워드임베딩과 텍스트 합성곱 신경망을 이용한 강의평가 감정 분석

Lecture Evaluation Sentiment Analysis using Word Embedding and Text-CNN

저자
(Authors)

임호준, 최강혁, 신수용
Hojun Lim, Ganghyeok Cho, Sooyong Shin

출처
(Source)

한국정보과학회 학술발표논문집 , 2017.12, 1953-1955 (3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07322762

APA Style

임호준, 최강혁, 신수용 (2017). 워드임베딩과 텍스트 합성곱 신경망을 이용한 강의평가 감정 분석. 한국정보과학회 학술발표논문집, 1953-1955.

워드임베딩과 텍스트 합성곱 신경망을 이용한 강의평가 감정 분석

임호준○, 최강혁, 신수용
경희대학교 컴퓨터공학부
lhj@oslab.khu.ac.kr, public6174@gmail.com

Lecture Evaluation Sentiment Analysis using Word Embedding and Text-CNN

Hojun Lim○, Ganghyeok Cho, Sooyong Shin
Computer Science and Engineering, Kyung Hee University

요 약

수강신청 시즌이 되면, 많은 대학생들이 보다 좋은 강의를 찾기 위해 다양한 노력들을 한다. 그러한 노력 중 하나로, 해당 강의를 수강한 다른 학우의 강의 평가를 열람하는 것이 있다. 강의 평가를 읽는 작업은 강의에 대한 정보를 확실하게 얻을 수 있으나, 수 많은 강의 평가를 일일이 읽어야 한다는 어려움이 있다. 본 논문은 이러한 점에 착안하여, 대학생들이 수강하고자 하는 강의에 대해 여러 강의 평가를 읽어보지 않더라도 해당 강의 평가에 대한 지배적인 감정을 효율적으로 얻을 수 있도록 강의 평가 텍스트 감정 분석 연구를 수행하고자 한다.

강의 평가 텍스트에서 의미 있는 정보를 추출하기 위해서, 최근 감정 분석 연구에서 다양하게 활용되고 있는 딥 러닝(Deep Learning)을 적용하였다. 확도 높은 감정 분석을 진행하기 위해 분산 표현 방식의 워드 임베딩(Word Embedding)과 최근 자연어 처리에서 그 효과를 입증하고 있는 합성곱 신경망(Convolutional Neural Networks)를 사용하였다. 실제 수집된 강의 평가 데이터를 사용한 실험에서 긍정, 부정 두 가지 감정을 80%이상의 정확도로 분류하는 것을 확인하였다.

1. 서 론

사람들이 일상 생활, 문화 생활등을 할 때 텍스트 데이터는 자연히 발생하게 된다. 그 예시로, 영화를 보고 해당 영화에 대한 감상 평을 남긴다거나, 제품의 후기를 남긴다거나, 페이스 북과 같은 SNS(Social Network Service)에 자신의 일상을 기록하는 것 등이 있다.

위와 같이 발생되는 텍스트 데이터에서 영화의 긍 부정 분석과 같은 작업을 수 작업으로 처리 하는 것이 아니라, 자동화 할 수 있다면 다양한 이점을 얻을 수 있으므로 최근 이를 위한 다양한 연구가 진행되고 있다.

본 논문에서도 텍스트 데이터에서 유 의미한 정보를 추출하여, 사용자에게 편의성을 제공할 수 있는 시스템을 구현하고자 한다. 여기서 말하는 사용자는 대학교에 재학 중인 학생과 복학할 예정인 학생을 말한다.

수강신청 시즌이 되면, 많은 대학생들이 보다 좋은 강의를 찾기 위해 다양한 노력들을 한다. 그러한 노력들로는 해당 강의를 수강한 선 후배에게 물어보거나, 대학생 커뮤니티 사이트에 다른 학생이 올리는 강의 평가를 열람하는 것 등이 있다.

이 중 커뮤니티 사이트에서 강의 평가를 열람하는 것은 강의에 대한 정보를 열람하는 것은 확실하고, 간편하게 얻을 수 있다. 그러나 한 강의만 하더라도 여러 개의 강의 평가가 존재하고, 각각의 강의 평가는 여러 개의 문장으로 이루어져 있어 적지 않은 시간을 투자해야 한다. 따라서 강의 평가 전체를 읽어보지 않더라도, 해당 강의 평가의 지배적인 감정, 의견을 얻을 수 있다면 강의 평가를 읽는데 시간을 투자하지 않더라도 효율적으로 정보를 얻을 수 있다.

강의 평가 텍스트로부터 작자의 감정(긍정, 부정)을 인식할 수 있도록 텍스트 데이터와 관련된 자연어 처리, 감정 분석에서 뛰어난 효과를 보이고 있는 딥 러닝(Deep Learning)을 적용하였다.

2. 관련 연구

텍스트 기반에서 작자의 감정을 분석하는 연구들이 [1-4]과 같이 수행되었다. [2-4]는 [1]1992년에 발표된 SVM(Support Vector Machine)를 사용하여 감성분석을 하였다. [2]는 140자 이내의 단문 메시지를 음절 커널에서 음운 커널로 수정한뒤, SVM을 적용하여 메시지 내의 긍정, 부정적 의미를 분류하였다. 그러나 문장 단위로 긍정, 부정의 감정을 분류하기 때문에, 본 논문에서처럼 전체 텍스트에서 지배적인 감정(긍정, 부정)을 분류하기에는 적합하지 않았다. [3]은 스탠포드 감성 트리 말뭉치의 모든 노드와 감성 태그를 추출한 뒤, 문장 단위로 SVM을 적용하여, 긍정, 부정의 2가지 감정을 분류하였다. 그러나 [3]에서는 교착어(agglutinative language)인 한글과는 구조가 다른 고립어(isolating language)인 영어를 대상으로 진행한 연구였고, 단어 사전을 구축하지 않았기에 본

논문과는 성향이 다르다. [4]는 트위터를 대상으로 표본을 추출하고, 형태소와 음절을 자질로 사용한 뒤, SVM을 사용해 감정을 분류하였다.

3. 강의 평가내의 감정 분석 구현

3.1 데이터 수집

본 논문에서는 실험을 위해 대학생 커뮤니티 사이트인 에브리타임에서 경희대 학생들이 강의를 듣고 작성한, 강의 평가 11,000개를 수집하였다. 강의평가는 작자가 작성한 텍스트와 작자가 평가한 해당 수업의 별점(1~5)로 이루어져있다.

3.2 데이터 전처리

실험데이터는 한글로 작성되어있다. 한글의 경우 조사 등의 복잡한 언어 구조 때문에 조사를 제거하는 것이 텍스트 기반 감정 분석에 효율적이다. KoNLPy 형태소 분석기를 사용하여 불필요한 조사를 실험데이터로부터 제거하였다.

지도학습을 진행하기 위해, 에브리타임에서 수집된 강의 평가 텍스트에 작자가 부여한 별점에 따라 1~2점의 경우 부정, 3점~5점인 경우에는 긍정인 것으로 라벨링(Labeling) 하였다.

인공신경망은 숫자로 이루어진 데이터를 입력 값을 사용한다. 따라서 딥 러닝 기반의 감정분석을 수행하기 위해서는 텍스트 데이터의 각 단어들을 숫자로 표현해야 한다. 이를 위해 수집된 강의 평가 텍스트 내에서 등장 빈도 수가 많은 단어 12,000개를 추출하여 1~12000까지의 인덱스(index)를 부여 단어 사전을 생성했다.

3.3 Word Embedding & Word2Vec

워드 임베딩(Word Embedding) 방법으로는 그림 1의 (a)국소표현(one-hot representation) 과 (b)분산 표현(distributed representation)이 있다[5].

$$w_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} w_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdots w_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

(a) 국소표현(one-hot representation)

$$w_1 = \begin{bmatrix} 0.3 \\ 0.9 \\ 0.8 \\ 0.5 \\ 0.1 \end{bmatrix} w_2 = \begin{bmatrix} 0.2 \\ 0.9 \\ 0.9 \\ 0.5 \\ 0.1 \end{bmatrix} \cdots w_n = \begin{bmatrix} 0.8 \\ 0.1 \\ 0.2 \\ 0.1 \\ 0.9 \end{bmatrix}$$

(b) 분산표현(distributed representation)

그림 1. 단어의 국소표현과 분산표현 예[5]

국소 표현은 간편하지만 단어 간의 의미 및 관계를 표현하지 못한다. 그러나 분산 표현은 실수 값으로 이루어진 일련의 벡터를 사용한다. 따라서 비슷한 단어는 비슷한 실수 값의 벡터를 가지게 되므로, 단어 사이의 관계를 나타낼 수 있다.

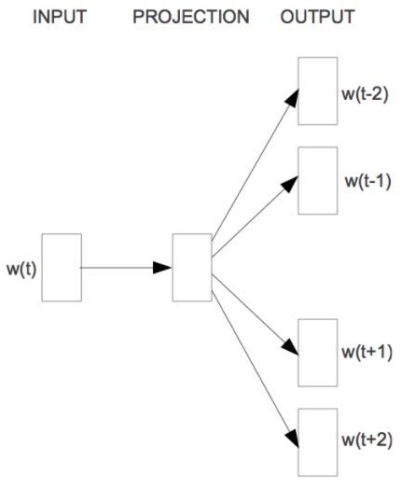

그림 2. skip-gram Architecture[6]

[6]워드 임베딩의 분산표현 모델 중 Word2Vec의 Skip-gram 모델은 특정 단어를 입력으로 하여 설정된 최대 거리만큼 전후의 단어들을 예측한 뒤, 학습 데이터와 출력 값과의 오차를 줄이는 방향으로 행렬이 학습된다. 학습 속도는 비교적 느리나, 다른 모델에 비해 뛰어난 성능을 보임으로 Skip-gram 모델을 사용하여 학습시켰다. 그림 2는 Skip-gram의 구조를 나타낸다.

3.4 Text-CNN

[8]에서는 Word2Vec을 사용하고, 하나의 레이어(layer)로 구성된 간단한 형태의 합성곱 신경망(Convolutional Neural Networks)이 텍스트 감정 분석에 뛰어난 성능을 보임을 입증하였다. 본 논문에서는 [8]에서 제시된 합성곱 신경망의 구조를 적용하여 입력 값으로 들어오는 텍스트로부터 긍정, 부정 두 가지 감정을 분류하기 위한 모델을 구현하였다. 그림 3은 구현에 사용된 합성곱 신경망의 구조를 나타낸다.

첫 번째 레이어는 입력 텍스트에 들어있는 단어들을 단어 사전과 학습된 Word2Vec을 통해 인덱스로 변환한 후, 벡터로 임베드하여 2차원 테이블을 구성한다.

두 번째 레이어는 입력 데이터로의 단어들의 묶음에서 특징을 추출하기 위해, 크기가 다른 다수의 필터를 이용하여 특징지도를 생성한다.

세 번째 레이어는 각 특징지도에서 최대 값을 갖는 하

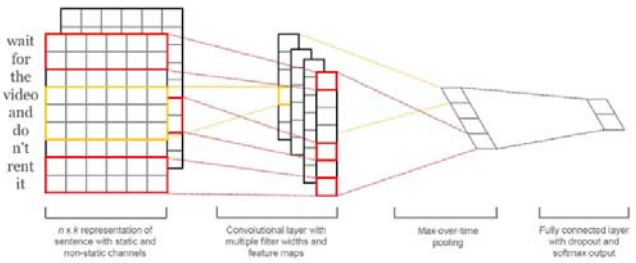

그림 3. 감성분석을 위한 합성곱신경망[8]

나의 특징을 선택한 뒤, 선택된 특징들을 연결한다.

네 번째 레이어는 연결된 특징들을 기반으로 완전 연결 레이어(Fully-Connected layer)를 구성한 뒤, 최종적인 감정을 분류한다.

마지막으로 과적합(Overfitting)을 막기 위해 L2 regularization과 Dropout[9]를 적용하였다.

4. 실험 및 결과

감정 분석 분류 모델을 평가하기 위해 3.1에서 수집한 데이터 셋을 사용하였다. 데이터 셋의 전처리는 형태소 분석기를 사용하여 조사만 제거하였다. ㅠㅠ, ㅎㅎ와 같은 이모티콘의 경우 작자의 감정을 분석하는데 도움이 될 거라 판단해 제거하지 않았다.

감정 분석 분류 모델인 Text-CNN은 Tensorflow[10]를 이용하여 구현하였으며, 워드 임베딩은 3.3에서 128차원으로 분산 표현하여 학습시킨 Word2Vec을 사용하였다.

데이터 셋이 11,000개로 많지 않으므로, cross validation (CV)를 이용하여 수행하였다. CV는 데이터 셋을 k의 같은 크기로 나눈 다음 k개를 각각 평가에 사용하여 평균을 내는 것을 말하며 본 논문에서는 k의 값을 10으로 하였다.

표 1 감정 분석 분류 결과의 Confusion Matrix

		Prediction	
		Positive	Negative
Label	Positive	8344	325
	Negative	1674	657

표 1은 감정 분석 실험 결과의 Confusion Matrix이다. 표 안의 숫자는 데이터의 수를 나타낸다. 합성곱 신경망을 적용한 감정 분석 분류 모델이 강의 평가 텍스트에서 긍정, 부정의 두 가지 감정을 분류했을 경우 10-CV로 계산된 정확도(accuracy)는 82.54%였다.

5. 결론

본 논문에서는 Word2Vec의 Skip-gram 기법을 이용하여 워드 임베딩을 진행한 뒤, 이를 사용하여 Text-CNN으로 강의 평가 텍스트 내의 감정 분석을 수행하였다. 실험 결과를 통해 최근 자연어 처리에서도 적용이 되고 있는 [7]합성곱 신경망의 성능을 다시금 확인하였다.

향후 연구로는 현재의 모델의 정확도 개선과 범용성을 위해 추가적인 학습 데이터를 수집하고, 파라미터(parameter)등의 수정을 통해 최적화를 진행할 것이다.

6. Acknowledments

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터(IITP)에서 지원하는 서울어코드활성화지원사업(2011-0-00883)과 SW중심대학지원사업(2017-0-00093)의 지원으로 수행되었음.

참 고 문 헌

[1] Bernhard E Boser, Isabelle M Guyon, andVladimir N Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” *Proc. The Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152, 1992.

[2] 김현우, 이승룡, “모바일 텍스트의 감성분류를 위한 SVM 기반 음운 커널 기법,” *정보과학회논문지 : 소프트웨어 및 응용* 제 40권 제 6호, 2013

[3] 이성욱, “스탠포드 감성 트리 말뭉치를 이용한 감성 분류 시스템,” *Journal of the Korean Society of Marine Engineering*, Vol. 39, No. 3 pp. 274~279, 2015

[4] 임좌상, 김진만, “한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교,” *Journal of Korea Multimedia Society* Vol. 17, No. 2, pp. 232-239, February 2014

[5] 서상현, 김준태, “딥러닝 기반 감성분석 연구동향,” *한국멀티미디어학회지* 제 20권 제 3호 2016년 9월

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, Jeff Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Advances in Neural Information Processing Systems* 26, Pages 3111-3119, 2013

[7] Rojas-Barahona LM., “Deep learning for sentiment analysis,” *Lang Linguist Compass*, pp.701-719, 2016

[8] Kim, Yoon. "Convolutional neural networks forsentence classification.", *arXiv preprint* arXiv:1408.5882, (2014).

[9] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research 15*, Pages 1929-1958, 2014

[10] M Abadi, et al, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed System,” arXiv:1603.04467, 2016