JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 12, pp. 2077-2087, Dec. 2021

Check for updates

일상적인 한국어 문장의 정치적 편향을 표현하는 키워드 추출에 대한 연구

강 형 석[1] · 양 장 훈[2*]
[1]서울미디어대학원대학교 인공지능 응용소프트웨어학과 연구원
[2*]서울미디어대학원대학교 인공지능 응용소프트웨어학과 부교

Study on the Methods of Keywords that Represent Political Bias in Ordinary Korean Sentences

Hyungsuc Kang[1] · Janghoon Yang[2*]

[1]Researcher, Department of AI Software Engineering, Seoul Media Institute of Technology, Seoul 05790, Korea
[2*]Associate Professor, Department of AI Software Engineering, Seoul Media Institute of Technology, Seoul 07590, Korea

[요 약]

본 연구에서는 일상생활에서 사용하는 단어에 담긴 숨겨져 있는 정치적 편향을 표현하는 키워드를 추출하는 방법을 제안한다. 제안 방법은 [10]의 정치 성향 추정 모델에 의해서 문장의 정치적 성향을 구분하고 동일한 정치적 성향으로 판단된 문장들로부터 키워드를 추출하는 방식으로, 형태소의 제거에 의해서 정치적 성향의 분류가 달라지는 키워드를 후보 키워드로 선정하고 형태소의 분포를 기반으로 키워드를 추출하는 방식으로 설계되었다. 제안 방식을 일상적인 대화로 구성되는 시나리오 데이터 셋에 대해서 4가지의 정치적 성향의 키워드를 추출한 결과 제안 방식은 기존 방식인 keyBERT와 TF-IDF를 적용했을 때와 일부 유사한 개념들이 추출되는 것을 확인할 수 있었다. 또한, 기존의 두 방식이 유사한 개념의 단어들을 주로 추출하는 것에 비해서, 제안 방식은 좀 더 다양한 의미의 키워드를 도출하고 있는 것도 관찰되었다.

[Abstract]

This study attempts to extract keywords that represent political bias in ordinary Korean sentences. The proposed method builds on the political-bias prediction model in [10], which classifies the political orientation of the document. The proposed methods consist of collecting sentences with political orientation, selecting candidate keywords which change the class of the sentence with the removal of that morpheme, and determining keywords from the distribution of the morpheme. To verify the efficacy of the proposed methods, the keywords from a movie scenario dataset were extracted and compared with the existing methods, keyBERT and TF-IDF. The analysis of the extracted keyword shows that all methods have some common keywords while the proposed methods extract keywords with more diverse meanings.

색인어 : 문서 분류, 키워드 추출, 정치적 편향, KeyBERT, TF-IDF

Keyword : Document classification, Keyword extraction, Political orientation, KeyBERT, TF-IDF

http://dx.doi.org/10.9728/dcs.2021.22.12.2077

Received 30 October 2021; **Revised** 23 November 2021
Accepted 10 December 2021

***Corresponding Author; Janghoon Yang**

Tel: +82-2-6393-3237
E-mail: jhyang@smit.ac.kr

Ⅰ. 서 론

인간의 가치관은 언어와 행동을 통하여 직간접적으로 표현이 된다. 행동 표현을 통하여 가치관을 추정하는 방법은 크게 비전기반, 웨어러블 센서 기반, 및 언어 기반으로 분류할 수 있는데, 가장 손쉽고 많은 데이터를 수집하는 방법이 언어 기반이다. 이렇듯 인간의 언어에는 개인이나 집단의 가치와 문화를 담고 있으며, 최근에 빠르게 발달하고 있는 인공지능 기술과 자연어 처리 기술을 활용하여 다양한 언어 데이터로부터 중요한 정보를 추출할 수 있는 가능성을 내포하고 있다.

인간의 가치관은 일상생활 중에 직간접적으로 표현된다. 국가라는 틀 안에서 존재하는 다양한 정치적 상황은 개인의 가치관에 영향을 주고, 이는 다양하게 일상생활에서 표현된다. 정치적 편향은 몇 가지로 대표적으로 분류될 수 있고, 정치적 관점에 따라서, 정치 집단과 각 정치 집단과 개인의 정치적 편향간의 심리적 유대 관계가 결정된다. 이러한 정치적 편향의 특징을 언어 분석을 통해서 도출한다면 보다 정치적 편향의 의미와 영향을 좀 더 구체적으로 분석할 수 있는 기회를 줄 뿐만 아니라, 일반적으로 인식하지 못했던 새로운 사물이나 행동에 대해서 정치적인 편향 해석할 수 있는 기회를 줄 수 있다. 또한, 정치적 편향을 담고 있는 단어들이 우리의 일상생활에서 어떻게 작용하고 있는지를 알기 위해서는 단순히 정치적인 글을 분석하는 것 보다는 일상 대화를 분석하고 일상 대화 속에서의 정치적 성향의 키워드를 뽑는 것이 보다 다양한 키워드를 뽑아낼 수 있는 기회를 제공할 것으로 예측된다. 따라서, 이 연구에서는 기존의 키워드 추출방법과, 글의 정치적 성향을 분석하는 방법을 결합하여, 일상 대화에서 존재하는 정치적 편향의 키워드를 추출하고자 한다.

자연어 처리의 주요 응용 분야 중에 대표적인 분야인 키워드 추출은 다양한 글로부터 중요한 정보를 추출하는 도구로서 활용될 수 있는 잠재력을 보유하고 있다. 키워드 추출(keyword extraction)은 정보 검색, 문서 분류, 문서 요약 등을 포함하는 텍스트 마이닝(text mining) 분야에서 흔히 사용되는 기술이다. 키워드 추출 방식에는 Total Keyword Frequency, TF-IDF(Term Frequency - Inverse Document Frequency), RAKE(Rapid Automatic Keyword Extraction), KP(Keyphrase)-Miner, YAKE(Yet Another Keyword Extractor), KeyBERT 등의 다양한 방식이 존재한다[1]. 문서 내부의 단어에 대한 중요도를 평가하기 위한 TF-IDF 가중치 모델이 전통적으로 많이 사용되었다[2, 3, 4]. 최신 단어 임베딩 기술인 BERT 임베딩을 활용하는 keyBERT[5]도 최근에 제안되었다. TF-IDF를 이용하는 기존의 연구[6, 7]는 주로 뉴스 기사에서 키워드를 추출했다. KeyBERT를 이용하는 최근의 연구는 특허 문서[8]와 정책 문서[9]에서 키워드를 추출했다.

저자들의 이전 연구[10]에서 한국어 신문 기사와 사설의 정치적 편향을 분석한 바 있다. 본 연구는 해당 연구를 기초로 해서, 일상적인 한국어 문장의 정치적 편향을 분석한다. 이를 위해, 정치적 편향을 표현하는 키워드를 추출한 후, 이런 키워드의 특성을 고찰하고자 한다. 문장의 정치적 편향을 판단하는 확률을 근거로, 해당 문장의 키워드를 추출하는 방식을 사용할 것이다. 즉, 특정 정치적 편향으로 판단된 문장에서 한 개의 키워드씩 제거하면서, 결과적인 문장의 정치적 편향이 변하는 경우 해당 키워드를 추출할 것이다. 이렇게 추출된 정치적 편향을 나타내는 키워드를 기존의 키워드 추출 방식(Key BERT 및 TF-IDF)으로 추출된 키워드와 비교할 것이다. 이를 통해, 키워드 추출 방식을 통해서 발견한 정치적 편향을 표현하는 단어들의 특징을 고찰하고 제안 방법과 기존 방법의 한계와 이 분야의 연구를 심화시킬 수 있는 연구 방향에 대해서 제시하고 한다.

본 논문의 나머지는 다음과 같이 구성된다. 제II장에서는 관련된 이전 연구를 간단히 소개한다. 제III장은 본 논문의 실험 데이터와 방법을 소개하며, 제IV장은 실험 결과를 분석한다. 마지막으로 제V장에서는 결론과 향후 과제가 제시된다.

Ⅱ. 관련 연구

2-1 Doc2vec 모델을 이용한 신문의 정치적 편향 분석

저자들의 이전 연구[10]에서 신문의 정치적 편향을 분석한 바 있다. 해당 연구의 기본적인 절차는 그림 1과 같다. 우선 특정 정당(2018년 당시 더불어민주당, 자유한국당, 바른미래당, 그리고 정의당)의 Facebook 공식 페이지에 “좋아요”를 표시한 Facebook 사용자의 모든 게시글(2018년 기준)을 학습 말뭉치롤 수집했다. 그리고 해당 정당의 Facebook 공식 페이지에 “좋아요”를 표시하지 않은 일반 Facebook 사용자의 게시글도 비교를 위해 수집했다. 문서 분류 기법 중 하나인 doc2vec 모델[11]을 이용해서 해당 Facebook 말뭉치를 학습함으로써, 각 게시글의 정치적 편향(PB: Political Bias)을 벡터로 표현하는 모델을 개발했다. 위의 4개 주요 정당(더불어민주당, 자유한국당, 바른미래당, 그리고 정의당)을 지지하는 정치적 편향은 각각 DPK, LKP, BP, 그리고 JP라는 라벨로 표시되었고, 특정한 정치적 편향을 보이지 않는 경우는 UP(UnPoltical)라는 라벨로 표시되었다. 수집된 각 게시글의 벡터는 로지스틱 회귀(logistic regression) 모델을 통해 분류된다.

그런 다음, 주요 신문사의 기사와 사설을 수집한다. 수집된 각 기사/사설을 위에서 학습한 doc2vec 모델을 통해 벡터로 변환한 후, 위에서 학습한 로지스틱 회귀 모델을 통해 해당 기사/사설의 정치적 편향이 결정된다. 본 연구에서는 [10]에서 사용된 doc2vec 모델을 일상적인 한국어 문장에 적용함으로써, 특정 정치적 편향을 표현하는 키워드를 추출하고자 한다.

2-2 Mecab-ko 태거

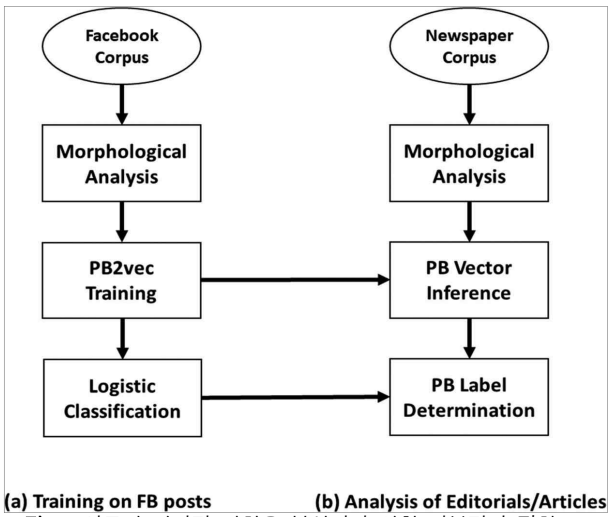


그림 1. 신문의 정치적 편향을 분석하기 위한 기본적인 절차
Fig. 1. Basic procedure for analyzing political bias of Newspapers

NLP(Natural Language Processing)의 전처리 과정으로 토큰화(tokenization)가 필수적이다. 일반적으로 영어 NLP에서는 띄어쓰기 단위로 토큰화가 처리된다. 하지만 일반적으로 한국어 문장은 조사/어미를 이전의 체언/어간에 붙여 쓰므로, 한국어 NLP에서는 형태소 분석기(morphological analyzer)/품사 태거(POS tagger)를 이용해서 토큰화하는 것이 바람직하다.

KoNLPy[12]는 한국어 NLP를 위한 파이썬 패키지이다. 해당 패키지에서는 몇 가지 형태소 분석기/품사 태거가 제공되는데, 이 중 mecab-ko 태거가 제일 성능이 우수하다[13]. 이전 연구[10]에서도 형태소 분석기/품사 태거로 mecab-ko 태거가 사용되었기 때문에, 본 연구에서도 mecab-ko 태거가 사용될 것이다. Mecab-ko 태거에서 사용되는 품사에 대한 약어(e.g. NNG = 일반 명사)는 구글 공유 문서[14]에 요약되어 있다.

2-3 KeyBERT

구글에서 공개한 인공지능 언어모델 BERT[15]는 최근 각광받고 있는 임베딩(embedding) 모델이다. 이 모델은 일부 성능 평가에서 인간보다 더 높은 정확도를 보이며, 범용 목적의 언어 이해(language understanding) 모델을 훈련시켜서, 다양한 NLP 태스크(e.g. 질문·응답 등)에 적용될 수 있다.

[16]는 BERT 임베딩을 활용해서, 특정 문서의 키워드를 간단히 추출할 수 있는 파이썬 패키지를 제공한다. 본 연구에서 추출된 키워드와 KeyBERT를 이용해서 추출된 키워드를 비교할 것이다.

2-4 TF-IDF

TF-IDF(Term Frequency - Inverse Document Frequency)는 여러 문서로 이루어진 문서군에 대해, 어떤 단어가 특정 문서에서 얼마나 중요한지를 나타내는 통계적 수치이다. 단어 빈도 TF는 특정 단어가 특정 문서에서 얼마나 자주 등장하는지를 나타내는 값이고, 역문서 빈도 IDF는 특정 단어가 발견되는 문서의 수에 대한 역수이다. TF-IDF는 TF와 IDF를 곱한 값이다.

단어 빈도 TF와 역문서 빈도 IDF를 구하는 방식은 여러 가지가 있지만, 본 연구에서는 아래와 같이 [17]에서 제시한 방식을 선택한다.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}} \quad (1)$$

단, $f(t,d)$ = 문서 d에서 단어 t가 언급된 총 빈도

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

단, D = 전체 문서의 집합, $|D|$ = 전체 문서의 수

Ⅲ. 실험 데이터 및 방법

3-1 일상적인 한국어 문장의 수집

우선 일상적인 한국어 문장 말뭉치를 구성하기 위해, 필름메이커스커뮤니티(www.filmmakers.co.kr)에 공개된 약 200개의 한국영화 시나리오가 수집되었다. 시나리오는 일상 대화를 연구하는데 있어서, 드라마 시나리오 말뭉치를 사용하여 일상 회화에서 수동문을 연구[18]하거나, 예고 표현이 일상 대화에서 주는 영향과 특징을 연구[19]하는 등 다양한 분야에서 분석 데이터로 사용되었다. 따라서, 본 연구에서도 신문 기사 같은 글과는 달리, 시나리오의 지문과 대사는 가장 일상적인 한국어 문장이라고 판단되기 때문에 일상적인 한국어 문장의 말뭉치로 사용하였다. 각 시나리오에서 5글자 이하의 문장은 제외하고, 마침표/물음표/느낌표 단위로 문장이 추출되었다.

3-2 한국어 문장의 정치적 편향에 대한 판별 및 문서 생성

이전 연구[10]에서 학습된 doc2vec 모델을 적용해서, 수집된 일상적인 한국어 문장의 문장 벡터를 구한다. 그런 다음, 이전 연구[10]에서 학습된 로지스틱 회귀 모델을 적용해서, 각 문장 벡터의 정치적 편향(즉, BM, DPK, JP, LKP, 그리고 UP)을 판단한다. 정치적 편향이 총 5가지이므로, 각 정치적 편향으로 판단될 확률의 합이 100%가 된다. 가장 높은 확률을 갖는 레이블에 대해서 해당 문서의 정치적 편향으로 판단한다. 이 모델을 문장별로 적용하여 UP를 제외한 각 정치적 편향(즉, BM, DPK, JP, LKP)으로 판단된 문장을 각 정치적 편향별 문서로 통합한다.

이 연구에서 사용된 정치 성향 분류기 [10]의 특징을 정성적으로 분석하기 위해서 Aihub (www.aihub.co.kr)에서 제공하는 KETI 감정 분류용 데이터 셋에 적용하여 문장들의 정치적 성향을 분류하였다. 각 정치 성향별로 가장 높은 확률로 뽑힌 문장과 높은 확률로 분류되는 문장들의 특징은 다음과 같이 정리될 수 있다. UP 에서 가장 높은 확률로 뽑힌 문장은 "다음주부터 사각턱이랑 광대 눈 상담받으러 상담투어 갑니다 ㅋㅋ "이고, 주로 외모나 남녀 관계에 관한 문장들이 높은 확률로 UP문장으로 분류되었다. MJ의 경우에는 "김앤장을세무감사 해보소그네는 죽어도 못 할걸"이고, 주로 민주당이나 민주당 관련 인사의 이름이 들어가 있는 문장들이 높은 확률로 MJ문장으로 분류되었다. LK에서 가장 높은 확률로 뽑힌 문장은 "솔직히 돼정은 멍멍이자식과 그 추종자 및 국내 종북좌빨, 북한으로 이 득취하려는 극우만 빼면 통일 못할것도 없지 뭐...."이고, 주로 북한이나 좌파를 부정적으로 표현하는 문장들이 높은 확률로 LK문장으로 분류되었다. BM의 경우에는 "소신 지키다 원내대표서 쫓겨나고 공천 못받고 그래도 옳은 소리하고 유승민 다시봤다. 무성이는 너무 계산기 두들긴다."이고, 주로 관련 정치인들이 들어간 문장이나 교육이나 기술과 관련된 문장들이 높은 확률로 BM문장으로 분류되었다. JP에서 가장 높은 확률로 뽑힌 문장은 " ㅋㅋㅋㅋ ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 닥그네아줌마 말하면서 수심이 그득하네 ㅋㅋㅋㅋ 간만에 보네 ㅋ"이고, 관련 정치인들이 포함된 문장이나 민족에 관련된 문장들이 높은 확률로 JP문장으로 분류되었다. 각 정치 성향별로 뽑힌 문장들은 일반적으로 받아들여지는 정치적인 색깔들을 대부분 표현하고 있는 것으로 판단되어, 문장의 정치적 성향을 분류하는데 적절하게 동작함을 정성적으로 확인할 수 있다.

3-3 키워드 후보 추출

형태소 단위 키워드 후보 추출을 위해서 다음과 같이 정치적 편향의 변화를 주는 형태소를 키워드 후보로 추출하는 방법을 제안한다. 각 정치적 편향으로 판단된 문장에서 형태소를 하나씩을 제거한 후 동일한 doc2vec 모델과 로지스틱 회귀 모델을 적용해서 정치적 편향을 다시 판단한다. 해당 정치적 편향으로 판단될 확률(형태소가 제거되기 전에는 원래 20% 이상)이 20% 미만으로 떨어지는 경우 해당 형태소를 결정적인 키워드로 간주한다. 즉, 특정 형태소가 제거되었을 때, 해당 문장의 정치적 편향 판단 확률이 20% 미만으로 떨어지는 경우, 해당 형태소는 정치적 편향을 표현하는 후보 키워드로 간주하도록 한다. 그리고 전체 문서에 존재하는 의미를 가지는 실질 형태소를 추출하기 위해서 문서에서 자주 등장하는 조사, 어미, 관형사, 대명사, 및 문장부호는 제외한다.

본 연구의 키워드 추출 방식을 수식으로 표현하면 다음과 같다. n개의 형태소로 이루어진 문장 S에 대한 키워드 k는 다음과 같이 구해진다.

$$S = \{morph_m | m = 1, 2, \dots n\} \tag{3}$$

$$S_i = \{morph_m | m = 1, 2, \dots n \text{ and } m \neq i\} \tag{4}$$

$$C = \{morph_i | P(S_i) < 0.2 \text{ for } i = 1, 2, \dots n\} \tag{5}$$

위 식에서 $P(S_i)$는 S_i에 대한 정치적 편향 확률을 의미하고, $morph_i$는 문장 S에서 m번째 형태소이다. 또한, C는 형태소의 제거에 의해서 문장의 정치적 편향이 바뀌는 형태소들의 모임이다.

3-4 제안 키워드 추출 방법

3-3에서 제안된 방식을 통하여 추출된 후보 키워드는 다양한 방법과 결합되어 정치적 편향을 나타내는 키워드를 추출하는데 사용될 수 있다. 본 논문에서는 TF-IDF에서의 개념을 확장한 세가지 방식의 키워드 추출을 제안한다.

첫 번째 방식은 먼저 [10]에서의 방법을 이용하여 특정 정치적 편향으로 판단된 문장의 집합을 P로 정의하고, 문장에 대해서 수식 (5)에 의해서 추출된 키워드 후보 집합을 C_P로 정의한다. C_P에 속하는 i번째 원소 c_i에 대해서 P속하는 문장 중 c_i를 포함하고 있는 문장의 수를 s_i, 이 문장들 중에서 c_i를 C에 포함하게 되는 문장의 수를 f_i라고 표기할 때에 f_i를 내림차순으로 정렬하여 상위 N개의 형태소를 추출하는 방법이다. 이 방법을 본 논문에서는 “분류 변경 빈도수 방법”이라고 부른다.

첫 번째 방식에서는 특정 형태소가 얼마나 자주 발생하는지는 고려하지 않고, 문장의 정치적 분류를 얼마나 자주 변화시키는지를 고려한다. 하지만, P에서 자주 발생하는 단어는 TF-IDF에서와 유사하게 중요성을 가질 수 있기 때문에 이를 고려하여, $c_i * f_i$ 를 내림차순으로 정렬하여 상위 N개의 형태소를 추출하는 방법을 정의하고, 이 방법을 본 논문에서는 “발생 빈도수/분류 변경 빈도수 방법”이라고 부른다.

마지막 방식은 3-3에서 제안된 방식을 통하여 추출된 후보 키워드내에 대해서 TF-IDF를 적용하여 상위 N개의 키워드를 추출하는 방식이다. 이 경우에 TF는 각 정당별 P에 대해서 구하지만, IDF를 구할 때에 문서의 범위를 어떻게 설정하느냐에 따라서 IDF의 값이 크게 바뀔 수 있다. 각 정당에 속하는 문장의 집합인 P를 하나의 문서로 고려할 때에 문서의 개수는 4개 밖에 되지 않아서, 많은 단어들이 다른 문서에서 중복적으로 발생하면서, IDF의 효과가 제한적으로 발생될 수 있다. 반대로, 현재 TF-IDF를 구하기 위한 P에 속하는 문장이외의 모든 문장들을 하나의 문서로 취급한다면 문서의 개수가 너무 많아지면서 IDF값이 과도하게 TF-IDF를 결정하는 요인이 된다. 따라서, 이를 해결하기 위해, 각 정당에 해당하는 문장들을 M개씩 묶어서 하나의 소 문서를 만들고 이를 기반으로 IDF를 계산하는 방법을 사용한다. 이 방법을 본 논문에서는 “후보 키워드 TF-IDF”라고 부른다.

Ⅳ. 실험 결과

4-1 수집된 문장의 정치적 편향 분석

약 200개의 한국영화 시나리오로부터 총 472,347개의 문장이 수집되었다. 이전 연구[10]의 doc2vec 모델과 로지스틱 회귀 모델을 적용해서, 각 문장의 정치적 편향을 판단한 결과는 그림 2와 같다. 충분히 예상할 수 있듯이, 대부분의 일상적인 한국어 문장(91.98%)은 UP(비정치적)인 것으로 판단되었다.

그림 3과 표 1은 [10]의 로지스틱 회귀 방법이 추정하는 각 정치적 편향으로 판단된 확률의 분포 및 그 통계를 표시한다. UP로 판단된 문장은 평균적으로 약 0.387의 확률로 해당 정치적 편향으로 판단된다. 그리고 LKP(자유한국당)로 판단된 문장은 평균적으로 약 0.32의 확률로, 나머지 정치적 편향으로 판단된 문장은 평균적으로 0.28미만의 확률로 해당 정치적 편향으로 판단된다. 즉, UP로 판단된 문장은 나머지 정치적 편향으로 판단된 문장보다 더 높은 확률로 UP로 판단된다. 그림 3과 표 1이 시사하는 바는 특정 문장이 BM(바른미래당)/DPK(더불어민주당)/JP(정의당)/LKP(자유한국당)로 판단되더라도, 그 확률은 0.32 이하로 그리 높지 않다는 점이다.

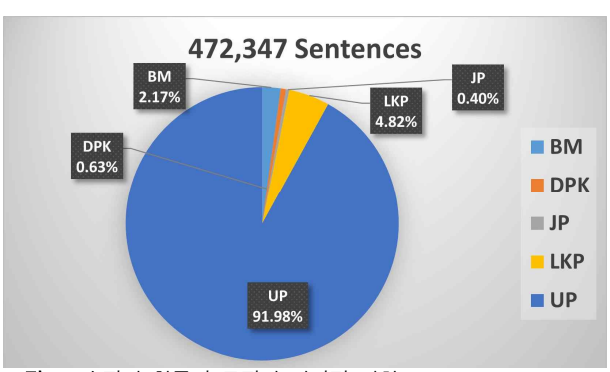


그림 2. 수집된 한국어 문장의 정치적 편향
Fig. 2. Political orientations of the collected Korean sentences

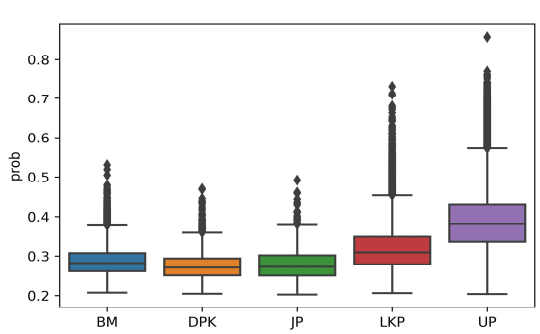


그림 3. 각 정치적 편향(PO)으로 판단된 확률의 분포
Fig. 3. Distribution of probabilities judged as each PO(Political Orientation)

표 1. 각 정치적 편향(PO)으로 판단된 확률의 통계
Table 1. Statistics of probabilities judged as each PO(Political Orientation)

Political Orient. (PO)	No. of Sentences	Probabilities			
		max.	min.	average	std.
BM	10,245	0.531	0.207	0.276	0.035
DPK	2,996	0.471	0.204	0.276	0.035
JP	1,866	0.492	0.202	0.279	0.038
LKP	22,784	0.731	0.206	0.320	0.058
UP	434,456	0.857	0.203	0.387	0.071

4-2 정치적 편향을 표현하는 후보 키워드

1) 각 정치적 편향으로 판단된 문장의 후보 키워드 추출

3-3섹션에서 기술한 방법에 따라서, UP를 제외한 정치적 편향(BM, JP, LKP, DPK)으로 판단된 문장에서 한 형태소씩 제거한 후 정치적 편향을 판단하는 확률을 0.2 미만으로 떨어뜨리는 해당 형태소(후보 키워드)를 추출했다. 추출된 후보 키워드의 통계는 표 2 및 그림 4와 같다. 그림 4에서 확인할 수 있듯이, 일부 키워드는 2가지 이상의 정치적 편향으로 판단된 문장에서 추출되었다. 이런 중복된 키워드를 제외하면, 각 정치적 편향(BM/DPK/JP/LKP)으로 판단된 문장에서 추출된 키워드의 개수는 각각 577/540/736/1,644개이다. 본 연구에서는 이런 중복되지 않은 키워드만을 분석할 것이다.

표 2. 각 정치적 편향(PO)으로 판단된 문장에서 추출된 후보 키워드(KW)의 개수
Table 2. Number of candidate keywords(KW) extracted from sentences judged as each PO(Political Orientation)

Political Orient. (PO)	BM	DPK	JP	LKP	Total
No. of KW (all)	822	746	968	1,957	3,955
No. of KW (exclusive)	**577**	**540**	**736**	**1,644**	**3,497**

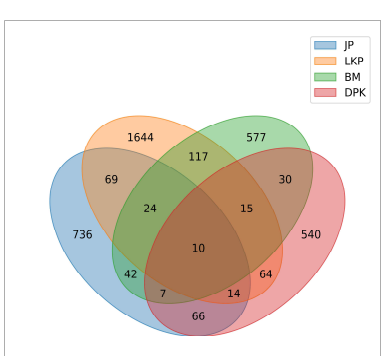


그림 4. 각 정치적 편향(PO)으로 판단된 문장에서 추출된 후보 키워드(KW)의 개수
Table 4. Number of candidate keywords(KW) extracted from sentences judged as each PO(Political Orientation)

2) 추출된 후보 키워드의 통계

그림 5는 각 정치적 편향(BM/DPK/JP/LKP)으로 판단된 577/540/736/1,644개 키워드의 f_i/s_i의 분포를 퍼센트로 나타내고 있다. 이 그림에서 보수정당의 지지성향(BM, LKP)에 비해 진보정당의 지지성향(DPK, JP)을 표현하는 키워드가 정치적 편향을 결정하는 데 더 큰 역할을 할 수 있음을 확인할 수 있다. 즉, 진보정당의 지지성향을 표현하는 키워드는 해당 키워드가 포함된 문장에서 더 높은 확률로 결정적인 역할을 한다고 볼 수 있다.

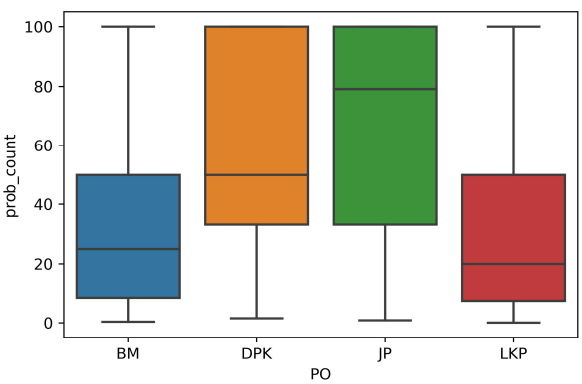


그림 5. 각 정치적 편향을 결정할 때 후보 키워드가 핵심 역할을 할 비중 (퍼센트)

Fig. 5. Percentage that a keyword plays a key role in determining each PO(Political Orientation)

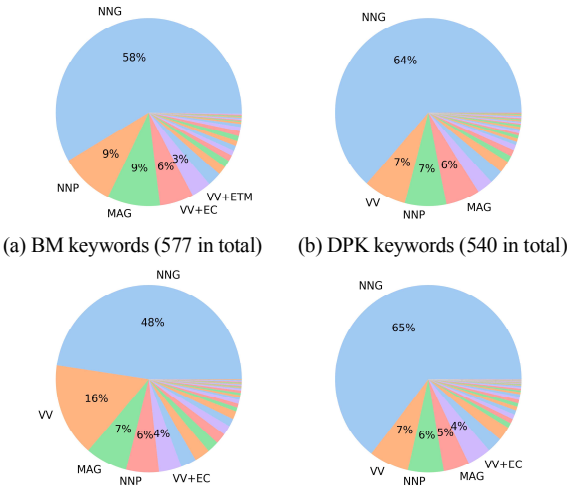
그림 6은 각 정치적 편향을 표현하는 총 3,497개 키워드에 대한 품사의 분포를 보여준다. 추출된 키워드는 각 정치적 편향별로 약간의 차이가 있긴 하지만, 전반적으로 일반명사(NNG), 고유명사(NNP), 일반부사(MAG) 및 동사 어간(VV)이 75% 이상을 차지한다. 특히 명사가 정치적 편향을 표현하는 주요 품사라고 볼 수 있는데, 이는 당연한 결과라고 할 수 있다.

3) 추출된 후보 키워드의 벡터공간상 분포

그림 7은 추출된 총 3,497개의 키워드 중 각 정치적 편향별로 상위 500개(기하평균 기준) 키워드의 단어 벡터를 t-SNE를 이용해서 2차원 공간에 표시한 그림이다. 이해의 편의를 위해, 각 정치적 편향별(BM/DPK/JP/LKP) 분포도 함께 표시했다. 그림 7이 보여주는 것처럼, 각 정치적 편향을 표현하는 키워드는 전체적으로 군집화되지는 않고, 일부 국지적으로 군집화된다고 할 수 있다. 즉, 추출된 키워드의 단어 벡터가 각 정치적 편향별로 명시적으로 군집화되지 않고 대부분 혼재되어 있음을 확인할 수 있다.

4-3 제안 방법으로 추출된 키워드

표 3은 분류 변경 빈도수 방법으로 후보 키워드로부터 추출된 상위 10개의 키워드를 보여주고 있다.



(a) BM keywords (577 in total) (b) DPK keywords (540 in total)

(c) JP keywords (736 in total) (d) LKP keywords (1,644 in total)

그림 6. 각 정치적 편향을 표현하는 키워드의 품사

Fig. 6. POS(Parts of Speech) of keywords that represent each PO(Political Orientation)

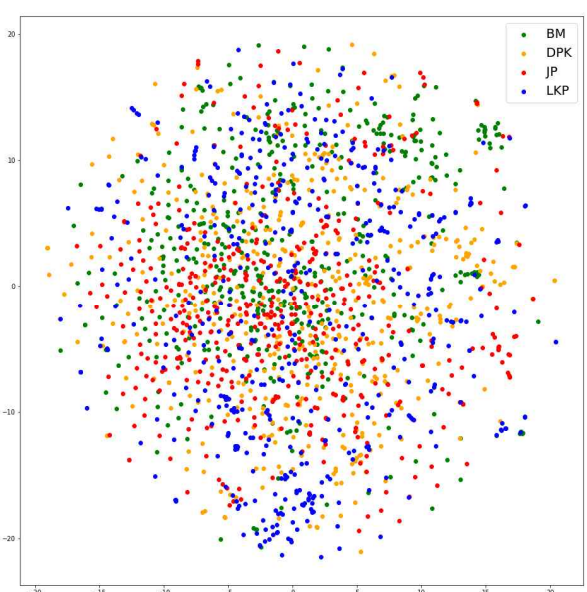


그림 7. 각 정치적 편향별 상위 500개 키워드의 단어 벡터에 대한 공간적 분포

Fig. 7. Spatial distribution of word vectors of top 500 keywords for each PO(Political Orientation)

우선 표 3의 키워드 중에 "*"으로 표시된 부분은 해당 정당명(미래/NNG, 바른/VA+ETM, 민주/NNG, 더불/VV)에 해당하거나, 특정 정치적 편향이 드러나는 경우(노동자/NNG, 피켓/NNG, 좌/NNG)이다. 문장의 정치적 성향을 분석하는데 사용한 [10]의 추정 방법은 Facebook의 포스트 데이터를 기반으로 학습되었기 때문에, 당연히 특정 정치적 편향을 가진 Facebook 게시글은 해당 키워드를 많이 언급하여, 이런 키워드가 추출되었을 것으로 추정된다.

그리고 "^"으로 표시된 부분은 특정한 인명에 해당하는 경우이다. 인명은 고유명사 NNP로 태깅되어야 하지만, 형태소 분석기 자체의 한계로 일반명사 NNG로 태깅된 경우기 자주 발생한다. 특히 LKP의 "정은/NNP" 또는 DPK의 "전두환/NNP"처럼, 특정한 인명이 Facebook 게시글에서 많이 언급되었을 가능성이 있다. 그런 경우, 해당 인명은 특정 정치적 편향을 판단하는 것으로 사용된 모델에서 학습되었을 수 있다.

나머지 단어들은 분류의 변화를 가져오는 문장 자체를 검토한 결과 특별한 시사점을 발견할 수가 없었다. 단지, BM에서의 "기숙사", "공부", 및 "재수"는 유사한 맥락에서 BM의 특징을 나타내는 키워드로 추정할 수 있다.

표 3. 분류 변경 빈도수에 의해 추출된 상위 10개 키워드
Table 3. Examples of top 10 keywords extracted from the class change frequency
(Note: The keywords in the table are written in Korean because it is inappropriate to translate them into English, which can lead to the loss of their real meanings.)

keyword	PO	f_i	keyword	PO	f_i
치성/NNG^	BM	45	진호/NNG^	LKP	118
미래/NNG*		22	정은/NNP^		94
재수/NNG		20	회장/NNG		50
바른/VA+ETM*		13	기훈/NNG^		31
다른/MM		12	주민/NNG		28
인적/NNG		12	문가/NNG		27
강두/NNG^		10	영주/NNG^		26
기숙사/NNG		9	좌/NNG*		24
공부/NNG		9	수호/NNP^		24
공길/NNP^		7	벌구/NNG^		24
사내/NNG	JP	25	민주/NNG*	DPK	38
연결부/NNG		14	진석/NNP^		16
정민/NNG^		13	더불/VV*		11
노동자/NNG*		10	반색/NNG		9
비상구/NNG		9	여사/NNG		8
구석/NNP		8	경수/NNG^		7
겨눈다/VV+EF		8	병호/NNP^		7
발전소/NNG		8	브리핑/NNG		6
상근/NNG^		7	명/XPN		6
피켓/NNG*		7	전두환/NNP^		6

표-4에서는 발생 빈도수/분류 변경 빈도수 방법에 의해 추출된 상위 10개 키워드를 보여주고 있다. 이 방식은 발생 빈도수를 같이 고려하기 때문에 분류 변경 빈도수가 낮지만 발생 빈도수가 높아서 표-3과 다르게 키워드로서 추출된 단어들이 표-4에서 적색으로 표시되었다. 새롭게 추출된 키워드들의 일부는 각 정당의 특징을 일부 표현하는 것으로 추정되는 단어들이 일부 뽑힌 반면에, 큰 의미없이 자주 문장에 등장하기 때문에 뽑힌 단어들이 있는 것으로 판단된다. BM과 JP에서 "*"으로 표시된 단어들은 대부분 일반 문장에서 자주 발생하는 단어들이 새롭게 추가된 것으로 보이는 반면에, LKP에서는 보수층을 대표하는 "노인"과 공권력을 대표하는 "경찰"이 키워드로 선정되고, DPK에서는 과거 민주화 운동에 참여한 사람들이 억울하게 감옥에 가게되는 것과 연관된 "수인"이 키워드로 선정된 것으로 추정된다.

표 4. 발생 빈도수/분류 변경 빈도수 방법에 의해 추출된 상위 10개 키워드
Table 4. Examples of top 10 keywords extracted from geometric mean of the class change frequency and the occurrence frequency
(Note: The keywords in the table are written in Korean because it is inappropriate to translate them into English, which can lead to the loss of their real meanings.)

keyword	PO	f_i	s_i	keyword	PO	f_i	s_i
치성/NNG	BM	45	142	진호/NNG	LKP	118	307
미래/NNG		22	113	정은/NNP		94	248
다른/MM		12	188	회장/NNG		50	450
재수/NNG		20	52	문/NNG*		9	1293
공부/NNG		9	79	석/NNG*		9	584
같/VA*		2	296	주민/NNG		28	179
큰/VA+ETM*		4	125	기훈/NNG		31	158
강두/NNG		10	49	경찰/NNG*		11	381
걸어가/VV*		3	153	노인/NNG*		21	178
희재/NNP*		6	76	영주/NNG		26	119
사내/NNG	JP	25	60	민주/NNG	DPK	38	52
정민/NNG		13	22	진석/NNP		16	53
구석/NNP		8	31	기/NNG*		2	133
연결부/NNG		14	16	더불/VV		11	22
노동자/NNG		10	15	대문/NNG		6	27
발전소/NNG		8	17	수인/NNG*		6	20
선생/NNG*		5	27	브리핑/NNG		6	17
서/VV+EC*		1	119	전두환/NNP		6	17
비상구/NNG		9	12	반색/NNG		9	11
나/VV*		2	46	여사/NNG		8	12

표-5에서는 후보 키워드 TF-IDF 방법에 의해서 추출된 상위 10개의 키워드를 보여주고 있다. 앞에서 설명한 바와 같이 IDF를 계산하기 위해서 TF-IDF를 계산하는 대상에 속하지 않은 문장들은 1과 전체 문장 개수인 472,347의 기하 평균에 가까운 1000개의 문장들을 묶어서 하나의 가상의 문서로 취급하여 TF-IDF를 계산하였고, 이 과정에 의해서 생성된 가상의 문서는 표-5에서 제시된 바와 같이 약 460여개의 가상 문서로 구성이 됨을 확인할 수 있다. 표-5에서의 "*"으로 표시한 부분은 표-3에 포함된 키워드를 표시하고 있고, 약

50%정도가 일치하는 것을 확인할 수 있다. 또한, BM의 키워드인 “인적”, “대학”, “교회” 등은 표-6에서 keyBERT에 의해서 추출된 키워드와 유사성이 높은 것 단어들이 추출된 것으로 확인되며, JP의 키워드인 “발전소”와 “방사능”은 원자력 발전소와 관련한 정의당의 활동과 관련이 있을 것으로 추정된다. LKP의 “인민군”, “정은”, “부대”, “경례”, “선동”, 및 “국군” 등은 표-6에서 keyBERT에서 추출된 LKP와 유사한 맥락을 가지면서 보다 다양하게 해당 개념들을 표현하는 키워드가 선정된 것을 확인할 수 있다. 이와 유사하게 DPK의 키워드인“독방”, “구치소”, “시민”, 및 “연행” 등도 표-6에서의 keyBERT에서 추출된 개념을 보다 확장된 개념으로 제시하는 키워드들이 선정된 것을 확인할 수 있다.

표 5. 후보 키워드 TF-IDF 방법에 의해 추출된 상위 10개 키워드
Table 5. Top 10 keywords extracted from candidate keyword TF-IDF
(|D| = 464 for BM, 473 for JP, 452 for LKP, 472 for DPK)
(Note: The keywords in the table are written in Korean because it is inappropriate to translate them into English, which can lead to the loss of their real meanings.)

keyword	PO	TF-IDF	keyword	PO	TF-IDF
인적/NNG	BM	50.17	인민군/NNG	LKP	120.67
대학/NNG		38.84	정은/NNP*		117.61
재수/NNG*		34.38	부대/NNG		109.12
바른/VA+ETM*		32.20	경례/NNG		101.31
교회/NNG		31.52	주민/NNG*		94.93
면접/NNG		27.44	비서/NNG		78.76
완득이/NNP		27.11	진호/NNG*		77.31
구조/NNG		25.81	벌구/NNG*		71.33
기숙사/NNG*		25.58	선동/NNG		70.92
치성/NNG*		25.25	국군/NNG		70.59
노동자/NNG*	JP	22.71	민주/NNG*	DPK	47.86
연결부/NNG*		19.02	더불/VV*		43.45
피켓/NNG*		16.82	전두환/NNP*		29.23
비상구/NNG*		14.73	강필/NNG		24.91
발전소/NNG*		12.17	브리핑/NNG*		18.41
미사/NNG		11.56	독방/NNG		16.59
토치카/NNG		11.43	구치소/NNG		15.87
투쟁/NNG		11.43	베/NNG		14.69
상근/NNG*		11.31	시민/NNG		14.20
방사능/NNG		10.83	연행/NNG		11.98

4-4 KeyBERT로 추출된 키워드

표 6는 각 정치적 편향으로 판단된 문장으로 구성된 4개의 문서에 대해 KeyBERT로 키워드를 추출한 결과이다. BERT 임베딩은 기본적으로 형태소 분석기를 사용하지 않는다. 하지만 키워드를 띄어쓰기 단위로 추출하면 조사/어미도 함께 추출되는 문제가 있어서, 형태소 분석기를 통해 조사와 어미를 띄어쓰기한 후에도 키워드를 추출해 보았다. BM의 경우 학교 관련 용어가 많이 추출되며, DPK의 경우 정당이나 선거 관련 용어가, JP의 경우 노동이나 범죄 관련 용어(형태소 분석기를 적용한 경우에 추출되는 키워드는 특이하게 동사가 많음), 그리고 LKP의 경우 군사 관련 용어가 많이 추출되었다. 이는 각 정치적 편향으로 결정된 문장들이 일정한 패턴을 갖고 있음을 반증한다. KeyBERT는 문서의 주제와 유사도가 높은 단어들을 키워드로 선택하는 특징이 있기 때문에 유사한 단어들이 키워드로 뽑힘을 확인할 수 있다. 이 관점에서 볼 때에 KeyBERT는 문서가 갖는 중요한 개념이나 주제가 무엇인지를 확인하는데 좋은 방법으로 판단되며, KeyBERT에서 뽑힌 단어들은 각 정당이 갖는 핵심 개념을 파악하는데 도움이 되는 키워드를 제공하고 있음을 확인할 수 있다. 하지만, 유사도를 기반으로 키워드를 뽑는 다는 특징 때문에, 소수개의 키워드를 뽑을 때에 특정 정당의 다양한 측면에서의 키워드를 선정하기에는 한계를 갖는 방식임을 알 수 있다.

표 6. KeyBERT로 추출된 키워드
Table 6. Keywords extracted by KeyBERT
(Note: The keywords in the table are written in Korean because it is inappropriate to translate them into English, which can lead to the loss of their real meanings.)

Political Orient. (PO)	Keywords extracted by KeyBERT
BM (not morphologically analyzed)	'학교생활', '선생으로서', '초등학교였네', '1학년보다', '등록금도', '학원에서', '학교이다', '학교에서는', '선생들도', '학원은'
BM (morphologically analyzed)	'교사', '가르쳐도', '교육학', '국민학교', '교육부', '학원', '교양', '유치원', '초등학생', '가르쳤'
DPK (not morphologically analyzed)	'급장선거에', '유권자들과', '부정선거에', '투표인단께', '선거에', '반장선거', '투표권을', '선거개입', '민주에게는', '후보자들은'
DPK (morphologically analyzed)	'민주당', '정치인', '선거인단', '민주', '유권자', '정치', '투표', '민주주의', '정당', '선출'
JP (not morphologically analyzed)	'노동자였습니다', '막노동꾼들을', '피해자였어요', '망치더니', '뜯더니', '부패한', '범인은', '도둑', '변명하듯', '살인범이란'
JP (morphologically analyzed)	'어둑', '짤린', '일자', '유괴범', '악보', '처박힌', '짤리', '짤렸', '버둥거리', '처박혀'
LKP (not morphologically analyzed)	'군사들', '군병력들이', '국방군', '의용군들이', '병력들도', '치곤국방부가', '국군장병', '군인들의', '인민군병사들의', '군병력도'
LKP (morphologically analyzed)	'국방군', '군인', '군부대', '군사', '미군부대', '포병대', '군대', '군복', '군사분계선', '병사'

4-5 TF-IDF로 추출된 키워드

후보 키워드 TF-IDF에 대비하여 기존의 TF-IDF를 적용하여 키워드를 추출하는 경우의 특징을 표 7에 정리하였다. 정치 특징 별로 구분된 문장들의 모임에 속하는 모든 형태소

에 대해서 TF-IDF를 구하여 그 값이 큰 상위 10개의 키워드를 구한 것이다. BM의 경우에는 표 5와 비교하여 새롭게 대체된 형태소들은 특별한 의미를 부여하기 어려운 형태소들이 추출된 것을 확인할 수 있다. 하지만, “야구” 같은 경우에는 BM의 정치 성향을 갖은 사람들이 즐겨하는 운동이라고 추정할 수도 있지만, 그 근거를 제시하기 위해서는 추가적인 보다 면밀한 연구가 필요할 것이다. JP의 경우에는 정의당 명에 해당되는 “정의”가 키워드로 추가되어 일차원적인 생각될 수 있는 키워드가 선정된 것을 확인할 수 있으며, LKP의 경우에는 새롭게 추가된 단어들이 주로 keyBERT에서 추정되는 LKP의 핵심 주제인 “군”에 관련된 단어들이 추가된 것을 확인할 수 있다. DKP의 경우에도 keyBERT에서 추정되는 DKP의 핵심 주제인 “선거”에 관련된 단어들이 추가된 것을 확인할 수 있다.

표 7. 전체 문서에서 TF-IDF로 추출된 상위 10개 키워드
Table 7. Keywords extracted by TF-IDF
(|D| = 464 for BM, 473 for JP, 452 for LKP, 472 for DPK)
(Note: The keywords in the table are written in Korean because it is inappropriate to translate them into English, which can lead to the loss of their real meanings.)

keyword	PO	TF-IDF	keyword	PO	TF-IDF
인적/NNG	BM	50.17	인민군/NNG	LKP	120.67
대원/NNG		43.81	정은/NNP		117.61
야구/NNG		40.89	대원/NNG		114.15
서류/NNG		40.57	부대/NNG		109.12
하선/NNP		39.98	경례/NNG		101.31
대학/NNG		38.84	금고/NNG		96.67
재수/NNG		34.38	주민/NNG		94.93
후보/NNG		33.98	병사/NNG		87.92
학/NNG		33.85	비서/NNG		78.76
바른/VA+ETM		32.20	군/NNG		78.24
노동자/NNG	JP	22.71	경선/NNG	DPK	74.73
연결부/NNG		19.02	민주/NNG		47.86
피켓/NNG		16.82	더불/VV		43.45
비상구/NNG		14.73	대통령/NNG		33.75
정의/NNG		12.36	의원/NNG		32.29
발전소/NNG		12.17	전두환/NNP		29.23
미사/NNG		11.56	후보/NNG		27.33
투쟁/NNG		11.43	강필/NNG		24.91
토치카/NNG		11.43	감방/NNG		22.37
상근/NNG		11.31	수사/NNG		22.08

Ⅴ. 결 론

본 연구에서는 일상 생활에서 사용하는 단어에 담긴 숨겨져 있는 정치적 편향을 표현하는 키워드를 찾기 위해서, 다양한 주제가 혼재되고, 일상의 언어를 주로 표현하는 시나리오 말뭉치를 이용하여 정치적 편향을 표현하는 키워드를 추출하고 분석하였다. 제안 방법은 [10]의 정치 성향 추정 모델에 의해서 문장의 정치적 성향을 구분하고 동일한 정치적 성향으로 판단된 문장들로부터 키워드를 추출하는 방식으로 구성되었다. 제안된 3가지 방법중에서 형태소의 제거에 의해서 정치적 성향의 분류가 달라지는 키워드를 후보 키워드로 선정하고 M개의 문장을 가상의 문서로 처리하여 TF-IDF를 적용하여 키워드를 구하는 후보 키워드 TF-IDF가 정성적으로 나머지 2방식 보타 의미 있는 키워드를 추출하는 것으로 판단된다. [10]에 의해서 추출된 문서에 대해서 기존 방식인 keyBERT와 TF-IDF를 적용했을 때에도 일부 유사한 개념들이 추출되는 것을 확인할 수 있었다. 기존의 두 방식이 유사한 개념의 단어들을 주로 추출하는 것에 비해서, 후보 키워드 TF-IDF는 좀더 다양한 의미의 키워드를 도출하고 있는 것도 관찰되었다. 이 결과는 [10]이 일반적인 대화에서 정치적 성향을 추출하는데 활용할 수 있는 잠재력을 가지고 있으며, 이렇게 추출된 문장에서 키워드를 뽑을 때에 다양한 각도에서 문장의 모임에 대한 개념을 도출할 수 있는 방법에 대한 추가적인 연구가 요청된다.

이 연구에 있어서 한계는 다음과 같다. 본 연구에서는 시나리오 데이터를 이용하여 일상 생활에서 사용되는 단어의 정치적 편향을 연구하였으나, 사용된 데이터가 제한적이어서, 실제 영화의 특징에 기반한 정치적 성향의 단어들이 일부 뽑힌 것을 확인할 수 있다. 따라서, 보다 다양하고 많은 데이터의 수집을 기반으로 한 보다 심층적인 연구가 요청된다. 또한, 이전 연구[10]에서 사용된 Facebook 말뭉치의 크기가 제한적이라는 태생적 한계가 존재한다. 따라서, 정치적 성향의 키워드를 추출하는 첫단계인 문장의 정치적 성향을 판단하는 부분에서 정확도에 있어서 한계가 발생한다. 또한, 이전 연구[10]에서는 문서 단위로 정치적 편향이 판단되었지만, 본 연구에서는 문장 단위로 정치적 편향이 판단되었다. 이는 정치적 편향을 판단할 때, 단어의 수가 훨씬 적은 문장의 경우 정확한 정치적 편향을 판단하기 어렵다는 한계를 의미한다. 마지막으로 mecab-ko 태거 자체의 한계로 인해, 고유명사가 일반명사로 분석되는 문제가 있었고, 이로 인해서 형태소의 분포를 기반으로 키워드를 추출하는 방식에 한계가 존재한다.

연구의 한계를 극복하기 위한 향후 연구는 다음과 같이 제시될 수 있다. 정치적 편향을 표현하는 키워드에서 고유명사가 갖는 의미가 크다는 점을 확인했다. 따라서 정치적 편향을 표현하는 정확한 키워드의 추출을 위해서는 NLP 태스크 중 하나인 개체명 인식(NER, Named Entity Recognition)이 중요하다고 할 수 있다. 향후 연구 과제로 형태소 분석기에 개체명 인식 기능을 추가하는 것을 고려할 수 있다. 또다른 방향은 keyBERT의 경우에는 추출된 키워드가 모두 유사한 개념이지만, POS 태거로부터 자유롭다는 점이다. 이는

BERT와 같은 딥러닝 언어 모델을 이용하여 문서의 형태로 구성된 벡터 공간상에서 군집화되고 각 군집화된 집단의 센트로이드를 키워드로 선출하는 방식과 같이 문서의 핵심개념과 유사한 단어를 추출하는 방식에서 문서의 핵심 요소 벡터들을 구성하는 단어를 추출하는 방식으로 키워드를 뽑는다면, 특정 정치 성향을 구성하는 보이지 않는 중요한 요소를 뽑는데 활용될 수 있을 것으로 추정된다.

감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업입니다(과제번호 : NRF-2017R1A2B4007398). 일부는 과학기술정보통신부 및 정보통신산업진흥원의 '고성능 컴퓨팅 지원' 사업으로부터 지원받아 수행하였습니다.

참고문헌

[1] K. Piskorski et al. "Exploring Linguistically-Lightweight Keyword Extraction Techniques for Indexing News Articles in a Multilingual Set-up," *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation,* 2021.

[2] S. E. Robertson, "Term specificity," *Journal of Documentation,* Vol. 28, No. 1, pp. 164-165, Jan. 1972.

[3] S. E. Robertson, "Documentation Note: Specificity and Weighted Retrieval," *Journal of Documentation,* Vol. 30, No. 1, pp. 41-46, Jan. 1974.

[4] S. E. Robertson, "The probability ranking principle in IR," *Journal of documentation,* Vol. 33, No. 4, pp. 294-304, Apr. 1977. https://doi.org/10.1108/eb026647

[5] M. Grootendorst. "KeyBERT: Minimal keyword extraction with BERT"[Internet]. Available: https://maartengr.github.io/KeyBERT/index.html

[6] S. Lee and H. Kim, "Keyword extraction from news corpus using modified TF-IDF," *The Journal of Society for e-Business Studies,* Vol. 14, No. 4, pp. 59-73, Nov. 2009.

[7] J. Li, Q. Fan, and K. Zhang, "Keyword extraction based on tf/idf for Chinese news document," Wuhan University Journal of Natural Sciences, Vol. 12, No. 5, pp. 917-921, Sept. 2007. https://doi.org/10.1007/s11859-007-0038-4

[8] Y. Yoo, D. Lim, and T. Heo. "Solar cell patent classification method based on keyword extraction and deep neural network," arXiv preprint arXiv:2109.08796, 2021.

[9] A. Chang, B. Hua, and D. Yu, "Keyword Extraction and Technology Entity Extraction for Disruptive Technology Policy Texts," *Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents,* 2021.

[10] H. Kang and J. Yang, "Quantifying perceived political bias of newspapers through a document classification technique," *Journal of Quantitative Linguistics,* pp. 1-24, June 2020. https://doi.org/10.1080/09296174.2020.1771136

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *International conference on machine learning,* PMLR, pp. 1188-1196, 2014.

[12] E. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," *Annual Conference on Human and Language Technology,* Chuncheon, Korea, October 2014.

[13] H. Kang and J. Yang, "The Analogy test set suitable to evaluate word embedding models for Korean," *Journal of Digital Contents Society,* pp. 1999-2008, Oct. 2018.

[14] Explanations of mecab-ko POS tags [Internet]. Available: https://docs.google.com/spreadsheets/d/1-9blXKjtjeKZqsf4NzHeYJCrr49-nXeRF6D80udfcwY/edit#gid=589544265

[15] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[16] Github project, KeyBERT [Internet]. Available: https://maartengr.github.io/KeyBERT/index.html

[17] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*, Cambridge: Cambridge University Press, pp.100-123, 2008.

[18] M. Kim, "Evaluating the Usage of Passive Expression in the Scripts of Japanese Soap Operas and Radio Dramas," *Japanese Language & Culture Association of Korea*, Vol. 38, pp.131-152, Apr. 2017.

[19] E. Noh, "A study on the types and functions predictive expression in conversation; focused on the drama data," *Journal of Textlinguistics*, Vol.20, pp. 46-77, June 2006.

강형석(Hyungsuc Kang)

1996년 : 연세대학교 전파공학과 (공학사)
1998년 : 한국과학기술원 대학원 전기 및 전자공학과 (공학석사)
2013년 : 숙명여자대학교 TESOL 대학원 TESOL 전공 (TESOL 석사)
2020년 : 서울미디어대학원대학교 뉴미디어학부 (미디어공학 석사)

1998년~2003년: 삼성전자 네트워크사업부 선임연구원
2018년~2020년: 서울미디어대학원대학교 뉴미디어학부 재학
2020년~현 재: 서울미디어대학원대학교 뉴미디어학부 연구원
※관심분야 : 인공지능(Artificial Intelligence), 자연어처리(Natural Language Processing), 기계학습(Machine Learning) 등

양장훈(Janghoon Yang)

2001년 : University of Southern California (공학박사)

2001년~2006년: 삼성전자, 책임연구원
2006년~2010년: 연세대학교, 연구교수
2010년~현 재: 서울미디어대학원 뉴미디어학부, 부교수
※관심분야 : 중재 기술, 감성 공학, 간사이 공학, 정보이론, 이종 시스템 제어, 무선통신, 무선 네트워크, 뇌공학