KIPS Trans. Softw. and Data Eng.
Vol.9, No.2 pp.45~52 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.2.45

Topic Analysis of the National Petition Site and Prediction of Answerable Petitions Based on Deep Learning

Woo Yun Hui† · Hyon Hee Kim††

ABSTRACT

Since the opening of the national petition site, it has attracted much attention. In this paper, we perform topic analysis of the national petition site and propose a prediction model for answerable petitions based on deep learning. First, 1,500 petitions are collected, topics are extracted based on the petitions' contents. Main subjects are defined using K-means clustering algorithm, and detailed subjects are defined using topic modeling of petitions belonging to the main subjects. Also, long short-term memory (LSTM) is used for prediction of answerable petitions. Not only title and contents but also categories, length of text, and ratio of part of speech such as noun, adjective, adverb, verb are also used for the proposed model. Our experimental results show that the type 2 model using other features such as ratio of part of speech, length of text, and categories outperforms the type 1 model without other features.

Keywords : National Petition, Topic Analysis, Topic Modeling, K-means Clustering, LSTM, Deep Learning

국민청원 주제 분석 및 딥러닝 기반 답변 가능 청원 예측

우 윤 희† · 김 현 희††

요 약

청와대 국민 청원 사이트가 개설된 이래로 많은 관심을 받고 있다. 본 논문에서는 국민 청원의 주제를 분석하고 딥러닝을 활용하여 답변 가능한 청원을 예측하는 모델을 제안하였다. 먼저, 추천순으로 1,500개의 청원글을 수집하였고, K-means 클러스터링을 적용하여 청원글을 군집하여 대주제를 정의하고, 보다 구체적인 세부 주제를 정의하기 위하여 토픽 모델링을 실시하였다. 다음으로는 LSTM을 활용한 답변 가능한 청원 예측 모델을 생성하여, 20만의 청원동의를 얻는 청원을 예측하기 위한 모델을 개발하였다. 이를 위해 글의 주제와 본문뿐만 아니라 글의 길이, 카테고리, 특정 품사의 비율이 영향을 미칠 수 있는지를 살펴보았다. 그 결과, 본문과 함께 글의 길이, 카테고리, 체언, 용언, 독립언, 수식언의 품사의 비율을 변수로 추가한 모델의 f1-score가 0.9 이상으로 글의 제목과 본문을 변수로 하는 모델보다 예측력이 높음을 알 수 있었다.

키워드 : 국민청원, 주제 분석, 토픽 모델링, K-means 클러스터링, LSTM, 딥러닝

1. 서 론

청와대 국민청원[1] 제도는 2017년 8월 17일 현 정부 출범 100일을 맞이하여 신설되었으며, 30일 동안 20만 명 이상의 국민들이 추천한 청원에 대해 정부 및 청와대 관계자가 응답하는 시스템이다. 국민청원은 음주 운전 처벌 강화, 윤창호법, 심신미약 감경 의무 폐지, 김성수법 등 국민청원으로 필요한 법을 재정하는 순기능을 가져다주기도 했으나, 무분별한 글 게시와 중복되는 글, 비방의 글도 여과없이 올라오며 이에 따른 개편 필요성이 대두되었다. 이에 국민청원은 올해 3월 31일부로 100명의 사전 동의를 받아야 청원글이 공개되도록 폭력적, 선정적, 집단 혐오적 표현과 명예훼손 내용이 들어있는 청원들은 삭제되도록 개편되었다.

본 논문에서는 먼저 청원의 내용과 키워드를 바탕으로 어떤 주제와 이슈들이 있었는지 계층적으로 분석하고 실제 카테고리와 비교하였다. 다음으로 딥러닝을 기반으로 청원동의 수가 20만을 넘어 답변을 받는 청원을 예측하는 모델을 개발하였다.

계층적 주제를 분석하기 위해 2019년 3월 31일 기준, 추천 순으로 정렬된 청원 중 상위 1500개의 글에 k-means 클러스터링 알고리즘을 적용[3, 4]하여 대주제를 정의하였다.

※ 이 논문은 동덕여자대학교 교내연구비 지원에 의하여 수행된 것임.
※ 이 논문은 2019년도 한국정보처리학회 춘계학술발표대회에서 'K-means 클러스터링과 토픽 모델링을 기반으로 한 국민청원 사이트의 카테고리 재구성'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 동덕여자대학교 정보통계학과 학사과정
†† 정 회 원 : 동덕여자대학교 정보통계학과 부교수
Manuscript Received : July 17, 2019
First Revision : October 25, 2019
Accepted : November 20, 2019
* Corresponding Author : Hyon Hee Kim(heekim@dongduk.ac.kr)

대주제에 속하는 핵심단어를 탐색하고 세부 분류가 필요한 경우, 토픽 모델링[5]을 추가로 실시하여 세부적인 주제를 탐색, 계층적으로 나타내었다. 그 결과, 주제 및 핵심이슈는 다음과 같이 나타났다. '난민, 체육, 외교, 고용/노동, 금융/경제, 사이버범죄, 의료/보건, 보육/교육, 동물, 성/성평등, 폭행, 지역사회, 정치, 기타 ' 14개의 대주제를 선정하였고, '보육/교육'의 세부 주제로는 '보육'과 '교육', '성/성평등'의 세부분류로는 '성매매', '성평등', '낙태법', 마지막으로 '지역사회'는 '건설/교통'과 '주거'로 구성된다.

다음으로 국민청원의 답변가능 청원 예측에서는 Long Short-term Memory (LSTM)을 사용하여 제목과 본문만을 사용한 모델과 카테고리, 글의 길이, 의미있는 품사 사용 비율을 함께 변수로 추가해준 모델을 Adam와 Adadelta Optimizer를 사용한 경우로 나누어 각각 비교하였다. 실험 결과 제목과 본문에 카테고리, 글의 길이, 의미있는 품사 사용 비율을 함께 변수로 추가해주었을 때 더 예측이 뛰어나며 f1-score기준 0.9를 넘는 성능을 보였다는 것을 발견하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 데이터 수집과 정제 및 자연어 처리과정에 대해 서술하고 제 3장은 k-means 클러스터링을 적용하여 대주제를 선정하고, 토픽 모델링으로 세부 주제를 정의하는 과정에 대해 설명한다. 제 4장은 답변 가능한 청원 예측 모델에 대해 자세히 설명하고, 마지막으로 제 5장은 결론 및 향후 연구를 제시한다.

2. 자료 수집 및 전처리

2.1 자료 수집 및 전처리

먼저 카테고리 추출을 위해서 모든 크롤링에는 파이썬의 BeautifulSoup 라이브러리와 함께 Selenium 라이브러리, 도구로는 chrome 브라우저를 사용하였다. 국민 청원 사이트의 현재 카테고리인 "분야 종합"에서 만료된 청원을 추천 순으로 정렬한 뒤, 3월 31일 13:00를 기준으로 상위 1500개의 청원글로부터 제목, 내용, 카테고리, 그리고 게시일시를 뽑아냈다.

글을 토큰화하기 위해서는 파이썬을 이용한 형태소분석기 Konlpy[6]를 사용한다. Konlpy에는 Kkma, Hannanum, Okt, Komoran, 그리고 Mecab의 다섯 가지 형태소 분석기가 포함되어있다. 형태소 분석기는 글에 따라 성능이 다르므로 어느 분석기가 국민청원에 가장 적합한지 비교하는 과정을 거친다. 비교에는 1,500개 청원의 제목을 사용하는데, 이는 청원의 제목이야말로 청원내용을 잘 함축하고 있어 분석기의 성능을 빠르게 파악하기 가장 적합하다고 판단했기 때문이다 제목 중에서도 주제와 핵심 단어를 찾는 데 필요한 '일반 명사'와 '고유명사'를 어느 분석기가 가장 잘 추출하는지 비교한다.

본 논문에서는 윈도우환경을 사용하므로 윈도우 환경을 지원하지 않는 Mecab과 보통명사와 고유명사를 분리할 수 있는 기능을 가지고 있지 않은 Okt와 Hannanum은 제외하였다. 3개의 분석기를 제외한 2개의 형태소 분석기 Kkma와 Komoran에 1,500개의 제목 리스트를 넣어 보통명사와 고유명사를 출력한다. 출력 결과는 아래 그림과 같다. 비교 결과 Komoran의 성능이 더 좋다는 것을 확인할 수 있으며, Kkma는 Komoran보다 사람 이름 분석과 오타 등에 취약하였다.

분석기	고유명사 인식 오류	오타 인식 불가	외래어나 신조어
Kkma	평창 : 평 / 창 이수역 : 이수 / 역 표창원 :표 / 창원 유정호: 유정 / 호 등 NNP를 NNG로 분리	ㄱㅐ : 'ㄱ'(NNG) 'ㅐ'(UN)	버닝썬: '버닝'(NNG) '썰'(VV) 'ㄴ'(ETD) 드루킹: '드'(XPN)'루'(NNG) '킹'(NNG) 등
Komoran	유정호: 유 / 정호 경강선: 경 / 강선 등 NNP를 NNG로 분리		버닝썬: '버닝썬'(NA,인식불가) 드루킹: '드'(NNP)'루킹'(NNG) 등

Fig. 1. Comparison Between Kkma and Komoran

분석기로 Komoran을 사용하기로 결정한 후, 이를 이용하여 제목과 내용에서 일반명사와 고유명사를 뽑았다. 그 중 Komoran이 잘못 분류한 고유명사와 외래어, 신조어들을 직접 사전 추가하여 올바르게 분류해주고, 비교적 잘못 분류된 부분이 많은 1음절 단어들을 다시 검토하여 명사를 제외하고 모두 제거하였다. 그 후, 제목과 내용에서 뽑은 명사들을 공백으로 묶어 하나의 글처럼 생성하였다

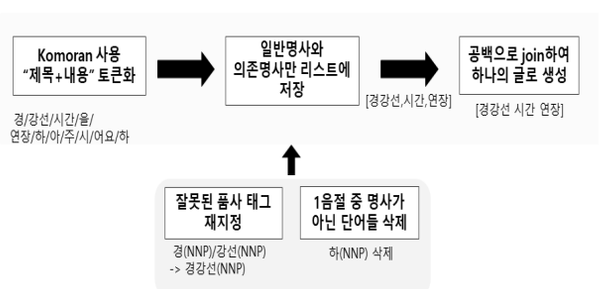

Fig. 2. Process of Preprocessing

2.2 TF-IDF를 통한 벡터화

Term Frequency-Inverse Document Frequency (TF-IDF)란 피처 벡터화 중 하나로, 문서에서 특정 단어가 얼마나 중요한 의미를 가지는지를 수치화하는 방법이다. 단순히 단어의 빈도 기반에서 한 단계 더 나아가, 중요도가 높은 단어의 가중치를 크게, 빈도수는 많지만 중요도는 낮은 단어의 가중치를 작게 보정해준다.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_{i,j}}) \quad (1)$$

TF-IDF 가중치는 수식 (1)과 같이 나타낼 수 있다. 문서 j에 속하는 키워드 i의 가중치 $w_{i,j}$는 tf값과 df값의 곱이다. 여기서 $tf_{i,j}$는 문서 j에 나타난 키워드 i의 빈도수이고, $df_{i,j}$

는 키워드 i를 포함하는 문서의 개수의 역수의 로그값이다. 키워드의 빈도수만을 고려하면 가장 빈도수가 높은 키워드들은 대부분 문서 집합에서 가장 일반적으로 사용되는 키워드들인 경우가 많다. 이러한 키워드들을 제거하고 더욱 구체적인 키워드들을 추출하기 위해서 키워드가 문서 집합에 등장한 횟수의 역수를 곱하면 문서 집합 전체에 많이 등장한 키워드들은 가중치 값이 줄어든다. 본 논문에서는 scikit learn[7]에서 제공하는 TfidfVectorizer 모듈을 사용하였다.

3. 계층적 주제 분석

3.1 주제 분석 프로세스

계층적 주제 분석 과정은 Fig. 3과 같다. k-평균 군집화를 적용하여 군집을 나눈 후 군집 별 키워드를 추출, 군집별로 키워드를 기반으로 Labeling이 가능한지의 여부를 살핀다. Labeling이 가능한 군집은 키워드를 기반으로 대주제를 설정하고 LDA를 실시하여 키워드에서 다시 Labeling이 가능하다면 세부주제를 설정, 아니라면 설정하지 않는다. k-평균 군집에서 핵심단어로 하나의 Labeling이 불가능한 군집은 LDA를 통한 키워드에서 Labeling이 가능한지의 여부를 따진다. 가능하다면 대주제를 설정한다.

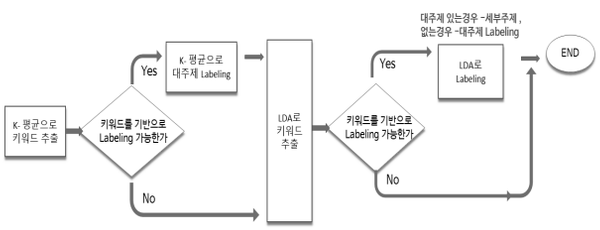

Fig. 3. Process of Finding Hierarchical Topics

3.2 대주제 정의

청원글의 주제를 파악하기 위해 먼저 TF-IDF로 생성한 벡터에 k-means 클러스터링 알고리즘을 사용하여 유사한 글끼리 그룹화한다. k-means 클러스터링 알고리즘은 k개의 군집 중심의 위치를 랜덤으로 선정한 뒤, 그 중심들을 기준으로 군집을 구성하고, 군집별 평균 위치로 중심 k를 이동시킨다. 이 과정은 k가 더 이상 이동하지 않을 때까지 계속된다. k-means 클러스터링 알고리즘은 단순하지만 성능이 비교적 좋으며, 구현이 쉬우나, 군집의 개수를 직접 설정해야 실시할 수 있다.

본 논문에서는 k-means 클러스터링 알고리즘의 구현을 위해 scikit-learn 라이브러리를 활용하였다. 또한 군집개수 설정을 위해, 군집개수를 8, 10, 12, 14개로 늘려가며 군집이 몇 개일 때 청원글이 가장 잘 분류되는지 실험을 실시하였으며 그 결과, 군집의 개수가 10일 때 청원글이 가장 효과적으로 분류되었다.

군집 개수가 10인 k-means 클러스터링을 실시한 후 군집별로 핵심 단어를 상위 10개씩 추출한 결과는 다음 Table 1과 같다. Table 1은 대주제를 정의하기 위해 k-means 클러스터링 알고리즘을 적용하고, TF-IDF 가중치가 높은 상위 10개의 키워드를 추출한 것이다.

cluster_0은 '난민, 이슬람, 선수, 제주도'의 난민에 관한 분야와 '선수, 빙상, 빙상연맹'의 체육 분야로 구성된다. 난민과 체육이 전혀 다른 주제라고 생각됨에도 같은 군집에 속하는 것은 난민문제와 체육 분야가 '외국인', '문화', '한국', '나라'와 같이 공통적인 단어들을 많이 사용하기 때문이다.

cluster_1은 '미세먼지, 중국, 일본, 우리나라'와 같은 미세먼지 및 외교 문제와 '기업, 회사, 근무, 노동자, 업체'와 같은 고용, 노동의 주제로 파악된다. 외교와 고용문제는 깊은 연관이 있다. 외교문제가 악화될 때 우리는 관련된 우리 기업의 타격을 우려한다.

Table 1. Main 10 Subjects and Keywords

n	Keywords
0	난민, 선수, 이슬람, 연맹, 빙상, 빙상연맹, 제주도, 외국인, 한국, 문화
1	미세먼지, 중국, 시간, 일본, 기업, 회사, 근문, 노동자, 업체, 우리나라
2	수사, 처벌, 공매도, 불법, 사건, 사이트, 판사, 범죄, 법, 주식
3	병원, 간호사, 환자, 의료, 치료, 의사, 수술, 보험, 아이, 병
4	교사, 어린이집, 보육, 학생, 교육, 원장, 시간, 아이들,보육교사, 학교
5	여성, 남성, 성매매, 성, 낙태, 지원, 여성가족, 평등, 사회, 인권
6	동물, 견, 강아지, 반려, 학대, 유기, 유기견, 반려동물, 보호소, 반려견
7	지역, 주민, 임대, 주택, 공공, 아파트, 신도시, 교통, 공사, 분양
8	가해자, 폭행, 피해자, 아이, 사건, 경찰, 처벌, 사람, 말, 생각
9	학생, 대통령, 학교, 국회의원, 정부, 반대, 국가, 교육, 사람, 의원

cluster_2는 '공매도, 주식, 사이트'의 단어들로 보아 금융, 화폐분야, 혹은 사이트에 관한 주제들로 파악된다. 금융 주제와 '사이트'라는 단어가 같은 군집에 속한 이유는 주식, 블록체인 등 금융의 많은 부분이 인터넷과 뗄 수 없는 관계이기 때문으로 파악된다.

cluster_3은 '병원, 간호사, 환자, 의료, 치료, 의사, 수술, 보험, 병'의 의료 관련 단어들로 구성된다. cluster_4는 '교사, 어린이집, 보육, 학생, 교육, 원장, 아이들, 보육교사, 학교'로 보아 보육과 교육 관련 주제이다. cluster_5는 '여성, 남성, 성매매, 성, 낙태, 지원, 여성가족, 평등, 사회, 인권'으로 성과 성평등에 관한 주제이다. cluster_6은 '동물, 견, 강아지, 반려, 학대, 유기, 유기견, 반려동물, 보호소, 반려견'으

로 동물에 관한 주제이다.

cluster_7은 '지역, 주민, 임대, 주택, 공공, 아파트, 신도시, 교통, 공사, 분양'의 단어들로 주거, 건축, 교통에 관한 주제들, 즉 지역사회 문제들로 구성된다. cluster_8은 '가해자, 폭행, 피해자, 사건, 경찰, 처벌'의 단어로 보았을 때 폭행사건과 관련된 주제이다. cluster_9는 '학생, 학교, 교육'의 교육과 관련된 분야와, '대통령, 국회의원, 정부, 국가, 의원'과 같은 정치에 관한 분야로 파악된다.

3.3 세부 주제 분석

k-means 클러스터링을 실시한 것 중 3.1절에서 정의된 주제가 명확하고 추가 분석이 필요없는 cluser_3의 의료, cluster_6의 동물, cluster_8의 폭행을 제외한 7개 군집에 토픽 모델링을 적용하여 군집별로 10개의 단어를 뽑아내었다. 토픽 모델링으로는 Latent Dirichlet Allocation (LDA)를 활용하였다. Table 2는 LDA를 활용하여 정의한 세부 주제를 나타낸다.

Table 2. Detailed Subjects Based on LDA

n	Topic	Keywords
0	refugee	난민, 외국인, 대한민국, 이슬람, 예멘, 문화, 문제, 제주도, 정부, 나라
	athletics	선수, 한국, 무슬림, 나라, 사람, 사회, 올림픽, 이민자, 문화, 아랍
1	diplomacy	미세먼지, 중국, 우리나라, 일본, 문제, 사람, 한국, 게임, 정부
	labor	시간, 근무, 업체, 회사, 노동자, 직원, 최저임금, 기업, 개선, 지역
2	finance	공매도, 시장, 주식, 금융, 불법, 정부, 투자자, 화폐, 블록체인, 가상
	site	처벌, 불법, 사건, 촬영, 수사, 범죄, 사이트, 피해자, 사랑, 일베
4	childcare	교사, 학생, 교육, 학교, 학대, 아이, 학부모, 문제, 수업, 의무
	education	교사, 어린이집, 보육, 시간, 아이들, 아이, 원장, 교육, 보육, 교원
5	gender	남성, 지원, 가족, 예산, 전용, 정책, 대한민국, 여성, 폐지, 의무
	equility	남성, 낙태, 사회, 임신, 평등, 차별, 임금, 현실, 문제, 여자
7	community	주민, 지역, 사업, 교통, 서울, 정부, 공사, 아파트, 문제, 건설
		임대, 공공, 분양, 주책, 10년, 아파트, 서민, 전환, 지역, 부동산
9	policy	학교, 학생, 교육, 장애인, 시간, 지원, 경찰, 제도, 생각, 가족
	politics	대통령, 정부, 사람, 국가, 생각, 의원, 청와대, 국회, 국회의원, 국채

cluster_0의 첫 번째 토픽은 난민, 두 번째는 체육과 난민에 관한 주제인데, k-means 클러스터링과 마찬가지로 공통된 단어들이 많아 명확히 분리되지 않았다. cluster_1은 미세먼지와 인근 국가들에 대한 주제와 노동, 고용에 대한 주제이다. 미세먼지는 군집의 개수를 늘려도 외교, 국가문제와 완벽히 분리되지 않았는데, 이는 미세먼지에 관한 글에 중국이라는 단어가 함께 쓰였기 때문이다.

cluster_2는 금융과 관련된 주제와, 불법촬영 및 유포, 사이버범죄와 관련된 주제로 구성된다. k-means 클러스터링에서는 사이버 범죄에 관련된 단어가 명확히 드러나지 않았는데, 토픽모델링으로 불법 촬영, 일베 등의 구체적인 내용이 확인되었다. cluster_4는 학교/교육과 보육으로 명확히 분류되었다.

cluster_5에서 여성가족부와 성평등에 관한 제도와 정책에 대한 주제로 파악되며, 두 번째는 성평등과 낙태/임신, 임금불평등에 관한 주제로 파악된다. 군집개수를 늘려도 임신과 임금이라는 단어는 분리되지 않는데, 이는 임신으로 인한 육아휴직, 휴가가 직장 내 임금 불평등과 관련이 있기 때문이다. 세 번째는 성매매와 관련된 주제이다.

cluster_7은 지역사회에 관한 내용으로, 개발/건설과 주거문제로 구성된다. 개발/건설은 신도시나 재개발문제 등의 이슈로 나타난다. 주거문제는 내 집 마련이나 부동산문제들로 인한 것으로 보인다. cluster_9는 학교와 정부에 관한 단어들로 구성되는데, cluster_9의 학교주제는 cluster_4의 학교/교육과 달리 교육 정책이나 제도에 관한 내용으로 파악된다. 두 번째 주제는 정치에 관련된 주제이다.

3.4 계층적 주제와 실제 카테고리의 비교

Fig. 4는 계층적으로 나타낸 청원 주제이다. 실제 카테고리와 비교하였을 때 내용 기반의 계층적 주제의 특징을 크게 네 가지로 정리할 수 있다. 첫째, '미래', '성장동력'과 같이 분류가 명확하지 않은 카테고리들이 주제에서는 삭제되었다.

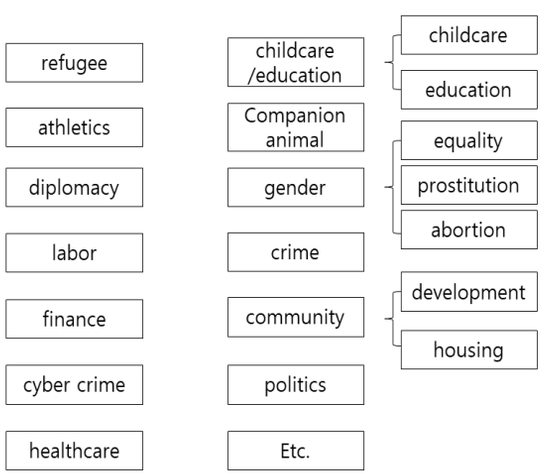

Fig. 4. Hierarchical Topic Structure

둘째, ‘사이버범죄’와 ‘폭행’ 과 같은 상위 주제가 생성되었다. 이는 기존 카테고리에서는 명확히 분류될 수 없는 내용까지 포함하는 것이다. 셋째, 하위 주제가 생성되었다. ‘보육/교육’, ‘성/성평등’, ‘지역사회’의 상위 주제에서 가장 핵심이 되는 하위 주제를 생성하여 청원 글들의 주요 내용을 한눈에 파악하고 세분화된 분류를 이끌어냈다. 넷째, 기존 카테고리에 존재하던 몇몇 주제들이 계층적 주제에는 존재하지 않거나, 혹은 통합되었다. ‘외교/통일/국방’은 ‘외교’로 바뀌었으며, ‘농산어촌’, ‘안전환경’, ‘저출산/고령화’, ‘행정’이 삭제되었다. 이는 삭제된 카테고리들이 신규 주제에서 남아있을 만큼 핵심 내용이 되지는 못했다는 것으로 해석된다.

4. 답변가능 청원 예측 모델

제 4장에서는 20만 이상의 청원동의를 받아 답변을 받을 수 있는 청원을 예측하기 위한 LSTM 기반의 답변 가능 청원 예측 모델을 제시한다. 만료된 청원을 추천 순으로 정렬 후, 상위 2,983개의 글들의 제목, 내용, 카테고리, 청원동의수를 추출하였다. 청원동의수가 20만이 넘는 경우 Class를 1, 아닌 경우 0으로 설정하여 답변을 받은 청원글과 아닌 청원글을 구분한다. 이진분류의 예측 모델로는 LSTM을 사용해준다. LSTM은 Gradient Vanishing문제를 해결하고 긴 시퀀스를 기억하기에 강점이 있어, 청원과 같이 한 주제에 대해 긴 문장이 이어지는 텍스트에 적용하기 적합하다.

상위 2,983개의 글에서 답변을 받은 청원이 118개, 아닌 것이 2,865개로 답변을 받은 글이 전체의 0.039(약 4%)밖에 되지 않아 데이터의 불균형을 맞춰줄 필요가 있다. 불균형한 데이터를 다루는 방법에는 Under Sampling과 Over Sampling, 그리고 이 둘의 혼합방법[8]이 있다. Under Sampling은 속도가 개선되지만 비교적 성능이 떨어지고, Over Sampling은 Under Sampling보다 성능은 좋으나 Over fitting의 위험이 있다. 본 논문에서는 Over Sampling 중에서도 과적합을 다소 완화하는 SMOTE를 사용한다. 그 결과, 답변을 받은 글과 아닌 글의 비율이 조정되며 데이터의 크기가 2,983개에서 3,839개로 확대되었다.

독립변수에는 청원동의에 영향을 줄 수 있는 모든 변수를 추가해준다. 청원 동의수가 20만을 넘겨 답변을 받는 글이 되기 위해서는 국민들에게 설득력 있고 논리 정연한 글의 짜임새가 중요할 것이라고 짐작할 수 있으나 이 외에도 많은 요소가 독립변수가 될 수 있다. 글의 길이가 청원의 동의를 받는 것과 관련이 있을 수 있으며, 해당 카테고리 또한 영향이 될 수 있다. 또, 명사나 형용사 등 사용된 품사의 비율이 의미를 가질 수도 있다.

본 논문에서는 먼저 글의 내용만을 활용하여 답변가능성을 예측해보고, 다음으로 글의 내용 외의 글의 길이, 카테고리, 특정 품사의 비율을 추가하여 글의 내용만을 변수로 사용할 때보다 예측이 향상되었는지 살펴본다. 모든 예측에는Early Stopping을 적용해주었다.Optimizer로는 Adam과 Adadelta를 사용한 결과만을 기록하였는데, 이는 많은 Optimizer를 사용한 결과 중 가장 성능이 좋았던 두 가지이다.

먼저, 글의 제목과 내용을 합친 글만을 LSTM에 적용하고 그 성능을 비교한다. 이 때 사용한 옵션을 Type1이라고 하자. Type1은 아래 Table 3과 같다.

Table 3. LSTM Type1 Options

options	values(functions)
x	title and text
activation	tanh
dropout	0.4
activation	softmax
loss function	binary crossentropy
optimizer	adadelta or adam
batch size	150
epochs	10

Type1 LSTM 중 Optimizer를 Adam으로 설정한 경우는 아래 Fig. 5, Fig. 6과 같다. 손실 부분에서는 epoch=4 이후로 Test의 손실이 Train과 격차가 벌어지며, 특히 7 이후로는 오히려 Test의 손실 값이 증가하였다. 또한 정확도의 경우 Train의 정확도가 Test보다 큰 폭으로 높아진 채 마무리되었다. 분류성능 평가에서는 평균 정밀도가 0.84, 평균 재현율이 0.83, 평균 및 전체 f1-score가 0.83을 기록하였다.

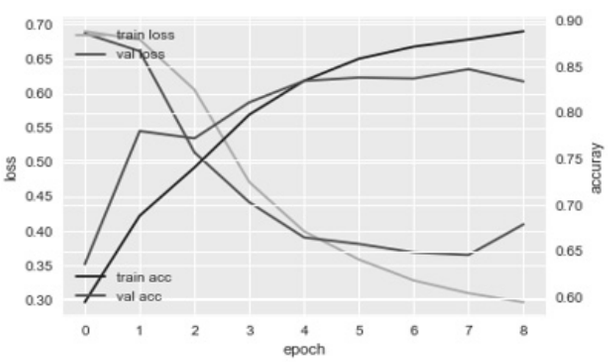

Fig. 5. Type1 (Adam) LSTM Loss & Accuracy

```
class    0    1  count  correct  correct_percent
    0  805  179    984      805             81.8
    1  134  773    907      773             85.2
------------------------------------------------------------
total count: 1891  correct: 1578(83.4%)  f1_score: 0.83
------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.86      0.82      0.84       984
           1       0.81      0.85      0.83       907

avg / total       0.84      0.83      0.83      1891
```

Fig. 6. Type1 (Adam) LSTM Classification Report

Type1 LSTM 중 Optimizer가 Adadelta일 경우는 아래 Fig. 7, Fig. 8과 같다. Train과 Test의 손실이 줄어들었으나 epoch=4 이후부터는 Test의 손실부분의 값이 증가하였다. 분류성능평가에서는 평균 정밀도가 0.83, 평균 재현율이 0.83, 평균 및 전체 f1-score가 0.83으로 동일하여 정밀도는 Adam의 경우보다 0.01 떨어지지만 0과 1을 분류하는 성능이 어느 한 쪽에 치우치지 않았음을 알 수 있다.

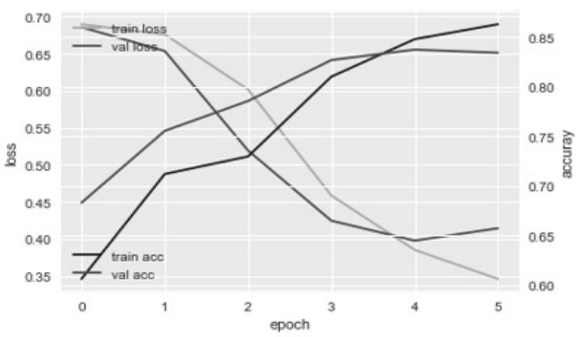

Fig. 7. Type1 LSTM(Adadelta) Loss & Accuracy

글의 제목과 본문을 독립변수로 설정한 Type1 LSTM은 0.83의 F1-score을 기록했다. 분류 결과는 Fig. 8과 같다.

class	0	1	count	correct	correct_percent
0	822	162	984	822	83.5
1	151	756	907	756	83.4

total count: 1891 correct: 1578(83.4%) f1_score: 0.83

	precision	recall	f1-score	support
0	0.84	0.84	0.84	984
1	0.82	0.83	0.83	907
avg / total	0.83	0.83	0.83	1891

Fig. 8. Type1 LSTM(Adadelta) Classification Report

Type2 옵션에서는 글의 제목과 본문 뿐만 아니라 카테고리, 본문의 총 길이와, 한나눔(Hannanum) 형태소 분석기를 사용하여 청원 글 내의 체언, 용언, 수식언, 독립언의 비율을 각각 변수로 추가하였다. Type2의 옵션은 Table 4와 같다.

Table 4. LSTM Type2 Options

Options	Values (Functions)
x	title, text, category, pos ,len of text
activation	tanh
dropout	0.4
activation	softmax
loss function	binary crossentropy
optimizer	adadelta or adam
batch size	150
epochs	10

Type2 LSTM에서 Optimizer를 Adam으로 설정해보았다. Train과 Test 모두 정확도가 0.9를 넘어섰으며, 손실도 0.3아래로 감소하는 결과를 보인다. Type1과 비교하여 loss와 Acc 모두 좋은 성능을 보여준다.

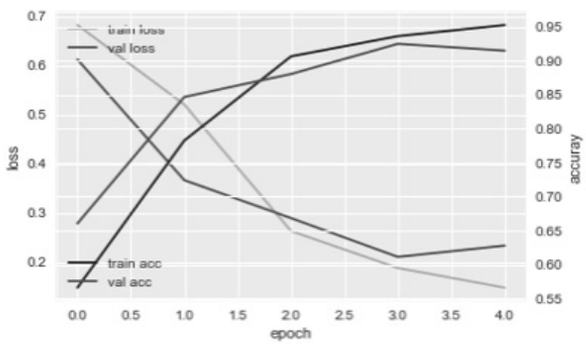

Fig. 9. Type1 LSTM(Adam) Loss & Accuracy

Cofusion Matrix에서도 0과 1 class에서 잘못 분류한 글의 개수가 100개를 넘지 않으며 평균 정밀도, 재현율, f1-score, 그리고 전체 f1-score에서 0.91을 기록하며 0과 1 class에서 고른 성과를 보인다.

class	0	1	count	correct	correct_percent
0	903	81	984	903	91.8
1	80	827	907	827	91.2

total count: 1891 correct: 1730(91.5%) f1_score: 0.91

	precision	recall	f1-score	support
0	0.92	0.92	0.92	984
1	0.91	0.91	0.91	907
avg / total	0.91	0.91	0.91	1891

Fig. 10. Type1 LSTM(adam) Classification Report

다음은 Type2 옵션 중 Adadelta를 Optimizer로 사용했을 경우이다. Train Data의 정확도가 Adam보다 원만하게 증가하는 것을 볼 수 있는데, 이와 달리 Test data에서는 epoch이 3을 넘을 때부터 0.91에 수렴한다. 또한 정확도와 손실 모두에서 Train보다 Test 값이 좋은 것을 확인할 수 있다.

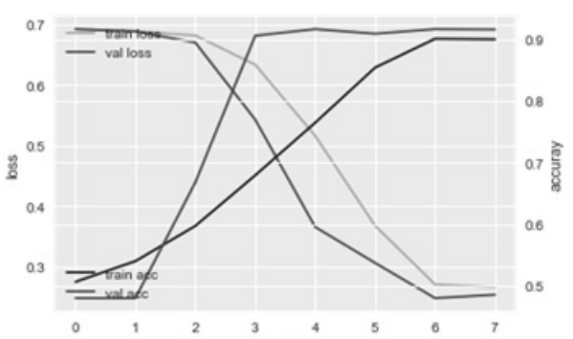

Fig. 11. Type2 LSTM (Adadelta) Loss & Accracy

Confusion Matrix에서는 1인 Class는 Adam의 2배에 달하는 158개 글을 잘못 분류하였지만, 0인 Class의 경우 모두 맞추는 성과를 보였다. 그 결과 0과 1의 평균 정밀도, 재현율, f1-score값은 Adam보다 높으나, 전체 f1-score는 Adam보다 0.01 작은 0.9를 기록하였다.

```
class    0    1  count  correct  correct_percent
    0  984    0    984      984            100.0
    1  158  749    907      749             82.6
-----------------------------------------------------------
 total count: 1891  correct: 1733(91.6%)  f1_score: 0.9
-----------------------------------------------------------
              precision    recall  f1-score   support

           0       0.86      1.00      0.93       984
           1       1.00      0.83      0.90       907

avg / total       0.93      0.92      0.92      1891
```

Fig. 12. Type1 LSTM(Adadelta) Classification Report

Type1과 Type2의 성능을 비교한 결과는 테이블 5와 같다. adam과 adadelta 모두에서, 제목과 본문만을 활용했을 경우보다 글의 길이, 카테고리, 특정 품사의 비율을 함께 사용했을 경우가 성능이 향상되었다는 사실을 알 수 있었다.

Table 5. Comparison between Type1 and Type2

	avg precision	avg recall	avg f1-score	total f1-score
type1 adam	0.83	0.83	0.83	0.83
type1 adadelta	0.84	0.83	0.83	0.83
type2 adam	0.93	0.92	0.92	0.9
type2 adadelta	0.91	0.91	0.91	0.91

Type2 중에서도 adadelta가 평균 정밀도와 재현율, f1-score가 adam보다 높았고, adam은 total f1-score에서 adadelta보다 높은 성능을 보였다.

5. 결 론

본 논문에서는 국민 청원 제도가 시작된 이래로 추천을 많이 받은 청원 내용의 주제를 분석하였다. 먼저 K-means 클러스터링 알고리즘을 적용하여 대주제를 선정한 다음, 좀 더 세분화 된 분류가 필용한 대주제에는 토픽 모델링을 적요하여 세부 주제를 추출하였다.

분석결과 청원글은 기타를 포함한 14개의 주제로 구성된다. 난민, 체육, 외교, 고용/노동, 금융/경제, 사이버 범죄, 의료/보건, 동물, 그리고 정치의 대분류는 세부 분류를 갖지 않으며, 보육/교육, 성/성평등, 그리고 지역사회는 토픽 모델링을 적용하여 다시 세부 주제를 정의하였다. 그 결과 보육/교육은 보육과 교육으로 성/성평등은 성평등, 성매매, 그리고 낙태법으로, 마지막으로 지역 사회는 개발/건설과 주거문제로 나누었다.

또한, 딥러닝 알고리즘 중 텍스트 분류 모델로 잘 알려진 LSTM을 적용하여 답변을 받을 수 있는 청원을 예측하는 모델을 개발하였다. 특히 대부분의 텍스트 분류에서 사용되는 제목이나 본문외에도 글의 길이, 카테고리, 특정 품사의 비율을 변수로 활용한 모델이 예측의 결과가 더 좋았다. 즉, 글의 길이가 길고 체언, 용언, 독립언, 수식언의 순서대로 비율이 높은 경우 많은 청원 동의를 받은 것을 알 수 있다.

또한 Optimizer로는 Adam을 사용했을 경우가 0.91의 f1-score기록하며 adadelta보다 높았으나 Adadelta가 답변을 받지 못할 청원(class=0)에 대해서는 완벽히 분류해내어 분류의 초점에 따라 두 가지 Optimizer 모두 사용될 수 있을 것으로 보인다.

References

[1] The Cheong Wa Dae National Petition Site [Internet], https://www1.president.go.kr/petitions

[2] K. Park, "Semantic Analysis of The Sub-Thematic Word in Big Data," *Journal of the Linguistic Society of Korea*, Vol. 65, pp. 89-109, 2013.

[3] D. Scully, "Web-scale K-means clustering," in *Proceedings of the 19th International Conference on WWW*, pp. 1177-1178, 2010.

[4] H. You, S. Lee, and Y. Ko, "Incremental Clustering and Multi-Document Summarization for Issue Analysis based on Real-time News," *Journal of KIISE*, Vol.45, No.4, pp.355-362, 2019.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[6] D. W. Ko and J. J. Yang, "Korean Natural Language Processing and Analysis Using KoNLPy and Word2Vec," in *Proceedings of the Korean Institute of Information Scientists and Engineers*, pp.140-142, 2018.

[7] Scikit-learn [Internet], https://scikit-learn.org/stable/

[8] G. U. Park and I. K. Jang, "Comparison of resampling methods for dealing with imbalanced data in binary classification problem," *The Korean Journal of Applied Statistics*, Vol.32, No.3, pp.349-374, 2019.

우 윤 희

https://orcid.org/0000-0003-0641-8128
e-mail : 1215yhui@gmail.com
2016년 ~ 현 재 동덕여자대학교
정보통계학과 학사과정
관심분야 : 자연어처리, 영상처리

김 현 희

https://orcid.org/0000-0002-7507-8342
e-mail : heekim@dongduk.ac.kr
1996년 이화여자대학교 컴퓨터학과(학사)
1998년 이화여자대학교 컴퓨터학과(석사)
2005년 이화여자대학교 컴퓨터공학과
(공학박사)
2005 ~ 2006년 LG전자 디지털미디어연구소 선임연구원
2006년 ~ 현 재 동덕여자대학교 정보통계학과 부교수
관심분야 : Big Data Analysis, Machine Learning, Deep Learning