http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.1, February 2020, pp.13-33
https://doi.org/10.7838/jsebs.2020.25.1.013

회선 신경망을 활용한 자모 단위 한국형 감성 분석 모델 개발 및 검증

Development and Validation of the Letter-unit based Korean Sentimental Analysis Model Using Convolution Neural Network

성원경(Wonkyung Sung)*, 안재영(Jaeyoung An)**, 이중정(Choong C. Lee)***

초 록

본 연구는 자모 단위의 임베딩과 회선 신경망을 활용한 한국어 감성 분석 알고리즘을 제안한다. 감성 분석은 텍스트에서 나타난 사람의 태도, 의견, 성향과 같은 주관적인 데이터 분석을 위한 자연어 처리 기술이다. 최근 한국어 감성 분석을 위한 연구는 꾸준히 증가하고 있지만, 범용 감성 사전을 사용하지 못하고 각 분야에서 자체적인 감성 사전을 구축하여 사용하고 있다. 이와 같은 현상의 문제는 한국어 특성에 맞지 않게 형태소 분석을 수행한다는 것이다. 따라서 본 연구에서는 감성 분석 절차 중 형태소 분석을 배제하고 초성, 중성, 종성을 기반으로 음절 벡터를 생성하여 감성 분석을 하는 모델을 개발하였다. 그 결과 단어 학습 문제와 미등록 단어의 문제점을 최소화할 수 있었고 모델의 정확도는 88% 나타내었다. 해당 모델은 입력 데이터의 비 정형성에 대한 영향을 적게 받으며, 텍스트의 맥락에 따른 극성 분류가 가능하게 되었다. 한국어 특성을 고려하여 개발된 본 모델이 한국어 감성 분석을 수행하고자 하는 비전문가에게 보다 쉽게 이용될 수 있기를 기대한다.

ABSTRACT

This study proposes a Korean sentimental analysis algorithm that utilizes a letter-unit embedding and convolutional neural networks. Sentimental analysis is a natural language processing technique for subjective data analysis, such as a person's attitude, opinion, and propensity, as shown in the text. Recently, Korean sentimental analysis research has been steadily increased. However, it has failed to use a general-purpose sentimental dictionary and has built-up and used its own sentimental dictionary in each field. The problem with this phenomenon is that it does not conform to the characteristics of Korean. In this study, we have developed a model for analyzing emotions by producing syllable vectors based on the onset, peak, and coda, excluding morphology analysis during the emotional analysis

이 논문 또는 저서는 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A5A2A01034324).

* First Author, Researcher, DTX Center, LG Electronics, Master, Graduate School of Information, Yonsei University(wonkyung.sung@lge.com)
** Co-Author, Doctoral Student, Graduate School of Information, Yonsei University(jaeyoungan@yonsei.ac.kr)
*** Corresponding Author, Professor, Graduate School of Information, Yonsei University(cclee@yonsei.ac.kr)
Received: 2019-09-16, Review completed: 2020-01-08, Accepted: 2020-01-21

procedure. As a result, we were able to minimize the problem of word learning and the problem of unregistered words, and the accuracy of the model was 88%. The model is less influenced by the unstructured nature of the input data and allows for polarized classification according to the context of the text. We hope that through this developed model will be easier for non-experts who wish to perform Korean sentimental analysis.

키워드 : 감성 분석, 자모 단위 기반 모델, 회선 신경망
Sentimental Analysis, Letter-Unit Based Model, Convolutional Neural Network

1. 서 론

소셜 네트워크 서비스 이용 증가에 따라 비정형 텍스트 데이터로부터 새롭고 유용한 정보 발견에 대한 중요성이 부각되고 있다. 이에 비정형 텍스트 데이터 활용을 위한 방법이 지속적으로 연구되고 있으며, 그중 대표적인 방법이 텍스트 마이닝(Text mining)이다. 텍스트 마이닝은 통계학 및 기계학습 등 기반으로 자연어 처리 기술을 활용하여 비정형 텍스트 데이터로부터 의미 있는 정보를 발견하기 위한 목적으로 활용되고 있다[14, 26]. 2011년부터 2014년까지 텍스트 마이닝 연구 현황을 확인한 결과 11% 증가하였으며[30], 지금까지도 텍스트 마이닝 연구는 끊임없이 증가되고 있다.

텍스트 마이닝 방법 중 감성 분석은 텍스트의 긍정, 부정에 대한 의견을 판단하는 분석 기법이다[41]. 감성 분석은 텍스트 마이닝에서 가장 많이 활용되고 있는 방법으로 컴퓨터 관련 분야뿐만 아닌 다양한 분야(사회과학, 의료, 마케팅 등)에서도 많이 사용되고 있다[30]. 하지만 영어 감성 분석에 비해 한국어의 경우 몇 가지 제한이 존재한다. 먼저, 공개된 한국어 감성 사전 미흡 및 알고리즘이 매우 빈약하다. 일반적으로 범용 감성 사전에서는 텍스트에 따라 언어별 감성이 변하는 언어 중의성 문제를 해결할 수 없으며[27], 특정 분야에서 사용하는 감성 사전 또는 알고리즘은 별도로 학습을 시켜야 한다. 영어 감성 분석에서는 'SentiWordNet'이라는 표준 감성 사전이 존재하지만, 언어 중의성 문제를 해결하기 위해 분야마다 사용하는 감성 사전이 조금씩 다른 상황이다[2, 16]. 한국어의 경우 Kim et al.[28]이 제안한 KOSAC(Korean Sentiment Analysis Corpus) 외 다양한 감성 사전이 개발되었지만, 표준 감성 사전으로 사용하는 데 한계가 있다[31]. 따라서 각 분야에 적합한 감성 분석을 위해서는 새로운 감성 사전 구축 또는 알고리즘을 학습시켜야 되는 번거로움이 있다[53]. 최근까지도 한국어 감성 분석에서는 각 분야에 맞는 감성 사전 구축이나 알고리즘이 용이하지 않은 상황이다. 이와 같은 상황은 전문 분야에 맞는 한국어 감성 사전을 구축하거나 알고리즘 학습 절차가 복잡하기 때문이다. 기존에 사용되었던 방법은 영어 기반 알고리즘이 개발되었기 때문에 이 알고리즘에 한국어를 그대로 적용하기에는 언어 특징의 차이가 있다[8]. 한국어는 교착어로 굴절어인 영어보다 단어의 구별이 어렵고 변형이 많으며[9, 15, 37], 영어와 다르게 형태소 구별이 복잡해 별도의 형태소 분석기 사전을 구축하여 분석 방법에 주로 사용했다[18]. 하지만 현재 대부분의 형태소 분석기 사전 기반이라 할 수 있는

‘세종 말뭉치 한글 사전’은 업데이트가 되지 않아 신조어 및 전문 용어가 포함되어 있지 않으며[39], 띄어쓰기, 오탈자, 속어 등 많은 비정형 데이터의 경우에도 잘 분석되지 않는다는 문제점을 가지고 있다[38]. 이런 문제를 해결하기 위해 본 연구의 목적은 감성 분석의 정확도 향상과 단순 단어 출현 횟수에 기반한 분석이 아닌 맥락을 반영하는 감성 분석 모델을 설계 및 검증을 하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 모델의 알고리즘에서 활용된 회선 신경망, 문자 기반 학습, 그리고 단어 임베딩에 대해 서술하였다. 또한 감성 분석의 정의 및 활용, 그리고 현재 연구 현황들에 관해 서술한다. 제3장에서는 한국어의 특성과 기존 영어권 모델들이 한국어에 적용될 때 발생하는 문제점들에 관해 서술하고, 회선 신경망과 문자 기반 모델, 그리고 단어 임베딩을 통한 새로운 감성 분석 모델을 제안한다. 제4장에서는 제안된 모델을 실제 데이터를 이용하여 기존 연구 내 모델들과 정량적, 정성적으로 비교하고자 한다. 마지막으로 제5장에서는 본 논문의 결론을 서술하고 향후 연구 방향에 대하여 논의한다.

2. 개념적 배경

2.1 회선 신경망

회선 신경망(Convolutional Neural Network, CNN)이란, 전방 전달 신경망(Feedforward Neural Network)의 한 종류로 하나 또는 여러 개의 회선 레이어(Convolution layer)에 인공 신경망 레이어를 올리는 구조를 뜻한다. 일반적으로 회선 레이어는 인공 신경망의 전처리 역할을 수행하며, 각 회선 레이어는 다양한 종류의 필터를 생성하여 학습을 완료하게 된다.

초기 회선 신경망은 이미지 처리에 국한되어 사용되었지만, 최근에는 자연어 처리(Natural Language Processing; NLP) 분야까지 확장되고 있다. 자연어 처리에 사용된 회선 신경망은 문장의 잠재적인 의미론적 표현(Semantic Representation)을 형성하기 위한 n-gram을 추출하는 기능을 갖추고 있다[12, 20, 32]. 텍스트 관련 논문을 살펴보면 회선 신경망을 활용한 의미 분석 연구[59], 문장 모델링 관련 연구가 있으며[20], 현재는 자연어 처리까지 연구 영역으로 넓혀가고 있다[12]. 국내에서도 회선 신경망을 활용한 텍스트 마이닝 연구가 활발히 진행되고 있으며, Kim et al.[21]은 회선 신경망을 활용한 단어 기반의 감성 분석 연구를 하였고, Shin et al.[51]은 자소 기반의 단어 객체명 인식 연구, Choi[8]는 음절 기반의 단어 학습 연구를 확인하였다.

2.2 단어 분산 표현 기법

자연어 처리에서 자연어를 벡터 값으로 변형할 때 사용되는 대표적인 방법은 원-핫 표현(One-hot representation) 방법과 분산 표현(Distributed representation) 방법이 있다. 원-핫 표현 방법은 단어 간 유사성을 계산할 수 없는 단점이 따르지만, 분산 표현 방법은 저 차원에 단어 의미를 여러 차원에 분산하는 장점이 있다[46]. 분산 표현 방법은 같은 맥락에서 단어의 비슷한 의미(Semantic meaning)를 공유한다는 분산 가정(Distributional Hypothesis)을 기반으로 발전해 왔다. 분산 표현 방법에는 주변 단어들을 활용하여 중간에 있는 단어를 추정하는

방법인 CBOW(Continuous Bag of Words)와 중심 단어를 기반으로 주변 단어를 추정하는 Skip-gram 두 가지 모델이 존재한다.

자연어를 벡터로 만들기 위해 word2vec을 사용한 것은 다음과 같은 장점이 있기 때문이다. 첫째, 텍스트를 벡터로 변환하여 텍스트 간 수치적인 연산이 가능하다. 둘째, 비슷한 문장에 기술된 단어는 비슷한 벡터를 가져 오타가 있어도 비슷한 벡터 값을 보유하기 때문에 차원의 저주(Dimensionality Reduction) 문제를 해결할 수 있다. 하지만 word2vec을 활용한 단어 임베딩(word embedding) 시 주의해야 할 점도 있다. 첫째, 빈도수에 매우 편향적이기 때문에 단어의 빈도수가 적을 경우 단어의 위치 추정이 정확하게 되지 않는다[42]. 둘째, 주변 텍스트 기반에 단어를 학습하기 때문에 상반된 감성의 단어이라도 유사한 표현으로 추정된다[52]. 예시로, 하나의 문장에서 '좋음'과 '나쁨'의 단어가 같이 포함되어 있을 때 단어 임베딩 시 이 두 단어는 유사한 표현으로 추정하게 된다. 따라서, word2vec은 감성 분석에서 대조적인 단어를 구별하는 모델의 성능 저하를 시키는 문제에 대해 야기되었으며[56], 한국어와 같이 적은 빈도 단어가 자주 출현하는 언어에서는 학습을 어렵게 만든다는 연구가 있었다[8]. 기존 선행연구에서는 형태소 단위에서 벡터 값을 사용하였기 때문에 본 연구에서는 자모 단위의 입력 데이터를 통해 벡터 값을 사용하기 위해 분산 표현 방식을 사용하였다.

2.3 문자 기반 모델

텍스트 마이닝 모델에서 입력 데이터의 단위는 단어 레벨(word-level) 또는 형태소 레벨(morpheme-level)인 경우가 일반적이다. 하지만 교착어(한국어, 터키어, 헝가리어, 핀란드어)와 같이 형태소의 활용이 다양해 언어의 형태가 복잡한 언어를 형태소 단위의 모델에 적용하였을 때 형태소 분석기에서 문제가 발생하여 오류가 발생될 여지가 있다[4, 36]. 또한 맥락에서 단어의 출현 빈도가 적어 단어 학습 시 추정 자체가 제대로 되지 않는 문제점이 존재한다[42]. 이러한 문제를 보완하기 위해 형태소 하위 레벨인 문자 기반(Character-level) 언어 모델이 최근 주목을 받고 있다[5, 11, 33, 35, 40]. 문자 기반 언어 모델의 경우 특정 자연어 처리 분야에서 형태적으로 복잡한 언어를 개선한 연구 결과들이 있다. Santos and Guimaraes[48]는 단어 임베딩과 함께 문자 수준 임베딩을 개체명 인식 문제에 적용해 포르투갈어와 스페인어 말뭉치에서 높은 수준의 결과를 보였으며, Peng et al.[45]의 경우 한자어 문자의 하위 단위인 라디칼 단위(Radical level)의 처리가 감성 분류 성능을 크게 개선할 수 있음을 입증하였다. 선행 연구 결과에 따르면 언어의 형태가 복잡한 언어에 딥러닝 기법을 적용하는 경우 단어 단위보다 문자 단위에서 임베딩을 하는 것이 더 좋은 성능을 나타내었다[62].

2.4 감성 분석

오피니언 마이닝(Opinion mining)의 일종인 감성 분석(Sentiment Analysis)은 텍스트에서 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하기 위한 기술을 의미한다[41]. 감성 분석은 텍스트 마이닝 연구에서 많이 활용되는 알고리즘이며[30], 이와 관련된 연구는 <Table 1>과 같다.

〈Table 1〉 A Study on the Use of Sentimental Analysis

Content	Author
Explore the impact of the results through sentimental analysis in the review of the product	Sen and Lerman[49]
Content Sentimental Score Measurements with Reviewed Data	Kim et al.[29]
The positive and negative aspects of the sentimental scorecard of title and review have an adjustment effect on the help of product purchase	Salehan and Kim[47]
The help of reviews depends on the sensitivity of the reviews, regardless of product type	Chua and Banerjee [10]
An Empirical Analysis of the New York Times News to Predict Stock Market Trends	Hong et al.[17]
To improve the accuracy of stock price forecasts, the analysis of the sensitivity of SNS and news articles and the prediction model of stock price using machine learning are presented	Kim et al.[22]
Propose a stock price fluctuation forecast model by combining LSTM based time series forecasts	Kim et al.[23]
Analysis of the sensitivity of keywords related words to suggest public opinion and policy implications related to the resort	Park[44]
Suggest that the change in feeling should be reflected in the policy proposal through the analysis of food safety-related sensitivities	Song[55]
Analyze online user comments to establish an sentimental analysis dictionary related to fashion design	An and Lee[1]

지금까지 활용되었던 감성 분석 사전은 범용이 아닌 별도의 감성 분석 사전을 구축하여 사용되었으며, 감성 사전을 구축하는 방법은 다음과 같다. 첫째, 리뷰와 같은 텍스트 데이터를 단어 기반의 문서 단어 행렬(Term-Document Matrix)과 같은 벡터 표현 형식으로 변환한다. 둘째, 단어의 사용 횟수를 독립 변수, 리뷰에서의 제품 평점(별점)을 종속 변수로 설정 후 로지스틱 회귀(Logistic regression) 분석을 통해 각 단어의 긍정과 부정 정도를 측정한다. 이때, 영향력이 적은 단어를 제거하거나 축소하는 정규화 과정을 통해 유효한 단어만을 추려 감성 사전을 구축한다. 마지막으로, 완성된 감성 사전을 통해 텍스트 내 포함된 감성 단어들의 총 감성 점수를 계산하여 감성 분석을 완료한다.

감성 사전 기반의 감성 분석은 텍스트의 의미 변화나 맥락의 차이, 동음이의어 등 문제를 해결하기 위해 다양한 감성 사전 구축과 감성 분석 정확도를 향상하기 위한 시도가 진행되었다[24, 25, 60]. 하지만 중의성 문제는 해결되지 못하고 있으며[27], 범용 감성 사전을 통한 감성 분석의 수행보다는 주제에 따라 특화된 감성 사전 사용이 감성 분석의 정확성을 향상한다는 연구 결과가 있었다[54]. 분석 분야가 달라질 경우 해당 분야에 맞는 감성 사전 또는 알고리즘을 학습시켜 감성 분석을 활용하는 양상을 띄고 있다. 영어 텍스트 감성 분석에서도 표준 감성 사전인 'SentiWordNet'이 존재하지만 감성 분석 연구에 따라 이에 맞는 새로운 감성 사전을 구축하여 사용하고 있다[2, 16]. <Table 2>와 같이 영어 기반의 감성 분석의 경우 딥 러닝(Deep Learning)을 통한 감성 분석 연구들이

〈Table 2〉 A Study on the Sentimental Analysis of English and Korea

Language	Algorithm	Accuracy	Author
English	Skip-Connected LSTM	81.47%	Bae and Lee[3]
	GRU-Attention	80.41%	Choi and Lee[7]
	CharSCNN	85.7%	Dos Santos and Gatti[13]
	CNN-static	89.6%	Kim[32]
	n-grams TFIDF	92.4%	Zhang and LeCun[61]
Korean	LSSVM	79%	Yang and Lee[58]
	Word-level CNN	80%	Kim et al.[21]
	Trigram-Signature hashing based sentimental analysis	80.7% (F-measure)	Chung et al.[6]
	Pattern based sentimental analysis	82.4%	Seo et al.[50]
	Sentiment corpus + SVM	82.5%	Kim et al.[28]
	Character-CNN	84.4%	Shin et al.[51]
	Word2vec + KNN	89.6%	Kwon[34]

활발하게 이루어지고 있으며, 국내에서는 감성 분석 시 정확도를 높이기 위한 연구들이 진행되고 있음을 확인하였다. 그 결과 각 분야에 적절하게 사용할 수 있는 한국어 감성 사전이 부재하기 때문에 감성 분석 시 자체적으로 사용할 수 있는 감성 사전 구축 또는 알고리즘 학습이 매 연구마다 진행되고 있음을 확인하였다.

3. 알고리즘 설계 및 개발

3.1 문제 정의 및 접근 방법

다양한 분야에서 용이하게 사용할 수 있는 표준 한국어 감성 분석 사전 및 알고리즘이 전무한 것은 감성 분석 사전 및 알고리즘 구축에 몇 가지 문제가 있기 때문이다. 첫째, 대부분 한국어 텍스트 전처리에서는 형태소 분석기 사전을 이용한 형태소 분석을 진행하는 것이 문제이다. 교착어인 한국어는 영어와는 달리 단어의 정확한 구별이 어렵고 변형이 많다는 특징이 있다[15, 37]. 영어 단어는 띄어쓰기로 구별이 되고 어근의 변형이 매우 제한적이다. 반면에 한국어는 조사, 어미의 형태로 인해 형태소 분석기에 사용되는 개수만 하더라도 수 십여 가지가 넘는다[15]. 따라서 한국어에서는 일반적으로 효율적인 형태소 분석을 위한 기분석(Full word morpheme) 사전을 구축 후 이 사전을 통해 형태소를 분석하는 방법이 주로 제안되어 왔다[18].

정부는 1998년부터 2007까지 10년간 진행하였던 '21세기 세종계획'을 통해 '세종 말뭉치'가 구축되었지만[19], 범용으로 사용되기에 몇 가지 문제가 있다. 첫째, 형태소 분석기에서 메신저 용어와 신조어의 품사 구별이 되지 않고 있다. 범용적인 사전 목적으로 구축이 되었지만 전문 용어를 포함하고 있지 않으며, 어근을 찾기 힘든 비속어, 줄임말, 이모티콘 등 형태 분석이 어렵다는 한계점을 지니고 있다[38]. 둘째, 한국어 단어

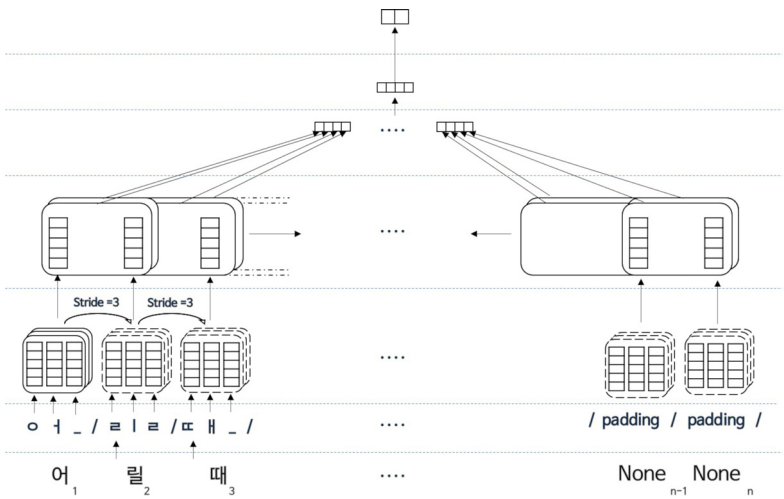
기반의 학습 문제이다. 현재 감성 분석을 비롯한 많은 텍스트 마이닝 연구에서 활용되고 있는 word2vec과 같은 단어의 분산 표현 기법은 빈도수에 편향적이기 때문에 희귀 단어 학습에는 어려움이 있다고 밝혀졌다[42]. 하지만 한국어는 굴절어로 저 빈도의 단어가 말뭉치에 더 자주 등장하는 특징을 보이는 언어이다[4, 42].

한국어가 저 빈도의 단어가 말뭉치에 더 자주 등장하는지 확인하기 위해 자모 단위의 임베딩을 위하여 사용된 181,852개의 기사 데이터를 트위터(Twitter) 형태소 분석기로 분석 후 각 단어별 빈도를 측정하였다. 그 결과, 추출된 225,816개의 고유 단어 중 빈도수가 10 이하인 단어는 147,413개, 5 이하 빈도의 단어는 126,586개로 확인되었다. 이와 같이 한국어는 저 빈도의 단어가 자주 출현하는 특징을 가지고 있기 때문에 일반적인 word2vec과 같은 단어 분산 표현의 학습에서 한계점을 가지고 있음을 알 수 있다[8].

따라서 본 논문에서는 한국어 감성 분석 전처리 시, 형태소 분석 단계에서 업데이트 미비로 인한 신조어 및 전문용어의 미반영 문제가 발생할 수 있음을 확인할 수 있었다. 한국어 단어 단위의 임베딩 단계에서는 단어의 대부분을 차지하는 저 빈도 단어에서 언더피팅(underfitting)의 문제가 발생할 소지가 있기 때문에 영어와 같은 학습 정도를 위하여 약 4배 더 많은 말뭉치가 필요하다는 문제점을 알 수 있었다[8]. 이를 해결하기 위해, 첫째, 단어 단위가 아닌 자모 단위로 전처리를 수행하였다. 이를 통해, 단어와는 달리 모든 문자가 충분한 빈도를 가질 수 있도록 하였다. 둘째, 감성 분석 내 형태소 분석을 배제할 수 있는 CNN 모델을 채택함으로써 미등록 단어(out of vocabulary) 문제인 신조어, 전문 용어 및 비속어의 형태소 분석이 제대로 이루어지지 않는 문제를 해결할 수 있도록 하였다.

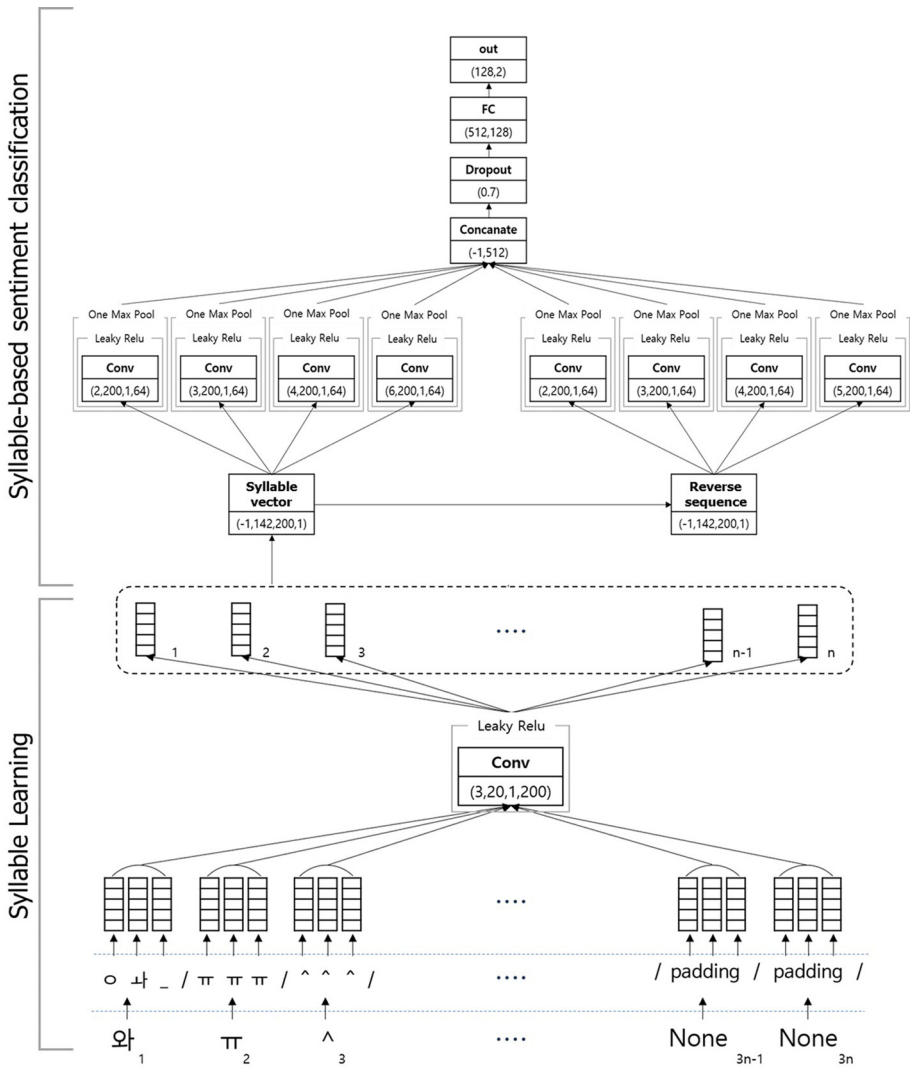
3.2 모델 제안 및 동작 과정

본 연구에서 제안하는 모델은 <Figure 1>과 같다. 회선 신경망에 자모 데이터를 기반으로



〈Figure 1〉 Proposed Model

음절 벡터를 생성하는 과정과 생성된 음절 벡터를 기반으로 감성을 분류하는 두 부분으로 구분하였다(<Figure 2> 참조). 먼저 자모 데이터를 기반으로 음절 벡터를 생성하기 위해 word2vec을 통하여 각 문자의 학습 후, 각 입력 데이터를 한국어 고유 특징인 초성(ㄱ, ㄲ, ㄴ, ㄷ, ㄸ, ㄹ, ㅁ, ㅂ, ㅃ, ㅅ, ㅆ, ㅇ, ㅈ, ㅉ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ), 중성(ㅏ, ㅐ, ㅑ, ㅒ, ㅓ, ㅔ, ㅕ, ㅖ, ㅗ, ㅘ, ㅙ, ㅚ, ㅛ, ㅜ, ㅝ, ㅞ, ㅟ, ㅠ, ㅡ, ㅢ, ㅣ), 종성(_, ㄱ, ㄲ, ㄳ, ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅄ, ㅅ, ㅆ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ)에 맞추어 배열하도록 하였다. 이때, 종성이 없을 경우 '_'로 처리하여 한 음절은 반듯이 3개의 문자로 구성이 되도록 배열하였다.



〈Figure 2〉 Overall Model Process

초성, 중성, 종성의 벡터 값은 하나의 회선 레이어를 거쳐 음절 벡터로 재 생성된 후 음절 벡터를 기반으로 본격적인 n-gram을 학습하여 단어가 아닌 음절 조합 방법을 통해 달라지는 의미에 대해 구분할 수 있도록 하였으며, 자모가 단독으로 쓰이거나 한국어가 아닌 경우에는 각 문자를 3번 반복하여 배열하였다. 입력 데이터 중 가장 긴 데이터가 총 S개의 한국어 음절과 k개의 단일 자소 및 비 한국어 문자로 이루어져 있고, 학습된 문자의 차원이 d라고 할 때 입력 데이터의 행렬은 $Q \in R^{d \times 3|s+k|}$가 되도록 구성하였다. 이후 각 입력 데이터의 길이를 맞추기 위하여 각 데이터의 모자란 부분은 0 삽입(Zero Padding)을 적용 후 데이터의 길이를 통일하여 입력 데이터의 전처리를 하였다. 이후 입력으로 들어온 데이터 S^t에 초성, 중성, 종성을 포함할 수 있는 폭이 3, 필터의 수 W인 컨볼루션 필터 $H \in R^{d \times 3 \times w}$에 건너뛰는 픽셀의 개수를 의미하는 폭을 [1, 3, 1, 1]에 맞추어 적용하여 필터의 수 W가 한 음절을 의미하는 피처 맵 $f^t \in R^{|s+k| \times w}$을 생성하였다. 피처 맵은 활성 함수 Leaky Relu를 거쳐 알고리즘 내에서 최종적으로 음절 벡터로 사용되었다(<Figure 2>의 Syllable Learning 참조).

<Figure 2>의 Syllable-based sentiment classification은 Syllable Learning에서 생성된 음절 벡터를 기반으로 감성 분류를 하였다. 각 음절 벡터에서 S개의 음절과 음절 벡터의 크기가 d라 할 때 $Q \in R^{d \times s}$가 음절 벡터의 행렬이 된다. 음절 벡터의 1차원을 기준으로 0 삽입(Zero Padding) 부분을 제외한 벡터를 역으로 배열한 역 순차(reverse sequence) 레이어를 생성한다. 그다음, 두 음절 벡터에 폭이 W인 회선 필터 $H \in R^{d \times w}$를 적용하여 피처 맵 $f^t \in R^{s-w+1}$을 생성한다. 본 연구에서는 정 방향 음절 벡터와 역 방향 음절 벡터에서 각 길이가 2, 3, 4, 6개의 필터와 2, 3, 4, 5개의 필터에 각각 64개의 레이어를 생성하여 적용하였다. 각 회선은 활성 함수 reaky relu과 회선 레이어에서 가장 큰 벡터 값만을 추출하는 최대 풀링(max pooling)을 진행하여 추출된 모든 벡터를 연결(concatenation)하였다. 이후, 네트워크 일부를 생략하는 드롭아웃(drop-out)과 fully connected layer를 통해 각 텍스트 데이터의 감성 분류를 할 수 있도록 하였다.

4. 성능 평가

4.1 비교 모델 선정 및 학습 방법

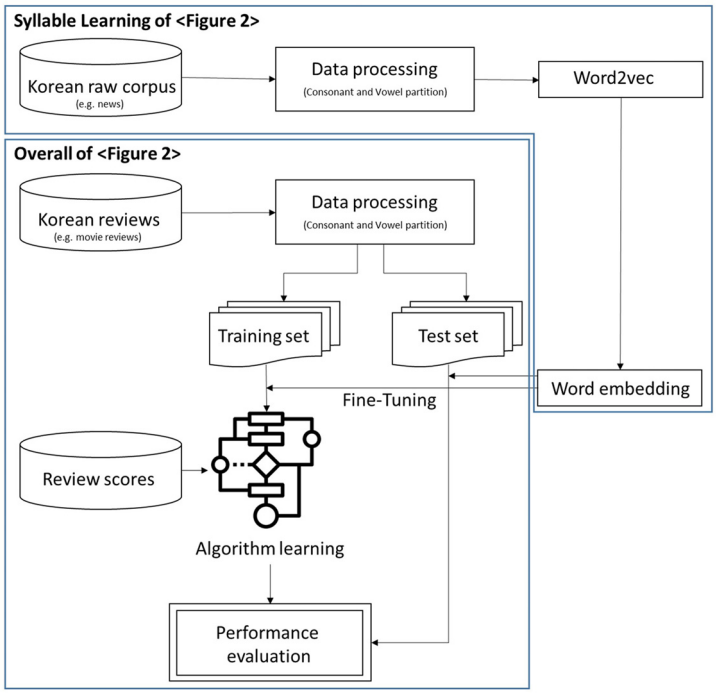
본 연구에서 제안한 모델 비교 대상으로 선정된 모델은 감성 분석에서 기본으로 사용되고 있는 Naive Bayes Classification, Logistic Regression과 영어 감성 분석에서 89.9라는 높은 성능을 기록하였던 Kim[32]의 CNN-static 모델이다. 감성 분석에서 기본으로 사용된 모델은 Park[43]이 제시한 절차에 따라 학습하였으며, Kim[32]의 경우 연구자가 공개한 모델에 Twitter 형태소 분석기와 Facebook에서 공개한 FastText 한국어 벡터를 사용하여 학습을 진행하였다(<Table 3> 참조). 정확도 평가를 위한 모델 간의 비교를 위한 절차는 <Figure 3>과 같다.

4.2 테스트 환경과 데이터 수집

본 연구 모델 개발을 위한 환경은 CPU(Intel core i7 8700k), RAM(64GB), Graphic Card

〈Table 3〉 Comparison of Learning Process among Referenced Researches

Logistic Regression Learning Process [43]				
Morphological analysis (Twitter)	Word2vec	Applying weight to TF-IDF	Model Training	Performance Test
Naive Bayes Classification Learning Process [43]				
Morphological analysis (Twitter)	Word2vec	Model Training	Performance Test	
CNN-static Learning Process [32]				
Morphological analysis (Twitter)	Word Embedding (Fast Text)	Model Training	Performance Test	



〈Figure 3〉 Overall Model Flow

(GeForce 1080ti), OS(Window 10 Pro), Program language(Python 3.6)이다. 모델 개발 및 데이터 전처리에 사용된 Python 라이브러리는 hgtk (0.1.1), numpy(1.13.3), pandas(0.20.3)이고, word2vec 및 모델 개발은 tensorflow(1.14), 모델 평가는 scikit-learn(0.19.1)을 사용하였다.

본 연구에서 개발한 모델과 감성 분석에서 보편적으로 사용되고 있는 모델(Naive Bayes Classification, Logistic Regression) 그리고 Kim[32] 모델의 성능을 비교하였다. 기본 모델은 sklearn(0.19.1)에 포함되어있는 모델을 사용하였으며, Kim[32]의 CNN-static 모델은 저

자가 GitHub에 공개한 모델을 사용하였다.

문자를 임베딩 하기 위한 한국어 말뭉치와 개발된 모델 학습을 위해 텍스트 마이닝을 수행하였다. 문자 임베딩을 위한 텍스트 마이닝은 다음과 네이버 뉴스(정치, 경제, 사회)에서 13만 건의 데이터를 수집하였고, 모델 학습을 위한 텍스트 마이닝은 네이버 영화 리뷰와 그 리뷰에 해당하는 별점을 3주간 수집하였다(<Table 4> 참조). 142자 이하로 작성된 영화 리뷰로 영화 당 100개씩 랜덤 하게 수집하였으며, 수집된 각 영화 댓글의 별점 1-4점은 부정, 9-10은 긍정으로 분류하였다. 긍정과 부정에 속하는 데이터를 각각 10만 개씩 샘플링하여 모델 학습에 사용할 데이터 셋을 최종적으로 완성하였다.

〈Table 4〉 Descriptive Analysis for Movie Reviews by Naver

Review Type	Count
Positive Reviews	100,000
Negative Reviews	100,000

4.3 문자 기반 임베딩

문자 임베딩을 위해 수집된 13만 건의 뉴스 데이터를 기반으로 한국어 자모 단위의 문자 임베딩을 수행하였다. 첫째, 문자 단위의 임베딩을 위한 데이터 전처리를 수행하여 URL 및 광고성 문구는 제거하였다. 이후 추출된 텍스트 데이터를 가지고 자모로 치환하였다. 텍스터 데이터 중 한국어 외 문자는 그대로 사용하였으며, 출현 빈도가 현저하게 낮은 한자는 모두 제거하였다. 최종적으로 한글 자모, 영어 대소문자, 일어, 특수문자 등 총 622개의 고유 문자를 가지고 있는 말뭉치를 생성하여 전처리 과정을 마쳤다. 둘째, 문자 단위 임베딩 학습 과정이다. 문자 벡터의 학습은 전처리된 말뭉치를 기반으로 word2vec의 skip-gram을 사용하여 진행하였으며, 0.001의 학습률로 epoch 500,000만 번을 수행하였다. 최종적으로 3.8의 average loss를 기록한 20차원의 문자 벡터를 추출하였다. 학습에서 사용된 모델 파라미터는 embeding size = 20, skip window = 16, skip number = 1, valid size = 16, valid window = 100, negative sample number = 256로 설정하였다.

4.4 모델 비교 및 성능 평가

본 연구에서 정량적 평가를 위한 정확도 비교는 K-겹 교차 검증(K-fold cross validation)을 수행하였다. K-겹 교차 검증은 전체의 데이터 셋을 K개로 나눈 뒤 하나의 데이터 셋은 평가를 위해 사용되고 나머지 데이터 셋은 모델 학습을 위한 데이터로 사용하는 방법이다. 본 연구에서는 전체 데이터를 10개의 데이터 셋으로 나누어 1개의 데이터의 셋은 평가를 위해 사용되었고 9개 데이터 셋은 모델 학습을 위한 데이터로 사용하였다. 학습의 총 횟수는 Wu[57]가 제안한 epoch = 3,000으로 설정하여 학습을 수행하였으며, 학습을 통해 생성된 10개 모델 성능이 일관성을 갖는지 확인한 결과 모두 동일한 성능을 가진 모델임을 확인하였다. 그 다음 본 연구에서 개발된 모델과 <Table 3>에서 언급한 모델과의 성능을 비교 하였다.

모델 성능 평가는 <Table 5>의 기준을 따랐으며, 예측 값이 실제 값과 같은 True Positive와 True Negative의 합에서 전체 데이터를 나눈 값을 정확도(Accuracy)로 하여 백분율로 나타

냄으로써 모델 별 성능 비교를 하였으며, 정확도를 구하는 공식은 TP+TN / TP+TN+FP+FN 이다. 4가지 모델에 대한 감성 분석 실험 결과는 <Table 6>과 같다. 본 연구에서 제안된 모델의 정확도는 88.0%로 매우 준수함을 알 수 있었다. 텍스트 분류에 보편적으로 사용되는 나이브 베이지안이나 로지스틱 회귀분석에서는 각각 80.4%, 78.7%로 양호한 정확도를 확인할 수 있었다. 반면에 Kim[32]이 제안한 모델에 한글 자모 기반의 데이터를 적용하였을 경우 정확도가 78.1%로 나타났다. 영어 기반으로 연구하여 나타난 88.9%의 정확도와는 다르게 한글에서는 10.2% 낮은 정확도가 확인된 것으로 언어에 따라 알고리즘 성능이 다르다는 것을 알 수 있었다.

〈Table 5〉 Evaluation Criteria (Confusion Metrix)

		Predicted Class	
		Positive	Negative
Actual Class	True	True Positive	True Negative
	False	False Positive	False Negative

〈Table 6〉 Model Performance Evaluation

Model	Accuracy
Logistic Regression	77.1%
CNN-static	78.1%
Naive Bayes classifier	80.4%
Proposed Model	88.0%

4.5 오타 증가에 따른 데이터 자질 별 분류 정확도 비교

본 연구에서 제안된 모델에 입력 데이터(영화 리뷰)에서 자모, 음절, 형태소에 따른 텍스트 내 비 정형성 증가에 정확도 감소율의 변화를 비교하기 위해 5글자(20%), 10글자(10%), 20글자(5%), 50글자(2%) 마다 한 번씩 자모를 랜덤하게 교체하여 의도적으로 오탈자를 생성 후 정확도의 감소율을 비교해 보았다. 입력 데이터에 따른 모델을 비교하기 위해 문자 임베딩을 자모, 음절, 형태소 3가지 유형으로 임베딩 학습을 하였다. 자소의 경우는 4.3절에서 만들어진 임베딩 모델을 사용하였으며, 음절의 경우 자모(초성, 중성, 종성)를 더해 60차원의 문자 벡터를 추출하였다. 형태소의 경우 100차원의 문자 벡터를 추출 한 후 문자 임베딩 학습을 word2vec 기반으로 하였다.

비교 결과 자모의 경우 음절, 형태소에 비해 적은 심볼 수를 가짐에도 불구하고 가장 우수한 정확도와 비 정형성에 따른 정확도 감소율이 나타났음을 확인 할 수 있었다. 또한 자모 기반의 모델과 음절 기반의 모델이 모두 비슷한 정확도를 기록하였지만 데이터의 비 정형성이 높아질수록 정확하지 않은 글자가 증가하면서 음절, 어절, 형태소의 분류 정확도가 모두 현저하게 낮아지는 것을 알 수 있다. 특히 형태소 기반 단어 입력의 경우, 오타 발생 시 원본 데이터에 비해 각각 1.1%, 2.7%, 5.4%, 12.2%의 분류 정확도 감소율을 보임으로써 자모나 음절 기반의 모델에 비해 비 정형성에 대하여 매우 취약함을 확인할 수 있었다.

4.6 정성적 모델 비교 및 성능 평가

본 연구의 정성적 분석은 선행 연구 한국어 감성 분석 모델들의 정확도와 비교를 진행하였으며, 두 번째로 리뷰의 성격에 따른 샘플을 제시 후 실험에서 제시하였던 모델들의 분석 결과를

〈Table 7〉 Reduction in Accuracy as Unstructured Text Increase

Input data	Movie review	20%	10%	5%	2%	Average reduction ratio
Letter	88.0%	83.4%	86.2%	87.0%	87.5%	1.98%
Syllable	87.1%	79.9%	83.5%	85.3%	86.3	3.38%
Morpheme	82.0%	69.8%	76.6%	79.3%	80.9	6.10%

비교해 보았다. 첫째, 기존 논문에서의 감성 분석 모델과 비교하였다. 일반적으로 자연어 처리에 많이 사용된 딥러닝 기법 모델들과 비교하였다. 일반적으로 텍스트 데이터에 사용되는 LSTM(Long Short-Term Memory)을 적용한 한국어 감성 분석의 경우 76.68%의 다소 낮은 정확도를 기록하였으며[34], Shin et al.[51]은 CNN을 활용한 한국어 감성 분석의 경우, 형태소, 음절, 자모로 각각 학습 시 84.6%, 84.4%, 83.6%의 준수한 정확도를 기록하였으나, 비 정형성에 따른 오타 감소율이 정확도와는 반대로 11%, 8%, 3%로 나타났음을 알 수 있다. 하지만 본 연구에서 제안하는 모델의 경우 정확도와 비 정형성에 따른 정확도 감소율이 모두 자모 단위에서 가장 우수한 것으로 나타났기 때문에 기존 모델과 큰 차이가 있는 것을 확인할 수 있었다.

다음으로 딥러닝 기반이 아닌 감성 분석 모델과의 비교를 하였다. 89.6%의 정확도를 기록한 word2vec과 KNN을 활용한 모델[34], 82.5%의 정확도를 기록한 현재 인터넷에 공개되어 있는 SVM과 한국어 감정 코퍼스 KOSAC을 활용한 모델[28], 80.7%의 정확도를 기록한 자모 단위의 전처리 후 Trigram-Signature를 해싱하여 감성 분석을 진행한 Chang et al.[6]의 모델과 비교해 보았을 때 해당 모델의 정확도가 전반적으로 우수한 편임을 알 수 있었다.

〈Table 8〉 Sensitivity Classification of Models by Review Category

Category	Review	Proposed Model	Logistic	Naive Bayes	CNN-static
Review	To be not funny(재미없다)	Negative	Negative	Negative	Negative
	To be funny(재미있다)	Positive	Positive	Positive	Positive
Contextual Review	It would be fun, but it's not fun (재미있을 줄 알았는데 재미없어요)	Negative	Negative	Negative	Negative
	It would not be fun, but it's fun (재미없을 줄 알았는데 재미있어요)	Positive	Negative	Negative	Positive
	Do you like this movie? I don't like it (이 영화 좋아? 난 싫어)	Negative	Negative	Negative	Negative
	You don't like this movie? I like it (너 이 영화 싫어? 난 좋아)	Positive	Negative	Negative	Positive
Review with new word	Very boring(핵노잼)	Negative	Negative	Negative	Negative
	Very interesting(핵잼)	Positive	Negative	Positive	Positive
	Annoyance(짱난다)	Negative	Negative	Positive	Positive
	Awesome(짱이다)	Positive	Positive	Positive	Positive

리뷰 샘플 별 모델 비교 결과는 <Table 8>과 같다. ‘일반적인 리뷰’, ‘맥락을 이해가 필요한 리뷰’, ‘신조어가 포함된 리뷰’ 총 세 가지 범주별 리뷰 예시를 생성 후, 학습한 모델 별로 분석 결과를 비교해 보았다. 맥락에 대한 이해가 필요하거나 신조어가 포함된 리뷰의 경우에서는 Naive Bayes classifier나 Logistic Regression 모두 제대로 된 판단을 하지 못하는 것으로 드러났다. 이러한 결과는 명사의 출현 횟수만을 기준으로 판단하는 알고리즘의 특성 때문이라고 볼 수 있다. 하지만 반면에 CNN을 기반으로 한 Kim[32]의 모델과 본 연구에서 제시한 모델의 경우, 맥락의 이해가 필요한 리뷰의 분석을 제대로 수행하는 것을 알 수 있었다. 하지만 Kim[32]의 모델 경우 Twitter 형태소 분석에 기반된 모델이므로 형태소 분석기에 포함되지 않는 신조어에 대한 분류는 제대로 수행되지 않는 것으로 확인되었다.

5. 결 론

본 연구는 한국어 감성 분석 모델의 한계점을 개선하기 위해 자모 단위의 회선 신경망을 활용한 한국어 감성 분석 모델을 제시 및 성능을 검증하였다. 기존 한국어 감성 분석에서 사용되는 형태소 분석을 배제하고, 문자를 자모 단위 기반으로 입력 데이터의 크기를 축소하였으며, 저 빈도 형태소에 따른 학습 문제를 방지할 수 있었다. 또한 기존 형태소 분석의 가장 큰 한계점이었던 미등록 단어 문제와 비 정형성에 따른 정확도 감소 문제를 개선하였다. 그 결과 기존 한국어 감성 분석 연구뿐만 아닌 다른 언어에서 사용되는 감성 분석 연구들과 비교해도 매우 우수한 정확도를 보이고 있음을 알 수 있었다. 또한 한국어의 특성인 초성, 중성, 종성을 고려한 알고리즘 설계로 이를 통해 음절 벡터를 추출할 수 있었다. 본 연구에서 설계된 모델은 기존 연구에서 소개된 모델과 비교하였을 경우 자소 단위에서 평균 3.4% 높은 정확도를 확인하였으며, 비 정형성에 따른 감소율에서 자모는 평균 1.98%, 음절은 평균 3.38% 그리고 형태소에서는 평균 6.10%로 자모 단위에서 성능이 더 좋다는 결과를 보이는 모델로 검증하였다.

본 연구의 결과는 학문적, 실무적 시사점을 가지고 있다. 학문적 시사점으로, 초성, 중성, 종성을 기반으로 한 한국어 감성 분석 모델을 설계하였다. 일반적으로 영어에서 사용하는 모델의 경우 알파벳 기반으로 단어의 벡터를 추출한 후 언어 모델(Language Model)이나 분류 모델(Classification Model)을 구축하였다. 하지만 이런 방법은 정확한 단어의 구별 및 추출이 복잡한 한국어에서는 문제가 발생하였다. 따라서 음절 그 자체로 의미를 갖는 한국어의 특성을 반영하여 자모 기반으로 음절 벡터를 추출 후 분류 모델을 통해 단어 구분 및 학습단계에서 발생될 수 있는 문제점을 개선하였다. 학술적으로 한국어 문자 단위(Character-level) 모델에서 단어 벡터의 추출 대신 음절 벡터의 추출로 전환하여 감성 분석을 하였을 시 더 좋은 성능을 보인다는 것을 확인하였다.

본 연구의 결과를 토대로 본 실무적 시사점은 다음과 같다. 첫째, 감성 분석에서 형태소 분석 및 단어 임베딩과 같이 문제가 되는 단계를 보완 또는 배제된 알고리즘으로 절차가 간소화되어 비전공자들도 쉽게 감성 분석을 할 수 있다. 감성 분석은 다양한 도메인에서 많이

사용하고 있지만, 공개된 한국어 감성 분석 알고리즘 및 사전으로 각 도메인에서 사용하기에는 어려움이 있었다. 또한 전문 분야에서 감성 분석 시 언어 중의성 문제도 발생했기 때문에, 사용되고 있는 감성 분석 모델 및 사전을 다른 전문 분야에서 사용하려면 그 분야에 맞게 재학습을 해야 하는 상황이었다. 따라서 본 연구에서 개발된 모델은 전문 분야에서 감성 분석에 필요한 신조어 및 전문 용어 업데이트를 위해 학습 시 비 전문가도 간단하게 사용할 수 있도록 알고리즘을 설계하였다. 둘째, 본 연구의 모델은 텍스트 마이닝에서 사용되는 리뷰 및 댓글과 같은 텍스트 데이터에 적합한 알고리즘이다. 본 모델은 자모 단위의 입력으로 형태소 분석을 배제하였기 때문에 기존 비정형 텍스트에서 흔히 나타나는 속어, 줄임 말, 오탈자 같은 데이터의 영향을 적게 받는다. 이런 결과는 감성 분석 시 데이터의 성향에 관계없이 일괄적인 성능을 보이고 있다는 것을 의미한다. 마지막으로 본 연구의 모델은 텍스트 데이터의 맥락을 반영한 결과를 도출할 수 있다. 단어의 출현 횟수에 기반하여 분석하는 로지스틱 모델이나 베이지안 모델과는 달리 데이터의 맥락을 반영할 수 있는 회선 레이어를 기반으로 개발되어 정확하고 다양한 텍스트 문맥을 이해할 수 있음을 검증하였다. 따라서 본 연구의 모델을 통하여 더 넓은 범주의 한국어를 정확하게 분석할 수 있을 것이다.

본 연구의 한계점과 향후 연구 방향은 다음과 같다.

본 연구에서 개발된 모델은 텍스트 마이닝에서 주로 다루는 짧은 문장(리뷰, 댓글)에서는 좋은 성능을 보이고 있지만 장문의 텍스트에서는 아직 미흡한 부분이 따른다. 따라서 장문의 텍스트에서도 좋은 성능을 보일 수 있도록 알고리즘을 개선해야 한다. 둘째, 본 모델은 텍스트 맥락에서 감성 분석을 하는데 국한되어 있다. 감성 분석에는 텍스트만 아닌 이모티콘에서도 감성을 파악할 수 있다. 따라서 텍스트에서만 얻는 감성 분석이 아닌 이모티콘까지 포함한 감성 분석 모델로 개선되어야 한다. 마지막으로 본 연구에서는 알고리즘을 개발과 성능을 평가하는데 그쳤다. 본 알고리즘을 기반으로 이용자들이 편리하게 사용할 수 있도록 그래픽 사용자 인터페이스(Graphical User Interface; GUI)를 적용해야 한다.

References

[1] An, H. and I. Lee, “Extraction of Fashion Sensibility Vocabulary for Globalization,” Journal of Basic Design & Art, Vol. 14, No. 3, pp. 135-141, 2013.

[2] Baccianella, S., Esuli, A., and Sebastiani, F., “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” Lrec, Vol. 10, pp. 2200-2204, 2010.

[3] Bae, J. and Lee, C., “Sentiment Analysis with Skip-Connected LSTM,” Korea Information Science Society, Vol. 2017, No. 6, pp. 633-635, 2017.

[4] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T., “Enriching word vectors with subword information,” Transactions of the Association for Computational Ling-

uistics, Vol. 5, pp. 135-146, 2017.

[5] Bojanowski, P., Joulin, A., and Mikolov, T., "Alternative structures for character-level RNNs," arXiv preprint arXiv: 1511.06303, 2015.

[6] Chang, D., Kim, D., and Choi, Y., "Opinion Mining Based on Korean Phoneme Trigram-Signature," The Korean Institute of Information Scientists and Engineers, Vol. 2015, No. 6, pp. 811-813, 2015.

[7] Choi, K. and Lee, C., "Sentiment analysis with GRU-Attention," Korea Information Science Society, Vol. 2015, No. 12, pp. 557-559, 2015.

[8] Choi, S., "The modeling and training methods for syllable-based Korean word embeddings," Seoul National University, Master Dissertation, 2017.

[9] Choi, S., Lee, J., and Kwon, O., "A Morphological Analysis Method of Predicting Place-Event Performance by Online News Titles," The Journal of Society for e-Business Studies, Vol. 21, No. 1, pp. 15-32, 2016.

[10] Chua, A. Y. and Banerjee, S., "Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality," Computers in Human Behavior, Vol. 54, pp. 547-554, 2016.

[11] Chung, J., Cho, K., and Bengio, Y., "A character-level decoder without explicit segmentation for neural machine translation," arXiv preprint arXiv:1603.06147, 2016.

[12] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P., "Natural language processing (almost) from scratch," Journal of machine learning research, Vol. 12, pp. 2493-2537, 2011.

[13] Dos Santos, C. and Gatti, M., "Deep convolutional neural networks for sentiment analysis of short texts," Proceedings of Coling 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 69-78, 2014.

[14] Durfee, A., "Text Mining Promise and Reality," AMCIS 2006 Proceedings, p. 187, 2006.

[15] Eun, Z. and Park, S., "A Classification of Endings for an Efficient Morphological Analysis of Korean," Journal of KIISE, Vol. 2000, No. 10, pp. 41-47, 2000.

[16] Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., and Poncelet, P., "Web Opinion Mining: How to extract opinions from blogs?," in Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, pp. 211-217, 2008.

[17] Hong, T., Kim, E., and Cha, E., "The Prediction of dow jones and S&P500 index using SVM and news sentiment analysis," The Journal of Internet Electronic Commerce Research, Vol. 17, No. 1, pp. 23-36, 2017.

[18] Hwang, H. and Lee, C., "Error correction in Korean morpheme recovery using deep learning," Journal of KIISE, Vol. 42, No. 11, pp. 1452-1458, 2015.

[19] Hwang, Y. and Choi, J., “The 21st century Sejong corpus properly-Using the Language Information Sharing Center,” National Institute of Korean Language, Vol. 26, No. 2, pp. 73-86, 2016.

[20] Kalchbrenner, N., Grefenstette, E., and Blunsom, P., “A convolutional neural network for modelling sentences,” arXiv preprint arXiv:1404.2188, 2014.

[21] Kim, D., Kim, K., and Kim, J., “Character-based multi-category sentiment analysis on social media using deep learning algorithms,” Korean Institute of Industrial Engineers, pp. 5082-5084, 2017.

[22] Kim, D., Park, J., and Choi, J., “A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles,” Journal of Information Technology Services, Vol. 13, pp. 211-233, 2014.

[23] Kim, G., Ock, K., and Lee, S., “Sentiment Dictionary Construction for Stock Fluctuation Prediction based on Security Company Reports,” The Korean Institute of Information Scientists and Engineers, pp. 1022-1024, 2016.

[24] Kim, H. D. and Zhai, C., “Generating comparative summaries of contradictory opinions in text,” in Proceedings of the 18th ACM conference on Information and Knowledge Management, pp. 385-394, 2009.

[25] Kim, J. O., Lee, S. S., and Yong, H. S., “Automatic Classification Scheme of Opinions Written in Korean,” Journal of KIISE: Databases, Vol. 38, No. 6, pp. 423-428, 2011.

[26] Kim, J. and Kim, D., “A Study on the Method for Extracting the Purpose-Specific Customized Information from Online Product Reviews based on Text Mining,” The Journal of Society for e-Business Studies, Vol. 21, No. 2, pp. 151-161, 2016.

[27] Kim, J., Oh, Y., and Chae, S., “The Construction of a Domain-Specific Sentiment Dictionary Using Graph-based Semi-supervised Learning Method,” Science of Emotion & Sensibility, Vol. 18, No. 1, pp. 103-110, 2015.

[28] Kim, M., Jang, H., Jo, Y., and Shin, H., “Korean Sentiment Analysis Corpus,” Korea Information Science Society, Vol. 2013, No. 6, pp. 650-652, 2013.

[29] Kim, M., Song, E., and Kim, Y., “A Design of Satisfaction Analysis System For Content Using Opinion Mining of Online Review Data,” Journal of Internet Computing and Services, Vol. 17, No. 3, pp. 107-113, 2016.

[30] Kim, S., Cho, H., and Kang, J., “The Status of Using Text Mining in Academic Research and Analysis Methods,” Journal of Information Technology and Architecture, Vol. 13, No. 2, pp. 317-329, 2016.

[31] Kim, S., Lee, Y. J., Shin, J., and Park, K. Y., “Text Mining for Economic Analysis,” BOK Economic Research Institute, Vol.

2019, No. 18, pp. 1-53, 2019.

[32] Kim, Y., “Convolutional neural networks for sentence classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1746-1751, 2014.

[33] Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M., “Character-aware neural language models,” Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[34] Kwon, S., “Sentiment Analysis of Movie Reviews using the Word2vec and RNN,” Graduate School of Dongguk University, Master dissertation, 2017.

[35] Lankinen, M., Heikinheimo, H., Takala, P., Raiko, T., and Karhunen, J., “A character-word compositional neural language model for finnish,” arXiv preprint arXiv:1612.03266, 2016.

[36] Lee, D. and Kim, K., “Web Site Keyword Selection Method by Considering Semantic Similarity Based on Word2Vec,” The Journal of Society for e-Business Studies, Vol. 23, No. 2, pp. 83-96, 2018.

[37] Lee, E., Kim, W., and Kim, S., “Korean Literature Dictionary,” Korean Dictionary Research History, 1998.

[38] Lee, J., Lee, H., and Lee, H., “Research on Methods for Processing Nonstandard Korean Words on Social Network Services,” Journal of the Korea Society Industrial Information System, Vol. 21, No. 3, pp. 35-46, 2016.

[39] Lee, S., AI seeds Hangul corpus in 2007 and landed reason. 2016; Available from: https://www.bloter.net/archives/260569.

[40] Ling, W., et al., “Finding function in form: Compositional character models for open vocabulary word representation,” arXiv preprint arXiv:1508.02096, 2015.

[41] Liu, B., “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, Vol. 5, No. 1, pp. 1-167, 2012.

[42] Mu, J., Bhat, S., and Viswanath, P., “All-but-the-top: Simple and effective post-processing for word representations,” arXiv preprint arXiv:1702.01417, 2017.

[43] Park, E., “Supervised Feature Representations for Document Classification,” Seoul National University, Doctoral dissertation, 2016.

[44] Park, S., “The media sentimental analysis of Yeongjong-do global MICE integrated resort,” International Journal of Tourism Management and Sciences, Vol. 31, No. 7, pp. 109-128, 2016.

[45] Peng, H., Cambria, E., and Zou, X., “Radical-based hierarchical embeddings for chinese sentiment analysis at sentence level,” The Thirtieth International Flairs Conference, 2017.

[46] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., “Learning representations by back-propagating errors,” Cognitive modeling, Vol. 5, No. 3, p. 1, 1988.

[47] Salehan, M. and Kim, D. J., “Predicting the performance of online consumer re-

views: A sentiment mining approach to big data analytics," Decision Support Systems, Vol. 81, pp. 30-40, 2016.

[48] Santos, C. N. D. and Guimaraes, V., "Boosting named entity recognition with neural character embeddings," arXiv preprint arXiv:1505.05008, 2015.

[49] Sen, S. and Lerman, D., "Why are you telling me this? An examination into negative consumer reviews on the web," Journal of interactive marketing, Vol. 21, No. 4, pp. 76-94, 2007.

[50] Seo, J., Jo, H., and Choi, J., "Design for Opinion Dictionary of Emotion Applying Rules for Antonym of the Korean Grammar," Journal of Korean Institute of Information Technology, Vol. 13, No. 2, pp. 109-117, 2015.

[51] Shin, H., Seo, M., and Byeon, H., "Korean Alphabet level Convolution Neural Network for Text Classification," Korea Information Science Society, Vol. 2017, No. 6, pp. 587-589, 2017.

[52] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D., "Semi-supervised recursive autoencoders for predicting sentiment distributions," Proceedings of the conference on empirical methods in natural language processing, pp. 151-161, 2011.

[53] Song, J. and Lee, S., "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews," Journal of KIISE: Software and Applications, Vol. 38, No. 3, pp. 157-168, 2011.

[54] Song, S. I., Lee, D. J., and Lee, S. G., "Identifying Sentiment Polarity of Korean Vocabulary Using PMI," Proceedings of the Korean Information Science Society Conference, pp. 260-265, 2010.

[55] Song, T., "Sentimatal Analysis on Food Safety Using Social Big Data," Korea Institute for Health and Social Affairs, Vol. 312, pp. 1-4, 2016.

[56] Wang, X., Liu, Y., Chengjie, S., Wang, B., and Wang, X., "Predicting polarities of tweets by composing word embeddings with long short-term memory," Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1, pp. 1343-1353, 2015.

[57] Wu, X., "A density adjustment based particle swarm optimization learning algorithm for neural network design," 2011 International Conference on Electrical and Control Engineering, pp. 2829-2832, 2011.

[58] Yang, S. and Lee, C., "Sentiment Analysis using Latent Structural SVM," Korea Information Science Society, Vol. 2015, No. 6, pp. 687-689, 2015.

[59] Yih, W. T., He, X., and Meek, C., "Semantic parsing for single-relation question answering," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 2, pp. 643-648, 2014.

[60] Yune, H. J., Kim, H. J., and Chang, J. Y., “An efficient search method of product reviews using opinion mining techniques,” Journal of KIISE: Computing Practices and Letters, Vol. 16, No. 2, pp. 222-226, 2010.

[61] Zhang, X. and LeCun, Y., “Text understanding from scratch,” arXiv preprint arXiv:1502.01710, 2015.

[62] Zheng, X., Chen, H., and Xu, T., “Deep learning for Chinese word segmentation and POS tagging,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 647-657, 2013.

저 자 소 개

성원경	(E-mail: wonkyung.sung@lge.com)
2016년	가톨릭대학교 문화콘텐츠학과 (학사)
2018년	연세대학교 정보대학원 비즈니스빅데이터분석 (석사)
2018년~현재	LG전자 DTX 센터 연구원
관심분야	Text Mining, Prediction Algorithm Development

안재영	(E-mail: jaeyoungan@yonsei.ac.kr)
2015년	대전대학교 컴퓨터공학과 (학사)
2017년	연세대학교 정보대학원 정보미디어전략 (석사)
2018년~현재	연세대학교 정보대학원 비즈니스빅데이터분석 박사과정
관심분야	Digital Business Transformation, Data mining, IT utilization in stone industry

이중정	(E-mail: cclee@yonsei.ac.kr)
1993년	University of South Carolina MIS (박사)
2000년	Salisbury State University 부교수
2000년~현재	연세대학교 정보대학원 교수
관심분야	IT Performance, IT Evaluation Measurement