ㅇ

학사학위논문

청와대 국민청원 제목과 참여율에 관한 연구

텍스트 마이닝 기반 Bag of Words 기법 중심으로

김보영

신연지

양지윤

정은지

한양대학교

2020년 6월

학사학위논문

청와대 국민청원 제목과 참여율에 관한 연구

텍스트 마이닝 기반 Bag of Words 기법 중심으로

지도교수 최 원 영

이 논문을 학사학위논문으로 제출합니다.

한양대학교

정보융합전공

조원 김보영 신연지 양지윤 정은지

이 논문을 　　의 학사학위 논문으로 인준함

이 논문을 　　의 학사학위 논문으로 인준함

이 논문을 　　의 학사학위 논문으로 인준함

이 논문을 　　의 학사학위 논문으로 인준함

2020년 8월

지 도 교 수 : 최 원 영 (인)

한양대학교

<목 차>

제 1 장 서론

현재 우리가 살고 있는 이 시대엔 인터넷이 매우 발달되어 국민들의 다양한 정치 참여가 증가하였다. 독일, 스코틀랜드, 영국, 미국 등 우리나라보다 앞장서서 전자청원 시스템을 도입한 나라를 쉽게 찾아볼 수 있다.

우리나라 대표적인 포털사이트 네이버, 다음 등만 들어가도 메인 화면에서 다양한 카테고리의 기사가 올라온다. 또한 「“마스크 써" 여성 택시기사 폭행·추행한 승객…靑 국민청원까지 등장」, 「점심 복귀 30분 늦어 사회복무 5일 연장…"억울" 국민청원」 등 기사 제목을 통해 현재 어떠한 이슈가 생겨났는지, 국민들이 어떤 이슈에 관심을 갖고 있는지 등을 손쉽게 접할 수 있다. 이러한 포털사이트에 올라오는 기사를 보면 “국민청원”이라는 단어가 기사 제목에 많이 포함되는 것을 느낄 수 있는데 그런 요소를 통해 국민들이 국민청원 시스템에 굉장히 많은 참여를 하고 있고, 국민청원의 참여를 통해 더 나은 세상으로 바뀌기를 원한다는 국민들의 간절한 염원을 느낄 수 있다. 하지만 이러한 국민들의 참여에도 불구하고 국민청원 시스템 제목에 대한 연구는 미비한 실정이다.

[그림 1] 제목 내 “국민청원” 단어 들어간 기사

현재 우리나라 국민청원 시스템은 국민청원 게시판에 올린 글 중 일정 기간 내에 20만 명 이상이 추천한 청원만이 정부의 공식적인 답변을 얻을 수 있다. 따라서 우리는 「청와대 국민청원 제목과 참여율에 관한 연구」를 통해 어떠한 키워드가 제목에 포함되었을 때 정부의 응답을 받을 확률이 높은지를 알아내고자 본 연구를 시작하였다.

제 2장 이론적 배경 및 선행연구 고찰

2.1 국내 및 국외 온라인을 통한 국민 의견 수렴

2.1.1 온라인 환경에서의 정치 참여

1993년, 국민과의 쌍방향 의사소통, 서비스 접근 향상을 목적으로 미국에서 전자정부가 시작되었고, 미국의 사례를 통해 다양한 나라에서 전자정부를 시행하였다.

그렇게 전자정부가 시행된 이후, 현재 우리가 사는 이 세상에선 특정 이슈에 대해 다양한 의견을 가진 사람들이 인터넷이라는 가상의 공간에서 만나 토론을 하고, 그것에 대해 이해하는 과정을 거치며, 이를 통해 더 나은 방향을 제시하여 점진적으로 발전하고자 하는 새로운 세상의 변화인 패러다임이 시작되었다.

2.1.2 국가별 온라인 청원 형태

(1) 스코틀랜드

영국 그레이트브리튼섬의 북부 지방의 스코틀랜드에서 2000년에 전자 청원이 처음으로 시작되었다. 스코틀랜드에서 시작된 전자 청원은 오늘 날 전자 청원 시스템의 시초라고 할 수 있다.

스코틀랜드의 전자 청원은 일반 시민들이 의회 홈페이지를 통해 쉽게 작성할 수 있으며, 청원이 접수된 이후로 6주 동안 전자 청원 홈페이지에 공개가 된다. 우리나라의 국민청원 시스템과 동일하게 홈페이지에 올라온 청원 글에 대해 동의하는 사람의 서명 받을 수 있지만, 가장 큰 차이점은 공개되는 모든 청원은 청원위원회에서 검토하므로 청원 동의자 수는 크게 중요한 요소가 아니라는 점이다. 청원이 완료되면 관련 자료가 의회에 교부가 되고 공공청원위원회에서 청원한 상정 여부를 최종적으로 결정한 뒤 본격적으로 청원 절차를 밟게 된다.

(2) 독일

독일의 청원 시스템은 2005년부터 시작되었으며, 독일 연방의회가 운영하는 청원실이 담당하고 있으며 비공개 원칙인 '개별 청원'과 입법 청원인 '공개 청원'으로 구성되어 있다.

온라인 청원은 E-메일 또는 홈페이지를 통해 청원서를 청원위원회에 제출하며, 공개청원가이드라인에 명시된 기준에 따라 선정이 될 경

우 전자 청원 플랫폼에 게시된다. 그렇게 게시된 청원은 게시된 날부터 4주간 공동 서명과 서명 기간 온라인 토론장에서 토론 개설 및 토론하는 것이 가능하다. 4주의 기간 동안 5만 명 이상의 서명을 받은 청원에 대해서는 청원인에게 청원위원회 출석 및 청원 내용 설명 권리가 부여된다. 그렇게 공청회를 열어 토론이 진행되며, 토론을 통해 청원이 반려되거나 통과되어 의회로 청원이 이관되는 둘 중 하나의 방식으로 진행된다.

(3) 영국

영국은 2014년 정부와 하원이 공동으로 전자 청원시스템을 운영하도록 하는 제안에 따라 2015년부터 정부와 하원의 청원위원회에서 공동으로 전자 청원 시스템을 운영하기 시작하였다. 영국의 전자 청원은 청원위원회에서 개별 청원 건들을 검토하는데 전자 청원을 받는 기간을 6개월로 설정하였으며, 6개월 미만이라도 서명이 10만이 넘으면 토론 대상이 될 수 있는지 고려한다. 서명은 개개인의 별도의 이메일 주소가 연동된 개별 서명이 맞는지 확인하는 시스템이 구축되어 있다. 그렇게 올라온 청원에 대해서 1만 명이 서명할 경우 정부의 답변을 받을 수 있으며, 10만 명이 서명 할 경우 영국 의회의 토론 대상이 될 수 있는지 검토를 하게 된다.

(4) 미국

미국의 전자 청원인 ‘위 더 피플(We the People)’은 2011년 9월 개설되어 2016년 12월까지 사용되었다. 위 더 피플에 청원을 올리면 청원자의 이메일로 동의자를 모집할 수 있는 링크가 전송되고, 해당 링크를 통해 150명의 동의를 받아야 ‘위 더 피플’을 통해 일반인에게 공개가 된다. 그렇게 공개된 청원은 30일 이내에 10만 명 이상의 동의를 얻으면 정부에게 응답 의무가 생기며, 정부는 60일 이내로 답변이 이루어질 수 있도록 노력을 해야 한다. 오바마 대통령의 임기 동안 ‘위 더 피플’에 약 48만여 건의 청원이 접수되어 268건이 10만 명 이상의 동의를 얻었으며, 이 중 227건만 정부의 답변을 받았다.

(5) 프랑스

다른 나라에 비해 늦은 2020년 1월 23일, 프랑스 상원에서 전자 청원 플랫폼이 시작되었다. 프랑스의 전자 청원은 법안에 관한 청원과 통제 임무에 관한 청원 두 가지로 나눠볼 수 있다. 법안에 관한 청원은 6개월간 10만 명의 서명 받으면 의장단의 검토를 거쳐 1명 또는

여러 명의 상원의원의 의원발의입법의 형태로 발의가 되며, 통제 임무에 관한 청원은 6개월간 10만 명의 서명 받으면 의장단의 검토를 거쳐 상원의 통제 임무가 신설된다.

2.2 국민청원 게시판과 온라인 정치참여

2.2.1 국민청원 게시판의 출범 배경

민주주의는 시대에 따라 끊임없이 진화해 왔다. 인공지능 등 정보통신분야의 신기술 활용이 보편화하고 데이터 처리 능력 및 컴퓨팅 역량이 비약적으로 증가하는 지능정보 시대를 맞아 전자민주주의 제도화 노력이 이루어지고 있다. 이와 같은 제도화 노력 중 하나가 청와대 홈페이지를 통해 운영되고 있는 국민청원이다.

청와대 홈페이지의 '국민 청원 및 제안' 게시판은 문재인 정부가 출범하고 100일을 맞아 국민과의 소통을 위한 노력을 반영하여 개편하여 열게 된 소통 게시판이다. 국민 청원 게시판은 특히 온라인을 중심으로 공중의 참여를 끌어 낸다는 점에서 '청와대 국민청원 사이트'에 대한 관심이 크게 대두되고 있다. 온라인 공간을 통한 국민청원 제도는 시민들의 의제 설정 권력을 강화하는 기제로 작동할 수 있다. 또한 정부가 국민과의 소통을 강화하고 소통의 기본이 되는 참여를 끌어내기 위해 활용되고 있다. 국민 청원 제도를 통한 참여의 확대가 합리적인 의사결정을 보장하는 것은 아니지만, 대의민주주의의 한계를 보완하고 직접민주주의를 강화하는데 긍정적인 역할을 하고 있다고 볼 수 있다. 즉 정보기술 발달로 기존의 대의제를 중심으로 한 간접민주주의가 직접민주주의화 되어가는 경향으로 인해 우리나라에서 등장한 대표적인 현상이 청와대 국민청원으로 이해할 수 있다.

게시판 상단의 '국민이 물으면 정부가 답한다'는 표어는 직접 소통을 위한 정부의 철학이 반영된 것으로 청원 게시판에 올린 글 중 30일 이내에 20만 명 이상이 추천한 청원은 정부 및 청와대 관계자가 답변하도록 운영 중이다.

청와대 국민청원이 갖는 가장 큰 의미는 그전에는 국민적 관심사에서 멀어졌던 청원권을 되살렸다는 데 있다. 청원권은 제헌헌법에서부터 인정됐던 헌법상 기본권 중의 하나였다. 1963년에는 청원권을 구체화한 '청원법'이 제정됐지만, 40여 년이 지난 2005년도에 이르러서야 1차 개정이 있을 정도였다. 사실상 청원권이라는 권리 자체가 유명무실했던 상황에서 청와대 국민청원 게시판 개설과 더불어 새롭게 주목받고 실현되고 있다.

청와대 국민청원 게시판 도입의 배경 중 하나로는 21세기에 접어들

어 주요 선진국에서 실시하고 있는 전자 청원제도(electronic petition) 즉, e-청원제도의 세계적 확산을 예로 들 수 있다. IT 기술을 접목한 전자 E-청원 시스템은 2000년 스코틀랜드 의회에서 시작되어 2002년 호주의 퀸즈랜드 의회가 도입하였다. 2005년에는 독일 의회가 스코틀랜드와 유사한 시스템을 도입하였고, 2006년에는 영국의 정부가 도입했다. 또한 2011년에는 미국의 오바마 정부가 '위더피플(We the People)'이라는 이름으로 e-청원 서비스를 개시하였다.

청와대 국민청원은 바로 오바마 정부의 '위더피플'을 벤치마킹한 것이다. 인터넷 통신 기술의 뛰어난 발전으로 정보가 신속하고 광범위하게 확산할 수 있었고, 실시간으로 동시성을 가지는 저비용 의사소통 기술 가능성이 확대되면서 시민의 정치 및 행정 참여에 대한 시공간의 제약이 줄어들었다.

2.2.2 국민 참여 추이와 그에 따른 사회적 영향력

국민 청원 출범 이후 16개월간 ('17.8.17. - '19.1.19.) 등록된 게시글은 총 38만 건을 돌파하였고, 하루 평균 약 735건, 시간당 약 30.6건의 청원 게시글이 신규로 등록되는 양상을 보인다.

또한 등록된 게시글은 접수된 청원 게시글이 아니며, 관리자가 삭제한 청원 수까지 모두 더한다면 이보다 더 많을 것으로 예상된다. 관리자가 임의로 삭제해버리는 게시글도 다수이기 때문이다. 제도 시행 후 불과 3개월여 만에 안정적으로 안착하면서 시민들에게 정부와 소통할 수 있는 효과적인 제도로 인식됨과 동시에 국민들의 여론을 보여주는 또 다른 소통의 창구가 되었다. 2020년 현재, 20만 명 이상의 국민 동의를 얻어 정부의 공식 답변을 얻은 국민 청원은 모두 153건이다. 아래 <일별 국민청원 게시판 글 등록 건수 및 게시판에서 이슈화된 초점 사건(focusing event)>를 보고 이슈화가 되면서 청원 수가 급증한 사례를 살펴보았다. 이 중 국민들의 원성을 사서 직접 입법에까지 영향을 끼치게 되면서 '구하라 법', '민식이 법', '해인이 법', '김성수법' 등 새로운 법에 대한 개정을 촉구하는 명칭들이 생겨나기 시작했다. 이런 이슈가 되었던 청원 중 '민식이 법'은 이미 개정이 완료되었으며 지난 3월 25일부터 실제로 시행되고 있다.

또한, 심신장애 탓에 흉악범들이 형을 감형 받는 것을 문제 삼은 청원 참여 인원들에 청와대 김형연 법무비서관은 심신미약 감형 의무조항을 폐지하는 형법 개정안, 이른바 '김성수법'을 통과시켰다.

이를 바탕으로 2019년 청원 중 참여 인원수가 천 명 이상인 것을 대상으로 분석한 결과를 보면, 참여 인원수가 천 명 이상인 청원 게시물은 모두 약 2,142건으로 한 달 평균 200여 건, 전체 청원 수의 약

2.9% 청원이 참여 인원이 천명 이상 되는 것으로 분석되었다. 청원 건수가 많았던 카테고리를 순서대로 살펴보면 정치개혁이 약 389건으로 제일 많았고, 그 다음 인권 및 성 평등, 안전 및 환경, 육아 및 교육 등에 대한 요구가 많았던 것으로 분석되었다.

[그림 2] 2019 국민청원 카테고리별 건수

제 3 장 연구 방법

3.1 연구 주제 설명

본 연구는 청와대 국민청원 게시판에 게시된 청원을 텍스트 마이닝하여 분석한다. 국가에서 답변해야 하는 기준 수인 20만 명 이상이 참여한 청원의 제목(표본 A)을 분석하고, 20만 명의 참여를 받지 못한 청원의 제목(표본 B)을 분석하여 추이를 확인하고자 한다. 이와 더불어 서론에서 살펴본 바와 같이 2019년 청원 건수가 가장 많았던 카테고리는 정치개혁이므로 이 사실을 바탕으로 하여 2019년 자료 및 2020년 1,2분기 자료를 종합한 후 '청원 제목에 정치 이슈와 관련된 단어가 들어가면 참여 인원이 20만 명을 넘을 것이다' 라는 가설을 세웠다.

비록 정치개혁 카테고리 건수가 많음에도 불구하고, 연구 결과는 그와 무관하게 드러난다. '청원 제목에 형사 사건 관련 단어나 가해자에 처벌에 대한 단어가 들어가면 참여 인원이 20만 명을 넘을 것이다'라는 결과가 도출되었고 본 논문에서 그 과정을 다룰 것이다.

3.2 표본 수집 방법 및 데이터 전처리

표본 수집 내용은 다음과 같다. 정확한 청원 수 파악을 위해 이미

만료된 2019년 05월 02일부터 2020년 05월 02일까지의 기간 동안 작성된 총 6,902건 게시글을 대상으로 연구를 진행한다.

웹크롤링은 국민청원 사이트에서 '제목', '청원 만료일', '참여 인원'을 대상으로 추출하며, 그 예시는 아래 [그림 3]에 기재되어 있다.

[그림 3] 국민청원 사이트 예시

웹크롤링을 위한 사전준비는 다음과 같다. 개발툴로는 Pycharm IDE(Integrated Development Environment)를 이용하며 언어는 python을 채택한다. python을 우선 설치한 후, Pycharm은 Jet Brains 공식 사이트에서 Community Edition version으로 설치한다. 이후 python이 설치된 절대 경로로 이동 후 "python -m pip install upgrade pip" 명령어를 입력하여 Python Pip을 최신 버전으로 업데이트한다. Pip(PyPI;The Python Package Index)은 타사 소프트웨어를 모아둔 저장소이며 Pip 통해 가장 안정적인 최신 버전의 플러그인을 설치한다. 라이브러리는 Beautifulsoup과 Selenium webdriver를 사용한다.

BeautifulSoup은 HTML page에서 데이터를 추출하기 위해 널리 사용되는 Python의 라이브러리이다. 단일 웹 페이지에만 국한되지 않으며 여러

웹 페이지에서 데이터를 추출할 수 있다. 대상 데이터를 얻기 위해서는 HTML 트리구조를 이용하여 데이터를 찾는 것이 중요하다. 이 과정에서 해당 데이터를 식별하는 방법을 보여주고, Beautifulsoup을 이용하면 데이터 추출을 위한 구문 분석 규칙을 효과적으로 작성할 수 있다. 추출을 위한 필요 코드를 파악하는 것이 선행되어야 한다.

Selenium WebDriver 라이브러리는 사용자가 먼저 사용하고자 하는 WebDriver 종류(브라우저)를 선택한다. 본 연구에서는 Chromedriver를 선택하여 테스트 도구로 사용한다. Selenium의 가장 큰 장점으로는 테스트 인스턴스를 선택한 브라우저와 통합한 후에 자동화 코드를 실행하여 소스 코드의 품질을 올리고 시간을 단축하여 웹크롤링을 보다 수월하게 진행할 수 있다.

웹크롤링 코드는 아래와 같다.

```
from bs4 import BeautifulSoup
from selenium import webdriver
#1. 셀레니움 설치 에러 시, python 설치된 폴더 가서 python -m pip install --upgrade pip 실행
import time

#2. 크롬드라이버 설치 사이트 : https://sites.google.com/a/chromium.org/chromedriver/downloads

#3. 크롬드라이버로 제어(반드시 절대경로 입력)

    soup = BeautifulSoup(driver.page_source, 'html.parser')

    result_list = []

#6. html 태그와 인덱스를
```

```
driver = webdriver.Chrome("C:/Users/BoyoungKim/AppData/Local/Programs/Python/Python36/chromedriver")

#4. 1부터 n 페이지까지 크롤링
#기간 : 2019-04-27 ~ 2020-05-02
for i in range(1,1219) :

#5. 링크 : 청원-만료된 청원
driver.get("https://www1.president.go.kr/petitions/?c=0&only=2&page="+str(i)+"&order=1")
data = i.find("div", class_="bl_agree").text[5:].strip() + "|" + i.find("div", class_="bl_date").text[7:].strip()+ "|" + i.find("div", class_="bl_subject").text[3:].strip()
        print(data.replace(" ",
```

```
참고하여 위치를 찾아 data 변수에 넣는다.
    for i in soup.select("#cont_view > div > div > div > div > div > div.ct_list1 > div.board.text > div.b_list.category.b_list2 > div.bl_body > ul > li > div"):
"|"))

    time.sleep(5)

driver.close()
```

웹크롤링으로 추출된 데이터는 다음과 같은 데이터 전처리 과정을 거친다. 데이터 전처리는 1. 데이터 정렬 2. 정제 3. 형태소 분석 4. 불용어 제거 과정을 따른다.

1) 데이터 정렬

: 분석의 용이한 데이터로 사용하고자 참여 인원을 기준으로 20만 명 이상, 20만명 이하의 데이터로 분류한 뒤, 분기별로 다시 분류를 진행하였다. 데이터 분류 기준 은 다음[표 1]과 같으며 이하 Dataset 이름은 아래 기호로 기재하였다.

[표 1] Dataset 설명

구분	20만 명 미만	20만 명 이상
2019 2분기	N_2019_2	O_2019_2
2019 3분기	N_2019_3	O_2019_3
2019 4분기	N_2019_4	O_2019_4
2020 1분기	N_2020_1	O_2020_1
2020 2분기	N_2020_2	O_2020_2

2) 정제(*.txt → List 형변환)

: 텍스트 파일로 저장된 Dataset을 'UTF8' 인코딩 세팅 후 라인별로 가져온 뒤, 리스트 형태로 저장한다. 저장된 리스트형 데이터를 더 작은 명사 단위로 나눈 상태에서 불용어를 찾아 제거하고 다시 문장 형태로 합치는 과정을 진행한다.

```
list = [line.rstrip('\n') for line in open('O_2020_2.txt', 'rt',
encoding='UTF8')]
newList = []
for x in range(len(list)):
    tempList2 = ''
    tempList = list[x]
    tempmini = hannanum.nouns(tempList)
    for y in tempmini:
        if y not in stopWords:
          tempList2 = tempList2 + y + ' '
    newList.append(tempList2)
```

3) 형태소 분석

: konlpy 패키지 안에 있는 Hannanum Class를 이용하여 형태소 분석을 진행하였으 며, 그중에서 nouns 메소드를 이용하여 명사 단위로 분석하였다.

```
from konlpy.tag import Hannanum
hannanum = Hannanum()

hannanum.analyze  # 구(Phrase) 분석
hannanum.morphs  # 형태소 분석
hannanum.nouns  # 명사 분석
hannanum.pos # 형태소 분석 태깅
```

4) 불용어 제거

: 유의미한 단어만 추출하기 위해 RANKS NL사에서 배포한 Korean Stopwords 불용어 사전을 사용했으며, 불용어 사전에는 없지만, 국민청원 제목에는 자주 발생하는 단어 중에 해당 연구에 이상값을 생성할 수 있는 단어는 자체적으로 불용어 사전에 추가하여 제거하였다.
(년도, 날짜, ~에 대한, ~로 인한, 위해, 대해 등)

```
stopWords = []
for line in open('bul.txt','rt', encoding='UTF8'):
    stopWords.append(line.split())
```

3.3 분석 방법

분석 방법은 텍스트 마이닝에서 자연어 처리에서 자주 사용되는 통계적 언어 모델(SLM: Statistical Language Model)인 단어 주머니(Bag of Words) 기법을 사용한다. 이는 입력 텍스트에 대해 각 단어가 얼마나 많이 등장하는지 나타내주는 방식으로 해당 문서에 주요 단어를 파악하는 데 효과적이다.

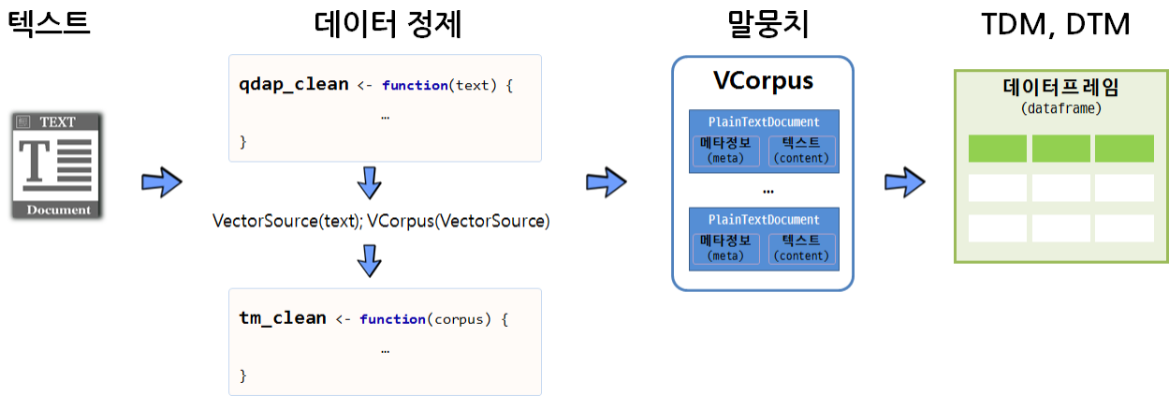
단어의 빈도는 TF-IDF를 통해 산출한다.

Term frequency(문서 빈도) : tf(t,d)
- 1 + log(tf)

Inverse document frequency(역문서 빈도) : idf(t,D)
- N : 코퍼스에 포함된 전체 문서 개수
- |{d∈D : t∈d}| : 단어 t가 존재하는 문서들의 개수

TF-IDF
- $-idf(t,d,D) = tf(t,d) \cdot idf(t,D)$

위 과정을 도식화 하면 다음과 같다.
(TDM : 단어문서행렬, DTM : 문서단어행렬, DTM은 TDM의 전치행렬)



[그림 4] TF-IDF 과정 도식화

표본 A와 표본 B에 대해 각각 단어주머니(Bag of Words) 기법 시행 후, 추출 단어를 기반으로 가설 검증을 진행한다.

제 4 장 연구 결과

4.1 분석 결과

TF-IDF의 정확한 결과 확인을 위해 빈도 분석도 같이 시행하였으며, 그 결과는 다음과 같다. 빈도 분석은 Dataset 10개에 대해 각각 시행했으며, Dataset 에서 빈번히 등장한 상위 30개 단어를 추출하였다.

빈도수 추출 코드는 아래와 같다.

```
from nltk import FreqDist
import re
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# 파일 불러오기
f = open("O_2020_2.txt", 'r', encoding='UTF8')
list = f.readlines()
a = ''
for i in list:
    a = a + i

# 데이터 정제(개행, 특수기호 제거 등)
a = a.replace('\n', ' ')
a = re.sub('\W+',' ', a)
newlist = []
newlist = a.split(' ')
```

```
# 불용어 제거
stopWords = []
for line in open('bul.txt', 'rt', encoding='UTF8'):
    a = line.replace('\n','')
    stopWords.append(a)
newlist_new = [x for x in newlist if x not in stopWords]

# 빈도수 상위 30개 추출
aee = FreqDist(newlist_new)
print(aee.most_common(30))
f.close()
```

아래 [표 2]는 빈번하게 발생한 상위 30개의 단어 중 최상위 5개 단어를 정리한 내용이다. 20만 명 미만 즉, 청원에 성공하지 못한 제목들은 '해주세요' '도와주세요' 등 '요청어'가 대다수였으며, '대한민국', '청와대' 등이 추출된 것으로 보아 특정 대상에 대한 청원보다는 정부 자체에 대한 청원으로 사료된다. 또한, 2020년 1, 2분기에 들어서는 '코로나', '마스크', '신천지' 등 사회면의 이슈를 담은 제목도 상위권에서 찾아볼 수 있었다.

반면, 20만 명 이상 즉, 청원에 성공한 제목들은 수집된 Dataset 규모가 작은 것으로 보아 2019년부터 2020년까지의 청원들은 대다수 청원에 실패한 것을 알 수 있다. 또한, 특이한 점은 앞선 청원 실패 제목과 비교해 특정 사건이나 인물을 제목에 많이 기재했으며, 사회 이슈보다 형사 사건에 대한 단어를 확인할 수 있었다.

[표 2] 빈도분석 결과

구분	20만 명 미만		20만 명 이상	
내용	단어	빈도수	단어	빈도수
	청원합니다	358	N번방	4
	주세요	347	문재인	3
	해주세요	242	그리핀(라이엇)	2
	코로나	191	텔레그램	2
	마스크	167	신상공개	2

위에서 진행한 빈도 분석에 문서 내 중요도를 고려한 TF-IDF 분석을 시행하였다. TF-IDF 추출 코드는 아래와 같다.

```
* |D| : 문서 집합 D의 크기, 또는 전체 문서의 수
* |{d ∈ D:t ∈ d}| : 단어 t가 포함된 문서의 수. (즉, tf(t,d) ≠ 0). 단어가 전체 말뭉치 안에 존재하지 않을 경우 이는 분모가 0이 되는 결과를 가져온다. 이를 방지하기 위해
1 + |{d ∈ D:t ∈ d}|로 쓰는 것이 일반적이다.
```

```
#
==============================
===============
# -- TF-IDF function
#
==============================
===============
def f(t, d):
    # d is document == tokens
    return d.count(t)

def tf(t, d):
    # d is document == tokens
```

```
def tfidf(t, d, D):
    return
tf(t,d)*idf(t, D)

def tokenizer(d):
    # return [ t for t
in d.split() if len(t) >
1 ]
    return d.split()

def tfidfScorer(D):
    tokenized_D =
[tokenizer(d) for d in
```

```
    return 0.5 + 0.5*f(t,d)/max([f(w,d) for w in d])

def idf(t, D):
    # D is documents == document list
    numerator = len(D)
    denominator = 1 + len([ True for d in D if t in d])
    return log10(numerator/denominator)
```

```
D]
    result = []
    for d in tokenized_D:

result.append([(t, tfidf(t, d, tokenized_D)) for t in d])
    return result

# 사용
if __name__ == '__main__':
    corpus = newList
    for i, doc in enumerate(tfidfScorer(corpus)):
        print(doc)
```

TF-IDF 분석 결과는 다음 [표 3]과 같다.

[표 3] TF-IDF 분석 결과

구분	20만 명 미만		20만 명 이상	
내용	단어	TF-IDF	단어	TF-IDF
	유시민	2.89	렌터카	0.88
	수의사	2.89	박사방	0.88
	피해자사망	2.89	가해자	0.88
	긴급재난	2.89	세월호	0.88
	민식이법	2.89	코로나19	0.88

20 만 명 미만 즉, 청원에 성공하지 못한 제목들은 '살인강도', '폭주' 등 강한 단어를 주로 사용하였고, '긴급재난', '민식이법' 등 사람들과 밀접한 관련이 있는 이슈 단어가 많이 사용된 것을 발견했다. 또한, 최근 이슈된 유튜브 스타에 관한 단어인 '수의사'도 찾아볼 수 있으며, 그 외에도 '온라인 개학 반대',

‘강제집행’, ‘KF94’ 등 코로나와 관련된 단어도 추출되었다. 특이한 점은 특정인의 이름이 거의 보이지 않던 해당 Dataset 에서 ‘유시민’ 이란 이름이 상위 랭크된 점이다. 반면에, 20 만 명 이상 즉, 청원에 성공한 제목들 역시 앞선 실패 제목과 동일하게 ‘강간’, ‘성폭행’, ‘여아살해모의’ 등 강한 단어를 사용하였고, 앞선 빈도 분석의 결과와 유사하게 형사 사건에 대한 특정 단어가 상위로 랭킹 되었다. 또한, 특이한 점은 20 만 명 미만의 제목은 피해자의 상황을 호소하는 단어가 많이 발견된 반면 20 만 명 이상의 제목은 ‘가해자’, ‘엄중’, ‘처벌’ 등 가해자에 관한 내용이 다수 확인되었다. 다음 [그림 4]와 [그림 5]는 해당 분석 내용을 ‘워드 클라우드’로 제작한 결과이다.

[그림 5] 빈도분석 결과 워드 클라우드

[그림 6] TF-IDF 결과 워드 클라우드

따라서, 위 결과를 종합했을 때 사전에 가정한 ‘청원 제목에 정치 이슈와 관련된 단어가 들어가면 참여 인원이 20 만 명을 넘을

것이다’ 라는 가설은 틀린 것으로 판명되었으며, ‘청원 제목에 형사 사건 관련 단어나 가해자에 처벌에 대한 단어가 들어가면 참여 인원이 20 만 명을 넘을 것이다’ 라는 사실을 도출하였다.

제 5장 결 론

본 논문은 청와대 국민청원 게시판의 2019년 5월 2일부터 2020년 5월 2일까지의 기간 동안 작성된 국민청원 제목을 분석한 것이다. 앞선 분석을 토대로 사전에 세웠던 가설인 ‘청원 제목에 정치 이슈와 관련된 단어가 들어가면 참여 인원이 20만 명을 넘을 것이다’ 는 빈도분석 결과인 [표 2]에서 '20만 명 이상' 에 나와 있는 단어들을 확인하였을 때에는 정당한 가설로 보여질 수 있으나, 빈도 분석 내에서도 중요도를 고려한 TF-IDF 분석을 거친 후 틀린 것으로 판명되었다.

그러나 그와 동시에 TF-IDF 분석 결과인 [표 3]의 '20만 명 이상'에서 볼 수 있듯이 TF-IDF 지수가 낮은 유의미한 단어들은 전부 형사 사건 혹은 관련 단어였음을 알 수 있다. 따라서 ‘청원 제목에 형사 사건 관련 단어나 가해자에 처벌에 대한 단어가 들어가면 참여 인원이 20만 명을 넘을 것이다’ 라는 새로운 사실을 도출한 것이다.

본 연구를 바탕으로 국민청원 게시판에 대한 향후 연구 방향에 대해 논의할 점을 서술하고자 한다.

첫째, 위의 도출한 사실을 바탕으로 국민 청원을 통해 가해자의 엄중 처벌에 대한 청원이 국민들의 동의를 받아 20만 명이 초과할 경우 과연 국민들의 바람대로 엄중 처벌이 이루어졌는지에 대하여 관련 통계를 범죄 카테고리별로 분석하는 연구를 진행한다면 더 의미 있는 사실을 도출해낼 수 있을 것이다.

둘째, 현대 우리 사회는 범죄에 관련된 청원이 상대적으로 많고, 또 이슈화 되고 있다. 향후 이와 관련된 다른 연구가 진행될 때에는 카테고리별로 정치개혁/경제민주화/인권 및 성평등 등의 범죄 분야를 제외한 기타 다른 카테고리별로 청원 데이터를 각각 수집하여 분석이 독립적으로 수행된다면 전혀 다른 결론에 도달할 수 있을 것으로 예상이 된다.

이러한 다른 카테고리별로 분석한 결과도 앞으로 지켜봐야 될 문제라 생각하며, 후속 연구가 필요할 것이다.

제 6장 참고문헌

1. 손형섭, 디지털 플랫폼과 AI에 의한 국회 전자 청원시스템 활성화 연구. <유럽헌법연>, 제31권 제1호, P.493-543 (2019년 12월),
2. 김주희 · 장혜영, 시민 정치참여의 제도화 : 독일의 e-청원 사례를 중심으로. <중앙대학교 국가정책연구소>, 제32권 제1호, p.1-19 (2018년 3월)
3. 장혜영, Two Track 모델 속 숙의 민주주의 시스템 비교 분석: 독일, 영국 및 한국 사례. <국가정책연구>, 제32권 제3호, p.177-203 (2018년 9월)
4. 정재환, 미국의 '위더피플' 사례를 통해 살펴본 청와대 국민청원의 개선방안. <NARS 현안분석>, Vol.27, p.1-11 (2018년 11월)
5. 정동재 · 박준 · 김은주, 국민청원제도 시행 16개 : 더 나은 제도운영을 위한 개선방안, p.3-9 (2019년)
6. 우윤희 · 김현희, 국민청원 주제 분석 및 딥러닝 기반 답변 가능 청원 예측, p.1-3 (2020년)
7. 김병록, 청와대 국민청원의 개선방안에 관한 연구, p.2-3 (2019 년)
8. 청와대 블로그, https://blog.naver.com/thebluehousekr/221275572142
9. https://www.spiderkim.com/post/데이터-분석-2019 년-청와대-국민청원-데이터-분석
10. 국민청원게시판, http://www.president.go.kr/
11. xwMOOC 자연어 처리 http://bitly.kr/NJNOyL2il!
12. Brian Okken, Python Testing with pytest: Simple, Rapid, Effective, and Scalable
13. Stone River eLearning, Python BeautifulSoup 中 Resource description page
14. Andy Craze, Beginning Selenium WebDriver Testing in Python
15. Alberto Boschetti and Luca Massaron, Python Data Science Essentials-Third Edition-102page
16. TF-IDF Function 소스코드 출처 블로그 발췌 https://url.kr/u6RrzA
17. NLTK 자연어 처리 패키지 https://url.kr/l54EI8
18. 형태소 분석 출처 https://cceeddcc.tistory.com/8

부록

[부록 표1] 전체 빈도 분석 결과

	1등	2등	3등	4등	5등
N2019_2	청원합니다', 68	'주세요', 55	'부동산', 38	'해주세요', 35	'김의겸', 35
N2019_3	주세요', 96	'청원합니다', 96	'도와주세요', 46	'해주세요', 45	'청원', 38
N2019_4	주세요', 48	'청원합니다', 31	'해주세요', 27	'고발합니다', 26	'청원', 24
N2020_1	코로나', 98	'청원합니다', 88	'주세요', 75	'마스크', 70	'해주세요', 65
N2020_2	마스크', 97	'코로나', 93	'청원합니다', 75	'코로나19', 74	'주세요', 73
O_2019_2	김무성', 1	'의원을', 1	'내란죄로', 1	'다스려주십시오', 1	'진주', 1
O_2019_3	청원합니다', 2	'해주세요', 2	'강화', 2	'일본', 2	'기밀누설죄를', 1
O_2019_4	그리핀', 2	'라이엇', 1	'코리아의', 1	'조', 1	'대표', 1
O_2020_1	문재인', 3	'탄핵을', 2	'촉구합니다', 2	'대통령님의', 1	'원하지', 1
O_2020_2	처벌해주세요', 2	'신상공개를', 2	'원합니다', 2	'N번방', 2	'청원합니다', 2

	6등	7등	8등	9등	10등
N2019_2	'청원', 30	'문재인', 28	'청와대', 28	'대한민국', 24	'바랍니다', 23
N2019_3	'주십시오', 38	'일본', 31	'바랍니다', 30	'요청합니다', 29	'대한민국', 28
N2019_4	'주십시오', 23	'도와주세요', 21	'막아주세요', 20	'아들', 19	'바랍니다', 17
N2020_1	'코로나19', 51	'신천지', 47	'요청합니다', 38	'우한', 38	'중국', 37
N2020_2	'해주세요', 70	'부탁드립니다', 32	'요청합니다', 30	'주십시오', 28	'바랍니다', 26
O_2019_2	'방화', 1	'살인', 1	'범죄자에', 1	'무관용', 1	'원칙이', 1
O_2019_3	'범한', 1	'윤석열', 1	'총장을', 1	'처벌해', 1	'주십시요', 1
O_2019_4	'김', 1	'현', 1	'DRX', 1	'감독의', 1	'징계에', 1
O_2020_1	'않습니다', 1	'대통령님을', 1	'응원', 1	'코로나19의', 1	'확산방지를', 1
O_2020_2	'텔레그램', 2	'n번방', 2	'렌트카', 1	'훔쳐', 1	'사망사고를', 1

	11등	12등	13등	14등	15등
N2019_2	'대통령님', 23	'3명', 22	'고발합니다', 21	'폐지', 21	'제발', 20
N2019_3	'고발합니다', 27	'제발', 25	'부탁드립니다', 23	'처벌을', 23	'요청', 22
N2019_4	'요청합니다', 17	'요청', 15	'검찰', 15	'촉구합니다', 14	'처벌을', 14
N2020_1	'청원', 36	'막아주세요', 36	'도와주세요', 35	'중국인', 32	'촉구합니다', 31
N2020_2	'요청', 26	'촉구합니다', 25	'청원', 25	'요구합니다', 25	'도와주세요', 24
O_2019_2	'필요합니다', 1	'대', 1	'수의대에서', 1	'실험중인', 1	'퇴역', 1
O_2019_3	'나경원', 1	'자한당', 1	'원내대표의', 1	'의혹에', 1	'특검', 1
O_2019_4	'재조사가', 1	'필요합니다', 1	'국가인권위가', 1	'조국', 1	'장관과', 1
O_2020_1	'애써주시는', 1	'문재인대통령님과', 1	'질병관리본부', 1	'정부부처', 1	'관계자분들께', 1
O_2020_2	'낸', 1	'10대', 1	'엄중', 1	'세월호', 1	'전면재수사', 1

	16등	17등	18등	19등	20등
N2019_2	'막아주세요', 19	'주십시오', 18	'요청합니다', 18	'처벌', 17	'반대합니다', 17
N2019_3	'처벌', 22	'막아주세요', 22	'요구합니다', 21	'처벌해주세요', 20	'처벌해', 19
N2019_4	'조국', 14	'윤석열', 14	'요구합니다', 13	'지켜주세요', 12	'부탁드립니다', 12
N2020_1	'반대합니다', 31	'바랍니다', 29	'주십시오', 28	'부탁드립니다', 28	'바이러스', 27
N2020_2	'대구', 24	'신천지', 24	'개학', 22	'n번방', 21	'지원', 20
O_2019_2	'탐지견을', 1	'구조해주십시요', 1	'소방공무원을', 1	'국가직으로', 1	'전환해주세요', 1
O_2019_3	'요청', 1	'언론사의', 1	'가짜뉴스의', 1	'강력한', 1	'처벌을', 1
O_2019_4	'가족', 1	'수사과정에서', 1	'빗어진', 1	'무차별', 1	'인권', 1
O_2020_1	'감사드립니다', 1	'한전', 1	'사업에', 1	'중국', 1	'기업의', 1
O_2020_2	'박사방', 1	'회원', 1	'여아살해모의한', 1	'공익근무요원', 1	'오늘', 1

	21등	22등	23등	24등	25등
N2019_2	'국회의원', 17	'요청', 17	'자유한국당', 15	'요구합니다', 15	'사건', 15
N2019_3	'불법', 18	'주십시요', 18	'촉구합니다', 18	'지켜주세요', 18	'있도록', 17
N2019_4	'장제원', 12	'처벌해주세요', 11	'대한민국', 11	'자유한국당', 10	'반대합니다', 10
N2020_1	'요구합니다', 26	'요청', 26	'신종', 26	'마스크를', 24	'코로나바이러스', 22
N2020_2	'어린이집', 19	'막아주세요', 19	'외국인', 19	'반대합니다', 19	'있게', 18
O_2019_2	'연합뉴스에', 1	'국민혈세로', 1	'지급하는', 1	'연', 1	'300억원의', 1
O_2019_3	'청와대는', 1	'조국', 1	'법무부장관', 1	'후보자의', 1	'임명을', 1
O_2019_4	'침해를', 1	'조사할', 1	'청원합니다', 1	'06년생', 1	'집단', 1
O_2020_1	'참여를', 1	'허락하는', 1	'것은', 1	'말도', 1	'안됩니다', 1
O_2020_2	'킬', 1	'KILL', 1	'한다', 1	'라며', 1	'술을', 1

	26등	27등	28등	29등	30등
N2019_2	'부탁드립니다', 15	'6명', 15	'도와주세요', 14	'처벌을', 14	'2명', 14
N2019_3	'억울함을', 16	'국회의원', 15	'갑질', 15	'수사', 14	'살려주세요', 14
N2019_4	'처벌', 10	'아파트', 10	'있도록', 10	'원합니다', 10	'주십시요', 10
N2020_1	'처벌해주세요', 22	'문재인', 21	'우한폐렴', 21	'폐렴', 20	'고발합니다', 18
N2020_2	'온라인', 17	'필요합니다', 16	'따른', 16	'제발', 16	'19', 16
O_2019_2	'재정보조금', 1	'제도의', 1	'전면', 1	'폐지를', 1	'청원합니다', 1
O_2019_3	'해주십시오', 1	'고', 1	'김성재님의', 1	'사망', 1	'미스테리를', 1
O_2019_4	'폭행', 1	'사건', 1	'언론의', 1	'세무조사를', 1	'명령한다', 1
O_2020_1	'국민청원', 1	'안', 1	'신천지', 1	'교주', 1	'즉각적인', 1
O_2020_2	'먹이고', 1	'딸을', 1	'합동', 1	'강간한', 1	'미성년자들을', 1

[부록 표 2] 전체 TF-IDF 분석 결과

2019년 2분기

1-1. 성공한 청원(상위 10개)

년도	분기	단어	TF-IDF
2019	2	김무성	0.477121
2019	2	진주	0.477121
2019	2	대	0.477121
2019	2	소방공무원	0.477121
2019	2	연합뉴스	0.477121
2019	2	안녕	0.477121
2019	2	전	0.477121
2019	2	방화	0.477121
2019	2	수의대	0.477121
2019	2	국가직	0.477121

2-1. 실패한 청원(상위 10개)

년도	분기	단어	TF-IDF
2019	2	주식공매도제도	2.947434
2019	2	토지신탁	2.947434
2019	2	수원시	2.947434
2019	2	에어컨	2.947434
2019	2	감독	2.947434
2019	2	김포학운2산업단지	2.947434
2019	2	백내장	2.947434
2019	2	자영업자	2.947434
2019	2	한부모가정의엄마	2.947434
2019	2	성남시	2.947434

2019 년 3 분기

1-1. 성공한 청원(상위 10개)

년도	분기	단어	TF-IDF
2019	3	의원님들	2.947434
2019	3	80세	2.947434
2019	3	허위광고	2.947434
2019	3	공산국가	2.947434
2019	3	중계화면	2.947434
2019	3	물적분할	2.947434
2019	3	보험사	2.947434
2019	3	'땅	2.947434
2019	3	습격	2.947434
2019	3	만6세	2.947434

2-1. 실패한 청원(상위 10개)

년도	분기	단어	TF-IDF
2019	3	조	2.903633
2019	3	댓글부대	2.727541
2019	3	대한간호조무사협회	2.903633
2019	3	대형마트	2.727541
2019	3	아베에	2.727541
2019	3	요양원	2.602603
2019	3	공양	2.903633
2019	3	피해자	1.94939
2019	3	서민형	2.727541
2019	3	범죄자	2.505693

2019 년 4 분기

1-1. 성공한 청원(상위 10개)

년도	분기	단어	TF-IDF
2019	4	국가인권위	0.30103
2019	4	06년생	0.30103
2019	4	조국	0.30103
2019	4	집단	0.30103
2019	4	언론	0.30103
2019	4	그리핀	0.30103
2019	4	장관	0.30103
2019	4	폭행	0.30103
2019	4	세무조사	0.30103
2019	4	가족	0.30103

2-1. 실패한 청원(상위 10개)

년도	분기	단어	TF-IDF
2019	4	가해자들	2.640978
2019	4	나경원	2.163857
2019	4	선생님	2.464887
2019	4	경비원폭행	2.640978
2019	4	입대후	2.640978
2019	4	박근혜	2.243038
2019	4	엄마	2.339948
2019	4	실효성	2.640978
2019	4	국민	1.686736
2019	4	병원	2.339948

2020 년 1 분기

1-1. 성공한 청원(상위 10개)

년도	분기	단어	TF-IDF
2020	1	코로나19	0.740363
2020	1	전	0.740363
2020	1	국민청원안	0.740363
2020	1	전자개표	0.740363
2020	1	저	0.740363
2020	1	중국인	0.740363
2020	1	윤석열	0.740363
2020	1	한국기독교총연합회	0.740363
2020	1	확산방지	0.740363
2020	1	사업	0.740363

2-1. 실패한 청원(상위 10개)

년도	분기	단어	TF-IDF
2020	1	난	2.886773
2020	1	가마니살인사건	2.886773
2020	1	친딸	2.886773
2020	1	신천지자산	2.886773
2020	1	수의사	2.886773
2020	1	김병욱	2.886773
2020	1	코로나바이러스예방	2.886773
2020	1	교습소	2.886773
2020	1	유사사이비종교	2.886773
2020	1	처지교인	2.886773

2020 년 2 분기

1-1. 성공한 청원(상위 10개)

년도	분기	단어	TF-IDF
2020	2	렌트카	0.875061
2020	2	세월호	0.875061
2020	2	박사방	0.875061
2020	2	"오늘	0.875061
2020	2	수출용	0.875061
2020	2	식	0.875061
2020	2	가해자	0.875061
2020	2	저희	0.875061
2020	2	코로나19로	0.875061
2020	2	중국인	0.875061

2-1. 실패한 청원(상위 10개)

년도	분기	단어	TF-IDF
2020	2	유시민	2.828982
2020	2	수의사	2.828982
2020	2	적폐	2.828982
2020	2	폭주	2.828982
2020	2	살인강도강간등	2.828982
2020	2	피해자사망	2.828982
2020	2	렌터카	2.828982
2020	2	청소년법개정처벌강	2.828982
2020	2	국회의원의임금	2.828982
2020	2	10조3항	2.828982