Journal of the Korean Institute of Industrial Engineers
Published Online, pp. 238-00, October 2017.
ISSN 1225-0988 | EISSN 2234-6457

<Original Research Paper>

워드임베딩과 그래프 기반 준지도학습을 통한 한국어 어휘 감성 점수 산출

서덕성 · 모경현 · 박재선 · 이기창 · 강필성†

고려대학교 산업경영공학부

Word Sentiment Score Evaluation based on Graph-Based Semi-Supervised Learning and Word Embedding

Deokseong Seo · Kyoung Hyun Mo · Jaesun Park · Gichang Lee · Pilsung Kang

School of Industrial Management Engineering, Korea University

Sentiment analysis plays an important role in both public and private sectors to understand consumers' responses to products or voters' reactions to policies. One of the most key success factors of sentiment analysis is to build an appropriate sentiment word dictionary. Many current existing approaches either heavily rely on the knowledge of domain experts or word co-occurrence statistics, the first of which causes low efficiency and high expenditure while the second of which suffers from incomplete data. In order to resolve these shortcomings, we propose a new domain-specific Korean word sentiment score evaluation method based on word embedding and graph based semi-supervised learning. First, words are embedded in a lower dimensional space by Word2Vec technique. Then, the word relation graph is constructed based on the similarity between words in the embedding space. Then, we assign sentiments to approximately 1% words utilizing some indicators like centrality measure. The sentiment scores of the other unlabeled words are automatically assigned by label propagation with semi-supervised learning. To verify our proposed method, we collect 1.98 million review comments from three movie review websites. Experimental results show that the proposed method achieves about 93% accuracy of polarity classification.

Keywords: Sentiment Analysis, Word Embedding, Graph-Based Semi-Supervised Learning, Label Propagation, Lasso Regression, Movie Review Evaluation

1. 서 론

감성분석(sentiment analysis)은 화자들이 특정 상품이나 사건, 인물에 대해 어떤 감정이나 의견을 가지고 있는지 인식하고 분류하는 것을 목적으로 하는 자연어 처리 및 텍스트 마이닝 분야의 대표 연구 분야이다(Liu, 2012; Jeong, 2016). 최근 제품과 서비스에 대한 고객의 반응을 살피고자 하는 기업들과 국민 여론 확인을 목적으로 하는 정부를 중심으로 감성분석에 대한 수요가 기하급수적으로 증가하고 있는 것으로 나타났다(Jo, 2012). 감성사전은 극성(polarity), 주관성(subjectivity) 등 감성을 지닌 단어들인 감성어휘(sentimental lexicon)들의 모음으로서 감성분석의 정확도를 결정짓는 핵심적인 요소이다. 이러한 중요성으로 인해 지금까지 감성사전을 구축하기 위한 다양한 방법론들이 제시되었다(Cambria, 2013). 연구 초기에는 연구자의 직관에 의존하는 사전들이 제안되었으며(Baccianella *et al.*, 2010), 이후엔 단어 빈도수를 활용한 통계적인 방식이 주

이 논문은 2016년도 정부(미래창조과학부 및 교육부)의 재원으로 한국연구재단과 정보통신기술진흥센터의 지원을 받아 수행된 기초연구사업임(No. 2015R1A2A2A04007359, NRF-2016R1D1A1B03930729, No. 2017-0-00349, QoE 정보를 이용한 머신러닝 미디어 스트리밍 최적화 시스템 개발).
† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3383, Fax : 02-929-5888, E-mail: pilsung_kang@korea.ac.kr
2017년 2월 7일 접수; 2017년 5월 23일 수정본 접수; 2017년 7월 11일 게재 확정.

류를 이루었다(Turney, 2002). 그러나 최근에는 대량의 문서 집합을 대상으로 기계학습 방법론을 적용하는 시도가 큰 흐름을 이루고 있는 추세이다(Rao *et al.*, 2009). 지금까지 연구된 대부분의 방법론들은 범용적으로 활용되는 사전 구축을 목표로 하는 관계로 분야(domain)에 따라서는 감성 분석의 정확도가 상대적으로 높지 않은 단점이 존재할 뿐만 아니라, 영어를 대상으로 한 연구들이 대다수를 차지하여 영어와는 문법, 의미 등이 매우 상이한 한국어에 바로 적용하기에는 상당한 어려움이 존재한다. Oh(2014)는 '두려움'이라는 보편적 감정도 한국, 미국 등 개별 문화권마다 독특한 의미 구조를 지닌다는 점을 논의하였다.

범용적인 감성사전 구축이 현실적으로 어려운 이유는 같은 단어라도 분야에 따라서 그 의미와 감정이 매우 다르게 사용되는 경우가 많기 때문이다. 예를 들어 "졸립다"라는 표현이 영화평에 등장했다면 매우 부정적인 뉘앙스를 풍기지만, 침대 제품평에 등장했다면 반대로 긍정적으로 사용되었을 가능성이 높다. 최근 자연어처리 분야에서는 이러한 단어의 문맥적인 의미를 보존하면서 기존의 bag-of-words 방식의 단점인 고차원(high dimensionality) 문제점을 해결하기 위해 다양한 방식의 분산 표상(distributed representation) 기법들이 제안되었다. bag-of-words 방식에서는 전체 단어 수만큼의 차원을 갖는 벡터에서 특정 단어의 값만 1을 갖고 나머지는 모두 0의 값을 갖는 one-hot-encoding 기법을 사용하여 단어를 표현한다. 이러한 방식은 하나의 단어를 표현하기 위해 문서 집합(corpus)에 존재하는 수만큼의 차원을 갖는 벡터가 필요할 뿐만 아니라 단어들이 모두 독립성을 갖는 것으로 취급되어 문맥상의 의미를 전혀 보존하지 못하는 문제점이 존재한다. 반면에 분산 표상에서는 하나의 단어가 미리 정의된 차원(예 : 100차원)에서 연속형의 값을 갖는 벡터로 표현(임베딩, embedding)되는데 이 벡터는 다른 단어들과 문맥상의 의미가 최대한 보존될 수 있도록 알고리즘에 의해 학습이 된다는 특징을 가지고 있다(Mikolov, Chen *et al.*, 2013). 따라서 분산 표상은 어떠한 분야에서 사용되는 문서들을 대상으로 학습했는가에 따라서 동일한 단어가 다른 벡터로 표현될 수 있다.

본 연구에서는 이러한 분산 표상의 특징을 활용하여 분야별로 최적화될 수 있는 한글 감성어휘 점수 산출 방법론을 제안하고자 한다. 우선 대상이 되는 분야에서 충분히 많은 수의 문서를 수집한 뒤 대표적인 분산 표상 기법인 Word2Vec 기법을 활용하여 저차원 공간에 임베딩을 수행한다. 임베딩된 공간은 해당 어휘들의 문맥적 의미를 반영하여 생성된 것이므로 특정 단어가 매우 긍정적이라고 하면 해당 단어와 가까운 위치에 존재하는 단어들 역시 긍정적 의미를 가질 가능성이 높다고 유추해 볼 수 있다. 따라서 본 연구에서는 임베딩된 공간에서의 단어간 유사도를 고려한 네트워크를 구축하여 단어들 간의 관계를 표현하였다. 이렇게 구축된 단어 네트워크에 그래프 기반의 준지도학습(graph-based semi-supervised learning)을 적용하여 개별 단어의 감성 점수를 산출하는 방법론을 개발하였다. 준지도학습이란 입력 변수와 종속 변수가 존재하는 관측치들만을 이용하여 예측 모델을 구축하는 지도학습(supervised learning)과는 달리 입력변수만 존재하는 관측치들도 학습 과정에 포함시켜 예측 모델의 성능을 향상시키고자 하는 대표적인 기계학습 기법이다(Zhu *et al.*, 2009). 입력변수는 충분히 확보할 수 있으나 출력 변수를 확보하는 것이 어려운 상황(이미지의 개체명 등)에서 준지도학습은 상당한 성능 개선 효과를 나타내는 것으로 알려져 있다(Chapelle *et al.*, 2006). 이러한 준지도학습 개념은 감성점수 산출에 유용하게 적용될 수 있다. 감성사전을 구축하기 어려운 이유는 매우 명백하게 긍정이거나 부정인 어휘는 쉽게 판별할 수 있는 반면에 중립에 가깝거나 상황에 따라 긍정적 또는 부정적 의미를 갖는 어휘들에 대한 감정 점수를 부여하는 것이 까다롭다는데 있다. 본 연구에서는 특정 분야에서 누구나 동의하는 긍정 어휘 집합과 부정 어휘 집합을 생성한 뒤, 이 어휘들과 긍정 혹은 부정이 할당되지 않는 다른 어휘들이 단어 네트워크에서 갖는 관계를 고려하여 그래프 기반 준지도학습을 통해 모든 어휘들에 대해 긍정 혹은 부정의 정도를 산출하는 방식을 제안한다. 제안하는 방법론의 효과를 검증하기 위하여 영화 리뷰 분야에 대한 감성 어휘 점수를 산출하였다. 다음(Daum) 등 3개 영화 리뷰 사이트에서 얻은 198만 개 영화평을 대상으로 Word2Vec을 적용하여 저차원으로 임베딩을 실시하였다. 이렇게 임베딩된 공간에서 단어 네트워크를 구축한 뒤 긍정/부정이 명백한 50개의 단어를 초기 표식(label)으로 정의하고 준지도학습을 통해 모든 단어에 긍정/부정에 대한 연속형 점수를 할당하였다. 할당된 긍정/부정 점수의 타당성은 실제 사용자가 부여한 평점과 해당 영화평에 사용된 어휘들의 긍정/부정 점수를 결합하여 추정된 평점을 비교하여 검증하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 감성사전 구축과 관련한 선행연구를 소개한다. 제 3장에서는 연구에서 사용된 Word2Vec 및 그래프 기반 준지도학습 방법론을 간략히 소개한다. 제 4장에서는 제안하는 방법론을 검증하기 위한 실험 설계에 대해 서술하고 제 5장에서 실험 결과를 분석하여 제안하는 방법론의 타당성을 검증한다. 마지막으로 제 6장에서 본 연구의 결론과 시사점을 서술한다.

2. 선행연구

감성사전 구축을 위한 대표적인 접근 방법론들은 크게 연구자 직관 중심의 감성사전 구축, 통계적 방식 기반 감성사전 구축, 기계학습 모델 기반 감성사전 구축의 세 가지로 구분될 수 있다. 우선 연구자 직관 중심의 감성사전 구축 방법론을 소개한다. Kim(2002)은 한국어 정서자동사를 안심, 만족, 흥분, 초조, 탄식, 감각, 즐거움, 불쾌함 등 22개로 분류했다. Kim *et al.*(2013)는 총 7,744개 문장의 감정표현을 주석한 '한국어 감정 및 의견 분석 코퍼스'를 직접 구축했다. 영어에선 Senti-WordNet(Baccianella

et al., 2010)이 대표적인 감성사전으로 각 어휘에 긍정, 중립, 부정 등 극성을 태깅했다. 연구자가 자연언어를 언어학 기반의 방법론으로 직접 분석한다는 점에서 상대적으로 정확하고 엄밀한 결과를 도출할 수 있다. 다만 이 방법론은 수작업에 크게 의존하기 때문에 사전 구축에 비용이 많이 들고, 단어 선정 시 연구자마다 다른 기준이 적용될 가능성이 있으며, 코퍼스에 없는 단어에 극성을 부여하기 어렵다.

통계적 접근으로는 대표적으로 Turney(2002)가 있다. Turney (2002)는 자동차, 영화 리뷰 등에서 명백한 감성을 지니는 단어 ("excellent", "poor" 등)과 동시에 출현하는 공기(co-occurrence) 어휘들을 추출해 이들의 감성을 추론했다. Beineke *et al*.(2004)는 감성단어를 추출하기 위해 나이브 베이즈 모델 등을 활용했다. Kaji *et al*.(2007)은 극성을 지닌 문장을 추려낸 뒤 이 문장에 등장한 단어의 빈도를 바탕으로 각 단어의 극성값을 계산하였다. 통계 기반의 방법론은 단어의 극성값을 정량적으로 얻을 수 있지만 특정 단어가 한 문장이나 문서 내에서 동시에 출현하는 빈도가 낮기 때문에 분석 대상 코퍼스 규모가 작을 경우 정확한 분석이 쉽지 않다는 특성을 갖고 있다.

최근엔 기계학습을 활용한 방법론이 주목받고 있다. 이 가운데 그래프 기반의 연구 결과가 속속 소개되고 있다. Rao *et al*. (2009)는 WordNet 등 영어 어휘의 의미 관계를 정리해 놓은 데이터를 바탕으로 그래프를 구축하고 Label Propagation 방법론으로 극성값을 전파했다. Li *et al*.(2012)는 기존 감성사전을 참고해 분야에 관계없이 포괄적으로 쓰일 수 있는 감성단어를 뽑고, 이를 바탕으로 Relational Adaptive Bootstrapping(RAP)라는 알고리즘을 정의하여 목표 분야의 감성사전을 구축하였다. 하지만 앞의 두 방법론은 연구자 직관 중심의 감성사전 구축 방법론을 일부 차용한 방식으로, 단어 의미관계를 정성적으로 구축해 놓은 데이터를 사전에 확보하지 못할 경우 그 적용이 쉽지 않다. Kim *et al*.(2015)는 영화 리뷰의 공기(co-occurrence) 정보를 토대로 단어 간 유사도를 계산한 뒤 LP를 활용해 극성값을 전파했다. 이 방법론은 WordNet 같은 데이터를 사전에 필요로 하지 않는다는 장점이 있지만, 단어 출현 순서까지 고려한 Word2vec 기법을 활용해 단어 간 유사도를 도출한 본 연구와 차이가 있다.

3. 방법론

3.1 Word2Vec

Word2Vec(Mikolov, Chen *et al*., 2013)은 전방 전달 신경망(feedforward neural network) 기반의 언어 모델(language model)을 활용한 대표적인 단어 임베딩 방법론이다. Word2Vec을 학습하기 위한 대표적 신경망 구조로는 <Figure 1>과 같이 continuous bag-of-words(CBOW) 구조와 continuous skip-gram(Skip-gram) 구조 두 가지가 있다(Mikolov, Sutskever *et al*., 2013). CBOW 방식에서는 문맥을 이용해 단어를 예측하는 과정에서 단어의 임베딩이 수행되며, Skip-gram 방식에서는 단어를 이용해 문맥을 예측하는 과정에서 단어의 임베딩이 수행된다.

CBOW의 단어 임베딩 절차는 다음과 같다. 먼저 문서 집합의 각 단어 $w(t)$를 one-hot-encoding으로 벡터화하고, 1개의 은닉층 h_j을 갖는 신경망을 구성한다. 이때, 은닉층의 노드 개수 d가 벡터화 하고자 하는 차원의 수가 된다. $w(t)$의 앞, 뒤 각각 k개의 단어를 입력층에, $w(t)$를 출력층에 둔다. 여기서 k는 Word2Vec 학습을 위해 사용자가 지정해야 하는 하이퍼 파라미터(hyperparameter)인 윈도우 크기(windows size)이다. 이때 특이한 점은 입력층의 각 단어와 은닉층을 이어주는 가중치 행렬 *W(V by d)*를 공유한다는 사실이다. 학습이 완료된 W가 바로 V개의 각 단어를 d차원 좌표에 임베딩한 결과물이다. Skip-gram은 입력층과 출력층만 다를 뿐 나머지 모든 절차는 CBOW와 같다.

3.2 단어 네트워크 생성

단어 네트워크는 각 단어를 노드(node)로 정의하고 두 단어가

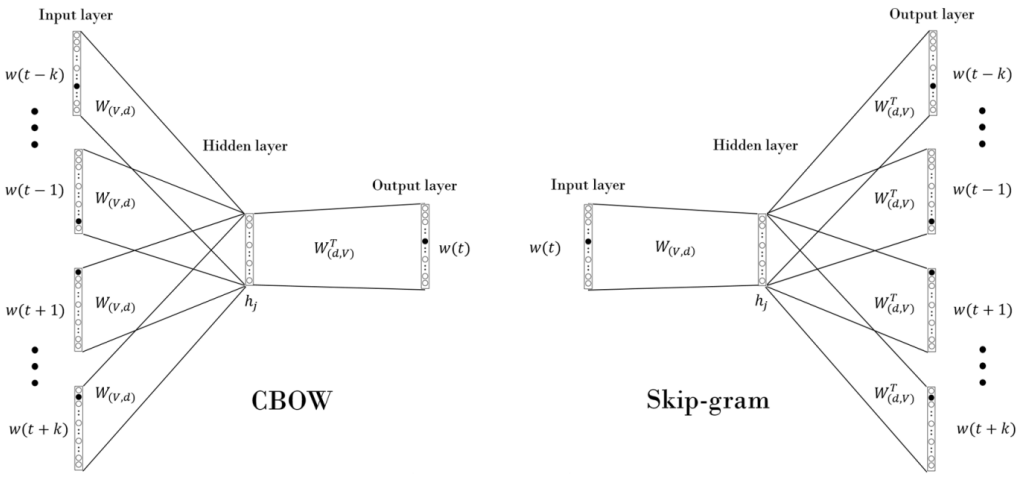


Figure 1. The Architecture of CBOW & Skip-gram(Seo *et al*., 2017)

충분히 유사할 경우 두 노드를 연결해주는 호(arc, edge)가 존재하는 그래프이다. 임베딩된 단어 공간에서 네트워크를 생성하는 것은 결국 어떤 기준으로 호를 연결해줄 것인가의 문제로 귀결된다. 호를 연결하는 대표적인 두 가지 방식으로는 ϵ-neighborhood 방식과 k-nearest neighbor 방식이 있다. ϵ-neighborhood 방식은 사전에 정의된 ϵ보다 작은 거리를 가지는 노드 쌍만을 연결하는 것이며, k-nearest neighbor 방식은 특정 노드와 가까운 k개 이웃을 연결하는 방식이다(Von Luxburg, 2007). ϵ-neighborhood 방식은 데이터의 분포와 밀도를 보다 정확하게 반영할 수 있는 장점이 있는 반면 연결 그래프(그래프의 임의의 두 노드를 연결할 수 있는 호의 집합이 존재하는 그래프)를 보장해주지 못하는 단점이 있다. 이에 반하여 k-nearest neighbor 방식은 연결 그래프는 보장을 하나 데이터의 분포와 밀도가 제대로 반영되지 못하는 단점이 존재한다. 그래프 기반 준지도학습에서 연결 그래프는 매우 중요한 필요조건이기 때문에 본 연구에서는 기본적으로 ϵ-neighborhood 방식을 사용하되, 연결이 되지 못하는 노드들에 대해서 1개 이상의 다른 노드와 연결될 수 있도록 최소 신장 나무(minimum spanning tree, MST)를 보완책으로 활용하였다.

3.3 그래프 기반 준지도학습

준지도학습은 출력 변수 값이 존재하는 데이터(labeled data)의 입력 변수 및 출력 변수와 출력 변수 값이 존재하지 않는 데이터(unlabeled data)의 입력 변수를 함께 사용하여 존재하지 않는 출력변수 Y_u를 예측하는 방법론이다. 본 연구에서는 명백한 긍정/부정이 존재하는 단어들을 labeled data로 정의하고 나머지 단어들을 unlabeled data로 정의한 뒤, 제 3.2절에서 제시된 방식으로 단어 간 네트워크를 생성하고 이 네트워크에 그래프 기반 준지도학습인 Label propagation과 Label spreading을 적용하였다.

Label propagation과 Label spreading의 수행 절차는 다음과 같다. 우선 제 3.2절에서 생성된 단어 네트워크에서 연결된 두 노드 간의 거리를 계산한 뒤, 이를 기반으로 다음과 같이 Radial basis function(RBF) kernel을 사용하여 두 노드 사이의 거리가 가까울수록 큰 가중치를 부여한다.

$$w_{ij} = \exp\left(-\frac{d(x_i,x_j)^2}{2\sigma^2}\right) \text{ where node } i \text{ and node } j \text{ adjacent} \quad (1)$$

$d(x_i, x_j)$: distance between node i and node j

σ : hyperparameter

이를 바탕으로 노드 i의 출력 변수 추정 값 $\hat{y}_i$는 다음과 같이 계산되며, 추정값들이 지속적으로 변화하므로 추정 값들이 수렴할 때까지 반복적인 추정을 수행하게 되며, Label propagation과 Label spreading은 이 반복을 수행하는 방식에서 차이가 존재한다.

$$\hat{y}_i = \frac{\Sigma_j w_{ij}\hat{y}_j}{\Sigma_j w_{ij}} \quad (2)$$

Label propagation 알고리즘은 <Figure 2>에 나타난 바와 같으며 학습을 충분히 반복하면 Y_u는 초기값에 상관없이 수렴한다는 사실이 증명되어 있다(Zhu, 2002; Chapelle *et al.*, 2006). Label spreading은 <Figure 3>과 같이 Label propagation의 지역적 비중을 그래프 전체 구조로 옮긴 알고리즘이며, α는 label의 고정 비율을 조절하는 하이퍼 파라미터이다(Zhou *et al.*, 2004; Chapelle *et al.*, 2006). 본 연구에서는 두 알고리즘을 통해 추정된 $\hat{Y}^{(\infty)}$를 각각 해당 단어의 긍정/부정 점수로 사용하였다.

Algorithm 1 Label Propagation

1. Compute the diagonal degree matrix D by $D_{ii} \leftarrow \Sigma_j w_{ij}$
2. Initialize $\hat{Y}^{(0)} \leftarrow (y_1, \cdots, y_l, 0, 0, \cdots, 0)$
3. Iterate
   Ⅰ. $\hat{Y}^{(t+1)} \leftarrow D^{-1}W\hat{Y}^{(t)}$
   Ⅱ. $\hat{Y}_1^{(t+1)} \leftarrow \hat{Y}_1^{(0)}$
4. Until convergence to $\hat{Y}^{(\infty)}$

Figure 2. Label Propagation Algorithm

Algorithm 2 Label Spreading

1. Compute the diagonal degree matrix D by $D_{ii} \leftarrow \Sigma_j w_{ij}$
2. Compute the normalized graph Laplacian $\mathcal{L} \leftarrow D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$
3. Initialize $\hat{Y}^{(0)} \leftarrow (y_1, \cdots, y_l, 0, 0, \cdots, 0)$
4. Iterate
   Ⅰ. $\hat{Y}^{(t+1)} \leftarrow \alpha\mathcal{L}\hat{Y}^{(t)} + (1-\alpha)\hat{Y}^{(0)}$
5. Until convergence to $\hat{Y}^{(\infty)}$

Figure 3. Label Spreading Algorithm

4. 실험설계

본 연구는 <Figure 4>에 나타난 절차를 따라 영화 분야에서의 한국어 감성 어휘 점수를 산출하고 이의 적절성을 평가하였다. 영화 리뷰 사이트에서 영화평 텍스트 데이터와 평점 데이터를 수집하고 전처리한 뒤 각 단어를 Word2Vec을 이용해 저차원으로 임베딩하였다. 각 단어를 그래프의 노드로 정의하고 임베딩된 공간에서의 거리 기반 단어 네트워크를 구축한 뒤 그래프 기반 준지도학습을 이용하여 단어의 긍정/부정 점수를 산출하였다. 산출된 단어의 긍정/부정 점수를 바탕으로 영화평에 대한 평점을 예측하고 실제 평점과의 비교를 통해 제안하는 방법론의 효과를 검증하였다. 실험 환경은 CPU i5, RAM 32GB이며 R과 Python을 이용하여 실험을 진행하였다.

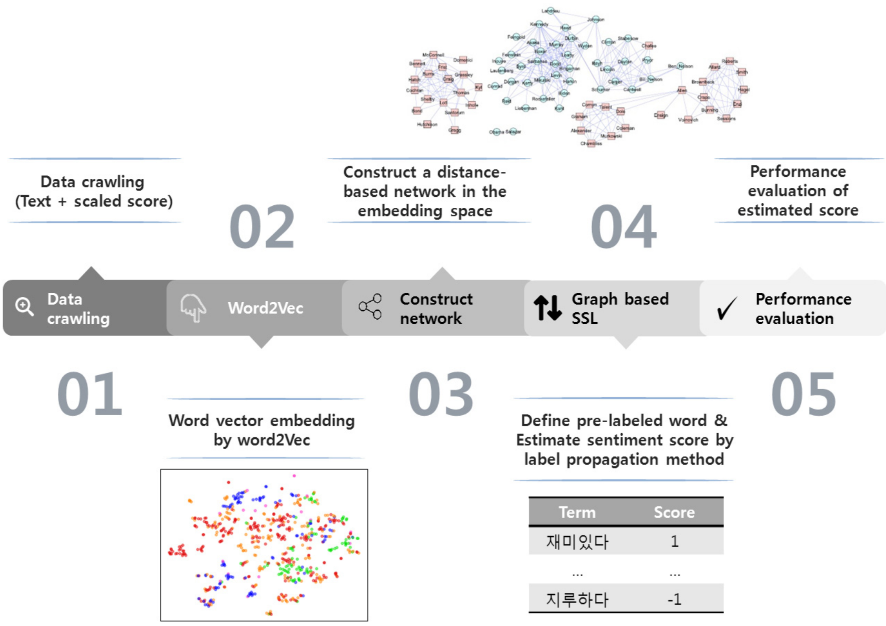


Figure 4. Framework of Proposed Method

4.1 데이터 수집 및 전처리

본 연구에서는 다음(Daum) 영화, 롯데시네마 홈페이지, 메가박스 홈페이지의 세 영화 관련 사이트에서 총 2,041,334개의 영화평 데이터를 수집하였다. 각 영화에 대한 영화평 본문과 평점을 포함하여 사이트 이름, 영화 제목, 영화 고유 ID, 영화평 작성자의 고유 ID, 날짜까지 7개의 정보를 수집하였다. 영화 평점의 경우 사이트마다 범위가 다른 관계로 최저 1점, 최고 5점을 갖도록 점수의 범위를 조정하였다. 각 사이트 별 리뷰의 개수와 영화 평점의 범위는 <Table 1>과 같다.

Table 1. Information about Each Movie Review Site

Movie review site	The number of Movies	The number of reviews	The range of review score
Daum movie	26,143	1,497,462	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
Lotte cinema	2,043	476,401	{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
Megabox	717	67,471	{1, 2, 3, 4, 5}

4.2 형태소 분석(part-of-speech tagging)

본 연구에서는 R의 KoNLPQ 라이브러리에 포함된 twitter 형태소 분석기를 사용하였다. 대표적인 한글 형태소 분석기인 꼬꼬마 형태소 분석기(http://kkma.snu.ac.kr/documents)나 한나눔 형태소 분석기(http://semanticweb.kaist.ac.kr/hannanum)의 경우 정제된 언어가 사용되지 않는 문서에 대한 형태소 분석 정확도가 높지 않은 문제점이 있다. 본 연구에서 사용한 영화평 데이터는 띄어쓰기 원칙이 지켜지지 않고 이모티콘이나 비표준어, 비속어 등이 많이 포함되어 있는 정제되지 않은 언어이며 이러한 데이터에 대해서는 인터넷 용어 분석에 강점을 갖는 twitter 분석기가 보다 적합하다고 할 수 있다(Kim *et al*., 2017).

<Table 2>는 꼬꼬마 형태소 분석기와 본 연구에서 사용한 twitter 분석기가 형태소 분석한 결과를 비교한 예시이다. 띄어쓰기를 전혀 하지 않은 글에 대해 꼬꼬마 형태소 분석기는 무한루프에 빠져 오류가 나는 반면 twitter 분석기는 적절히 분해한다. 또한 twitter 분석기는 배우 이름 등의 고유명사 분석 성능이 비교적 좋다는 점을 확인할 수 있다.

Table 2. Examples of Kkma and KoNLPQ Pos tagger

Example of movie review	Kkma(python)	KoNLPQ(R)
굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿 굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿 굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿 굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿 굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿 굿굿굿굿굿굿굿굿굿굿굿굿굿굿굿	(무한루프)	굿굿*/Noun
이성재와 배두나를 보는것만해도 즐거운 영화 ㅎ	(‘이성’, ‘NNG’), (‘재와’, ‘MAG’), (‘배‘, ‘NNG’), (‘두’, ‘JX’), (‘나’, ‘NP’), (‘를’, ‘JKO’), (‘보’, ‘VV’), (‘는’, ‘ETD’), (‘것’, ‘NNB’), (‘만’, ‘JX’), (‘하’, ‘VV’), (‘어도’, ‘ECD’), (‘즐겁’, ‘VA’), (‘ㄴ’, ‘ETD’), (‘영화’, ‘NNG’), (‘ㅎ’, ‘NNG’)	이성재/Noun 와/Josa 배두나/Noun 를/Josa 보다/Verb 만해/Noun 도/Josa 즐겁다/Adjective 영화/Noun ㅎ/KoreanParticle

본 연구에서는 한국어 어휘에 대한 감성 점수 산출이 목표이므로 twitter를 통해 얻어진 형태소 중 알파벳, 숫자, 구두점, 외국어, 한자 및 기타기호의 네 가지 품사는 분석에서 제외하였다. <Table 3>은 수집된 영화평 수와 위 규칙에 따라 제거된 영화평 수를 정리한 표이다.

Table 3. The Number of Reviews Before and After Preprocessing

Movie review site	The number of total reviews	Reviews excluding Korean morpheme	Reviews including Korean morpheme
Daum movie	1,497,462	51,910	1,445,552
Lotte cinema	476,401	9,322	467,079
Megabox	67,471	494	66,977

4.3 단어 임베딩

다음 영화 등 3개 영화 리뷰 사이트에서 수집한 데이터에 대해 Python의 Gensim 모듈이 제공하는 Word2Vec을 이용하여 100차원 공간에 임베딩을 실시하였다. CBOW와 Skip-gram 두 방식을 모두 사용했으며 윈도우 크기는 2, 신경망 반복 횟수(epoch)는 50번을 적용하여 실험을 진행하였다. 데이터 내의 총 단어 수는 132,395개이며 이 중 출현 빈도(term frequency)가 100 이상인 단어 10,215개만 사용하여 임베딩을 수행하였으며 그 결과의 일부는 Table 4에 나타난 바와 같다.

Table 4. Coordinates of Words in 100 Dimensions

Word	d_1	d_2	⋯	d_{100}
명작	0.234	0.891	⋯	0.963
아쉽다	0.477	0.443	⋯	0.564
⋯	⋯	⋯	⋯	⋯
연기력	0.119	0.863	⋯	0.787

4.4 임베딩된 공간에서의 거리기반 단어 네트워크 구축

각 단어 벡터를 노드로, 단어 간의 거리를 통해 구한 가중치를 호로 하는 그래프를 구축하였다. 가중치는 제 3.3절의 식 (1)을 기본으로 하여 σ는 1로 설정한 다음 식을 사용하였다.

$$w_{ij} = \exp\left(-\frac{d(x_i,x_j)^2}{2}\right) \text{ where node } i \text{ and node } j \text{ adjacent} \quad (3)$$

이후 ϵ-neighborhood 방식과 최소 신장 나무 기법을 적용하여 완전 그래프에 비해 가벼운 그래프를 구축하되 고립된 노드가 생기지 않도록 하였다. 이렇게 구축된 그래프는 그래프 기반 준지도학습 알고리즘 적용시 labeled data의 출력변수 값 Y_1이 모든 단어에 전파될 수 있게 해준다. 본 연구에서는 단어 간 거리의 제 3, 4, 5, 10, 15, 20, 30, 40 분위수를 ϵ로 설정하고 실험을 수행하였다.

4.5 명확한 긍정 및 부정 어휘 집합 생성

본 연구에서는 그래프 기반 준지도학습을 수행하기 위해 label이 존재하는 초기 노드로 긍정과 부정이 비교적 명확한 단어들을 선정하였으며, 이를 pre-labeled word라 정의하였다. 총 550개의 긍정 및 부정 단어 집합들 중 긍정 단어로는 "재밌다", "최고", "명작" 등 331개가 선정되었으며 부정 단어로는 "지루함", "졸리다", "애매" 등 219개가 선정되었다. <Table 5>는 pre-labeled word의 일부를 나타낸 것이다.

Table 5. Samples of Positive and Negative Pre-Labeled Words

Positive pre-labeled words			Negative pre-labeled words		
좋다	역시	연기력	억지	지루	때우다
훌륭하다	즐겁다	아름답다	아쉬움	약하다	부담
재밌다	기대하다	긴장감	절대	흠	부끄럽다
사랑	좋아하다	매력	짜증	진부	고통
재미있다	멋지다	짱	지루함	싫다	난해
감동	대박	빠지다	밉다	별루	거슬리다
기대	웃다	멋있다	이상하다	심하다	망치다
최고	눈물	울다	킬링타임	식상하다	애매
재미	추천	코믹	필요없다	까다	거지
괜찮다	슬프다	오랜만	어이	졸리다	질질

4.6 Label Propagation을 통한 단어 감성 전파

550개의 Pre-labeled word 중에서 100개를 그래프 기반 준지도학습에 사용하여 10,215개의 단어의 감성사전을 구축하였으며 나머지 450개는 성능 평가에만 사용하였다. 학습 단어를 선정한 기준은 단어 네트워크에서 노드간 가중치와 노드 출현 빈도를 모두 고려한 중심성 지표(식 (4) 참조)가 가장 높은 긍정 및 부정 상위 50개이다.

$$Centrality_i = (\Sigma_j w_{ij}) \times Frequency_i \quad (4)$$

Word2Vec으로 임베딩한 단어가 총 10,215개라는 점을 고려하면 학습에 사용된 pre-labeled word는 전체의 0.98%에 해당하는 매우 낮은 비율이라고 할 수 있다. 학습 및 성능평가에 사용한 pre-labeled word의 개수는 <Table 6>과 같다.

Table 6. The Number of Pre-Labeled Words of Learning and Testing

Sentiment	Training	Evaluation	Total
Positive	50	281	331
Negative	50	169	219

사전에 긍정 및 부정이 미리 정의된 100개의 pre-labeled word의 정보를 바탕으로 그래프기반 준지도학습을 수행하면 긍정 및 부정이 정의되지 않은 나머지 10,115개 단어에 대한 긍정

혹은 부정의 정도를 산출하게 된다. 이것이 해당 단어의 감성 점수이며 -1부터 +1사이의 실수 값을 갖는다. 감성 점수는 -1에 가까울수록 부정적이며 +1에 가까울수록 긍정적인 의미를 갖는 단어라는 것을 의미한다. 실험에는 Python의 sklearn 모듈이 제공하는 Label Propagation을 사용했다.

4.7 파라미터 탐색 범위 및 성능 평가 지표

본 연구의 절차를 수행하기 위한 파라미터는 <Table 7>에 나타난 바와 같이 총 다섯 가지이며, 본 실험에서는 각 파라미터별로 최소 2개, 최대 8개의 후보군에 대한 실험을 수행하였다. 따라서 총 128가지의 조합이 탐색되었으며, 각 파라미터에 따른 성능의 민감도를 분석하였다.

Table 7. Parameter Search Space

No.	Parameter	Search space
1	Word2Vec	CBOW, Skip-gram
2	Embedding dimension	50, 100
3	Distance measure	Euclidean distance, Angular distance
4	Graph based SSL	Label propagation, Label spreading
5	ϵ(percentile of distance)	3, 4, 5, 10, 15, 20, 30, 40

본 연구에서 제안하는 감성 어휘 산출 방법론에 대해 두 가지 관점에서 성능을 평가하였다. 우선 학습에 사용하지 않은 Pre-labeled word의 긍정/부정을 정답으로 두고 제 4.6절에서 추정한 해당 어휘의 감정 점수의 부호를 사용하여 예측한 긍정/부정이 정답과 얼마나 일치하는지를 통해 성능을 평가하였다. 비교 대상으로는 감성사전 구축시 사용되는 지도학습 방법론 중 대표적인 벌점화 회귀분석 알고리즘인 라쏘 회귀분석(Lasso regression)을 사용하였다(Tang *et al*., 2014; Pröllochs *et al*., 2015). 영화 평점을 긍정, 부정으로 나누어 라쏘 분류문제로 감성사전을 구축할 경우 정보 손실이 생기므로, 라쏘 회귀분석을 진행하였다. 라쏘 회귀분석에서도 제안하는 방법론에서 다루는 10,215개의 단어에 대해 1,979,608개 리뷰 모두를 사용하여 감성사전을 구축하였다. 회귀 계수가 양수인 단어를 긍정 단어로 판별하였고, 회귀 계수가 음수인 단어를 부정 단어로 판별하였다. 라쏘 회귀분석을 통해 회귀계수가 0이 된 단어는 평가 대상에서 제외하고 혼동행렬을 도출한 뒤 성능을 확인하였다. R의 glmnet 패키지에서 제공하는 cv.glmnet 함수는 교차 검증(cross validation)을 통해 최적의 λ를 제공하며, 이를 이용하여 라쏘 회귀분석을 적합하였다.

두 번째로는 다섯 개 이상 단어로 이뤄진 1,695,005개 영화평의 실제 평점과 본 연구에서 산출된 어휘들의 감성 점수들을 토대로 예측한 평점을 비교하여 성능을 평가하였다. 예측 평점은 각 영화평의 어휘들에 산출된 감성점수를 적용하고 이를 평균 낸 값으로 정의하고 1점부터 5점으로 표준화 했다. <Table 8>은 한 리뷰에 대해 감성점수를 적용하고 평균 낸 예시이다. 실제, 예측 평점 모두 5점 척도이므로 3점 미만인 경우 부정으로 3점 이상인 경우 긍정으로 판별하는 분류 문제로 변환하여 제안된 방법론의 효과를 검증하였다. 라쏘 회귀분석을 통한 감성사전에 대해서도 같은 방식으로 적용하여 예측 평점을 얻었다. 영화 평점을 학습에 사용하는 라쏘 회귀분석의 경우에는 일반화 성능을 확인하기 위해 데이터를 학습데이터 70%, 검증데이터 30%로 나누어 성능을 확인했으며, 30회 반복실험을 통해 평균과 분산을 확인하였다.

두 관점 모두 혼동행렬(confusion matrix)을 만들 수 있으며, 혼동행렬에서 성능 평가지표인 단순정확도(Accuracy, ACC), 재현율(Recall, REC), 정밀도(Precision, PRE), F1-지표(F1) 및 균형정확도(Balanced correction rate, BCR)을 계산하여 성능을 확인했다. <Table 9> 및 아래 식 (5)~식 (9)는 혼동행렬 및 각 지표를 계산하는 방법이다.

Table 9. Confusion Matrix

		Predicted label	
		Positive	Negative
Actual label	Positive	a	b
	Negative	c	d

$$ACC = \frac{a+d}{a+b+c+d} \tag{5}$$

$$REC = \frac{a}{a+b} \tag{6}$$

$$PRE = \frac{a}{a+c} \tag{7}$$

$$F1 = \frac{2\times(REC\times PRE)}{REC+PRE} \tag{8}$$

$$BCR = \sqrt{\frac{a}{a+b}\times\frac{d}{c+d}} \tag{9}$$

Table 8. Example of Applying Word Sentiment Score

									Mean
Original string	명감독의 재기발랄한 데뷔작								
POS tagging	명	감독	의	재기	발랄	한	데뷔	작	
Sentiment score	0.0027	0.0094	0.0315	0.0661	0.0894	0.0263	0.0267	0.0149	0.0334

5. 실험 결과

5.1 감정 점수 시각화

<Figure 5>는 제안하는 방법론에서 가장 좋은 성능을 보인 파라미터 조합의 실험에 대한 긍정 및 부정단어 100개(pre-labeled word 제외)씩을 추출하여 시각화한 것이다. 청색은 긍정, 적색은 부정을 의미하며 동그라미가 클수록 긍정 또는 부정의 강도가 높다. 대표적인 긍정 단어로는 “철철”, “이뿌”, “방울”, “머시”, “슬펏”, “먹먹” 등이 있으며 부정적 단어는 “안달”, “현저”, “오산”, “별루였어”, “딸리”, “고리타” 등이 있다.

5.2 단어의 감성에 대한 성능 평가

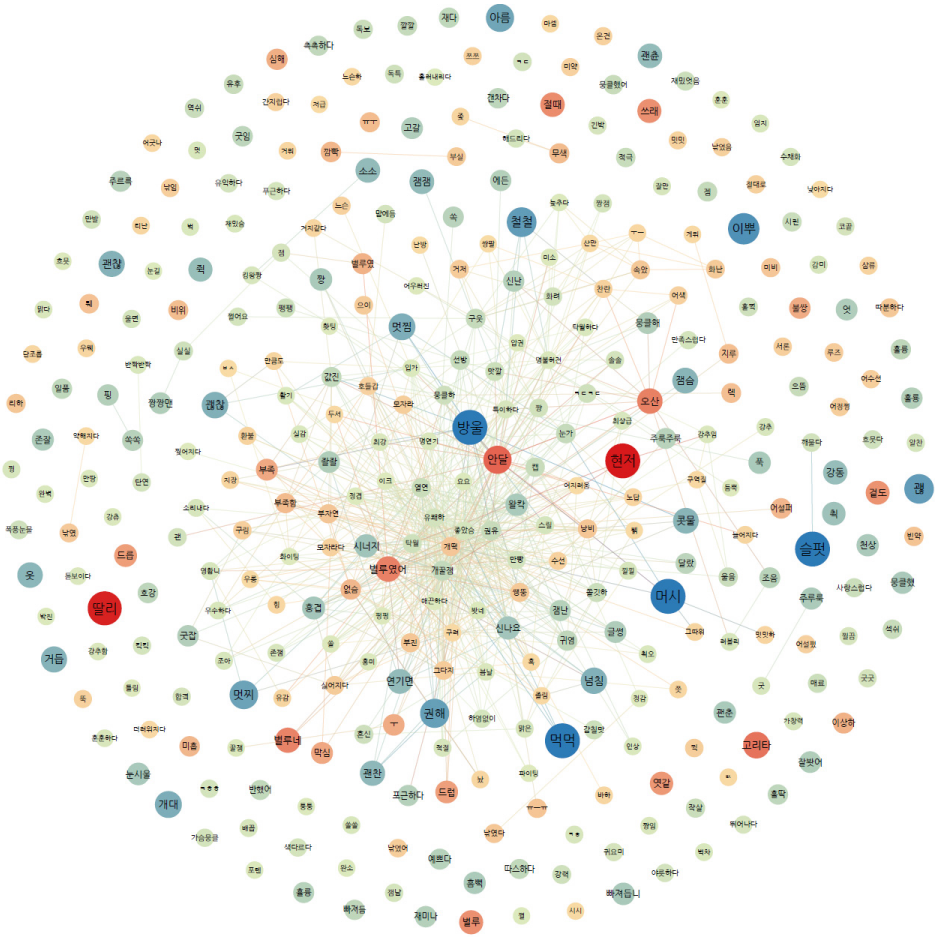
Pre-labeled word 중 전파에 사용되지 않은 총 450개 단어에 대해 예측된 감성 점수를 바탕으로 구축된 혼동행렬은 <Table 10>에 나타난 바와 같다. 이 결과는 여러 파라미터 조합 가운데 BCR 기준으로 가장 좋은 성능을 보인 파라미터 조합(Skip-gram, 100차원 임베딩, Euclidean distance, Label propagation, ϵ : 20)에 대한 결과이다. 이를 바탕으로 라쏘 회귀분석과 감정 어휘 판별 성능을 비교한 결과는 <Table 11>에 나타난 바와 같다. 제안하는 방법론은 모든 평가 지표에 대해서 최소 0.9288에서 최대 0.9596의 높은 정확도를 나타내나, 비교 대상인 라쏘 회귀분석 기반의 방식은 상대적으로 매우 부정확한 결과를 나타내는 것을 알 수 있다.

Table 10. Confusion Matrix between the Polarity of the Pre-Labeled Words and the Predicted Polarity based on the Proposed Method

		Sentiment scores of words	
		Positive	Negative
Pre-labeled word	Positive	261	20
	Negative	11	158

Table 11. Comparison between the Proposed Method and Lasso Regression

	ACC	REC	PRE	F1	BCR
Proposed method	0.9311	0.9288	0.9596	0.9439	0.9319
Lasso regression	0.6327	0.3958	0.9333	0.5558	0.6262



(parameter : Skip-gram, 100 dimensions embedding, Euclidean distance, Label propagation, ϵ : 20)

Figure 5. Visualization of Sentiment Dictionary

Table 12. Estimated Polarity Scores for the Pre-Labeled Words Whose Labels are not Utilized During the Label Propagation

Positive		Negative	
Words	Sentiment Score	Words	Sentiment Score
슬펏	1	딸리	-0.9576
먹먹	1	별루였어	-0.5458
이쁘	0.8282	겉도	-0.4673
괭찮	0.5956	드럽	-0.4140
잼잼	0.5126	부족	-0.3713
신나요	0.4357	심해	-0.3449
눈시울	0.4089	어설퍼	-0.3123
짱	0.4002	미흡	-0.3062
왈칵	0.4001	지루	-0.2728
존잘	0.3851	빈약	-0.2531

Table 13. Confusion Matrix between the Actual Polarity of Review Texts and Predicted Polarity Based on the Proposed Method

		Predicted polarity	
		Positive	Negative
Actual polarity	Positive	1,139,882	256,503
	Negative	110,963	187,657

Table 14. Comparison between the Proposed Method and Lasso Regression (In case of Lasso regression, mean (standard deviation) values for 30 trials, respectively)

	ACC	REC	PRE	F1	BCR
Proposed method	0.7832	0.8163	0.9113	0.8612	0.7162
Lasso regression	0.8201 (0.0083)	0.9998 (0.0001)	0.8202 (0.0083)	0.9011 (0.0050)	0.0128 (0.0087)

<Table 12>는 전파에 사용하지 않은 pre-labeled word 중 긍정 점수 상위 15개, 부정 점수 상위 15개 단어에 대한 예시이다. 주목해야 할 점은 긍정 점수가 매우 높게 나타난 어휘들 중 "슬펏", "먹먹", "눈시울", "왈칵"과 같은 단어는 일반적으로 슬픔을 표현하는 단어들이나 영화평에 해당 단어들이 등장하면 그 영화는 매우 긍정적으로 평가됨을 확인할 수 있다. 즉, 분야에 따라 어휘의 감성 점수가 달라질 수 있다는 본 연구의 가정이 다시 한 번 뒷받침되는 예시라고 할 수 있다.

5.3 영화 평점에 대한 성능 평가

실제 영화 평점의 긍정 및 부정에 대해 가장 우수한 성능을 나타내는 파라미터 조합(Skip-gram, 100차원 임베딩, Euclidean distance, Label Propagation, ϵ : 40)의 예측 결과는 <Table 13>에 나타난 바와 같으며, <Table 14>은 제안한 방법론과 라쏘 회귀분석 결과를 비교한 표이다. 영화평 데이터는 부정적인 평가보다 긍정적인 평가가 상대적으로 비율이 높은 특징을 가지고 있기 때문에 긍정보다는 부정을 정확하게 예측하는 것이 상대적으로 어려운 문제가 된다. BCR은 긍정과 부정 범주를 동일한 가중치로 고려한 정확도 지표인데, 라쏘 회귀분석의 경우 대부분의 영화 리뷰를 긍정으로 예측하여 BCR이 0.0128로 매우 낮게 나타나는 것을 알 수 있다. 반면, 본 연구에서 제안하는 방법론은 라쏘 회귀분석보다 향상된 값을 나타내어 기존 방법론에 비해서 상대적으로 매우 우수한 예측을 수행할 수 있음이 입증되었다.

5.4 파라미터 변화에 따른 성능 변화

<Figure 6>은 각 파라미터를 변경해가며 진행한 실험을 단어의 감성에 대한 BCR지표와 영화 평점에 대한 BCR지표에 대해 상자 그림을 그린 것이다. 단어의 감성에 대한 실험에 사용된 파라미터 중 Word2Vec 방법론은 Skip-gram, 임베딩 차원은 100차원, 거리 척도는 Euclidean distance measure를 사용하는 것이 좋은 성능을 보이는 것을 확인할 수 있다. 그래프 기반 준지도학습 기법은 Label propagation이 최고점에서는 좋은 성능

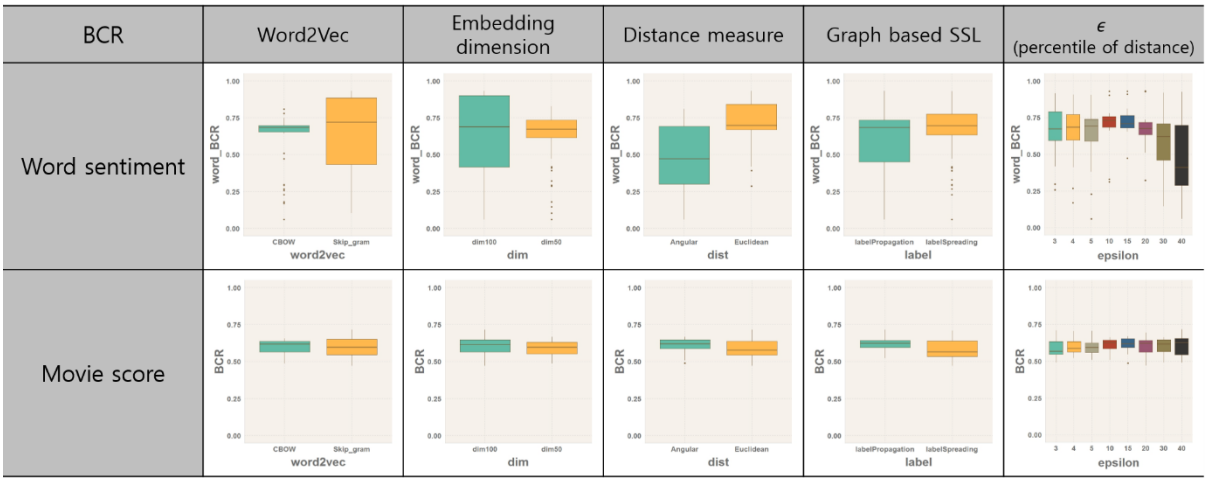


Figure 6. Boxplots of Performance

을 보이는 것을 확인할 수 있지만, 그 분포를 고려했을 때 두 차이가 명확히 있다고 하기는 힘들다고 생각된다. 또한 그래프 구축에 사용된 ϵ의 경우 10보다 작게 할당한 경우에는 성능이 좋지 않으며, 10에서 20분위수 정도를 정해 주었을 때 안정적인 성능 확보가 가능하고, 30 이상의 경우 성능은 비슷하지만, 시간 복잡성을 고려했을 때 성능이 좋다고 볼 수 없다. 영화 평점에 대한 실험은 5개 파라미터에 대한 성능이 전반적으로 비슷한 경향을 보이는 것을 확인할 수 있다.

6. 결론 및 활용방안

본 연구는 한국어 감성사전 구축 방법을 제안하였다. 다음 등 3개 사이트에서 얻은 영화 리뷰의 각 단어를 Word2Vec으로 저차원 벡터공간에 임베딩하고 거리 기반의 그래프를 구축하였다. 이후 비교적 명확한 극성을 지닌 Pre-labeled word를 seed node로 하는 그래프 기반 SSL을 적용해 label이 없는 단어에 감성점수를 부여함으로써 감성사전을 자동 구축했다. 제안된 방법론의 성능을 평가하기 위해 pre-labeled word의 극성과 개별 리뷰 평점의 예측 정확도를 각각 측정했다. 라쏘 회귀분석과 비교하여 두 관점 모두 좋은 성능을 보임을 확인하였다.

본 연구는 도메인 기반 한국어 감성사전 구축 방법론을 제안한다. 이를 통해 특정 도메인의 텍스트 데이터를 통해 해당 도메인의 감성사전 구축이 가능할 것이다. 구축된 감성사전을 이용하여 최종적으로는 도메인에서 각 문서의 감성을 보다 정밀하게 파악하여 여론의 의견을 수렴한 의사결정을 내릴 수 있을 것이라 기대한다. 실험에 사용된 파라미터 중 Word2Vec 방법론은 Skip-gram, 임베딩 차원은 100차원, 거리 척도는 Euclidean distance measure의 조합이 비교적 높은 성능을 보임을 확인하였다. 향후 연구자들의 연구 설계에 도움이 될 수 있을 것이라 기대한다.

본 연구에서 사용한 Word2Vec은 주변 단어의 맥락을 통해 의미 기반으로 단어를 임베딩한다. 이는 엄밀하게는 온전히 의미 유사도를 담고 있을 수 없으므로 본 연구는 임베딩시 한계점이 있다. 이를 개선하기 위해서는 최근 Facebook에서 발표한 fastText (Bojanowski *et al*., 2016)를 한글 기반으로 연구하여 적용한다면 개선의 여지가 있을 것이라 생각한다. 또한 영화 도메인에서 진행한 실험이므로 다른 도메인에도 실험을 진행하여 일반적으로 좋은 성능을 보임을 확인할 필요가 있다. 마지막으로 한글 범용 사전인 KOSAC(Kim et al., 2013)을 기반으로 특정 도메인의 감성사전을 구축한다면 더욱 정교한 감성사전이 나올 것이라 기대하고 있다.

참고문헌

Baccianella, S., Esuli, A., and Sebastiani, F. (2010), *SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, LREC.

Beineke, P., Hastie, T., and Vaithyanathan, S. (2004), *The sentimental factor : Improving review classification via human-provided information*, Proceedings of the 42nd annual meeting on association for computational linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016), *Enriching word vectors with subword information*, arXiv preprint arXiv: 1607.04606.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013), New avenues in opinion mining and sentiment analysis, *IEEE Intelligent Systems*, **28**(2), 15-21.

Chapelle *et al*. (2006), *Semi-Supervised Learning*, **6**, 193-196.

Jeong, B. and Yoon, J. (2016), Identifying product development opportunities using topic modeling and sentiment analysis, *Journal of the Korean Operations Research Society*, **4**, 4817-4822.

Jo, E. (2012), The Current State of Affairs of the Sentiment Analysis and Case Study Based on Corpus, *The Journal of Linguistics Science*, **61**, 259-282.

Kaji, N. and Kitsuregawa, M. (2007), *Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents*, EMNLP-CoNLL.

Kim, E. (2002), *A Wordfield Study of Korean Sentiment Verbs*, Sejong Press.

Kim, J., Seo, D., Kim, H., and Kang, P. (2017), Facebook Spam Post Filtering based on Instagram-based Transfer Learning and Meta Information of Posts, *Journal of Korean Institute of Industrial Engineers*, **43**(3), 192-202.

Kim, J., Oh, Y., and Chae, S. (2015), The Construction of a Domain-Specific Sentiment Dictionary Using Graph-based Semi-supervised Learning Method, *Korean Journal of the Science of Emotion & Sensibility*, **18**(1), 103-110.

Kim, M., Jang, H., Jo, Y., and Shin, H. (2013), KOSAC: Korean Sentiment Analysis Corpus, *Korea Information Science Society*, **6**, 650-652.

Li, F., Pan, S. J., Jin, O., Yang, Q., and Zhu, X. (2012), *Cross-domain co-extraction of sentiment and topic lexicons*, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, *Association for Computational Linguistics*, **1**, 410-419.

Liu, B. (2012), Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, **5**(1), 1-167.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013), *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013), *Distributed representations of words and phrases and their compositionality*, Advances in neural information processing systems, 3111-3119.

Rao, D. and Ravichandran, D. (2009), *Semi-supervised polarity lexicon induction*. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics.

Oh, S. (2014), A Study on the Conceptualization of FEAR in Korean and English, *Korean Semantics*, **44**, 141-170.

Pröllochs, N., Feuerriegel, S., and Neumann, D. (2015), *Generating Domain-Specific Dictionaries using Bayesian Learning*, ECIS.

Tang, J., Alelyani, S., and Liu, H. (2014), Feature selection for classification: A review, *Data Classification : Algorithms and Applications*, 37-64.

Turney, P. D. (2002), *Thumbs up or thumbs down? : semantic orientation applied to unsupervised classification of reviews*, Procee-

dings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics.

Von Luxburg, U. (2007), A tutorial on spectral clustering, *Statistics and Computing*, **17**(4), 395-416.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004), Learning with local and global consistency, *Advances in Neural Information Processing Systems*, **16**, 321-328.

Zhu, X. and Ghahramani, Z. (2002), Learning from labeled and unlabeled data with label propagation, Citeseer.

Zhu, X. and Goldberg, A. B. (2009), Introduction to semi-supervised learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **3**(1), 1-130.