언론의 보도가 선거에 미치는 영향

20180431 김수지

주제

제20대 대선 후보별 기사 제목 텍스트 분석

연구배경

2022년 3월 9일 제20대 대통령 선거가 치러졌다. 선거 운동 기간 내내 이재명, 윤석열 두 후보의 지지도는 치열했으며, 그 결과 이재명 47.83%, 윤석열 48.56% 약 0.73%의 표차로 윤석열 후보가 당선되었다. 박빙의 선거가 치러진 만큼 그 승패요인에도 귀추가 주목되고 있다.

오래전부터 많은 학자들은 미디어가 선거에 미치는 영향에 대해 고찰해왔다. 미디어를 통해 전달되는 정보들이 유권자들의 선택에 영향을 미친다는 매커니즘이 존재한다는 것이다. 바로 의제설정효과(Agenda-setting), 프레이밍 효과(Framing), 프라이밍 효과(Priming)이다. 먼저 의제설정효과란 언론을 비롯한 많은 미디어에서 특정 이슈를 반복적으로 노출하고 강조하면서 유권자들의 우선순위에 영향을 줄 수 있고 이는 투표 선택에 차이를 가져올 수 있다는 것이다. 다음으로 프레이밍 효과란 특정 이슈의 맥락, 상황을 제시하거나 특정 측면을 강조함으로써 이슈를 어떤 방식으로 이해하고 적용하느냐에 차이를 가져올 수 있음을 의미한다. 미디어가 특정 상황을 어떤 식으로 보도하냐에 따라서 그 이슈에 대한 입장을 기준으로 후보나 정책을 평가하도록 이끌 수 있다는 것이다. 마지막으로 프라이밍 효과는 기준점을 제공하고 그에 기반해서 후보나 정당에 대해 평가를 내리도록 하는 것을 뜻한다. 이는 프레이밍 효과와 엄밀하게 구분하기 어렵지만 선택하는 기준의 변화에 영향을 미치려고 하는 시도이다. 이렇게 세가지 매커니즘을 통해 알 수 있듯 정보들이 어떤 미디어에서 어떻게 전달되느냐에 따라 유권자들의 인식과 선택에 영향을 미칠 수 있다. 미디어의 범주는 기존 TV, 신문, 라디오 등에서 인터넷 뉴스, 유튜브, SNS 등으로 그 범위를 점차 넓혀가고 있다. 미디어의 소비 행태는 더 쉽고 간편하게 접근할 수 있도록 바뀌고 있기 때문에 미디어 특히 인터넷 뉴스를 통한 의제 설정, 프레이밍 효과가 더 광범위하고 손쉽게 일어날 수 있는 것으로 보인다.

따라서 본인은 다양한 미디어 중 언론 특히 인터넷 뉴스가 선거에 미치는 영향에 대해 주목하여 제20대 대선에 언론이 미친 영향을 살펴보고자 한다. 그 방법으로는 비정형데이터분석 중 하나인 텍스트 마이닝과 TF-IDF 지수를 통해 지난 20대 대선에서 뉴스미디어가 '이재명'과 '윤석열' 후보를 어떤 키워드를 중심으로 보도했는지 분석하고자 한다. 특히 공식 선거 기간인 22일 중 유권자들이 가장 집중하는 마지막 이틀을 중점적으로 살펴보며 어떤 키워드들이 이번 선거에 영향을 미쳤는지 알아보고자 한다.

관련연구

텍스트 마이닝

텍스트 마이닝이란 대규모의 비정형 텍스트 데이터를 형태소 분석과 자연어 처리 기술을 이용해 유용한 정보를 추출 및 가공하는 빅데이터(Big Data) 분석 기법 중 하나를 의미한다. 인터넷 문서, doc, PPT 등 다양한 포맷의 대규모 문서에서 추출한 유용한 정보들을 효과적으로 표현함으로써 거시적 관점에서 텍스트 맥락을 기반으로 한 분석결과를 도출할 수 있다(Liu, Liang, & Wishart, 2015). 또한 기존의 설문조사나 전문가 조사를 탈피해 연구 영역에서 생산된 텍스트를 객관적인 연구 프로세스를 통해 연구 주제를 탐색할 수 있다(조수곤, 조재희, 김성범, 2015; 정효정, 2016). 텍스트 마이닝은 DM을 생성하는 방식으로 데이터를 구조화해 자료의 패턴을 검색 및 분석하거나 감성분석(Sentiment Analysis)이나 오피니언 마이닝(Opinion Mining) 등을 활용해 분석 결과에 의미를 더한다. 우선 연구에 필요한 텍스트 문서를 선정해 수집한 후, 자연어 처리 기술 (Natural Language Processing, NLP)을 이용해 전처리 과정을 거친다. 'KoNLPy'패키지를 이용해 한국어 자연어 처리를 하며, 형태소분석기를 이용해 문서 내 어절을 의미를 갖는 최소 단위인 형태소로 분리하고 품사를 찾아내는 형태소 분석을 수행한다(곽수정, 김보겸, 이재성, 2013).

TF-IDF

TF IDF는 단어 빈도(Term Frequency)와 역문서 빈도(Inverse Document Frequency) 가 결합된 값으로, TF는 특정 단어의 문서 내 등장 횟수값, IDF는 역문서 빈도를 의미한다. TF-IDF는 여러 문서로 이루어진 문서 군이 있을 때, 단어가 특정 문서 내에서 얼마나 중요한지를 나타내는 통계적 수치로, TF 값과 IDF 값을 각각 구해 둘을 곱함으로써 구할 수 있다. TF 값은 단어 빈도를 의미하며 특정 단어가 문서 내에서 얼마나 자주 등장하는지를 나타내는 값이다. 단순한 단어의 빈도수는 조사나 관사와 같은 무의미하지만 빈번하게 등장하는 단어들을 중요한 단어로 판단할 수 있기 때문에, TF-IDF 방법에서는 TF 값에 역문서 빈도인 IDF 값을 곱해줌으로써 일종의 페널티를 부여하고 전체적으로 정말 유의미한 단어들의 중요도를 계산해 낸다. TF-IDF 값이 크다는 것은 TF 값과 IDF 값이 모두 크다는 것이며, 이는 특정 문서 내에서 많이 사용되는 단어이지만 여러 문서에서 동시에 나타나는 단어는 아니기 때문에 전체적인 맥락에서 의미 있는 단어라고 해석할 수 있다.

TF값	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: 단어 t_i가 문서 d_j에서 출현한 횟수 $\sum_k n_{k,j}$: 문서 d_j에서 모든 단어가 출현한 횟수
IDF값	$idf_i = \log \frac{\lvert D \rvert}{\lvert d_j \rvert t_j \in d_j \rvert}$ $\lvert D \rvert$: 문서집합에 포함되어 있는 문서의 수 $\lvert d_j \rvert t_j \in d_j \rvert$: 단어 t_j가 등장하는 문서의 수
TF-IDF 가중치	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

제안방법론

연구 순서는 다음과 같다.

1. 설정 기간동안 게시된 후보자 관련 기사 제목 크롤링
2. 형태소 분석
3. 막대그래프 및 워드클라우드를 통한 시각화
4. TF-IDF 지수
5. 결론

실험내용

1. 설정 기간동안 게시된 후보자 관련 기사 제목 크롤링

먼저 공식 선거 운동 기간 22일 중 어느 정도를 설정 기간으로 둘 것인지를 결정하였다. 모든 일자의 기사를 수집하기에는 시간과 데이터 크기의 문제가 존재하므로 유권자들의 관심이 제일 높은 마지막 이틀 (2022년 3월 7일 – 2022년 3월 8일)에 게시된 두명의 후보자 관련 기사 제목을 각각 크롤링하여 수집하였다. Pandas와 BeautifulSoup, re, requests 라이브러리를 사용하여 설정 기간동안 이재명, 윤석열 후보 각각을 검색하였을 때 나오는 기사를 크롤링하였다. 기사의 작성일과 기사 제목, 기사 주소를 수집하여 저장하였다.

이재명 후보의 기사는 284600개, 윤석열 후보의 기사는 110600개가 수집되었으며 그 수를 맞추기 위해 실험에서는 두 후보의 기사 모두 110600개를 사용하였다.

2. 형태소 분석

텍스트 마이닝을 위해 저장된 기사 제목을 Twitter를 이용하여 형태소 분석을 진행하였다. 본 연구는 언론의 언급에서 언급한 키워드를 중점으로 보기 때문에 명사로 된 단어들을 추출한다. 기사 제목에서 추출된 명사들은 각 빈도를 측정하였다.

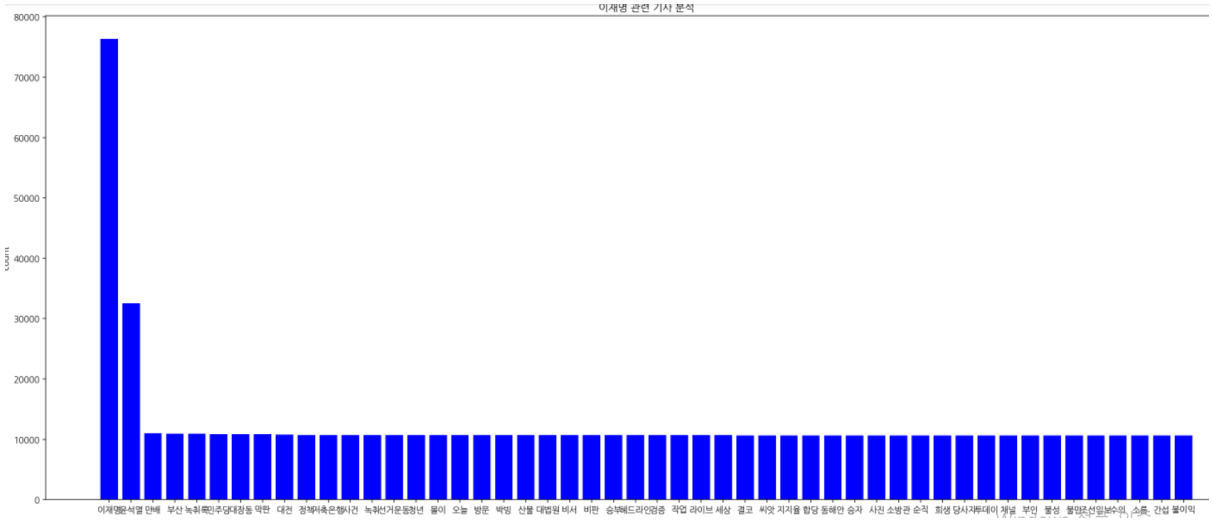
두 후보 각각 높은 빈도를 가지는 단어 50, 200개를 추출하여 저장하였다.

3. 막대그래프 및 워드 클라우드를 통한 시각화

형태소 분석에서 추출된 단어들을 쉽게 확인하기 위해 막대그래프 및 워드 클라우드를 사용하여 시각화를 진행하였다.

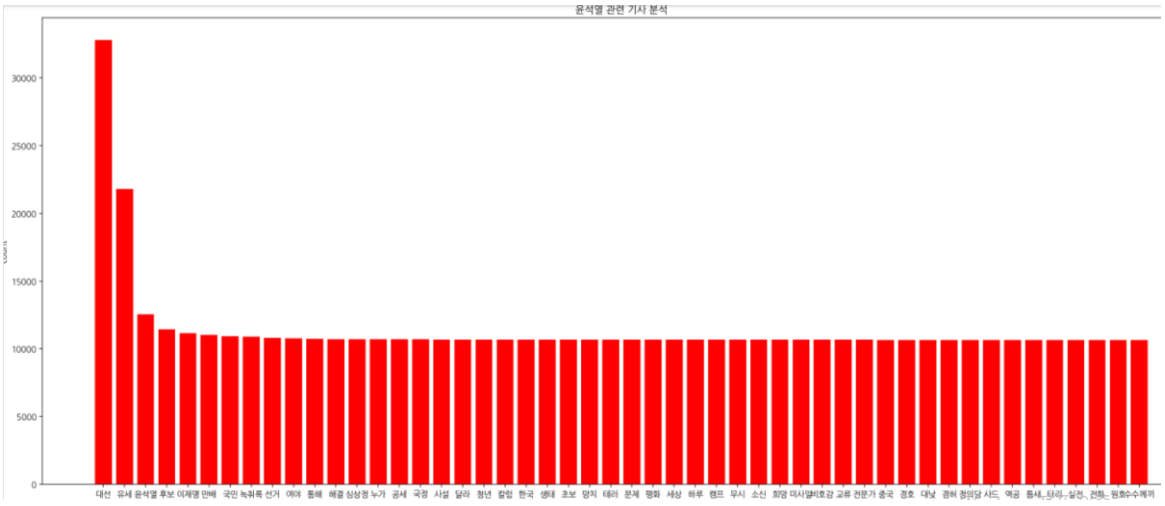
- 막대그래프

막대그래프의 경우 높은 빈도를 가지는 50개 단어들을 나타낸다.



위 그래프는 이재명 후보의 기사 제목에서 추출된 단어들이다.

후보들의 이름이나 기본 단어들을 제외하고 (김)만배, 부산, 녹취록, 민주당, 대장동 등이 높은 빈도로 이재명 후보의 기사 제목들에서 언급되었다.



위 그래프는 윤석열 후보의 기사 제목에서 추출된 단어들이다.

후보들의 이름이나 기본 단어들을 제외하고 (김)만배, 국민, 녹취중, 선거, 여야, 해결 등이 높은 빈도로 윤석열 후보의 기사 제목들에서 언급되었다.

- 워드 클라우드

워드 클라우드의 경우 높은 빈도를 가지는 200개의 단어들을 나타낸다.

위 그림은 이재명 후보의 기사 제목에서 추출된 단어들이다.

후보 이름을 제외하고 눈에 띄는 단어들은 대장동, 저축은행, 녹취록, 민주당, 대법원, 몰이, 청년 등이 있다.

위 그림은 윤석열 후보의 기사 제목에서 추출된 단어들이다.

후보 이름을 제외하고 눈에 띄는 단어들은 유세, 심상정, 청년, 녹취록, 사설, 국정, 초보, 희망 등이 있다.

4. TF-IDF 지수

먼저 두 후보의 일자별(3월 7일, 3월 8일) 기사들을 나누고 랜덤하게 기사 제목 5개씩을 뽑아 실

험을 진행하였다.

이재명 후보의 랜덤으로 추출된 3월 7-8일 기사 제목 10개의 정규화된 TF-IDF 지수를 살펴보면 아래와 같다.

대선후보	선거운동하는	세상	세운	수도권	승부	승자의	안돼	어제	이재명
0.867498	0	0	0	0	0	0	0	0	0.49744
0	0	0	0.70711	0.70711	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.9898	0	0.43817
0.687499	0	0	0	0	0	0	0	0	0.49744
0	0	0	0	0	0.66851	0.66851	0	0	0.32587
0	0	0	0	0	0	0	0	0.89889	0.43817
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	0.89889	0	0	0	0	0	0	0	0.43817

표를 확인하였을 때 '세상', '이재명', '안돼', '어제', '대선후보', '선거운동하는' 등이 높은 TF-IDF지수를 가짐을 볼 수 있다.

윤석열 후보의 랜덤으로 추출된 3월 7-8일 기사 제목 10개의 정규화된 TF-IDF 지수를 살펴보면 아래와 같다.

대선	안돼	안방	안해	언론은	역공	외신기자	유세	유세... 安과	후보
0	0	0	0.57735	057735	0	0.57735	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0.70711	0	0	0	0	0	0.70711	0
0.687499	0	0	0	0	0	0	0	0	0.49744
0.70711	0	0	0	0	0	0	0	0	0.70711
0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0.54362	0	0	0	0	0	0	0.63949	0	0.54362

표를 확인하였을 때 '안돼', '후보', '대선', '후보', '유세' 등이 높은 TF-IDF지수를 가짐을 볼 수 있다.

실험 한계점

실험을 진행하며 여러가지 한계점이 존재했다. 먼저 선거 기간 22일 중 단 2일밖에 데이터를 수집한 것이 가장 큰 한계점이다. 시간과 데이터의 방대한 크기 때문에 더 많은 기사들을 수집할 수 없었다. 또한 기사의 특성상 같거나 비슷한 기사들이 여러 개 존재하는데 이를 완벽하게 처리하지 못하여 동일한 기사들이 많이 수집되었다. 이는 형태소 분석이나 TF-IDF지수를 살펴볼 때에도 좋지 않은 영향을 끼쳤을 것이라 생각한다. 그리고 TF-IDF 지수를 살펴볼 때 모든 기사 제목들을 살펴보기엔 한계가 있어 일자별로 랜덤하게 5개씩 뽑아서 진행하였는데 더 의미 있는 결과를 보여주는 방법이 있을 것이라 생각한다.

결론

본 연구에서는 언론의 언급이 선거에 영향을 미치는가에 대하여 제20대 대통령 선거 운동 기간 중 게시된 기사 제목의 텍스트분석을 통해 알아보았다.

2022년 3월 7일 – 3월 8일 게시된 기사들을 크롤링하여 형태소 분석을 한 결과 후보들의 이름이나 기본 단어들을 제외하고 (김)만배, 부산, 녹취록, 민주당, 대장동, 저축은행 등이 높은 빈도로 이재명 후보의 기사 제목들에서 언급되었다. 윤석열 후보의 경우 (김)만배, 국민, 선거, 여야, 해결, 초보 등이 높은 빈도로 언급되었다. 두 후보 기사 제목에서 공통적으로 언급된 김만배, 녹취록, 대장동 세가지 키워드는 이재명 후보가 성남시장으로 있을 당시 대장동 개발과 관련하여 이재명의 비리를 언급하는 공격적 기사의 네거티브 키워드로 보인다. 윤석열 후보의 기사에서 언급된 이 세가지의 키워드는 윤석열 본인에 대한 네거티브 키워드라기보다는 이재명 후보를 공격하는 의도의 기사에서 네거티브적인 키워드들이 발견되었다고 볼 수 있다. 이는 대선을 바로 앞둔 이틀동안 보도된 뉴스들이 이재명 후보에 대한 공격적 네거티브 언론보도가 많았던 것으로 확인된다. 이러한 언론 보도는 뉴스를 소비하는 사람들로 하여금 이재명 후보에게 대단한 실격사유, 도덕적 결함이 있는 것처럼 프레임 효과를 조장하며, 이것이 중요한 대선의 의제로 떠오르게끔 의제 설정을 하였다고 볼 수 있다. 실제로 언론이 선거에서 어떤 의제가 중요한지 언론이 직접 설정하는 경향이 있고 이는 유권자들이 실제로 그러하다고 받아들일 수 있다는 것으로 보인다. 이러한 가설은 2022년 3월 9일 실시된 대선 결과 이재명 후보가 윤석열 후보에게 패배했다는 것을 확인함으로써 입증 가능하다.

본 연구에서는 형태소 분석과 TF-IDF 지수를 통한 텍스트 분석 기술을 이용하였다. 실험에서 진행한 형태소 분석으로 뉴스 제목의 키워드들을 추출할 수 있었으며 그 빈도수를 확인할 수 있었다. TF-IDF 지수는 랜덤으로 추출된 기사 제목들에서 어떤 단어들이 의미 있는 단어인지 확인할 수 있었다. 본 연구를 통해 본인은 생각보다 더 다양한 분야에서 텍스트 분석을 활용할 수 있으며 의미 있는 결과를 도출할 수 있다는 것을 확인하였다.