LSTM 기반의 감정분석을 통한 상품평 자동분류

공민정*, 김상원**, 김기천*
*건국대학교 컴퓨터공학과
**건국대학교 컴퓨터정보통신공학과
e-mail : kminamy@konkuk.ac.kr, llllssss94@konkuk.ac.kr, kckim@konkuk.ac.kr

Sentiment Analysis of Product Reviews using LSTM

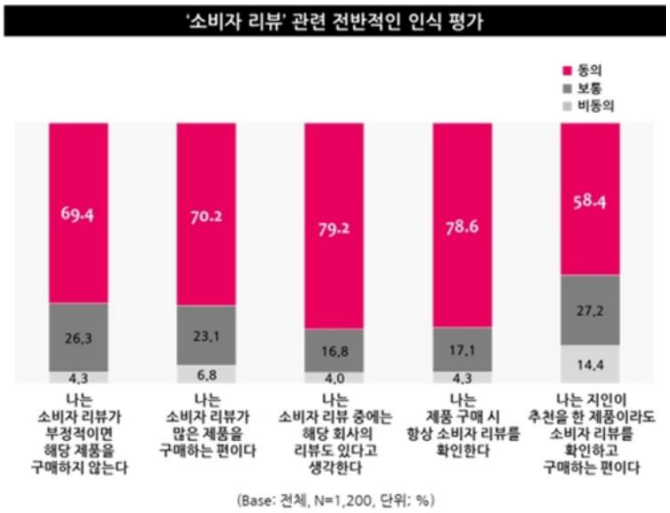
Minjeong Kong*, Sangwon Kim**, Keecheon Kim*
*Department of Computer Science and Engineering, Konkuk University
** Department of Computer, Information & Communications Engineering, Konkuk University

요 약

인터넷 기술의 발전에 힘입은 전자상거래의 급격한 발전에 따라 소비자들의 소비습관은 오프라인에서 온라인으로 빠르게 바뀌었다. 이에 따라, 구매한 상품에 대한 평가를 작성하는 것 또한 만연해지면서 소비자들에게 구매 결정의 중요한 요인으로 작용하기 시작하였고 실제 판매량에도 직접적인 영향을 끼치기 시작하였다. 그러나, 현재 전자상거래 시스템에서는 상품에 대한 평가를 한눈에 알아볼 수 있는 기능이 부재하고 있어 소비자의 소비 전략과 판매 전략측면에서의 비효율을 야기하고 있다. 따라서, 본 논문에서는 LSTM 을 기반으로 한 딥러닝 모델을 이용해 감정분석을 하여 온라인 상품평을 긍정/부정에 따라 자동으로 분류하고자 한다. 이를 통해, 효율적인 반응 분석을 위한 기술 개발의 기반을 마련하여 소비자와 판매자 모두에게 더 나아진 전략 수립의 기회를 제공할 것으로 기대한다.

1. 서론

지난 20 여년간, 전자상거래의 발전에 따라 소비자들의 소비습관이 온라인으로 상품을 구매하는 것으로 바뀌면서 구매한 상품에 대한 평가를 작성하는 것이 만연해졌다. 이에 따라, 상품평은 소비자에게 상품 구매 결정의 중요한 요인이 될 뿐만 아니라, 판매량에도 직접적으로 영향을 미치게 되었다.[1]



(그림 1) '소비자 리뷰' 관련 전반적인 인식 평가

[그림 1]은 소비자 리뷰가 제품구매에 결정적인 역할을 한다는 것을 잘 보여주고 있다.[2] 따라서, 소비자와 판매자는 상품평을 통해 전반적인 만족도, 상품의 장단점 등을 효과적으로 통찰하기를 원한다. 하지만, 현재의 온라인 쇼핑몰 시스템에선 상품평이 많을수록 일반적으로 긍정 및 부정의 평가가 무작위로 섞여 한눈에 종합적인 판단이 어렵고, 이는 소비자와 판매자 모두에게 비효율적이다. 이 문제를 해결하기 위해 본 논문에서는 LSTM(Long Short-Term Memory)을 기반으로 감정분석을 수행하여 온라인 상품평을 긍정/부정에 따라 자동으로 분류하려고 한다. 이를 통해 소비자는 상품에 대한 구매결정에 걸리는 시간을 단축시키고, 판매자는 리뷰에서의 피드백을 받아들임으로써 더 나은 방향으로 개선시키리라 기대한다.

본 논문의 구성은 다음과 같다. 서론에 이어 제 2 장에서는 단어 임베딩을 소개하고 제 3 장에서는 LSTM 기반의 모델을 설계한다. 마지막으로 제 4 장에서는 연구의 결론 및 향후 연구과제를 기술한다.

2. 관련 연구

2.1 데이터 셋

세계적인 전자상거래 사이트인 아마존닷컴에는 수많은 상품에 대한 다양한 리뷰가 존재하기 때문에 이를 활용한 공개된 데이터셋의 수가 많아 데이터의 사

용이 편리하다. 따라서, 본 논문에서는 Kaggle 대회에서 사용된 긍정과 부정이 섞여 있는 아마존 상품평 데이터셋을 이용하였다. 학습할 데이터는 2017 년 9 월부터 2018 년 10 월까지의 데이터로, 그 수는 5000 개이고 평균 글자수는 161.5 개이다. 테스트할 데이터는 2019 년 2 월부터 2019 년 4 월까지의 데이터로, 그 수는 28322 개이다. 학습과 테스트할 모든 리뷰는 영어로 작성되었다.

<표 1> 아마존 상품평 분류 예시

문장	긍정/부정
"We've ordered these a few times cause we've been so happy with product. These are good quality. We are also happy with price. I would recommend these."	긍정
"Worthless, except as a regular echo and a poor excuse for video chat."	부정
"Great battery just as good as the name brands."	긍정
"Very cheap and was not impressed at all never again"	부정

2.2 단어 임베딩(Word Embedding)

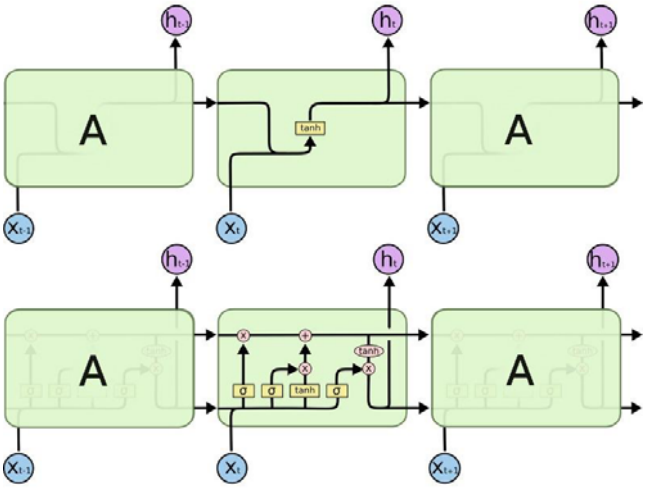
머신 러닝의 데이터 전처리 방식 중 하나인 one-hot encoding 은 단어 집합의 크기를 N 이라 할 때, 이를 N 차원의 벡터로 표현하고, 단어가 포함된 인덱스에 1 을, 다른 인덱스에는 0 을 부여하는 부호화 방법이다. 하지만, 단어를 벡터로 변환시킬 경우에 단어 간 의미의 차이를 벡터에 표현하지 못하기 때문에 이 방법은 단어 간의 유사성을 파악하지 못한다. 또한, 벡터의 길이가 총 단어의 수와 동일하기 때문에 단어의 개수가 늘어날수록 벡터의 차원도 마찬가지로 늘어나면서 이로 인해 신경망에서 성능이 좋지 못하게 되는 한계를 지니고 있다. 따라서, 본 연구에서는 자연어 처리를 위하여 보다 적절한 단어 임베딩 방식을 사용하고자 한다.

단어 임베딩이란, 컴퓨터가 이해하고 처리할 수 있도록 문장 내의 각 단어들을 실수 벡터로 변환시키는 것을 의미한다. one-hot encoding 과는 달리 단어 임베딩은 단어의 의미를 최대한 벡터에 담을 수 있도록 하여 단어 간의 유사성을 파악할 수 있도록 한다. 이는 embedding layer 를 훈련시키는 대신, 사전에 훈련된 단어 임베딩(pre-trained embedding)을 하여 embedding layer 로 전달함으로써 분류 모델을 훈련시키는데 있어 시간을 절약시켜준다. 왜냐하면 이와 같이 데이터들을 사전에 학습시키면 이후에 더 적은 데이터로도 미리 학습된 임베딩을 사용하여 성능을 높일 수 있기 때문이다. 이를 위해 대문자를 소문자로 바꾸고 구두점을 제거하고 stop word 를 제외시키고 단어 토큰들을 형성하는 등 데이터를 알맞게 처리하여 임베딩을 할 수 있는 알맞은 text corpus 를 준비한다. 이러한 작업은 Word2Vec 을 이용하여 진행한다.

Word2Vec 모델은 데이터셋에서의 각 문장을 취해 학습을 진행하고, 특정 단어가 있는 문장에서의 문맥을 파악함으로써 단어를 벡터화한다. 이에 따른 결과는 훈련한 corpus 에 각각의 단어마다 이에 해당하는 벡터를 포함한 embedding matrix 이다. 이러한 과정을 통해, embedding layer 의 input 으로서 사용 가능해진 단어 벡터들을 얻는다.[3]

3. 딥러닝 모델 설계

3.1 RNN 과 LSTM



(그림 2)RNN 셀 구조(위)와 LSTM 셀 구조(아래)[4]

RNN 의 큰 특징은 은닉층의 결과가 동일한 은닉층의 입력으로 받아들여지도록 연결되어 시간과 순서적인 측면을 고려할 수 있다는 점이다. 이러한 순서를 고려할 수 있는 특성은 sequence data 를 처리할 때 유용하게 활용될 수 있다. 왜냐하면 sequence data 의 가장 대표적인 예인 문장을 고려해 보았을 때, 문장은 문장 내의 단어들은 앞서 나왔던 단어들 간의 관계를 통해서 해석되기 때문이다. 하지만, sequence 가 길어질수록 문장에서 입력된 데이터와 이전에 나왔던 참고할 데이터의 위치의 차이가 커질 수 있기 때문에 RNN 으로서는 두 정보의 맥락을 파악하기 힘들어지고 성능 또한 저하된다. 즉, RNN 은 짧은 sequence 에서는 효과적이지만, sequence 가 길어질수록 비효율적이다. 이런 RNN 의 취약점을 극복하기 위해 본 논문에서는 이러한 문제를 해결하기 위해 Hochreiter 와 Schmidhuber 의 연구 결과를 바탕으로 RNN 의 종류 중 하나인 LSTM 을 사용한다.[5][6]

LSTM 이 RNN 과 구별되는 점은 상호작용을 하는 4 개의 layer 의 구조가 반복되는 모듈을 가지고, RNN 의 뉴런 대신에 cell state 구조를 사용한다는 것이다. cell state 는 정보들을 선별하여 흘러갈 수 있도록 하는 gate 를 활용하여 정보를 더하거나 제거하여 정보가 지속해서 다음 단계에 전달할 수 있도록 한다. 이는 긴 sequence 들을 기억하는데 유용하여 장기 의존성 문제를 해결할 뿐만 아니라, 이전의 데이터를 통해 미래 데이터를 예측 가능하게 한다. 또한, LSTM 은 RNN 에 forget gate 를 추가하여 RNN 의 vanishing gradient 와 exploring gradient 문제를 보완한다.

3.2 모델 설계

모델은 input layer, embedding layer, LSTM layer, output layer 로 총 4 개의 layer 로 이루어져 있다. embedding layer 에서는 가장 빈번하게 사용되는 단어의 크기인 input_dim, 단어가 임베딩될 벡터공간의 차원인 output_dim, 최장길이인 input_length 파라미터를 필요로 한다. 이 계층에서 훈련 및 테스트할 데이터를 벡터로 변환하는 것이다. embedding layer 를 통해 전환된 embedding 벡터들은 LSTM layer 에 input 으로서 더해진다. 이때, 이 feature 들은 이전 time step 의 결과가 LSTM 셀에 input 으로써 다시 사용된다. 마지막으로, output layer 에서는 sigmoid activation function 을 사용하여 0 또는 1 을 예측을 한다. 값이 0 일 경우에는 미래의 결과에 지장을 주지 않지만, 1 일 경우에는 미래 결과를 예측하는데 영향을 끼치도록 한다. 이 때문에 sigmoid 함수는 로지스틱 회귀분석이나 인공신경망의 이진분류를 할 경우에 사용되는데, 상품평을 긍정 또는 부정으로 분류하는 것은 후자에 해당하므로 본 연구에 이용하기 적합하다. RNN 에서 자주 발생하는 overfitting 문제를 완화하기 위해 embedding layer 와 LSTM layer 의 사이와 LSTM layer 와 output layer 사이에 dropout 을 넣어 준다. 이는 훈련하는 동안 인공신경망의 일부를 임의적으로 생략하게 하는데, 이러한 생략은 학습의 결과나 진행에 영향을 끼치지 않는다. [7]

4. 결론

잠재적 구매자와 상품 판매자 모두에게 상품평은 구매를 결정하거나 상품의 질이나 배송 등을 개선시키는 데 있어 중요한 역할을 해왔다. 하지만, 긍정적이거나 부정적인 상품평이 무작위로 섞여 있는 경우에는 이러한 점을 제대로 활용하지 못하게 된다. 따라서, 본 논문은 LSTM 을 기반으로 한 모델을 제안하여 상품평을 긍정적인지 부정적인지 자동으로 분류하도록 하였다. 상품평을 자동적으로 분류함으로써 판매자와 소비자 모두 리뷰를 효과적으로 사용할 수 있을 것이라 기대한다. 현재까지 감정분석과 연관된 연구들은 영어 등의 서양 중심의 언어로 진행되어왔기 때문에, 서양에 비해 어순이 상대적으로 중요하지 않고 교착어의 특징을 가진 한국어는 추가적인 연구를 필요로 한다. 따라서, 향후에는 한국어를 기반으로 감정분석을 할 수 있도록 하는 연구를 하고자 한다.

Acknowledgement

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 「대학혁신지원사업」의 연구결과입니다.

참고문헌

[1] Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. Management science, 54(3), 477-491.

[2]김종민,” '소비자 후기', 제품구매에 결정적.10 명중 7 명 "광고보다 더 신뢰"” ,2017.09.10, http://www.newsis.com/view/?id=NISX20170908_0000090409

[3] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

[4]” 텐서플로(TensorFlow)로 LSTM 구현하기, 한빛출판네트워크,2018 년 03 월 21 일 수정,2019 년 9 월 18 일 접속, http://www.hanbit.co.kr/channel/category/category_view.html?cms_code=CMS6074576268

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[6] Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02), 107-116.

[7] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.