Word2vec을 이용한 오피니언 마이닝 성과분석 연구

Performance Analysis of Opinion Mining using Word2vec

어 균 선*, 이 건 창**†
성균관대학교 일반대학원 경영학과*,
성균관대학교 경영대학 글로벌경영학과/삼성융합의과학원**
†교신저자 (kunchanglee@gmail.com)

Kyun Sun Eo*, Kun Chang Lee**†
SKK Business School, Sungkyunkwan University*
SKK Business School/SAIHST (Samsung Advanced Institute of Health Sciences & Technology)
Sungkyunkwan University**
†Corresponding Author

요약

본 연구에서는 Word2vec을 머신러닝 분류기를 이용해 효율적인 오피니언 마이닝 방법을 제안한다. 본 연구의 목적을 위해 BOW(Bag-of-Words) 방법과 Word2vec방법을 이용해 속성 셋을 구성했다. 구성된 속성 셋은 Decision tree, Logistic regression, Support vector machine, Random forest를 이용해 오피니언 마이닝을 수행했다. 연구 결과, Word2vec 방법과 RF분류기가 가장 높은 정확도를 나타냈다. 그리고 Word2vec 방법이 BOW방법 보다 각 분류기에서 높은 성능을 나타냈다.

키워드: 워드임베딩; 오피니언 마이닝; 분류기; BOW 방법

Abstract

This study proposes an analysis of the Word2vec-based machine learning classifiers for the sake of opinion mining tasks. As a bench-marking method, BOW (Bag-of-Words) was adopted. On the basis of utilizing the Word2vec and BOW as feature extraction methods, we applied Laptop and Restaurant dataset to LR, DT, SVM, RF classifiers. The results showed that the Word2vec feature extraction yields more improved performance.

Keywords: Word2vec; opinion mining; classifiers; BOW (Bag-of-words)

1. 서론

오피니언마이닝(Opinion mining)은 텍스트에 내포된 긍정적인 의견 또는 부정적인 의견을 분석하는 연구분야이다. 인터넷 뉴스, 블로그, 소셜네트워크 서비스의 발전과 더불어 사용자가 작성한 콘텐츠가 폭발적으로 증가하게 되었다[1]. 그 중 텍스트는 콘텐츠에서 상당 부분 차지하고 있기 때문에 오피니언마이닝은 점점 중요해지고 있다. 오피니언마이닝은 감성분석의 한분야로서 가장 활발한 연구 분야중 하나이다[2].

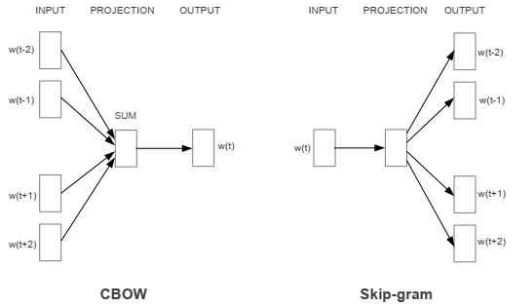
감성 분석(Sentiment analysis)은 상품 및 서비스에 대한 리뷰를 긍적적 또는 부정적인지 자동으로 분류하는 것으로 머신러닝을 이용해 높은 성과를 나타내고 있다. 상품에 대한 반응을 파악하는 것은 기업의 입장에서는 매우 중요하다고 할 수 있다. 상품에 대해 긍정적인 반응이면 장점을 극대화 할 수 있고 부정적인 반응이면 개선안을 마련할 수 있다. 이에 대해 기존의 감성분석은 BOW(Bag-of-words) 방법을 이용해 N-gram의 조합을 만들어 속성을 구성했다. 하지만 이 경우는 문장의 의미론적(Semantic) 특성을 반영하지 않고 분류 모델을 구축한다는 단점이 있다. 최근연구는 벡터공간에 단어를 표현하는 워드임베딩(Word embedding)이 활발히 연구된다[3]. 본 연구에서는 Wor2vec(Word to vector)방법을 이용해 의미론적 특성을 반영할 수 있는 오피니언마이닝 방법을 제안한다.

2. 관련연구

2.1. Representation of Word2vec

Word2vec(Word embedding to vector)은 Tomas Mikolov와 그의 팀원들에 의해 연구된 딥러닝 방법을 이용한 단어 임베딩 학습 모델이다[3]. 머신러닝 분류기 모델을 적용하기 위해서는 단어를 벡터로 나타내는 과정이 필요하다. 단어를 벡터로 표현하는 방법 중에는 텍스트 내에 있는 단어를 포착해 해당 단어가 존재하면 1로 표현하고 존재하지 않으면 0으로 표현하는 BOW방법이 있다. 나아가 단어의 빈도수를 측정하는 TF(Term frequency)와 , 문서 내 단어 중요도를 나타내는 의 TF-IDF(Term frequency-Inverse document frequency)를 측정해 단어벡터를 구성한다. 하지만 이 BOW 방법은 단어 사이의 관계를 파악할 수 없고, 텍스트 전체의 단어를 이용해 속성(feature)를 생성하기 때문에 벡터의 크기 증가 및 희박성(sparsity) 문제로 견고한 머신러닝 모형을 만드는 것은 어렵다. BOW와 달리 Word2vec을 이용해 단어를 학습할 경우, 문맥상 비슷한 의미를 가진 단어들은 서로 가까운 공간벡터를 가진다. Word2vec은 워드임베딩을 표현 하는 2가지 방법을 제공한다. 모델은 그림1과 같다. 첫 번째는 CBOW(Continuous bag-of-words) 모델이고, 두 번째는 skip-gram모델이다. CBOW모델은 전체 텍스트에서 단어의 주변 단어들을 이용하여 단어를

예측하는 모델 구조이고, skip-gram 모델은 주어진 단어를 통해 주변 단어를 예측한다.



▶▶ 그림 1. Word2vec 모형

3. 연구방법

본 연구는 Word2vec과 머신러닝 분류기를 적용해 효율적인 오피니언 마이닝 방법을 제안하고자 한다. 본 연구는 다음과 같은 단계로 구성된다.

1. 데이터의 전처리 단계로 리뷰 데이터에 있는 문장을 단어로 분리하고 분석에 불필요한 텍스트를 제거한다.
2. 전처리된 텍스트를 벡터화 한다. 본 연구에서는 BOW과 Word2vec 두가지 벡터화 방법을 사용했다.
3. 벡터화 된 데이터 셋을 입력변수로 지정하고 긍정, 부정으로 된 감성변수를 출력변수로 사용한다.
4. BOW와 Word2vec 속성으로 된 벡터표현간의 분류기 성능을 측정한다.

3.1. 데이터

본 연구는 SemEval2014 데이터를 사용했다[4]. SemEval2014 데이터는 Laptop, Restaurant으로 두가지 주제로 이루어져 있으며 Laptop은 총 1853건, Restaurant는 2969건이다. 두 도메인 모두 긍정적인 리뷰, 부정적인 리뷰로 구성되어 있다. 본 연구에서는 분류성능 평가를 위해 10 fold cross validation을 사용해 분류 정확도(Accuracy)를 측정했다.

4. 연구 결과

본 연구에 사용한 BOW 방법과 Word2vec 방법에 대한 분류기 별 성능 평가 결과는 다음 표와 같다. 연구에 사용한 분류기는 다음과 같다. 의사결정나무(Decision tree, DT), 로지스틱 회귀분석(Logistic regression, LR), 서포트 벡터 머신(Support vector machine, SVM), 랜덤 포레스트(Random forest, RF)이다. Laptop 데이터의 BOW 방법과 Word2vec 방법중 Word2vec 방법의 RF의 정확도가 81.43%로 가장 높았다. BOW 방법에서는 SVM이 73.91로 가장 높았다. BOW방법의 DT가 68.69%로 가장 낮은 정확도를 나타냈다. Laptop 데이터에서 모든 분류기가 BOW방법보다 Word2vec 방법이 높은 정확도를 나타냈다. Restaurants 데이터의 BOW 방법과 Word2vec 방법에서 RF기 82.59%로 가장 높은 정확도를 나타냈다. BOW 방법에서는 SVM이 74.88%로 가장 높았다. BOW 방법의 DT가 73.45%로 가장 낮은 정확도를 나타냈다. Restaurants 데이터의 모든 분류기가 BOW 방법보다 Word2vec 방법이 높은 정확도를 나타냈다.

표 1. Accuracy of representation

Laptop	DT	LR	SVM	RF
BOW	68.69	73.82	**73.91**	75.00
Word2vec	76.69	77.61	78.69	**81.43**

Restaurants	DT	LR	SVM	RF
BOW	73.45	74.71	**74.88**	74.00
Word2vec	78.92	79.42	79.15	**82.59**

5. 결론

본 연구는 오피니언 마이닝을 위한 효율적인 감성 분류 방법을 제안하는 것을 목적으로 한다. 본 연구의 목적을 위해 SemEval 2014데이터 셋의 Laptop 도메인과 Restaurants 도메인에 대해 BOW 방법과 Word2vec 방법을 이용해 속성 데이터 셋을 구성했다. 구성된 데이터 셋에 분류기를 사용해 감성분류를 수행했다. 연구 결과는 다음과 같다. Laptop 도메인과 Restaurants 도메인에서 동일하게 Word2vec 방법과 RF분류기를 사용한 정확도가 가장 높았다. 그리고 전반적으로는 BOW 모델을 사용했을 때보다 Word2vec 방법이 모든 분류기에서 높은 정확도를 나타냈다. 이것은 BOW 모델보다 Word2vec 모델이 텍스트의 단어 사이의 의미론적 관계가 잘 반영된 것으로 볼 수 있다. 본 연구에서는 RF분류기가 가장 높은 정확도는 나타냈다. 이것은 텍스트 감성분류에서 Word2vec 방법과 RF분류기가 가장 높은 성능을 가진 조합으로 볼 수 있다.

본 연구의 한계는 다음과 같다. 연구에 사용된 데이터는 Laptop 도메인과 Restaurants 도메인이기 때문에, 다른 도메인에 적용하는데 문제가 발생할 수 있다. 다른 영역의 도메인에 대해서도 감성분류를 위한 방법을 연구할 필요가 있다. 그리고 본 연구에서 사용한 데이터는 영어 데이터이므로 한국어 데이터에 적용하기 위해서는 추가적인 연구가 필요하다.

Acknowledgment: 이 성과는 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2017R1A2B4010956).

■ 참 고 문 헌 ■

[1] Kang, M., Ahn, J., & Lee, K. "Opinion mining using ensemble text hidden markov models for text classification", Expert Systems with Applications, Vol. 94, pp. 218-227, 2018.

[2] Mantyla, M. V., Graziotin, D., & Kuutila, M. "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers", Computer Science Review, Vol. 27, pp. 16-32, 2018.

[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013.

[4] Poria, S., Cambria, E., & Gelbukh, A. "Aspect extraction for opinion mining with a deep convolutional neural network", Knowledge-Based Systems, Vol. 108, pp. 42-49, 2016.