|  |
| --- |
| **BERT 인코더를 통한 영화 리뷰에서의 감성 인식** |
| **(Sentiment Analysis On Movie Reviews Using BERT)** |
| 2014104136 이형진 (경희대학교 컴퓨터공학부) |
| **요 약**  영화 리뷰 데이터를 Word2Vec을 사용해 벡터공간에 표현하고, CNN 등의 뉴럴 네트워크 모델에 학습시키는 것에 더 나아가 구글에서 개발한 bidirectional한 Pre-trained BERT 모델에 Fine Tuning을 함으로써 더욱 정확한 감성 분류기를 만드는 것을 목표로 한다. | | |

1. **서 론**

빅 데이터 내의 감정 정보를 추출하여 사용자들이 특정 대상에 대해 갖고 있는 인식을 파악하고자 하는 노력이 활발히 이루어지고 있다.

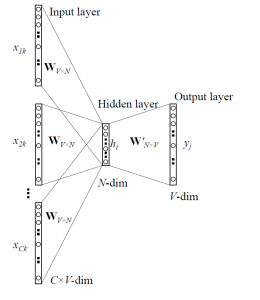
이를 감성 분석(sentiment analysis)이라 하며, 분석을 위한 첫 단계인 데이터 수집이 가장 중요하다고 할 수 있다. 인터넷 상에 있는 방대한 양의 데이터를 기반으로 그 속에서 positive, negative, neutral 등의 극성을 탐지해 내려면 수치가 필요하다. 따라서 수치와 plain text의 정보를 포함하고 있는 데이터가 최적의 형태라고 판단하였고, 이에 맞는 데이터 셋을 검색하기 시작하였다.

수치와 평서문을 지닌 적절한 데이터 형식은 ‘영화리뷰’ 이다. 웹크롤링을 통하여 손쉽게 데이터를 수집할 수 있고, 학습 과정에서 이를 패턴화 시키거나 극성을 분석하기도 어렵지 않다는 점이 장점이라 생각해서 연구 방향을 결정하게 되었다. 연구 결과가 상용화된다면 학습된 모델을 바탕으로 수치가 표기되지 않은 유튜브나 인스타그램 댓글에서 실시간으로 감성추출을 함으로써 기업의 마케팅 전략으로 사용될 것으로 기대된다.

1. **기존 연구**
   1. **감성사전**

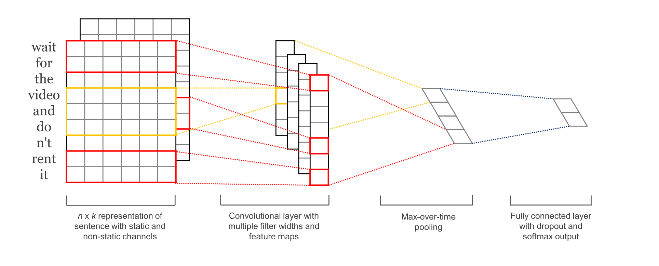
감정분석에 대한 연구결과는 여러 분야에서 사용되고 있는데, 대다수가 감정에 대한 사전을 구축하고 그 사전을 기반으로 점수를 계산하는 방법을 취하고 있다. 프린스턴대학 연구소에서는 WordNet이라는 단어사전을 만들었고, 논문에서는 그 중 감정에 SNS를 분석하여 SentiWordNet이라는 사전을 구축하여 활용하고 있다 기존 연구 1에서는 SentiWordNet과 비교하여 한글화 된 사전을 구축하고 감정 사전을 위한 데이터베이스를 활용하여 작성된 글을 점수화 하는 시스템을 제안한다.

감정 사전의 경우, 1차원적으로 그 단어가 가진 의미만을 판단하여 감정을 인식한다. 이는 중의적, 다중적 의미를 가진 단어나 역설적, 반어적 의미를 지닌 문장이 있을 때에 대해서는 제대로 인식이 되지 못한다. 예를 들어, ‘bank’라는 단어는 때에 따라서 ‘은행’이라는 뜻으로 쓰일 때도 있고, ‘둑방’ 이라는 뜻으로 쓰일 때도 있다. SentiWordNet은 정확성이 떨어지는 문제도 있다. 예를 들어, 동사 ‘blame’은 부정적인 견해를 드러내는 단어임이 틀림없어 보이는데도 SentiWordNet에서는 이들의 긍정/부정 값이 모두 0으로 기록되어 있다. 또 다른 예로, 해부학 용어로 서 중립적으로 여겨지는 ‘pia mater(연뇌막)’ 라는 어휘는 부정값은 없이 긍정값만 1.0만점에 0.5에 이른다. 또한 SentiWordNet의 경우 영어 단어를 데이터로 갖는데, 한글로 1대1 대응되지 못한다는 단점이 있다. 또한 한글 데이터를 수집한 경우 영어에는 없는 형태소라는 문법적 개념이 반영되지 않는다는 단점이 있다.

* 1. **Word2Vec**

구글에서 개발한 워드 임베딩 라이브러리이다. 워드 임베딩이란 문서 안의 단어들을 벡터 공간에 표현하는 기법이다. 이를 구현하려면 각 단어마다 특징 값이 필요하며 이 값을 추출하기 위해 여러 방법을 거친다. 가장 단순한 방법이 각 단어들 사이의 코사인 유사도를 구해서 이 값들을 feature로 보관하는 것이다. 또한 문서 단어 행렬이나 TF-IDF 행렬을 만들어서 문서의 유사도를 구하는 방법도 있다. TF(Term Frequency)는 문서 안에 단어가 등장하는 횟수이고, IDF(Inverse Document Frequency)는 단어 자체가 전체 문서 군 내에서 사용되는 빈도인 DF 값에 역수를 취한 값이다. 코사인 유사도 측정 방법과 TF-IDF방법은 단순히 빈도 수나 거리 값으로만 유사도를 측정해서 임베딩 한다는 점에서 오차가 존재한다. 또한, Word2Vec는 저 차원 벡터공간에 임베딩된 단어 벡터 사이의 유사도를 측정하는 데에는 어느 정도 좋은 성능을 띄지만, 사용자가 지정한 윈도우 내에서만 분석이 이루어지기 때문에 말뭉치 전체의 co-occurrence는 반영하기 어렵다.

1. **시스템 모델**
   1. **기존 연구와 차이점 및 해결방안**

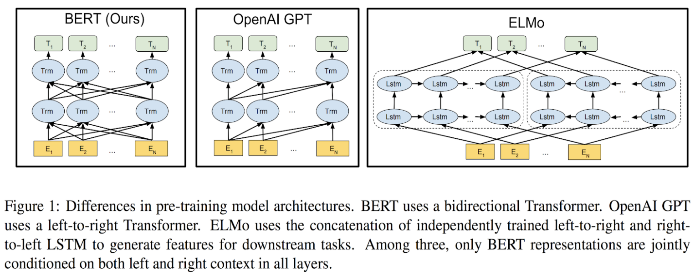
****

기존의 연구는 다음의 3가지 과정으로 진행되었다.

1. 정답 레이블이 있는 'Rotten Tomato Movie Review' 데이터 17만 문장에 대해서 데이터 전 처리

2. 전체 말 뭉치에 있는 단어를 정규화 시켜 벡터로 표현(표현 과정에서는 word2vec를 사용)

3. 표현된 단어들과 정답 레이블(0 : negative, 1 : positive) 을 CNN에 넣어서 Fully connected layer와 Softmax함수를 이용해 분류

기존 연구에서는 데이터 전처리 과정에서 감성을 포함하지 않는다고 판단되는 단어는 제외하여 CNN 모델에 학습시켰다. 임의의 영화 리뷰 문장 하나에 있어서는 91.4%로 성능이 좋았으나 말뭉치를 대상으로는 67.9%로 정확도가 다소 낮았다. 이는 학습 데이터의 양이 부정으로 치우쳐져 있어서 일어난 현상이며 데이터 셋을 균등하게 맞추면 보다 더 정확도가 상승할 것으로 보고 있다. 

본 연구에서는 BERT 기술을 활용하여 입력값을 의도적으로 왜곡함으로써 이런 문제를 해결하는 데에 착안한다. 일반적으로 Neural Network에서 hidden layer의 수가 input parameter의 수 보다 많아질 경우 출력값이 입력값과 같아질 가능성이 높다. 이 현상은 모델에서 부정확한 결과값을 야기할 수 있다. 이 문제를 해결하기 위해서 BERT는 일반적인 뉴럴 네트워크의 표현 방식과 다른 두 가지 특징을 가지고 있다.

첫 번째로, Masked LM이라는 기법을 사용한다. 즉 입력 단어 배열에서 무작위적으로 단어를 선택해서 80%의 확률로 [MASK]로 치환하거나 10%의 확률로 임의의 단어로 치환하거나 나머지 10%의 확률로 바꾸지 않고 그대로 사용하게 된다. BERT 논문에서는 마스킹을 하여도 결과값에는 큰 영향을 미치지 않는다고 언급하고 있다. 이는 연구를 통하여 확인하도록 하겠다.

두 번째로, Bidirectional한 모델을 사용한다. 논문에서 발췌한 위 Figure 1에서는 BERT, OpenAI GPT, ELMo 각각에서 사용된 Pre-training 모델 아키텍처를 도식화해서 보여주고 있다. OpenAI GPT에서는 left-to-right Transformer를, ELMo에서는 독립적인 left-to-right, right-to-left 모델을 결합해서 사용하는데 반해서 BERT에서는 하나의 bidirectional 모델을 사용해서 좌우 컨텍스트를 동시에 보는 구조를 가지고 있다. 이 특징을 통해 영화 리뷰에서 앞 뒤 문맥을 고려하여 감성을 더욱 정확히 분석할 수 있을 거라 기대가 된다.

1. **프로젝트 내용 및 현재까지의 진행 상황**

네이버 영화 리뷰를 수집하여, 주관성 탐지를 통해 감성분류 모델을 학습시키는데 필요한 요소, 즉 positive, negative, neutral과 같은 감성과 관련된 정보만 남긴다. 마지막으로 극성 분석을 통해 레이블링 하고, lexical meaning을 고려한 문맥까지 판단하여 기존의 연구보다 좀 더 정확한 감성 분석을 이끌어 내는 것을 목표로 한다. 위 프로젝트를 구현하는 데에 사용되는 방법은 BERT의 Bidirectional한 장점을 극대화하는 것에 초점을 둔다. 현재 BERT와 관련된 선행 연구들을 참고하여 네이버 영화 리뷰를 처리할 수 있는 감성분류기를 구축하였고, GPU 서버에서 학습 중이다. 모델 레이어, 하이퍼 파라미터 등의 변화에 따른 분류 정확도의 구체적인 결과값은 최종 보고서에 추가할 예정이다.

1. **결론 및 향후 연구**

본 프로젝트는 영화 리뷰에서의 감성추출에 관한 기존 연구의 문제점을 찾아내고 그것을 보완한 모델을 개발하는 것을 목표로 한다. 여러 방법을 고안하여 오차율을 줄이고 최적화된 감성 분석 알고리즘을 가진 분류 모델을 만들어 낸다면 평점과 같은 수치가 없는 데이터, 예를 들어 영화 관련SNS 댓글이나 영화 관련 컨텐츠의 실시간 유튜브 댓글에서의 감성 분석을 할 수 있다.

**참 고 문 헌**

[1] Learning Word Vectors for Sentiment Analysis, Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, Stanford University

Stanford, CA 94305, 2011

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. Cornell University, 11 Oct 2018

[3] 최석재, 권오병 (2014). 빅데이터 분석을 위한 한국어 SentiWordNet 개발 방안 연구. 한국전자거래학회지, 19(4), 1-19.

[4] 조태민, 이지형 (2015). LDA 모델을 이용한 잠재 키워드 추출. 한국지능시스템학회 논문지, 25(2), 180-185.

[5] 이상민, 이도훈 (2014). 소셜 네트워크 서비스 문장의 사전 기반 감정 분류 기법. 한국정보과학회 학술발표논문집, 255-257.

[6] 배장성, 이창기, 임수종 외 1명 (2019). BERT를 이용한 한국어 의미역 결정. 한국정보과학회 학술발표논문집, 512-514.