

CNN-LSTM 조합모형을 이용한 영화리뷰 감성분석*

박호연

동국대학교_서울 일반대학원
경영정보학과
(hoyeonpark@dongguk.edu)

김경재

동국대학교_서울 경영대학
경영정보학과
(kjkim@dongguk.edu)

인터넷 기술과 소셜 미디어의 빠른 성장으로 인하여, 구조화되지 않은 문서 표현도 다양한 응용 프로그램에 사용할 수 있게 마이닝 기술이 발전되었다. 그 중 감성분석은 제품이나 서비스에 내재된 사용자의 감성을 탐지할 수 있는 분석방법이기 때문에 지난 몇 년 동안 많은 관심을 받아왔다. 감성분석에서는 주로 텍스트 데이터를 이용하여 사람들의 감성을 사전 정의된 긍정 및 부정의 범주를 할당하여 분석하며, 이때 사전 정의된 레이블을 이용하기 때문에 다양한 방향으로 연구가 진행되고 있다. 초기의 감성분석 연구에서는 쇼핑몰 상품의 리뷰 중심으로 진행되었지만, 최근에는 블로그, 뉴스기사, 날씨 예보, 영화 리뷰, SNS, 주식시장의 동향 등 다양한 분야에 적용되고 있다. 많은 선행연구들이 진행되어 왔으나 대부분 전통적인 단일 기계학습기법에 의존한 감성분류를 시도하였기에 분류 정확도 면에서 한계점이 있었다. 본 연구에서는 전통적인 기계학습기법 대신 대용량 데이터의 처리에 우수한 성능을 보이는 딥러닝 기법과 딥러닝 중 CNN과 LSTM의 조합모형을 이용하여 감성분석의 분류 정확도를 개선하고자 한다. 본 연구에서는 대표적인 영화 리뷰 데이터셋인 IMDB의 리뷰 데이터 셋을 이용하여, 감성분석의 극성분석을 긍정 및 부정으로 범주를 분류하고, 딥러닝과 제안하는 조합모형을 활용하여 극성분석의 예측 정확도를 개선하는 것을 목적으로 한다. 이 과정에서 여러 매개 변수가 존재하기 때문에 그 수치와 정밀도의 관계에 대해 고찰하여 최적의 조합을 찾아 정확도 등 감성분석의 성능 개선을 시도한다. 연구 결과, 딥러닝 기반의 분류 모형이 좋은 분류성적을 보였으며, 특히 본 연구에서 제안하는 CNN-LSTM 조합모형의 성과가 가장 우수한 것으로 나타났다.

주제어 : CNN, LSTM, 딥러닝, 조합모형, 영화리뷰, 감성분석

논문접수일 : 2019년 12월 5일 논문수정일 : 2019년 12월 5일 게재확정일 : 2019년 12월 16일
원고유형 : 일반논문 교신저자 : 김경재

1. 서론

분류분석 연구는 전통적으로 정형 데이터에 적용되어 왔으나 최근 텍스트, 이미지, 오디오 데이터 등 비정형 데이터의 양산으로 비정형 데

이터의 분류분석에도 많은 연구가 진행되고 있다. 그 중, 텍스트 분석을 다루는 NLP (Natural Language Processing) 연구가 활발하며, NLP 연구 중 감성분석은 온라인 텍스트에서 감성, 주관적인 생각, 감정 등을 식별하고, 서비스에 대한

* This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03032770).

느낌을 분석할 수 있다. 즉, 감성분석은 텍스트 시퀀스를 사전 정의된 범주를 이용하여 분류하는 분석방법이다. 감성분석 중 문장 관점에서의 분류분석을 극성분석이라 하며, 극성분석은 문서 내 의견이 긍정인지 부정인지 중립인지를 분류하는 것이다.

영화 리뷰와 같은 비정형 데이터의 극성분류는 정형 데이터보다 더 정교한 분류를 가능하게 하고 리뷰 내의 문장의 순서 등을 고려할 수 있는 기법이 필요하다. 선행연구에서는 주로 인공 신경망, SVM 등의 단순 기계학습기법을 활용하여 왔으나 전술한 특성 때문에 분류 성과가 미흡한 경우가 많았다. 본 연구에서는 이러한 한계점을 보완하고자 최근에 기계학습 분야에서 활발하게 연구되고 있는 대표적인 딥러닝 기법과 본 연구에서 영화 리뷰의 감성분석에 적합할 것으로 생각되어 CNN-LSTM 조합모델을 제안한다.

CNN은 주로 얼굴인식이나 이미지 분류에서 사용되는 알고리즘이지만, NLP에서 Bag of words와 유사하게 사용되며, LSTM은 주어진 단어를 미리 예측할 수 있도록 순차적으로 배열할 수 있는 장점이 있기에 주로 챗봇과 텍스트 번역에서 유용하게 이용되는 알고리즘이다. 따라서 CNN-LSTM의 조합모델을 이용하게 되면 양자의 장점을 활용할 수 있으며, 이를 통해 리뷰 데이터 감성분석 성능을 개선할 수 있을 것으로 기대된다.

한편, 딥러닝 기반 모델들은 NLP 등 여러 분야에서 전통적인 기계학습기법보다 뛰어난 성능을 갖고 있는 것을 알려져 있다. 하지만 실제 응용에 있어서는 과적합이 되지 않도록 견고성을 확보하는 것이 중요하다. 특히, 딥러닝 기법은 이미지 인식 등에서 활발하게 이용되고 있으나 영화 리뷰 등의 텍스트 분석 사례는 적기 때문에

본 연구에서 영화 리뷰 데이터의 극성 분류문제에 이용해 보고자 한다.

본 연구의 순서는 다음과 같다. 2장에서는 감성분석과 관련된 선행연구를 소개하고, 3장에서는 제안하는 CNN-LSTM 조합모델이 어떻게 구현되는지를 설명한다. 4장에서는 논문의 분석과정과 결과를 설명하며, 5장에서는 결론과 향후 연구방향을 제시한다.

2. 문헌연구

2.1 기계학습기법을 이용한 감성분석

감성분석은 텍스트에서 제품이나 상황 별 속성에 대한 주관적인 의견과 감정을 분류하고 식별하는 분석방법이다. 기본적인 감정분석은 긍정과 부정에 대한 빈도수 기반으로 제한되지만, 감정분석을 기계학습기반의 분류문제로 모델링할 경우, 사용자의 의도와 반응을 도출할 수 있다.

나이브베이지 (NB, Naive Bayes), 의사결정나무, kNN, 인공신경망, SVM (support vector machines), 최대 엔트로피 등이 지도학습기법 중 감성분석에 많이 활용되는 기법이다 (Appel et al., 2015). Pang et al.(2002)은 영화 리뷰 데이터 셋에 n-gram을 활용하여 감성분석을 진행하였으며, 나이브베이지, SVM, 최대 엔트로피로 분류 분석한 결과, SVM이 가장 우수한 정확도를 보였다. Ferguson et al.(2009)의 연구에서는 감성분석에 기계학습기법을 적용한 결과, 다항 나이브베이지가 SVM보다 우수한 성능을 보였다고 보고하였으며, Pak and Paroubek (2010)에서도 유사한 결과가 보고되었다.

한편, 텍스트 분류에서 짧은 텍스트를 처리할

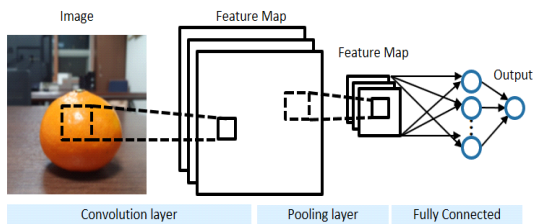
때에는 다항성과 동의어 문제에 직면하게 되는데, 이 문제를 해결하기 위하여 Lochter et al.(2016)는 각종 소셜네트워크에 게시된 9개 테마의 짧은 메시지를 대상으로 앙상블 기법과 결합하여 감성을 분석하였다.

일반적으로 감성분석은 영화 리뷰 데이터에 자주 적용되어 왔는데 그 이유는 영화 리뷰 데이터에는 일반적으로 영화에 대한 사용자의 의견 뿐만 아니라 평점 데이터가 존재하기 때문에 평점을 기반으로 리뷰의 극성을 판단하는데 유용하기 때문이다. 따라서 감성분석 연구 중 많은 연구들이 영화 리뷰 데이터를 활용하고 있다 (Lee et al., 2016; Kim and Song, 2016; Hyun et al., 2019).

2.2 딥러닝 기법

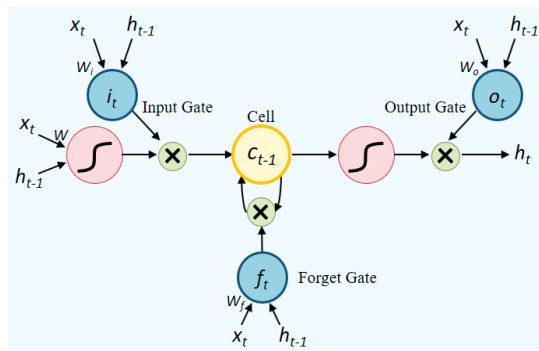
딥러닝은 일반적으로 기계학습기법의 하나인 신경망의 계층을 심화시킨 알고리즘을 의미한다. 전통적인 기계학습기법은 분류에 대한 특징 집합을 따로 추출하여야 하지만, 딥러닝은 특징 집합을 추출하는 과정뿐만 아니라, 복잡한 특징을 자동으로 추출할 수 있다 (Liu et al., 2016; Xu et al., 2016; Nguyen and Le, 2019).

딥러닝 기법 중 대표적인 기법인 CNN (Convolutional Neural Network)은 convolution layer와 pooling layer로 구성되어 있는데, Convolution layer는 가중치와 편향을 적용하고, pooling layer는 활성화를 수행한 convolution layer에서 벡터를 형성하여 값을 가져 온다 (Tang et al., 2019). CNN은 오타자가 있어도 감성분석에 탁월하기 때문에 사전 없이 사용되기도 한다. <Figure 1>은 일반적인 CNN의 구조를 그림으로 표현한 것이다.



<Figure 1> CNN structure

한편, LSTM(Long short-term memory)은 장기 의존성 문제를 해결하기 위해 고안된 모델 (Bengio et al., 1994)으로, 입력 게이트, 출력 게이트, 망각 게이트 등 3개의 gate 구조를 가지고 있으며 이는 <Figure 2>와 같다. 입출력 시점의 상태를 LSTM에 전달하면, 현재시점의 출력이 완전 연결 계층에 전달되어 업데이트 된 시점을 출력할 수 있게 된다(Graves, 2012; Xu et al., 2016; Rao et al., 2018; Abid et al., 2019).



<Figure 2> LSTM model

본 연구에서는 분류모형의 성능을 향상시키기 위해 RNN의 일종인 LSTM을 이용하여 IMDB의 감성분석 정보를 추출하고, 추출된 감성분석의 극성 예측 모델의 성능 평가를 실시한다.

3. CNN-LSTM 조합모델

이 연구에서 제안하는 모델은 감성분석의 긍정 및 부정의 특징 추출을 위해 CNN과 LSTM으로 이루어진 모델이다.

기존 기계학습과 달리 CNN은 convolution layer를 적용하여 특징의 자동 추출이 가능하고, 대규모 병렬 처리가 가능하다. 한편, LSTM은 CNN과 달리 대규모 병렬처리가 가능하지 않지만, RNN과는 다르게 원하는 시기에 진행 및 제어 할 수 있는 입력, 출력, 망각 게이트가 있다. LSTM의 게이트는 메모리 블록을 은닉노드에 배치할 수 있는 장점이 있는데, 메모리 블록이 모든 데이터를 기억할 수는 없지만, CNN의 장기 의존성 문제를 해결할 수 있다. 게다가 LSTM을 CNN의 pooling layer에서 이용할 경우, end-to-end 구조를 띄기 때문에 공간 및 시간적인 특징을 동시에 고려할 수 있다. 그 뿐만 아니라, LSTM은 단어를 예측할 때, 동일하게 sequence 벡터를 모델링할 수 있기 때문에 정확도를 향상시킬 수 있다.

먼저 word embedding을 수행하는데 이는 단어를 표현하기 위해서 사용되는 NLP 작업 중 하나이며, 감성분석에서 문장 간 유사성과 긍정, 부정의 레이블을 훈련시킨다. Word embedding의 출력을 통해 텍스트 매트릭스 벡터를 만드는 과정에서 CNN을 도입한다. CNN의 convolution layer의 목적은 kernel의 크기를 조정하여 다른 단어와 함께 사용될 때 다른 단어를 이해하게 할 수 있는 구조가 되게 하는 것이다. 이 구조를 이용하여 CNN이 쉽게 지역 특징을 추출할 수 있다. CNN에서는 각 단계마다 그래디언트 소멸 문제를 완화하기 위하여 ReLU를 활성화함수로 이용한다.

각 과정에서 ReLU를 진행 한 후, 데이터 중 전체 단어로 구성되어 있는 부분을 CNN이 인지하고 pooling을 적용한다. Max pooling은 pooling의 한 방법으로 단일출력으로 변환하여 지정한 영역의 값들을 한 곳으로 모을 수 있도록 하는 역할을 한다. Max pooling의 과정에서 convolution layer의 결과가 축소되고 다운 샘플링 된다. 이와 같이, convolution layer와 max pooling layer를 거치고 나면 감성분석에 필요한 특징이 추출되게 된다. Max pooling 이후에는 과적합을 방지하고 특정 입력에 집중하지 않도록 하기 위해 레이어의 입력 일부를 임의로 설정하는 dropout 과정을 거친다.

LSTM layer의 경우, CNN의 장기 의존성 문제를 보완하기 위하여 이용된다. LSTM은 순차 데이터 특성으로 사용되며, cell state라는 변수를 두어 메모리에 기억을 한다. LSTM은 입력, 출력, 망각 게이트를 갖고 있기 때문에 가변적으로 데이터 상황에 따라 제어할 수 있다. 다음으로 Dense layer라고 불리는 완전히 연결된 계층에서 sigmoid 함수를 이용하여 단일 값을 출력한다.

4. 분석과정

데이터 셋은 Keras의 Scikit-learn 라이브러리에서 제공된 IMDB(imdb.com) 영화 리뷰 데이터를 사용하였으며, 그 중 이미 긍정 및 부정으로 분류된 25,000개의 데이터 셋을 대상으로 진행되었다. 데이터 예시는 다음 <Table 1>과 같다.

<Table 1> IMDB sentiment data

No	Review	Sentiment
0	I recently bought the DVD, forgetting just how...	0
1	Nathan Detroit (Frank Sinatra) is the manager ...	1
2	Leave it to Braik to put on a good show. Final...	1
3	Nathan Detroit (Frank Sinatra) is the manager ...	1
...

4.1 연구 데이터와 전처리

25,000개의 데이터 셋 중 training은 80% (20,000개), validation은 20% (5000개)로 두었다. IMDB 데이터 셋은 리뷰 데이터를 바탕으로 긍정은 “1”, 부정은 “0”으로 레이블 구성 되어 있는 데이터지만, 기계학습과 딥러닝 모델에 최적화하기 위하여 텍스트 전처리가 필요하다. 텍스트 전처리가 진행되는 과정의 예는 <Table 2>와 같다.

<Table 2> Data cleaning example

No	Step	Example
1	Review	***SPOILER*** Do not read this, if you think a...
2	No punctuation	spoiler do not read this if you think about wa...
3	Tokenized data	[spoiler, do, not, read, this, if, you, think,...
4	Stopwords	[spoiler, read, think, watching, movie, although...
5	Lemmatization	spoiler read think watching movie although...

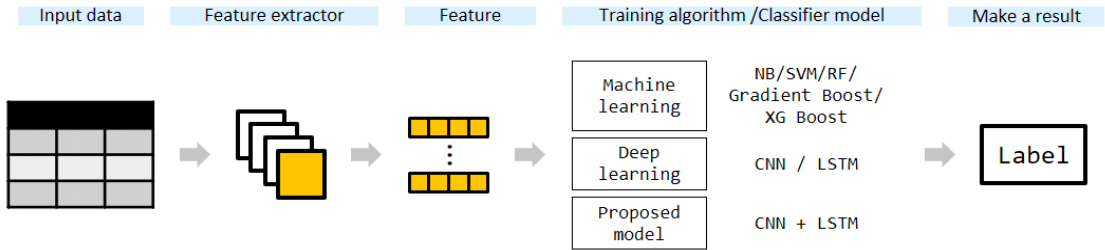
또한, 텍스트 리뷰 데이터에서 부호(punctuation mark)를 제거한 후, 토큰화(tokenization)를 진행하였다. 토큰화 라이브러리는 자연어 어휘를 많이 보유하고 있는 Stanford NLP를 이용하였다. Stanford NLP를 이용하여 단어를 분류한 후에는 대명사나 전치사 등의 stopwords를 제거하여 전처리를 하였다. 그 다음 단계에서는 텍스트 분석을 보다 정확하게 하기 위하여, 비슷한 의미끼리 lemmatization을 진행하였는데, Lemmatization은 문맥을 고려하여 비슷한 의미의 기본 단어로 변경해주는 것이다.

이상의 단어 전처리 이후에는 딥러닝과 기계학습에 용이하게 진행할 수 있도록 수치화해야 한다. 수치화하기 위해서, 분류된 단어는 숫자 텐서로 변형하는 벡터화를 적용하였다. 어휘의 고유 token은 one hot encoding을 통해 고차원 벡터를 구성하였으며, 벡터화 방식은 BoW와 TF-IDF 벡터로 나누었다. Bag of words는 단어가 몇 번 나오는지 계산하는 방법으로 나오는 순서는 고려하지 않는다. 반면에 TF-IDF는 문서와 전체 코퍼스를 벡터화하는 방법으로 텍스트 분류에 맥락을 많이 반영할 수 있는 벡터화 방법이다. 본 논문에서는 두 가지 벡터화를 모두 사용하여 분석을 진행하였다.

4.2 분석 절차

제안하는 모형의 성능을 테스트하기 위해서 선행연구의 검토를 통해 비교 기법을 선택하였다. 전체적인 분석과정과 비교 모형은 <Figure 3>와 같다.

먼저, 비교모형으로 사용하기 위하여 감성분석과 유사한 텍스트 분류 알고리즘에서 높은 수준의 분류성능을 보이는 기계학습기법인 NB,



(Figure 3) Proposed experimental process

SVM, XGboost를 선정하였고, 그 외에 RF (random forests)와 Gradient Boost를 비교 모형으로 채택하였다. 한편, 딥러닝에서 텍스트 데이터는 주로 CNN 알고리즘을 통해 텍스트 분류를 진행하기 때문에 따로 벡터화할 필요가 없지만, 전통적인 기계학습기법에서 사용하기 위해서 텍스트를 숫자로 벡터화해야 한다. 벡터화하는 방법 중에서는 Bag of Words(단어 주머니)와 TF-IDF가 많이 쓰인다.

BoW(Bag of Words)는 텍스트 문서의 문법과 단어 순서를 무시하여 코퍼스 전체를 기반으로 어휘 목록을 작성한다. 이 어휘목록들은 각각 숫자벡터로 표기되며, 데이터 확장성과 시계열 분류(Liu and Chen, 2015), 텍스트 처리(Baydogan et al., 2013) 등 여러 영역에서 효율적으로 모델링할 수 있으며, 이미지와 관련된 연구(Shuang et al., 2018)에서도 많이 활용되고 있다. 현재는 비정형 데이터의 기계학습이나 딥러닝 등을 이용한 분류문제에서 새로운 접근법을 제시할 때 쓰이고 있다.

TF-IDF(Term Frequency-Inverse Document Frequency)는 각 문서에 대한 상대 용어 빈도를 이용하는 방법으로 TF는 문서 내의 단어의 빈도를 나타내며, IDF는 동일한 단어 내에서의 또

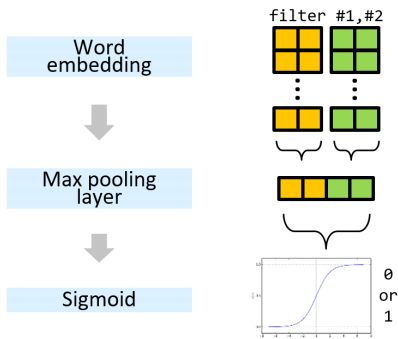
다른 문서별 빈도수를 정의한다. TF-IDF가중치가 높은 단어는 TF-IDF가중치가 낮은 단어보다 중요하며, 가중치가 높을수록 문서의 단어 의미 중요도가 높다. 본 연구에서는 데이터 전처리 후 텍스트를 벡터로 변환할 때, BoW와 TF-IDF를 사용하여 문서에 대한 특징을 추출하여 알고리즘을 비교할 수 있도록 하였다.

두 번째, 딥러닝의 대표적인 기법인 CNN, LSTM을 딥러닝 단일 비교 모형으로 채택했다. 마지막으로, 본 연구에서 제안하는 모형인 CNN과 LSTM을 조합한 모형을 제시한다. CNN-LSTM 조합모델은 각 기법의 약점을 서로 보완할 수 있고, 특히 LSTM의 end-to-end를 이용하여 layer 별로 학습성능을 향상시킬 수 있는 장점이 있다.

딥러닝 분석을 위해, lemmatization된 텍스트를 대상으로, 단어별 토큰화를 다시 진행하였다. 단어별 토큰화를 진행하면, 데이터를 벡터화처럼 숫자로 변화할 수 있을 뿐 아니라 CNN과 LSTM의 딥러닝에 용이하게 사용할 수 있는 sequence 데이터로 만들 수 있게 된다.

CNN만 단일 적용한 경우, one hot vector를 표현하는 word embedding 값들이 1차원으로 변경된다. 이 때, 임베딩 벡터 크기는 문장길이에 따라 결정되며, 패턴 내 단어 개수를 임베딩 벡터

크기의 합성곱 형태로 나타냄으로써 max pooling layer를 형성할 수 있도록 한다. 이 과정은 <Figure 4>으로 표현하였다.



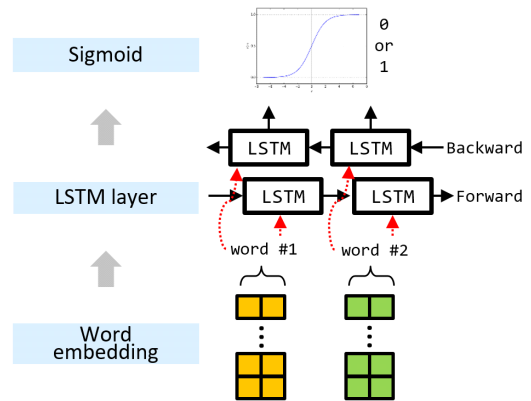
<Figure 4> Computation step of CNN

이 분석에서 형태소로 전처리된 문장의 최대 길이는 1191이기 때문에 리뷰 문장의 길이 차를 고려하여 단어 벡터의 최대 길이를 1200으로 재구성하여 word embedding 과정을 진행하였다. 또한, GPU의 성능을 고려하여 dimension은 100으로 구성하였다. 100 × 1200으로 구성된 word embedding을 토대로 convolution layer를 구성하였다.

Convolution layer는 filter를 1개씩 할당하는 convolution layer 1D를 256개로 구성하였으며, kernel은 2, 활성화함수는 ReLU, stride는 1로 정하였다. 전역으로 max pooling layer를 사용하여 kernel node 중 최댓값을 구하도록 하였다. Feature map을 설정하기 위하여 256개의 filter에 활성화함수 ReLU를 적용한 후, 1개씩 filter에 활성화함수 Sigmoid를 적용하여 분류기 함수에 근사 값을 갖게 하였다. 학습률을 설정할 수 있는 알고리즘으로 Adam(Adaptive moment)를 이용하여 속도벡터와 그라디언트 누적벡터를 계산한

후 다음단계에서 정확도를 측정할 수 있도록 하였다. Epoch은 30, batch size를 32로 설정하여 정확도를 측정하였다.

한편, LSTM만 적용한 경우에는 CNN과 비교할 수 있도록 word embedding 값을 동일하게 설정한 후 진행하였다. 최대길이는 1200, dimension은 100으로 구성한 후, 양방향 LSTM layer를 구성하였다. 이 과정은 <Figure 5>과 같다.

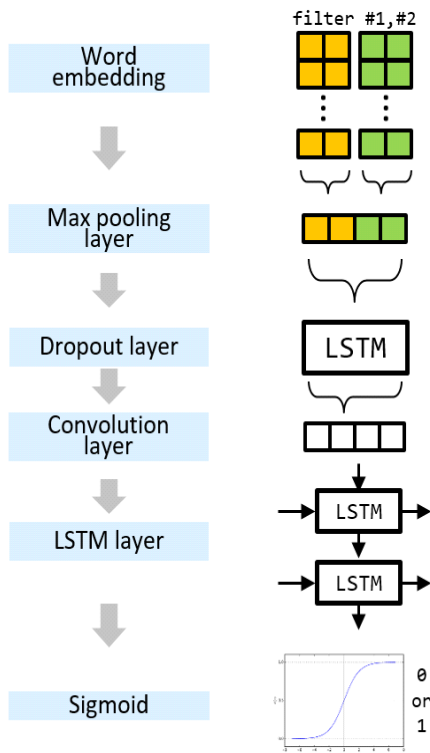


<Figure 5> Computation step of LSTM

LSTM layer의 설정은 sequence 분류 문제에서 모델 성능을 향상할 수 있도록, 입력 게이트는 forward로 설정하고 출력 게이트는 backward로 설정하였다. 단방향보다 양방향으로 설정한 이유는 sequence가 진행될 때, 더 빠른 속도로 완전한 학습이 가능하기 때문이다. 128개의 뉴런을 양방향 LSTM layer로 생성한 후, 25%의 dropout layer를 삽입하여 LSTM 반복학습을 유지할 수 있도록 지정하였다. Dropout layer 이후에는 CNN처럼 감성을 0이나 1로 모델링할 수 있도록 sigmoid를 활성화함수로 이용하였다.

본 연구의 제안 모형인 CNN-LSTM의 조합모델의 경우에는 CNN, LSTM의 단일모델과 같이

최대길이는 1200, dimension은 100으로 word embedding 값을 동일하게 설정하였다. CNN으로 설정할 때는 filter는 128개, kernel size는 5, feature map을 구성하기 위한 pool size를 4로 구성한 뒤, LSTM과 동일하게 25%의 dropout layer를 생성하였다. 이후에는 Adam으로 진행하였다. 이 과정은 <Figure 6>과 같다.



<Figure 6> Computation step of CNN-LSTM

4.3 평가 방법

제안하는 기법들을 이용한 감성분석의 결과들을 비교하기 위하여 Accuracy, Precision, Recall, F-measure 등의 평가지표를 이용하였다. 이 지표들은 분류분석에서 흔히 사용하는 것들로서 이

미 많은 선행연구에서 사용되고 있는 지표들이다. 제안된 평가지표의 산식은 다음과 같다.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F\text{-measure} = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (4)$$

4.4 분석 결과

감성분석을 위한 기계학습 분석 결과는 <Table 3>, 딥러닝에 대한 분석 결과는 <Table 4>와 같다.

전통적인 단일 기계학습기법에서 BoW는 82.23%, TF-IDF는 82.29%의 평균 정확도를 갖고 있었다. 기계학습기법 중 SVM 기반의 분류 모델이 BoW와 TF-IDF 두 데이터 세트에 대하여 딥러닝 수준의 높은 정확도를 갖고 있으며 안정적이었다. 다른 기계학습기법들은 SVM보다 정확도가 떨어졌으며, NB는 BoW 모델의 무작위성 때문에 분류 정확도가 가장 낮은 76.44%를 기록하였다. TF-IDF 모델에서는 Gradient Boost가 77.96%로 가장 낮은 정확도를 보였다. BoW와 TF-IDF의 벡터화 환경을 다르게 했을 때, SVM, RF, XG Boost의 경우, 1~2% 차이를 보였다. 다른 벡터환경에 대해서 NB가 약 5.56%, Gradient Boost는 약 7.8% 정확도 차이가 있었다.

단일 CNN 분석 결과, 89.04%의 정확도가 측

〈Table 3〉 Experimental results of machine learning techniques

Training algorithm	Vector	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Naive Bayes	Bag of Words	76.30	76.24	76.27	76.44
	TF - IDF	82.61	81.83	82.22	82.00
Support Vector Machine	Bag of Words	84.32	88.68	86.45	86.88
	TF - IDF	90.65	87.37	88.98	88.86
Random Forest	Bag of Words	81.38	82.17	81.77	82.00
	TF - IDF	85.62	78.62	81.97	81.40
Gradient Boost	Bag of Words	83.41	87.80	85.55	85.80
	TF - IDF	76.58	79.02	77.78	77.96
XG Boost	Bag of Words	76.62	86.71	81.35	80.05
	TF - IDF	78.09	87.40	82.48	81.23

〈Table 4〉 Experimental results of the proposed model and deep learning techniques

Training algorithm	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
CNN	88.41	89.69	89.04	89.04
LSTM	89.20	85.40	87.26	86.88
CNN-LSTM	91.22	89.77	90.48	90.33

정되었으며, 단일 LSTM 분석의 결과는 86.88%의 정확도가 측정되었다. 한편, CNN-LSTM 조합모델의 결과는 90.33%의 정확도가 측정되어 가장 우수한 분류 정확도를 보이는 것으로 나타났다.

5. 결론

이 연구에서는 영화 리뷰의 감성분석을 딥러닝 기법 알고리즘인 CNN과 LSTM을 결합하여 제안하였다. 그 결과, CNN-LSTM 조합모델의 정확도가 90.33%로 다른 전통적인 기계학습기법과 딥러닝 단일 분류 모형보다 더 안정적이고 정확하였다. 기계학습기법 중 SVM 모델도 정확도

88.86%로 우수한 성과를 보였지만, SVM은 sequence 모델이 아닌 vector형으로 변형해야 하는 단점이 있다. Sequence 모델이 아닌 vector형 모델을 사용할 경우, 단어별 출현 횟수에 따라 만들어진 특징 벡터가 벡터의 차원이 지나치게 커질 수 있는 문제인 “차원의 저주”에 해당할 수 있으며, 많은 정보가 유실될 수 있다. 이에 비해 CNN은 특징을 추출할 때, sequence 모델을 통해 네트워크 순서, convolution 커널 크기를 word embedding에 따라 구축할 수 있다. Sequence를 이용할 경우, “차원의 저주”를 극복할 수 있을 뿐 아니라, 원하는 문장을 자유롭게 분절 및 병렬 corpus화하여 훈련할 수 있다. 또한, sequence를 max pooling하여 가장 관련성 높은 문서로 표현할 수 있다. 한편, 조합모델에서는 LSTM을 합

게 사용하는데 이는 3개의 게이트를 이용하여, 입력과 출력 시기를 조정할 수 있다. 특히, 이를 이용함으로써 새로운 문장을 응용할 때, 다음 단어를 미리 예측할 수 있는 방식으로 학습할 수 있다. LSTM이 학습할 때, 확률분포 뿐 아니라 말뭉치로부터 긍정과 부정으로 예측된 다음 단어를 미리 예측하여 감성분석을 진행할 수 있는 장점 때문에 비교모형보다 정교한 결과를 나타낸 것으로 판단된다.

영화 리뷰와 같은 온라인 리뷰는 공개적으로 정보 수집을 편하게 할 수 있을 뿐 아니라 비즈니스에도 영향을 미친다. 그러나 온라인 리뷰는 항상 좋은 글만 있는 것은 아닐뿐더러 완벽하게 식별하기엔 어려움이 따른다. 그래서 이 연구에서는 리뷰를 최대한 자동화 분류를 할 수 있도록, 기계학습은 Bag of Words, TF-IDF의 벡터화를 이용하고, 딥러닝에서는 sequence data화하여 word embedding을 통해 텍스트를 수치화하여 분석하였다. 본 연구의 결과, CNN-LSTM 조합모델의 연구 결과가 몇몇 기계학습기법과 비교하여 주목할 정도로 정확도가 개선되지는 않았다. 향후 연구에서는 feature selection을 도입하여 훈련 데이터를 벡터화하여 분류하고, 매개변수의 영향을 비교하여 정확도를 개선할 수 있을 것이다.

본 연구의 결과에 대한 실무적 시사점은 딥러닝에 기반한 감성분석이 감성 예측을 더욱 정교하게 하여 흥행요인 분석 등에 활용될 수 있을 것이다. 정교한 영화 흥행요인의 분석은 영화 산업의 발전에 기여할 수 있을 것이다. 또한, 영화 리뷰에 대한 텍스트 분석은 영화 추천시스템에 활용될 수 있으며, 사용자의 평점 데이터에만 의존하는 기존의 추천시스템에 비해 감성분석의 결과를 고려하여 보다 정교한 추천 결과를 도출할 수 있을 것으로 기대된다.

참고문헌(References)

- Abid, F., M. Alam, M. Yasir, and C. Li, "Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter," *Future Generation Computer Systems*, Vol.95(2019), 292~308.
- Appel, O., F. Chiclana and J. Carter, "Main concepts, state of the art and future research questions in sentiment analysis," *Acta Polytechnica Hungarica*, Vol.12, No.3(2015), 87~108.
- Baydogan, M. G., G. Runger, and E. Tuv, "A bag-of-features framework to classify time series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.11(2013), 2796~2802.
- Bengio, Y., P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, Vol.5, No.2(1994), 157~166.
- Ferguson, P., N. O'Hare, M. Davy, A. Bermingham, P. Sheridan, C. Gurrin and A. F. Smeaton, "Exploring the use of paragraph-level annotations for sentiment analysis of financial blogs," *Proceedings of WOMAS 2009-Workshop on Opinion Mining and Sentiment Analysis*, (2009).
- Graves, A., "Long short-term memory," *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, Berlin, Heidelberg, 2012
- Hyun, J., S. Ryu, and S.-Y. Lee, "How to improve the accuracy of recommendation systems: Combining ratings and review texts sentiment scores," *Journal of Intelligence and Information Systems*, Vol.25, No.1(2019), 219~239.

- Kim, Y., and M. Song, "A study on analyzing sentiments on movie reviews by multi-level sentiment classifier," *Journal of Intelligence and Information Systems*, Vol.22, No.3(2016), 71~89.
- Lee, S.-H., J. Cui, and J.-W. Kim, "Sentiment analysis on movie review through building modified sentiment dictionary by movie genre," *Journal of Intelligence and Information Systems*, Vol.22, No.2(2016), 97~113.
- Liu, P., X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv 1605.05101*, 2016.
- Liu, S., and J.-H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Systems with Applications*, Vol.42, No.3(2015), 1083~1093.
- Lochter, J. V., R. F. Zanetti, D. Reller, and T. A. Almeida, "Short text opinion detection using ensemble of classifiers and semantic indexing," *Expert Systems with Applications*, Vol.62, (2016), 243~249.
- Nguyen, H., and N. M. Le, "An ensemble method with sentiment features and clustering support," *Neurocomputing*, Vol.370(2019), 155~165.
- Pak, A. and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, (2010), 1320~1326.
- Pang, B., L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Proceedings of the ACL-02 conference on Empirical methods in Natural Language Processing*, 2002.
- Rao, G., W. Huang, Z. Feng, and Q. Cong, "LSTM with sentence representations for document level sentiment classification," *Neurocomputing*, Vol.308(2018), 49~57.
- Sailunaz, K., and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *Journal of Computational Science*, Vol.36(2019), 101003.
- Shuang K., Z. Zhang, H. Guo, and J. Loo, "A sentiment information Collector - Extractor architecture based neural network for sentiment analysis," *Information Sciences*, Vol.467 (2018), 549~558.
- Tang, F., L. Fu, B. Yao, and W. Xu, "Aspect based fine-grained sentiment analysis for online reviews," *Information Sciences*, Vol.488(2019), 190~204.
- Xu, J., D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification," *arXiv preprint arXiv:1610.04989*, 2016.

Abstract

Sentiment Analysis of Movie Review Using Integrated CNN-LSTM Model

Ho-yeon Park* · Kyoung-jae Kim**

Rapid growth of internet technology and social media is progressing. Data mining technology has evolved to enable unstructured document representations in a variety of applications. Sentiment analysis is an important technology that can distinguish poor or high-quality content through text data of products, and it has proliferated during text mining. Sentiment analysis mainly analyzes people's opinions in text data by assigning predefined data categories as positive and negative. This has been studied in various directions in terms of accuracy from simple rule-based to dictionary-based approaches using predefined labels. In fact, sentiment analysis is one of the most active researches in natural language processing and is widely studied in text mining. When real online reviews aren't available for others, it's not only easy to openly collect information, but it also affects your business. In marketing, real-world information from customers is gathered on websites, not surveys. Depending on whether the website's posts are positive or negative, the customer response is reflected in the sales and tries to identify the information. However, many reviews on a website are not always good, and difficult to identify. The earlier studies in this research area used the reviews data of the Amazon.com shopping mal, but the research data used in the recent studies uses the data for stock market trends, blogs, news articles, weather forecasts, IMDB, and facebook etc. However, the lack of accuracy is recognized because sentiment calculations are changed according to the subject, paragraph, sentiment lexicon direction, and sentence strength. This study aims to classify the polarity analysis of sentiment analysis into positive and negative categories and increase the prediction accuracy of the polarity analysis using the pretrained IMDB review data set. First, the text classification algorithm related to sentiment analysis adopts the popular machine learning algorithms such as NB (naive bayes), SVM (support vector machines), XGboost, RF (random forests), and Gradient Boost as comparative models. Second, deep learning has demonstrated discriminative features that can extract complex features of data.

* Department of Management Information Systems, Graduate School, Dongguk University_Seoul
** Corresponding Author : Kyoung-jae Kim
Business School, Dongguk University_Seoul
30 Pildong-ro 1gil, Chung-gu, Seoul 100-715, Korea
Tel: 02-2260-3324, Fax: 02-2260-3684, E-mail: kjkim@dongguk.edu

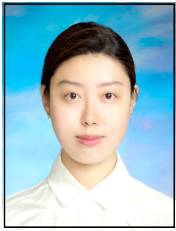
Representative algorithms are CNN (convolution neural networks), RNN (recurrent neural networks), LSTM (long-short term memory). CNN can be used similarly to BoW when processing a sentence in vector format, but does not consider sequential data attributes. RNN can handle well in order because it takes into account the time information of the data, but there is a long-term dependency on memory. To solve the problem of long-term dependence, LSTM is used. For the comparison, CNN and LSTM were chosen as simple deep learning models. In addition to classical machine learning algorithms, CNN, LSTM, and the integrated models were analyzed. Although there are many parameters for the algorithms, we examined the relationship between numerical value and precision to find the optimal combination. And, we tried to figure out how the models work well for sentiment analysis and how these models work. This study proposes integrated CNN and LSTM algorithms to extract the positive and negative features of text analysis. The reasons for mixing these two algorithms are as follows. CNN can extract features for the classification automatically by applying convolution layer and massively parallel processing. LSTM is not capable of highly parallel processing. Like faucets, the LSTM has input, output, and forget gates that can be moved and controlled at a desired time. These gates have the advantage of placing memory blocks on hidden nodes. The memory block of the LSTM may not store all the data, but it can solve the CNN's long-term dependency problem. Furthermore, when LSTM is used in CNN's pooling layer, it has an end-to-end structure, so that spatial and temporal features can be designed simultaneously. In combination with CNN-LSTM, 90.33% accuracy was measured. This is slower than CNN, but faster than LSTM. The presented model was more accurate than other models. In addition, each word embedding layer can be improved when training the kernel step by step. CNN-LSTM can improve the weakness of each model, and there is an advantage of improving the learning by layer using the end-to-end structure of LSTM. Based on these reasons, this study tries to enhance the classification accuracy of movie reviews using the integrated CNN-LSTM model.

Key Words : CNN, LSTM, Deep Learning, Integrated Model, Movie Review, Sentiment Analysis

Received : December 5, 2019 Revised : December 5, 2019 Accepted : December 16, 2019

Publication Type : Regular Paper Corresponding Author : Kyoung-jae Kim

저 자 소개



박호연

동국대학교에서 컴퓨터공학을 전공하여 공학사, 경영정보학을 전공하여 경영학석사를 취득하였으며, 현재 동교에서 경영정보학을 전공하여 박사과정을 수료하였다. 주요 관심분야는 딥러닝, 빅데이터, 비즈니스 애널리틱스, 텍스트마이닝 등이다



김경재

현재 동국대학교 경영대학 경영정보학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, 연구 관심분야는 비즈니스 애널리틱스, CRM, 추천기술, 빅데이터 분석 등이다.