Journal of KIIT. Vol. 20, No. 2, pp. 21-28, Feb. 28, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.2.21

텍스트 마이닝을 활용한 SMU 감성 사전 구현 및 신조어 감성 분석 연구

박혜림*[1], 조소정*[2], 한현빈*[3], 유석종**

Text Mining-based Sentiment Analysis of Newly-Coined Words and Implementation of SMU Sentimental Dictionary

Hye-Lim Park*[1], So-Jeong Jo*[2], Hyun-Bin Han*[3], and Seok-Jong Yu**

요 약

인터넷의 발달과 함께 신조어의 생성·소멸 속도는 급속도로 증가하고 있다. 신조어는 세대 간 의사소통에 걸림돌이 된다는 사회적인 문제점을 가진다. 따라서 본 연구는 신조어에 익숙하지 못한 집단에게도 원활한 소통의 기회를 주고자 신조어 분석 시스템을 구축하였다. 본 연구의 분석 시스템은 인터넷 커뮤니티, 소셜 미디어 등에서 발생하는 신조어에 대해 동시 빈출 단어, 카테고리 분석, 감성 값 시각화를 제공한다. 특히 감성 값 분석에 있어 기존에 구축된 감성사전을 신조어에 대해 확장하는 방법론을 제공하며 SMU 감성사전을 구축하였다. 신조어를 포함하는 문장에 대해 본 연구에서 제시하는 SMU 감성사전은 기존 감성사전보다 더 정확한 감성 분석을 가능하게 하였다.

Abstract

With the development of the Internet, the rate of creation and extinction of newly coined words is increasing rapidly. Newly-coined words have the social problem of being an obstacle to inter-generational communication. Therefore, this study established a system for analyzing coined words to give opportunities for smooth communication to groups who are not familiar with them. The analysis system of this study provides simultaneous frequent words, category analysis and emotional value visualization for newly coined words that occur in the Internet community, social media, etc. In particular, Sookmyung Women's Univ (SMU) emotional dictionary was established by providing a methodology for extending the existing emotional dictionary to newly coined words in analyzing emotional values. The SMU emotional dictionary presented in this study for sentences containing newly coined words has enabled more accurate emotional analysis than the existing emotional dictionaries.

Keywords

sentiment analysis, newly-coined word, machine learning, text mining

* 숙명여자대학교 소프트웨어학부
- ORCID[1]: https://orcid.org/0000-0002-7385-5398
- ORCID[2]: https://orcid.org/0000-0003-3567-9044
- ORCID[3]: https://orcid.org/0000-0001-5019-8439
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Nov. 09, 2021, Revised: Dec. 06, 2021, Accepted: Dec. 09 2021
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Womens's University,
Cheongpa-ro 100, 47-gil, Cheongpa-ro, Yongsan-gu, Seoul, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

인터넷의 발달과 함께 언어의 사용 양상이 크게 변화하면서 신조어라는 새로운 형태의 언어가 나타났다. 국립국어원에 따르면 '신조어'는 "새로 생긴 말. 또는 새로 귀화한 외래어" 라고 정의되는 개념이다[1]. 신조어는 표현이 간결하고 전달하고자 하는 바가 뚜렷하여 직관적이며 결합된 독특한 양상 때문에 흥미롭고, 말이 쓰이는 상황까지 담고 있다는 점에서 매우 효율적인 형태의 언어이다. 하지만 신조어는 사전적으로 정의된 의미가 없기 때문에 뜻을 유추하기 어려우며 급속도로 생성·소멸된다는 점에서 신조어의 의미를 굳혀진 단어로 인식하기엔 오류성이 크다. 특히 의미를 모르고 사용할 경우 집단 간 언어 갈등을 불러일으킬 수 있는 위험성을 가진다.

이러한 사회적 문제를 해결하기 위해 신조어의 의미를 유추하고 올바른 사용을 장려하도록 도와주는 연구가 요구된다. 따라서 본 연구는 신조어 의미 분석 연구의 한 방법론으로써 고안한 시스템에 대해서 설명하고자 한다.

감성 분석을 수행하는 방법은 크게 사전을 이용한 방법과 머신러닝을 이용한 방법 두 가지로 구분된다. 기존에 구축되어 널리 활용되는 한국어 감성 사전에는 KOSAC[2]과 KNU 감성 사전[3]이 대표적이다. KOSAC은 한국어 감성 분석 연구에 널리 활용되고자 하는 목적으로 서울대학교에서 개발된 감성 사전이다. 세종 말뭉치로부터 개발한 감성 주석 스키마, KSML을 사용하여 감성 어휘 목록을 선별하고 이를 사전 형태로 제공한다. 약 17,600개의 감성 어휘가 실려 있으나 감성 어휘가 추출된 문장이 특정 도메인에 치우쳐져 있으며 신조어의 비중이 적은 편이다.

본 연구에서는 디시인사이드, 클리앙과 같이 비교적 신조어 사용성이 높은 인터넷 커뮤니티들을 선정해 문장 데이터를 수집하고, 신조어를 다각적으로 분석한 뒤, 이를 총합하여 의미를 유추할 수 있게 하는 시스템을 제공하고자 한다.

구축한 시스템은 뜻풀이가 사전에 등재된 표준어나 비교적 널리 알려진 신조어에 대해선 국립국어원 우리말 샘 API[1]를 이용한 기본적인 사전의 기능을 수행하며, 그렇지 않은 단어들, 즉 신조어에 대해서는 의미 분석 기능을 수행한다. 이때 분석 요소는 동시 출현 단어 분석, 신조어 카테고리 분석과 예시 문장 감성 분석으로 구성된다. 이 중 예시 문장 감성 분석은 감성 사전 확장 방법론을 이용하여 구축한 SMU(Sookmyung Women's Univ.) 감성 사전을 활용해 신조어를 포함한 문장에 있어 기존 감성 사전 기반 감성 분석 기법보다 정확도를 더욱 높이고자 하는데 초점을 둔다.

이를 위한 본 논문의 구성은 다음과 같다. 2장에서는 기존 신조어 관련 서비스와 감성 분석 관련 연구에 대해 살펴보고 각 한계점에 대해 논한다. 3장에서는 본 논문에서 제안하는 신조어 분석 시스템과 해당 시스템에서 제공하는 신조어 동시 빈출 단어, 카테고리 값, 감성 분석에 대해 설명하고 사용한 기술을 소개한다. 4장에서는 감성 분석에 있어 제안하는 SMU 감성 사전의 구현 과정과 기존 감성 사전과의 성능 비교 및 분석을 논하며 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

2.1 신조어 관련 서비스 - 네이버, 위키백과

신조어의 뜻풀이를 제공하고자 하는 대표적인 시스템에는 '네이버 open 사전'과 웹 페이지 목록으로 정리된 '위키백과 대한민국 인터넷 신조어 목록'이 있다. 위 두 시스템은 정적으로 구축된 사전 데이터베이스에서 보여주는 식으로 사전의 기능에 충실히 뜻풀이를 수행한다. 네이버 open 사전은 네이버 사전에서 확장된 사용자 참여형 사전이다. 사용자들끼리 기존 사전 범위에 없는 단어들에 대해서 질문하고 단어의 뜻을 등록, 수정, 뜻풀이 추천, 반대하는 등의 기능을 수행한다. 위키백과 신조어 목록은 가나다순으로 신조어의 의미를 추가, 삭제, 수정 할 수 있게 한 역시 사용자 참여형 사전이다. 하지만 정리된 신조어의 양이 200여개로 적고 단어의 의미에만 초점을 두어 신조어가 어떻게 사용되는지 비슷한 단어에는 무엇이 있는지 신조어 사용 현황을 해석하는데 한계점을 가진다.

2.2 표준어 감성 사전

KNU 감성 사전은 군산대학교에서 개발한 감성 사전이다. 이 사전은 표준 국어 대사전의 뜻풀이 감성을 Bi-LSTM을 활용하여 뜻풀이의 감성 값을 분류한 뒤, 긍정으로 분류된 뜻풀이에서는 긍정에 대한 감성 어휘를 부정으로 분류된 뜻풀이에서는 부정에 대한 감성 어휘를 n-gram으로 추출하여 표 1과 같이 약 14,800개의 감성 어휘를 구축하였다. 각 단어는 매우 부정(-2) 부정(-1) 중립(0) 긍정(1) 매우 긍정(2)의 감성 값을 가진다. KOSAC 감성 사전과 달리 도메인에 영향 받지 않는 어휘들로 구성하였다는 점에서 감성 분석의 정확도를 높이고자 하였다. 또한 표준 국어 대사전 외에 감성 단어사전, 위키피디아 신조어 목록, 위키피디아 이모티콘 목록을 참조하여 다양한 단어의 감성 값을 고려하였다. 하지만 KNU 감성 사전 역시 표준 감성 어휘를 중점으로 하고 부수적으로 신조어를 고려하였기 때문에 급속도로 생성·소멸되는 신조어들을 수용하지 못한다는 점에서 한계를 지닌다.

표 1. KNU 감성사전 통계
Table 1. KNU sentimental dictionary statistics

Sentiment level	words
Very positive (2)	2,597
Positive (1)	2,266
Neutral (0)	154
Negative (-1)	5,029
Very negative (-2)	4,797
Total	14,843

2.3 신조어 및 비속어 감성 분석

[4]는 기존 감성 사전에 정의가 안 된 비속어의 감성을 분석하기 위해 비속어 학습 모델을 전체 시스템에 추가하여 구성하였다. 비속어를 해석하기 위해 댓글들을 학습하여, 문장 내에서 유사한 관계에 있는 단어들을 찾아내는 word2vec 알고리즘을 사용하여 기존 감성 사전을 확장하였다. 하지만 비속어 사전을 구축할 때, word2vec 알고리즘을 통해 구한 단어들의 유사도에 따른 가중치를 두지 않아 감성을 섬세하게 표현할 수 없었다. 또한, 기계학습에 이용되는 댓글 데이터가 매순간 업데이트되어 양이 방대해지는 것을 고려하지 않아 모델을 학습시키는 시간이 오래 걸려 사용자에게 결과를 도출해 내는 과정에 긴 시간이 소요되었다.

[5]는 음식점 평점 예측을 위하여 KNU 감성 사전을 이용하였다. 한국어로 작성된 음식점 리뷰를 대상으로, 감성분석을 수행하여 평가 항목별로 세분화된 평점을 제공 가능한 예측 방법론을 제안한다. 한국어 리뷰를 KNU 감성사전을 통해 수치화하여 평점 예측에 직접 반영함으로써 평점 예측 성능을 개선하였으나 신조어 및 축약어 등을 포함한 다양한 감성 어휘들을 반영한 추가 연구의 필요성을 언급하였다.

III. 신조어 감성 분석 시스템

3.1 시스템 구조

신조어의 생성·소멸 속도는 매우 빨라 단어가 사전에 등재되는 과정 중에 소멸되거나 뜻풀이가 변화할 가능성이 높다. 이러한 신조어의 특징을 고려하여 본 시스템에서는 신조어 사용이 활발한 인터넷 커뮤니티의 데이터를 활용한 의미 분석 시스템을 제안하고자 한다.

그림 1은 본 논문에서 제안하는 신조어 분석을 위한 시스템의 전체 구조를 나타낸다. 해당 시스템은 다음과 같은 과정을 거친다.

웹서버는 커뮤니티에서 데이터 크롤러(Data crawler)를 이용하여 문장데이터를 수집한다.

1) 수집한 문장데이터를 전처리한 후, 신조어 추출 및 필요한 모델 학습에 사용한다.

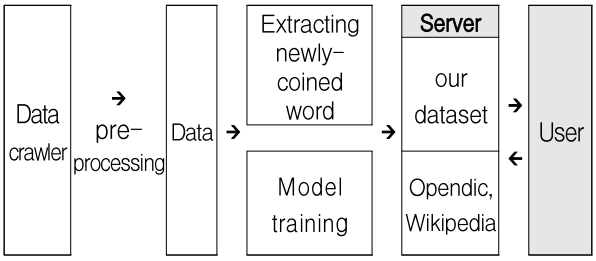

그림 1. 신조어 의미 분석 시스템 구조
Fig. 1. Architecture of newly-coined words semantic analysis system

2) 웹서버에 신조어 분석 데이터(JSON)를 저장한 후, 사용자 요청에 따라 출력해준다.
3) 요청 받은 단어에 우리말 샘 사전 API[1]나 위키백과 신조어 목록에 등재된 경우, 그 의미를 함께 출력해준다.

3.2 데이터 수집 및 전처리

다양한 연령대가 사용하는 신조어를 고루 추출하기 위해 디시인사이드, 클리앙, 유튜브 3개의 사이트를 선정하였다. Python Selenium 프레임 워크, Beautiful soup 패키지를 이용해 각 웹 사이트에서 제목 및 댓글 문장 데이터를 수집한다. 수집한 문장 데이터에 대한 정보는 표 2와 같다.

표 2. 수집한 문장 데이터 정보
Table 2. Information of collected sentence data

Sites	Clien	Dcinside	Youtube
Period	2019.08 ~ 2020.07	2020.03 ~ 2020.06	2020.08.03/ 2020.08.10
Size	43.7MB	45.8MB	12.3MB
Lines	695,771 lines	954,251 lines	112,353 lines
Total	100.8 MB (1,762,375 lines)		

디시인사이드의 게시물 제목은 대부분 완전한 문장의 형태로 끝나지 않고 'ㅋ'음절이 다수 나타나는 반면 클리앙의 경우 'ㅋ'음절의 반복이 적게 나타났다. 수집한 데이터에 대해 데이터 분석 라이브러리인 Pandas와 Numpy를 활용하여 먼저 'ㅋ'음절과 특수 기호를 제거한 뒤, 2차적으로 알파벳, 정치 및 종교 관련 단어, 비속어 등과 같은 금칙어 제거 과정을 거쳐 총 59MB의 전처리된 문장 데이터로 취합한다.

3.3 신조어 판별 및 추출

신조어를 판별하고 추출하기 위해선 3.2.1에서 수집한 데이터 셋을 토크나이징 하는 작업이 필요하다. 본 연구에서는 비지도 학습 기반인 Soynlp[6]의 MaxScore는 WordExtractor와 KoNLPy의 Okt를 활용한 방법을 통해 토크나이징을 진행하였다. 후에 Sejong Corpus[7]와 Hunspell 맞춤법 검사기를 이용하여 표준성을 검사한뒤, 해당 단어를 신조어 후보로 채택한다.

3.4 카테고리 분류

신조어가 생성되는 원리를 바탕으로 '합성어', '줄임말', '초성어', '기타'의 4가지 범주로 카테고리 값을 범주화 하였다.

3.3에서 추출된 후보 단어에 대해 ~력, ~템, ~주의 등과 같은 결합 어미를 가진 단어일 경우, 결합 어미의 뜻을 카테고리 값으로 부여한다.

그 외의 후보 단어들에 대해서는 지도 학습을 통해 분류한다. 지도 학습을 위한 라벨 값은 KoNLPy의 Kkma 형태소 분석기를 이용해 토크나이징 하였다. 단어의 품사가 명사 + 명사인 경우 '합성어'로, 단어의 길이와 토크나이징 된 길이가 같다면 '줄임말'로, 자음 또는 모음만으로 이루어진 단어인 경우 '초성어'로, 그 외의 단어들에 대해 '기타'로 1차 분류 한 뒤, 정확성을 높이기 위해 2차로 수작업 라벨링을 거친다. 라벨링된 모든 단어들의 Pos Tagging 값을 TF-IDF으로, 라벨 값은 One-hot Encoding을 통해 벡터화한 뒤 Keras의 1D-CNN 층 기반의 모델에 주입하여 분류 모델을 만든다.

표 3은 생성 JSON 파일 내 단어들을 1D-CNN 모델에 주입한 결과 라벨링 된 데이터와 비교하여 나타낸 모델의 정확도를 나타낸다.

표 3. 1D - CNN 모델의 정확도
Table 3. Accuracy of 1D-CNN models

Word set	941
Correct words	786
Not correct words	155
CNN model accuracy : 83.5	

모델의 초기 정확도는 0.75였으나 합성어, 줄임말, 기타 클래스 분류의 균형을 위해 class간 가중치를 0.2, 0.6, 0.2로 부여한 뒤 모델의 정확도가 0.835로 상승하였다.

3.5 동시 빈출 단어

본 시스템에서는 신조어의 의미 파악을 위한 동시 빈출 단어를 시각화하여 이미지 형태로 제공한다. 동시 빈출 단어 추출을 위해 페이스북의 단어 임베딩 라이브러리인 FastText를 사용한다. FastText는 3.2와 같이 데이러 수집 및 전처리를 거친 약 50MB의 문장 데이터를 통하여 학습시켰다. 학습시킨 FastText의 Most_Similar 함수를 이용하여 유사도 상위 10개 단어를 WordCloud를 이용하여 시각화하였고, 사용자에게 이미지 형태로 출력한다.

3.6 감성 분석

본 연구에서는 신조어 및 비속어의 사용이 활발하고 비정형화 된 인터넷 텍스트의 감성 분석 정확도를 높이기 위해 SMU 감성 사전을 구축하였다. SMU 감성 사전은 KNU와 같이 기존에 구축된 감성 사전을 기반으로 미등록 언어(신조어, 비표준어)들의 감성 값을 도출해내는 확장 방법론을 기초로 한다. SMU 감성 사전 구축에는 Naver 영화 감상평 데이터 셋을 사용했으며, 구현 과정은 다음과 같다.

1) Naver 영화 감상평 데이터 전처리
2) 신조어 추출 및 “신조어 - 예시 문장” 형태 저장
3) KNU 감성 사전을 통한 예시 문장 감성 값 추출
4) 신조어 당 예시 문장 긍·부정 개수 통계
5) 4)를 통한 신조어 감성 값 부여
6) SMU 감성 사전 생성

IV. 실험 설계 및 성능 평가

4.1 데이터 셋 전처리

본 연구에서는 신조어와 같이 기존 감성사전에 미등록된 단어의 감성 값을 구하기 위해 SMU 감성 사전을 구축하였다. 이는 신조어나 비표준어를 포함한 문장에 있어 더 정확한 감성 분석을 하기 위함에 있다. SMU 감성 사전의 감성 분석 성능 평가를 위해서 본 연구에서는 감성 분석의 정답 값이 존재하는 Naver 영화 감상평 데이터 셋[8]을 이용하였다. 해당 데이터 셋에는 감성 정답 값이 존재하며, 인터넷 게시물 특성상 신조어 및 비표준어의 사용이 빈번하다. 데이터의 원래 크기와 전처리 후 크기는 표 4와 같다. 전처리 과정에서 숫자, 영어, 특수문자를 제거하고 20자 미만의 길이 문장 데이터를 제거하였다.

표 4. Naver 영화 감상평 데이터 통계
Table 4. Naver movie review data statistics

File	Lines	Lines after pre-processing
Naver_movie.txt	200,000 lines	120,010 lines

4.2 감성 분석

Naver 영화 감상평 데이터를 3.3의 신조어 판별 및 추출을 통해 KNU에 존재하지 않으며(감성 값이 존재하지 않으며), 신조어 734개를 추출하였다. 신조어와 각각의 예시 문장(신조어 포함 문장)에 대해 KNU 감성 사전을 이용하여 식 (1) 과 같이 감성 값을 추출하였다. 해당 공식에서 추출된 감성 값이 0 보다 클 경우 ‘긍정’으로, 0 보다 작을 경우 ‘부정’으로, 0일 경우 ‘zero’로 그 개수를 저장하였다.

$$Sentiment = \frac{\sum_{i=1}^{k} Polarity(T_i)}{k} \quad (1)$$

$Sentiment$ = 해당 예시문장의 감성 값
$Polarity$y = 감성 어휘의 강도 (-2, -1, 0, 1, 2)
T_i = 문장 내 i번째 감성 어휘
k = 예시 문장에 출현하는 감성 어휘의 개수

총 데이터에서 신조어 마다 총 출현 수, 긍정 문장 개수, 부정 문장 개수, zero(감성 값 0)의 개수를 저장하고, 해당 개수를 통하여 SMU 감성사전에 등재될 감성 값을 표 5와 같은 기준으로 설정하였다.

표 5. SMU 감성 값 설정 기준
Table 5. SMU sentiment value setup criteria

r = postive lines / negative lines	
r >= 2	2
r >= 1	1
r <= 1	-1
r <= 0.5	-2
Else	0

해당 방법으로 구축한 SMU 감성 사전에 등재된 단어의 예시 및 통계 자료는 표 6과 같다. 총 734개의 신조어 중 데이터 내 출현 수 상위 7개에 대한 통계자료이며, 왼쪽부터 차례로 데이터 내 신조어의 총 출현 수, KNU로 문장의 감성 값을 계산하였을 때 긍정(0 보다 큰 수)인 문장 개수, 부정 문장 개수, 0인 문장 개수, Naver 정답 값과의 감성 일치 횟수, 불일치 횟수, 일치도(일치 횟수 / 총 출현 수)로 구성된다.

예를 들어, KNU에는 'ㅋㅋ', '알바', '코믹', '초딩'과 같은 단어의 감성 값이 등재되지 않았으므로 zero문장 개수가 많고 일치도 또한 낮다. SMU는 'ㅋㅋ'와 같은 신조어의 감성 값을 얻기 위해서 KNU의 '긍정 문장 개수'와 '부정 문장 개수'를 이용하여 표 6과 같은 공식을 통해 'ㅋㅋ'의 감성 값을 구하며, 표에 따르면 'ㅋㅋ'의 감성 값은 1에 해당한다. 이와 같이 구한 감성 값은 SMU 감성사전에 등재된다.

본 연구에서 구축한 SMU 감성사전을 이용한 감성 분석은 그림 2의 순서와 계산식을 이용하며, KNU와 같이 734개의 신조어에 대해 계산된 결과는 표 6의 파란색 부분에 해당한다. KNU와의 감성분석 정확도 비교는 정답 레이블인 네이버 감성 사전의 감성 값과의 일치 개수를 통하여 도출하였다. (일치도 = 일치 개수/총 출현 수) 구축한 SMU 사전을 이용한 감성 분석 과정은 그림 2와 같은 순서도로 나타낼 수 있다.

4.3 실험 결과

Naver 영화 감상평 데이터에 대해 KNU와 SMU의 정확도(정답과의 일치도) 비교 결과는 표 7과 같다. 네이버 영화리뷰 데이터에 대해 기존 감성사전인 KNU를 사용하였을 때 보다 SMU 감성사전을 사용할 경우 신조어를 포함한 문장에 대해 정확도가 17.3% 증가한 것으로 나타난다. 이는 기존 감성사전에 포함되지 않는 신조어에 의미 있는 감성 값을 부여하여 미등록 문제, OOV(Out Of Vocabulary)를 해결하였기 때문이라고 볼 수 있다.

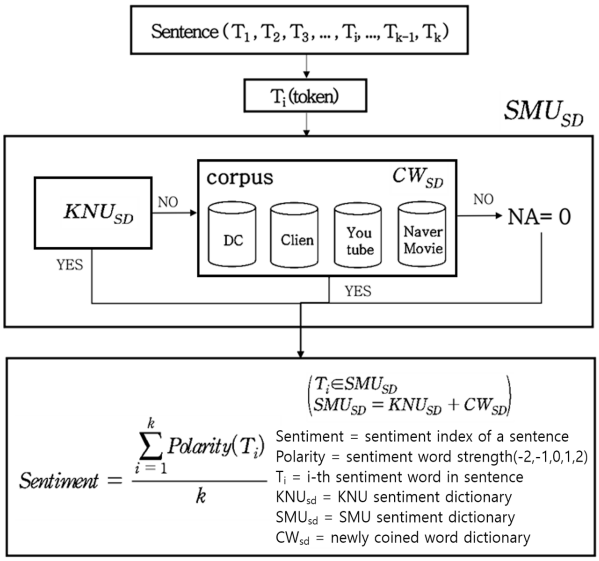

그림 2. SMU 감성 사전 구축 순서도 및 감성 계산 식
Fig. 2. SMU sentimental dictionary deployment flowchart and sentiment equation

표 6. SMU 등재 단어 및 감성 값 (데이터 내 총 출현 수 상위 7개)
Table 6. SMU registered words and sentiment values (Top 7 total appearances in data)

		KNU					SMU				
Newly coined words	Frequency	Pos. lines	Neg lines	Zero lines	Matches (Matched/ freq.)	KNU sentiment index	Pos. lines	Neg. lines	Zero lines	Matches (Mached/ freq.)	SMU sentiment index*
ㅋㅋ	3766	1825	1013	928	0.54	NULL	3182	451	133	0.61	1
알바	580	200	213	167	0.4	NULL	130	409	41	0.68	-1
코믹	578	293	159	126	0.58	NULL	471	71	36	0.6	1
초딩	448	166	175	107	0.51	NULL	122	293	33	0.61	-1
좀비	318	102	138	78	0.48	NULL	69	227	22	0.62	-1
ㄷㄷ	258	98	97	63	0.46	NULL	192	52	14	0.63	1
힐링	171	116	34	21	0.67	NULL	162	6	3	0.84	2
망작	170	53	63	54	36	NULL	42	115	13	0.67	-1

표 7. Naver 리뷰 정답과의 일치도 비교
Table 7. Sentimental pre-accuracy comparison

Sentiment dic.	KNU	SMU
Accuracy	52.1 %	61.6 %

SMU 감성사전을 이용한 신조어 분석 시스템의 구현 결과는 그림 3과 같다. 예를 들어, 신조어 '크보'가 포함된 문장에 대한 긍정 비율이 66.7%, 부정 비율이 33.3%를 차지하였다.

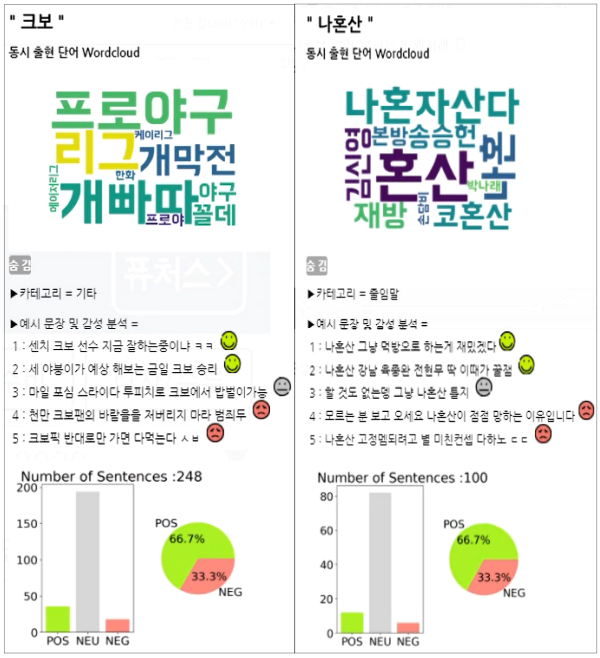

그림 3. 신조어 분석 시스템 구현 결과
Fig. 3. Results of implementation of the newly-coined words analysis system

표 8. 신조어 분석 시스템 데이터 셋 통계
Table 8. New word analysis system dataset statistics

JSON analysis result	
JSON total size	957 KB
Total words	1,066
Newly coined words	771
Non-coined word	295
Coined word rate	0.72

Category	Examples	Count
Compound	국뽕, 기레기, 꿀팁, 치트키, 혼밥	628
Shortened	개콘, 나혼산, 디씨, 몰카, 뮤비	376
Consonant	ㄱㄱ, ㄷㄷ, ㄱㅊ, ㅅㅅ, ㅎㄷ	39
Etc	군머, 크보, 능지	18

3절에서 소개한 바와 같이, 사용자는 본 시스템을 통해 신조어의 동시빈출단어와 카테고리 값, 그리고 예시문장에 있어 SMU 감성사전을 통해 더 정확한 감성 값을 확인할 수 있다. 본 시스템의 데이터 셋에 등재된 신조어의 예시는 표 8과 같다. 데이터셋의 신조어 비율이 72%이고, 합성어, 줄임말, 초성어 형태의 신조어의 순서로 분포하였다.

V. 결론 및 향후 과제

본 논문에서는 기존에 시도된 바 없던 한국어 신조어 의미 분석 시스템을 제안 및 구현하고 미등록 언어 감성 분석 시스템을 구현하였다는데 연구 의의를 찾을 수 있다.

또한 신조어나 비속어를 포함한 문장에 있어서 기존의 감성사전 기반 방법의 한계점을 해결할 수 있는 방법론으로 SMU 감성사전의 구축하였고 개선된 정확도를 확인하였다. 본 연구에서는 KNU 감성사전과 네이버 영화 감상평 문장 데이터를 통하여 신조어, 비표준어에 대한 감성 분석의 정확도를 높였지만, 해당 방법론은 주어진 문장 데이터와 감성사전을 기반으로 신조어 감성 사전을 구축하므로 일반 문장 데이터에서도 사용이 가능하다. 특히 신조어의 사용이 활발하고 혐오 표현이 잦은 인터넷 커뮤니티에서 활용도가 높다고 할 수 있다.

본 연구의 한계점은 다음과 같다. 연구 과정에서 신조어 후보를 해치지 않고 토크나이징하기 위해 데이터 전처리 과정 중 한글을 제외한 한자, 영어, 특수 문자 등을 제거하였다. 그 결과, 의미 생성 범위가 한글에만 국한되어 나타난다는 한계점이 생겼다. 또 다른 한계점은, 감성 분석 과정에서 신조어를 포함된 문장(예, '후회하지 않는다')이 이중 부정 문인 경우 부정 값이 상쇄되는 것이 아니라 누적된다는 점이다. 따라서 추후, 문맥상 문장이 강한 부정인지 긍정의 의미를 나타내는 것인지 파악할 수 있도록 보완이 필요하다.

References

[1] National Institute of Korean Language, https://stdict.korean.go.kr/main/main.do. [accessed: Sep. 20,

2020]
[2] KOSAC, Korean Sentiment Analysis Corpus , http ://word.snu.ac.kr/kosac/. [accessed: Oct. 15, 2020]
[3] S. M. Park, C. W. Na, M. S. Choi, D. H. Lee, and B. W. On, "KNU Korean Sentiment Lexicon: Bi-LSTM-based Method for Building a Korean Sentiment Lexicon", Journal of Intelligence and Information Systems, Vol. 24, No. 4, pp. 219-240, Dec. 2018. https://doi.org/10.13088/jiis.2018.24.4.219.
[4] S. J. Kim, J. E. Kim, W. Y. Seong, and Y. H. Kim, "Design of Video Advertisement Analysis via Analysis of Internet Term Sensitivity", Journal of KIISE, Vol. 46, No. 9, pp. 919-925, Sep. 2019. https://doi.org/10.5626/JOK.2019.46.9.919.
[5] J. S. So and P. S. Shin, "Rating Prediction by Evaluation Item through Sentiment Analysis of Restaurant Review", Journal of The Korea Society of Computer and Information, Vol. 25 No. 6, pp. 81-89, June. 2020. https://doi.org/10.9708/jksci.2020.25.06.081.
[6] Unsupervised Korean Natural Language Processing Toolkits, https://pypi.org/project/soynlp. [accessed: Sep. 12, 2020]
[7] Sejong Corpus,, https://github.com/coolengineer/sejong-corpus. [accessed: Nov. 12, 2020]
[8] Naver Sentiment Movie Corpus, https://github.com/review/NSMC_Sentimental-Analysis. [accessed: Nov. 10, 2020]

저자소개

박 혜 림 (Hye-Lim Park)

2017년 2월 ~ 현재 :
숙명여자대학교 소프트웨어학부
관심분야 : 머신러닝, 자연어처리, 영상처리

조 소 정 (So-Jeong Jo)

2017년 2월 ~ 현재 :
숙명여자대학교 소프트웨어학부
관심분야 : 프로그래밍, 머신러닝, 자연어처리

한 현 빈 (Hyun-Bin Han)

2018년 2월 ~ 현재 :
숙명여자대학교 소프트웨어학부
관심분야 : 프로그래밍, 데이터마이닝

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 컴퓨터과학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수
관심분야 : 데이터마이닝, 추천시스템, 정보시각화