이화여자대학교 대학원
2017학년도
석사학위 청구논문

감성사전 기반 Word2vec 자질을 이용한 감성 분류 시스템

빅 데 이 터 분 석 학 협 동 과 정
명 은 진
2018

감성사전 기반
Word2vec 자질을 이용한
감성 분류 시스템

이 논문을 석사학위 논문으로 제출함

2018 년 7 월

이화여자대학교 대학원

빅 데 이 터 분 석 학 협 동 과 정 명 은 진

명 은 진 의 석사학위 논문을 인준함

지도교수 신 경 식 _____

심사위원 김 은 갑 _____

민 대 기 _____

신 경 식 _____

이화여자대학교 대학원

목 차

표 목 차

그 림 목 차

논 문 개 요

감성분석(Sentiment Analysis)은 리뷰, 뉴스, 블로그 등의 다양한 텍스트 데이터로부터 특정 대상의 감성의 극성이 긍정, 부정 혹은 중립인지 찾는 것을 목표로 하는 자연어 처리(Natural Language Processing) 분석 기법이다. 감성분석의 기법으로는 크게 사전 기반 기법(lexicon based approach)과 기계학습 기법(machine-learning approach)이 있다. 사전 기반 기법의 경우 범용 감성사전 혹은 연구자가 직접 제작한 사전을 이용하여 감성분석을 수행하게 되며, 대량의 문서들을 대상으로 할 경우, 기계학습 기법을 적용하는 것이 일반적인 흐름이다. 이와 관련한 연구로는, 기계학습 기법의 중간단계인 자질 선택(feature selection)단계에서 분석 정확도를 높이기 위해 다양한 통계적 방법을 이용하는 연구가 있었다. 그리고 분류단계에서 필요한 분류기를 다양하게 적용하여 감성분석 모형의 성능을 비교하거나, 최근에는 대표적인 워드임베딩 방법론인 Word2vec을 이용해 기계학습 기법의 자질 선택 단계에 적용하여 분류를 수행한 연구들이 많이 있었고, 딥러닝을 이용한 감성분석에서도 자질 선택 단계에서 Word2vec을 적용하고 있는 추세이다.

Word2vec은 신경망 기반의 연속 워드임베딩(continuous word embedding)로서, Word2vec을 이용해 단어들을 학습할 경우 워드임베딩 공간에서 비슷한 문맥을 가진 단어들은 서로 가까운 공간 분포를 가지게 되는데, 어떠한 분야에서 사용되는 문서들을 대상으로 학습했는가에 따라 단어가 공간 분포는 달라지게 된다. Word2vec을 적용한 감성분석 연구로는 자질 선택 단계에서 Word2vec의 고차

원 문제를 해결하기 위해, 클러스터링 기법을 적용하여 차원을 축소시키는 것과 관련한 연구가 주류를 이루었고, Word2vec 모델이 문장 내에서 단어의 통사적 정보와 같은 쓰임새 정보만 학습하여, 감성적으로 극성이 다른 단어도 유사한 워드임베딩 벡터로 표현되는 한계점이 있는데 이는 감성분석의 정확도를 떨어뜨리는 문제점을 야기하기에 단어의 의미정보를 결합한 워드임베딩 방법을 제안하는 연구도 계속해서 등장하고 있다.

본 연구에서는 감성사전 정보를 결합한 Word2vec을 자질을 감성분석에 적용하는 새로운 감성 분류 모형을 제안하고자 한다. 곧, 연구자가 구축한 감성사전을 Word2vec 모델 학습 단계에 결합하여 감성사전 기반 Word2vec 모델을 구축하고 이를 통해 분석하고자 하는 문서의 자질을 구축하여 감성분석을 시행하였다. 본 연구에서 제안한 방법으로 감성분석을 시행했을 때, 기존 방법보다 분류 정확도가 약 2.5% 향상된 것을 확인하여, 제안모형의 유효성을 확인하였다.

본 연구는 기존의 감성분석 기법인 사전 기반 기법과 기계학습 기법을 혼합하여 감성분석을 수행하므로 연구자가 구축한 감성사전을 활용하여 감성의 극성 정보를 반영함으로써 분석 성과를 높인데 의의가 있었다. 또한, Word2vec을 적용한 감성분석 연구의 관점에서 바라봤을 때, 워드임베딩 기법이 단어 표현 방법으로는 우수하지만, 감성분석에 있어서 문장 구조적 통사적 정보를 학습하여 Word2vec이 감성의 극성 정보를 예민하게 반응하지 못하는 점이 있기에, 감성사전을 결합하여 기존의 Word2vec 자질의 한계점을 보완할 수 있는 연구를 수행했다는 점에서 의의가 있다.

Ⅰ. 서론

A. 연구의 배경 및 목적

현대 사회의 스마트기기와 소셜네트워크서비스(SNS) 매체의 발달로 온라인에서 사람들이 생산하는 데이터는 기하급수적으로 증가하고 있다. 데이터 형태도 정형 데이터뿐 아니라 비정형 데이터(텍스트, 동영상, 이미지 등)로 다양해졌다. 이에 따라 기업, 정부, 언론사 및 각종 단체에서는 데이터들을 이윤 창출, 국민 여론 조사 등의 필요한 목적에 맞게 활용하고자 데이터를 분석한다. 소비자들이 온라인에서 표현하는 반응 중 객관적인 정보 외에 소비자의 감정을 포함한 주관적 정보(subjective information)도 많은 사람들에게 확산될 경우, 잠재적 고객의 심리에 큰 영향을 주기에, 서비스를 제공하거나 물품을 판매하는 기업 중 일부는 소비자의 감정을 파악하여 이를 알맞게 활용할 수 있는 실시간 시스템 구축에도 많은 노력을 기울이고 있다. 사람들이 표현한 의견이 포함된 텍스트 데이터를 분석하는 것을 오피니언 마이닝(Opinion mining)이라고 하며, 그 중에서도 주관적 정보, 즉 감정에 관한 정보를 분석하는 것을 감성분석(Sentiment Analysis)라고 한다.(Liu et al., 2010; Dmitriy Bespalov et al., 2011)

감성분석은 텍스트에 표현된 감성이 긍정, 부정 혹은 중립인지 감성을 찾아 분류하는 것을 목표로 한다.(Dmitriy Bespalov et al., 2011; Maria Giatsoglou et al,. 2017) 감성분석 기법으로는 크게 사전 기반 기법(lexicon based approach)과 기계학습 기법(machine-learning approach)이 있다. 대량의 문서들을 대상으로 할 경우 감성정보를 학습시켜 구축된 감성분석 모형을 통해 감성분류를 수행하므로, 기계학습 기법을 적용하는 것이 일반적인 흐름이다(Rao et al., 2009). 이와 관련

한 연구로는, 기계학습 기법의 중간단계인 자질 선택(feature selection)단계에서 분석 정확도를 높이기 위해 TF(Term Frequency, 단어 빈도), TF-IDF(Term Frequency - Inverse Document Frequency, 역문서 빈도) 혹은 PMI(Pointwise Mutual Information) 기법을 이용하거나, n-gram을 이용해 단어의 순서를 고려한 어절의 조합의 개수에 따라 자질을 선택하는 통계적 방법을 이용하는 연구가 있었다. 그리고 분류단계에서 필요한 분류기(classifier)로 SVM, Naïve Bayes, ANN 등을 이용하여 감성분석 모형의 성능을 비교한 연구(Rodrigo Moraes et al., 2013)가 있었다. 최근에는 대표적인 워드임베딩 방법론인 Word2vec을 이용해 기계학습 기법의 자질 선택 단계에 적용하여 분류를 수행한 연구들이 많이 있었고(Maria Giatsoglou et al., 2017), 딥러닝을 이용한 감성분석에서도 자질 선택 단계에서 Word2vec을 적용하고 있는 추세이다.(Mario et al., 2017)

Word2vec은 기존의 단어 표현(Word Representation)방식인 'one-hot encoding'에서 문맥상의 의미를 전혀 보존하지 못했던 문제점을 해결하는 단어 표현 방식으로 신경망 기반의 연속 워드임베딩(continuous word embedding)이다. Word2vec을 이용해 단어들을 학습할 경우 워드임베딩 공간에서 비슷한 문맥을 가진 단어들은 서로 가까운 공간 분포를 가지게 되는데, 어떠한 분야에서 사용되는 문서들을 대상으로 학습했는가에 따라 동일한 단어가 다른 공간 분포를 가지게 되어, 다른 벡터로 표현될 수 있다.(서덕성 et al., 2017) Word2vec을 적용한 감성분석 연구로는 자질 선택 시 Word2vec의 고차원 문제를 해결하기 위해, 클러스터링 기법을 적용하여 차원을 축소시키는 것(Dhruv Mayank et al., 2016)과 관련한 연구가 주류를 이루었고, Word2vec 모델이 문장 내에서 단어의 통사적 정보와 같은 쓰임새 정보만 학습하여, 감성적으로 극성이 다른 단어도 유사한

워드임베딩 벡터로 표현되는 한계점이 있는데 이는 감성분석의 정확도를 떨어뜨리는 문제점을 야기하게 된다. 따라서 단어의 의미정보를 결합한 워드임베딩 방법을 제안하는 연구도 계속해서 등장하고 있다.((Duyu Tang et al., 2014; Maria Giatsoglou et al., 2017; 임미영 et al., 2017)

본 연구에서는 감성사전 정보를 결합한 Word2vec을 자질 선택 단계에 적용하는 새로운 감성 분류 모형을 제안하고자 한다. 연구자가 구축한 감성사전을 중심으로 Word2vec을 이용해 자질을 구성하여 감성분석을 시행한 후, 일반적인 방법으로 Word2vec 자질을 구성하여 실험한 감성분석 결과와 어떤 차이가 있는지 비교하였고, 제안모형으로 실험할 경우 분류 정확도가 기존 방법보다 약 2.5% 향상된 것을 확인하였다.

연구순서로 먼저 Python을 이용해 크롤링한 영화평 데이터를 추출하여 Word2vec 학습용 데이터와 감성분석 모형 학습 및 평가 데이터를 구축하였으며, 일부 데이터를 이용하여 감성사전을 구축한 뒤, 감성사전 기반한 Word2vec 모델을 구축하여 벡터 자질을 구성하였다. 이를 자질로 선택하여 연구자가 제안한 모형을 생성하였고, 감성분석을 시행한 후 감성분류 결과의 성능을 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서 선행연구로 국외와 국내에서 이루어진 감성분석 연구 및 워드임베딩에 관한 연구 현황에 대해 살펴보고, 3장에서는 본 연구에서 사용된 연구기법에 대해 서술하며, 4장에서 본 연구에서 제안하는 제안모형과 실험설계에 대해 설명하고자 한다. 그리고 5장에서 실험결과를 분석한 후, 마지막으로 6장에서 결론을 제시한다.

Ⅱ. 선행연구

A. 온라인 리뷰(Online Review)를 이용한 감성분석

전자 상거래가 활성화되면서 상품을 구매하거나 서비스를 이용하는 구매자들은 구매 후기, 즉 온라인 리뷰(Online Review)를 작성하게 되었다. 이는 구매예정자들이 상품을 구매하거나 서비스를 이용하기 전에, 참조할 수 있는 중요한 정보이지만, 데이터의 양이 급격이 증가하면서 많은 양의 리뷰를 다 확인할 수 없기에, 사이트에서는 이용자들에게 효과적으로 리뷰 정보를 제시하는 방법을 고안하게 되었다. 그러한 방법 중의 하나로 온라인 리뷰를 작성할 때, 미리 사용자가 만족도에 관한 점수를 별점(예, 1점~5점)형식으로 매기도록 하여, 후에 구매예정자들에게 전체 리뷰의 별점 통계를 요약적으로 제시하여 참조하게 하였다. 온라인 리뷰를 이용한 감성분석은 이러한 리뷰와 별점 즉, 점수 데이터를 이용하여 텍스트에 대한 감성분류를 자동화하는 연구이다.

감성분석은 감성의 결과값을 찾는 대상의 단위에 따라 문서 단위(Document-level) 분석, 문장 단위(Sentence-level) 분석, 속성 단위(Aspect-level) 분석으로 분류할 수 있다. 문서 단위 분석에서는 한 문서에서 하나의 감성을 표현한다고 사실을 가정하고 수행하는 감성분석으로 문서의 최종 감성 값을 찾는 것이며, 문장 단위와 속성 단위 분석은 각 문장과 특정 개체의 속성에서 하나의 감성을 표현한다고 가정하고 분석을 수행하게 된다. 아래 표1은 분석 단위에 따라 감성분석 선행 연구를 분류한 것이다.

(Bo Pang et al, 2004)의 경우, 텍스트 분류(text-categorization)기법을 문서의

주관적인(subjective) 영역에만 적용시키는 새로운 기계학습(machine-learning) 기법을 소개했다. 먼저, 문장 단위(sentence-level) 감성분석을 시행해, 감성이 없는 객관적인 문장(objective sentence)은 제거하여, 감성이 있는 문장(subjective sentence)만을 추출한 것을 통합하여 문서 단위(Document-level) 감성분석을 시행하였다. (Rodrigo Moraes et al., 2013)은 위의 (Bo Pang et al., 2004) 연구의 영화평 데이터를 벤치마크 하여, 각기 다양한 기계학습 기법(SVM, ANN)을 적용하여 감성분석 비교 연구를 수행하였다. (Theresa Wilson et al., 2005)의 경우 다차원의 질문에 대한 응답이나 요약을 해야 할 경우, 상품 리뷰의 분석은 문장 단위(Sentence-level) 혹은 구 단위(phrase-level)의 감성분석까지도 필요함을 주장하며, 많은 양의 문장 표현에서도 전체의 맥락이 나타내는 감성의 극성이 어떠한 지를 분류하는 새로운 방법론을 제시하였다. (Yohan J et al., 2011)에서는 리뷰 분석을 수행할 때 각기 다른 개체들의 대한 의견과 감성이 어떻게 표현되었는지 발견하는 것이 중요하다고 주장하며, SLDA(Sentence-LDA)라는 확률적 생성모형(probabilistic generative model)과, 각기 다른 개체들을 향한 감성들을 모델링하는 개체와 감성 단합모델 ASUM(Aspect and Sentiment Unification Model)을 제시하여 속성 단위(Aspect-level) 감성분석을 수행하였다.

한편 감성분석 연구는 분석단위 뿐 아니라 분석 방법론에 따라, 사전 기반(lexicon-based) 기법과 기계학습(machine-learning)기법을 적용한 연구로 분류하거나, 기계학습 기법 안에서 지도학습, 비지도학습 혹은 강화학습 기법을 적용한 연구로 분류할 수 있다. 이 외에 기존의 방법론을 병합한 하이브리드 기법을 적용한 감성분석을 연구하기도 하였다.

표1. 분석단위에 따른 감성분석 연구 분류

	연구자	연도	데이터
문서 단위 (Document-level) **감성분석**	Bo Pang et al	2004	영화평 리뷰
	Andrew L et al	2011	IMDB 영화평 리뷰
	Rodrigo Moraes	2013	영화평 리뷰
	Xing Fang et al	2015	아마존 상품 리뷰
	Ashna M.P et al	2017	영화평 리뷰
	Liang-Chih Yu et al	2018	Twitter 데이터, 영화평 리뷰
문장 단위 (Sentence-level) **감성분석**	A Meena et al	2007	자동차 리뷰
	Andrew L et al	2011	IMDB 영화평 리뷰
	Xing Fang et al	2015	아마존 상품 리뷰
	Orestes Appel et al	2016	Twitter 데이터, 영화평 리뷰
	Maria Pontiki et al	2016	7개 도메인 리뷰
	Ashna M.P et al	2017	영화평 리뷰
속성 단위 (Aspect-level) **감성분석**	Theresa Wilson et al	2005	MPQA 코퍼스
	Yohan J et al	2011	아마존 전자기기/Yelp 식당 리뷰
	Maria Pontiki	2016	7개 도메인 리뷰
	I. K. C. U. Perera et al	2017	식당 리뷰

(1) 사전 기반(lexicon-based) 감성분석

감성사전을 이용한 연구는 단어의 극성 정보를 포함된 감성사전을 이용해 문서에 출현하는 사전에 포함된 단어를 기반으로 분류하는 방법이 있고, 기존에 구축된 범용 감성사전이 아닌 도메인에 특화된 사전을 구축하여 실험한 연구가 있는데, (송종석 et al., 2011)에서는 도메인별로 제품 특징에 대한 서술어의 의미방향을 고려한 긍정/부정 사전을 자동으로 구축하여, 범용 감성사전을 활용해 감성분석을 시행한 결과보다 감성분석의 결과 정확도가 향상됨을 확인하였다. (Santanu Mandal et al., 2017)에서는 온라인 리뷰를 통해 기존의 사전 기반 감성분석 연구와 차별화 하여, 감성분석에 세 가지의 비교정도(the three degrees of comparison)로 기본/비교/최상급 단어를 사전에 적용하였고, 부정어 단어를 따로 적용하여 성능이 향상됨을 보였다. 이렇게 연구자가 직접 분석하여 감성사전을 구축하게 되면 정확하고 세밀한 결과를 도출할 수 있지만, 수작업에 의존하기에 시간적 비용이 들고 연구자마다 다른 기준으로 적용된다는 문제점도 있다. 한편, 감성사전 구축은 이렇게 연구자 직관을 중심으로 구축하는 것도 있지만, 그 외에도 크게 통계적 방식 기반 감성사전 구축, 기계학습 모델 기반 감성사전 구축의 방법도 있다.

통계적 방식을 기반하여 감성사전을 구축한 연구로 (이상훈 et al., 2016)에서 감성사전 구축을 위해 PMI(Point-wise Mutual Information)기법을 사용하였다. 즉 감성사전을 데이터의 특성에 맞게 적절하게 변형하여 구축하는 방법을 시도한, 맞춤형 감성사전 구축을 위한 핵심 기법 SO-PMI (Semantic Orientation from Point-wise Mutual Information)을 통해, 기존의 범용 감성사전 대비 예측 정확도가 통계적 유의 수준에서 향상된 것을 확인하였다. (Kaji et al., 2007)에서는 극

성을 가진 문장을 분류하여 문장 속 TF(Term Frequency, 단어 빈도)를 바탕으로 각 단어의 극성값을 계산하였다. (김승우 et al., 2014)에서도 TF를 기반으로 감성 지수를 계산하여 감성사전을 구축하는 것을 소개하였다. 이러한 통계 기반 방법론은 단어의 극성값을 정략적으로 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 있기에 (서덕성 et al., 2017)에서는 기계학습 방법론을 바탕으로 개별 단어의 감성 점수를 산출하여 감성사전 구축에 기초가 되는 연구를 수행하였다. 영화평 리뷰 데이터로 신경망 모델인 Word2vec을 학습 시켜 워드임베딩 공간을 형성하였고, 임베딩된 공간에 어휘들이 문맥적 의미를 반영하여 위치함을 가정하여, 임베딩된 공간에서 단어간의 유사도를 고려해 네트워크를 구축하여 단어들 간의 관계를 표현하였다. 이렇게 구축된 단어 네트워크에 그래프 기반의 준지도학습(graph-based semi-supervised learning)을 적용하여 개별 단어의 감성어휘 점수를 산출하는 방법론을 개발하였다. 이 외에도 다양한 기계학습 기법을 활용해 감성사전을 구축한 연구들이 있었다. (Rao et al., 2009; Li et al., 2012)

(2) 기계학습 기반(machine-learning) 감성분석

기계학습 기반 감성분석 선행연구로는 (Bo Pang et al., 2004)에서 주관적인 감성이 포함된 문장만을 선별하여 분류기(classifier)로 SVM(Support Vector Machine, 서포트벡터머신)과 NB(Naïve Bayes, 나이브베이즈)를 선택해 기계학습 기법을 적용한 문서 단위 감성분석 연구를 수행하였는데, 이후 대부분의 많은

연구들이 이와 같은 방향으로 기계학습 기반 감성분석 연구가 진행되었으며, 분류 정확도를 높이기 위해 다양한 분류기를 적용해 감성분석 결과를 비교 연구를 하거나, 효과적인 자질 선택(feature selection)을 위해 다양한 기법을 적용하는 등의 연구를 진행하였다.

기존 통계적 언어처리에서 기본적인 자질 선택 기법인 TF-IDF(Term Frequency - Inverse Document Frequency, 역문서 빈도) 기법을 사용한 (임미영 et al., 2017) 연구와, 문장의 품사정보인 POS(Part of speech) 태그와 단어 빈도수(Term Frequency)를 적용한 (Oaindrila Das et al. 2014) 연구가 있었으며, (Manvee Chauhan et al.,2015)에서는 기계학습 기법을 적용한 상품 리뷰 감성분석을 시행하였고, NB 모델과 SVM 모델을 분류기로 적용해 리뷰의 극성을 계산하였다.

(Alvaro Ortigosa et al., 2013)에서는 기존 감성분석에서 분류기로 SVM을 적용한 것은 많았지만, ANN(Artificial Neural Network, 인공신경망)을 적용한 것은 적었기에 감성분석에 SVM과 ANN를 적용하여 비교 연구를 진행하였고, 문서의 분류 정확도를 높이는 요구사항을 분석하였다. 일부 불균형 데이터를 제외하고는 ANN을 적용했을 때, SVM 보다 성능이 우수했고, 특히 영화 리뷰 데이터에서 우수했음을 실험결과를 통해 확인하였다.

(이동엽 et al. 2017)에서는 워드임베딩 공간을 이용한 기계학습 기반 감성분석 모델을 제안하였고, 데이터 570만건을 학습하여 형성된 워드임베딩 공간을 이용하여 긍정, 부정 리뷰 데이터의 개수에 따라 총 3개의 SVM 분류기 모델을 학습하는 방식으로 진행하였다. 긍정 리뷰 데이터 5만건, 부정 리뷰 데이터 5만건을 이용하여 SVM 분류기로 학습하였을 때, 가장 높은 정확도를 나타냈음을 보였다.

(3) 하이브리드 기법을 적용한 감성분석

두 기법 이상을 혼합한 하이브리드 기법을 적용한 감성분석 연구로는, 먼저 (Alvaro Ortigosa et al., 2014)에서 페이스북 정보를 추출하여 개인화된 지도 시스템(Adaptive e-learning system)에 적용하고자, 사전 기반 기법과 기계학습을 결합한 하이브리드 감성분석 모형과 감정 변화를 탐지하는 모형을 제시한 연구가 있었다. 3000개의 페이스북 상태 메시지 데이터에 데이터 분석 도구 Weka를 이용하여 단어벡터(Word Vector)로 변환하여, 자질선택 단계에서 CFS(Correlation-based Feature Selection)을 적용하여 자질의 크기를 줄였으며, 이후 기계학습 기법을 적용하였을 때와 사전 기반 기법을 적용한 후에, 기계학습으로만 감성분석을 시행하였을 때의 결과를 비교하였다. 그 결과 연구자가 제시한 하이브리드 방식으로 감성분석을 시행했을 때, 정확도가 더 높아졌음을 확인할 수 있었다.

(Orestes Appel et al., 2016)에서는 트위터와 영화평 리뷰 데이터를 이용해 문장 단위 감성분석을 시행하였고, 퍼지이론(Fuzzy sets)과 SentiWordNet을 이용한 감성사전 기법을 결합한 하이브리드 기법 감성분석을 시행하였고, NB(Naïve Bayes, 나이브베이즈)와 ME(Maximum Entropy, 최대엔트로피)를 이용한 감성분석과 비교하여 연구자가 제안한 모형의 결과가 더 정확함을 보였다.

(임미영 et al., 2017)와 (Maria Giatsoglou et al., 2017) 연구에서는 워드임베딩을 통한 감성분석을 시행하였는데, 워드임베딩을 통한 자질 선택 단계에서 기존 워드임베딩 벡터와 의미 정보를 표현한 벡터를 결합하여 하이브리드 자질을 구축하여 사용하였고, 기존 모델과 감성 분류 결과를 비교하여, 연구자가 제안한 모형의 우수성을 확인하였다.

B. 워드임베딩을 적용한 감성분석

기계학습 기법을 적용하여 감성분석을 수행하기 위해서, 자질 선택(feature selection) 단계가 있으며, 감성분석 결과에 효과적인 자질은 전체 문서의 감성값을 잘 반영할 수 있어야 한다. 그 시점에서 단어 표현(Word Representation)단계가 필요한데, 이는 기계가 이해할 수 있도록 텍스트를 벡터화 하는 것을 의미한다. 단어 표현은 비단 감성분석 분야 뿐만 아니라 텍스트 마이닝(Text Mining)의 전반적인 분야에서도 벡터화한 텍스트를 사용해야하기에 중요한 요소이다. 기존의 단어를 벡터화하는 방법 중 하나로 'one-hot encoding' 방식이 있다. 이 방식은 사용된 단어를 '1', 사용되지 않은 단어를 '0'으로 표현해 총 단어의 개수 n만큼의 길이를 가진 벡터를 나타낸다.

이 방식을 적용한 감성분석 초기 연구로는 (Bo Pang et al., 2002)가 있었고, 이후 같은 방법으로 더 좋은 감성분류 결과를 얻기 위해 다양한 연구들이 수행되었다.(Bo Pang et al., 2008; Bing Liu et al., 2012; Ronen Feldman et al., 2013). 하지만 이러한 방식은 단어들 간의 복합적인 관계를 포함하지 못했다. 단어 하나를 표현하기 위해 문서 집합(corpus)에 존재하는 단어 개수만큼의 차원을 갖는 벡터가 필요하며 단어 간의 조합으로 갖는 문맥상의 의미를 나타내지 못하는 문제점이 생겼다. 즉, 단어의 존재여부 외에 단어와 단어 사이의 관계가 어떠한 지에 대한 연관성을 결과에 반영할 수 없다. 하지만 워드임베딩(Word Embedding)이라는 신경망 기반 모델이 등장하므로 이러한 한계점을 극복할 수 있었다. 워드임베딩 혹은 단어임베딩이란, 단어를 벡터로 표현할 때, 유사한 단어들은 유사한 값의 분포를 가지는 벡터들로 표현되도록 하여 어휘 의미를 표현하는 기술이다.(임미영 et al., 2017)

국외에 워드임베딩을 적용한 감성분석 연구로는, 처음으로 감성분석에 워드임베딩 기법을 적용시킨 (Dmitriy Bespalov et al., 2011) 연구가 있다. 이 연구에서는 문서 단위의 감성분석을 진행하였는데, 연구자는 기존의 BOW(Bag Of Words)의 한계를 극복하기 위한 n-gram을 이용한 자질 선택 방법에서 n의 값이 3 이상이 될 경우 일어나는 계산복잡도 문제점을 지적하였고, 이러한 문제점을 해결하기 위해서 고차원의 n-gram 사용을 위한 새로운 n-gram을 임베딩(embedding) 기법인 다차원 임베딩(multi-level embedding) 감성분석 모형을 제시하였다. 분류기로 로지스틱 회귀(Logistic Regression)를 선택하여 실험했을 때, 가장 좋은 성능을 보임을 증명하였다.

(Q. Le et al., 2014)에서 단락, 문장 그리고 문서라는 텍스트의 길이에 따라 벡터를 표현하는 단락 벡터 Doc2Vec(Paragraph Vector document to vector approach)을 제시하였고, (A. Tripathy et al., 2016)은 워드임베딩 적용 후, n-gram과 각기 다른 기계학습 분류기들(SVM, Naïve Bayes Maximum Entropy) 비교 연구를 수행해 유니그램(unigram)과 SVM을 사용했을 때 가장 좋은 성능이 보임을 나타내었다.

또한, 기계학습 기법을 통해 감성분석을 수행할 경우 자질 선택 시, 워드임베딩 기법인 Word2vec의 고차원 문제를 해결하기 위해 클러스터링 기법을 적용하여 차원을 축소시키는 것(Dhruv Mayank et al., 2016, Eissa M.Alshari et al., 2017)에 대한 연구와 각 문서(document)를 구성하고 있는 단어들의 벡터값들을 평균화하여 자질로 구성하는 연구가 있었다.(R. K. Bayot et al., 2016; Y. Belinkov et al., 2015)

(Duyu Tang et al., 2014)에서는 워드임베딩이 문법의 구조적인 단어의 맥락만을 모델링하기에, 'good'과 'bad'의 경우 정반대의 감성값을 가져도 문법 구조상으

로 비슷한 맥락을 가지기에, 워드임베딩 후에 비슷한 공간에 위치하는 것을 예시로 제시하며, 기존 워드임베딩을 적용하기에는 감성분석에 효과적이지 못한 부분이 있다고 주장하며 감성에 특화된 워드임베딩 SSWE(sentiment specific word embedding)을 학습하는 것을 제시하였다.

(Maria Giastoslou et al., 2017)에서는 위의 연구에서와 같이, 의미 정보를 강화하는 워드임베딩 연구를 수행하였고, 일반적인 데이터에 Word2vec을 적용하여 얻은 자질과 감성사전을 이용해 얻은 자질을 벡터로 변환하여 결합하였고, 이를 자질로 선택해 감성분석을 수행하였다. 영화평 및 휴대폰 상품평 데이터를 이용해 도메인별 감성분석을 수행한 결과 대부분 성능이 향상된 것을 확인하였다.

국내에 워드임베딩을 적용한 감성분석 연구로는 (이동엽 et al., 2017)에서 워드임베딩을 이용해 아마존 패션 상품 리뷰로 감성분석을 시행하였다. 아마존 리뷰 데이터로 워드임베딩 공간을 형성한 뒤, 패션 리뷰 데이터에 긍정, 부정, 리뷰 데이터의 개수(긍정 부정 데이터가 각각 1만, 2만, 5만건)에 따라 총 3개의 SVM 분류기 모델을 학습하였고, 감성분석 실험결과 긍정, 부정 리뷰 데이터를 각각 5만 건일 때와, SVM 분류기를 학습했을 때 가장 높은 분류 정확도를 보였음을 나타냈다.

(임미영 et al., 2017)에서는 (Duyu Tang et al., 2014) 연구에서와 같이, 기존의 워드임베딩 연구의 한계점을 지적하였다. 곧, 기존 워드임베딩에서는 단어의 통사적 정보와 같은 쓰임새 정보만을 학습하고 있기에 쓰임새는 같으나 의미가 반대인 ‘good’과 ‘bad’가 유사한 벡터로 표현되기에 단어의 쓰임새와 도메인별 단어의 감성 정보를 결합한 워드임베딩 방법을 제안하며, 이를 이용한 감성분류 시스템을 제안하였다. 도메인간에 공유할 수 있는 데이터를 이용해 구축한 Word2vec 통사 네트워크와 특정 도메인 데이터로 학습시킨 감성분류 네트워크

를 결합하였고, 두 네트워크를 통해 생성된 워드임베딩 벡터들을 연결하여 로지스틱 회귀(Logistic Regression)를 분류기로 사용하여 실험하였다.

표2. 워드임베딩을 적용한 감성분석 국외/국내 연구

	연구자	연도	기법
국외 연구	Dmitriy Bespalov et al	2011	SLNA, LSI, 로지스틱 회귀
	Xavier Glorot et al	2011	Denoising Auto-encoders, SVM
	Dmitriy Bespalov et al	2012	다항 로지스틱 회귀
	Duyu Tang et al	2014	SVM, NB
	A. Tripathy et al	2016	SVM, NB, Maximum Entropy
	Dhruv Mayank et al	2016	SVM
	Eissa M.Alshari et al	2017	로지스틱 회귀, SVM
	Maria Giatsoglou et al	2017	Lexicon, SVM
국내 연구	이동엽 et al	2017	SVM
	임미영 et al	2017	로지스틱 회귀

감성분석의 자질 선택에서 워드임베딩을 사용할 경우 텍스트 데이터를 이용해 워드임베딩 공간을 형성시키는 워드임베딩 모델 학습단계가 필요하다. 워드임베딩

모델인 Word2vec을 사용한 대부분의 연구자들은 모델 학습단계에서 아래 **표3**에서와 같이 감성이 포함된 리뷰 데이터를 학습 데이터로 선정하였으며, 사전 기반 기법과 기계학습 기법을 결합한 감성분석을 수행한 (Maria Giastoslou et al., 2017) 연구에서는 워드임베딩 벡터(Word embedding-based features)와 감성 사전을 이용한 벡터(Lexicon-based features)를 결합한 벡터(Hybrid features)를 최종자질로 선택하였는데, 워드임베딩벡터 학습 데이터로는 데이터로 감성이 배제된 위키피디아 기사를 이용하였고, (임미영 et al., 2017) 연구에서도 도메인 간에 공유할 수 있는 데이터를 이용한 Word2vec 네트워크를 구축하기 위해서, 감성 데이터 뿐만이 아니라 세종 원시 말뭉치 데이터도 포함하여 워드임베딩 모델을 학습시켰다.

표3. Word2vec을 적용한 감성분석 연구 Word2vec 모델 학습데이터

	연구자	연도	Word2vec 모델 학습데이터
국외 연구	Andrew L. Maas et al Duyu Tang et al Maria Giatsoglou et al	2011 2014 2017	IMDB 영화평 리뷰(중립 포함) Twitter 데이터(긍정, 부정) 위키피디아 기사
국내 연구	이동엽 et al 서덕성 et al 임미영 et al	2017 2017 2017	아마존 상품 리뷰 영화평 리뷰 세종데이터 원시말뭉치, 영화평 리뷰

단어 표현에 효과적인 Word2vec을 감성분석 단계에 적용한 연구는 계속 늘어나고 있는 추세이지만 Word2vec을 이용할 시, 감성의 극성이 다른 단어도 벡터에서 가까운 공간에 위치하는 문제 또한 끊임없이 제기되고 있어, 이를 보완할 수 있는 연구도 필요하다. 하지만 국내에서는 워드임베딩을 적용한 감성분석 연구와 특히 기존의 워드임베딩 학습의 한계점을 보완할 수 있는 연구모형 제안도 부족한 실정이다.

따라서 본 연구에서는 워드임베딩 기법인 Word2vec을 적용하여 감성분석 수행 시에 나타나는 한계점을 보완할 수 있는 연구모형을 제안한다. 즉 감성사전 정보를 결합한 Word2vec 자질을 이용하여 감성 분류를 수행하고자 한다. 기존의 Word2vec 모델만을 이용하여 감성분석을 수행할 경우와 본 연구에서 제안하는 모형을 이용할 경우, 감성분석 결과에 어떤 변화가 있는지를 조사하고자 하여 제안모형의 유효성을 입증하고자 한다.

Ⅲ. 연구기법

A. 감성분석

1. 감성분석 개요

감성분석(Sentiment Analysis)이란, 텍스트에 표현된 개체와 그 속성에 대한 의견, 감성, 평가, 태도 등을 분석하는 것으로 감성 마이닝(sentiment mining), 주관성 분석(subjectivity)라고 부르기도 한다. 감성은 정의하기에 따라 달라지는데, 하나의 문장에 감성의 대상이 하나일 수도 있고, 여러 대상이 있을 수도 있다. 감성분석의 목적은 대상에 대한 감성이 무엇인지 파악하는 것이며(Duyu Tang et al., 2014), 감성의 결과값을 찾는 대상의 단위에 따라 분석단위(Levels of Analysis)는 크게 세 가지로 분류한다.

첫번째는 문서 단위(Document-level)로 개별 문서에 대한 전체적인 감성이 무엇인지 분석하는 것이고, 두번째는 문장 단위(Sentence-level)로 문서 안에 속한 여러 문장이 있을 경우 문장을 단위로, 각각의 개별 문장에 대한 감성을 분석하는 것이며, 마지막으로 속성 단위(Aspect-level)는 감성의 대상이 되는 개체와 개체의 속성까지 감성과 함께 직접적으로 분석하는 것을 의미한다. 문서 단위의 감성분석의 경우 주어진 문서에서 감성의 대상이 되는 개체에 대해 개별 문서의 전체적 감성이 무엇인지 파악하는 것으로 전통적인 텍스트 분류에서의 감성분류를 다루므로 가장 단순한 단위의 감성분석이라고 할 수 있다. 본 연구에서는 문서 단위의 감성분석을 분석의 단위로 정하여 연구하고자 한다.

또한 감성의 유형(Type of Opinions)을 두 가지 방법으로 분류할 수 있다. 아래의 **표4-1, 4-2**를 참조하면 감성을 일반형 감성(Regular Opinion)과 비교형 감성(Comparative Opinion)으로 분류하거나, 주관적 감성(Subjective Opinion) 혹은 객관적 감성(Fact-Implied Opinion)으로 분류하기도 한다. 이하 연구에서는 특정 대상을 향한 감성을 표현하는 일반형 감성을 감성의 유형으로 정의하여, 이를 중심으로 설명하고자 한다.

표4-1. 감성의 유형(Type of Opinions)1

일반형 감성 **(Regular Opinion)**	**비교형 감성** **(Comparative Opinion)**
특정 개체 또는 개체의 속성에 대한 감성을 표현하며, 일반적으로 일반형 감성을 분석대상으로 함	개체 간 공유하는 속성의 비교를 통해 감성을 전달함

표4-2. 감성의 유형(Type of Opinions)2

주관적 감성 **(Subjective Opinion)**	**객관적 감성** **(Fact-Implied Opinion)**
주관적 문장에서 드러나는 일반형 감성 또는 비교형 감성	객관적(사실 전달) 문장에서 드러나는 일반형 또는 비교형 감성

감성분석에서 추가적으로 고려해야 할 사항은 감성분석의 도메인에 따라 감성어휘가 전달하는 감성이 달라질 수 있다는 것이다. 아래 예시 문장들을 보면, 똑같이 '무섭다' 혹은 '졸립다'라는 감성어휘가 포함된 문장이라도, 하나의 도메인에서는 문장의 감성이 긍정이고, 다른 쪽에서는 부정으로 표현되는 것을 볼 수 있다. 또한, 문장이나 문맥에 따라 감성이 달라지는 경우가 있다는 것을 알 수 있다.

(1) 집으로 돌아가는데, 길이 너무너무 무서웠어.
(2) 어제 그 공포영화 봤는데, 진짜 진짜 무서운 거 있지.

(3) 수업시간에 너무 졸렸어.
(4) 이 침대는 눕자마자 바로 졸린다니까

위와 같이 감성분석에서는 특정한 감성어휘가 전달하는 감성이 항상 일정하지 않다는 점에서, 도메인 혹은 문장의 내용에 의존적인 것을 알 수 있다. 따라서 대부분의 연구에서 특정 도메인 내에서 감성분석을 수행하거나, 도메인 간의 감성어휘 사전을 따로 구축하여 접근하는 방법을 시도하고 있다.

(Orestes Appel et al., 2016)에서는 감성분석의 기본적인 핵심 절차를 아래의 **그림1**로 표현하였다. 감성이 포함된 문서로부터 분석을 실시할 때, 분석의 대상(Object/Feature)과 감성을 표현하는 주체(Opinion holder)를 추출하여 감성분류를 수행한다. 문서 단위 감성분석에서는 한 문서에서 감성을 표현하는 주체와 대상이 하나라고 가정하기에, 감성분석 대상 및 주체 추출의 과정은 생략한다.

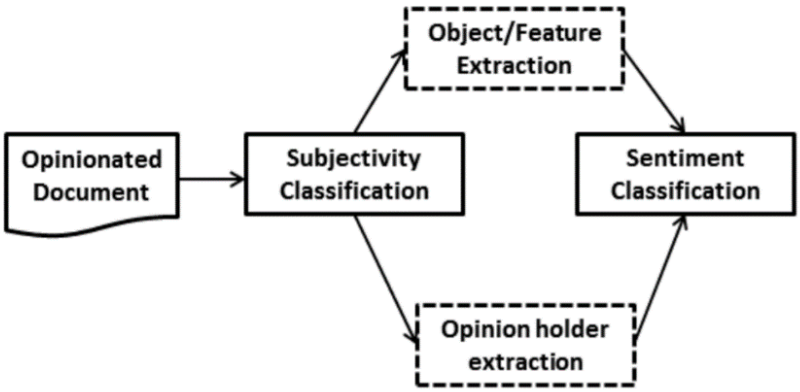

그림1. 감성분석 기본절차

2. 감성분석 기법

감성분석의 기법으로는 크게 사전 기반 기법(lexicon-based approach)과 기계학습 기법(machine-learning approach)이 있으며, 기법 간에 감성분석 절차가 조금씩 차이가 있다. 최근에는 이 두 기법을 다양한 방식으로 결합한 하이브리드 기법도 연구자에 의해 제안되고 있는 추세이다.

2.1 사전 기반 기법(lexicon-based approach)

사전 기반 기법(lexicon-based approach)은 연구자가 직접 감성분석에 필요한 감성사전을 구축하여, 분류하고자 하는 문서에 구축한 감성사전에 포함된 단어가 있을 경우 점수(+/-)를 합산하여 최종 감성을 분류하는 방식이다. 즉, 주요 어휘의 감성 극성이 미리 정의된 감성사전을 구축한 후, 새로 주어진 문서에 출현

한 어휘의 감성의 극성에 따라 문서전체의 감성 극성을 분류하게 되는데(김승우 et al., 2014) 감성사전은 극성(polarity), 주관성(subjectivity)등 감성을 지닌 단어들인 감성어휘(sentimental lexicon)들의 모음으로서 감성분석의 정확도를 결정짓는 핵심적인 요소라고 할 수 있다.(서덕성 et al., 2017)

도메인에 의존적인 감성분석의 특징에 따라, 기존 범용 사전을 사용해서 적절한 최종 감성의 값을 찾기 어려울 때가 있기에, 연구자가 도메인 별로 특화된 감성사전을 따로 구축하여 연구를 진행하기도 한다.

그림2은 (Khin Zezawar Aung et al., 2017)에서 학생들의 리뷰를 분석하는 방법으로 제시한 사전 기반 감성분석 모델로, 연구자가 구축한 감성사전(Sentiment word Database)을 기반으로 분석을 수행하게 된다. 이는 일반적인 사전 기반 기법의 프로세스를 잘 보여주고 있다.

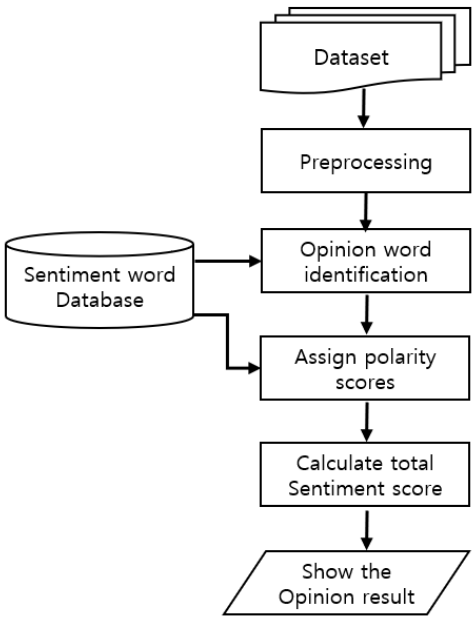

그림2. 사전 기반 분석 절차

2.2 기계학습 기법(machine-learning approach)

기계학습(machine-learning)이란 인간의 뇌가 자연스럽게 수행하는 '학습' 능력을 컴퓨터로 구현하는 방법으로 기존의 데이터가 가진 특징을 벡터로 변환시키는데 이 같은 변환을 자질 선택 혹은 특징추출(feature selection)이라고 한다. 이러한 학습을 기반으로 분류기(classifier)를 통해 분류 예측 등을 수행하게 된다.

감성분석에서 기계학습 기법은 대량의 문서 데이터에 감성라벨(긍정/중립/부정)이 있을 시 일반적으로 사용하는 방법으로, 데이터 전처리(Pre-Processing)를 거친 뒤, 문서와 라벨을 학습시켜 모형을 구축하는데, 구축된 모형으로 감성분류(classification)를 수행하게 된다.

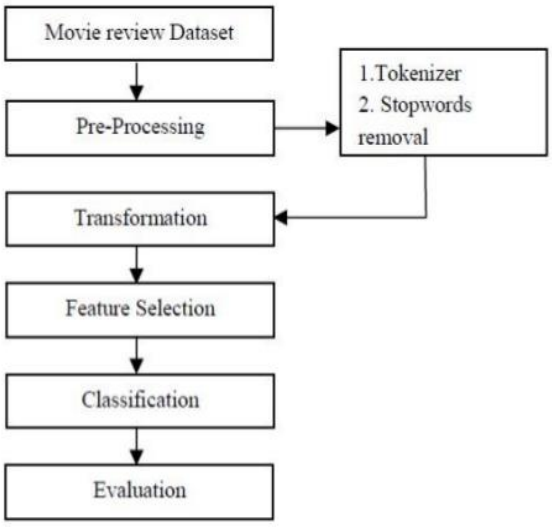

그림3. 기계학습 분석 절차

그림3는 (Sangeeth Nagarajan, 2014)에서 기계학습 분석 모델로 제시한 것으로 감성분석에서의 일반적인 기계학습 분석 절차를 잘 나타내고 있다. 선행연구로는 기계학습 기법의 자질선택 단계에서 다양한 통계적 방법 이용하거나, 분류기로 SVM, Naïve Bayes, ANN 등을 이용하여 모형 성능을 비교한 연구들이 있었다.

B. Word2vec

앞서 기계학습 기법에서 자질 선택 단계에서 데이터가 가진 특징을 벡터로 변환한다고 설명한 것처럼, 감성분석에서 자질들은 어휘 곧 단어들이 된다. 이러한 자질들을 벡터로 변환한다는 것은 단어 표현(Word Representation) 단계를 거치는 것을 의미한다. 단어 표현은 많은 자연어 처리 시스템(Natural Language Processing Systems)에서 중요한 요소이다(Andrew L.Maas et al., 2011). 컴퓨터가 단어에 대해 인지할 수 있도록 하려면 수치적인 방법으로 단어를 표현할 수 있어야 하는데, NLP(Natural Language Processing, 자연어 처리)에서는 아래와 **그림4**과 같은 'one-hot encoding' 방식을 사용하였다.

단어1	100000000000000
단어2	010000000000000
단어3	001000000000000
…	…
단어14	000000000000010
단어15	000000000000001

그림4. 'one-hot encoding' 방식으로 변환된 단어 표현 예시

‘one-hot encoding’ 방식은 전체 단어의 개수가 N이라면 길이가 N인 벡터를 만들고, 해당 단어 자리에만 숫자 ‘1’을 넣고 나머지 단어들을 표현하는 자리에는 숫자 ‘0’을 넣는 것이다. 이와 같은 단어 표현은 단어의 존재여부만을 반영하는 독립성을 갖기에, 단어 간의 관계는 반영하지 못하고, 하나의 단어를 표현하는데 문서 집합(corpus)에 존재하는 수만큼의 차원을 갖는 벡터가 필요한 단점이 존재한다. 감성분석에서는 특정 단어들의 조합이 감성을 결정하는데 영향을 줄 수 있기에, 단어 간의 관계를 파악하는 것은 중요한 부분이다. 즉, 어떤 단어가 함께 등장하고, 문장 구조 정보가 어떠한 지에 따라 문장 전체 맥락이 달라지기 때문에 문장 혹은 문서가 표현하는 감성의 값이 달라질 수 있다.

기존의 단어 표현 방식인 ‘one-hot encoding’의 한계를 극복하기 위해, 연구자들은 단어 자체가 가지는 의미를 다차원 공간에 ‘벡터화’하는 방식을 고안하게 되었다. ‘비슷한 분포를 가진 단어들은 비슷한 의미를 가진다’는 언어학의 가정에 입각하여 1990년대부터 여러 모델이 제안되었으며 2000년대에 와서 neural network의 학습 원리에 기반을 둔 ‘NNLM(Neural Network Language Models)’(Y. Bengio et al., 2003)이 만들어졌다. ‘NNLM’은 n-gram의 언어모델을 신경망을 이용하여 구현한 후 목표단어의 앞뒤 단어를 입력 받아 목표단어들과 의미적으로 연관성을 갖도록 학습시킨다. 그러나 NNLM의 학습에는 많은 시간이 필요하며 대부분의 시간이 hidden layer와 output layer 사이의 계산에서 소요되는 문제가 있었다. 2013년에 제안된 Word2vec은 hidden layer를 감축하고 속도와 정확도를 높일 수 있는 forward/backward propagation 개념을 도입하여 신경망 구성의 단순함에 비해 학습된 단어의 벡터 표현에 대한 우수한 성능을 보여준다.(김나리 et al., 2017)

즉, 각 단어가 비슷한 문맥을 가진 것인지 단어 사이의 관계까지 나타내는 단어

표현을 구현한 것이 바로 신경망 기반의 연속 워드임베딩(continuous word embedding) 학습 모델인 Word2vec이다. Word2vec을 이용하여 각 단어들을 학습할 경우, 워드임베딩 공간에서 비슷한 문맥을 가진 단어들은 서로 가까운 공간 분포를 가지게 된다. Word2vec은 연속 워드임베딩을 표현하는 방법으로 두 가지의 모델 구조를 가지는데, Continuous Bag-Of-Words(CBOW)와 Skip-gram 모델 구조이다.

CBOW는 예측하려는 단어의 주변 단어들을 이용하여 단어를 예측하고, Skip-gram은 현재 주어진 단어를 이용하여, 그 단어 주변의 단어들을 예측하는 모델으로서 같은 양의 데이터에서 Skip-gram이 CBOW보다 학습량이 많기에, Word2vec을 Skip-gram 모델로 사용하는 경우가 많다. 본 연구에서도 Skip-gram 모델을 사용하고자 한다.

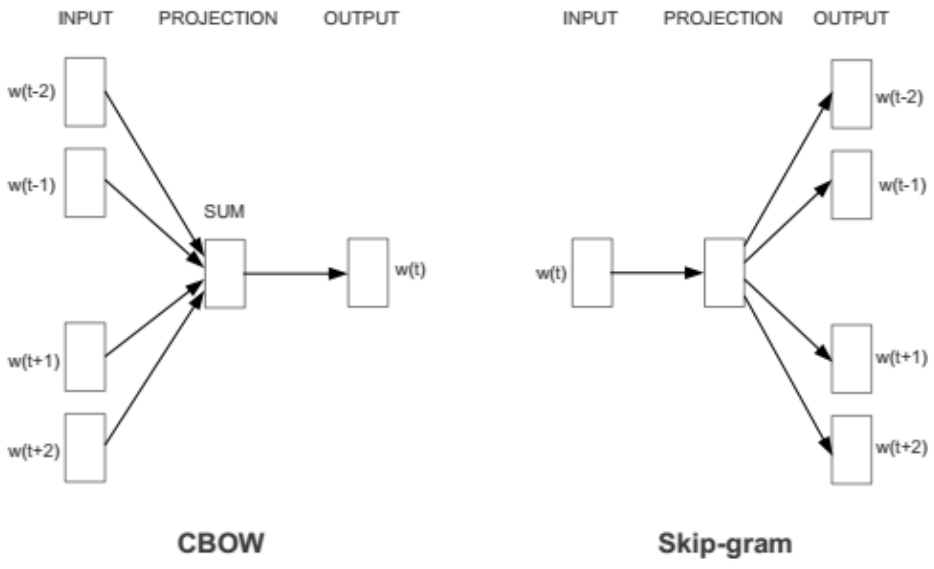

그림5. Word2vec의 두 가지 모델 구조

Skip-gram 모델

Skip-gram 모델은 주어진 단어를 이용하여 그 단어 주변의 단어들을 예측하는 신경망 모델로서 입력데이터를 이용해 1개의 은닉층(hidden-layer)만 갖는 간단한 신경망 훈련을 통해 은닉층에서의 가중치를 학습시켜 단어벡터를 얻는다. 실제 신경망 훈련은 최종 결과인 신경망은 사용하지 않고, 목적한 단어벡터를 얻기 위해 수행하게 되며, 모든 단어에 대해 주변 단어에 대한 확률을 학습시키게 된다. 윈도우 사이즈(window-size)를 지정할 수 있는데, 곧 해당 단어를 기점으로 주변의 몇 개의 단어까지 학습하는지를 나타내는 변수이다. 윈도우 사이즈를 '5'로 설정한다면, 해당 단어 앞에 있는 단어 5개와 뒤에 있는 단어 5개, 총 10개의 단어까지 학습한다는 의미이다. 신경망 학습은 훈련용 문서에 있는 단어집합을 만들어 학습을 시키게 된다.

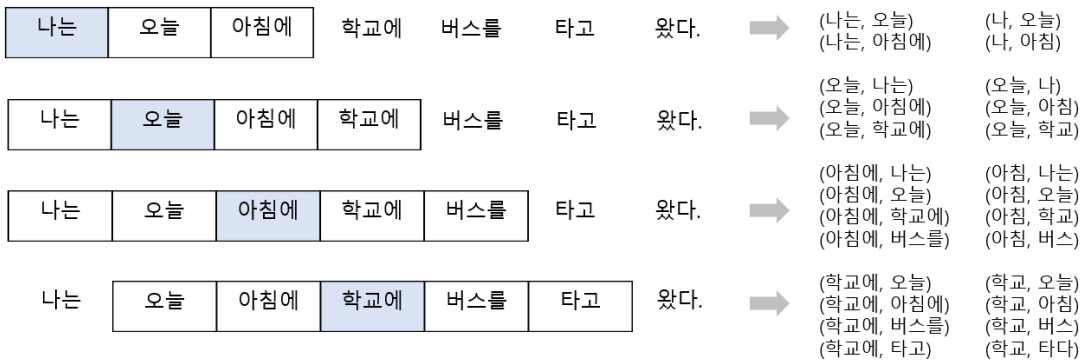

그림6. 좌측 데이터로부터 훈련샘플을 만들어 내는 Skip-gram 모델 예시

위의 **그림6**은 “나는 오늘 아침에 학교에 버스를 타고 왔다.” 문장에서 취할 수 있는 몇 가지 학습 샘플을 보여준다. 윈도우 사이즈가 2로 설정되었을 때, 해당 단어의 앞뒤로 있는 2개의 단어를 해당 단어와 함께 학습하게 된다. 한국어 문법 구조의 특성상 텍스트 전처리 단계인 형태소 분석 단계를 거치게 되는데, 이 때 조사나 어미를 제거한다. 가장 오른쪽 단계가 형태소 분석 단계를 거친 모습이다.

학습 후에 특정단어를 입력값으로 주면, 모든 단어들이 입력값과 얼마나 가까운지 확률로 출력값을 가지게 되어 많이 학습될수록 더 높은 확률을 출력하게 된다. 위의 예시 훈련샘플 학습을 통해 신경망이 만들어질 경우, 입력값이 ‘학교’일 때, 출력단어 ‘아침’과 ‘오다’ 단어 중에서는 ‘아침’ 단어가 윈도우 사이즈에 포함되어 더 많이 학습되었기에 ‘오다’ 단어보다 더 높은 확률값을 출력할 것이다. 위의 신경망 훈련과정을 신경망의 구조적으로 표현하면 아래 **그림7**과 같다.

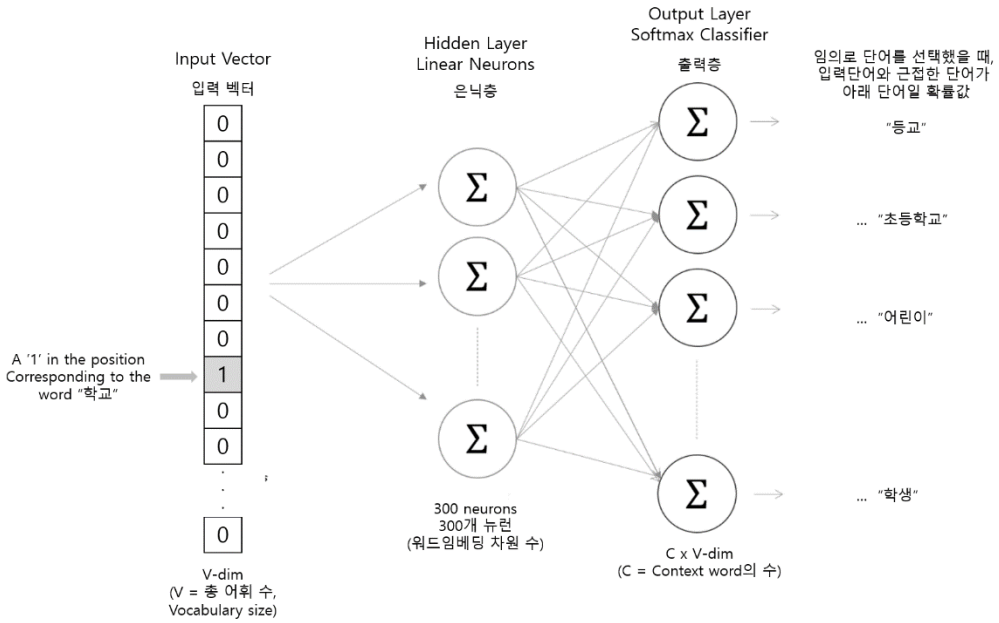

그림7. Word2vec 모델 신경망 학습 구조

Word2vec 모델을 만드는 신경망 구조는 크게 입력층(input layer)과 은닉층(hidden layer) 그리고 출력층(output layer)으로 구성되며, 입력층에서 입력값은 문서 내의 단어벡터값이며, 위의 그림7에서 예제는 입력값 단어를 “학교”라고 했을 때, ‘one-hot encoding’방식으로 만들어진 입력벡터인 ‘one-hot vector’이다. 신경망의 출력값은 문서의 모든 단어에 대해, 해당 단어가 입력단어 근처에 있을 확률을 표현하는 단일 벡터값이다. 위의 예시에서는 10,000개의 단어세트가 있기에 입력값 또한 10,000 차원의 벡터값을 가지고 있으며, 출력값은 문맥단어 수 X 단어세트의 차원이 된다. 또한, 300개의 특성(feature)을 가진 단어벡터를 학습한다고 가정할 때, 은닉층은 단어 전체의 개수인 10,000개의 행과 은닉층의 뉴런의 개수인 300개의 열이 있는 가중치 행렬로 표시된다. 여기서 특성의 수는 워드임베딩의 차원 수에 해당하며, 연구자가 정하는 하이퍼파라미터이다. 더 좋은 결과를 얻기 위해 실험적으로 최적값을 찾아야 한다.

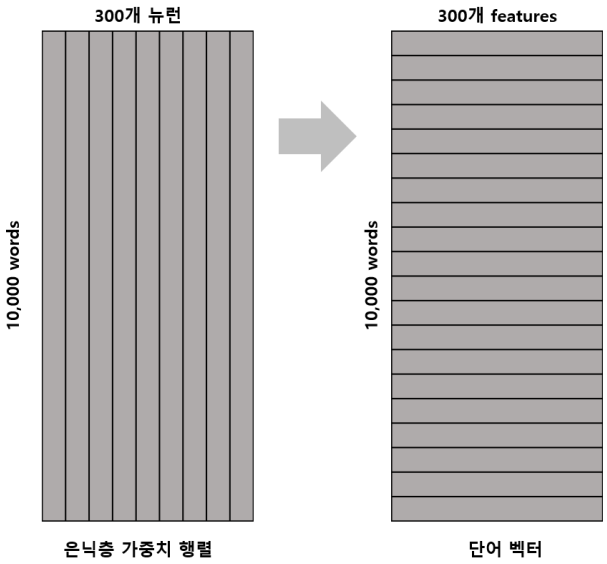

그림8. 은닉층 가중치 행렬에서 단어벡터로 변환하는 과정

위의 **그림8**은 가중치 행렬에서 나중에 구해야 하는 단어벡터로 변환되는 것을 나타낸 것이다. 위의 그림과 같이 가중치 행렬에서 변환시킨 단어벡터를 구할 경우 신경망 학습목표에 도달했기에 출력층은 필요하지 않는다. 은닉층에서 얻어진 단어벡터는 최종적으로 입력층의 단어벡터가 된다.

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

그림9. 입력층, 변환된 단어벡터, 출력층 행렬 예시

예시로 1x300 차원의 '학교'라는 단어벡터가 있을 경우, 단어벡터는 출력층에 주어지며, 출력층에서는 소프트맥스 회귀 분류기(Softmax Regression)를 사용하게 된다. 소프트맥스 회귀 분류에서는 각 출력층의 뉴런이 0과 1 사이의 값을 갖고, 모든 출력의 값의 합은 1이 되는데 특히, 각 출력층의 뉴런은 가중치 벡터를 가지게 된다. 가중치 벡터는 은닉층의 단어벡터에 곱해져, exp(x)를 거친다. 아래 **그림10**은 단어 '수업'의 출력 뉴런의 출력값을 계산하는 예이다.

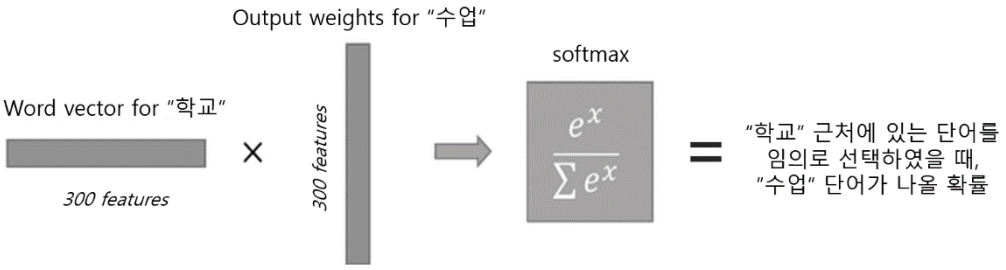

그림10. 단어 '수업'의 출력 뉴런 출력값을 계산하는 예

신경망은 입력 단어와 출력 단어의 상대적 위치에 대한 정보는 포함되지 않았기에 특정 단어가 앞 혹은 뒤에 올 확률의 차이는 학습하지 않는다. 또한 단어벡터를 이용하여 학습했기에, 가까이 출현하는 두 단어가 있다면, 단순히 가까운 정도만이 아니라 비슷한 문맥에 있는 경우 신경망은 두 단어에 대해 유사한 단어벡터를 학습할 가능성이 높아지게 된다.

C. SVM(Support Vector Machine)

SVM이란 기계학습 기법 중의 하나이다. 다양한 연구가 활발히 진행되어 매우 높은 인식 성능을 발휘하는 기법 중의 하나로 선을 구성하는 매개변수를 조정해서 요소들을 구분하는 선을 찾고 이를 기반으로 패턴을 인식하는 방법이다.
예를 들어, □와 ○ 두 가지 패턴이 있을 때, □와 ○의 패턴을 구분하는 방법을 찾는 것이 패턴 인식의 목표라면 이를 벡터로 나타내 평면 상에서 구분선을 그릴 수 있는데, 패턴의 경계가 되는 것이 '식별 평면'이다. 구분선을 정하는 기준이 있

으면, 새로운 패턴이 나타났을 때도 쉽게 분류할 수 있다.

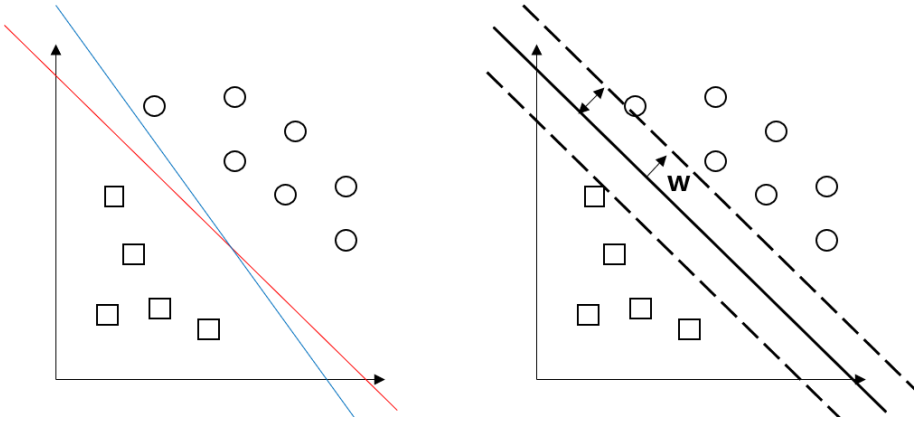

그림11. 선형 SVM

위의 **그림11**에서의 왼쪽 도표와 같이, 두 종류의 패턴 사이에 구분선이 여러 개 나올 수가 있는데, SVM은 이 패턴을 구분하는 중간을 지나는 선을 결정할 때, 위 오른쪽 도표와 같이 식별 평면에서 패턴들과의 거리(마진)를 최대로 만드는 것을 목표로 한다. 즉, 두 개의 클래스(class)를 분류하는 실선을 초평면이라고 하고, w를 초평면의 법선 벡터(mormal vector)라고 할 때, 이 초평면과 제일 가까운 두 클래스의 각 벡터들을 서포트 벡터(support vector)라 한다.

D=(xi,yi)라 하고 yi는 1 또는 −1의 값으로 xi가 어떤 클래스에 속해 있는지를 나타내는 값이라고 할 때, 만약 yi가 선형적으로 분리 가능하면 초평면은 아래와 같은 식으로 나타낼 수 있다. 곧, 마진의 최대값을 구하려면 w를 최소화 하는 최적화 문제를 해결하는 것이다.

$$w \cdot x - b = 0$$

$$yi(w \cdot x - b) \geq 2, \text{for all } 1 \leq i \leq n$$

Ⅳ. 제안모형 및 실험설계

본 연구에서는 자연어 처리를 위한 라이브러리인 Gensim을 사용하여, Word2vec을 구현하였고, 감성분류 평가는 python 기계학습 라이브러리인 scikit-learn의 SVM을 사용하였다.

A. 실험 데이터

실험 데이터로는 Python 라이브러리 BeautifulSoup를 이용하여 크롤링한 네이버 영화평 166,445개를 사용하였다. 영화평 리뷰 약 15만개 중 약 11만개는 Word2vec 모델 학습용 데이터로 사용하였고, 4만개는 감성분석 모형 학습 및 평가를 위한 데이터로 사용하였다. 실험 데이터 요약은 아래 **표5**와 같다.

표5. 실험 데이터 요약

	개수
Word2vec 학습용 데이터	네이버 영화평 116,351 개
감성분석 모형 학습 및 평가 데이터	네이버 영화평 리뷰 40,000 개 (긍정 부정 비율 1:1) -긍정 데이터 (10 점) 20,000 개 -부정 데이터 (1~4 점) 20,000 개

먼저, 영화평 약 11만개를 이용해 Word2vec 학습용 데이터세트를 구성하였으며, 이는 긍정 부정 비율이 약 7:3으로 나타났다. 아래 **표6**은 Word2vec 모델 학습용 데이터세트의 내용이다. 영화평 리뷰 데이터의 내용은 140자 이내로 평가자가 작성한 문장으로 이루어지며, 워드임베딩 공간을 형성하는 Word2vec모델 학습용 데이터에서는 평가자의 별점 라벨이 없는 데이터(unlabeled data)로만 구성하여 학습시킨다. 리뷰 1개는 문서 1개를 의미하며, 대부분의 리뷰의 경우 약 1~3문장으로 구성된다.

표6. Word2vec 모델 학습용 데이터 구조(raw data structure)

	내용
Word2vec 모델 학습용 데이터 (영화평 116,351 개)	완전 재밌어요!!! 10점만점!!! 굿굿 또 봐도 질리지 않는 영화! … 시간과 돈이 아까운 영화, 노잼 ㅡㅡ 휴...안타깝네요..정말... 흥행1위라 봤는데 ㅠㅠ …

감성분석을 수행할 학습 및 평가 데이터는 영화평 리뷰 4만개이며, 학습 데이터와 평가 데이터의 비율을 각각 75%와 25%로 구성하였다. 또한 본 실험 결과에 대해 신뢰성을 확보하기 위해 10-Fold 교차검증을 수행하였다. 학습 및 평가 데이터의 구조는 아래 **표7**와 같다. 별점 10점으로 이루어진 긍정 데이터 2만개와 별점 1~4점으로 이루어진 부정 데이터 2만개로 구성되며, 별점을 감성 값 라벨

(긍정:1, 부정:0)로 변환할 때, 연구자의 판단으로 별점 1~4점을 부정 라벨로, 10점을 긍정 라벨로 치환하여 사용하였다.

표7. 감성분석 학습 및 평가 데이터 구조 (labeled raw data structure)

리뷰내용	별점	라벨 (부정:0 긍정:1)
휴...안타깝네요..정말...	1	0
시간과 돈이 아까운 영화	1	0
…		0
흥행1위라 봤는데 ㅠㅠ	2	0
후우..돈이아깝다...	2	0
…	…	0
지루하기 짝이 없었음	3	0
별로…	3	0
…		0
에휴…	4	0
언제 끝나냐 시간만 쳐다봤다 ㅡㅡ	4	0
…		
완전 재밌어요!!!	10	1
10점 만점!!!	10	1
굿굿!	10	1
또 봐도 질리지 않는 영화!	10	1
역시!! 보는 내내 흥미진진 했어요	10	1
계속 보고싶은 영화 위로되는 영화 정말 재밌게봤다	10	1
완전 무서웠다ㅠㅠ	10	1

B. 제안모형

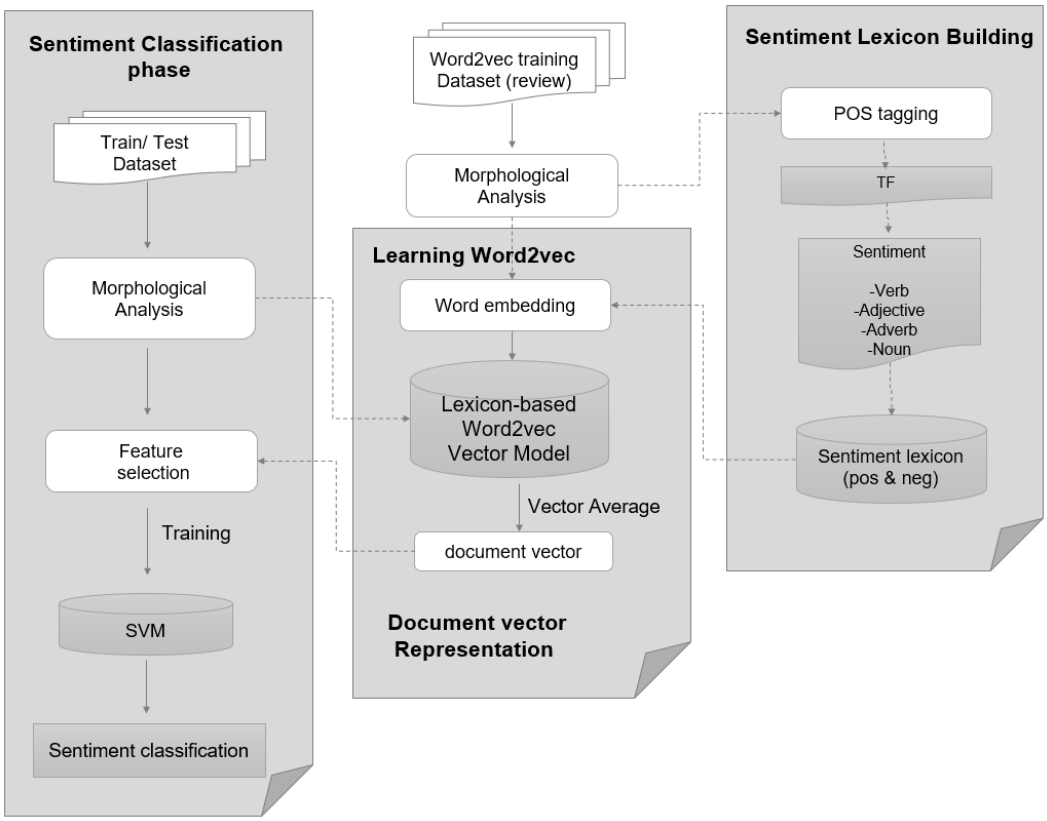

그림12. 제안모형(Overview of proposed model)

본 연구에서 제안하는 모형은 **그림12**와 같으며, 다음과 같은 3단계로 구성된다.

1. 감성사전 구축
2. 감성사전 기반 Word2vec 모델 구축 및 문서 벡터 구현
3. 감성 분류

Word2vec 학습 데이터에서 형태소 분석 단계를 거친 후 감성사전을 구축하고, 구축된 사전을 워드임베딩 학습 단계에 활용해 감성사전 기반 Word2vec 모델을 구축한다. 이후 구축한 모델을 이용하여 문서 자질(document feature)을 추출하여 이를 감성분석 단계에서 자질로 선택하고, 분류기(classifier) SVM을 사용하여 감성 분류를 진행한다.

C. 실험설계

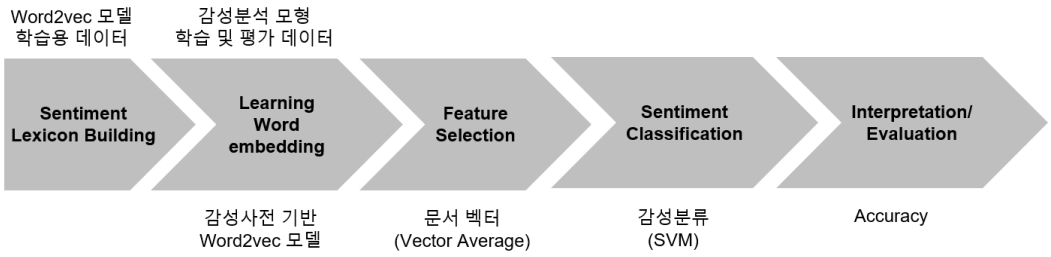

그림13. 감성분석 프로세스(Process of the Sentiment Analysis)

본 연구에서 진행되는 실험설계는 위의 **그림13**과 같다.

1. 감성사전 구축

감성사전의 경우 표5,6에서 소개한 Word2vec 학습용 데이터를 이용하였으며, 1차로 형태소 분석 단계를 거친 이후, POS tagging(Part-Of-Speech tagging, 품사 부착) 및 TF(단어 빈도)를 계산하여, 상위 TF 기준으로 품사가 동사/형용

사/부사/명사인 어휘 2만개를 추출하였다. 이 중 감성의 극성이 긍정 혹은 부정인 어휘 약 10,000개를 선택하여 아래 **표8-1**과 같이 감성 어휘 사전을 구축하였다.

표8. 감성 어휘 사전 구성 예시

긍정 어휘 (Verb/Adjective/Adverb/Noun)	**부정 어휘** (Verb/Adjective/Adverb/Noun)
재밌다 / 좋다 /	자다 / 아깝다 /
재미있다 / 무섭다 /	아쉽다 / 지루하다 /
잘 / 괜찮다 /	재미없다 / 버리다 /
잘하다 / 기대하다 /	미치다 / 힘들다 /
좋아하다 / 대단하다 /	나쁘다 / 뻔하다 /
생각나다 / 느껴지다 /	어렵다 / 부끄럽다/
웃다 / 멋지다/	멍청하다 / 피곤하다 /
완전하다 / 끌리다 /	쓸데없이/ 식상하다 /
빛나다 / 설레다 /	어둡다 / 틀리다 /
사랑스럽다 / 최고 /	별로 / 알바 /
재미 / 느낌 /	최악 / 실망 /
기대 / 대박 /	쓰레기 / 노잼 /
추천 / 굿 /	미끼 / 의심 /
여운	억지

2. 감성사전 기반 Word2vec 모델 구축 및 문서 벡터 구현

형태소 분석된 Word2vec 모델 학습용 데이터로부터 감성사전에 있는 어휘만 추출하여 Word2vec 모델을 학습시켰고, 이후 구축 Word2vec 모델로부터 감성 분석 학습 및 평가 데이터의 문서 벡터를 구현하였다.

2.1 형태소 분석

문서를 어절단위로 분리하고, 어미/조사/구두점은 제외하였으며, 형태소 분석이 수행된 후 **표9**에서와 같이 단어 단위로 나열된 것을 볼 수 있다.

표9. 형태소 분석

문서	짜증나고 지루하고 재미도없고돈아깝고 시간버리고...								
	짜증	나다	지루하다	재미	없다	돈	아깝다	시간	버리다

2.2 감성사전 기반 Word2vec 모델 구축

일반적인 Word2vec 모델은 학습용 데이터를 형태소 분석하여 어절 분리된 모든 단어들에 대해 신경망 학습을 하게 된다. 본 연구에서는 모델 구축 시, Skip-gram 방식을 이용하였고, 모델 구축을 위한 학습 파라미터는 아래 **표10**과 같다.

벡터 차원 수(size)는 200으로, 최소 단어 빈도수(min_count)는 5로 고정하였고, 학습 윈도우 크기(window)는 3과 5로 설정하여 비교 실험하였다.

표10. Word2vec Skip-gram 모델 학습 파라미터

size	window	min_count
200	3, 5	5

감성사전 기반 Word2vec모델 구축 과정은 아래 **그림14**와 같은 알고리즘으로 표현할 수 있다. 미리 구축한 감성사전을 이용하여, 형태소 분석된 데이터세트에서 감성사전에 포함된 어휘들을 조사하여, 감성사전에 포함된 어휘들만 따로 추출한 뒤, 모든 단어가 아닌 감성사전에 등록된 단어들만 Word2vec 신경망 학습에 사용함으로써 감성사전 기반 Word2vec을 생성하였다. 본 연구는 모든 어휘들을 학습에 사용한 일반 Word2vec의 단어벡터 값을 자질로 사용한 SVM 모형의 결과와 본 연구에서 제안한 감성사전 기반 Word2vec의 단어벡터 값을 사용한 결과를 비교하는 데 연구의 초점이 있다.

Lexicon based Word2vec model algorithm

```
review <- review text in review data
lex <- lexicon data

while True:
        line <- review.line()
        if not line do break
        mor_line <- twitter.pos(line)
        for each w in mor_line do
          for lexicon in lex:
              if w in lexicon:
                  model_dataset<-review_morpheme.write(w)
              end if
          end for
        end for

lexicon based word2vec model <- Word2Vec(model_dataset)
```

그림14. 사전기반 Word2vec 모델 구축 알고리즘

2.3 문서 벡터 구현

제안모형을 기반하여 감성분석 단계에서 필요한 각 문서의 자질(document feature)을 구성해야 한다. 문서는 곧 단어로 이루어져 있는데, 단어 표현을 통해 문서를 표현할 수 있는 방법을 생각해야 한다. 이는 곧 문서 표현(Document representation)이 된다. Word2vec 모델을 구축하였기에, 각 단어는 벡터 값으로 모두 표현 가능하며, 문서 또한 벡터로 표현하는 것을 고려할 수 있다.

Word2vec을 이용해 문서 자질을 구성하는 방법으로는 각 문서를 구성하고 있는 단어들의 벡터 값들의 평균을 이용한 연구가 있었다.(Roy Bayot et al., 2016;

Yonatan Belinkov et al., 2015; 이동엽 et al., 2017) 문서를 구성하는 단어 개수가 n개이고, 각 단어의 벡터값을 v(i)라고 표현할 때, 문서 벡터는 아래와 같은 식으로 표현할 수 있다.

$$\frac{1}{n}\sum_{i=1}^{n} v(i)$$

본 연구에서도 이와 같은 방법을 이용하여 감성분석에 필요한 문서 자질을 구성하고자 한다. 아래 **표11,12,13**은 본 연구에서 사용된 데이터를 이용해 문서 벡터를 구축하는 과정을 예시로 나타낸 것이다

표11. 자질 선택1 : 문서 내 단어벡터 구현

	d_1	d_2	d_3	d_4	…	d_{198}	d_{199}	d_{200}
짜증	2.95	2.51	6.30	0.01	…	0.01	−0.37	8.83
나다	−10.38	−3.92	−0.82	−5.42	…	−1.09	−5.25	0.79
지루하다	−1.36	0.86	−2.16	0.52	…	4.85	1.84	2.14
돈	−4.88	6.37	−1.81	−4.69	…	3.63	6.00	−2.23
아깝다	0.91	−4.04	−2.28	1.07	…	−3.94	3.26	−7.45
시간	−1.36	0.86	−2.16	0.52	…	4.85	1.84	2.14
버리다	−4.88	6.37	−1.81	−4.69	…	3.63	6.00	−2.23

표12. 자질 선택2 : 단어벡터의 평균을 통한 문서 벡터 구현

	d_1	d_2	d_3	d_4	…	d_{198}	d_{199}	d_{200}
avg	0.11	−0.5	−0.28	0.13	…	−0.49	0.40	−0.93

표13. 자질 선택3: 총 문서벡터 자질(feature) 구현

	d_1	d_2	d_3	d_4	…	d_{198}	d_{199}	d_{200}
문서1	0.11	−0.5	−0.28	0.13	…	−0.49	0.40	−0.93
문서2	2.95	2.51	6.30	0.01	…	0.01	−0.37	8.83
문서3	−10.38	−3.92	−0.82	−5.42	…	−1.09	−5.25	0.79
문서4	−1.36	0.86	−2.16	0.52	…	4.85	1.84	2.14
문서5	−4.88	6.37	−1.81	−4.69	…	3.63	6.00	−2.23
문서6	0.91	−4.04	−2.28	1.07	…	−3.94	3.26	−7.45

3. 감성 분류

위의 단계를 거쳐 최종적으로 감성분류를 진행한다. 본 연구에서 제안하는 모형이 타당한지 확인하고자 모형1, 모형2 감성 분류를 진행하여 정확도를 비교한다. 모든 실험에서 사용되는 감성 분류 모형 학습 및 평가 데이터는 동일하며, SVM 분류기를 사용한다. 모형1에서는 일반적인 Word2vec을 적용한 자질을 이용한 감성 분류를 수행하며, 모형2에서는 제안모형인 감성사전을 이용한 Word2vec 자질을 이용한 감성 분류를 수행한다.

표14. 감성 분류 실험 모형 분류

모형1 **w2v**	Word2vec 자질을 이용한 감성 분류
모형2 **Lexicon based w2v**	감성사전 기반 Word2vec 자질을 이용한 감성 분류

D. 평가

감성분석 모형 평가를 위한 지표로는 아래와 같은 TP/TF/FP/FN에서 정확도를 계산하여 이를 기준으로 모형 간 결과를 비교하였다.

TP : 정답이 True 인데 True로 예측한 경우

TF : 정답이 False인데 False로 예측한 경우

FP : 정답이 False인데 True로 예측한 경우

FN : 정답이 True인데 False로 예측한 경우

정확도(Accuracy) $\frac{TP+TN}{TP+TN+FP+FN}$

Ⅴ. 연구결과 및 해석

표5,6에서 설명한 것과 같이 감성 분류 모형 학습 및 평가 데이터는 총 4만개의 리뷰 데이터이며, 긍정 부정 리뷰 데이터가 각각 2만개로 구성된다. 모형1 w2v(Word2vec 자질을 이용한 감성 분류)과 모형2 Lexicon based w2v(감성사전 기반 Word2vec 자질을 이용한 감성 분류)의 성능을 감성 분류 정확도(Accurracy) 지표로 비교한 결과는 아래 **표15**와 같았다.

표15. 모형1과 모형2의 분류 정확도(%) 비교

Fold	모형1 w2v		모형2 Lexicon based w2v	
	window = 3	window = 5	window = 3	window = 5
1	64.1	63.21	66.32	66.25
2	63.96	62.33	66.03	64.89
3	62.99	62.55	65.93	65.31
4	63.55	62.24	66.45	65.92
5	63.98	62.01	65.74	66.63
6	63.04	63.83	66.03	64.74
7	62.8	62.63	65.8	64.36
8	62.89	63.03	65.49	64.4
9	63.19	62.18	65.71	65.18
10	64.24	63.47	66.09	65.34
평균	**63.47**	62.75	**65.96**	65.3

최적의 감성 분류 시스템을 구축하기 위해 모형별로 윈도우의 크기를 3과 5로 조정해가면서 실험을 하였고, 모형1에 대한 평균 정확도는 윈도우 크기가 3일 때, 약 63.47%로 나타났고, 윈도우 크기가 5일 때, 약 62.75%를 보였다. 모형2의 평균 정확도의 경우, 윈도우 크기가 3일 때, 약 65.96%이며 윈도우 크기가 5일 때, 약 65.3%를 보임으로, 두 모형 모두 윈도우의 크기가 5일 때보다, 3인 경우에 조금 더 높은 성능을 보여 윈도우 크기에 상관없이 본 연구에서 제시한 모형2가 모형1보다 약 2.5% 더 높은 성능을 보임을 알 수 있었다.

또한 감성 분류 모형 학습 및 평가 데이터의 수에 따라 감성 분류 결과가 어떻게 달라지는지 확인하기 위해 학습 및 평가 데이터의 개수를 총 3만 개(긍정 리뷰 1만 5천 개/ 부정 리뷰 1만 5천 개)로 조정하여 추가 실험을 진행하였다. 아래 **표16**은 학습 및 평가 데이터의 개수에 따른 감성 분류 평균 정확도에 대한 실험 결과이다. 학습 및 평가 데이터의 개수가 3만개의 경우, 4만개일 때와 마찬가지로 윈도우 크기가 3일 때, 모형1과 모형2에서 가장 높은 감성 분류 정확도인 65.52%와 67.98%를 보였으며 윈도우의 크기에 상관없이 모형2에서 모형 1보다 약 2.5% 더 높은 성능을 보였다. 또한 전체적으로 평균 정확도가 학습 및 평가 데이터의 개수가 4만개였던 이전 실험보다 약 2% 향상된 모습을 나타냈다.

표16. 학습 및 평가 데이터의 개수에 따른 모형 분류 평균 정확도(%)

학습 및 평가 데이터 개수	모형1 w2v		모형2 Lexicon based w2v	
	window = 3	window = 5	window = 3	window = 5
3만	**65.52**	64.97	**67.98**	67.62
4만	**63.47**	62.75	**65.96**	65.3

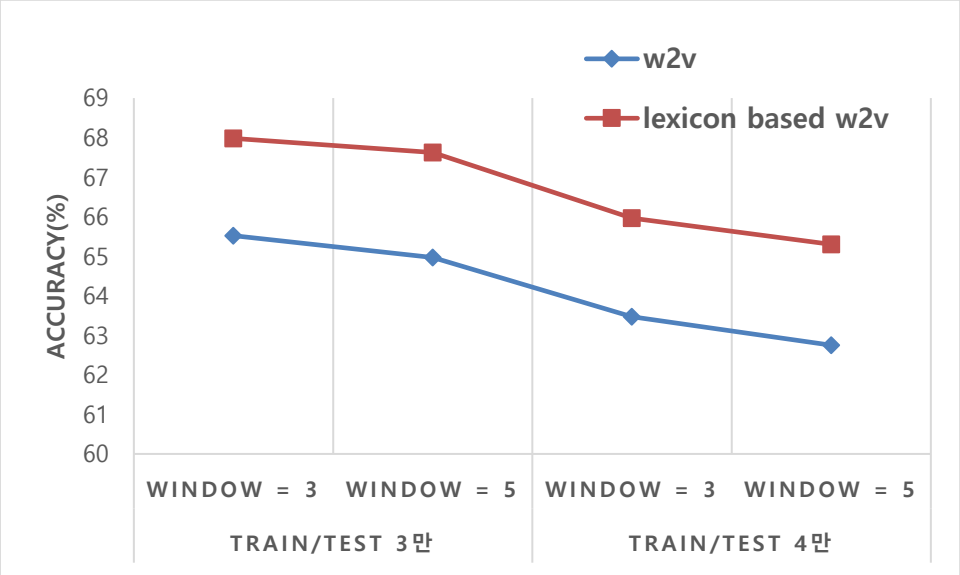

그림15. 감성 분류 정확도 비교

그림15는 위 두 실험에 대한 결과를 도표로 나타낸 것이다. 감성 분류 학습 및 평가 데이터의 개수에 따라 분류 정확도에 변화가 있었지만, 데이터의 개수나 윈도우 크기에 상관없이 모형1보다 모형2에서 꾸준히 더 높은 감성 분류 정확도가 나타난 것을 볼 수 있었기에 본 연구의 제안 모형이 감성 분류에 있어서 더 좋은 성능을 나타냄을 확인할 수 있었다. 또한 두 모형 모두 데이터의 개수에 상관없이 윈도우 크기가 더 작을 때, 감성 분류 정확도가 높았던 것을 확인할 수 있었다.

한편, 추가 실험에서는 감성 분류 정확도가 이전 실험보다 전체적으로 약 2%가 증가하였는데, 이는 실제 데이터를 확인해 본 결과 학습 및 평가 데이터의 개수가 증가할수록, 문서 데이터 중 길이가 긴 리뷰 데이터의 개수도 같이 증가함을 볼 수 있었다. 길이가 짧은 문서의 경우 단일 문장 혹은 단어로 '좋아요', '꿀잼'처

럼 명확하게 감성이 표현되어 있었지만, 길이가 긴 문서의 경우 **그림 16-1**처럼 감성의 극성을 분명하게 표현하기 보다 여러 감성을 섞어서 표현하거나,

> “7점은 아니고 7.5 판틴의 노래가 너무 좋았음 ㅎㅎ 러닝타임 내내 한장면도 지루하지 않고 집중해서 봤는데 아쉬운점은 장발장이 쫓긴 그 긴세월을 가볍게 다룬게 아쉽다.. 얼핏보면 혁명군영화인듯 ㅋ 혁명장면은 몸에 전율을 돋게하지만 스토리가 중구난방.. ”

그림16-1. 영화 리뷰 예시

그림 16-2처럼 한 문서 내에 다양한 개체에 대한 개별 감성이 존재하므로 문서 단위 감성 분류 모형을 통해 감성을 예측하기 어려워 더 낮은 분류 정확도를 보였음을 확인할 수 있었다.

> “앤 헤서웨이 연기력과 가창력은 최고였음!! 아만다는 그냥 이쁨 ㅋㅋ 휴잭맨은 역시 잘하고.. 의외로 러셀크로우가 안 어울린다는 느낌.. 아만다 나오기 전까지만 잘 보다가 급 루즈해짐.. 볼만하지만 조금 지루한 감이 있음. 노래들은 멋있었어요. ”

그림16-2. 영화 리뷰 예시

Ⅵ. 결론 및 논의

본 연구에서는 기존의 Word2vec 모델이 단어 표현 기법으로 우수하여 텍스트 분석 분야에서 각광을 받고 있지만, 감성분석 분야에 Word2vec 모델을 적용할 경우 통사적인 단어의 쓰임새 정보만 학습하기에 실제로 감성의 극성이 다른 단어도 워드임베딩 공간에 사상시키는 어려움이 있다는 점을 주목하였다. 이러한 문제를 완화하기 위해, 의미 정보를 결합한 Word2vec을 감성분석에 자질로 사용하는 감성사전 기반 Word2vec 자질을 이용한 감성 분류 시스템을 제안하였고, 이는 일반 Word2vec을 이용한 자질의 문제점을 보완할 수 있을 것이란 가정 하에 실험을 진행하였다.

리뷰 데이터로부터 TF를 계산하여 상위 TF 2만개를 기준으로 형태소 분석 및 POS태깅을 거쳐 감성 어휘를 분류하였고, 1만개의 감성 어휘를 포함하는 연구자 직관 중심의 감성사전을 구축하여, 이를 Word2vec을 이용한 문서 자질을 만드는데 이용하였다. 제안모형을 통해 실험한 결과, 일반적인 Word2vec 자질을 이용한 감성 분류 모형과 비교했을 때, 학습 및 평가 데이터의 개수에 상관없이 제안모형에서 약 2.5%이상 더 높은 분류 정확도를 보임으로 제안모형의 성능이 더 우수함을 확인할 수 있었다.

본 연구의 한계점으로는 모형 학습 및 평가 데이터가 증가할수록 제안모형의 분류 정확도가 떨어진다는 점으로, 길이가 긴 리뷰 문서에서 여러 개체 및 그에 대한 감성이 개별적으로 존재하기에 문서 단위 감성 분류 모형인 본 연구의 제안모형을 통해서는 예측하기가 어려운 부분이 있다는 점이다. 따라서 향후 길이가 긴 문서에 한해 속성 단위 분석을 추가적으로 진행한다면, 이러한 부분을 더욱 보완할 수 있을 것이라 기대한다.

본 연구는 기존의 감성분석 기법인 사전 기반 기법과 기계학습 기법을 혼합하여 감성분석을 수행하므로 연구자가 구축한 감성사전을 활용하여 감성의 극성 정보를 더욱 반영함으로써 분석의 정확도를 개선시킨 것에 의의가 있었다. 또한, Word2vec을 적용한 감성분석 연구의 관점에서 바라봤을 때, 워드임베딩 기법이 감성분석에 있어서 문장 구조적 통사적 정보를 학습하여 Word2vec이 감성의 극성 정보를 예민하게 반응하지 못하는 점이 있기에, 감성사전을 결합하여 기존의 Word2vec 자질의 한계점을 보완할 수 있는 연구를 수행했다는 점에서 의의가 있다.

참 고 문 헌

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.79-86.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ACL p.417-424.

Dave, K., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. WWW p.519-528.

Bengio, Y., Ducharme, R., & Vincent, P.(2003). A neural probabilistic language model. Journal of Machine Learning Research, vol. 3, pp.1137-1155

Pang, B., & Lee, L. (2004). A Sentiment Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL no. 271.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. HLT/EMNLP, p.347-354.

Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. Advances in knowledge discovery and data mining, pp301-311.

Meena, A., & Prabhakar, T. (2007). Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis. ECIR 2007, LNCS 4425, pp. 573 - 580.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1-135.

Martineau, J., & Finin, T. (2009). Delta TFIDF : An Improved Feature Space for Sentiment Analysis. Proceedings of the Third International ICWSM Conference.

Rao, D., & Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. Proceedings of the 12th conference of the European Chapter of the Association for Computational Liguistics.

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing (pp. 627-666). Chapman and Hall: CRC Press.

Andrew, L, M., Raymond, E., Daly, P., Huang, D., Andrew, Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. ACL, p142-150.

Jo, Y., & Oh, A. (2011). Aspect and Sentiment Unification Model for Online Review Analysis. ACM 978-1-4503-0493-1.

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

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4):82-89.

Moraes, R., Francisco, J., & Neto, W. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40 (2013) 621-633.

Xue, B., Fu, C., & Shaobin, Z. (2014). A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec. IEEE.

Ortigosa, A., Martin, J., & Carro, R. (2014). Sentiment analysis in Facebook and

its application to e-learning. Computers in Human Behavior 31 (2014). 527-541.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. ACL, p.1555-1565.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data 2:5.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2016). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications 42(4).

Mayank, D., & Padmanabhan, K. (2016). Multi-Sentiment Modeling with Scalable Systematic Labeled Data Generation via Word2Vec Clustering. IEEE.

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems 108 (2016) 110-124.

Tripathy, A., Agrawal, A., & Kumar, S. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications 57, 117-126.

Giatsoglou, M., Vozalis, M., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. (2017). Sentiment analysis leveraging emotions and word Embeddings. Expert System with Applications 69 (2017) 214-224.

Zhang, X., & Yu, Q. (2017). Hotel Reviews Sentiment Analysis Based on Word Vector Clustering. IEEE.

Alshari, E., & Azman, A. (2017). Improvement of Sentiment Analysis based on Clustering of Word2Vec Features. Expert Systems Applications.

Paredes-V, M., Colomo-P, R. Salas-Z, M. & Valencia-G, R. (2017). Sentiment Analysis in Spanish for Improvement of Products and Services : A Deep Learning Approach. Hindawi Scientific Programming 1329281, 6p

Yu, L., Wang, J., & Lai, K.(2018). Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. IEEE 26-3.

김승우, & 김남규. (2014). 오피니언 분류의 감성사전 활용 효과에 대한 연구. 한국지능정보시스템학회 학술대회논문집, 2013-11 :121-128.

정의석, & 박전규. (2016). 워드 임베딩과 품사 태깅을 이용한 클래스 언어모델 연구. KIISE Transactions on Computing Practices, Vol.222, No.7, pp.315-319.

서덕성, 모경현, 박재선, 이기창, & 강필성. (2017). 워드임베딩과 그래프 기반 준지도 학습을 통한 한국어 어휘 감성 점수 산출. Journal of the Korean Institute of Industrial Engineers.

이동엽, 조재춘, & 임희석. (2017). 워드임베딩을 이용한 아마존 패션 상품 리뷰의 사용자 감성 분석. 한국융합학회 vol8-4, pp1-8.

임미영, & 강신재. (2017). 의미 정보가 강화된 워드 임베딩을 통한 감성 분석. Asia-pacific Journal of Multimedia Services Convergent Vol.7. no.2, pp 321-329.

김나리, & 김형중. (2017). 연관법령 검색을 위한 워드 임베딩 기반 Law2vec 모형 연구. 디지털콘텐츠학회논문지 Vol. 18, No. 7, pp. 1419-1425.

ABSTRACT

A sentiment classification system using Lexicon-based Word2vec features

Myeong EunJin

Major in Big Data Analysis

The Graduate School of Ewha Womans University

Sentiment Analysis is Natural Language Processing that aims to distinguish positive, negative, and neutral sentiments of a specific object from various text data such as reviews, news, and blogs. Sentiment analysis techniques are largely classified into lexicon-based approach and machine-learning approach. In the case of lexicon-based approach, general-purpose sentiment lexicon or a lexicon created by a researcher is used for sentiment analysis. On the other hand, when a large amount of documents are targeted, machine-learning approach is generally applied. In the relevant studies, there have been various statistical explorations to improve the analysis accuracy in the feature selection stage, which is an intermediate stage of the machine-learning approach. Also, there have been studies that compare the performance of sentiment analysis models by applying various classifiers required in the classification stage, or more recently, studies that apply and classify Word2vec, a representative word embedding methodology, to the feature selection stage of the machine-learning approach.

Word2vec is continuous word embedding based on a nerve network. In the case of word learning using Word2vec, words with similar contexts in the word-embedding space have close

spatial distribution, which makes differences in the spatial distribution of the words depending on the kinds of documents used in a field. The majority of the sentiment analysis studies applying Word2vec are to reduce the dimension by applying the clustering technique to solve the high dimension problem of Word2vec in the feature selection stage. The Word2vec model only learns the usage information such as the syntactic information of the word in a sentence, Thus, the limitations of expressing sentiment with different polarity in a similar word embedding vector can cause a problem of impairing the accuracy of sentiment analysis. Therefore, studies suggesting word embedding that integrate the semantic information of a word have emerged.

This study aimed to propose a new sentiment classification model applying Word2vec which integrates sentiment lexicon information to sentiment analysis. In other words, the sentiment lexicon constructed by the researcher was combined with the Word2vec model learning stage to establish the sentiment lexicon-based Word2vec model, and sentiment analysis was performed by building the features of the documents to be analyzed. When the sentiment analysis was performed by the method proposed in this study, the accuracy improved by about 2% compared to the existing method. As a result, the validity of the proposed model was confirmed.

By combining the lexicon-based technique and machine learning technique to conduct sentiment analysis, this study has implications for enhancing the analytical performance to reflect the polarity information of sentiments using the sentiment lexicon constructed by the researcher. Moreover, from the perspective of sentiment analysis study using Word2vec, the word embedding method is excellent as a word expression method, however, it does not sensitively react to sentiment polarity information to learn sentence structural syntactic information in sentiment analysis Accordingly, this study has implications in that it can complement the limitations of the existing Word2vec features by combining the sentiment lexicon.