The Journal of the Convergence on Culture Technology (JCCT)
Vol. 7, No. 2, pp.149-156 May 31, 2021. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2021.7.2.149
JCCT 2021-5-16

정치 도메인에서 신조어휘의 효과적인 추출 및 의미 분석에 대한 연구

Study on Effective Extraction of New Coined Vocabulary from Political Domain Article and News Comment

이지현*, 김재홍**, 조예성***, 이민구****, 최혜봉*****

JihyunLee*, Jaehong Kim, Yesung Cho***, Mingu Lee****, Hyebong Choi*****

요 약 정치적 사안에 대한 대중의 의견과 인식을 객관적으로 이해하기 위한 방법으로 텍스트 마이닝을 통한 빅데이터 분석을 수행할 수 있다. 기존 어휘 사전에 기반한 텍스트 마이닝 알고리즘은 신조어와 같이 사전에 수록되지 않은 어휘를 분석하는데 한계가 나타난다. SNS를 통해 나타나는 사용자들의 의견은 많은 경우 신조어와 비속어를 포함하는데, 이러한 어휘들을 효과적으로 분석하지 못한다면 정확한 대중의 인식과 의견을 파악하기 어렵게 된다. 본 논문은 정치 섹션의 뉴스 댓글로부터 정치적 의미성을 지니는 신조어와 비속어를 효과적으로 추출하는 방법을 제안하고, 추출한 신조어휘들의 의미와 맥락을 이해하기 위한 다양한 방법을 제시하였음.

주요어 : 텍스트 마이닝, 인터넷 댓글, 형태소 분석, 신조어 분리, 뉴스 댓글 분석

Abstract Text mining is one of the useful tools to discover public opinion and perception regarding political issues from big data. It is very common that users of social media express their opinion with newly-coined words such as slang and emoji. However, those new words are not effectively captured by traditional text mining methods that process text data using a language dictionary. In this study, we propose effective methods to extract newly-coined words that connote the political stance and opinion of users. With various text mining techniques, I attempt to discover the context and the political meaning of the new words.

Key words : Text mining, Online comment, Tokenize, Neologisms tokenize, News comment analysis

*준회원, 한동대학교 전산전자공학부 학부과정
**준회원, 한동대학교 커뮤니케이션학부 학부과정
***준회원, 한동대학교 ICT 창업학부 학부과정
****준회원, 한동대학교 커뮤니케이션학부 학부과정(교신저자)
*****정회원, 한동대학교 ICT 창업학부 조교수
접수일: 2021년 3월 2일, 수정완료일: 2021년 4월 3일
게재확정일: 2021년 4월 18일

Received: March 2, 2021 / Revised: April 3, 2021
Accepted: April 18, 2021

*Corresponding Author: 21400525@handong.edu
Dept. of Coummincation Arts, Handong Univ, Korea

Ⅰ. 서 론

텍스트 마이닝은 평서문(plain text)으로 구성된 데이터로부터 유용한 정보와 지식을 추출하는 방법이다. 최근 텍스트 마이닝을 이용하여 사회 관계망 서비스(SNS:Social Network Service)로부터 정치적 사안에 대한 대중의 인식과 의견을 분석하는 연구가 진행되고 있다. Jung et al.(2020)은 2016년 SNS(트위터)로부터 세월호사건에 대해 서로 다른 정치성향의 사용자들이 작성한 글을 분석하여 두 집단 사이의 사용어휘와 의견의 차이점을 분석하였다[1]. Ahn et al.(2017)은 대선 후보들에 관한 대중의 인식과 여론을 포털 뉴스 기사와 댓글로부터 추출하여 각 후보들에 대한 유권자들의 의견을 분석하였다[2]. 또한, Han et al.(2019), Kang et al.(2017)과 Lee et al.(2020)은 포털 뉴스 댓글에서 나타나는 정치적 편향성을 분석하고 파악하는 방법을 제안하였다[3][4][5].

포털 뉴스 댓글을 이용하는 사람들은 자신의 정치적 입장과 의견의 표현에 정치적 의미를 지니는 신조어와 비속어를 포함하는 경향이 있다[6]. 따라서, 텍스트 마이닝 기법으로 뉴스 댓글을 분석하여 대중의 정치적 인식과 의견을 파악하기 위해서는 문맥 속에서 사용된 신조어와 비속어를 효과적으로 추출하고 정치적 의미를 이해하여야 한다.

텍스트 마이닝에서 평서문을 단어나 어휘와 같은 의미의 최소단위로 분할하는 과정을 분절화(tokenization)라고 한다. 분절화된 단어를 토큰(token)이라고 하는데, 주어진 평서문을 토큰의 집합으로 보고 개별 토큰들의 의미관계, 분포, 빈도 등을 분석하여 원글의 맥락과 의미를 파악하게 된다. 전통적인 분절화 방법들과 형태소 분석기들은 어휘 사전을 기반으로 분절화를 수행한다[7]. SNS와 뉴스댓글을 사용하는 인터넷 사용자들은 어휘 사전에 등록되어 있지 않은 신조어와 비속어를 사용하여 의견을 표현하는 경우가 많다. 따라서 이러한 어휘들은 기존 어휘 사전에 기반한 분절화 방법으로 효과적으로 추출하지 못하는 문제가 발생한다.

본 연구에서는 인터넷 정치 신조어와 비속어를 효과적으로 추출하기 위한 분절화 방법을 제시하며, 이를 통해 정치 뉴스 댓글 작성자들의 정치적 인식과 의견을 분석할 수 있게 한다. 또한, 동시출현 단어분석 방법(Co-word analysis)과 워드 임베딩방법(Word embedding)을 활용하여 추출된 신조어와 비속어의 의미와 맥락을 파악하고 효과적으로 정치 신조어가 추출되었음을 검증하였다.

Ⅱ. 데이터 수집 및 처리 과정

1. 뉴스 댓글 데이터 수집

정치 관련 신조어와 비속어를 추출하기 위해, 19대 정부가 출범한 2017년 6월을 기준으로 약 3년 7개월, 2020년 12월까지의 뉴스 댓글을 수집하였다. 국내에서 가장 많은 사용자를 갖고 있는 대표적인 뉴스 포털 네이버와 다음에서 각각 85,000개 가량의 댓글을 수집하여 총 170,000개의 댓글에 대하여 분석을 수행하였다. Hyun et al.(2020)의 연구에 따르면 뉴스를 직접 생산하지 않는 포털은 정파적 편향성을 갖지 않지만, 포털을 통해 뉴스를 소비하는 이용자들은 그들의 정파성에 따라 선호하는 포털이 다르게 나타난다[8]. 서로 다른 정치 성향을 갖는 이용자들의 사용 어휘를 균일하게 수집하기 위하여 두 포털에서 동일한 수준의 뉴스 댓글을 수집하였다.

본 연구는 '정치' 관련 신조어와 비속어 추출에 목적을 두고 있다. 따라서 정치적 의미를 갖는 어휘와 일상 어휘를 구분하기 위해 정치 섹션 뉴스기사와 비정치 역역의 뉴스기사로부터 각각 댓글을 수집하였다. 수집된 약 170,000개 댓글 중 약 50%는 정치 분야 섹션의 뉴스기사로부터 수집하였으며, 나머지 50%는 정치 외 섹션(IT, 생활, 문화, 사설칼럼, 국제) 뉴스 기사에서 수집하였다. 데이터 수집에는 Python 3.6의 Selenium 3.141.0 버전과 BeutifulSoup 4.9.0 버젼의 모듈을 사용하였다[9].

2. 신조어 비속어 추출

1) 띄어쓰기 기반 분절화

전통적인 한국어 텍스트 마이닝에서는 형태소 단위로 분절화가 이루어진다. 형태소는 의미를 가지는 가장 작은 말의 단위를 의미하며 전통적인 한국어 텍스트 마이닝에서는 한국어 어휘 사전을 기반으로 이루어진다. 굴절어인 영어와 달리 조사, 한국어의 경우 다양한 어미변화로 나타나는 복잡한 구조가 존재한다. 따라서 한글 형태소 분석을 위한 다양한 방법들이 독립

적으로 연구되고 있다[10]. 기존 어휘사전에 기반한 분절화 방법은 사전에 나타나지 않는 신조어와 비속어를 효과적으로 추출하지 못하는 한계점이 나타난다. 본 연구에서는 어휘사전 기반 분절화 대신 띄어쓰기를 기준으로 분절화를 수행하여 신조어와 비속어를 효과적으로 추출하려고 하였다. 그림 1은 어휘사전 기반 분절화 방법 중 하나인 Komoran 패키지를 사용하여 분절화를 진행한 방식과 띄어쓰기를 기준으로 분절화를 진행한 방식을 비교한 그림이다[11].

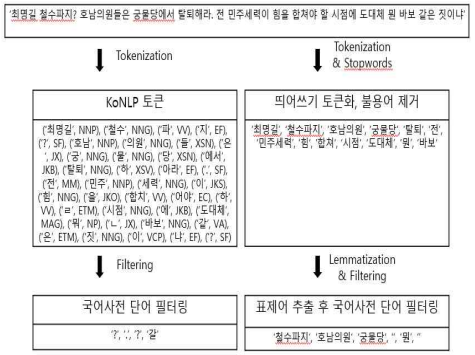

그림 1. KoNLP기반 분절화와 띄어쓰기 기반 분절화 결과 비교

Figure 1. Comparing tokenization by KoNLP packages and tokenization by word spacing

그림 1을 보면, 어휘사전 기반 분절화 알고리즘을 사용하여 분절화를 진행한 경우 사전에 등록되어 있지 않은 '궁물당'과 같은 단어를 하나의 독립된 명사 단위로 인지하지 못하고 '궁'(명사), '물'(명사), '당'(접미사)이라는 독립된 단위로 분절화 되어 원래의 의미를 잃어버린다. 그러나, 띄어쓰기로 분절화하는 경우, '철수파지', '호남의원', '궁물당' 등과 같은 신조어들이 분할되지 않고 원래의 의미를 유지한다.

2) 한국어 토큰 전처리

불용어(stop-word)는 영어에서 정관사, 부정관사, Be 동사처럼 빈번하게 사용되며 의미적 표현보다는 문법적 구성요소로 사용되는 단어들을 의미한다. 이러한 불용어 들은 텍스트의 의미를 더욱 분명히 파악 위해서 전처리 과정에서 제거되곤 한다. 그림 2는 띄어쓰기 단위로 문장을 분절화한 이후 신조어와 비속어를 추출하는 과정을 표현하고 있다.

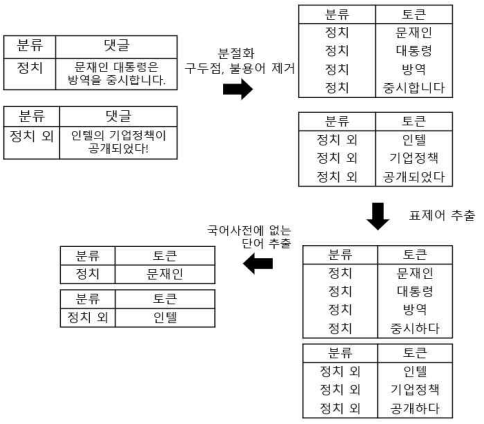

그림 2. 뉴스 댓글 전처리 과정

Figure 2. Process of preprocessing news comment

띄어쓰기 단위로 문장을 토큰(token)으로 분할한 후 토큰 뒤에 붙는 '는 ', '을', '죠' 와 같은 조사/지시대명사 등 184개를 불용어로 지정하여 이를 제거하였다. 이후 불용어를 제거한 단어를 표제어로 재구성하여(중시합니다 -> 중시하다) 토큰들을 기본 사전형으로 변환하였다. 토큰을 기본 사전형으로 변환하기 위해 soylemma 패키지의 Lemmatizer를 사용하였다[12].

사전형으로 변환된 단어 중 신조어와 비속어를 구분하기 위해 추출한 단어로부터 국어사전에 등재된 단어를 제거하는 과정을 거쳤다. 국어사전은 국립국어원 언어정보나눔터의 전자사전을 사용하였다[13]. 국어사전에 등록된 단어를 삭제하는 과정을 거쳐 131,821개의 신조어와 비속어를 추출하였다.

3. 정치 관련 신조어, 비속어 추출

뉴스 댓글로부터 정치적 사안에 대한 대중들의 인식과 의견을 분석하기 위해서는 정치적 의미성이 높은 신조어와 비속어를 구분하여 추출할 필요가 있다. 이를 위해, 본 논문에서는 정치 분야에서 상대적으로 유의미한 차이를 두고 많이 출현하는 어휘를 정치적 의미를 지니는 신조어로 정의하였고, 이러한 특징을 가지는 단어를 추출하는 과정을 거쳤다.

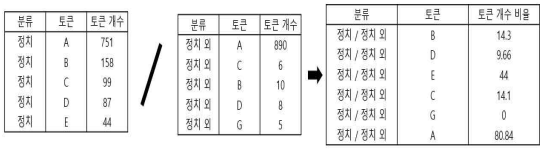

그림 3. 정치 관련 토큰 추출 과정
Figure 3. Political token extraction process

그림 3은 이전 단계에서 추출한 신조 어휘를 대상으로 정치 분야 뉴스와 정치 외 분야 뉴스에서의 출현 빈도 비율을 계산하는 과정이다. 계산 식은 아래와 같다.

$$\text{출현빈도비율} = \frac{n_{\text{정치 분야}}}{n_{\text{정치 외 분야}} + 1}$$

수식 1. 정치 외 분야 대비 정치 분야 단어 출현빈도비율
Expression 1. Percentage of frequency of word appearance in politics comared to non-politics fields

출현빈도비율 값은 해당 단어가 정치 분야에서 정치 외 분야 대비 빈번하게 출현하는 정도를 나타낸다. 예를 들어, '나베'라는 단어가 정치 분야에서 400번 정치 외 분야에 53번이 들어간 경우, 비율 값은 약 7.42가 나오게 된다. 반면, '주구장창'이라는 단어의 경우 정치 분야에 30번, 정치 외 분야에 16번 정도 들어가게 되면 비율 값은 약 1.82가 나오게 된다.

한 분야에서만 출현하는 단어의 경우, 비율을 계산할 때 분모가 0이 되는 문제가 발생하므로 이를 방지하기 위해서 추출한 모든 신조어에 대해서 출현빈도 값에 1을 더한 후 비율 계산을 수행하였다.

본 연구에서는 정치 분야에서 등장한 빈도수가 정치 외 분야에서 등장한 빈도수보다 두 배 이상 많은 경우 유의한 정치 신조어 및 비속어라고 판단하여 비율 값이 2 이상인 단어들을 추출했다. 표1의 좌측 단어들은 제거된 단어(비율 값이 2 미만)의 표본이다. 표1은 뉴스 댓글로부터 추출한 단어들과 단어의 출현빈도비율을 나타낸 것이다. 우측은 출현빈도비율이 2 이상인 단어들로 일반 뉴스 댓글에 비해서 정치 뉴스 댓글에서 2배 이상 높은 빈도로 출현한 단어이다. '홍영표', '비대위원장'와 같은 정치적 의미를 갖는 어휘들이 나타남을 알 수 있다. 좌측은 출현빈도비율이 2 미만인 단어들로 일반 뉴스 댓글과 정치 뉴스 댓글에서의 출현 빈도가 크게 다르지 않은 단어이다. '미치광', '이자슥'처럼 정치적 의미가 비교적 적은 단어들로 나타난다. 본 연구에서는 정치적 의미를 갖는 어휘를 분석 대상으로 하기 위해 출현빈도비율이 2 이상인 15,726개의 신조 어휘들을 추출하였다.

표 1. 비율 값이 2 미만인 단어(좌), 비율 값이 2 이상인 단어(우)
Table1. Words with a proportion value under 2(left), Words with a proportion value more than 2(right)

어휘	비율	어휘	비율
로멘스	1.8	공수처법	26
남탓	1.78	비대위원장	15
미치광	1.71	금의원	13
이자슥	1.67	친일매국당	12
멍멍멍	1.67	홍영표	11.67
컴푸렉스	1	해골찬	9
글렀엌	1	특별재판부	8
띵호아	1	금뱃지	7
쏴라잇네	1	적폐보수	3

또한 일부 사용자에 의해서 발생한 오타와 같은 어휘를 신조어로 인식하지 않기 위해 정치 분야에 출현빈도가 10회 이상인 단어들만 분석 대상으로 추출하였다. 위 과정을 통해 총 1,511개의 단어가 추출되었다. 본 연구에서 제시한 방법론으로 추출한 최종 단어들 중, 상위 300개를 그림 4으로 시각화하였다. 그림 4의 글씨 크기는 정치 외 대비 정치 분야에 댓글에 사용된 비율 값과 비례한다. 특정 의원들을 속되게 이르는 '국쌍', '홍재앙' 등이 주로 추출되었으며, '김여정', '김두관'과 같이 정치인이거나 정치권과 밀접한 관련이 있는 인물들의 이름이 추출되었다.

그림 4. 추출된 정치 신조어 / 비속어
Figure 4. Extracted political newly-coined word

III. 정치 신조어 및 비속어 관계성 분석

뉴스 댓글 작성자의 정치적 입장과 인식을 효과적으로 파악하기 위해서는 추출한 신조어와 비속어의 의미를 파악하는 것이 필요하다. 이를 위해 추출한 1,511개의 단어에 대해 단어 관계성 분석을 진행하였다. 분석에는 동시출현 빈도 분석방법과 워드임베딩 기법을 사용하였다. 두 기법은 신조어와 비속어의 문장 내 동시 출현하는 빈도 및 분포를 통해 어휘들 사이의 의미적 관계성을 파악하는데 효과적이다.

1. 동시출현 단어 빈도분석

동시출현단어 빈도 분석은 텍스트의 코퍼스(corpus) 내에서 동시 출현하는 단어의 패턴을 분석하는 텍스트 마이닝 기법이다[14]. 동시 출현 빈도 분석은 말뭉치들의 추세나 주제를 파악할 수 있고, 출현한 단어들의 상관성을 분석할 수 있다는 장점으로 인하여 논문의 동향을 파악하는 메타 논문 연구에서 주로 사용되고 있다[15].

본 연구에서는 정치 뉴스 댓글에 사용된 정치 관련 신조어와 비속어들 사이의 의미적 연관 관계를 파악하기 위해, 정치 섹션 뉴스에 작성된 84,000개의 댓글을 대상으로 동시출현 분석을 수행하였다. 각 댓글을 하나의 단위로 설정하여 댓글 안에서 짝을 이뤄 등장하는 단어들의 빈도를 단어 문서 행렬(Term-Document Matrix)로 나타내었다. 본 연구에서는 정치 신조어 비속어의 의미를 파악하는 목적에 따라 분절화로부터 추출한 1,511개의 정치 신조어를 대상으로 행렬을 구성하였다.

표 2. 정치어 동시 출현 매트릭스 예시
Table 2. Example of political word co-occurrence matrix

	자한당	문빠	나베	공수처	개누리	이명박근혜
자한당	-	2	6	9	5	8
문빠	2	-	1	1	0	0
나베	6	1	-	1	0	0
공수처	9	1	1	-	0	0
개누리	5	0	0	0	-	1
이명박근혜	8	0	0	0	1	-

가령 댓글 (가)와 (나)라는 2개의 뉴스 댓글에서, 댓글 (가)에 A, B, C라는 어휘가 나타나고 댓글 (나)에 A, C, D라는 어휘가 나타난다고 가정하자. 단어 쌍 (A, C)의 경우 댓글 (가)에도 등장하고 댓글 (나)에도 같이 동시에 등장하므로 동시 출현 빈도는 2가 된다. 단어 쌍 (A, B)와 (A, D)의 경우 각각 댓글 (가), 댓글 (나)에서만 한번씩 출현하므로 동시 출현 빈도는 1로 계산된다. 이와 같은 방법으로 어휘 사이의 동시 출현 빈도는 표 2과 같이 표현될 수 있다.

정치 뉴스 댓글에서 어휘의 출현 비율이 높았던 '자한당', '공수처', '나경원'을 중심으로 동시출현 분석을 활용해 맥락을 분석해 보았다. 표 3는 이들 어휘와 동시 출현 빈도가 높은 상위 10개의 단어를 나타낸 것이다.

표 3. '자한당', '공수처', '나경원'에 대한 상위 10개 단어 쌍
Table 3. Top 10 word pairs for '자한당', '공수처', '나경원'

기준단어	동시출현빈도가 높은 상위10개 단어
자한당	나경원, 홍준표, 바미당, 김성태, 토착왜구, 공수처, 새누리당, 한국당, 헛소리, 이명박근혜
공수처	자한당, 검찰개혁, 나경원, 윤석열, 정치검찰, 국개의원, 패스트트랙, 검언유착, 한국당, 바미당
나경원	황교안, 자한당, 장제원, 김성태, 윤석열, 홍준표, 전희경, 이은재, 자유한국당 나베

자유한국당을 의미하는 '자한당'과 같이 자주 등장한 정치 단어는 자유한국당의 소속 의원이거나 소속 의원이었던 '나경원', '홍준표' 등이며 고위 공직자 범죄 수사처를 의미하는 '공수처'는 검찰 총장인 '윤석열'이나 '검찰개혁' 등과 함께 등장했다. 이를 통하여 특정 정치 신조어, 비속어가 어떠한 맥락에서 자주 활용되고 있는지 파악할 수 있다.

2. 워드임베딩을 이용한 유사단어 탐색

워드임베딩은 분산 단어 표현(distributed word expression)이라고도 하며, 단어를 의미(syntactic) 및 구문(sementic)관계를 표현하는 벡터로 변환하는 방법이다[15][16]. 벡터로 표현된 어휘들은 벡터 연산을 사용해 다양한 의미 관계를 분석할 수 있다. 본 논문에서는 워드임베딩의 방법 중 하나인 FastText 방법을 사용하여 단어를 벡터화 하였다. FastText는 하나의 단어를 n개의 문자(character)로 나누어 임베딩하는 방법으로 한국어와 같이 다양한 어미의 활용을 가지고 있

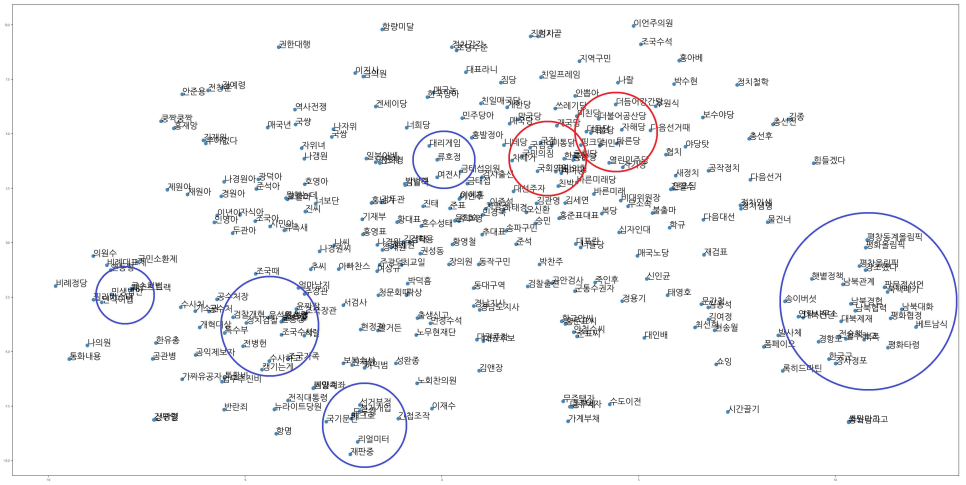

그림 5. t-SNE를 이용한 단어 벡터 시각화
Figure 5. Word vector visualization using t-SNE

는 언어에서 뛰어난 성능을 보이며, 학습하지 않은 OOV(Out of Vocabualary)단어 대해서도 단어를 n 개의 문자로 나누어 뜻을 유추할 수 있다는 장점이 있다[17]. 그림 5에서는 FastText를 이용하여 정치 외 단어 대비 정치 단어 비율값이 큰 상위 300개의 정치어를 벡터로 나타내었다. FastText의 학습 시 vector의 크기는 50, 학습 반복 횟수는 100으로 하여 학습하고, 2차원의 공간에 벡터를 나타내기 위하여 t-SNE를 이용하여 벡터를 2차원으로 축소하였다[18]

그림 5에서 같은 정당에 속한 인물들이나, 하나의 대상을 지칭하는 단어가 여러개인 경우 단어들이 비슷한 곳에 위치하였다. 예를 들어, 붉은 원으로 표시된 부분에서 '더불어민주당'을 의미하는 '더불당', '더민주당' 등의 단어가 비슷한 곳에 위치해 있고, '국민의 힘' 당을 의미하는 '국힘당', '핑크당' 등의 단어가 비슷한 좌표에 위치해 있다. 또한 동일한 정치적 사건과 관련된 단어들이 비슷한 곳에 위치했다. 그림5 에서 파란 원을 확인해 보면 대리게임과 류호정, 검찰개혁과 정치검찰, 햇볕정책과 대북문제 등, 정치적인 사건을 중심으로 연관 단어가 비슷한 곳에 위치한 것을 볼 수 있다.

표 4. Fasttext 를 이용하여 찾은 유사단어 표
Table 4 Similar words table using Fasttext

단어	유사단어1	유사단어2	유사단어3	유사단어4
드루킹	김경수	댓글조작	여론조작	매크로
나베	나경원	나씨	장제원	토착왜구당
공안검사	민주투사들	군사독재의	군사정권	유신독재
빨갱이	종북	종북좌파	간첩	공산당
황교활	황교안	김무성	교활이	김성태
국민의짐	국짐당	미통당	자한당	국민의힘
국개의원	국회의원	국개	비례대표	의원

표 4은 신조어와 비속어 중 일부를 선택하여 전체 단어 집합 내의 유사한 단어들을 나타낸 표이다. 표 4에서는 '드루킹'이라는 단어가 댓글조작, 여론조작과 관련된 단어라는 것을, '국민의 짐'은 국민의 힘 당을 다르게 부르는 말 중 하나라는 것을 .보여준다.

이와 같이, 벡터화 된 정치 신조어, 비속어와 가까운 거리의 단어를 찾는 것으로 신조어와 비속어의 뜻을 유추 할 수 있다.

Ⅳ. 결 론

본 연구에서는 대중의 정치적 인식과 의견을 효과적으로 분석하기 위해 정치적 의미를 지니는 신조어와 비속어를 추출하고 의미를 파악하는 방법을 제안하였다.

정치 댓글 내에 있는 신조어와 비속어를 추출하기 위해 사전기반의 형태소 분석기가 아닌 띄어쓰기를 기반으로 문장을 분절하였으며, 이를 통해 확보한 정치 단어 집합에서 정치 외 단어 집합을 제거하는 방법으로 정치 신조어와 비속어를 추출하였다. 또한 추출 된 단어들의 의미를 파악하기 위하여 단어 간의 동시 출현 빈도분석과 워드 임베딩을 사용하여 단어들 간의 관계성을 파악하였다.

이는 신조어와 비속어로 인해 소그룹화 되는 인터넷 커뮤니케이션을 연구하는 것에 기여 할 수 있을 것으로 보이며, 정치 분야의 뿐만 아니라 다른 도메인의 텍스트에서 신조어와 비속어를 추출하기 위한 방법으로도 적용가능할 것으로 기대한다.

References

[1] H.J. Jung, J.H. Bae, S.L. Hong, C.U. Park, M. Song, “Analysis of Twitter Public Opinion in Different Political Views : A Case Study of Sewol Ferry Accident”, Korean Society For Journalism And Communication Studies, Vol. 60, No. 2, pp. 269-302, 2016.

[2] E.H. An, J.K. An. “An Analysis of the 2017 Korean Presidential Election Using Text Mining”, Vol. 11. No. 5, pp. 199-207. 2020. DOI: https://doi.org/10.15207/JKCS.2020.11.5.199

[3] J.Y. Han, Y.I. Lee., J.B. Lee, M..Y. Cha. “The fallacy of echo chambers: Analyzing the political slants of user-generated news comments in Korean media”, In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019),pp. 370-374. 2019. DOI: https://doi.org/10.18653/v1/D19-5548

[4] H. Kang, D.K. Kang, “Long Short Term Memory based Political Polarity Analysis in Cyber Public Sphere”, International Journal of Advanced Culture Technology, Vol. 5, No.4, 2017. DOI: https://doi.org /10.17703/IJACT.2017.5.4.57 DOI:https://doi.org /10.17703/IJACT.2017.5.4.57

[5] Y.I. Lee, J.Y. Han, M.Y. Cha, “Building a Political Bias Classifier for News Comments using User Labeling”, The Korean Institute of Information Scientists and Engineers, pp. 1643-1645, 2020.

[6] H.B. Choi, J.H. Kim, J.H. Lee, M.G. Lee. “Political Information Filtering on Online News Comment”, The Journal of the Convergence on Culture Technology, Vol. 6, No. 4, pp. 575-582, 2020. DOI: https://doi.org/10.17703/JCCT.2020.6.4.575

[7] J.W. Kim, J.W Jeong, M.Y Cha, Automatic New Korean Words Extraction Using Portal News Headlines, The HCI Society of Korea, ,pp. 163-166, 2020

[8] K.D. Hyun, N.W. Jung, M.H. Seo, “Examining the Effects of Perceived Partisan Slants of News and User Comments from Portal News Sites on Portal News Trust, Third Person Perception and Selective Exposure : Comparisons of Conservative and Progressive Users”, Korean Society For Journalism And Communication Studies, Vol. 64, No. 4, pp. 247-288, 2020. DOI: https://doi.org/10.20879/kjjcs.2020.64.4.007

[9] https://pypi.org/project/beautifulsoup4/

[10] https://github.com/shineware/KOMORAN

[11] J.P. Hong, J.W. Cha, “A New Korean Morphological Analyzer using Eojeol Pattern Dictionary”, The Korean Institute of Information Scientists and Engineers, Vol. 35, pp. 279-284, 2008.

[12] “Korean Lemmatizer”, https://github.com/lovit/korean_lemmatizer

[13] National Istitute of Korean Language, “전자사전 전체파일”, https://ithub.korean.go.kr/user/total/database/electronicDicManager.do, 2017.

[14] Q. HE, “Knowledge Discovery Through Co-Word Analysis”, LIBRARY TRENDS, 1999.

[15] H.J. Kim, M. Song, “A Study on the Research Trends in Domestic/International Information Science Articles by Co-word Analysis”, Journal of the Korean society for information management, Vol. .31, No. 91, pp. 99 - 118, 2014. DOI: http://dx.doi.org/10.3743/KOSIM.2014.31.1.099

[16] S. Lai, L. Kang, L. Xu, J. Zhao, “How to generate a good word embedding”, IEEE Intelligent Systems, Vol. 31, No. 6, pp. 5-14. 2016.

[17] T. Mikolov, W.T. Yih, G. Zweig. “Linguistic regularities in continuous space word representations”, In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 746-751, 2013.

[18] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, “Enriching word vectors with subword information.” Transactions of the Association for Computational Linguistics", Vol.5, pp. 135-146. 2017. DOI: https://doi.org/10.1162/tacl_a_00051

[19] L. Van der Maaten, G. Hinton. “Visualizing data using t-SNE. Journal of machine learning research”, Journal of Machine Learning Research, Vol. 9, No. 11. 2008.