석사학위논문

핵심어 추출 및 데이터 증강기법을 이용한 텍스트 분류 모델 성능 개선

2023년 2월 22일

전 북 대 학 교 대 학 원

통계학과

이 강 철

핵심어 추출 및 데이터 증강기법을 이용한 텍스트 분류 모델 성능 개선

Improving the performance of text classification models using keyword extraction and data augmentation techniques

2023년 2월 22일

전 북 대 학 교 대 학 원

통계학과

이 강 철

핵심어 추출 및 데이터 증강기법을 이용한 텍스트 분류 모델 성능 개선

지도교수 안 정 용

이 논문을 이학 석사 학위논문으로 제출함.

2022년 10월 19일

전 북 대 학 교 대 학 원

통계학과

이 강 철

이강철의 석사학위논문을 인준함.

위 원 장 전북대학교 교 수 양성준 (인)

부위원장 전북대학교 교 수 최혜미 (인)

위 원 전북대학교 교 수 안정용 (인)

2023년 1월 3일

전 북 대 학 교 대 학 원

목 차

표 목차

그림 목차

ABSTRACT

Improving the performance of text classification models using keyword extraction and data augmentation techniques

Lee, Gang Cheol
Department Of Statisitics
The Graduate School
Jeonbuk National University

Topic modeling aims to identify and categorize topics latent in documents, and is useful for exploring core topics of each document and the characteristics of the topics. However, a problem with interpreting topics this technique is that common terms often appear near the top of multiple topics, making it hard to extract keywords identifying the topics. Another weakness is that this technique can lead to loss of information when synonyms are excluded from keywords, and high performance often depends on the size and quality of data. To improve these problems, we propose a method that utilizes relevance and word embedding techniques for extracting keywords. In addition, we use the EDA(easy data augmentation) techniques to increase the size of the data, and then apply the KoBERT model for boosting performance on text classification tasks. As a result of data analysis, it was possible to grasp the specific characteristics of the topics based on the discriminating keywords. The results also showed that using the augmented data sets, the text classifier model has higher accuracy than the original data sets with a score of 0.94 and 0.85, respectively.

Keywords : topic model, relevance, word embedding, data augmentation, text classification

제1장 서론

기계학습 및 자연어 처리 분야에서 토픽 모델링(topic modeling)은 수집된 문서들을 비슷한 특성을 가지는 개체들끼리 군집화하는 비지도 학습(unsupervised learning) 방법이다 (Redondo & Sandoval, 2016). 토픽 모델링은 문서에 출현하는 각 단어들이 통계적으로 특정 토픽에 포함될 확률을 파악하여 문서의 토픽(주제)을 추정하는 기법으로, 가장 널리 알려진 일반적인 모형은 LDA(Latent Dirichlet Allocation) 모형이다 (Blei, Ng, & Jordan, 2003).

LDA 모형은 문서에서 관측되는 단어가 잠재된 토픽을 생성하는 과정에서 확률적으로 관측되었다고 가정하기 때문에 생성적 확률 모형(generative probabilistic model)이라 부르기도 하며, 사후확률을 사전확률과 데이터의 혼합으로 가정하므로 3층 위계적 베이지안 모형(three-level hierarchical Bayesian model)으로 불리기도 한다 (Blei, Ng, & Jordan, 2003). 토픽 모델링 기법으로 LDA 모형 이외에도 CTM(Correlated Topic Model)과 STM(Structural Topic Model)이 제안되어 있다. CTM은 토픽 사이의 연관 관계를 적극적으로 반영한 모델이며 (Blei & Lafferty, 2006), STM은 문서의 메타데이터 정보를 추정할 수 있는 모형이다 (Roberts, Stewart, & Airoldi, 2016). 이러한 모형 중에서 LDA 모형은 다른 모형들에 비해 훈련 과정(training process)이 단순하다는 특징과 함께 데이터의 차원을 축소하는 데 유용하며, 의미적으로 일관성이 있는 주제들을 추출하는 데 장점을 가지고 있다 (Mimno & McCallum, 2008; 양연희, 2021).

LDA 모형은 현재 다양한 분야의 많은 연구들에서 활용되고 있는데, 국외의 대표적인 연구로 Geletta, Follett & Laugerman(2019), Sun, Yu & Yang(2020), Moss & Rohrmeier(2021), Nastiti, Hidayatullah & Pratama(2021) 등을 들 수 있다. 국내에서도 LDA 모형을 활용한 연구들이 많이 진행되었다. 김성근, 조혁준과 강주영(2016)은 학술논문 검색사이트 DBPia에서 텍스트 마이닝을 주제로 한 학술논문들을 수집해 LDA 모형을 이용하여 분석하였으며, 임영재 등(2019)은 감성적 교재수준에 따라 토픽

모델링의 신뢰도 비교 연구를 수행하였다. 남승주 등(2020)은 LDA 모형을 사용하여 공항산업의 동향을 분석하였으며, 한지영과 허고은(2021)은 코로나바이러스 감염증으로 인하여 변화된 대학 강의만족도 영향요인을 파악하기 위하여 온라인 커뮤니티인 <에브리타임>의 강의평가 데이터에 LDA 모형을 적용하였다.

그러나 이러한 많은 활용에도 불구하고 LDA 모형을 적용한 토픽 모델링은 여러 가지 문제점을 가지고 있다. 대표적인 문제점은 서로 다른 토픽 내에서 동일한 단어들이 상위 비중을 갖는 경우, 토픽 간 변별력이 있는 핵심어 추출이 어렵다는 것이다. 예를 들어, 어떤 단어가 토픽 A와 토픽 B 모두에서 상위 비중을 갖는다면 해당 단어는 토픽 간 변별력을 낮추게 되는 결과를 유발한다. 이러한 문제는 패널티(penalty) 척도와 textrank를 이용하면 어느 정도 개선할 수 있으나 여전히 빈도수에 기반한 핵심어 추출 기법의 한계가 존재한다 (김은회와 서유화, 2020). 예를 들어, 직업군인의 급여에 관한 기사(문서)가 있을 때 이 기사에서 빈도수에 기반에 핵심어를 추출한다면 '군인' 이라는 단어가 핵심어로 추출될 확률이 높을 것이다. 그러나 '군인' 이라는 단어와 의미적 유사성이 있는 단어인 '직업군인' 이 핵심어로 추출되지 못한다면 해당 기사가 일반 '일반병사' 의 급여에 관한 기사인지 '직업군인' 의 급여에 관한 기사인지 판단하기가 어려울 수 있다. 또 다른 문제점은 새로운 데이터가 입력되었을 때 이 데이터를 적절한 토픽으로 분류하기 위해서 토픽 모델링을 다시 수행해야 한다는 단점이 있다.

이러한 문제점을 개선하기 위하여 본 연구에서는 핵심어 추출 시 연관성 척도(relevance)와 워드 임베딩(word embedding) 기법을 적용하는 방법을 제안한다. 이 방법의 적용 단계는 다음과 같다. 첫 번째 단계는 LDA 모형을 이용하여 토픽을 찾고, 출현 빈도수에 기반하여 1차 핵심어를 추출한다. 두 번째 단계에서는 1차 핵심어에 relevance 척도를 적용하여 1차 핵심어를 포함한 문서들을 추출하고, 이 문서들로부터 단어 집합을 생성한다. 마지막 단계에서는 워드 임베딩 기법을 이용하여 1차 핵심어와 의미적 유사성이 있는 단어를 추출 후 각 토픽의 최종 핵심어 집합을 구축한다.

또한, 본 연구에서는 새로운 데이터의 분류 성능을 개선하기 위하여 데이터 증강 기법(data augmentation technique)인 EDA(easy data augmentation) 기법을 이용하여 데이터를 양적으로 보강하고, 데이터를 분류하기 위하여 KoBERT 모델을 적용한다. EDA 기법은 학습 데이터가 부족할 때 데이터를 변형시켜 그 양을 늘리는 기법으로 외부 데이터나 모델 훈련 없이 성능향상을 기대할 수 있다 (Wei & Zou, 2019). KoBERT는 구글에서 공개한 BERT 모델(Devlin 등, 2019)을 한국어 기반으로 제작한 것으로 한국어에 대해 많은 사전 학습이 이루어져 있고, 다중 분류가 가능한 강점이 있다 (황상흠과 김도현, 2020). KoBERT 모델의 분류 정확도는 상당히 높다고 알려져 있으나 학습 데이터가 부족할 경우에는 정확도가 떨어질 수밖에 없기 때문에 본 연구에서는 데이터 증강 기법을 이용하여 분류 정확도를 개선하는 방법을 사용한다.

본 연구에서 다루고자하는 연구 문제는 핵심어의 변별력 증가 및 데이터 증강 기법의 효용성이며, 구체적인 연구 질문은 첫째, relevance 척도와 워드 임베딩 기법을 적용해 추출한 핵심어가 기존의 기법으로 추출한 핵심어에 비해 토픽 간 변별력 있는가? 둘째, 토픽의 최종 핵심어 집합에 기반하여 명명한 토픽 레이블(label)을 통해 해당 토픽에서 다루는 주제를 직관적으로 파악할 수 있는가? 셋째, EDA를 통해 증강 시킨 데이터로 학습한 분류 모델이 그렇지 않은 경우에 비해 개선된 성능을 보이는가? 등이다. 2장에서는 본 연구에 사용된 기법들을 소개하고, 3장에서는 데이터 분석 방법 및 결과에 대해 설명한다. 4장에서는 결론 및 한계점, 향후 연구방향에 대해 기술한다.

제2장 관련연구

2.1 토픽 모델링

토픽 모델링은 문서 집합들로부터 토픽을 추출하는 기법이다. 토픽이란 문서가 가지는 주제라고 할 수 있다. 토픽 모델링은 문서 집합들을 군집화하기 위한 계층적 확률모델로 여러 분야에서 사용되고 있다. 특히, 컴퓨터 비전 분야에서 많이 이용하고 있으며, 주석(comment) 또는 레이블(label)이 부착된 이미지 분류 등과 같은 다양한 문제를 처리하기 위해 사용된다 (Blei, Carin, & Dunson, 2010). 토픽 모델링의 기본 가정은 다음과 같다.

(가) 모든 문서는 토픽들의 혼합체이다.

각 문서는 몇 가지 토픽에서 나온 단어가 특정 비율로 포함되어 있다고 가정한다. 예를 들어, <표 1>과 같이 2개의 토픽을 가지는 문서가 있다고 가정할 때, 각 문서는 관측 단어들의 빈도수에 기반하여 토픽 비율을 계산한 후 해당 문서가 어떤 토픽에 더 치중하고 있는지 결정한다.

<표 1> 토픽 비율에 따른 문서의 토픽 결정

문서 번호	토픽 비율	토픽
1	토픽 A : 90%, 토픽 B : 10%	A
2	토픽 A : 30%, 토픽 B : 70%	B

(나) 모든 토픽은 단어들의 혼합체이다.

문서는 단어들로 구성되어 있고 단어는 토픽을 반영한다. 예를 들어, 정치와 음식이라는 두 가지 토픽이 존재한다고 가정하자. 정치 토픽에서 가장 흔히 사용되는 단어는 '대통령', 'TV', '국회의원' 등이 될 수 있으며, 음식 토픽에서 많이 사용되는 용어는 '맛집', 'TV', '불고기' 등이 될 수 있을 것이다.

다음 <표 2>와 같이 음식이라는 토픽으로 분류된 문서가 있을 때, ‘국회의원’이라는 단어는 정치 토픽, ‘맛집’, ‘식사’, ‘식당’ 등의 단어는 음식 토픽과 밀접하게 연관되어 있고, ‘TV’라는 단어는 정치 토픽과 음식 토픽 모두에 비슷하게 연관되어 있다. 이러한 연관 비중을 <표 3>과 같이 정리할 수 있는데, 많은 단어들이 비중의 차이는 있지만 여러 토픽을 중첩(overlap)하여 반영하는 것을 알 수 있다. 특히, ‘TV’라는 단어는 정치 토픽과 음식 토픽 양쪽을 비슷하게 반영하고 있는데, 이 단어가 양쪽 토픽에서 상위 비중을 갖는다면 토픽의 변별성이 떨어지게 되는 문제가 발생한다.

<표 2> 음식으로 분류된 문서

문서 번호	토픽 비율	토픽
1	국회의원 A씨가 TV에서 맛집으로 소문난 B라는 식당에 들려서 식사를 맛있게 했습니다. 또 ~~	음식

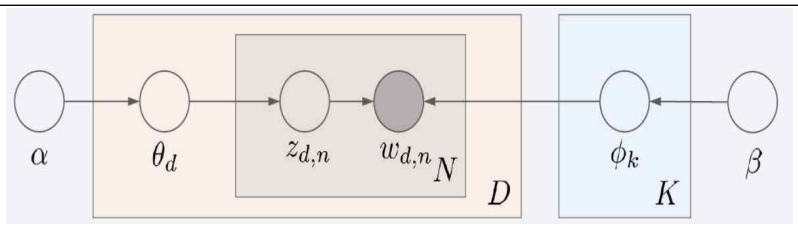
<표 3> 단어-토픽 연관성 비중

토픽 / 단어	정치	음식
TV	0.4	0.4
국회의원	0.9	0.1
식당	0.1	0.9
맛집	0.0	0.8

2.2 LDA (Latent Dirichlet Allocation) 모형

LDA는 토픽 모델링 기법 중 가장 많이 이용되는 모형으로 문서에 출현하는 단어들의 정보를 이용하여 각 문서에 어떤 토픽들이 존재하는지를 추정한다. LDA는 토픽별 단어의 분포, 문서별 토픽의 분포를 모두 추정하

며, LDA 모형은 <그림 1>과 같이 표현된다.



여기서 K : 토픽의 수
D : 문서의 수
N : 단어의 수
α : 토픽의 Dirichlet prior weight, θ값 결정 파라미터
β : 토픽별 단어의 Dirichlet prior weight, ϕ값 결정 파라미터
θ_d : d번째 문서의 토픽 비중
$z_{d,n}$: d번째 문서에서 n번째 단어의 토픽
$w_{d,n}$: d번째 문서에서 n번째 단어
ϕ_k : 토픽 k의 단어 w의 생성확률

<그림 1> LDA 모형

<그림 1>에서 관찰 가능한 변수는 d번째 문서에 출현한 n번째 단어 $w_{d,n}$이 유일하며, 이 정보만을 가지고 사용자가 지정하는 파라미터 α와 β를 제외한 모든 잠재 변수를 추정해야 한다. 이를 위해 LDA는 토픽의 단어분포와 문서의 토픽 분포의 결합으로 문서 내 단어들이 생성된다고 가정한다. $w_{d,n}$을 이용하여 잠재변수를 역으로 추정하는 과정은 실제 관찰 가능한 문서 내 단어를 가지고 알고 싶은 토픽의 단어분포, 문서의 토픽 분포를 추정하는 것이다. 즉, 관측값 $w_{d,n}$이 주어졌을 때, 토픽의 단어분포와 문서의 토픽 분포의 결합 사후확률 $p(z,\phi,\theta|w)$를 최대로 만드는 z,ϕ,θ를 찾는 것이다.

사후확률을 계산하려면 분모에 해당하는 $p(w)$를 구해야 되는데, $p(w)$는 잠재변수 z,ϕ,θ의 모든 경우의 수를 고려한 각 단어(w)의 등장 확률을 가

리킨다. 그러나 z, ϕ, θ는 직접 관찰하는 것이 불가능하고, $p(w)$를 구할 때 z, ϕ, θ의 모든 경우를 감안해야 된다. 따라서 $p(w)$를 계산하는 것이 쉽지 않기 때문에 MCMC(Markov Chain Monte Carlo)의 한 방법인 깁스 샘플링 (Gibbs sampling) 같은 기법을 사용한다.

2.2.1 깁스 샘플링 (Gibbs sampling)

깁스 샘플링은 MCMC 알고리즘의 한 예로, 두 개 이상의 확률변수의 결합 확률 분포로부터 일련의 표본을 생성하는 확률적 알고리즘이며 결합 확률분포나 그에 관련된 확률 계산을 근사하기 위해 사용된다. 깁스 샘플링은 다음번 생성될 샘플이 현재 샘플에 영향을 받는다는 점에서는 MCMC와 같지만, 나머지 변수는 그대로 두고 한 변수에만 변화를 준다는 점이 다르다.

LDA에서 깁스 샘플링은 사후확률 $p(z, \phi, \theta | w)$를 구하기 위해 사용된다. 정확하게 표현하면 나머지 변수는 고정시킨 채 한 변수만을 변화시키되, 불필요한 일부 변수를 샘플링에서 제외하는 붕괴된 깁스 샘플링(collapsed Gibbs sampling) 기법을 이용한다. LDA에서는 $p(z, \phi, \theta | w)$를 구할 때 ϕ, θ를 계산에서 생략하고, 다음 수식 (1)과 같은 확률을 구한다.

$$p(z_{d,i} = j | z_{-i}, w) \tag{1}$$

위 수식에서 w는 관측 단어 값이고, z는 각 단어가 어떤 토픽에 할당되어 있는지를 나타내는 변수이다. z_{-i}는 i번째 단어의 토픽 정보를 제외한 모든 단어의 토픽 정보를 나타낸다. 따라서 수식 (1)은 w와 z_{-i}가 주어졌을 때 문서의 i번째 단어의 토픽이 j일 확률을 나타낸다. 잠재변수 z에 대한 확률을 구할 때 사용되는 수식은 아래와 같다.

$$p(z_{d,i} = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha} \tag{2}$$

수식 (2)에서 $n_{-i,j}^{(\cdot)}$ 는 j번째 토픽에서 i번째 단어를 제외한 모든 단어의 빈도수, $n_{-i,j}^{(w)}$ 는 j번째 토픽의 단어 w의 빈도수이다. 즉, 수식 (2)의 오른쪽 항목에서 첫 번째 비율 부분은 j번째 토픽에서 임의의 단어 w가 어느 정도 비중을 차지하고 있는지를 나타낸다. 두 번째 비율 부분은 문서 d에서 j번째 토픽이 어느 정도 비중을 차지하고 있는지를 나타낸다. 파라미터 α와 β는 두 비율 부분이 0이 되는 것을 방지하며, V와 K는 문서 내 고유 단어 수와 설정한 토픽 수를 의미한다.

2.2.2 혼란도 (perplexity)

LDA에서 토픽 수 K는 연구자가 지정하는 파라미터이다. 토픽 모델링에서는 단어 집합 w에 대한 로그-우도 함수값을 이용하여 수식 (3)과 같이 혼란도를 계산해 최적의 토픽 수 K를 결정한다.

$$Perplexity(w) = \exp\left[\frac{-\log\ p(w)}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{jd}}\right] \tag{3}$$

여기에서 $\log\ p(w) = \sum_{d=1}^{D}\sum_{j=1}^{V} n^{jd} \log[\sum_{k=1}^{K} \theta_k^d \phi_k^j]$ 이다. 토픽 모델링 관점에서 우도함수는 잠재변수 θ와 ϕ가 주어졌을 때 단어의 분포에 대한 설명력을 나타내는 지표라고 할 수 있다. Perplexity는 로그-우도 함수 $\log p(w)$에 음수를 취한 형태이다. 따라서 $\log p(w)$의 값이 클수록 혼란도의 값이 작아지고, 최적의 K에 가깝다고 할 수 있다. 위 수식을 요약하면 다음 수식 (4)와 같다.

$$Perplexity(w) = \exp\left[-\sum_{d=1}^{D}\sum_{j=1}^{V}\sum_{k=1}^{K} \theta_k^d \phi_k^j\right] \tag{4}$$

앞에서 설명한 바와 같이, 수식 (4)에서 D는 총 문서의 수, V는 고유

단어의 수, K는 토픽 수이다. θ_k^d는 d번째 문서에서 k번째 토픽의 비중을 의미한다. ϕ_k^j 는 k번째 토픽에서 j번째 단어의 비중을 의미한다. 토픽의 개수를 정할 때 K값을 변화해 가면서 토픽 수에 따른 혼란도를 계산한 뒤 과적합을 고려하여 최종 토픽 수를 선택한다. 과적합을 고려하는 이유는 새로운 데이터가 입력되었을 때 최적의 토픽 수가 변경될 수 있기 때문이다.

2.2.3 연관성 (relevance)

LDA를 이용하여 산출된 토픽-단어 확률인 ϕ는 토픽 내 단어 출현 빈도수에 기반한 척도이다. 이러한 빈도수에 기반한 척도는 서로 다른 토픽 내에서 동일한 단어들이 상위 비중을 갖는 경우, 토픽 간 변별력이 있는 핵심어 추출이 어렵다는 문제점을 가지고 있다. 예를 들어, <그림 2>와 같은 내용을 갖는 문서를 생각해 보자.

> 오늘 날씨는 어젯밤의 온 폭설로 인하여 도로가 많이 미끄러울 것으로 예상됩니다. 특히 어젯밤 서울시에는 교통사고가 급격히 증가했습니다. 또한, 한파로 인해 외출 시 방한 도구들을 꼭 챙기시기를 바랍니다.

<그림 2> 예제 문서

<그림 2>의 문서가 주어졌을 때 빈도수에 기반하여 추출한 $\phi_{\text{날씨}}$, $\phi_{\text{교통사고}}$ 등을 살펴보면 상위 비중 값을 가지는 단어들이 비슷한 단어들로 구성되어 있을 가능성이 높다. 이런 경우, ϕ에서 높은 비중을 가지는 단어들을 고려하여 토픽 이름을 명명한다면 토픽 간에 변별성이 없어진다. 이러한 문제를 개선하기 위한 하나의 방법은 수식 (5)와 같이 표현할 수 있는 연관성(relevance) 척도를 이용하는 것이며 (Sievert & Shirle, 2014), 연관성은 단어와 토픽 간의 관계를 나타낸다.

$$r(w,k \mid \lambda) = \lambda \log(\phi_{k,w}) + (1-\lambda) \log(\frac{\phi_{k,w}}{p_w}) \quad (5)$$

수식 (5)에서 p_w는 전체 문서 집합 내에서 해당 단어의 등장 확률이며, $\phi_{k,w}$는 k번째 토픽에서 단어 w의 비중이다. 또한, λ는 0~1 사이의 값을 가지는 파라미터이다. λ 값이 1 이라면 기존의 토픽-단어 확률인 ϕ값의 상위 단어를 기반으로 토픽 이름을 명명한다는 의미이고, λ 값이 1보다 작을 경우 해당 토픽에서 많이 출현한 단어라도 다른 토픽에서 자주 등장하는 단어라면 relevance 값을 감소시키는 역할을 한다. Sievert와 Shirle (2014)는 λ의 최적값으로 2/3를 사용하였다.

2.3 워드 임베딩

2.3.1 원-핫 인코딩(one-hot encoding)

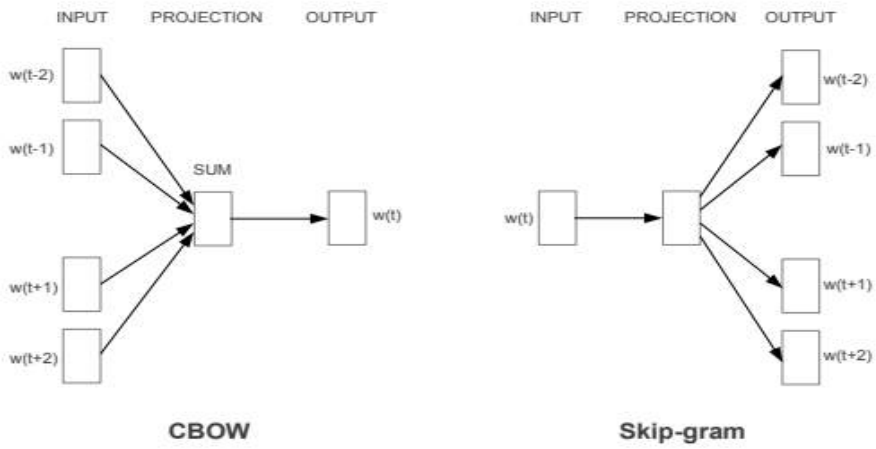
문자를 컴퓨터가 이해할 수 있는 숫자로 바꾼 결과 또는 그 과정을 임베딩(embedding)이라고 하며, 가장 간단한 형태의 임베딩 방법은 원-핫 인코딩(one-hot encoding)이다. 원-핫 인코딩은 단어 집합의 크기를 벡터의 차원으로 하고, 표현하고 싶은 단어의 인덱스에 1의 값을 부여하고, 다른 인덱스에는 0을 부여하는 방식이다. 이렇게 표현된 벡터를 원-핫 벡터(one-hot vector)라고 한다.

원-핫 벡터는 대부분의 원소가 0 값을 갖게 되는 희소벡터이고, 단어의 개수가 늘어날수록 벡터를 저장하기 위해 필요한 공간이 계속 늘어난다는 단점이 있다. 또 다른 문제는 단어의 유사도를 표현하지 못한다는 단점이 있다. 이러한 단점을 해결하기 위해 단어의 잠재 의미를 반영하여 다차원 공간에 벡터화하는 기법이 두 가지 종류가 있다. 첫째, 단어의 출현 빈도수 기반의 벡터화 방법인 LSA(latent semantic analysis; Landauer 등, 1998), HAL(hyperspace analogue to language; Lund & Burgess, 1996) 등

이 있다. 둘째, 예측 기반으로 벡터화하는 NNLM(feedforward neural network language model; Bengio 등, 2003), RNNLM(recurrent neural network language model; Mikolov 등, 2010), Word2Vec(Mikolov 등, 2013), FastText(Bojanowski 등, 2017) 등이 있다.

2.3.2 Word2Vec

앞에서 언급한 바와 같이 원-핫 벡터는 희소벡터이고, 단어 벡터 간 유의미한 유사도를 계산할 수 없다는 단점이 있다. 이를 해결할 수 있는 대표적인 방법이 Word2Vec이다. Word2Vec의 학습방식은 CBOW(continuous bag of words)와 Skip-Gram 두 가지가 있다(Mikolov 등, 2013). CBOW 모델은 주변에 있는 단어들을 활용하여 중간에 있는 단어들을 예측하는 방법이며, Skip-Gram 모델은 중간에 있는 단어들을 활용하여 주변 단어들을 예측하는 방법이다. 이 방법들을 도식화하면 <그림 3>과 같다.



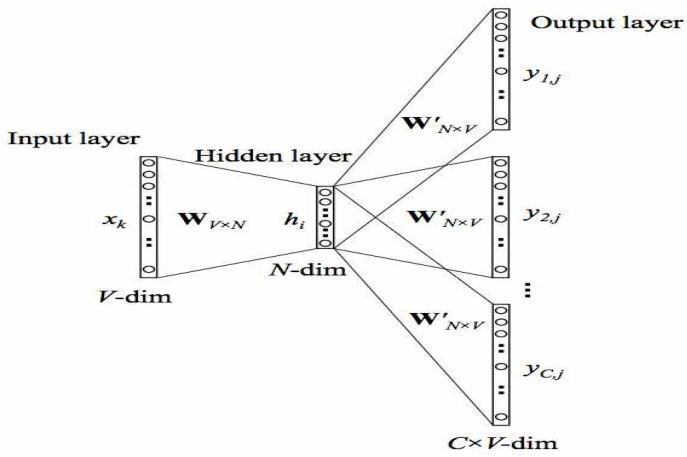
<그림 3> CBOW와 Skip-Gram의 구조

두 모델을 비교해보면, 학습 속도 측면에서는 CBOW 모델이 더 효율적이나 단어 분산 표현의 정밀도 측면과 말뭉치가 커질수록 저빈도 단어나

유추 문제의 성능 측면에서 Skip-Gram 모델이 더 좋은 성능을 발휘한다 (Mikolov, 2013). 또한, 한국어 문서에 적용했을 경우에도 Skip-Gram 모델의 성능이 더 우수한 것으로 알려져 있다 (강형석과 양장훈, 2019).

2.3.3 Skip-Gram

Skip-Gram은 <그림 4>와 같이 은닉층이 하나인 간단한 구조의 신경망이다. <그림 4>에서 x_k 는 k번째 단어의 원-핫 벡터이며 W와 W'은 단어 집합의 가중치 행렬이다. Skip-Gram은 중심 단어를 이용하여 주변 단어들을 좀 더 잘 예측하기 위해 W와 W'을 학습시킨다. 학습 진행 과정에서 중심 단어가 몇 개의 단어를 예측할 것인가를 사전에 설정해야 하는데 이를 window-size라고 부른다. 예를 들어, window-size를 2로 설정하면 중심 단어 좌측과 우측에 2개의 단어를 참조하여 손실함수를 계산 후 중심 단어 벡터를 업데이트한다.



<그림 4> Skip-Gram의 구조

Word2Vec의 Skip-Gram은 수식 (6)과 같은 다중 클래스 분류에 사용되는 소프트맥스 함수 출력값을 최대화하는 것을 목표로 한다. 소프트맥스 함수는 중심 단어를 이용하여 주변 단어들을 예측할 때 주변 단어들을 정

답으로 예측할 확률이 얼마인지 출력하는 함수이다.

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \tag{6}$$

수식 (6)의 소프트맥스 함수는 중심 단어(c)가 주어졌을 때, window-size 범위 내에 존재하는 주변 단어(o)가 나타날 확률을 의미한다. 우변의 v는 입력층과 은닉층을 연결하는 가중치 행렬 W의 행벡터, u는 은닉층과 출력층을 잇는 가중치 행렬 W'의 열벡터이다.

$$v_c^{t+1} = v_c^t + \alpha(u_o^t - \sum_{w=1}^{W} P(w \mid c) \cdot u_w) \tag{7}$$

중심 단어(c) 벡터의 업데이트 과정은 수식 (7)과 같이 계산된 그래디언트 값과 학습률(learning rate) α값 통해 업데이트된다. 그러나 이러한 과정은 문서의 수가 많아질수록 계산량이 매우 많이 증가한다는 단점이 있다. 이러한 문제점을 해결하기 위해 negative-sampling 기법이 제안되었다 (Mikolov 등, 2013).

2.3.4 Negative-Sampling

Negative-Sampling은 소프트맥스 함수의 출력값을 계산할 때 전체 단어를 대상으로 하지 않고 일부 단어만 이용한다. 사용자가 지정한 window-size 내에 등장하지 않은 단어(negative sample)를 일부(일반적으로 5~20개)만 선택하여 소프트맥스를 계산한다. 예를 들어, window-size를 3, negative-sample을 20으로 설정했다면 23개 단어의 가중치만 갱신하면 된다. 이러한 방식을 사용하여 성능은 유지하고 계산 시간은 단축할 수 있다.

2.4 KoBERT

자연어 처리 분야에서 딥러닝 기반의 대표적인 문서 분류 모델은 순환 신경망 모델인 RNN(recurrent neural network)이다. 그러나 RNN 모델은 입력 데이터에 대한 병렬 처리가 불가능하고, 구조적으로 고정된 크기의 벡터에 모든 정보를 압축하기 때문에 정보 손실 문제가 발생한다(Vaswani 등, 2017). 이러한 문제점을 개선하기 위해 트랜스포머(transformer) 기반의 BERT(bidirectional encoder representations from transformers) 모델이 제안되었다 (Devlin 등, 2019). BERT 모델은 RNN 모델과는 다르게 참조할 단어의 위치 정보에 기반한 단어 관계 정보를 이용함으로써 입력 데이터에 대한 병렬 처리가 가능하여 학습 시간을 크게 줄일 수 있다.

그러나 BERT 모델은 영어 문서에 대해 정확도는 높지만 한국어 문서에 대해서는 영어 문서보다 정확도가 떨어진다. 따라서 BERT 모델을 한국어 데이터에도 잘 활용할 수 있도록 KoBERT 모델이 개발되었다. KoBERT는 위키피디아, 뉴스 등에서 수집한 한국어 데이터를 추가로 학습시켰으며, 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 데이터 기반 토큰화(tokenization) 기법을 적용하여 성능을 향상시켰다. 이러한 장점에도 불구하고, 학습 데이터가 부족할 경우에는 문서의 분류 정확도가 떨어지기 때문에 데이터 증강 등을 통하여 정확도를 개선하는 방법이 필요하다.

2.5 EDA

EDA(easy data augmentation)는 학습 데이터가 부족한 상황에서 데이터를 증강하는 기법이다. 이 기법은 유의어 교체(synonym replacement), 임의 삽입(random insertion), 임의 교체(random swap), 임의 삭제(random deletion) 등과 같은 규칙을 사용하여 데이터를 증강한다.

2.5.1 유의어 교체(SR)

유의어 교체는 문장 내 임의의 단어를 동의어로 교체하는 방법이다. 문장 내에서 임의의 단어를 선택 후, 같은 의미의 단어를 사전에서 찾아 교체한다. 예를 들어, “이번 태풍의 범위는”이라는 문장에 유의어 교체 기법을 적용하면 “이번 태풍의 영역은”이라는 문장이 생성될 수 있다.

2.5.2 임의 삽임(RI)

임의 삽입은 문장 내에서 임의의 단어를 선택하여 동의어를 찾고, 이를 문장 내 임의의 위치에 삽입한다. 예를 들어, “이번 태풍의 범위는”이라는 문장에서 “범위”라는 단어가 선택되고, 이 단어의 동의어가 “영역”이라면 이를 문장 내 임의의 위치에 삽입하여 “이번 태풍의 영역 범위는”이라는 문장이 생성될 수 있다.

2.5.3 임의 교체(RS)

임의 교체는 문장 내에서 임의로 두 개의 단어를 선택 후 두 단어의 위치를 변경하는 방법이다. 예를 들어, “이번 태풍의 범위는”이라는 문장에서 “태풍”이라는 단어와 “범위”라는 단어가 선택되었다면 “이번 범위의 태풍은”이란 문장이 생성될 수 있다.

2.5.4 임의 삭제(RD)

임의 삭제는 문장 내에서 각 단어가 삭제될 확률을 지정하여 임의의 단어를 삭제하는 방법이다. 예를 들어, “이번 태풍의 범위는”이란 문장에서 “범위”라는 단어가 삭제 단어로 선택되었다면 “이번 태풍은”이라는 문장이 생성될 수 있다.

제 3장 연구 방법 및 분석 결과

3.1 연구 방법

3.1.1 데이터 수집

본 논문에서 사용한 데이터는 공공누리[1]에서 제1유형으로 제공하는 청와대 국민청원 데이터로 2017년과 2018년 데이터 중에서 5%를 임의 추출한 18,067건의 데이터이다. 해당 데이터는 깃허브[2]에서 csv 파일 형식으로 제공되며 데이터 형태는 <표 4>와 같다.

<표 4> 데이터 형태

article_id	category	title	content
58	일자리	국토교통부와....	안녕하세요? 존경하고 지지하는 문재인 대통령님 저는 성남시..
63	보건복지	살려주세요	안녕하십니까? 저는 올해 63세된 홀로 사는 늙은.....
...	...	...	...

category는 국민청원을 등록할 때 선택하는 청원의 범주이며 일자리, 보건복지, 행정, 경제민주화, 반려동물, 해꼬지, 행정, 정치개혁, 저출산, 육아/교육, 미래, 외교 등 총 17개로 구성되어 있다. title은 청원의 제목, content는 청원의 본 내용이다. 본 논문에서는 청원의 제목만으로 해당 청원의 주제를 파악하기 어렵다고 판단하여 청원의 본 내용인 content를 이용하여 데이터를 분석하였다.

3.1.2 데이터 전처리

한국어 텍스트 데이터의 전처리 단계에서 형태소 분석을 할 때 가장 많

1) https://www.kogl.or.kr/info/license.do
2) https://github.com/akngs/petitions

이 쓰이는 패키지는 KoNLP 패키지이다. 그러나 KoNLP의 경우 문장의 수가 많아질수록 데이터 처리 속도가 느려지는 단점이 있고 신조어, 은어에 대한 품사 분류 성능이 떨어져 본 연구에서는 Rmecabko 패키지를 사용하였다 (김준혁, 2017). 이 패키지는 한국어 텍스트에 대한 품사 태깅(tagging) 및 토큰화(tokenization) 기능을 제공하며, 띄어쓰기가 잘못된 경우에도 형태소 분석에 문제가 없고 전처리 속도가 빠르다는 장점이 있다. 또한, 신조어 또는 줄임말 등이 포함된 경우에도 좋은 성능을 보인다고 알려져 있다 (한지영, 허고은, 2021).

〈표 5〉 불용어 목록

불용어
“문재“, “안녕“, “대통령“, “필요“, “경우“, “때“, “국민“, “사람“, “생각“, “나라“, “문제“, “자신“, “사실“, “개인“, “현실“, “해결“, “발생“, “정부“, “국가“, “청원“, “이상“, “시간“, “이유“, “사회“, “사건“, “가능“, “이번“, “상황“, “사용“, “내용“, “이용“, “방법“, “운영“, “조사“, “관리“, “마음“, “결과“, “본인“

〈표 6〉 전처리 전과 후의 데이터 변화

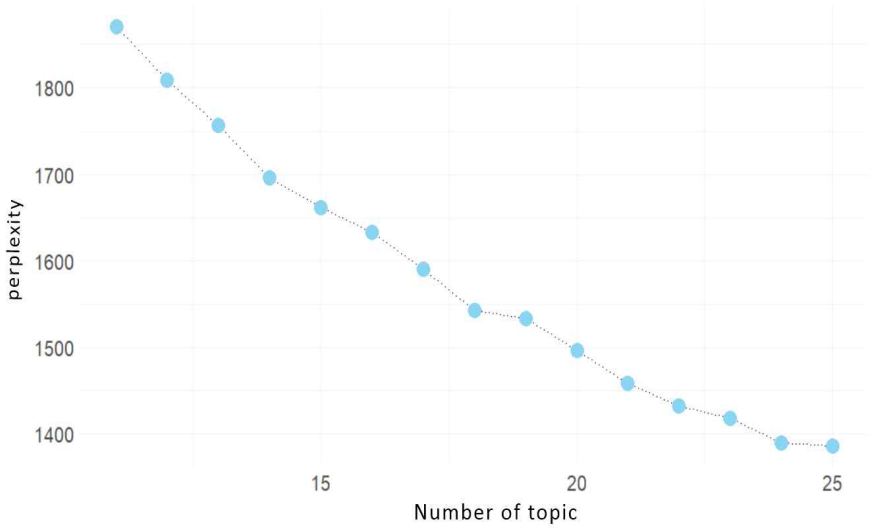
전처리 전	안녕하세요? 존경하고 지지하는 문재인 대통령님!\n저는 성남시 분당구 정자동 주택전시관 입점 업체의 임차인 입니다.\n주택전시관 한 업체로써, 절절한 심정을 호소합니다.이들이 숨기는게 없고 떳떳하다면 정당하게 자료공개를 했을것입니다.\n하지만 자꾸 관련도 없는 지자체에 민원을 이송한다는것은 크나큰 잘못과 책임회피라고 생각이 듭니다.\n부디 철저한 조사와 감사를 통해 힘든 상인과 국민에게 힘이 되어주시기를 간절히 부탁드립니다~~~~~~~(이하 생략)
전처리 후	존경 지지 주택 전시관 입점 업체 임차인 주택 전시관 업체 심정 호소 강제 철거 업체 계약 관계 사전 통보 피눈물 외면 만연 직권 남용 고발 철저 감사 업체 업체 불법 전대 합법 이행 업체 보상 불법 이행 업체 보상 추측 감사 업체 동안 무상 부장 당시 주택 전시관 책임자 누락 임대료 청구 묵인 직권 남용 사료 철저 감사....(이하 생략)

전처리 수행 후 형태소 분석을 통해 명사만 추출하였으며, 전체 문서에서 5회 미만 출현 단어는 분석에서 제외하였다. 또한, 문서에서 출현 빈도가 빈번하나 토픽 형성 시 변별력 저하를 유발하는 단어를 선별하여 불

용어로 선정 후 분석에서 제외하였다. <표 5>는 문서 내에서 제외할 단어인 불용어 목록이며 <표 6>은 전처리 수행 전후 데이터의 변화를 보여준다.

3.1.3 토픽 모델링

전처리가 완료된 데이터를 이용하여 최적의 토픽 수(K)를 찾기 위해 11에서 25까지 K의 수에 따른 혼란도(perplexity) 지표를 살펴보았다. 원 데이터는 17개의 범주로 구성되어 있으나 각 범주에 포함되어 있는 데이터가 소수의 토픽으로 결합될 가능성과 세부 토픽으로 분류될 가능성을 고려하여 토픽 수 K의 범위를 11~25로 선정하였다.



<그림 5> 토픽 수에 대한 perplexity

<그림 5>에서 보는 바와 같이 토픽의 수가 25개일 때 혼란도가 가장 낮다. 그러나 추가적인 데이터의 입력 시 과적합 문제를 고려하여 최종 토픽의 수를 23개로 선정하였다.

한편, 본 연구에서는 토픽 모델링을 위해 topicmodels 패키지에서 제공하는 LDA 함수를 사용하였다. $\alpha = 0.05$, $\beta = 0.05$, $K = 23$ 을 함수의 파라

미터로 설정한 후, ϕ, θ을 추론하기 위해 collapsed Gibbs sampling 방법을 이용하였다. 모수를 추론하기 위해 반복 수(iteration)는 1,000으로 설정하였다.

3.2 데이터 분석 결과

3.2.1 LDA 모형에 의해 추출된 핵심어

먼저, 토픽 간 변별력 있는 핵심어가 추출되는지를 검토하기 위해 LDA 모형을 이용하여 토픽에서 상위 비중($\phi_{k,d}$)을 차지하는 단어를 살펴보았다. <표 7>은 토픽 1과 토픽 14에서 상위 비중을 차지하는 15개의 단어이다. 토픽 1과 토픽 14의 경우, 범죄와 관련된 동일 단어들이 상위 비중을 차지하고 있어 토픽 간 변별력을 제공하는 핵심어가 추출되지 못하는 것으로 나타났다.

<표 7> 토픽 1과 토픽 14에서 상위 비중 단어 비교

토픽 1	토픽 14
경찰	**경찰**
범죄	**조직**
처벌	불법
피해자	사찰
청소년	**범죄**
부패	살인마
폭행	부정부패
가해자	빨갱이
폭력	경찰청
인권	신고
살인	경찰서
범죄자	감시
조직	**살인**

<표 8>은 토픽 5과 토픽 13에서 상위 비중을 차지하는 15개의 단어를 보여준다. 앞의 예에서와 마찬가지로 부모, 어머니, 아버지, 엄마 등 동일 의미의 단어들이 상위 비중을 차지해 토픽 간 변별력 있는 핵심어 추출이 어려웠으며, 그 이외에 다른 토픽들에서도 주제를 파악하기 위한 구체적인 핵심어가 추출되지 못하는 것으로 나타났다.

<표 8> 토픽 5와 토픽 13에서 상위 비중 단어 비교

토픽 5	토픽 13
장애	**아이**
가족	**부모**
아버지	어린이집
어머니	교사
정신	유치원
전화	지원
나이	**엄마**
친구	출산
동생	가정
아들	**자녀**
부모	보육
얘기	결혼
이야기	**아동**

3.2.2 relevance 척도와 skip-gram 기법을 이용한 핵심어 추출

토픽 간 변별력 있는 핵심어 집합을 구성하기 위해 본 연구에서는 relevance 척도와 skip-gram 기법을 이용하여 토픽 내 핵심어를 추출하였다. relevance 척도를 이용하여 임의의 단어가 다른 토픽에서 차지하는 비중만큼 패널티를 부여하였으며, 각 토픽에서 relevance 값 기준 10개의 상위 단어를 1차 핵심어로 선정하고 해당 단어를 포함한 문서들을 추출하였다. 그 결과, 총 8,218개의 문서가 추출되었으며 추출된 문서들의 고유 단어 집합을 생성해 skip-gram 모델의 입력 데이터로 사용하였다.

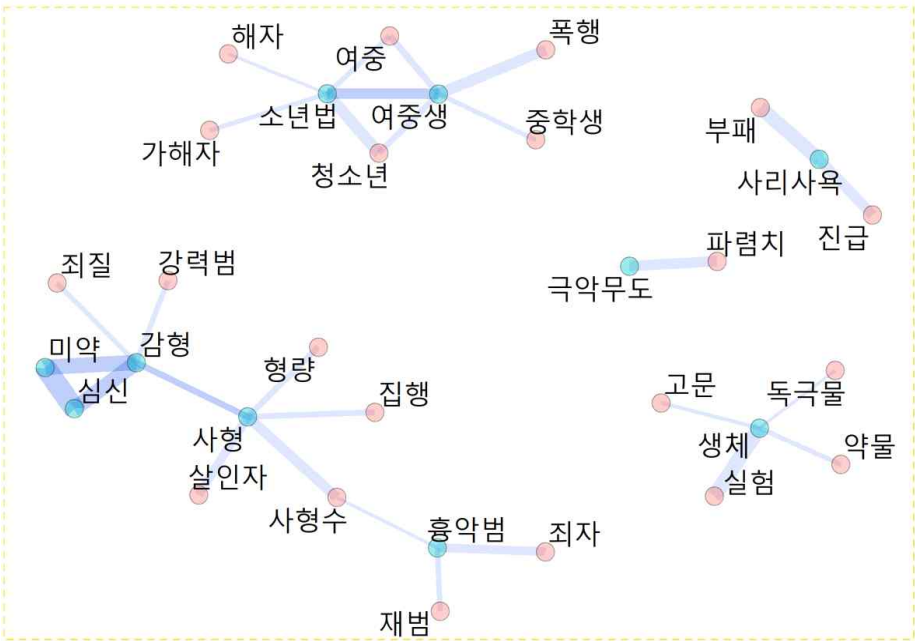
모델을 학습하기 위하여 gensim 모듈을 이용하였고, 학습 파라미터 갱신을 위한 반복 epoch 100회, 임베딩 차원 수 50, window-size 5, 네거티

브 샘플 수 15를 적용하였다. 또한, 각 토픽의 일차 핵심어와 임의의 단어 간 유사성을 측정하기 위해 수식 (8)의 코사인 유사도를 사용하여 핵심어를 재구성하였다.

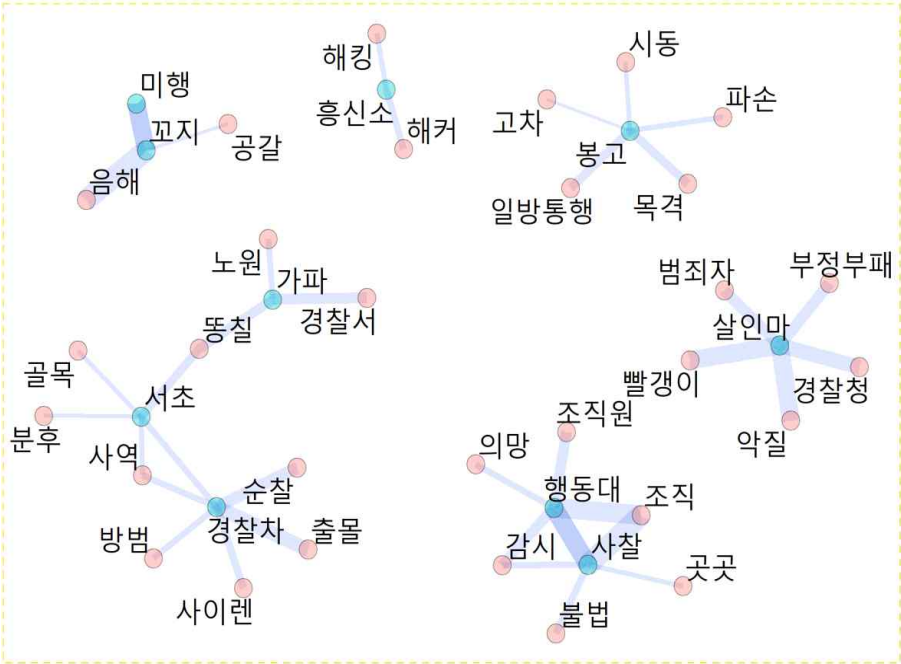
$$\cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum(x_i \times y_i)}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \quad (8)$$

□ 핵심어 재구성

<그림 6>과 <그림 7>은 3.2.1절에서 토픽 간 핵심어의 변별력이 미흡한 예로 제시했던 토픽 1과 토픽 14의 핵심어를 재구성하고 의미적 유사성을 시각화한 것이다. 그래프에서 파란색 노드는 일차 핵심어를 뜻하며 간선의 두께는 의미적 유사도를 나타낸다. 토픽 1의 경우, 기존 토픽 내 상위 비중인 ϕ 척도에서 관찰되지 않았던 강력범, 살인자, 독극물, 약물, 중학생 등의 단어가 관측되었다. 토픽 14의 경우에서도 미행, 행동대, 순찰, 사찰, 방범, 사이렌 등의 단어들이 관측되었다.



<그림 6> 토픽 1의 핵심어-연관어 그래프



〈그림 7〉 토픽 14의 핵심어-연관어 그래프

〈표 9〉 토픽 1과 토픽 14의 핵심어 변화 비교

토픽 1		토픽 14	
ϕ	$r + sg$	ϕ	$r + sg$
경찰	**소년법**	경찰	**사찰**
범죄	**감형**	조직	**경찰차**
처벌	**미약**	불법	**꼬지**
피해자	**심신**	사찰	**행동대**
청소년	**사리사욕**	범죄	**미행**
부패	**여중생**	살인마	**고차**
폭행	**사형**	부정부패	**노원**
가해자	**사형수**	빨갱이	**해커**
폭력	**음주**	경찰청	**조직원**
인권	**폭행**	신고	**방범**
살인	**출소**	경찰서	**칼질**
범죄자	**고문**	감시	**해킹**
조직	**강력범**	살인	**사역**

<표 9>와 <표 10>은 LDA 모형의 토픽-단어 상위 비중인 ϕ 척도 기반의 핵심어와 relevance 척도와 skip-gram 기법을 적용하여 재구성한 (r + sg) 척도 기반의 핵심어를 비교한 것이다. 2개의 표에서 볼 수 있는 바와 같이 ϕ 척도 기반의 핵심어에 비해 재구성된 핵심어 집합이 해당 토픽의 구체적인 내용을 담고 있었으며, 각 토픽이 어떠한 주제를 다루고 있는지 비교적 명확하게 파악할 수 있다.

토픽 1의 경우는 '강력 범죄'에 관한 주제를 다루고 있었으며 토픽 14의 경우에는 '방범'에 관한 주제를 다루고 있다고 할 수 있다. 토픽 5와 토픽 13의 경우도 ϕ 척도 기반의 핵심어로는 변별력을 확보하기 어려웠으나 재구성한 핵심어를 기반으로 살펴본 결과, 토픽 5는 '가족 경조사', 토픽 13은 '유아 교육'에 관한 주제를 다루고 있음을 알 수 있다.

<표 10> 토픽 5와 토픽 13의 핵심어 변화 비교

토픽5		토픽 13	
ϕ	r + sg	ϕ	r + sg
장애	**어머니**	아이	**어린이집**
가족	**동생**	부모	**유치원**
아버지	**아버지**	어린이집	**보육**
어머니	**언니**	교사	**양육**
정신	**오빠**	유치원	**아이**
전화	**아버님**	지원	**유아**
나이	**남동생**	엄마	**폐원**
친구	**올케**	출산	**양육자**
동생	**누나**	가정	**시험관**
아들	**부조금**	자녀	**아기**
부모	**장인어른**	보육	**육아**
애기	**친족**	결혼	**하원**
이야기	**장례식**	아동	**병설**

□ 토픽들의 핵심어 비교

<표 11>은 전체 토픽들의 ϕ 척도 기반의 핵심어와 relevance 척도와 skip-gram 기법을 적용하여 재구성한 (r + sg) 척도 기반의 핵심어를 비교한 것이다. 토픽 1은 강력범죄 재판에 관한 청원이며, 토픽 2는 채무, 토

픽 3은 온라인 화폐거래, 토픽 4는 종교재판에 관한 청원임을 알 수 있다.

또한, 토픽 5와 13의 경우에는 동일 단어들이 상위 비중을 차지해 변별력 있는 핵심어 추출이 어려웠으나 재구성한 핵심어를 기반으로 토픽 5는 가족 경조사, 토픽 13은 유아 교육에 관한 청원임을 알 수 있었다. 토픽 20의 경우에는 상위 비중의 단어들을 살펴보았을 때 자녀 교육에 관한 청원임을 알 수 있었으나 재구성한 핵심어를 살펴본 결과 대학 입시에 관한 핵심어들이 관측되어 구체적인 주제를 파악할 수 있었다.

토픽 22의 경우에도 ϕ 척도 기반의 핵심어를 살펴보았을 때 성별에 따른 차별과 범죄에 관한 청원임을 알 수 있었으나 구체적인 내용을 파악할 수 없었다. 그러나 핵심어를 재구성한 결과 '몰래카메라', '꽃뱀', '페미니스트', '노르딕' 등의 단어를 통해 해당 토픽에서 다루는 구체적인 내용을 파악할 수 있었다.

〈표 11〉 토픽들의 핵심어 비교

ϕ	$r + sg$	ϕ	$r + sg$	ϕ	$r + sg$
토픽 1		토픽 2		토픽 3	
경찰	**소년법**	회사	**연체**	화폐	**화폐**
범죄	**감형**	업체	**신용**	시장	**주식**
처벌	**미약**	판매	**카드**	가상	**가상**
피해자	**심신**	전화	**쇼핑**	주식	**공매도**
청소년	**사리사욕**	고객	**불량자**	투자	**거래소**
부패	**여중생**	카드	**고객**	거래	**코인**
폭행	**사형**	신고	**통화료**	금융	**투자자**
가해자	**사형수**	기업	**환불**	기업	**암호**
폭력	**음주**	금액	**약정**	거래소	**블록체인**
인권	**폭행**	제품	**타사**	공매도	**대주주**
살인	**출소**	영업	**판매**	투자자	**증권사**
범죄자	**고문**	구매	**채무자**	회사	**분식**
조직	**강력범**	수수료	**수수료**	규제	**채굴**
토픽 4		토픽5		토픽 6	
수사	**판사**	장애	**어머니**	임금	**근로자**
판결	**판결**	가족	**동생**	근무	**임금**
검찰	**대법관**	아버지	**아버지**	최저	**시급**
판사	**재판**	어머니	**언니**	근로자	**정규직**
법원	**대법원장**	정신	**오빠**	기업	**비정규직**
처벌	**법관**	전화	**아버님**	일자리	**최저**

재판	**기소**	나이	**남동생**	회사	**계약직**
검사	**사법부**	친구	**올케**	고용	**잔업**
변호사	**판검사**	동생	**누나**	근로	**근로**
헌법	**대법원**	아들	**부조금**	직원	**노동자**
경찰	**소법**	부모	**장인어른**	노동자	**고용**
비리	**신천지**	얘기	**친족**	청년	**임금법**
사법	**법원**	이야기	**장례식**	급여	**용역**
토픽 7		토픽 8		토픽 9	
국회의원	**난민**	주택	**집값**	선수	**선수**
정치	**국회의원**	부동산	**부동산**	평화	**축구**
난민	**선거**	서민	**청약**	역사	**출전**
선거	**체류**	아파트	**임대**	통일	**단일팀**
국회	**비자**	정책	**무주택**	대표	**월드컵**
의원	**체자**	대출	**무주택자**	전쟁	**무장**
외국인	**자국민**	임대	**주택**	세계	**아이스하키**
투표	**외노**	집값	**세대주**	축구	**남북한**
정권	**조선족**	분양	**디딤돌**	민족	**평화**
세금	**소환제**	가격	**신혼**	경기	**핵무기**
적폐	**개헌**	세금	**폭등**	회담	**스키**
반대	**유권자**	투기	**전세가**	스포츠	**선수단**
자유	**총선**	소득	**세대원**	국제	**메달**
토픽 10		토픽 11		토픽 12	
연금	**연금**	병원	**환자**	정책	**기상청**
공무원	**기회균등**	보험	**진료**	경제	**사업단**
세금	**고갈**	환자	**병원**	기업	**북극**
일반	**다복**	의료	**수술**	기술	**수출**
행정	**유익**	의사	**의료**	개발	**창업**
제도	**기도문**	치료	**의료인**	지원	**반도체**
소득	**폐직**	수술	**간호사**	사업	**지표**
복지	**노령**	건강	**치료**	발전	**분석**
건강	**수령자**	진료	**병동**	일자리	**총체**
지원	**호세**	질병	**경화증**	산업	**글로벌**
기초	**수급**	간호사	**의료진**	세계	**일맥상통**
당장	**입법권**	사고	**국소**	제도	**전망**
지급	**수령**	기록	**피보험자**	미래	**경제**
토픽 13		토픽 14		토픽 15	
아이	**어린이집**	경찰	**사찰**	공사	**건축**
부모	**유치원**	조직	**경찰차**	지역	**건축물**
어린이집	**보육**	불법	**꼬지**	주민	**준공**
교사	**양육**	사찰	**행동대**	건물	**건축법**
유치원	**아이**	범죄	**미행**	아파트	**건축주**
지원	**유아**	살인마	**고차**	허가	**서안**
엄마	**폐원**	부정부패	**노원**	계약	**콘크리트**
출산	**양육자**	빨갱이	**해커**	건설	**허가**
가정	**시험관**	경찰청	**조직원**	사업	**공사**
자녀	**아기**	신고	**방범**	분양	**축물**

보육	**육아**	경찰서	**칼질**	건축	**수분**
결혼	**하원**	감시	**해킹**	도시	**살균기**
아동	**병설**	살인	**사역**	개발	**경지**
토픽 16		토픽 17		토픽 18	
시험	**복무**	담배	**흡연**	민원	**민원인**
의무	**병역**	소음	**에어컨**	공원	**공원**
군대	**예비군**	전기	**금연**	답변	**효력**
복무	**징병제**	환경	**층간**	처리	**상방**
병역	**입대**	설치	**흡연자**	결정	**회신**
군인	**국방**	원전	**소음**	계획	**지법**
국방	**현역**	흡연	**누진세**	민원인	**귀하**
제도	**징병**	쓰레기	**더위**	신청	**시과**
훈련	**군복**	건강	**공해**	기관	**도시공원법**
기간	**입영**	발전소	**계류장**	접수	**민원**
양심	**예비역**	공항	**연료**	고시	**이송**
거부	**모병**	항공기	**비흡연자**	이전	**사고자**
공무원	**민방위**	공기	**방사능**	국토	**투숙**
토픽 19		토픽 20		토픽 21	
차량	**운전자**	학생	**학생**	업무	**포상**
사고	**차량**	학교	**수능**	기관	**국기원**
안전	**보행자**	교육	**수시**	공무원	**태권**
운전	**운행**	대학	**입시**	위원회	**태권도**
버스	**버스**	동물	**학교**	감사	**정무직**
주차	**번호판**	교사	**고등학교**	직원	**집행부**
기사	**화물차**	공부	**학교장**	규정	**근정**
택시	**횡단**	선생	**사교육**	관련	**지침**
도로	**사납금**	초등	**정시**	은행	**하위직**
자동차	**좌회전**	수업	**학기**	협회	**단증**
단속	**정차**	고등학교	**재학**	장관	**청조**
교통	**시내버스**	학년	**학년**	단체	**상훈**
설치	**운전**	시험	**대입**	인사	**기획부**
토픽 22		토픽 23			
여성	**여성**	방송	**지상파**		
남성	**남성**	언론	**방송**		
남자	**페미니스트**	게임	**시청자**		
여자	**페미니즘**	사이트	**컨텐츠**		
인권	**남녀**	기사	**만화**		
평등	**꽃뱀**	뉴스	**방송사**		
차별	**몰래카메라**	인터넷	**안테나**		
처벌	**우월주의**	댓글	**케이블**		
가족	**동성애자**	동의	**애니메이션**		
피해자	**무고죄**	정보	**댓글**		
성범죄	**성범죄**	보도	**포르노**		
남녀	**노르딕**	삭제	**팝콘**		
성매매	**성별**	조작	**유행어**		

<표 11>과 같이 재구성된 핵심어를 기반으로 각 토픽이 다루는 주제를 파악하고, 각각의 토픽명을 붙일 수 있다. <표 12>는 각 토픽의 이름과 재구성된 핵심어를 나타낸 것이다.

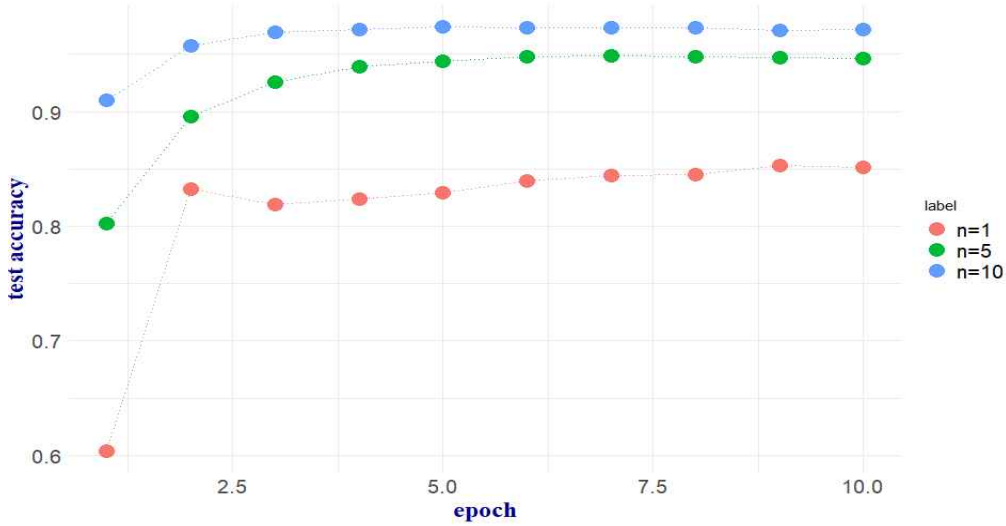
<표 12> 각 토픽의 이름 및 핵심어

토픽 번호	토픽명	핵심어
1	**강력 범죄**	소년법, 감형, 사형, 사형수, 음주, 폭행, 독극물 등
2	**채무**	신용, 불량자, 채무자, 파산, 수수료, 위약금 등
3	**mts**	화폐, 주식, 가상, 코인, 블록체인, 채굴 등
4	**종교 재판**	교주, 신천지, 사이비, 판사, 판결, 대법원장, 이단 등
5	**가족 경조사**	장례식, 장인어른, 부조금, 애인, 아내, 중환자실 등
6	**근로법**	근로자, 임금, 시급, 정규작, 임금법, 용역 등
7	**외국인 체류**	난민, 체류, 비자, 체자, 자국민, 조선족, 외노 등
8	**부동산**	집값, 임대, 무주택, 주택, 세대주, 폭등, 전세가 등
9	**남북 외교**	단일팀, 남북한, 월드컵, 평화, 핵무기, 선수단 등
10	**연금**	연금, 기회균등, 고갈, 기도문, 노령, 수령자, 수급 등
11	**보건 · 복지**	환자, 진료, 병원, 의료인, 간호사, 병동, 경화증 등
12	**성장동력**	북극, 수출, 기상청, 반도체, 지표, 글로벌, 경제, 전망 등
13	**유아 교육**	어린이집, 유치원, 보육, 양육자, 육아, 하원 등
14	**방범**	행동대, 미행, 방범, 사이렌, 순찰, 경찰차 등
15	**건축**	콘크리트, 건축법, 건축주, 공사, 축물, 경지 등
16	**군복무**	병역, 예비군, 입대, 국방, 현역, 징병, 군복, 민방위 등
17	**이웃갈등**	흡연, 에어컨, 층간, 누진세, 더위, 공해, 비흡연자 등
18	**도시 환경**	도시공원법, 민원인, 공원, 효력, 지법, 시청 등
19	**교통 · 안전**	운전자, 차량, 은행, 버스, 화물차, 보행자 등
20	**대학 입시**	수능, 수시, 입시, 학교장, 사교육, 정시, 재학, 대입 등
21	**국기원 포상**	태권도, 정무직, 단증, 상훈, 국기원, 포상 등
22	**성 평등**	여성, 남성, 페미니즘, 꽃뱀, 몰래카메라, 우월주의 등
23	**방송**	지상파, 방송, 컨텐츠, 댓글, 유행어, 촬영장, 예능 등

3.2.3 토픽 분류

국민청원 데이터에 대한 토픽 분류를 수행하기 위해 본 연구에서는 KoBERT 모델을 이용하였다. 적절한 분류 성능에 도달하기 위한 epoch를 살펴본 결과 epoch 7 이후에는 미미한 변화를 보여 최종 epoch를 10으로 설정하였으며 분류 모델의 성능 개선을 위해 데이터 증강기법 EDA를 적용하였다. EDA의 유의어 교체, 임의 삽입, 임의 교체 등을 통해 변경되는 단어의 수는 $n = \max(1, \alpha \times l)$ 로 정해진다. l은 각 문장의 토큰 수이며, α는 한 문장에서 변경되는 토큰의 비율이다. 본 연구에서는 사전 연구를 참고하여 α값으로 0.1을 설정하였다(Wei, J., Zou, K. 2019).

각각의 데이터(증강하지 않은 데이터, 5배 증강한 데이터, 10배 증강한 데이터)를 입력 데이터로 사용하여 비교 분석을 진행하였으며, 훈련 데이터(training data)와 테스트 데이터(test data)를 8:2로 구성하였다. <그림 8>은 3개의 데이터 각각에 대한 분류 정확도를 나타낸 것이다.



<그림 8> 분류 정확도

데이터를 5배, 10배 증강시킨 경우 각각의 예측 정확도는 94%, 97%를 보여(epoch를 10으로 설정한 경우) 데이터를 10배 증강시킨 경우의 정확도가 가장 높게 나타났다. 이는 데이터를 증강하지 않았을 경우보다 각각

7%, 12% 정도 높은 결과이다. 그러나 과적합 문제로 인해 새로운 문서가 적절치 못한 토픽으로 분류될 가능성을 고려해 94%의 예측 정확도를 보인 5배 증강한 데이터 모델을 최종 모델로 선정하였으며, 원 데이터와 비교를 위해 각 토픽에 대한 정밀도(precision)와 재현율(recall), F1 score를 살펴보았다.

<표 13>은 원 데이터와 5배 증강 시킨 데이터의 분류 성능을 비교한 것이다. 원 데이터의 경우 다수의 토픽에서 낮은 성능을 보였다. 특히 토픽 14의 경우 정밀도 0.33, 재현율 0.25, F1-score 0.28로 가장 낮은 성능을 보였다. 또한, 토픽 5, 토픽 18, 토픽 21의 경우에도 다른 토픽들에 비해 낮은 성능을 보였다. 데이터를 5배 증강시켜 학습한 모델의 경우, 토픽에 대한 정밀도, 재현율, F1 score 값이 토픽 14를 제외하고 0.9 이상으로 나타났다.

<표 13> 데이터 증강 전/후의 성능 비교

label	precision		recall		f1-score		n	
	증강전	**증강후**	증강전	**증강후**	증강전	**증강후**	증강전	**증강후**
토픽 1	0.921	0.955	0.884	0.958	0.902	0.957	204	937
토픽 2	0.832	0.938	0.792	0.933	0.811	0.936	146	702
토픽 3	0.920	0.951	0.858	0.944	0.887	0.947	132	657
토픽 4	0.797	0.944	0.853	0.927	0.824	0.935	162	912
토픽 5	**0.604**	**0.920**	**0.608**	**0.914**	**0.606**	**0.917**	**122**	**653**
토픽 6	0.897	0.947	0.927	0.949	0.912	0.948	214	946
토픽 7	0.823	0.932	0.849	0.946	0.836	0.939	279	1413
토픽 8	0.928	0.951	0.964	0.975	0.946	0.963	162	905
토픽 9	0.912	0.961	0.902	0.968	0.907	0.965	237	1138
토픽 10	0.836	0.966	0.895	0.938	0.864	0.952	112	628
토픽 11	0.842	0.958	0.878	0.956	0.860	0.957	63	410
토픽 12	0.734	0.920	0.659	0.945	0.694	0.932	127	641
토픽 13	0.841	0.940	0.819	0.946	0.830	0.943	139	666
토픽 14	**0.331**	**0.810**	**0.258**	**0.931**	**0.288**	**0.866**	**19**	**87**
토픽 15	0.749	0.967	0.785	0.951	0.767	0.959	73	426
토픽 16	0.894	0.949	0.814	0.947	0.852	0.948	119	584
토픽 17	0.891	0.961	0.833	0.939	0.861	0.950	150	740
토픽 18	**0.700**	**0.930**	**0.683**	**0.930**	**0.692**	**0.930**	**36**	**143**
토픽 19	0.886	0.939	0.934	0.956	0.909	0.947	145	679
토픽 20	0.919	0.974	0.914	0.966	0.916	0.970	183	843
토픽 21	**0.525**	**0.902**	**0.584**	**0.916**	**0.553**	**0.909**	**64**	**333**
토픽 22	0.857	0.964	0.825	0.940	0.841	0.952	142	681
토픽 23	0.803	0.950	0.857	0.938	0.829	0.944	169	869

토픽 14의 경우, 정밀도 0.81, 재현율 0.93, F1-score 0.87로 나타나 원 데이터 보다는 높은 성능을 보였으나, 다른 토픽들과 비교하면 상대적으로 낮게 나타났다. 이에 대한 원인을 파악하기 위해 데이터를 검토해 본 결과, 토픽 14에 해당하는 데이터의 수가 전체 데이터의 0.5% (전체 15,993개의 데이터 중 87개의 데이터) 정도의 적은 수였기 때문인 것으로 판단된다.

제 4 장 결론

본 논문에서는 깁스 샘플링을 이용한 lda 토픽 모델링을 수행하여 수집된 문서 집합을 23개의 토픽으로 분류하였다. 또한, relevance 척도와 워드 임베딩 기법을 적용하여 토픽 간 변별력 있는 핵심어 집합을 재구성하였으며 재구성한 핵심어를 기반으로 각 토픽의 이름을 명명해 해당 토픽의 주제를 직관적으로 파악하였다.

추가적인 데이터가 입력되었을 때 알맞은 토픽으로 분류하기 위해 KoBERT 모델을 제안하였으며 제안한 모델의 성능을 개선하기 위해 EDA 기법을 적용하였다. EDA 기법을 적용하여 5배 증강한 데이터로 학습한 KoBERT 모델을 최종 분류 모델로 선정 후 성능을 증강 전 데이터로 학습한 모델과 비교한 결과 약 9% 높은 정확도를 보였으며 정밀도, 재현율, F1-score 값을 통해 성능 개선에 대한 구체적인 근거를 확보하였다.

본 연구를 고객 상담 이력 데이터에 적용 시 주제에 따른 고객 분류가 가능하며 만족도 변수를 추가할 경우 만족도에 기인하는 구체적 요인을 파악할 수 있을 것으로 기대된다. 또한, 잠재 고객들을 자동 분류하여 잠재 고객들에게 맞춤형 서비스를 제공할 수 있다.

본 연구의 한계점은 다음과 같다. 첫 번째 모델 적합 시 문서 집합에서 명사만을 대상으로 모델링을 하였다. ‘나쁜’, ‘착한’ 등의 감정적 의미의 형용사들은 해당 문서의 성격을 어느 정도 표현할 수 있기에 정보의 누락 문제가 의심된다. 두 번째 단어에 대한 분포 고려 시 단일 단어 분포만 고려하였다. 예를 들어 ‘대학 입시’라는 한 가지 단어가 존재할 경우 ‘대학’과 ‘입시’ 두 가지 단어로 모델을 적합해 정보의 왜곡이 발생할 수 있다.

향후 연구에서 명사뿐 아니라 형용사, 부사 등 다양한 품사를 적용해 토픽을 형성한다면 각 토픽에서 전달하고자 하는 정보를 명확하게 전달할 수 있을 것이다. 또한, 단어에 대한 분포 고려 시 bigram, trigram 등과 같이 결합분포를 고려한다면 본 연구에서 제안한 방법보다 토픽 간 변별력 있는 핵심어를 추출할 수 있을 것이며 언급한 문제점들을 해결할 수 있을 것이다.

참고문헌

[1] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137-1155.

[2] Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022.

[3] Blei, D.M., and Lafferty, J.D. (2006). Dynamic topic models, Proceedings of the 23rd international conference on Machine learning, 113-120.

[4] Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. IEEE signal processing magazine, 27(6), 55-65.

[5] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[6] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186.

[7] Mimno, D.M. & McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression, Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence,

411-418.

[8] Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

[9] Lund, K., Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical cooccurrence. Behavior Research Methods, Instruments Computers, 28, 203-208

[10] Redondo, T., Sandoval, A.M. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence, International Journal Of Interactive Multimedia And Artificial Intelligence, 3, 57-64

[11] Roberts, M.E., Stewart, B.M., Airoldi, E.M. (2016). A Model of Text for Experimentation in the Social Sciences, Journal of the American Statistical Association, 111, 1-49.

[12] Sievert, C., Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces. 63-70.

[13] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. https://doi.org/10.48550/arXiv.1301.3781

[14] Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S. (2010). Recurrent neural network based language model. Proceedings of the Annual Conference of the International Speech Communication Association, 2703-2707.

[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Proceedings of the International Conference on Neural Information Processing Systems, 2, 3111-3119.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. Polosukhin, I. (2017). Attention is All you Need. Proceedings of the Annual Conference on Neural Information Processing Systems, 5998-6008.

[17] Wei, J., Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 6382-6388.

[18] 강형석, 양장훈 (2019). 한국어 단어 임베딩을 위한 Word2vec 모델의 최적화, 디지털콘텐츠학회논문지, 20(4), 825-833.

[19] 김성근, 조혁준, 강주영 (2016). 학술연구에서의 텍스트 마이닝 활용 현황 및 주요분석기법. 정보화연구, 13(2), 317-329.

[20] 김은회, 서유화 (2020). 토픽 레이블링을 위한 토픽 키워드 산출 방법. 디지털산업정보학회논문지, 16(3), 25-36.

[21] 남승주, 최솔샘, 김준환, 김진기 (2020). 공항 산업 동향 분석을 위한 텍스트 애널리틱스 모델에 관한 연구. 경영과학, 37(1), 61-74.

[22] 한지영, 허고은 (2021). 토픽 모델링 기반 비대면 강의평 분석 및 딥러닝 분류 모델 개발. 한국문헌정보학회지, 55(4), 267-291.

[23] 황상흠, 김도현 (2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델, The Journal of Society for e-Business Studies, 25(1), 203-214.

국문초록

핵심어 추출 및 데이터 증강기법을 이용한 텍스트 분류 모델 성능 개선

이강철
전북대학교 통계학과

토픽 모델(topic model)은 문서와 단어로 구성된 행렬(document-term matrix, DTM)을 기반으로 문서에 잠재되어 있는 토픽의 출현확률을 추정하는 기법이다. 이 기법은 문서-토픽(documents-topics) 비중과 토픽-단어(topics-terms) 비중을 통하여 각 문서의 핵심 토픽과 각 토픽의 특성을 직관적으로 파악할 수 있다는 장점이 있다. 그러나 서로 다른 토픽에서 동일 단어가 상위 비중을 차지하는 경우, 토픽 간 변별력이 있는 핵심어(keywords) 추출이 어렵다는 문제점이 있다. 또한, 이 기법은 단어의 출현 빈도수에 기반한 방법이기 때문에 핵심어와 의미적 유사성이 있으나 핵심어로 채택되지 못한 단어들이 존재하는 경우 정보의 누락이 발생한다. 이러한 문제점을 개선하기 위하여 본 연구에서는 핵심어를 추출할 때 연관성 척도(relevance)와 워드 임베딩(word embedding) 기법을 적용하는 방법을 제안한다. 또한, 분류 성능을 개선하기 위해 EDA(Easy Data Augmentation) 기법을 이용하여 데이터를 5배 증강한 후 KoBERT 모델을 적용하여 데이터를 분류하였다. 데이터 분석 결과, 토픽 간 변별력 있는 핵심어를 추출하여 해당 토픽의 구체적인 내용을 파악할 수 있었으며, 94% 정확한 분류 결과를 얻어 데이터 증강기법을 적용하지 않은 경우에 비해 9% 정도 개선된 결과를 얻을 수 있었다.

주요어 : 토픽 모델, 연관성 척도, 워드 임베딩, 데이터 증강, 텍스트 분류