한국 정치에 관한 LDA와 BERTopic 비교 분석: '탄핵'을 중심으로

진민준

경북대학교

jmj1569@knu.ac.kr

Comparative Analysis of LDA and BERTopic on Korean Politics: Focusing on Impeachment Data

Min Joon Jin

Kyungpook National University

요 약

대규모 텍스트 데이터에서 유의미한 정보를 추출하기 위해 토픽 모델링이 널리 사용되고 있다. 대표적인 방법으로는 확률적 생성 모델인 LDA와 사전 학습된 언어 모델인 BERTopic이 있다. LDA는 계산이 비교적 간단하고 해석하기 쉬운 장점이 있지만, 토픽의 일관성이 낮다는 단점이 있다. 반면, BERTopic은 높은 토픽 일관성을 제공하지만, 계산량이 많아 복잡하다는 특징이 있다. 본 연구에서는 빅카인즈에서 제공하는 '탄핵'을 검색어로 하여 추출된 기사 데이터를 중심으로 LDA와 BERTopic의 토픽 모델링 결과를 비교 분석하였다. 각 방법에서 발견된 토픽은 탄핵 원인과 결과에 대한 유의미한 관계가 있었으며, 두 방법의 특성과 상호 보완 가능성을 파악하기 위해 정량적 분석과 정성적 평가를 수행하였다.

1. 서 론

빅데이터 시대의 도래와 함께 방대한 양의 데이터와 정보가 빠르게 유입되고 있다. 이러한 환경에서 텍스트 데이터를 분석하고 핵심 주제를 도출하는 토픽 모델링은 점점 더 주목받고 있다. 토픽 모델링은 문서 집합(corpus) 내 주제 정보의 분포를 파악할 수 있게 함으로써 정보 탐색 과정에서 시간과 노력을 효과적으로 절감하는 데 중요한 역할을 한다.

최근 비상계엄령 및 탄핵소추안 가결과 관련하여 짧은 기간 동안 다수의 기사와 칼럼이 쏟아져 나왔다. 이러한 방대한 정보 속에서 논의되고 있는 주제들을 효과적으로 파악하는 것은 매우 중요한 과제이다. 이를 위해 토픽 모델링은 대규모 데이터를 요약하고 주요 주제 및 키워드를 제공함으로써 유용한 도구로 작용한다. 본 연구는 LDA와 BERTopic을 활용하여 빅카인즈에서 제공하는 '탄핵'관련 기사 데이터를 대상으로 토픽 모델링 결과를 비교 분석하는 것을 목적으로 한다.

2. 이론적 배경

전체 문서 집합에서 나타나는 의미론 구조를 탐색하기 위해 사용되는 텍스트 마이닝 기법인 토픽모델링은 문서 집합 내에 숨겨져 있는 토픽을 다양한 통계적인 방법을 활용하여 표현하는 분석 방법이다[1]. 토픽모델링 기법에는 잠재 디리클레 할당(LDA), 잠재의미분석(LSA), 확률적 잠재의미분석(pLSA)등이 있다[2].

3. 연구 절차 및 방법

3.1. 연구 절차

본 논문에서는 대통령 탄핵 소추안 및 탄핵 연구 동향을 분석하고자 [그림 1]과 같은 절차로 연구를 수행하였다.

그림 1. 연구 절차

3.2. 연구 방법

3.2.1. 데이터 수집

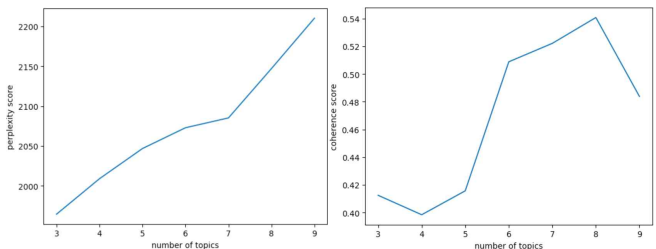
분석대상은 빅카인즈 서비스(https://www.bigkinds.or.kr/)에서 2024년 12월 1일부터 2025년 1월 08일까지 '탄핵'을 검색어로 하여 추출된 기사 데이터로 한정하였다. 기사의 일자, 언론사, 기고자, 제목, 본문등의 데이터를 수집하였다. 분석 시점인 2024년 12월 1일부터 2025년 1월 08일까지 기사를 대상으로 하였으며, 총 5,068건의 기사가 수집되었다.

3.2.2. 데이터 전처리

수집된 기사를 대상으로 파이썬을 기반으로 프로그래밍하여 특정 기호나 빈번하게 발생하는 '탄핵', '대통령', '윤석열', '방송', '앵커' 단어들을 불용처 처리하여 제거하였으며 기사의 제목과 본문을 결합한 후 okt 형태소분석을 활용해 명사, 동사, 형용사만을 추출하여 토픽모델링을 진행하였다.

3.2.3. LDA 토픽모델링 분석

5,068개의 기사에서 전처리한 데이터셋을 활용하여 LDA 기반의 토픽 모델링을 수행하여 탄핵에 관한 정치적 연구 동향을 분석하였다. 토픽모델링을 수행하기 위해 파이썬 기반의 gensim의 LDA 모듈을 사용하였다. 최적화된 토픽 수를 찾기 위해 토픽 수 3개부터 9개까지 하나씩 증가시키면서 coherence 값과 perplexity 값을 확인한 결과 [그림 2]와 같았다.



(a) perplexity 값 (b) coherence 값

그림 2. 토픽 수에 따른 perplexity 값과 coherence 값

LDA 알고리즘을 통해 6개의 토픽을 추출하고 시각화하였으며 토픽과

관련된 단어를 확인하였다.

3.2.4. BERTopic 토픽모델링 분석

5,068개의 기사에서 전처리한 데이터셋을 활용하여 BERTopic 토픽모델링을 수행하였다. Konlpy에서 제공하는 한국어 형태소 분석기 okt를 사용하여 명사, 동사, 형용사만을 추출하는 형태소 분석 및 토큰화를 진행하였다. CountVectorizer를 사용하여 텍스트 데이터를 토큰화를 진행한 후, 단어의 빈도수를 확인하였다. 토픽수는 10으로 설정하고 각 토픽에서 상위 10개의 단어들을 추출하였다.

4. 연구 결과

4.1. LDA 토픽모델링 결과 분석

LDA 토픽모델링을 활용하여 토픽의 개수를 6개로 설정하였으며, 6개의 토픽이 다양한 주제를 가지고 있음을 알 수 있다. 토픽 1은 한덕수 권한대행 체제에서 내란 특검법 공포 미집행에 대한 주제, 토픽 2는 계엄령 선포 및 계엄령 해제등 계엄령에 관한 주제, 토픽3 은 국회의원들의 탄핵소추안이 국회를 통과에 관한 주제, 토픽4는 최근 여론에서 탄핵을 찬성과 반대하는 시민들의 집회에 관한 주제, 토픽5는 계엄령에 대한 수사 및 조사에 관한 주제, 마지막으로 토픽6은 탄핵을 고려해서 조기대선이 치러지면 차기 대권주자에 대한 주제이다. 토픽별 키워드와 연구주제는 <표 1>과 같다.

표 1. LDA 토픽별 연구주제 및 주요 키워드

토픽	개수	연구주제	주요 키워드
1	521	한덕수 권한대행 특검법 공포 미집행	대행, 권한, 특검법, 한덕수, 내란, 본회의, 김건희
2	828	계엄령 선포 및 해제	선포, 계엄, 담화, 해제, 대국민, 정치, 한국
3	543	탄핵소추 국회 통과	표결, 대표, 한동훈, 추안, 여당, 퇴진, 찬성
4	253	탄핵 찬성과 반대 집회	집회, 계엄, 시민, 정치, 촉구, 지역, 정부
5	381	계엄령에 대한 수사 및 조사	내용, 진행, 뉴스, 바랍니다, 수사, 확인, 출연
6	926	탄핵을 고려한 조기대선에 대한 여론	여론조사, 이재명, 지지율, 대선, 찬성, 대표, 결과

4.2. BERTopic 토픽모델링 결과 분석

BERTopic 활용 토픽모델링으로 기사를 분할 때 10개의 토픽으로 설정하였고 각 토픽에 해당하는 기사 수가 높은 상위 10개의 주제이며 주요 키워드는 <표 2>와 같다.

표 2. BERTopic 토픽모델링 논문 수 상위 10개 주제

번호	주제	논문수	비율	주요 키워드
0	-1	2152	42%	국민, 국회, 의원, 비상계엄
1	0	1616	31%	대표, 한동훈, 민주당, 대행
2	1	988	19%	계엄, 비상계엄, 사태, 선포
3	2	209	4%	정치, 심판, 사태, 퇴진
4	3	31	0.6%	내일, 표결, 국회, 본회의
5	4	27	0.5%	불확실, 금융시장, 경제, 금리
6	5	20	0.3%	보고,본회의, 발의, 국회
7	6	10	0.19%	트럼프, 배넌, 미국, 당선인
8	7	8	0.15%	언론, 시위, 선포, 포고령
9	8	6	0.11%	촛불,집회,시민,광장

전체 기사의 42%에 해당하는 주제 -1은 탄핵연구와 관련된 기사들 중 어느 주제에만 치우치지 않고 전체적으로 포괄하는 기사를 모아 놓은 것이다. 전반적으로 '국민', '국회', '의원', '비상계엄'등의 키워드가 자주 등장함을 알 수 있다.

다음으로 주제 0은 '대표', '한동훈', '민주당', '대행' 등의 키워드를 포함하여 계엄령 사태에 대한 한동훈 대표의 입장, 민주당의 입장, 권한대행 체제와 관련 있는 내용임을 알 수 있다. 다음으로 많이 나오는 주제 1은 '계엄', '비상계엄', '사태', '선포' 등의 키워드가 빈번하게 등장하여 계엄령 선포에 관한 기사 내용임을 알 수 있다. 주제 2는 '정치', '심판', '사퇴', '퇴진' 등의 키워드를 포함하여 계엄령에 관해 대통령에 대한 사퇴 및 심판과 관련 있는 내용으로 구성됨을 알 수 있다.

그 밖에는 전체의 4%미만에 해당하는 토픽으로 빈도가 낮고 중요한 부분은 아니지만 계엄령과 대통령 직무정지로 인해 금융시장 불확실, 경제 불안정과 관련된 연구의 흐름을 파악할 수 있다.

10개의 토픽간 관계성을 나타내는 토픽 간 거리 지도를 살펴보면 다음과 같이 4개의 그룹이 가깝게 형성되는 것을 볼 수 있다. [그림 3]은 각 그룹을 구성하는 토픽들을 의미하는 것으로 토픽간의 관계를 잘 파악할 수 있다.

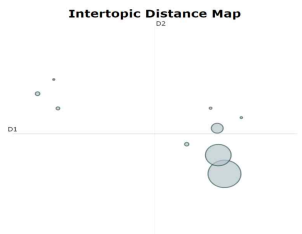


그림 3. BERTopic 토픽 간 거리 지도

5. 결론

LDA와 BERTopic은 레이블이 없는 텍스트 데이터에서 토픽을 추출하기 위한 대표적인 비지도 학습 기법이다. 특히, LDA는 확률적 생성 모델에 기반하여 간단하고 구현이 용이하다는 장점을 가진다. BERTopic 모델은 사전 학습된 언어 모델을 활용하기 때문에 문맥정보를 반영한 토픽 모델링을 통해 높은 토픽 일관성을 제공할 수 있다. 빅카인즈에서 수집한 '탄핵' 기사 데이터에 대한 토픽모델링 결과 실제 계엄령부터 탄핵 소추안 가결까지의 상황을 토픽이 인지하였음을 확인하였다. 특히 LDA는 계엄령, 탄핵, 수사, 조기대선까지 다양한 토픽을 추출하였으며 BERTopic은 계엄령, 탄핵에 관한 주제에 집중해서 토픽을 추출하였다.

참 고 문 헌

[1] 박종도(2019). 토픽모델링을 활용한 다문화 연구의 이슈 추적 연구. 한국문헌정보학회지, 53(3), 273-289.

[2]] 문진주. 토픽모델링을 활용한 회계교육 연구 동향 분석. 경영교육연구, 38(1), 67-88. 2023.