BERT 기반 감정 분석기 구성

명정연

고려대학교 컴퓨터정보통신대학원
audwjd@korea.ac.kr

Complementing the limitations of conversation data BERT-based sentiment analysis model

Jeongyeon Myeong
Korea University Graduate School of Computer & Information Technology

요 약

감정분석은 인간의 언어에서 감정을 읽어내는 기술로서 기업과 정치에서 수요가 급증하고 있다. 본 논문에서는 잘 알려진 텍스트 데이터인 네이버 영화 리뷰와 Friends를 이용하여 감정분석을 진행하는 모델을 구성하고자 했다. 사전 훈련된 embedding으로 모델의 성능을 향상시킬 수 있는 BERT를 이용하여 텍스트 데이터 분석을 진행하였다. 한 줄 리뷰에 대해서만 분석하는 한국어 데이터 셋과 달리 영어 데이터 셋은 대화로 구성되어 있으므로 화자의 정보 및 대화 안에서 문맥의 흐름이 중요하며 이를 반영하여 데이터 셋을 재구성하여 모델 학습을 진행했다. 본 논문에서는 기본적인 모델의 구성했을 때와 성능을 비교하였으며 문맥의 정보를 반영했을 때 데이터 셋이 적은 emotion에서 약간의 성능향상이 이루어졌다. 그 이외 화자의 정보 반영 미흡 및 데이터 셋과 tokenizer 구성에 대한 여러 한계점들이 존재했다. 이러한 점들이 보완된다면 챗봇 혹은 상품 리뷰에 대한 모니터링 서비스 등 다양한 산업분야에 확장되어 활용될 것으로 예상된다.

키워드 : 감정 분석, BERT, 화자의 정보, 문맥의 흐름, tokenizer 구성

1. 서론

감정분석은 인간의 언어에서 감정을 읽어내는 기술로서 기업과 정치 분야에서 수요가 급증하고 있으며 SNS 발달로 분석 데이터가 증가함에 따라 고객 리뷰 분석, 서비스 평가, 마케팅 모니터링 등 다양한 분야에서 사용되고 있다. 감정 분석의 성공 사례로서 대선 결과 예측과 같이 빅데이터에서 특정 후보에 대한 지지/반대 여론을 추론해 낼 수 있다[1]. 딥러닝의 발전과 단어 임베딩 기술로 만들어진 벡터들을 딥러닝 입력 값으로 활용하면서 점차적으로 감성분석 기술의 수준이 상승하고 있다. 텍스트 데이터만 있다면 자연어처리를 통해 화자의 감정을 파악하고 다양한 분야에 적용되어 활용할 수 있는 범위가 높아졌다.

본 논문에서는 잘 알려진 영어와 한국어 텍스트 데이터 셋을 이용하여 감정분석을 진행하는 모델을 구현하고자 했다. 영어는 Friends 데이터와 한국어는 네이버 영화 리뷰 데이터를 이용하여 화자의 감정을 분석하는 연구를 진행했다. 만들어진 감정 분석 모델은 실생활에서 여러 분야에 적용될 수 있다. 소비자가 구매 리뷰에 대한 긍정과 부정적인 평가를 모두 세세하게 분석하기 전에 감정 분석 결과로 종합적인 평가를 할 수 있다는 장점이 있다. 그리고 대화 속에서 화자의 의도를 파악하기 어려울 때, 감정 분석 결과를 기반으로 고객의 문의에 대응하는 챗봇 서비스 제공에 활용될 수 있다. 이와 같이 다양한 분야에 널리 활용될 수 있는 감정 분석 모델의 수요는 점차 증가될 것이며 활용 시 산업 현장에 긍정적인 영향을 미칠 것으로 예상된다.

2. 관련 연구

최근의 감정 인식 연구들은 학습 기반 방법에 초점이 되어있다[1]. 여러 모델 중에서 CNN, LSTM, Transfer모델이 있다.

문장 정보를 추출하는데 널리 사용되는 텍스트 분류 CNN 모델은 문장의 순서 정보와 자질정보를 잘 반영할 수 있지만 대화 내에서 상황에 맞는 감정 흐름의 인식이 부족하다는 단점[2] 있다. 그리고 Transformer보다 long range dependency의 낮은 정확도를 갖고 있으며, LSTM와 Transformer모델은 높은 정확도를 나타낸다[3].

Transformer 모델은 병렬로 attention연산이 수행되는 점과 layer마다 총 계산의 복잡성이 줄어들어 적은 수행시간이 소요되는 장점이 있다[4]. 본 연구에서는 기존 recurrent 모델들보다 Transformer 모델의 장점들을 이용하여 감정 분석하고자 한다.

3. 제안하는 방법

본 연구에서 사전 훈련된 embedding을 통해 과제의 성능을 향상시킬 수 있는 BERT모델을 이용하여 데이터 분석을 진행하였다. 이전의 토큰들에만 영향을 받는 단방향으로 처리하는 문제에서 벗어나 BERT는 양방향 transformer로 이루어진 모델이다. 자연어 추론 및 문장 간의 관계를 예측하는데 사전에 훈련된 언어 모델이라 자연어 처리의 향상에 효과적이다[5]. 이와 같은 장점으로 본 논문의 자연어 처리 모델로 이용하여 문장 안에서의 감정을 분석하였다. 데이터의 특성에 따라 BERT에 적용되는 사전의 훈련 모델이 나누어진다. 영어 데이터는 대규모 영어 데이터 코퍼스에 대해 사전 훈련된 모델인 bert-base-uncased을 사용했다. 한국어 데이터는 동일한 모델에 다국어 데이터의 대규모 코퍼스에 대해 사전에 학습된 모델 bert-base-multilingual-cased을 사용하여 모델에 적용하였다.

한 줄 리뷰에 대해 분석하는 한국어 데이터 셋과 달리 영어 데이터 셋은 대화의 특징으로 입력문장에서 문맥의 흐름과 화자의 특징을 파악할 필요가 있다. 먼저, 사람의 성격 및 성향에 따라 주로 느끼는 감정이 다르기 때문에 사람의 마다 감정의 패턴이 다를 것이라 예상되었다. 그리고 대화 안에서 화자의 감정이 앞의 대화와 지속적으로 이어져 나타나는 경향이 있으므로 대화 안에서 감정의 흐름 파악이 중요하다. 이 두 가지 요소를 중점적으로 고려하여 모델을 구성하였다.

4. 실험

4.1 데이터 셋

영어 데이터 셋은 Friends 데이터로 화자, 대화, 대화의 감정으로 구성되어 있으며 감정은 'joy', 'surprise', 'non-neutral', 'sadness', 'disgust', 'fear', 'anger', 'neutral' 8가지로 나타내었다. 데이터의 자세한 설명은 http://doraemon.iis.sinica.edu.tw/emotionlines/index.html에서 확인할 수 있으며 그림 1과 같은 형식으로 구성되어 있다. 한국어 데이터 셋은 네이버 영화의 리뷰에서 수집되어 문장, 영화리뷰에 대한 긍정/부정을 나타내었다. https://github.com/e9t/nsmc.git에서 ratings_train.txt와 ratings_train.txt 데이터 셋으로 모델 학습 및 테스트를 진행하였으며 데이터 셋은 그림 2과 같은 형식으로 구성되어 있다.

그림 1. 영어 데이터 셋 예제

	id	document	label
0	6270596	굳 ㅋ	1
1	9274899	GDNTOPCLASSINTHECLUB	0
2	8544678	뭐야 이 평점들은.... 나쁘진 않지만 10점 짜리는 더더욱 아니잖아	0
3	6825595	지루하지는 않은데 완전 막장임... 돈주고 보기에는....	0
4	6723715	3D만 아니었어도 별 다섯 개 줬을텐데.. 왜 3D로 나와서 제 심기를 불편하게 하죠??	0
5	7898805	음악이 주가 된, 최고의 음악영화	1

그림 2. 한국어 데이터 셋 형식

아래 표 1과 표 2는 훈련 데이터에서 emotion과 label의 분포에 대해 나타내었다.

표 1. 영어 데이터 셋

Emotion	*데이터의 수*
joy	1406
surprise	1371
non-neutral	2231
sadness	413
disgust	263
fear	214
anger	598
neutral	5243
총 합	11739

표 2. 한국어 데이터 셋

Label	*데이터의 수*
1(긍정)	74827
0(부정)	75173
총 합	150000

주요한 데이터 셋의 특징은 영어 데이터의 emotion 중 sadness, disgust, fear, anger은 데이터 수가 다른 emotion에 비해 상당히 적은 수준이며 총 데이터 수 또한 1만개 밖에 되지 않는다. 한국어 데이터의 경우 긍정과 부정 label의 데이터 수의 분포가 균등하며 총 데이터 수는 15만개로 상당히 많은 데이터를 갖고 있다. 1만개의 데이터로 학습을 진행하기에는 매우 적은 학습 데이터로 사전에 훈련된 모델인 BERT를 사용하는 것이 다른 모델보다 우수한 성능을 나타낼 수 있다.

4.2 실험 절차

본 논문에서 적용한 BERT모델의 가장 기본적인 방법은 전처리 단계에서 input embedding layer 생성 후 input 값을 BERT Model에 넣어 모델을 구성할 수 있다.

전처리 단계에서는 입력된 문장에서 [CLS] 토큰으로 시작하여 [SEP]로 문장을 분리할 부분을 넣어주며 맨 마지막에 [PAD]를 넣어 padding할 부분을 나타낸다. BERT Tokenizer로 토큰들을 넣은 문장을 tokenize한다면 ['[CLS] 정말 많이 울었던 영화입니다. [SEP] [PAD]']

Role	Utterance	Emotion
Rachel	Oh okay, I'll fix that to. What's her e-mail address?	Neutral
Ross	Rachel!	Anger
Rachel	All right, I promise. I'll fix this. I swear. I'll-I'll'll-I'll talk to her	Non-neutral
Ross	Okay!	Anger
Rachel	Okay!	Neutral

문장이 ['[CLS]', '정', '##말', '많이', '울', '##었던', '영화', '##입', '##니다', '.', '[SEP]', '[PAD]']로 나타날 수 있다. Tokenize 진행한 데이터에 padding을 적용시킨다. 그림 3과 같이 문장의 최대 길이 분포를 확인하여 Padding size는 영어 데이터는 200, 한국어 데이터는 160으로 설정하였다. BertTokenizer을 이용하여 숫자형태의 id로 바꾸는 작업을 진행한 후 BERT모델 적용한다. 추가적으로 attention mask을 추가하여 padding이 적용된 부분에 attention이 적용되지 않도록 했다.

영어 데이터의 경우 문장 맨 앞에 화자의 정보를 추가하여 텍스트 분석을 진행했다. 그리고 대화 안에서 화자의 감정 흐름이 일관되게 이어지도록 현재 문장의 화자 정보가 바로 앞 문장 혹은 전전 문장의 화자 정보와 동일한 경우, 현재 문장의 끝 부분에 이전 문장의 정보를 덧붙여서 학습했다. 두 개의 문장을 구분하기 위해서 [SEP]토큰을 이용하여 BERT모델 훈련 시 segment embedding에 값을 넣었다.

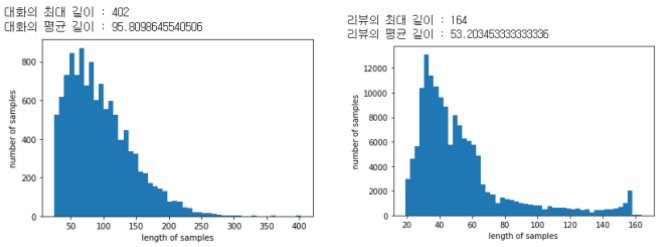

그림 3. 왼쪽 영어 데이터 오른쪽은 한국어 데이터

모델 훈련 시, 영어 데이터는 적은 수의 데이터 셋으로 batch size는 4, 8, 16으로 설정하여 성능이 높은 모델을 찾고자 하였고 한국어 데이터는 높은 성능으로 알려진 32개의 batch size로 설정하여 훈련하였다. 한국어 데이터는 label에 따라 데이터 수가 균등하지만 영어 데이터의 emotion이 sadness, disgust, fear, anger일때 다른 emotion에 비해 상대적으로 매우 적기 때문에 CrossEntropyLoss 손실함수에 weight를 부여하여 데이터 불균형 문제로 인한 발생할 수 있는 분류 문제를 해결하고자 했다.

4.3 실험 결과 및 실험 분석

BERT 모델로 학습을 진행한 결과 성능은 아래 표 3와 같다. batch size에 바꿔가며 훈련을 진행한 후 해당 모델의 테스트 정확도를 나타내는 F-score로 표현하였다.

표 3. 영어 및 한국어 테스트 데이터의 최종 모델 성능

	영어/ 4batch	**영어/ 8batch**	영어/ 16batch	**한국어/ 32batch**
Mean F-score	0.54	**0.577**	0.52	**0.84**

표를 보면 영어 데이터의 경우 bathch size가 8인경우 가장 높은 정확도를 나타내어 최종 모델로 선택되었다.
같은 모델을 이용하여 영어와 한국어 데이터를 학습했지만 영어 데이터가 한국어 데이터에 비해 상당히 낮은 정확도를 보인다. pre-trained된 모델을 이용해도 label의 범주가 한국어 데이터는 2개이지만 영어 데이터는 8개로 다양하며 데이터의 개수 또한 한국어 데이터의 1/15만큼 되는 데이터로 영어 데이터 학습이 이루어졌기 때문이다. 자세히 정확도를 분석해보면 아래 표 4와 같다. non-neutral, sadness, disgust, fear, anger 5개의 감정은 다른 emotion에 비해 precision과 recall이 상당히 낮다. 전체적인 모델의 성능을 낮게 만드는 요인이 될 수 있다.

표 4. 영어 데이터의 emotion의 precision과 recall (DEV Data)

Emotion	*joy*	*sur prise*	*non-neutral*	*sadness*
precision	0.75	0.70	0.22	0.47
recall	0.51	0.43	0.30	0.30
Emotion	*disgust*	*fear*	*anger*	*neutral*
precision	0.29	0.25	0.24	0.65
recall	0.13	0.11	0.54	0.81

아래는 표 5, 표 6, 표 7 은 화자의 정보만 추가했을 경우와 문맥의 흐름을 추가했을 경우, 어떤 정보도 반영하지 않은 경우의 recall 과 precision 의 값을 나타내었다. 결과를 비교하면 문맥의 정보와 화자의 정보를 반영해도 모델의 성능향상이 크지 않은 것으로 보인다. 화자의 정보만 반영했을 경우 다른 경우들보다 성능이 더욱 낮게 나타났다. 문맥의 흐름을 반영했을 경우가 fear 와 sadness, anger 의 recall 과 precision 이 약간 상승되었고 joy, surprise, non-neutral 의 precision 값이 상승되었음을 확인할 수 있다.

표 5. 어떤 정보도 반영하지 않았을 경우(DEV Data)

Emotion	*Joy*	*sur prise*	*non-neutral*	*sadness*
precision	0.75	0.64	0.18	0.45
recall	0.52	0.48	0.34	0.24
Emotion	*disgust*	*fear*	*anger*	*neutral*
precision	0.30	0.15	0.32	0.70
recall	0.13	0.07	0.40	0.78

표 6. 문맥 흐름만 반영한 경우(DEV Data)

Emotion	*joy*	*sur prise*	*non-neutral*	*sadness*
precision	0.81	0.65	0.22	0.51
recall	0.46	0.47	0.30	0.33
Emotion	*disgust*	*fear*	*anger*	*neutral*
precision	0.27	0.12	0.44	0.62
recall	0.22	0.083	0.37	0.81

표 7. 화자의 정보만 반영한 경우(DEV Data)

Emotion	*joy*	*sur prise*	*non-neutral*	*sadness*
precision	0.63	0.66	0.22	0.36
recall	0.60	0.41	0.30	0.19
Emotion	*disgust*	*fear*	*anger*	*neutral*

precision	0.26	0.18	0.39	0.67
recall	0.23	0.06	0.38	0.79

4.4 보완방향 제안

학습 결과 영어 데이터 모델의 성능이 낮은 부분에 대한 아쉬움이 크다. 본 논문에서는 모델의 성능을 높이기 위해 영어 데이터에서 상대적으로 적은 수의 emotion에 대한 데이터 셋을 추가 및 emotion 재정의, 한국어 데이터에 적용하는 tokenizer을 보완 총 2가지로 나타낼 수 있다.

첫 번째는 영어데이터의 경우 emotion을 이해하고 분류하는 것이 명확하지 않으며 특정 emotion에 해당하는 데이터 수가 적어 모델의 정확도가 낮다. non-neutral은 다른 감정에 비해 어떤 감정인지 직관적으로 이해하기 어렵다. 데이터 수가 많지만 precision과 recall이 낮은 원인은 감정을 이해하기 어렵기 때문에 분류하는데 정확도가 떨어지는 결과가 나타난 것으로 판단된다. 또한 실제 상황에서 sadness, disgust, fear, anger 이 4가지의 감정이 혼합하여 느끼는 경우가 많다. 예를 슬프고 화가 나는 상황이나 두렵기도 하면서 슬픈 상황이 있다. 복합적인 감정을 세분화하여 구분하면서 분류의 오류가 발생할 수 있는 가능성은 높아진다. 그리고 sadness, disgust, fear, anger의 데이터 수 또한 매우 적다. 해당하는 emotion에 대해 더 많은 데이터를 수집하고 해당 감정을 명확하게 구분 지을 수 있는 emotion을 재정의하여 감정 인식에 대한 분류 정확도를 높일 수 있기를 제안한다.

두 번째는 tokenizer의 사용이다. 영어 데이터에서 tokenize 진행 시, ['My duties? All right.']에서 ['[CLS]', 'my', 'duties', '?', 'all', 'right', '.', '[SEP]', '[PAD]']로 나타내어 단어 단위로 토큰화가 잘 되어 있음을 확인할 수 있다. 하지만 동일한 모델을 한국어 데이터에 적용 시 ['포스터만 보고 기대했다가 실패했어.']라는 한국어 문장에 대해서는 ['[CLS]', '포', '##스터', '##만', '보고', '기', '##대', '##했다', '##가', '실', '##패', '##했', '##어', '.', '[SEP]', '[PAD]']로 tokenize가 된다. 단어를 제대로 인식하여 적절한 의미 단위의 Tokenize가 진행되지 않았다는 것을 확인할 수 있다. 새롭게 한국어 기반으로 단어를 정의한 tokenizer을 이용하여 적절한 의미 단위로 tokenize가 진행되어 성능을 높일 수 있는 시도가 필요하다.

5. 결론

본 논문에서는 사전 훈련된 embedding을 통해 성능을 향상시키는 BERT의 기본적인 감정 분석 모델에서 벗어나 대화로 구성된 Friends데이터에 대해 향상된 성능을 갖는 모델을 구성하고자 했다. 대화의 흐름을 반영하고자 동일한 화자에 대해 이전의 문장 정보를 추가하고 문장의 앞 부분에 화자의 정보를 반영하여 모델을 구성하였다.

모델 분석 결과, 한국어 데이터의 경우 성능은 84%로 높은 편이었지만 영어 데이터의 모델 성능은 57%로 낮게 나타내었다. 원인으로는 한국어 데이터의 범주는 2개이지만 영어 데이터는 8개로 다양하여 emotion을 명확하게 분류하기 어려운 상황이 발생할 수 있다. 데이터의 개수 또한 한국어 데이터의 1/15만큼 되는 영어 데이터로 학습이 이루어졌다. 화자의 정보를 대화 내용 안에 넣어 감정분석을 진행했지만 오히려 기존 모델보다 성능이 떨어지는 문제점이 발생하였으며 문맥의 흐름을 반영했을 경우에는 몇 개의 emotion에 대해 약간의 성능향상이 보여졌다.

감정 분석 모델을 구성한 결과 위와 같은 한계점들이 나타나며 영어 데이터에서 상대적으로 적은 emotion에 대해 데이터 추가 및 emotion 재정의, 한국어 데이터에 적용하는 tokenizer을 보완이 필요하다. 추가적으로 화자의 정보를 효과적으로 반영하여 감정을 예측하는 성능을 높일 수 있는 방법에 대한 연구가 진행되어야 할 것이다. 향후, 모델에 대한 보완이 이루어진다면 리뷰 및 대화 분석에도 강점을 갖는 자연어 처리 감정분석 모델로서 여러 산업 분야에서 신속하게 고객 서비스를 제공하는 모델로 활용될 것으로 예상된다.

참고문헌

[1] Dong-a business Review, SR2. 감성분석 잠재력과 한계, 261호 (2018년 11월 issue 2), https://dbr.donga.com/article/view/1101/article_no/8892/ac/a_view

[2] Chao-Chun Hsu. EmotionLines: An Emotion Corpus of Multi-Party Conversations. arXiv:1802.08379, (2018).

[3] Gongbo Tang. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. arXiv:1808.08946, (2018).

[4] Ashish Vaswani. Attention Is All You Need. arXiv:1706.03762, (2017).

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).