

텍스트마이닝으로 읽는 북한의 담론 변화: 노동신문·조선중앙통신 분석 사례를 중심으로

임수진 | 국가안보전략연구원 부연구위원 | happysjlim@gmail.com / sjlim@inss.re.kr

I. 들어가며

북한 연구에서 노동신문과 조선중앙통신 같은 공개 매체는 정권의 공식 입장과 대내외 선전 전략을 확인할 수 있는 사실상 유일한 창구다. 그러나 이 매체들은 객관적 사실 전달보다 주민 설득과 대외 선전이라는 목적을 위한 선전물이기 때문에 문자 그대로 읽을 경우 정권의 실제 의도를 파악하기는 어렵다. 지면에 반복적으로 등장하는 상투적 수사와 실제로 강조되는 정책 사이에 괴리가 존재하기 때문에, 북한의 공식 담론이 갖는 맥락과 함의를 해석할 때 연구자의 정성적 문헌 독해와 판단이 무엇보다 중요하게 인식되어 왔다. 특히 특정 표현이 등장한 정치적 맥락과 발화 주체, 표현의 강도와 변화 등은 연구자의 통찰을 통한 정성적 분석이 핵심적인 것으로, 이는 제한적으로 공급되는 국내 북한 문헌 연구에서 거의 유일무이한 분석 방법이었다 할 수 있다.

그러나 이러한 정성적 접근은 연구자의 경험과 전문성에 크게 의존한다는 점에서 분석 결과의 객관성과 재현 가능성에 일정한 한계를 갖는다. 동일한 자료를 대상으로 하더라도 연구자의 관점이나 문제의식에 따라 중요하게 포착되는 표현과 해석이 달라질 수 있으며, 분석 대상이 대규모·장기간의 텍스트로 확대될수록 개별 연구자의 독해만으로 담론의 변화 양상을 체계적으로 파악하기도 어렵다. 이러한 한계를 보완하기 위해 북한 연구에서도 텍스트를 정량적으로 처리하는 연구방법이 활용되기 시작했다. 특히 노동신문 등 주요 매체에 대한 국내 디지털화가 수행되면서, 대표적으로 텍스트마이닝(text mining) 분석 방법이 활용될

수 있게 되었다. 텍스트마이닝은 방대한 문헌에서 단어의 빈도와 관계, 시기별 증감과 같은 반복적 패턴을 계량적으로 추출함으로써 연구자 판단의 근거를 보다 체계적으로 확인하고 검증하는 분석 도구로 활용되고 있다.

본고에서는 이러한 문제의식 아래, 노동신문·조선중앙통신 기사를 대상으로 단어빈도(TF)와 역문서빈도(TF-IDF), 워드클라우드, 바이그램(bi-gram) 네트워크 분석, 그리고 감성분석 등 비교적 접근성이 높은 텍스트마이닝 기법을 적용해 김정은 시대 대내외 담론의 어조 변화와 정권의 의도를 분석한 네 건의 사례 연구를 소개한다. 네 사례 모두 필자가 수행한 연구로, 대상 시기와 세부 주제는 다르지만 표면적 수사와 실질적으로 강조되는 내용 사이의 간극을 텍스트마이닝을 통해 확인하려 했다는 문제의식에서 공통점을 지닌다.

이는 몇 해 전 본지에서 발표한 최지영·이혜진(2023)의 논의와도 맥이 닿아 있다. 이들 연구 역시 북한의 공식 문헌을 1차 자료로 삼아, 정성적 독해만으로는 놓치기 쉬운 단어 간 관계와 시기별 변화를 정량적으로 검증하였다. 다만, 이 논의의 경우 LASSO, DOC2VEC과 같은 방법을 통해 경제 제도의 변화를 추적했다면, 이번에 소개하는 사례들은 북한 정치외교 및 사회 분야에서 생산된 선전 담론을 중심으로, TF-IDF, 워드클라우드, 네트워크 분석, 감성분석 등 비교적 단순한 기법을 통해 유의미한 정책적 함의를 도출할 수 있었다는 점을 강조하고자 한다. 다시 말해 텍스트마이닝의 유용성이 가설 설계와 정성적 해석 역량과 상당 부분 연계되어 있다는 점을 논하고자 한다.

본고는 다음과 같이 구성된다. II장에서는 네 개의 사례 연구 각각이 어떤 자료와 기법을 사용해 어떤 결론을 내렸는지 소개한다. III장에서는 네 사례에서 사용된 텍스트마이닝 분석에서 도출된 방법론적 시사점과 고려사항을 설명한다. 끝으로 IV장에서는 향후 연구과제를 제안한다.

II. 텍스트마이닝을 활용한 북한 담론 분석 사례

1. 김정은 시대 대외 적대담론의 변화: ‘미제’·‘일제’·‘남조선괴뢰’ 언급빈도에 대한 시계열 분석(임수진, 2022)

첫 번째 사례는 김정은 정권 출범 이후 10년간(2012~21년) 노동신문 기사 제목에 나타난

‘미제’, ‘일제’, ‘남조선괴뢰’라는 세 전통적 적대어의 언급 빈도를 연도별로 집계하고, 이를 시계열로 살펴봄에 김정은시대 북한 대외 적대 담론의 구조적 변화를 추적한 연구다. 동 연구는 형태소 분석과 같은 통계적 가중치 기법을 사용하지 않고, 사전에 설정한 핵심어(‘미제’ · ‘반미’ · ‘항미’, ‘일제’ · ‘반일’ · ‘항일’, ‘남조선괴뢰’ 및 정부비판 관련 어휘 등)가 기사제목에 등장한 횟수를 그대로 집계하는 빈도분석 방식을 사용했다. 동 연구의 경우 필자의 박사학위 논문 일부를 보완 · 발전시킨 것으로, 북한의 집단주의(또는 전체주의)적 특성에 기초한 공동체적 타자인식과 그 적대성 변화에 주목하여 분석을 진행하였다.

연도별 빈도를 보면, 북한의 미국에 대한 적대성을 함의하는 어휘로서 ‘미제’(반미 · 항미 포함)에 대한 언급은 2017년 148건으로 정점을 찍은 뒤 2018년 이후 가파르게 줄었는데, 2019~21년 3년 연속 기사 제목에서 ‘미제’ 어휘가 단 한 차례도 등장하지 않았다는 점을 특이점으로 볼 수 있다(표 1). 이는 김정은 정권이 2018년 국제 사회 진출을 도모하는 과정에서, 이념적 · 부정적 함의를 가진 ‘미제’의 사용을 최소화함으로써 미국과 일정한 거리를 유지하면서도 향후 관계 변화의 가능성을 열어두는 전략적 담론 기조를 취했음을 보여준다.

남한에 대한 적대성을 내포한 ‘남조선괴뢰’ 표현의 경우, 2018~19년 남북정상회담 국면까지는 보수 정치세력과 군부를 지칭할 때 제한적으로 사용되었으나, 2019년 하노이 북미정상회담 결렬 이후 다시 증가하기 시작했고, 특히 2020년 3월 김여정의 담화를 계기로 그 사용이 두드러짐이 확인되었다. 이 과정에서 주목되는 변화는 대남 비난의 대상이 보수정권이나 군부를 넘어 유화적인 대북정책을 펼친 진보성향 정부로까지 확대되었다는 점이다.

〈표 1〉 노동신문 기사 제목 내 ‘미제’ · ‘반미’ · ‘항미’ · ‘미국’ 언급 빈도(2012~21)

연도	미제	반미	항미	미국
2012	70	34	0	340
2013	92	38	0	425
2014	70	27	0	394
2015	60	27	0	365
2016	47	27	0	278
2017	148	75	0	559
2018	28	20	0	481
2019	0	12	0	452
2020	0	4	1	347
2021	0	0	0	2

이는 향후 북한의 대남 적대화정책의 강화 가능성을 살펴볼 수 있는 하나의 단서를 제공해 준다.

아울러 노동신문과 조선중앙통신에 게재된 김여정의 대남 담화를 비교한 결과, 대내 매체인 노동신문에서는 조선중앙통신에 비해 대남 비난 담화가 상대적으로 제한적이고 선별적으로 나타나는 차이가 확인되었다. 이는 북한의 대남 담론이 모든 매체에서 동일하게 전달되는 것이 아니라, 대내외 독자와 선전 목적에 따라 담화의 내용과 노출 수준이 조정되고 있음을 보여준다. 이는 특정 적대 어휘의 빈도 변화뿐 아니라 해당 담론이 어떤 매체를 통해, 어떤 독자를 대상으로 전달되는지를 함께 분석할 필요가 있음을 시사한다.

동 연구는 북한 문헌에서 자주 사용되는 소수의 핵심어에 대한 문맥상의 이해와 통찰을 통해 이를 장기간에 걸쳐 시계열적으로 비교·분석하는 것만으로도 정권의 담론 전략 변화를 포착할 수 있음을 보여준다. 다만, 이러한 방식은 연구자가 사전에 선정한 핵심어를 중심으로 분석이 이루어진다는 점에서, 분석 범위가 연구자의 판단에 의해 제한될 수 있다는 한계도 상정하고 있다. 이에 연구는 핵심어의 빈도 변화만을 해석하는 데 그치지 않고 관련 기사의 내용을 함께 검토함으로써, 기존의 특정 대상을 지칭하는 표현(예를 들어, ‘미제’)이 감소하는 한편, 특정 대상을 명시하지 않는 표현(예를 들어, ‘제국주의자’)이 2018년 이후 증가하는 현상을 함께 포착하기도 하였다. 즉, 빈도분석은 주요 표현의 증가와 감소를 직관적으로 보여주는 데 유용하지만, 특정 표현의 감소나 소멸이 어떠한 표현으로 대체되는지를 파악하기 위한 연구자의 정성적 해석이 함께 이루어질 필요가 있는 방식이라고 할 수 있다.

2. ‘적대적 두 국가론’ 이후 대남 보도 변화 분석(임수진, 2025)

두 번째 사례는 2023년 말 북한이 ‘적대적 두 국가론’을 공식화한 이후 노동신문과 조선중앙통신의 남한 관련 보도가 어떻게 변화했는지를 2024년 1월부터 2025년 3월까지의 기사를 대상으로 분석한 연구다. 동 연구는 기사 제목 속 ‘한국’ 또는 ‘남조선’ 언급 빈도를 연도별·월별로 집계하는 단순 빈도분석에서 출발해, 특정 시기(예: 2024년 10월, 2025년 1월)에 집중적으로 게재된 기사를 대상으로 워드클라우드와 바이그램 네트워크 분석을 수행하는 방식으로 진행되었다.

이를 통해 뚜렷한 변곡점들이 확인되었다. 노동신문 기사 제목에서 ‘한국’ 또는 ‘남조선’ 언급 빈도는 2023년 한 해 동안 급격히 감소했으나, ‘적대적 두 국가론’ 선포 이후인 2024년에는 급격히 증가하는 양상이 확인되었다. 이후 2024년 10월 국내 정치 갈등이 고조된 시점과

〈표 2〉 노동신문 ‘한국’ 관련 기사의 월별 키워드 순위(2024.10~2025.1)

연월	1순위	2순위	3순위	4순위
2024. 10	윤석열	괴뢰	한국	전쟁
2024. 11	윤석열	탄핵	괴뢰	전쟁
2024. 12	윤석열	미국	괴뢰	지역
2025. 01	괴뢰	윤석열	미국	세계

2024년 12월 계엄·탄핵 국면을 전후해 보도량이 크게 요동쳤는데, 특히 계엄 선포 이후 오히려 관련 기사 수가 감소하는 현상이 관찰되어, 단순한 ‘갈등 고조 = 보도량 증가’라는 직관적 가설만으로는 설명되지 않는 패턴이 나타났다.

이 지점에서 워드클라우드와 바이그램 네트워크 분석이 보완적 역할을 했다. 특정 월의 기사 제목·본문에서 자주 함께 등장하는 단어들을 시각화한 결과, 시기가 지날수록 ‘한국’과 함께 등장하는 핵심어가 국내 정치 이슈(탄핵, 정권)에서 미국·군사·전쟁 관련 어휘로 옮겨 가는 흐름이 확인되었다. 즉, 단순 빈도만으로는 언급량이 늘거나 줄었다는 사실만 파악할 수 있지만, 함께 등장하는 단어들의 네트워크 구조를 살펴봄으로써 무엇에 대해, 누구를 겨냥해 언급량이 변화했는지까지 포착할 수 있었다. 이러한 분석을 통해 동 연구는 북한의 대남 보도가 남한 정부에 대한 직접적 비난에서 벗어나 한미 동맹과 미국의 군사적 역할을 비판하는 방향으로 초점을 옮기고 있으며, 이 과정에서 남한의 존재감을 희석시키려는 ‘전략적 무관심’의 기조가 자리 잡고 있다는 근거를 제시할 수 있었다.

〈표 2〉는 2024년 10월부터 2025년 1월까지 노동신문 기사 제목 속 ‘한국’이 언급된 기사 속에서 빈도가 가장 높았던 상위 키워드를 월별로 정리한 것이다. 표에서 확인되듯 2024년 10월, 11월에는 ‘윤석열’·‘탄핵’ 등 국내 정치 관련 어휘가 상위권을 차지했으나, 12월 이후로는 ‘미국’·‘지역’·‘세계’ 등 대외·안보 관련 어휘의 비중이 빠르게 커졌다. 불과 두 달 사이에 상위 키워드의 성격이 국내 정치에서 대외 안보로 옮겨간 것으로, 이는 앞서 서술한 워드클라우드·네트워크 분석 결과와 같은 방향을 가리킨다는 점에서 상호 검증적 의미를 지닌다.

3. 대내 선전 담론의 이면: ‘대동강’ 기사에 대한 TF-IDF 분석

세 번째 사례는 김정은 집권 이후 노동신문에 게재된 ‘대동강’ 관련 기사(2015~23년, 총 1,695건)를 대상으로 단어빈도(TF)와 역문서빈도(TF-IDF) 분석 결과를 비교해, 대동강 개발 담론의 표면적 내용과 실제 강조점 사이의 괴리를 확인한 연구다.

동 연구의 핵심적인 방법론은 TF 분석과 TF-IDF 분석을 교차 비교했다는 점에 있다. TF 분석이 전체 텍스트에서 반복적으로 등장하는 어휘의 빈도를 통해 담론의 전반적인 관심 대상과 주요 주제를 파악하는 데 초점을 둔다면, TF-IDF 분석은 특정 문서에서 상대적으로 두드러지는 어휘에 가중치를 부여함으로써 개별 시기나 정책 국면을 특징짓는 변별력 높은 표현을 식별하는 데 집중한다. 이에 동 연구는 두 분석 결과를 비교함으로써 담론 전반에서 지속적으로 반복되는 핵심어와 특정 시기 또는 사건을 계기로 부각되는 특징적 핵심어를 구분하고, 단순 빈도분석만으로는 포착하기 어려운 내용상의 강조점과 그 변화를 살펴보았다.

단순 단어빈도(TF) 기준으로는 대동강 관련 기사에서 가장 많이 언급된 단어는 ‘인민’이었고, 이는 ‘인민을 위한 도시 개발’이라는 정권의 공식 서사와 일치하는 결과였다. 그러나 단순한 출현 빈도가 아니라, 각 연도의 문서에서 상대적으로 강조되는 단어를 식별하는 TF-IDF 분석을 적용하자 결과는 달라졌다. 연도별 TF-IDF 상위 단어는 ‘공장’, ‘생산’, ‘건설’, ‘농장’ 등으로, 이는 생산활동과 관련된 표현들로 판단할 수 있었다.

즉, ‘인민’의 경우 모든 대동강 기사에 공통적으로 등장하는 관용구에 가까워 TF-IDF 분석을 통해 볼 때 변별력이 낮은 어휘로 평가된 반면, ‘공장’·‘생산’ 등은 시기별로 변화가 나타나는 실질적 핵심어로 판단할 수 있었다. 여기에 바이그램(bi-gram) 네트워크 분석을 더해 ‘인민’ 관련 어휘군과 ‘공장’·‘생산’·‘건설’ 관련 어휘군이 문장 구조상 서로 분리된 군집을 이루고 있다는 점까지 확인하였다. 이에 동 연구는 노동신문의 ‘대동강’ 관련 기사가 ‘인민을 위한 여가 공간’이라는 표면적 이미지 아래 실제로는 노동력 동원을 정당화하고 있다는 분석 결과를 제시할 수 있었다. 이는 단순 빈도분석에서는 특정하기 힘든 상투어와 핵심어를 구분하여 분석함으로써, 표면적으로 반복되는 선전적 표현을 넘어 북한의 선전 서사와 그 이면에 내재된 동원의 내용과 방향을 구체적으로 확인할 수 있었다는 점에서 의의가 있었다 할 수 있다.

〈표 3〉 노동신문 ‘대동강’ 관련 기사 연도별 TF-IDF 상위 키워드(2015~23년)

연도	1순위	2순위	3순위	4순위
2015	아이	공사	생산	공장
2016	장군	공장	생산	건설
2017	공장	장군	생산	사업소
2018	공장	장군	생산	통일
2019	공장	대학	건설	인민
2020	공장	당원	공사	장군
2021	총비서	공장	인민	생산
2022	인민	총비서	민족	공장
2023	농장	공장	농사	생산

4. 감성분석을 통한 대남 담화의 어조 변화 분석

마지막 사례는 조선중앙통신이 발표한 북한의 대남 담화 및 성명 476건(2011년 12월~2026년 7월)을 대상으로 감성분석을 실시해, 김정은 집권 이후 남북관계 주요 국면과 대남 담화의 어조 변화를 교차 분석한 연구다. 앞선 사례들이 단순 빈도 기반 분석을 수행한 것과 달리, 동 연구는 긍정·부정 어휘를 5단계(강한 부정-약한 부정-중립-약한 긍정-강한 긍정)로 분류한 감성사전(총 449개 어간)을 자체 구축하고, kiwi 형태소 분석기로 담화 원문을 형태소 단위로 정규화한 뒤 문서별 감성점수를 산출하는 방식을 취하였다.

동 연구의 경우 북한어 특유의 비대칭성, 즉 부정 어휘(297개)가 긍정 어휘(136개)보다 약 2배 많은 사전적 긍정·부정 어휘 구성의 불균형을 감안하여 z-score 표준화를 통해 보정 작업을 추가 수행하였다. 표준화를 수행하지 않은 어휘사전을 통한 감성분석을 통해 답론의 긍정·부정 기조를 비교할 경우 구조적으로 부정 기조 쪽에 유리하게 치중될 왜곡 가능성이 존재하므로, 문서별 감정어 밀도(형태소 1,000개당 출현 수)를 말뭉치 전체 평균·표준편차 기준으로 표준화한 뒤 재합산하는 방식으로 이를 보정하였다. 실제로 보정 전 기준으로는 전체 476건의 문서 중 부정(강한 + 약한) 문서가 66.4%에 달했으나, z-보정 이후에는 34.0%로 크게 줄어들어, 사전 구성의 비대칭성이 분석 결과에 상당한 영향을 미칠 수 있음을 확인할 수 있었다.¹⁾

동 연구에서 중요한 질문은 감성점수의 변화가 실제 남북관계의 주요 국면과 어느 정도

1) z-보정의 경우, 긍정·부정 기조의 편차 그 자체보다, 답론 기조의 변화 양상에 주목하여 분석 결과를 살펴보는 것에 초점이 맞추어져 있다. 자세한 내용은 임수진(2026)의 글을 참조.

〈표 4〉 조선중앙통신 대남 담화 연도별 평균 감성점수(z-보정) 주요 국면

연도	감성 점수(z-보정)	주요 국면 · 사건
2011	-0.24	김정일 조문 불허, 집권 초기 대남 강경 기조
2012	-0.46	핵 · 미사일 도발, 정전협정 백지화 위협
2018	+1.50	평창올림픽 참가, 판문점 선언
2019	+0.72	싱가포르 북미정상회담, 하노이 회담(결렬 이전 포함)
2022	-0.24	‘담대한 구상’ 거부, 인신공격성 담화 증가
2023	-0.07	‘적대적 두 국가론’ 공식화 준비
2024	-0.40	적대적 두 국가론 선포, 오물풍선 · 무인기 사건(최저치)
2025	+0.05	확성기 · 대남 소음방송 상호 중단
2026	+0.10	긴장관리 국면 지속

일치하는지, 그리고 최근 대남 담화의 부정 기조가 과거와 비교해 어느 수준에 있는지를 확인하는 것이었다. 분석 결과, 2018~19년 남북 · 북미 대화 국면에서는 긍정적 어조가 뚜렷하게 확대된 반면, 2022년 이후에는 부정적 어조가 강화되었고 ‘적대적 두 국가’ 기조가 본격화된 2024년에는 가장 낮은 감성점수를 기록하는 등 감성점수의 변화가 전반적으로 남북관계의 주요 국면과 일정한 연관성을 보였다.

동 연구의 주요한 발견은 긍정 담화와 부정 담화가 서로 다른 시계열적 성격을 갖는다는 점이다. 부정 담화는 ‘적대적 두 국가론’과 같은 정책 선언 이후 이를 지속 · 강화하는 방향으로 나타나는 반면, 긍정 담화의 확대나 부정 어조의 완화는 오히려 대화 재개나 유화조치에 선행하는 경향을 보였다. 예컨대 2018년 1월 평창올림픽 참가 표명 직전인 2017년 11~12월에 감성점수가 이미 상승했던 반면, ‘적대적 두 국가론’ 선언(2023년 12월) 이후 2024년 1월부터 감성점수가 본격적으로 하락하는 등, 긍정 신호는 정책 전환에 선행하는 ‘조기지표’로, 부정 신호는 이미 선포된 기조를 뒤따르는 ‘반응지표’로서 기능할 가능성을 시사했다.

또한 동 연구는 2024년의 부정 강도가 김정은 집권 초기(2011~12년)의 대남 강경 국면보다 전면적으로 심화되었다고 보기는 어렵다는 점을 함께 지적한다. 이는 담화에 나타난 적대성의 정도를 실제 남북관계의 긴장 수준과 동일시하기보다, 장기적인 차원의 시계열적 분석과 함께 정책 및 행동의 변화를 살펴보며 종합적으로 해석할 필요가 있음을 시사한다. 따라서 감성점수를 통한 담론 분석은 시계열적 변화와 실제 행동 및 정책 변화와의 교차 분석을 통해, 남북관계의 국면과 변화 방향을 파악하고 향후 관계 변화 가능성을 관찰하는 주요한 지표로서 활용될 수 있다.

5. 종합: 표면적 서사와 실질적 강조 사이의 간극을 확인하기

제시한 네 가지 분석 사례를 통해 공통적으로 확인할 수 있는 점은, 북한 공식 매체가 표면적으로 제시하는 담론과 실제로 강조하는 내용 사이의 간극을 텍스트마이닝을 통해 구체적으로 포착할 수 있다는 점이다. 대외 적대 담론에서는 ‘미제’와 같은 직접적인 이념적 표현이 감소하는 동시에 ‘제국주의자’와 같이 대상을 특정하지 않는 표현이 증가하면서 적대 담론의 표현 방식이 변화하는 양상이 나타났다. 대남 보도에서는 남한을 직접적인 비난 대상으로 삼는 보도가 감소하는 한편, 남한을 미국의 군사적 위협과 결부시키는 방식으로 대남 적대성을 재구성하는 변화가 확인되었다. 대동강 관련 기사에서는 ‘인민을 위한 개발’이라는 공식적 서사가 강조되는 가운데 실제 기사에서는 건설과 개발을 위한 노동력 동원과 주민의 역할이 반복적으로 부각되었다. 마지막으로 대남 담화 감성분석에서는 ‘적대적 두 국가론’이라는 정책 기조가 유지되는 상황에서도 담화의 부정적 어조가 일부 완화되는 양상이 나타나, 정책 기조와 담화의 표면적 정서가 반드시 동일한 방향으로 움직이지 않을 수 있음을 보여주었다.

이러한 결과는 북한 공식 담론을 분석할 때 특정 표현이나 정책 구호의 표면적 의미만으로 정책 방향을 판단하기 어렵다는 점을 시사한다. 오히려 어떤 표현이 감소하고 무엇으로 대체되는지, 비난의 대상과 논리구조가 어떻게 재구성되는지, 공식적으로 제시된 목적과 실제로 반복 강조되는 내용이 무엇인지, 그리고 정책 기조와 담화의 정서적 변화가 일치하는지를 함께 살펴볼 필요가 있다. 이러한 간극을 체계적으로 가시화한다는 점에서 텍스트마이닝은 북한 공식 담론의 표면적 서사를 넘어 그 내부의 강조점과 변화 양상을 포착하는 분석 수단으로 활용될 수 있다.

〈표 5〉 분석 사례 종합

구분	분석 대상	주요 방법론	주요 발견
사례 1	『노동신문』 ‘미제’·‘일제’·‘남조선괴뢰’ 관련 기사 (1950~2021, 특히 2012~21)	장기 시계열 단순 빈도분석 정성적 내용분석	‘미제’는 소멸, ‘미국’으로 대체되는 한편, ‘남조선괴뢰’는 보수·진보 정부 모두로 대상 확대
사례 2	『노동신문』·『조선중앙통신』 남한 관련 기사(2024. 1~2025. 3)	빈도분석, 워드클라우드, 바이그램 네트워크	대남 비난 초점이 국내 정치에서 한미·군사 이슈로 이동
사례 3	『노동신문』 ‘대동강’ 관련 기사 (2015~23, 1,695건)	TF·TF-IDF 비교, 워드클라우드, 바이그램 네트워크	표면적 ‘인민’ 서사 이면에 노동력 동원 담론이 실질적 강조점
사례 4	『조선중앙통신』 대남 담화·성명 (2011. 12~2026. 7)	감정사전 기반 감성분석, z-score 표준화, 시계열 비교	긍정 어조는 관계 개선에 선행, 부정 어조는 정책 선언을 지속·강화(담론-행동 비대칭)

이러한 간극은 우연이 아니라 선전물이라는 자료의 근본적 성격에서 비롯되는 것으로 볼 수 있다. 프로파간다는 수사법·신화·상징이라는 세 층위를 통해 수용자를 설득하도록 구성되는데(오쇼네시 저, 박순석 역, 2009), 표면에 드러나는 상징적 수사와 그 이면에서 실제로 관철되는 정책적 목적은 애초에 일치하도록 설계되지 않는 경우가 많기 때문이다. 이 점에서 텍스트마이닝의 다양한 분석 방법은 방대한 선전 텍스트 속에서 이 간극을 경험적으로 포착하는 데 특히 유용한 도구라 할 수 있다.

Ⅲ. 북한 문헌 텍스트마이닝의 분석 전략과 고려사항

1. 북한 문헌의 선전 특성을 고려한 분석 방법의 적용

제시한 사례 모두 분석의 출발점은 노동신문·조선중앙통신 기사를 대상으로 분석 단위를 확정하고 불필요한 요소를 걸러내는 전처리 작업이었다. 대부분의 사례에서 형태소 단위로 문장을 분해하고 조사·어미 등 불용어를 제거하는 절차를 거쳤다. 북한어는 한자어 결합이 많고 남한과는 다른 맞춤법·띄어쓰기 규범을 사용하기 때문에, 남한어 기준으로 구축된 형태소 분석기를 그대로 적용할 경우 명사 추출 정확도가 떨어지는 문제가 있다. 그중 세 사례는 NKPro와 같은 북한 매체 전용 크롤링·데이터베이스 도구를 활용해 원문을 수집하고, 이를 형태소 분석을 통해 정제하는 과정을 거쳤다. 첫 번째 사례의 경우 형태소 분석 대신, 사전에 정의한 핵심어가 기사 제목에 등장하는지 여부만을 확인하는 단순 검색 방식을 취해 전처리 부담을 크게 줄였다는 차이가 있다. 마지막 사례(감성분석)의 경우, kiwi 형태소 분석기로 담화 원문을 형태소 단위로 정규화한 뒤, 북한어 특유의 어휘를 반영해 자체 구축한 5단계(강한 부정~강한 긍정) 감정사전 449개 어간을 매칭하는 방식으로 전처리와 분류를 결합했다는 점에서 나머지 사례들과 구별된다.

또한 북한 선전물의 특성상 최고지도자에 대한 상투적 찬양구나 체제의 우월성을 강조하는 관용적 표현이 시기와 주제에 관계없이 문서 전반에서 반복적으로 등장한다는 점도 고려해야 한다. 북한의 공식 담론은 사실이나 정보를 전달하는 기능뿐 아니라 지도자의 권위와 정당성을 재확인하고 주민들에게 특정한 가치와 행동규범을 지속적으로 주입하는 정치적 기능을 수행하기 때문에, 일정한 어휘와 표현이 반복적으로 사용되는 경향이 강하다. 이로 인해 단순 빈도분석

을 적용할 경우 ‘위대한’, ‘경애하는’, ‘령도’, ‘인민’, ‘사회주의’ 등 북한의 공식 문헌에서 관행적으로 사용되는 어휘가 분석 시기나 정책적 중요성과 관계없이 상위에 나타날 가능성이 높다. 문제는 이러한 고빈도 어휘가 북한 선전 담론의 기본적인 특성을 보여주는 데에는 의미가 있지만, 이를 통해 특정 시기에 새롭게 부각된 정책적 관심이나 주민들에게 강조하고자 하는 동원의 내용, 선전 서사의 변화를 파악하는 데에는 한계가 있다는 점이다. 특히 시기별 담론 변화를 분석하는 경우에는 단순히 ‘어떤 단어가 많이 등장했는가’뿐 아니라 ‘평소에도 자주 사용되는 단어인지, 아니면 특정 시기에 상대적으로 두드러지게 등장한 단어인지’를 구분하는 작업이 중요하다.

따라서 반복적으로 등장하는 상투어의 영향력을 낮추고 특정 문서나 시기에 상대적으로 중요성이 높아진 어휘를 식별하기 위해서는 TF-IDF와 같은 가중치 기법을 병행할 필요가 있다. 이를 통해 단순 빈도만으로는 드러나기 어려운 시기별 선전 서사의 변화와 그 이면에서 강조되는 정책적 메시지 및 동원의 방향을 보다 구체적으로 파악할 수 있다. 세 번째 사례에서 확인할 수 있듯이, TF와 TF-IDF는 서로 다른 차원의 정보를 제공한다는 점에서 두 분석을 비교·병행할 때 해석의 폭을 넓힐 수 있다. TF 분석이 특정 시기나 주제에서 가장 빈번하게 사용된 단어를 통해 담론의 전반적인 구성과 반복되는 서사를 보여준다면, TF-IDF 분석은 여러 문서와 시기에 걸쳐 반복적으로 등장하는 어휘의 영향력을 상대적으로 낮추고 특정 시기나 문서군에서 두드러지게 나타나는 특징적 어휘를 식별하는 데 유용하다. 따라서 두 분석 결과를 나란히 비교하면 북한이 선전물에서 반복적으로 제시하는 주요 서사와 특정 시기에 상대적으로 강조하는 정책적 메시지가 어떻게 같거나 달라지는지를 확인할 수 있으며, 이를 통해 시기별 선전의 강조점과 동원 방향의 변화를 보다 입체적으로 파악할 수 있다.

2. 시계열 분석을 통한 정책적 전환의 포착

앞서 제시한 분석 사례 모두는 특정 시점(김정은 집권 초기, 2018년 남북·북미 정상회담, 2019년 하노이 회담 결렬, 적대적 두 국가론 선포, 계엄·탄핵 국면 등)을 전후한 연도별·월별 비교를 핵심적인 분석 틀로 삼았다. 특히 마지막 사례는 2011년부터 2026년까지 15년에 이르는 연도별 감성점수 추이를 통해, 담화 기조의 전환점이 실제 정책·사건의 발생 시점에 앞서는지 뒤따르는지를 월 단위로 세밀하게 비교했다는 점에서 시계열 비교의 밀도를 한층 높인 사례라 할 수 있다. 이는 유의미한 정량적 차이가 나타나는 시점을 사전에 하나로 확정하기보다 여러 후보 시점을 두고 이들을 비교함으로써 실제 전환점을 찾아가는 방식이

효과적이라는 공통된 시사점을 준다.

실제 연구 사례를 살펴보면, 분석 초기 설정한 시기 구분이나 비교 대상 단어가 실제 분석에서도 그대로 유의미한 결과로 이어지는 경우는 많지 않았다. 첫 번째 사례에서는 김정은 집권 초기를 기준으로 한 구분과 2018년 국제 사회 진출 시점을 기준으로 한 구분을 함께 살펴본 끝에, ‘미제’ 언급이 사라지는 시점과 ‘남조선괴뢰’ 언급이 다시 증가하는 시점이 서로 다르다는 사실을 확인할 수 있었다. 두 번째 사례의 경우 계엄 선포 시점, 탄핵소추안 가결 시점, 헌법재판소 결정 시점 등 여러 후보 국면을 두고 비교한 끝에, 계엄 이후 보도량이 오히려 감소한다는 반직관적 패턴을 확인할 수 있었다.

세 번째 사례에서도 김정은 집권 초기(2011~12년)를 기준으로 한 시기 구분과, ‘인민대중제 일주의’ 구호가 전면화된 시점을 기준으로 구분을 각각 시도한 뒤에야 TF와 TF-IDF 결과의 괴리가 뚜렷하게 드러나는 지점을 특정할 수 있었다. 네 번째 사례 역시 초기에 설정한 시기 구분만으로는 감성점수의 변화와 남북관계 국면 간의 관계를 충분히 설명하기 어려웠다. 이에 연도별 감성점수의 추이를 주요 남북관계 사건 및 정책 변화와 반복적으로 교차 검토한 결과, 2018~19년 대화 국면에서 긍정적 어조가 확대되고 2022년 이후 부정적 어조가 강화되며 2024년 감성점수가 가장 낮은 수준에 이르는 흐름을 확인할 수 있었다. 또한 2024년 담화에서 나타나는 높은 부정 기조(낮은 감성점수)의 경우 김정은 집권 초기 남북 사이 강경 국면과 비교할 때 그 정도와 의미를 신중하게 해석할 필요가 있음을 함께 확인하였다. 이는 사전에 설정한 시기 구분에 분석 결과를 맞추기보다, 실제 데이터에서 나타나는 변화 지점을 확인하고 이를 주요 사건 및 정책 변화와 다시 대조하는 반복적인 분석 과정이 중요하다는 점을 보여준다. 이는 최지영·이혜진(2023)이 여러 정책적 분기점(2002년 7.1 조치, 2012년 우리식 경제관리 방법, 상업은행법 제정·개정 시점 등)을 두루 확인한 끝에 김정은 집권 전후가 가장 뚜렷한 차이를 보이는 기준임을 확인했던 과정과 본질적으로 다르지 않다. 즉, 유의미한 정량적 결과에 도달하기까지 분석 대상 시기와 비교 단어를 반복적으로 조정하는 시행착오는 분석기법의 종류와 무관하게 텍스트마이닝 연구 전반의 공통된 과정이라 할 수 있다.

3. 방법의 복잡성보다 중요한 것

제시된 사례에서 활용된 기법은 단순빈도분석, TF, TF-IDF, 워드클라우드, 바이그램 네트워크 분석, 감성분석 등 비교적 접근성이 높은 분석방법들이다. 그럼에도 불구하고 각 연구는 북한의 정책 변화와 선전 서사, 동원 방향, 대남 담화의 변화 등에서 유의미한 정책적

함의를 도출하였다. 이는 텍스트마이닝 연구의 성과가 반드시 분석기법의 기술적 복잡성에 의존하는 것은 아니라는 점을 설명한다. 즉, 중요한 것은 어떤 현상을 설명하고자 하는지에 대한 명확한 연구질문을 설정하고, 그 질문에 적합한 분석 단위와 비교 기준을 구성하는 과정이 분석기법의 선택보다 선행되는 것이라 할 수 있다. 특히 시기 구분, 비교 대상, 핵심어 선정 등을 고정된 기준으로 적용하기보다 실제 분석 결과를 확인하면서 이를 조정하고 재분석하는 과정이 중요하다. 결국 텍스트마이닝 기법은 그 자체로 결론을 산출하는 도구라기보다, 연구자가 설정한 질문을 계량적으로 확인하고 기존의 정성적 분석만으로는 포착하기 어려웠던 패턴과 변화 지점을 발견하는 수단으로 이해할 필요가 있다.

이러한 점은 북한 문헌 연구에서 더욱 중요하다. 북한 문헌은 우리가 접하는 일상적인 텍스트와 달리 정치적 목적과 선전·동원의 기능을 전제로 생산되며, 동일한 표현의 반복, 관용적 수사, 정책적 용어의 의도적 사용 등 텍스트 자체의 특수성이 강하다. 따라서 특정 단어의 출현 빈도나 감성점수와 같은 계량적 결과만으로 그 의미를 곧바로 해석하기보다는, 해당 어휘가 어떠한 정책적 맥락에서 등장했는지, 이전 시기와 비교해 무엇이 달라졌는지, 실제 정책과 행동의 변화가 어떻게 나타났는지를 함께 검토해야 한다. 이러한 맥락적 해석이 결합될 때 북한이 반복적으로 제시하는 선전 서사와 특정 시기에 새롭게 부각하는 정책적 강조점을 구분하고, 담론 변화가 실제 정책 변화와 어떠한 관계를 갖는지까지 분석할 수 있다. 따라서 북한 텍스트마이닝 연구에서는 적절한 연구질문의 설계, 분석 결과에 대한 반복적인 검증, 그리고 북한의 정책·담론에 대한 정성적 이해를 계량적 결과와 결합하는 역량이 연구의 설명력과 정책적 활용도를 결정하는 핵심 요소라고 할 수 있다.

IV. 향후 연구과제

첫째, 제시된 분석 사례 모두 노동신문과 조선중앙통신이라는 한정된 매체를 분석 대상으로 삼았다. 북한은 대내용 매체(노동신문, 조선중앙TV)와 대외용 매체(조선중앙통신, 조선신보 등)의 서술 전략을 달리하는 경향이 있는 만큼, 매체별 비교 분석을 통해 대내·대외 담론의 차이를 함께 살펴보는 보다 많은 연구들이 필요하다.

둘째, 텍스트마이닝으로 도출된 정량적 결과를 그 자체로 정책적 결론으로 해석하는 위험을 경계할 필요가 있다. 선전물이라는 자료의 특성상 표면적 언급 빈도가 실제 정책 우선순위와

반드시 일치하지 않으며, 앞서 설명한 연구 사례가 보여주었듯 오히려 TF-IDF와 같은 가중치 지표를 적용해야 실질적 강조점을 드러낼 수 있는 경우도 있기 때문이다. 첫 번째 사례가 보여주었듯 특정 용어의 소멸 자체가 유의미한 정책 신호일 수도 있고, 마지막 사례가 제시한 바와 같이 표면적 감성점수의 회복이 정책 실질의 변화를 의미하지 않을 수도 있다. 따라서 텍스트마이닝 결과를 해석하는 과정에서는 해당 시기 북한의 정책 환경에 대한 연구자의 정성적 분석과 판단이 반드시 함께 수행되어야 한다.

셋째, 형태소 분석의 정확도 문제를 보완할 필요가 있다. 우리와 다른 북한식 신조어, 축약어, 붙여쓰기는 우리가 쓰는 기존 형태소 분석기를 통한 분석의 정확도를 떨어뜨리는 요인으로, 형태소 분석을 사용한 네 연구 모두 수작업 검수를 병행할 수밖에 없었다. 마지막 사례에서 구축하였던 감정사전 역시 ‘위대’·‘경애’와 같은 지도자 찬양 어휘가 북한 자신의 지도자를 향한 것인지, 문맥에 따라 다른 대상을 향한 것인지 구분하기 위해 별도의 문맥 판별 절차를 두어야 했는데, 이는 감정사전 기반 접근이 형태소 분석 이상의 문맥적 정교화를 요구한다는 점을 보여준다. 첫 번째 사례처럼 형태소 분석 없이 핵심어를 직접 검색하는 방식은 이러한 오류를 피할 수 있지만, 1950~80년대 신문의 디지털화·OCR 품질이 고르지 않아 장기 시계열 자료일수록 결측·오독 가능성을 별도로 검토해야 한다는 과제를 남긴다. 북한 문헌에 특화된 형태소 사전과 분석 도구를 지속적으로 보완·공유하는 작업이 필요하다.

넷째, 향후에는 텍스트뿐 아니라 노동신문·조선중앙TV에 실린 이미지·영상 자료를 함께 분석하는 멀티모달 접근, 그리고 최근 발전하고 있는 대규모언어모델(LLM) 기반 담론 분석기법을 결합해 문맥적 함의까지 포착하는 방법론적 확장을 모색할 필요가 있다.

다섯째, 이번에 소개한 네 사례 모두, 북한 문헌을 기계가 읽을 수 있는 형태로 가공하는 전처리 단계에 상당한 시간과 인력이 소요되었다는 공통된 어려움을 안고 있다. 이는 개별 연구자 단위로는 반복적으로 재투자될 수밖에 없는 비효율적인 과정으로, 정제된 북한 공식 문헌 텍스트 데이터를 축적하고 연구자 간에 공유할 수 있는 공동 플랫폼이 마련된다면 관련 연구의 저변을 크게 넓힐 수 있을 것이다. 아울러 이번 사례들에서 확인된 ‘표면적 서사와 실질적 강조점의 간극’이라는 패턴이 현재 세계적으로 확장되고 있는 권위주의 체제의 선전물 일반에서도 공통적으로 나타나는지, 북한과 이들 사이에 어떠한 차이가 있는지 확인하기 위해, 비교 가능한 사회의 공식 매체를 추가하여 북한 담론과 이들 사이의 비교연구를 수행하는 것도 흥미로운 후속 과제가 될 수 있다.

정리하면, 텍스트마이닝은 북한 공식 문헌이라는 방대하지만 여전히 제한적인 자료를 경험적으로 다루는 데 유용한 도구라 할 수 있다. 그러나 앞선 논의에서 제시하였듯 북한

텍스트에 대한 정량적 분석은 그 자체로 완결된 해석을 제공하지는 않는다. 이번에 소개한 사례들에서 공통적으로 제시하듯, 텍스트마이닝을 통한 분석의 성과는 결국 연구자가 어떤 질문을 던지고, 어떤 시기와 어떤 단어를 비교 대상으로 설정하며, 도출된 정량적 결과를 어떤 정성적 맥락 속에서 해석하느냐에 달려 있기 때문이다.

참고문헌

- 니콜라스 잭슨 오쇼네시(N. J. O'Shaughnessy), 박순석 역, 『대중을 유혹하는 무기(정치와 프로파간다)』, 파주: 한울아카데미, 2009.
- 양문수, 「북한 문헌, 어떻게 읽을 것인가: 경제연구의 사례」, 『현대북한연구』, 제12권 제2호, 2009, pp.7~61.
- 임수진, 「‘적대적 두 국가론’ 이후 북한의 남한 관련 보도 변화 및 시사점」, 『INSS 전략보고』, 서울:국가안보전략연구원, 2025.
- 임수진, 「김정은 시기 북한의 국면별 대남담화 변화 및 전략적 시사점」, 『INSS 전략보고』, 서울:국가안보전략연구원, 2026.
- 임수진, 「김정은시대 북한의 대외 적대담론 연구: 『노동신문』 속 ‘미제’, ‘일제’, ‘남조선괴뢰’를 중심으로」, 『통일과 평화』, 제14집 제2호, 2022, pp.163~201.
- 임수진, 「북한 ‘대동강’ 기사를 통해 본 김정은정권 ‘인민’ 동원 담론의 변용: TF-IDF 분석을 중심으로」, 『도시연구: 역사·사회·문화』, 제37권, 2024, pp.75~99.
- 최지영·이혜진, 「텍스트 마이닝을 이용한 북한경제 연구」, 『KDI 북한경제리뷰』, 제25권 제6호, 한국개발연구원, 2023, pp.52~61.