BERT 를 활용한 문장 감정 분석 연구

이한범, 구자환, 김응모
성균관대학교 소프트웨어대학
seaman95@skku.edu, jhkoo@skku.edu, ukim@skku.edu

A Study on Emotion Analysis on Sentence using BERT

Hanbum Lee, Jahwan Koo, Ung-Mo Kim
College of Software,
Sungkyunkwan University

요 약

소셜 네트워크 서비스 등의 발전으로 인해 개인이 다수에게 의견을 표출하는 통로가 활성화되었다. 게시물에 드러난 감정을 통해 특정 주제에 대한 여론을 도출할 수 있다. 본 논문에서는 BERT 를 통한 문장 분석 기술, 그 중에서도 감정 분석을 하는 방법을 분석하고, 이를 일반화된 문장에 적용시키기 위한 데이터 셋 구성에 관하여 연구를 진행하였다.

1. 서론

소셜 네트워크 서비스의 사용자수가 증가하고 다양한 형태의 플랫폼에서 자신의 의견을 표출하는 사용자의 게시물이 꾸준히 게시되고 있다. 사용자가 게시하는 내용 중엔 현재 이슈가 되는 사건, 인물 등에 대한 의견을 드러내는 게시물이 존재하고, 이를 분석한다면 해당 사건, 인물에 대한 대중의 의견을 알아낼 수 있다. 정책의 찬반여부부터 대권 후보의 지지율까지의 사회의 주요 쟁점이 되는 시민의견을 현재의 설문조사 형태의 여론조사에서 벗어나 보다 큰 범위의 표본을 사용하여 더 정확한 결과를 얻어낼 수 있다.

이러한 평문 데이터를 통해 여론을 얻어내려면 해당 게시물의 문장에서 의견을 도출해 낼 수 있어야 한다. 현재 자연어처리 기술 중 하나인 BERT[1]를 활용하여 문장을 분석하고 의견을 도출해 낼 수 있다. 감정이 라벨링된 문장으로 학습시킨 언어모델에 문장을 입력하면 해당 문장에 맞는 라벨을 제시하게 설계한다. 성공적인 모델을 구축하려면 감정이 라벨링된 문장, 데이터 구축이 무엇보다 중요하다고 할 수 있다.

데이터 구축을 할 때 고려되어야 할 것은 어떤 데이터를 선정해서 학습시켜야 이후에 성능이 높게 나오는지 파악하는 것이다. 이를 위해 특정 분야에서 학습을 시키는 것 보다 다양한 분야의 데이터를 섞어서 학습을 시키면 더 높은 성능을 얻을 수 있을 것이라 가설을 세웠고 이를 확인하는 과정을 거쳤다.

이번 연구의 목적은 다양한 source 의 데이터를 사용했을 경우 특정 분야의 데이터를 사용했을 경우보다 높은 성능을 가지는지 여부를 확인하는 것이다. 이를 통해 감정 분석을 위한 모델을 설계할 때 더 성공적으로 데이터를 수집하여 학습시킬 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구로서 자연어 분석 기술의 동향 및 BERT 임베딩 및 문장 감정 분석 방법에 대해 기술한다. 3 장에서는 기존 모델의 한계와 이를 극복하기 위한 데이터 구성 연구 설계에 대해 기술한다. 4 장에서는 연구의 결과와 보완할 점, 향후 연구 과제에 대해 기술한다.

2. 관련 연구

2-1) BERT[1]

자연어 문장을 기계가 이해하기 위해선 문장 분석을 수행해야 한다. BERT 는 “Attention is all you need(Vaswani et al., 2017)” [2]논문에서 소개한 Transformer 구조에서 language representation 을 담당하는 부분이다. BERT 는 이러한 문장 분석을 수행하기 위해 wiki 등에서 발췌한 unlabeled data 를 대규모로 학습하고, 특정 task 에서 역할을 수행하기 위해 task-specific 한 fine-tuning 을 진행한다.

기존 연구는 좌측에서 우측, 우측에서 좌측 등의 단방향의 분석을 수행하여 문맥을 파악한다. 또한 task 에 따라 모델 설계를 다시 해야 한다는 단점이 존재한다. 이에 반해 BERT 는 양방향의 분석을 수행하며 task 가 바뀌어도 parameter 들을 조금씩 바꿔 task 를 수행하는 fine-tuning 의 방식을 진행한다.

BERT 는 단방향이 아닌 양방향으로 분석을 진행하여 문장 벡터화를 보다 효율적으로 수행한다. 양방향 학습을 위해 BERT 는 두 가지 방법을 사용하여 Training 을 진행한다.

Masked LM: 문장에서 일부(15%)의 토큰에 masking 을 하여 해당 토큰을 predict 하게 한다. 이 방법을 통

해 기존 단방향 분석이 아닌 양방향 분석을 할 수 있다.

Next Sentence Prediction: 문장 사이의 관계를 추론한다. 두 문장을 주고 두 번째 문장이 첫 번째 문장 바로 다음에 오는지 여부를 확인한다.

여기서 문장 감정 분석을 수행하기 위해 주목해야 할 항목은 Masked LM 이다. 일부 토큰을 마스킹해 해당 토큰을 추론하는 과정을 거치며 문맥을 파악하며 학습을 하게 된다.

이러한 BERT 모델은 실제 영어, 한국어 데이터에서 효과적으로 작동하여 SQuAD 와 KorQuAD 질문답변 score 에서도 상위 순위자의 BERT 응용 사례를 찾아볼 수 있다[3][4].

2-2) SKT KoBERT[5]

BERT 에는 영어 뿐만 아니라 다양한 언어를 학습시킨 외국어 모델이 존재한다(base multilingual cased model). 이 모델보다 정확도를 높인 모델을 만들기 위해 SKTBrain 팀이 연구한 KoBERT 는 자체 제작한 한국어 테스트셋에서 기존 base_multilingual_cased 모델의 정확도인 0.875 보다 높은 0.901 을 기록했다[5]. 한국어 위키 문장 500 만개, 단어 5400 만 개를 사용했고, 한국어 뉴스 문장 2000 만개, 2 억 7000 만 단어를 사용해 training 을 진행하였다. 이 모델의 경우 새로운 알고리즘을 만들었다기 보다는 BERT 알고리즘을 활용하여 데이터와 parameter 를 한국어 문장에 맞게 설정해 정확도를 높인 모델이다.

2-3) ETRI KorBERT[6]

ETRI 에서는 BERT 의 원리를 이용하여 새로운 모델을 설계하였다. BERT 외국어 모델에서는 문장을 음절 단위로 토큰화시켜 이 토큰을 벡터화하여 문장을 인식한다. 영어에서는 글자 그대로 잘라 토큰화를 시켜도 문장이 정상적으로 나누어 질 수 있으나 한국어에선 글자대로 잘라도 정확한 단위로 나누어지지 않는다. 한국어는 영어와 달리 명사 또는 동사가 조사 또는 어미와 결합하여 어절을 구성하고 이 어절이 모여 문장이 된다[6]. ETRI 에서는 이런 한국어의 형태소라는 특성을 이용하여 형태소 단위로 문장을 나누어 토큰화하였다.

KorBERT 모델은 백과사전 및 신문기사 데이터 23GB 를 사용하여 47 억개의 형태소를 학습하였다. KorBERT 모델은 기존 BERT 외국어 모델과 상대적 평가를 위해 다섯 개의 task 를 통해 실험을 진행하였다. 해당 실험 결과 5 개의 task 에서 KorBERT 형태소 모델은 기존 BERT 외국어 모델에 비해 평균 4% 높은 정확도를 기록하였다[6].

2-4) 한국어 문장 감성 분석

BERT 를 사용하기 이전에 진행됐던 문장 감정분석 연구는 문장에서 seed 표현을 포착해 대입하는 형식이다[7]. 이 연구에서는 감정분석을 위해 감정분석 corpus 를 제시하였다. 해당 연구에서 제시한 감정 분류 기준은 다섯 가지로 분류된다.

2-5) Naver 영화 리뷰 분석[8]

BERT 가 발표되고 문장 단위로 감정을 학습하여 모델을 설계하는 방식이 연구되었다. 해당 연구에서는 네이버 영화 리뷰 데이터를 활용하여 BERT 를 학습시켜 문장의 감정을 분석한다. 리뷰에는 별점과 문장이 모두 포함되어 별점이 높은 문장은 긍정적, 별점이 낮은 문장은 부정적 평가임을 알 수 있다. 해당 모델에서는 1~10 점의 평가 중 8~10 점의 점수를 가진 리뷰를 긍정, 1~4 점의 점수를 가진 리뷰를 부정 데이터로 학습시켰다. 긍정데이터는 tag 를 1 로, 부정데이터는 tag 를 0 으로 주고 리뷰문장과 짝을 지어 fine-tuning 을 하였다.

해당 모델은 BERT 외국어 모델인 base_multilingual_cased 모델을 통해 pre-train 된 모델을 얻고, 리뷰데이터를 통해 fine-tuning 을 진행하여 영화 리뷰를 입력받으면 해당 리뷰의 긍정/부정 여부를 판단하는 모델로 설계되었다.

3. 연구 설계

이전 영화 리뷰데이터로 학습한 모델은 영화 리뷰를 입력받아 해당 리뷰가 긍정적인 평가를 하는지, 부정적인 평가를 하는지 판단한다. 하지만 이를 일반적인 문장에 활용해 해당 문장의 긍정/부정 여부를 판단하려면 영화 리뷰데이터만 가지고는 정확도가 낮을 가능성이 있다. SNS 에 게시된 문장의 감정을 분석하여 여론을 알아볼 때 영화 리뷰와 전혀 관계가 없는, 또는 리뷰에서 쓰인 표현의 의미가 일상 표현에서의 의미와 다른 경우 의미 있는 조사가 이루어지지 못할 가능성이 있다.

따라서 본 연구에서는 일상용어를 포함한 데이터를 통해 모델을 구축하고 해당 모델이 기존의 영화리뷰 데이터로 이루어진 모델보다 높은 정확도를 가지는지 여부를 관찰하여 모델 구축 시 데이터 수집의 방향을 찾도록 한다.

일상데이터는 AI HUB 의 '한국어 감정 정보가 포함된 단발성 대화 데이터셋'[9]을 사용하였다. 해당 데이터는 기쁨, 슬픔, 놀람, 분노, 혐오, 중립의 7 개 감정이 라벨링된 28594 개의 문장으로 이루어져 있다. 이를 라벨링에 따라 묶어 긍정, 부정으로 분류하고 중립적인 데이터를 제외하여 2 만 7 천개의 데이터를 생성하였다. 해당 데이터와 이전의 영화 리뷰데이터를 합쳐 데이터 셋을 구성하였다.

챗봇 데이터를 가공하기 위해 라벨링 된 챗봇데이터를 분류하였다. 중립 데이터를 제거하고 부정을 0, 긍정을 1 의 값으로 바꾸고 가공된 데이터를 기존의 영화 리뷰 데이터셋과 합쳐 혼합 데이터셋을 만들어 이를 train 데이터와 test 데이터로 나누었다.

가공된 데이터를 학습시키기 위해 BERT 모델을 구축하기 위해 colab 을 사용해 진행하였으며 TPU 를 사용하여 연산 시간을 감소시켰다. BERT 의 외국어 모델인 base_multilingual_cased 모델을 설치하여 pretrain 까지의 과정을 진행하였다.

데이터셋을 나누어 90%의 데이터를 학습데이터로, 10%의 데이터를 테스트 데이터로 선정하여 학습을

진행한다. 이때 10%의 테스트 데이터는 학습에 포함되지 않도록 하였다.

<표 1> model1 영화 리뷰 test 결과

	precision	recall	f1-score
부정	0.86	0.88	0.87
긍정	0.88	0.86	0.87
accuracy	0.87	0.87	0.87

precision = 실제 true/true 판정, recall = true 판정/전체 true

<표 1>은 기존 영화리뷰 데이터로 구성한 모델(model1)이다.학습을 모두 영화리뷰 데이터로 하였고 테스트도 모두 영화리뷰 데이터로 하였을 때의 결과이다.

<표 2> model1 (기존 영화리뷰 15 만 데이터)

	precision	recall	f1-score
부정	0.79	0.91	0.85
긍정	0.92	0.81	0.86
accuracy	0.85	0.86	0.85

<표 2>는 위와 동일한 모델로 테스트 데이터만 일상데이터를 추가한 결과이다. 기존 영화리뷰 테스트 데이터 5 만 문장에 일상 대화 테스트데이터 2 천개를 추가하였다. 정확도가 조금씩 하락한 결과를 확인할 수 있다.

<표 3> model2 (영화리뷰 15 만+ 일상 대화 2.6 만)

	precision	recall	f1-score
부정	0.89	0.83	0.86
긍정	0.85	0.90	0.87
accuracy	0.87	0.87	0.87

영화리뷰 데이터와 일상 대화 데이터 모두를 사용하여 학습을 진행한 모델(model2)이다. 테스트는 <표 2>의 테스트 데이터와 동일하게 영화데이터와 일상데이터를 합쳐서 진행한 결과이다. 정확도가 상승한 모습을 확인할 수 있다.

기존의 영화 데이터 만으로 일상 데이터가 섞인 테스트에서 85%이상의 높은 정확도를 가진 모델을 만들 수 있었다. 하지만 2.6 만개의 일상 대화 데이터를 추가하여 학습을 시켜 더 높은 정확도를 가진 모델을 만들 수 있었다. 조금의 정확도를 높이기 위하여 더 많은 데이터를 생성해 학습을 시켜야 하는 인공지능 분야에서는 더 효율적인 데이터 셋이 필요하다. 한 가지 분야의 데이터로 학습하는 것 보다 여러 분야의 데이터를 골고루 모아 학습한다면 동일한 데이터 수를 가지고 학습했더라도 여러 분야가 섞여있는 일반적인 사용에서 더 높은 정확도를 가진 모델을 만들 수 있을 것이다.

4. 결론

한 분야의 데이터만 집중적으로 학습하여 높은 정확도를 얻을 수 있었지만 일부라도 다른 분야의 데이터를 섞는다면 정확도가 상승한다는 것을 확인할 수 있었다.

물론 짧은 길이의 텍스트만으로 일상 대화의 감정을 완벽히 분석해 내기엔 한계가 있었다. 보다 많은 데이터를 수집하여 더 다양한 분야의 데이터를 통해 충분한 학습을 시킨다면 더 높은 정확도를 얻어낼 수 있을 것이다. 다양한 분야의 학습을 위해 감정 라벨링이 된 일상 대화 데이터 셋이 요구된다. 추가적인 데이터를 수집한다면 더 높은 정확도를 가진 모델을 구축할 수 있을 것이다.

데이터의 종류도 추가적으로 늘릴 수 있을 것이다. 영화리뷰 뿐만 아니라 앱스토어, 배달 음식 앱, 쇼핑몰 리뷰 등에서도 별점과 리뷰 데이터를 가지고 있다. 해당 데이터에서도 추가적인 데이터 셋을 생성할 수 있을 것이다.

긍정/부정으로 두 개로 나뉘는 감정평가에서 발전하여 기쁨/슬픔/분노 등 다양한 데이터를 BERT 를 이용하여 도출해 낼 수 있는 모델도 설계될 것이라 예상하고 실제로 서론에서 제시한 SNS 등을 통한 매체로 대중의 여론을 수집하는 모델도 설계될 것이라 기대한다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A1B07049464).

참고문헌

[1]Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

[2]Ashish Vaswani “Attention is all you need”, arXiv

[3]“KorQuAD2.0 The Korean Question Answering Dataset” https://korquad.github.io

[4]“SQuAD2.0 The Stanford Question Answering Dataset.”, https://rajpurkar.github.io/SQuAD-explorer/

[5]Korean BERT pre-trained cased (KoBERT), “https://github.com/SKTBrain/KoBERT”

[6] 임준호, 김현기, 김영길, “딥러닝 사전학습 언어 모델 기술 동향”,한국 정보통신연구원(2020)

[7]김문형, 장하연, 조유미, 시효필, “KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스”, 한국정보과학회(2013)

[8]”Naver sentiment movie corpus v1.0”, https://github.com/e9t/nsmc

[9]AI HUB Open 데이터, https://aihub.co.kr/keti_data_board/language_intelligence