

빅데이터 기반 실시간 소비예측지표 개발 - 2026. 9. -

이 연구는 국회예산정책처의 연구용역사업으로 수행된 것으로서,
보고서의 내용은 연구용역사업을 수행한 연구자의 개인 의견이며,
국회예산정책처의 공식 견해가 아님을 알려드립니다.

연구진

연구책임자
공동연구원

이나경
이승희
유승주

서울여자대학교
한국개발연구원
계명대학교

목 차

제1장 서론	1
I. 연구 배경 및 목적	1
II. 연구 내용 및 범위	3
III. 연구방법 및 분석체계	5
제2장 선행연구 검토	7
I. 뉴스 기반 소비예측 관련 선행연구	8
II. 선행연구의 종합 및 본 연구에의 시사점	24
제3장 뉴스 빅데이터를 활용한 소비동향 기초분석	27
I. 분석 목적 및 방법	28
II. 소비 관련 뉴스의 규모와 변화	43
III. 소비 관련 뉴스의 주요 내용과 특징	49
IV. 기초 뉴스분석의 시사점 및 한계	57
제4장 실시간 소비예측지표의 개발 및 검증	61
I. 사전기반 소비 방향성 분류	62
II. 딥러닝 기반 문맥분류 모형 활용	76
III. 주간 뉴스지수 기반 소비지표 예측	98
제5장 결론	111
I. 주요 연구결과	112
II. 연구의 시사점	116
III. 연구의 한계 및 향후 과제	119
참고문헌	121

제1장 서론

I. 연구 배경 및 목적

- 민간소비는 경제활동의 주요 구성요소로서 경기의 흐름과 가계의 경제적 여건을 파악하는 데 중요한 정보를 제공함
 - 민간소비 동향은 소매판매액지수 등 공식 통계를 통해 파악할 수 있으나, 이러한 통계는 일정한 작성·공표 주기를 가지므로 실제 경제상황의 변화와 통계가 확인되는 시점 사이에 시차가 존재함
 - 특히 소비는 경기상황뿐 아니라 물가, 소비심리, 정책 변화, 금융여건 및 각종 사회·경제적 사건의 영향을 받을 수 있어, 공식 통계가 공표되기 이전에 관련 변화를 보다 신속하게 파악할 수 있는 보조적 정보의 활용 필요성이 있음
- 디지털 환경에서 지속적으로 생산·축적되는 뉴스자료는 경제상황과 경제주체의 심리 및 기대 변화를 비교적 신속하게 반영할 수 있다는 점에서 고빈도 경제정보의 주요 자료원으로 활용되고 있음
 - 기존 연구에서는 경제·정책·불확실성 관련 키워드의 출현빈도, 뉴스기사의 긍정·부정 감성, 단어 간 의미적 관계, 토픽구조 및 딥러닝 기반 문맥분석 등 다양한 방법을 활용하여 뉴스자료로부터 경제적 정보를 추출하고 이를 경기상황의 측정과 예측에 활용해 왔음
 - 이러한 연구는 뉴스자료가 공식 통계를 대체하기보다는 기존 통계가 제공하지 못하는 고빈도 정보를 보완하고, 경제여건과 경제주체의 인식 변화를 보다 신속하게 파악하는 보완자료로 활용될 수 있음을 보여줌
- 다만 뉴스자료를 소비동향 분석과 예측에 활용하기 위해서는 단순한 기사량이나 특정 단어의 출현 여부만으로 소비 관련 정보를 판단하기보다 뉴스에 포함된 내용과 문맥을 구분할 필요가 있음
 - 소비 관련 뉴스에는 실제 상품·서비스의 판매와 수요뿐 아니라 물가, 소비자

원정책, 금융소비자 보호, 산업활동, 기업실적 등 다양한 경제적 내용이 함께 포함될 수 있음

- 또한 동일한 ‘소비’ 관련 표현이라도 기사에서 사용되는 맥락에 따라 실제 민간소비의 확대·위축과 직접 연결되지 않을 수 있으며, 내구재·준내구재·비내구재 등 소비 품목군별로 소비행태와 뉴스의 보도 특성에도 차이가 존재할 수 있음
- 따라서 대규모 뉴스자료를 소비예측에 활용하기 위해서는 먼저 소비 관련 뉴스의 규모와 내용구조를 파악하고, 이후 기사에 포함된 소비 관련 문맥과 방향성을 보다 정교하게 식별하는 단계적 접근이 필요함

□ 이에 본 연구는 국내 경제뉴스 빅데이터를 활용하여 소비 관련 정보를 고빈도로 추출하고, 이를 바탕으로 민간소비의 단기적 흐름을 파악하는 데 활용할 수 있는 뉴스 기반 소비예측지표를 개발·검증하는 데 목적이 있음

- 먼저 한국언론진흥재단 BIGKinds 경제뉴스를 활용하여 소비 관련 뉴스의 규모와 시기별 변화, 주요 특성추출어 및 의미기반 토픽구조를 분석함으로써 소비 관련 뉴스의 기본적인 특성과 내용구조를 파악함
- 다음으로 소비 품목군별 관련 기사와 소비 방향성을 사전기반 방식으로 식별하고, 이를 바탕으로 딥러닝 기반 문맥분류모형을 적용하여 뉴스에 포함된 소비 관련 정보를 보다 정교하게 분류함
- 마지막으로 품목군별 뉴스정보를 주간 단위의 고빈도 지표로 구성하고 공식 소비지표와의 관계 및 예측성과를 검토함으로써, 뉴스 빅데이터를 활용한 단기 소비예측 보완지표의 활용 가능성을 검토하고자 함

II. 연구 내용 및 범위

- 본 연구는 국내 경제뉴스 빅데이터를 활용하여 소비 관련 정보의 특성을 파악하고, 이를 바탕으로 뉴스 기반 소비예측지표를 구축·검증하는 것을 주요 연구내용으로 설정함
 - 연구는 크게 뉴스 빅데이터를 활용한 소비동향 기초분석과 실시간 소비예측 지표의 개발 및 검증의 두 단계로 구성함
 - 첫 번째 단계에서는 전체 경제뉴스 가운데 소비 관련 뉴스의 규모와 시기별 변화를 파악하고, 주요 특성추출어와 의미기반 토픽분석을 통해 소비 관련 뉴스의 내용구조와 주요 의제의 변화를 탐색함
 - 두 번째 단계에서는 내구재·준내구재·비내구재 등 소비 품목군별 관련 기사와 소비 방향성을 식별하고, 딥러닝 기반 문맥분류모형을 활용하여 뉴스기사에 포함된 소비 관련 정보를 보다 정교하게 분류한 뒤 이를 주간 단위의 뉴스지표로 구축함
 - 구축된 뉴스지표는 공식 소비지표와의 관계를 검토하고 혼합주기 및 별점화 예측모형 등을 활용하여 소비지표에 대한 단기 예측 가능성을 검토함
- 분석자료는 한국언론진흥재단 BIGKinds에서 수집한 국내 경제뉴스를 중심으로 구성함
 - 뉴스자료는 BIGKinds 통합분류상 ‘경제’ 분야와 그 하위분류에서 수집된 기사자료를 활용하며, 기사 제목, 본문, 발행일자, 언론사, 뉴스 식별자, 키워드, 특성추출어 등 분석 목적에 필요한 기사 단위 정보를 활용함
 - 기초 뉴스분석에서는 전체 경제 기사를 기준집단으로 활용하고, 이 가운데 기사 제목에 ‘소비’가 명시적으로 포함된 기사를 소비 관련 기사로 선별하여 기사 규모와 내용구조를 분석함
 - 소비예측지표 개발에서는 뉴스기사의 제목과 본문에 포함된 소비 품목군 및

방향성 관련 정보를 활용하여 사전기반 분류와 딥러닝 기반 문맥분류를 수행함

□ 분석기간은 2022년 1월 1일부터 2026년 4월 30일까지로 설정함

- 기초 뉴스분석과 소비예측지표 개발·검증에 동일한 분석기간을 적용함으로써 뉴스자료의 특성 분석과 이후 지표 구축·예측결과를 동일한 시기적 범위에서 연계하여 해석할 수 있도록 함
- 기초 뉴스분석에 활용된 최종 전체 경제기사는 5,838,352건이며, 이 가운데 기사 제목에 '소비'가 포함된 소비 관련 기사는 76,301건으로 전체 경제기사의 약 1.31%에 해당함
- 2026년 자료는 1월부터 4월까지의 부분기간이므로 연도별 기사 규모나 토픽 구성을 비교할 때에는 완전한 연도와 동일하게 해석하지 않음

□ 본 연구의 범위는 뉴스자료를 활용한 소비 관련 정보의 식별과 소비예측 보조지표의 개발 가능성을 검토하는 데 한정함

- 뉴스기사의 수, 비중, 특성추출어 및 토픽구조는 언론보도의 상대적 규모와 내용구성을 나타내는 것으로 실제 민간소비의 수준이나 증감을 직접 측정하는 지표로 해석하지 않음
- 또한 뉴스 기반 지표는 공식 소비통계를 대체하기 위한 것이 아니라 공식 통계와 함께 활용할 수 있는 고빈도 보완지표로서의 가능성을 검토하는 데 초점을 둠
- 이에 따라 본 연구에서는 뉴스정보가 실제 소비지표와 어떠한 관계를 보이는지와 단기 예측에 활용될 수 있는지를 중심으로 평가하며, 뉴스기사의 변화 자체를 실제 소비증가나 특정 정책의 인과효과로 해석하지 않음

Ⅲ. 연구방법 및 분석체계

- 본 연구는 뉴스 빅데이터의 기초적 특성을 파악한 뒤, 소비 관련 문맥을 식별하여 뉴스 기반 소비지표를 구축하고 이를 공식 소비지표와 비교·검증하는 단계적 분석체계로 구성함
 - 먼저 뉴스 기반 경제지표와 관련된 국내외 선행연구를 검토하여 뉴스자료를 활용한 경제정보 추출방법의 발전과 주요 특징을 살펴보고, 본 연구의 분석 방향을 설정함
 - 선행연구에서는 키워드 출현빈도에 기초한 경제정책 불확실성지수(EPU), 뉴스 감성분석, 의미기반 분석, 토픽모형 및 Transformer 기반 문맥분석 등 뉴스자료를 경제상황의 측정과 예측에 활용한 주요 접근을 검토함

- 첫 번째 분석단계에서는 BIGKinds 경제뉴스를 활용하여 소비 관련 뉴스의 규모와 내용구조를 기초적으로 분석함
 - 전체 경제기사와 소비 관련 기사의 연도별·월별 규모 및 비중을 산출하여 소비 관련 보도가 전체 경제뉴스에서 차지하는 상대적 규모와 시기별 변화를 파악함
 - BIGKinds가 제공하는 특성추출어를 기사 단위로 재집계하여 소비 관련 뉴스에서 반복적으로 나타나는 주요 표현을 확인하고, BERTopic을 활용하여 기사 제목의 의미적 유사성에 기초한 주요 토픽과 시기별 의제 변화를 탐색함
 - 이를 통해 단순한 기사량이나 개별 단어의 출현빈도만으로는 구분하기 어려운 소비 관련 뉴스의 이질적인 내용구조를 확인하고, 이후 문맥기반 소비정보 식별을 위한 기초정보를 구축함

□ 두 번째 분석단계에서는 소비 품목군별 관련 기사와 소비 방향성을 식별하고 이를 뉴스 기반 소비지표로 구축함

- 내구재·준내구재·비내구재 등 소비 품목군별 관련 표현과 소비활동·방향성 표현을 활용한 사전기반 분류를 통해 소비 관련 기사 후보군과 소비 상승·하락 등의 방향성을 일차적으로 식별함
- 사전기반 분류결과를 기초로 딥러닝 기반 문맥분류모형을 적용하여 단순한 키워드 출현 여부를 넘어 기사 내 소비 관련 표현의 의미와 문맥을 반영한 분류를 수행함
- 분류결과를 주간 단위로 집계하여 품목군별 뉴스 기반 소비지표를 구축하고, 공식 소매판매 지표와의 관계 및 선행 가능성을 검토함

□ 마지막으로 구축된 고빈도 뉴스지표의 소비예측 활용 가능성을 검토함

- 주간 뉴스지수와 월별 공식 소비지표 간 자료주기의 차이를 고려하여 Bridge, MIDAS, U-MIDAS 등 혼합주기 모형과 별점화 회귀를 결합한 예측모형을 적용함
- 예측에는 계절조정 소매판매액지수와 뉴스지수의 차분값을 활용하였으며, 전 통적 시계열 모형을 기준모형으로 설정하고 확장창 표본외 예측을 통해 모형별 예측성과를 비교함
- 특히 다지수·다시차 뉴스정보를 활용하기 위해 Adaptive Elastic-Net을 적용한 AEN U-MIDAS 모형을 검토하고, 모형 간 예측력 차이는 Diebold-Mariano-West 검정을 통해 추가적으로 살펴봄
- 이를 통해 뉴스지표가 기존 소비통계를 보완하여 당월 및 단기 소비예측에 활용될 수 있는지를 검토함

제2장 선행연구 검토

I. 뉴스 기반 소비예측 관련 선행연구

1. 경제정책 불확실성지수(EPU)¹⁾

가. 연구 개요

- Baker, Bloom, Davis(2016)는 뉴스기사에 나타난 경제정책 관련 불확실성을 정량적으로 측정하기 위해 경제정책 불확실성지수(Economic Policy Uncertainty Index: EPU)를 개발함
 - 기존의 경제 불확실성 연구는 금융시장 변동성이나 전문가 설문 등에 주로 의존하였으나, EPU는 언론보도에 나타난 정책 불확실성을 직접 측정함으로써 경제주체가 인식하는 불확실성을 보다 시의성 있게 포착하고자 하였음
 - 특히 EPU는 뉴스기사에 포함된 경제(Economy), 정책(Policy), 불확실성(Uncertainty) 관련 표현의 출현 빈도를 활용하여 지수를 구축한 대표적인 뉴스 기반 경제지표로 평가됨
 - 이후 미국뿐 아니라 영국, 독일, 프랑스, 일본, 한국 등 다양한 국가로 확장되면서 뉴스기반 경제지표 구축 연구의 대표적 사례로 활용되고 있음

나. 자료 및 분석방법

- EPU는 미국 주요 10개 신문을 활용하여 구축되었으며, 경제·정책·불확실성 관련 키워드가 동시에 등장하는 기사 빈도를 기반으로 산출됨
 - 분석대상 신문은 USA Today, Miami Herald, Chicago Tribune, Washington Post, Los Angeles Times, Boston Globe, San Francisco Chronicle, Dallas Morning News, New York Times, Wall Street

1) 본 절은 Baker, S. R., Bloom, N., and Davis, S. J.(2016), "Measuring Economic Policy Uncertainty," The Quarterly Journal of Economics, 131(4), pp.1593-1636을 바탕으로 작성하였음.

Journal로 구성됨

- 연구진은 1985년 이후 각 신문의 디지털 아카이브를 대상으로 경제(Economy), 정책(Policy), 불확실성(Uncertainty) 관련 키워드가 동시에 포함된 기사를 추출하여 분석에 활용함²⁾
 - 구체적으로 EPU는 불확실성 범주의 “uncertainty”, “uncertain”, 경제 범주의 “economic”, “economy”, 정책 범주의 “Congress”, “deficit”, “Federal Reserve”, “legislation”, “regulation”, “White House” 등의 표현이 모두 포함된 기사를 경제정책 불확실성 관련 기사로 분류함
 - 이때 “uncertainties”, “regulatory”, “the Fed” 등 파생형 및 유사 표현도 함께 포함함
 - 특히 정책 키워드는 연구자가 임의로 설정한 것이 아니라, 12,009개 기사에 대한 human audit을 통해 regulation, budget, spending, policy, deficit, tax, federal reserve, war, white house, house of representatives, government, congress, senate, president, legislation 등 15개 후보어를 도출한 뒤, 약 32,000개 조합을 비교하여 false positive와 false negative를 최소화하는 기준으로 선정함³⁾
 - 그 결과 최종 기준 정책 키워드로 regulation, deficit, federal reserve, white house, congress, legislation이 채택됨

□ 기사 추출 이후에는 신문별 기사 수 차이를 보정하기 위한 표준화 과정을 수행함

- 특정 월의 관련 기사 수를 해당 월 전체 기사 수로 나누어 비율화
- 신문별 시계열을 표준화한 후 10개 신문 평균을 산출

2) Baker, Bloom, and Davis(2016: 1598-1599), “We search the digital archives of each paper from January 1985 to obtain a monthly count of articles that contain the following triple: ‘uncertainty’ or ‘uncertain’; ‘economic’ or ‘economy’; and one of the following policy terms: ‘congress’, ‘deficit’, ‘Federal Reserve’, ‘legislation’, ‘regulation’ or ‘white house’.”

3) Baker, Bloom, and Davis(2016: 1607-1610), “The auditors reviewed 12,009 articles from 1900 to 2012...”; “We chose the P term set that minimizes the gross error rate.”

- 최종적으로 1985~2009년 평균이 100이 되도록 정규화하여 EPU 지수를 구축함

□ 키워드 체계의 타당성을 검증하기 위해 대규모 human audit을 수행함

- 약 12,000건의 기사에 대해 연구보조원이 직접 기사 내용을 검토
- 정책 불확실성과 관련된 기사 여부를 판정한 뒤 자동 분류 결과와 비교
- 이를 통해 정책 키워드 체계를 수정·보완하였으며 높은 수준의 일치도를 확인함

다. 주요 결과

□ EPU는 주요 정책 및 경제 충격 시기에 급격히 상승하는 것으로 나타남

- 걸프전(Gulf War), 9·11 테러, 글로벌 금융위기, 미국 부채한도 협상, 재정절벽(Fiscal Cliff) 논쟁 등 주요 정책 이벤트 시기에 지수가 큰 폭으로 상승함
- 정책 불확실성의 증가는 투자 감소, 고용 감소, 생산 감소 등 거시경제 변수와 유의한 관련성을 보이는 것으로 나타남
- 이후 EPU는 미국뿐 아니라 영국, 독일, 프랑스, 일본, 한국 등 다수 국가로 확장되어 활용되고 있음

라. 본 연구에의 시사점

□ EPU는 뉴스자료를 활용하여 경제 관련 심리 및 불확실성을 측정할 수 있음을 보여준 대표적인 사례임

- 비교적 단순한 키워드 기반 접근만으로도 장기간 시계열 구축과 국가 간 비교가 가능함을 제시함
- 또한 대규모 뉴스자료에서 특정 주제와 관련된 기사만을 선별하여 지수를

구축할 수 있음을 보여줌

- 다만 EPU는 키워드 출현 여부를 중심으로 분석하기 때문에 기사의 실제 논조나 문맥을 충분히 반영하지 못한다는 한계를 가짐
- 이러한 한계는 이후 감성사전 기반 분석, 의미기반 분석, 기계학습 기반 분석으로 발전하는 계기가 되었으며, 본 연구 역시 단순 키워드 빈도를 넘어 소비 관련 문맥과 방향성을 식별하는 방향으로 확장하고자 함

2. 뉴스 감성지수(DNSI)⁴⁾

가. 연구 개요

- Shapiro, Sudhof, Wilson(2020)은 경제뉴스의 감성을 측정하는 방법론을 제시하였으며, 이를 바탕으로 FRBSF(Federal Reserve Bank of San Francisco)의 Daily News Sentiment Index(DNSI)가 구축·공개됨
- DNSI는 뉴스기사에 포함된 긍정적·부정적 표현을 활용하여 경제주체의 심리 변화를 측정하는 뉴스 기반 경제심리지표임
- 기존 EPU가 특정 키워드의 출현 빈도를 활용하여 정책 불확실성을 측정하였다면, DNSI는 기사 자체의 논조(sentiment)를 직접 측정한다는 점에서 차별성을 가짐
- 연구진은 뉴스기사에 포함된 감성 정보가 현재의 경제상황을 반영할 뿐 아니라 향후 경기변화를 예측하는 데에도 활용될 수 있다는 관점에서 일별 및 월별 뉴스 감성지수를 구축하였음
- 이후 Buckman et al.(2020)은 동일한 방법론을 활용하여 COVID-19 시기의 뉴스심리 변화를 분석하고 뉴스기반 심리지표의 활용 가능성을 검증함

4) 본 절은 FRBSF, “Daily News Sentiment Index,” <https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/> 및 Buckman, S. R., Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020), “News Sentiment in the Time of COVID-19,” FRBSF Economic Letter 2020-08, April 6, 2020의 내용을 바탕으로 재구성·정리하였음.

나. 자료 및 분석방법

- DNSI는 미국 주요 신문에 게재된 경제 관련 기사를 활용하여 구축됨
 - 초기 연구 및 COVID-19 적용 연구에서는 LexisNexis 기반 16개 미국 주요 신문을 활용하였으며, 이후 FRBSF가 공개하는 DNSI는 Factiva 기반 24개 미국 주요 신문 자료로 확대·업데이트됨⁵⁾
 - 분석대상은 경제(Economics) 분야이면서 미국(United States)을 대상으로 하는 기사로 한정하였으며, 기사 길이는 200단어 이상인 경우만 포함함⁶⁾
- 기사별 감성점수 산출을 위해 감성사전(lexicon)을 활용한 분석을 수행함
 - 공개적으로 활용되는 감성사전과 연구진이 구축한 뉴스 특화 감성사전을 결합하여 기사별 sentiment score를 산출함
 - 단순 키워드 출현 빈도를 집계하는 방식이 아니라 기사 전체의 긍정성(positivity)을 반영한 sentiment score를 산출함
 - 산출된 기사별 감성점수는 일별 및 월별 수준으로 집계하여 시계열 지표로 변환함
- 지수의 안정성을 확보하기 위해 다양한 보정 과정을 적용함
 - 개별 기사의 감성점수를 집계하는 과정에서 신문사별 보도 성향과 기사유형에 따른 체계적 차이를 통제하고, 특정 시점의 순수한 뉴스심리 효과(day fixed effect)를 추정함⁷⁾

5) FRBSF, Daily News Sentiment Index, updated June 8, 2026. “Shapiro, Sudhof, and Wilson (2020, hereafter SSW), construct sentiment scores for economics-related news articles from 24 major U.S. newspapers compiled by the news aggregator service Factiva.” <https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/>

6) FRBSF, Daily News Sentiment Index, updated June 8, 2026. “SSW use articles with at least 200 words where Factiva identified the article’s topic as ‘economics’ and the country subject as ‘United States.’” <https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/>

7) FRBSF, Daily News Sentiment Index (updated June 8, 2026), “The units are the day fixed effects from the following regression, indexed by article.

- 또한 기사 발행 이후 시간이 경과함에 따라 가중치가 감소하는 지수가중평균(trailing weighted average)을 적용하여 단기 변동성을 완화함
- 연구진은 무작위로 추출한 800개 기사에 대해 연구보조원이 직접 1~5점 척도로 감성을 평가한 결과와 모형 산출 결과를 비교하여 지수의 타당성을 검증함

다. 주요 결과 및 시사점

□ DNSI는 소비자심리지수 및 주요 경기지표와 높은 관련성을 보이는 것으로 나타남

- DNSI는 미시간대학교 소비자심리지수 및 Conference Board 소비자신뢰지수와 역사적으로 높은 상관관계를 보이며, 경기침체기에는 하락하고 경기확장기에는 상승하는 경기순응적 움직임을 보이는 것으로 나타남⁸⁾
- 또한 소비지출, 고용, 산업생산 등 주요 거시경제 변수와도 유의한 관련성을 나타냄
- 일부 시기에는 설문조사 기반 심리지표보다 경기 변화를 선행적으로 포착하는 것으로 분석됨

□ Buckman et al.(2020)은 COVID-19 시기 뉴스 감성의 급격한 악화를 분석함

- 분석 결과 COVID-19 확산 초기 뉴스 감성지수는 소비자심리지수보다 약 2개월 빠르게 심리 악화를 포착한 것으로 나타남
- 또한 경제 관련 기사 중 “coronavirus” 또는 “COVID-19”를 포함한 기사 비율을 별도로 산출한 결과, 관련 보도는 1월 20일경부터 나타나기 시작해 3월 말에는 경제 관련 뉴스의 약 95%에 도달하였으며, 뉴스 감성지수의 하락

<https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/>

8) Buckman et al.(2020), “The news sentiment index correlates strongly with the survey-based consumer sentiment measure, and both are strongly procyclical, dipping during recessions and rising during economic expansions.”

은 이러한 COVID-19 관련 보도 증가와 함께 나타난 것으로 분석됨⁹⁾

□ DNSI는 뉴스기사의 논조 자체를 활용하여 경제심리를 측정한 대표적 사례로 평가됨

- DNSI는 1980년부터 현재까지 일별 빈도로 구축되어, 기존 설문 기반 심리지표가 갖는 낮은 빈도와 짧은 시계열의 한계를 보완할 수 있는 고빈도 경제심리지표로 활용될 수 있음.
- 특히 단순 키워드 출현 빈도보다 기사의 감성적 방향성을 직접 반영할 수 있다는 장점을 가짐
- 뉴스 데이터만을 활용하여 고빈도 경제심리지표를 구축할 수 있음을 보여주었다는 점에서 의의가 있음
- 다만 감성사전에 기반한 접근은 사전 구성 방식과 문맥 해석의 적합성에 영향을 받을 수 있다는 한계를 가지며, 이후 의미기반 분석 및 기계학습 기반 감성분석 연구와 함께 확장·활용되고 있음
- 본 연구 역시 뉴스기사에 나타난 소비 관련 방향성을 식별하고자 한다는 점에서 DNSI의 접근과 유사성을 가지며, 소비활동과 방향성 표현을 결합하여 소비 신호를 추출한다는 점에서 확장된 적용 사례로 볼 수 있음.

3. 의미기반 뉴스 감성지수(IMF)¹⁰⁾

가. 연구 개요

□ Huang, Simpson, Ulybina, Roitman(2019)은 뉴스기사를 활용하여 금융위기 조

9) Buckman et al.(2020), "Articles mentioning the coronavirus and COVID-19 began around January 20 and then rapidly increased. By late March, the percentage of economics-related news articles mentioning the virus reached an astounding 95%."

10) 본 절은 Huang, C., Simpson, S., Ulybina, D., and Roitman, A.(2019), "News-Based Sentiment Indicators," IMF Working Paper WP/19/273, International Monetary Fund, December 2019의 내용을 바탕으로 재구성·정리하였음.

기경보지표(Early Warning Indicator)를 구축하기 위해 뉴스 기반 감성지수 (News-Based Sentiment Indicators)를 개발함

- 해당 연구는 국제통화기금(IMF) Working Paper로 발표되었으며, 뉴스에 나타나는 위험 인식과 심리 변화를 정량화하여 금융위기 발생 이전의 위험 신호를 포착하는 데 목적을 둠
- 기존의 키워드 기반 접근이나 감성사전 기반 접근이 사전에 정의된 단어 목록에 의존하는 것과 달리, 단어 간 의미적 관계를 활용하는 의미기반 (semantic) 분석 방법을 적용하였다는 점에서 차별성을 가짐
- 연구진은 뉴스기사에 포함된 감성 표현이 금융위기 발생 이전에 변화할 수 있다는 점에 주목하여 국가별 감성지수를 구축하고 조기경보지표로서의 활용 가능성을 검토함

나. 자료 및 분석방법

- 연구진은 Financial Times(FT)의 기사 약 311만 건을 활용하여 감성지수를 구축함
 - 분석대상 기간은 1980년부터 2019년까지이며, 총 20개국을 대상으로 국가별 월간 감성지수를 산출함
 - FT는 글로벌 경제·금융 관련 보도가 집중적으로 이루어지는 대표적인 경제 매체로서 금융시장 및 거시경제 관련 뉴스의 주요 정보원으로 활용됨
 - 국가별로는 기사 제목, 초록, 기사 서두에 특정 국가명이 포함된 경우 해당 국가 관련 기사로 분류함
- 감성 측정을 위해 word2vec 기반의 의미기반 분석기법을 적용함
 - 연구진은 Word Vector Representation 기법을 활용하여 단어 간 의미적 유사성을 벡터 공간상에서 계산함

- 구체적으로는 word2vec 기반 벡터 공간 모형을 구축하고, 의미적으로 유사한 단어군을 자동으로 추출하는 semantic clustering 기법을 활용함
 - 분석에는 Python 기반의 gensim 패키지를 사용하였으며, vector dimension은 200, context window size는 5로 설정함
- 감성 개념별로 seed word를 설정하고 이를 의미적으로 확장하는 방식을 활용함
 - 연구진은 fear, risk, hedging, opinion, crisis 등 다섯 개 핵심 감성 범주를 설정함
 - 예를 들어 fear 감성은 fear, concern, afraid, anxious 등의 seed word를 기준으로 의미적으로 유사한 단어들을 확장하여 하나의 semantic cluster를 구성함¹¹⁾
 - 이외에도 positive sentiment와 negative sentiment를 별도로 구축하고, 다섯 개 감성 범주와 긍정·부정 감성지수를 조합하여 “Negative Sentiment+”, “All Sentiment” 등 합성지수를 구성함
- 감성지수는 특정 감성군에 속하는 단어의 상대적 빈도를 활용하여 산출함
 - 특정 국가의 특정 월에 나타난 semantic cluster 단어 수를 해당 월 전체 기사 단어 수로 나누어 표준화함
 - 이를 바탕으로 국가별·월별 감성지수를 구축하고 시계열 분석을 수행함

다. 주요 결과 및 시사점

- 분석 결과 뉴스 기반 감성지수는 금융위기 발생 이전에 상승하거나 급등하는 경향을 보이는 것으로 나타남

11) Huang et al.(2019: 10-11), “To construct a ‘fear’ sentiment index we consider terminology and language associated with the following ‘seed’ terms: ‘fear’, ‘concern’, ‘afraid’ and ‘anxious’.”

- 특히 fear, risk, crisis 관련 감성지수는 금융위기 발생 이전부터 점진적으로 증가하는 경향을 보였으며, 위기 발생 직전에는 급격한 상승이 관측됨
- 연구진은 이러한 결과를 바탕으로 뉴스 감성지수가 금융위기 조기경보지표로 활용될 가능성이 있다고 평가함
- 또한 다양한 감성 범주를 결합한 종합지수는 해석 가능성이 다소 낮아질 수 있으나, 금융위기 이전의 이상신호를 보다 폭넓게 포착할 가능성이 있다고 제시됨

□ 본 연구는 단어의 단순 출현 빈도가 아니라 의미적 관계를 활용하여 감성을 측정하였다는 점에서 기존 연구와 차별성을 가짐

- 동일한 감성 개념이라 하더라도 다양한 표현 방식이 존재할 수 있다는 점을 고려하여 의미적으로 유사한 단어들을 함께 활용함으로써 감성 측정의 범위를 확장함
- 이는 사전에 정의된 제한적 키워드에 의존하는 방식보다 보다 풍부한 의미 정보를 반영할 수 있다는 장점을 가짐
- 다만 의미기반 접근은 분석과정이 복잡하고, 특히 여러 감성 범주를 결합할 수록 이상신호 포착 가능성은 높아질 수 있으나 개별 감성 요인의 해석 가능성은 낮아질 수 있다는 한계가 있음
- 본 연구의 경우 소비 관련 기사에서 나타나는 방향성 표현을 식별하고자 한다는 점에서 IMF 연구와 문제의식을 공유하지만, 금융위기 감성 측정보다는 소비활동의 증가·감소·보합 신호를 구분하는 데 초점을 둔다는 점에서 차이가 있음
- 또한 본 연구는 의미기반 접근의 장점을 참고하되, 초기 단계에서는 규칙기반 분류를 활용하여 학습자료를 구축하고 향후 딥러닝 기반 문맥분류모형으로 확장하고자 함.

4. 경제 뉴스 구조(Bybee et al., 2024)¹²⁾

가. 연구 개요

- Bybee et al. (2024)은 경제 상태를 비즈니스 뉴스의 텍스트 분석으로 측정하는 접근법을 제안
 - 새로운 접근법을 통해 기존 텍스트 분석 기반 실증 거시경제학 연구의 한계를 극복
 - Manela and Moreira(2017)는 WSJ 1면 기사의 헤드라인·초록만 활용하였고, Baker et al.(2016)의 EPU는 사전 정의된 소수 키워드의 출현 빈도 계산에 국한된 반면, Bybee et al.(2024)는 신문 전문(full text)을 활용
 - 뉴스가 무엇에 주목(attention allocation)하는지 그 자체가 경제 상태의 충분통계량에 가깝기 때문에 뉴스 데이터에서 획득 가능한 주제별 배분이 측정 대상
 - 이러한 뉴스 관심도를 바탕으로 광범위한 경제활동을 추적하고 주식시장 수익률을 예측

나. 자료 및 분석방법

- 1984~2017년 Wall Street Journal 기사 전문 약 80만 건을 분석 대상으로 설정
- 토픽 모형인 Latent Dirichlet Allocation(LDA)를 기반으로 총 180개 토픽을 분석하고 토픽별 단어 분포와 토픽에 대한 뉴스 관심도를 정량화
 - 토픽별 단어 분포는 토픽의 해석가능한 의미를 부여 (예: “recession”, “record high”, “problems”, “options/VIX” 등의 라벨¹³⁾)

12) 본 절은 Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024), “Business News and Business Cycles,” Journal of Finance의 내용을 바탕으로 재구성·정리하였음.

13) 개별 토픽들은 거시·경기 관련 토픽과センチ먼트에 가까운 토픽, 시대 국지적(time-localized) 토픽, 금융시장 토픽 등으로 분류 가능

- 기사별로 각 토픽에 할애된 텍스트 비중을 추정하고, 이를 통해 시점별로 각 토픽에 대한 뉴스 관심도(news attention)를 측정

□ 토픽별 관심도를 거시 및 금융 변수와 연결

- 180개 토픽에 대한 관심도를 산업생산, 고용, 주가 변동성 등 광범위한 경제활동 지표에 회귀하였고, 특정 토픽(특히 recession 토픽)이 경기순환을 밀접하게 추적함을 확인
- 180개 토픽 관심도를 LASSO 등 벌점화 회귀 계열 분석 방법을 바탕으로 거시 변수(산업생산 증가율, 고용 증가율 등) 예측에 투입 및 예측력 평가

다. 주요 결과 및 시사점

□ 분석 결과 뉴스 데이터가 경제 상태를 추적하고 거시 예측 및 금융시장 예측에 유용하게 활용될 수 있음을 입증

- 뉴스 관심도가 광범위한 경제활동을 밀접하게 추적하였고 주식시장에서의 시장 수익률 상당 부분(약 25%)를 설명
- 토픽 모형이 뉴스 텍스트로부터 신뢰할 만한 마켓 타이밍 신호를 추출함을 보였고 기존 거시 지표와 결합했을 때 뉴스 데이터에서의 정보가 예측력 제고에 유의미한 역할을 수행함을 확인
- 아울러 Babii et al.(2021) 후속 연구를 통해 다양한 토픽에 대한 뉴스 관심도가 전통적 거시·금융 데이터를 보완하여 미국 GDP 성장률 나우캐스팅에 유용함을 확인

5. 국내 뉴스심리지수(NSI)¹⁴⁾

가. 연구 개요

- 한국은행은 경제주체의 심리 변화를 실시간에 가깝게 측정하기 위해 뉴스심리지수(News Sentiment Index: NSI)를 개발함
 - 기존 소비자동향지수(CSI), 기업경기실사지수(BSI) 등은 설문조사를 기반으로 작성되므로 조사비용과 시차가 발생하는 한계가 존재함
 - 이에 한국은행은 대규모 뉴스 기사를 활용하여 경제심리를 측정할 수 있는 뉴스심리지수를 구축하고, 경제상황을 보다 신속하게 파악할 수 있는 보조지표로 활용하고자 함
 - NSI는 국내 언론사에서 생산되는 경제 관련 뉴스를 활용하여 경제심리를 측정한 대표적인 국내 사례로 평가됨
 - 특히 키워드 빈도나 감성사전에 의존하던 기존 접근과 달리 Transformer 기반 딥러닝 모형을 활용하여 뉴스 문맥을 분석한다는 점에서 차별성을 가짐

나. 자료 및 분석방법

- NSI는 국내 주요 언론사에서 생산된 경제뉴스를 활용하여 구축됨
 - 분석대상은 2005년 이후 인터넷 포털사이트에서 경제분야로 분류된 뉴스 기사임
 - 수집대상 매체는 중앙일간지, 경제신문, 방송사, 통신사, 지역언론 등 약 50여 개 언론사로 구성됨
 - 2005년에는 평일 기준 일평균 약 2,000건 수준이었으나, 2021년에는 평일

14) 본 절은 서범석·이영환·조형배(2022), 「기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용」, 『국민계정 리뷰』 2022년 제1호, 한국은행, pp.68-90의 내용을 바탕으로 재구성·정리하였음.

기준 일평균 약 4,000건 수준으로 증가한 것으로 나타남

- 동일 기사 반복으로 인한 왜곡을 방지하기 위해 직전 30일 이내 동일 기사 원문이 존재하는 경우 중복 기사로 간주하여 제거함

□ 감성분석은 기사 단위가 아니라 문장 단위 기반의 딥러닝 모형을 활용하여 수행됨

- 일반적으로 뉴스기사는 긍정적 내용과 부정적 내용을 동시에 포함하는 경우가 많기 때문에 기사 전체를 하나의 논조로 분류하는 방식은 효과적이지 않음
 - 이에 한국은행 NSI는 뉴스 기사를 구성하는 개별 문장을 입력 단위로 설정하고, 각 문장의 논조를 긍정·부정·중립으로 분류함
- 연구진은 2005년부터 2021년까지의 경제기사에서 약 44만 개 문장을 추출하여 학습데이터를 구축함
- 16명의 통계조사원이 각 문장을 긍정, 부정, 중립으로 직접 분류하여 정답자료를 생성함
- 또한 문장이 국내 경제를 의미하는지, 해외 경제를 의미하는지, 또는 양자를 모두 포함하는지에 대해서도 추가 분류를 수행함
 - 모형 적용 전에는 뉴스문장을 형태소 단위로 분해하고, 이를 정수 변환하여 통계모형이 인식할 수 있는 수치형 벡터로 전환함
 - 문장별 수치형 벡터의 최대 길이는 80으로 설정하였으며, 이는 학습데이터 기준 99%의 문장이 포함되는 길이를 고려한 것임

□ 뉴스 문맥을 반영하기 위해 Transformer 기반 감성분류모형을 적용함

- Transformer 모형은 multiple-head Attention 구조를 활용하여 입력 벡터 전체의 맥락을 파악할 수 있다는 장점이 있으며, 한국은행 NSI는 이를 활용하여 뉴스문장의 감성을 긍정·부정·중립 확률로 분류함
- Random Forest, SVM, Attention 기반 모형 등 다양한 알고리즘을 비교한

결과 Transformer 모형이 가장 우수한 분류 성능을 보이는 것으로 나타남

- 2021년 기준 일별 수집 뉴스는 약 4,000개 기사, 이를 문장 단위로 환산하면 약 7만 문장 수준이므로 전체 문장에 모형을 적용할 경우 컴퓨팅 비용이 크다고 판단함
- 이에 일별 문장 모집단에서 1만 개 문장을 임의추출하여 일별 뉴스데이터베이스를 구성하고, 각 문장에 감성분류 모형과 국내외분류 모형을 적용함
- 최종 모형은 표본추출된 개별 문장을 긍정·부정·중립 확률로 분류하고, 국내외분류 모형을 통해 국내 경제 관련 문장을 식별함
 - 뉴스심리지수는 국내로 분류된 문장 중 긍정문장 수와 부정문장 수를 카운트한 뒤, 긍정문장과 부정문장의 상대적 비중을 이용하여 산출함
 - 일별 지수는 변동성을 고려하여 해당일 기준 직전 7일간 발간된 뉴스 기사를 기준으로 작성하고, 월별 지수는 해당월 중 발간된 뉴스 기사를 기준으로 작성함
 - 산출된 지수는 2005년부터 직전 연도까지의 평균과 분산을 활용하여 평균 100, 표준편차 10이 되도록 표준화함

다. 주요 결과 및 시사점

- NSI는 주요 경제심리지표 및 실물경제지표와 높은 관련성을 보이는 것으로 나타남
 - 뉴스심리지수는 소비자심리지수(CCSI), 소비자동향지수(CSI), 기업경기실사지수(BSI), 경제심리지수(ESI) 등과 전반적으로 유사한 움직임을 보임
 - 특히 글로벌 금융위기와 코로나19 위기 시기에 경제심리의 급격한 악화를 신속하게 포착하는 것으로 확인됨
 - 분석 결과 월별 NSI는 주요 경제심리지표에 대해 약 1~2개월 정도 선행하는 특성을 보이는 것으로 나타남

□ 또한 실물경제지표와도 유의한 관련성을 보임

- KOSPI 전년동기대비 증가율, 선행종합지수 순환변동치, 분기 GDP 증가율 등과도 일정 수준의 상관관계를 보이는 것으로 분석됨
- 이는 뉴스기사가 경제주체의 기대와 심리 변화를 비교적 빠르게 반영한다는 점을 시사함

□ NSI는 국내 뉴스자료를 활용한 고빈도 경제심리지수 구축의 대표적인 사례라는 점에서 의미가 있음

- 대규모 뉴스데이터와 딥러닝 기반 자연어처리 기법을 활용하여 경제심리를 실시간에 가깝게 측정할 수 있음을 보여주었음
- 또한 뉴스 데이터가 기존 설문조사 기반 지표를 보완할 수 있는 유용한 정보원임을 확인하였음
- 본 연구 역시 대규모 뉴스자료를 활용하여 소비 관련 신호를 추출한다는 점에서 NSI와 문제의식을 공유함
- 다만 NSI가 경제 전반의 심리를 측정하는 데 초점을 둔 반면, 본 연구는 내구재·준내구재·비내구재 등 소비 품목군별 소비 방향성을 식별하고 고빈도 소비지표를 구축한다는 점에서 차별성을 가짐.

II. 선행연구의 종합 및 본 연구에의 시사점

1. 뉴스 기반 경제지표 구축 방법론의 발전

- 선행연구들은 뉴스 데이터를 활용하여 경제상황과 경제주체의 심리를 측정할 수 있음을 공통적으로 보여주고 있음
 - EPU는 경제·정책·불확실성 관련 키워드의 출현 빈도를 활용하여 정책 불확실성을 측정하였으며, 비교적 단순한 구조를 통해 장기간 시계열 구축과 국가 간 비교가 가능함을 제시하였음
 - DNSI는 뉴스기사의 긍정·부정 논조를 직접 측정함으로써 뉴스 감성과 경제 심리 간의 관계를 분석하였으며, 뉴스 기반 심리지표의 속보성과 예측 가능성을 확인하였음
 - IMF의 News-Based Sentiment Indicators는 단어 간 의미적 관계를 활용하는 semantic clustering 기법을 도입하여 기존 키워드 중심 접근의 한계를 보완하고자 하였음
 - 한국은행 NSI는 Transformer 기반 딥러닝 모델을 활용하여 문장 수준의 감성을 분석함으로써 뉴스기반 경제심리지수 구축의 정교성을 한층 높인 사례로 평가됨

2. 뉴스 기반 경제지표 구축의 주요 특징

- 선행연구들은 뉴스 데이터가 경제심리 및 경기변동을 신속하게 반영할 수 있는 유용한 정보원임을 보여주고 있음
 - 뉴스 데이터는 설문조사 기반 지표와 달리 실시간으로 축적되므로 경제주체의 기대 변화와 심리 변화를 보다 빠르게 포착할 수 있음
 - 실제로 DNSI와 NSI는 소비자심리지수 및 기업경기실사지수 등에 선행하는

특성을 보였으며, IMF 연구에서도 금융위기 이전에 감성지수가 상승하는 경향이 확인됨

- 또한 뉴스 데이터는 기사 단위, 문장 단위, 의미군집 단위 등 다양한 수준에서 분석이 가능하여 연구 목적에 따라 유연한 활용이 가능함

□ 반면 뉴스 데이터 활용에는 해석상의 한계도 존재함

- 단순 키워드 기반 접근은 문맥을 충분히 반영하지 못할 수 있으며, 감성사전 기반 접근 역시 동일한 표현이 문맥에 따라 서로 다른 의미를 가질 수 있다는 한계를 가짐
- 의미기반 분석과 딥러닝 기반 분석은 보다 정교한 문맥 분석이 가능하지만 대규모 학습자료 구축과 추가적인 검증 과정이 요구됨

3. 본 연구에의 시사점

□ 선행연구 검토 결과, 뉴스 데이터는 소비와 관련된 경제활동의 변화를 고빈도로 포착할 수 있는 잠재력을 갖는 것으로 판단됨

- 다만 기존 연구들은 주로 경제심리, 정책 불확실성 또는 금융위험을 측정하는 데 초점을 두고 있으며, 소비 품목군별 소비 방향성을 직접 식별하는 연구는 상대적으로 제한적임
- 특히 민간소비는 내구재, 준내구재, 비내구재 등 품목별 특성이 상이하고, 동일한 기사 내에서도 소비 증가감소·가격 변화 등이 혼재되어 나타날 수 있으므로 보다 세분화된 접근이 필요함

□ 이에 본 연구는 선행연구의 방법론을 참고하여 소비 품목군별 소비 방향성 식별에 초점을 두고자 함

- 우선 규칙기반(lexicon-based) 분류체계를 활용하여 소비 관련 기사와 방향

성 신호를 선별하고, 이를 바탕으로 소비 방향성에 대한 1차 분류라벨을 구축함

- 이후 해당 결과를 딥러닝 기반 문맥분류모형의 학습자료로 활용함으로써 소비 관련 문맥을 보다 정교하게 식별할 수 있는 체계를 구축하고자 함
- 이를 통해 기존 경제심리지수 연구를 소비 분야로 확장하고, 뉴스자료를 활용한 고빈도 민간소비 보조지표 구축 가능성을 검토하고자 함.
- 아울러 각 품목군에 해당하는 품목별 정보를 활용한 민간소비 보조지표 구축 가능성을 검토하고자 함.

제3장 뉴스 빅데이터를 활용한 소비동향 기초분석

I. 분석 목적 및 방법

1. 분석 목적 및 범위

가. 분석 목적

- 본 장은 한국언론진흥재단 BIGKinds 경제뉴스 자료를 활용하여 소비 관련 뉴스의 규모와 시기별 변화, 주요 내용구조를 기초적으로 파악하는 데 목적이 있음
 - 전체 경제기사 가운데 제목에 ‘소비’가 명시적으로 포함된 기사를 선별하고, 기사 수와 전체 경제기사 대비 비중의 변화를 분석함
 - 이를 통해 소비 관련 보도가 전체 경제뉴스에서 차지하는 상대적 규모와 시기별 변화를 파악함
- 소비 관련 뉴스에서 반복적으로 나타나는 주요 내용과 시기별 의제 변화를 탐색함
 - BIGKinds가 제공하는 특성추출어를 기사 단위로 재집계하여 소비 관련 뉴스에서 빈번하게 나타나는 주요 표현을 확인함
 - 기사 제목의 의미적 유사성을 활용한 BERTopic 분석을 통해 주요 토픽을 탐색하고, 토픽별 비중의 시기적 변화를 살펴봄
 - 특정 시기에 소비 관련 보도가 증가한 경우에는 주요 특성추출어와 토픽의 변화를 함께 검토하여 기사량의 변화가 어떠한 뉴스 내용과 관련되어 나타나는지 확인함
- 본 장의 분석결과는 실제 민간소비의 수준이나 증감을 직접 측정하거나 예측하기 위한 것이 아니라, 소비 관련 뉴스의 규모와 내용구조를 파악하기 위

한 기초적·탐색적 분석으로 활용함

- 기사 수와 비중, 특성추출어의 빈도 및 토픽별 비중은 언론보도의 상대적 규모와 내용구성을 나타내며, 실제 소비지출의 크기나 경제적 영향력을 직접 의미하지 않음
- 단순 기사량과 어휘빈도만으로는 기사의 문맥과 실제 소비상황을 충분히 식별하기 어렵다는 한계를 함께 검토함
- 이러한 기초분석을 통해 뉴스자료에 포함된 소비 관련 정보의 특성과 한계를 파악하고, 제4장에서 기사 문맥을 활용하여 소비 관련 정보를 보다 정교하게 식별하고 소비예측지표를 개발·검증하기 위한 분석적 기초를 제공함

나. 분석대상 및 기간

- 분석대상은 한국언론진흥재단 BIGKinds의 통합분류상 ‘경제’ 분야에 포함된 국내 뉴스기사로 설정함
 - 별도의 검색어는 입력하지 않고 통합분류 ‘경제’와 그 하위분류 전체에서 수집된 기사자료를 활용함
 - 전체 경제기사 자료는 소비 관련 기사의 규모와 전체 경제뉴스 대비 상대적 비중을 산출하기 위한 기준집단으로 활용함
- 분석기간은 2022년 1월 1일부터 2026년 4월 30일까지로 설정함
 - 분석기간을 일치시킴으로써 제3장의 기초 뉴스분석과 제4장의 소비예측 분석이 동일한 시기적 범위에서 해석될 수 있도록 함
 - 2026년 자료는 1월부터 4월까지의 부분기간이므로 기사 수와 토픽 구성의 연도별 비교에서는 완전한 연도와 동일하게 해석하지 않음
- 원자료 통합, 뉴스 식별자 기준 중복 제거 및 BIGKinds의 ‘분석제외 여부’를 반영하여 최종 분석대상을 확정함

- 누락일 추가수집을 포함한 정제자료를 기준으로 분석기간의 전체 경제기사 5,838,352건을 최종 분석대상으로 사용함
- 동일한 뉴스 식별자가 중복된 경우 하나의 기사만 유지하고, BIGKinds의 '분석제외 여부'가 공란인 기사만 주요 분석에 포함함
- 분석기간 1,581일 전체에 기사가 존재하여 별도의 결측일 또는 결측월 처리는 적용하지 않음

□ 전체 경제기사 가운데 기사 제목에 문자열 '소비'가 명시적으로 포함된 76,301건을 소비 관련 기사로 선별함

- 소비 관련 기사는 최종 전체 경제기사의 약 1.31%에 해당함
- 제목에 동일한 문자열이 포함되었는지를 기준으로 적용하여 선별기준을 단순하고 재현 가능하게 구성함
- 다만 소비지출, 소매판매 또는 가계구매와 관련된 기사라도 제목에 '소비'가 포함되지 않으면 분석대상에서 제외될 수 있음
- 반대로 에너지 소비와 같이 가계의 상품·서비스 구매와 직접 관련되지 않은 의미의 '소비'가 포함된 기사도 일부 포함될 수 있음
- 따라서 본 분석의 소비 관련 기사 자료는 모든 민간소비 관련 뉴스를 포괄하는 자료라기보다, 제목에서 '소비'가 명시적으로 부각된 기사를 일관된 기준으로 선별한 분석집단으로 해석함

□ 분석은 기사 수와 전체 경제기사 대비 소비기사 비중, BIGKinds 특성추출어 및 기사 제목 기반 토픽구조를 중심으로 실시함

- 기사 수와 비중은 연도 및 월 단위로 집계하여 소비 관련 보도의 규모와 시기별 변화를 파악함
- 특성추출어와 BERTopic 분석은 소비 관련 뉴스의 내용구조와 주요 의제의 시기별 변화를 탐색하는 데 활용함

- 분석결과는 실제 소비지출 또는 민간소비의 수준을 직접 측정한 결과가 아니라 소비 관련 뉴스의 보도규모와 내용구성을 분석한 결과로 해석함

2. 자료 및 분석방법

가. 자료 구축 및 전처리

- 한국언론진흥재단 BIGKinds에서 수집한 경제뉴스 원자료를 기사 단위로 통합하여 분석용 정제자료를 구축함
 - 별도의 검색어를 입력하지 않고 BIGKinds 통합분류 ‘경제’와 그 하위분류 전체를 대상으로 자료를 수집함
 - 기존 수집자료는 총 640개의 엑셀 파일, 9,316,591개의 기사기록으로 구성됨
 - 기사 제목, 발행일자, 언론사, 뉴스 식별자, 통합분류, 키워드, 특성추출어 및 분석제외 여부 등 기사 단위 분석에 필요한 정보를 유지함
 - 원본 파일은 별도로 보존하고 분석용 정제자료를 별도로 구축하여 자료처리 과정의 재현성과 검증 가능성을 확보함
- 자료수집 당시 BIGKinds에서 선택 가능한 총 104개 언론사를 대상으로 하였으며, 구체적인 분석대상 언론사는 <표 3-1>과 같음
 - 기사 제목, 발행일자, 언론사, 뉴스 식별자, 통합분류, 키워드, 특성추출어 및 분석제외 여부 등 기사 단위 분석에 필요한 정보를 활용함

<표 3-1> BIGKinds 분석대상 언론사 목록

구분	언론사 수	언론사
전국일간지	12	경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 아시아투데이, 조선일보, 중앙일보, 한겨레, 한국일보
경제일간지	13	대한경제, 매일경제, 머니투데이, 메트로경제, 브릿지경제, 서울경제, 아시아경제, 아주경제, 이데일리, 이투데이, 파이낸셜뉴스, 한국경제, 헤럴드경제
지역일간지	45	강원도민일보, 강원일보, 경기신문, 경기일보, 경남도민일보, 경남신문, 경남일보, 경북도민일보, 경북매일신문, 경북일보, 경상일보, 경인일보, 광남일보, 광주매일신문, 광주일보, 국제신문, 금강일보, 기호일보, 남도일보, 대구신문, 대구일보, 대전일보, 동양일보, 매일신문, 무등일보, 부산일보, 새전북신문, 영남일보, 울산매일, 울산신문, 인천일보, 전남일보, 전라일보, 전북도민일보, 전북일보, 제민일보, 제주일보, 중도일보, 중부매일, 중부일보, 충북일보, 충청일보, 충청타임즈, 충청투데이, 한라일보
지역주간지	5	당진시대, 설악신문, 영주시민신문, 평택시민신문, 홍성신문
방송사	5	KBS, MBC, OBS, SBS, YTN
기타전문지	10	기자협회보, 디지털타임스, 미디어오늘, 소년한국일보, 시사IN, 일요신문, 전자신문, 주간한국, 한겨레21, 환경일보
스포츠지	3	스포츠서울, 스포츠월드, 스포츠한국
인터넷신문	11	EBN, PD-저널, 노컷뉴스, 뉴스핌, 뉴스핌, 데일리안, 브레이크뉴스, 비즈워치, 쿠키뉴스, 프레시안, 헬로디디

주: 분석대상 언론사는 BIGKinds 자료수집 당시 선택 가능한 매체 가운데 본 분석의 자료수집 대상으로 설정한 언론사임.

자료: 한국언론진흥재단 BIGKinds 자료수집 설정을 기준으로 연구진 정리.

□ 분석기간 중 실제 기사가 수록된 언론사 범위와 기사량의 집중도를 연도별로 추가 확인함

- 전체 경제기사의 경우 2022년에는 100개 언론사에서 기사가 수록되었으며, 2023년 이후에는 분석대상 104개 언론사 모두에서 경제기사가 확인됨
- 소비기사의 수록 언론사 수는 2022년 94개에서 2025년 103개까지 증가하였으며, 2026년 1~4월에는 99개로 나타남
- 기사 수 기준 상위 10개 언론사의 비중은 전체 경제기사에서 약 44~46%, 소비기사에서는 약 34~38% 수준으로 나타남
 - 이는 분석자료가 특정 소수 언론사의 기사만으로 구성된 것은 아니라는 자료구성상의 특성을 보여주지만, 언론사별 기사 생산량의 차이나 표본의 대표성을 통계적으로 검증한 결과로 해석하지 않음

<표 3-2> 연도별 수록 언론사 수 및 상위 10개 언론사 비중

연도	전체 경제기사 수록 언론사 수	상위 10개 비중(%)	소비기사 수록 언론사 수	상위 10개 비중(%)
2022	100	46.28	94	35.61
2023	104	45.22	98	35.28
2024	104	44.45	96	33.51
2025	104	44.88	103	34.33
2026.1~4	104	45.86	99	38.03

주: 1) 수록 언론사 수는 해당 연도에 분석대상 기사가 1건 이상 존재한 언론사의 수임. 2) 상위 10개 언론사 비중은 해당 연도 기사 수 기준 상위 10개 언론사의 기사 수가 전체 기사에서 차지하는 비율임. 3) 2026년은 1~4월 부분기간임.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

□ 뉴스 식별자를 기준으로 원자료 간 중복기사를 제거함

- 동일한 뉴스 식별자가 여러 파일에 포함된 경우 하나의 기사만 유지함
 - 뉴스 식별자가 결측된 8개 기록은 발행일자·언론사·제목 등 주요 기사정보도 확인되지 않아 유효한 기사기록으로 보기 어려운 자료로 판단하여 제

외함

- 파일 간 중복으로 확인된 초과기록 6,025건을 제거한 결과
9,310,558개의 고유한 뉴스 식별자를 확보함
- 중복 제거는 BIGKinds가 개별 기사에 부여한 ‘분석제외 여부’와 구분하여 적용함
 - 파일 간 중복 제거는 동일한 뉴스 식별자가 복수의 원자료 파일에 포함된 경우를 처리하는 절차임
 - BIGKinds의 ‘분석제외 여부’는 개별 기사에 대해 BIGKinds가 별도로 표시한 분석 제외 정보를 반영하는 절차임

□ 분석기간 전체에 대해 일별 기사자료의 수집상태를 점검하고 누락자료를 보완하여 최종 분석자료의 완전성을 확인함

- 기존 정제자료를 일별로 점검한 결과 일부 날짜의 기사자료 누락이 확인되어 해당 기간의 BIGKinds 원자료를 추가 수집하고 기존 자료와 통합함
- 추가수집 자료의 뉴스 식별자 결측 및 기존 자료와의 중복 여부를 점검한 후 정제자료에 반영하였으며, 그 결과 2022년 1월 1일부터 2026년 4월 30일까지 총 1,581일 모두에서 경제기사가 관측되어 별도의 결측일 또는 결측월 처리를 적용하지 않음

□ BIGKinds가 제공하는 ‘분석제외 여부’를 반영하여 최종 분석대상을 확정함

- 뉴스 식별자 기준 중복 제거와 누락일 추가수집을 완료한 자료에서 분석기간을 2022년 1월 1일부터 2026년 4월 30일까지로 제한함
- 이 가운데 ‘분석제외 여부’가 공란인 기사만 본 분석에 포함함
 - 해당 자료는 이후 소비기사 선별, 연도·월별 기사 규모 산출 및 내용분석의 기준자료로 활용함

□ 기사 발행일을 기준으로 연도 및 월 단위의 집계자료를 구축함

- 각 기간의 전체 경제기사 수와 소비기사 수를 각각 합산하여 기사 규모를 산출함
- 소비기사 비중은 동일 기간의 소비기사 수를 전체 경제기사 수로 나누어 산출함
 - 일별 비중의 단순평균이나 이동평균은 사용하지 않음
 - 소비기사 비중은 실제 소비 수준이나 소비증가율이 아니라 해당 시기 전체 경제뉴스 가운데 소비 관련 보도가 차지하는 상대적 비중으로 해석함

나. 소비기사 선별 및 내용분석

□ 전체 경제기사 가운데 기사 제목에 문자열 ‘소비’가 명시적으로 포함된 기사를 소비 관련 기사로 선별함

- 최종 전체 경제기사 가운데 제목에 ‘소비’가 포함된 76,301건을 소비 관련 기사로 분류함
- 동일한 문자열 포함 여부를 기준으로 적용하여 기사 선별과정을 단순하고 재현 가능하게 구성함
 - 제목에 ‘소비’가 포함되지 않은 소비지출·소매판매 관련 기사는 제외될 수 있으며, 반대로 에너지 소비 등 민간소비와 직접 관련되지 않은 기사도 일부 포함될 수 있음
 - 따라서 소비 관련 기사 자료는 모든 민간소비 관련 뉴스를 포괄하기보다 제목에서 ‘소비’가 명시적으로 부각된 기사를 일관된 기준으로 선별한 자료로 해석함

□ 소비 관련 뉴스의 규모와 시기별 변화를 파악하기 위해 기사 수와 전체 경제기사 대비 비중을 산출함

- 연도 및 월별 소비기사 수를 집계하고, 같은 기간의 전체 경제기사 수 대비

소비기사 비중을 산출함

- 소비기사 비중은 시기별 전체 경제뉴스 생산량의 차이를 고려하여 소비 관련 보도의 상대적 규모를 비교하기 위한 지표로 활용함
 - 해당 비중은 실제 소비활동이나 소비증가율을 나타내는 지표가 아니라 전체 경제뉴스 가운데 소비 관련 보도가 차지하는 비율을 의미함

□ 소비 관련 뉴스의 주요 내용을 파악하기 위해 BIGKinds가 기사 본문을 기반으로 산출하여 제공하는 특성추출어 정보를 활용함

- BIGKinds는 개별 기사에 대해 본문의 주요 개체·기관·정책·사건·경제현상 등을 나타내는 특성추출어를 가중치순으로 제공하며, 본 분석에서는 기사별 상위 특성추출어를 최대 50개까지 활용함
- 전체 소비기사 76,301건을 대상으로 기사별 특성추출어를 결합한 뒤, 각 표현이 몇 개의 기사에 주요 특성추출어로 포함되었는지를 기사 단위로 집계함
- 동일한 특성추출어가 하나의 기사에서 중복되어 나타나는 경우 기사 단위에서는 한 번만 포함하여, 해당 표현이 포함된 기사 수를 기준으로 빈도를 산출함
 - 이에 따라 특성추출어의 빈도는 해당 표현의 가중치 합계나 경제적 중요도가 아니라 전체 소비기사 가운데 해당 표현이 주요 특성추출어로 포함된 기사 수를 의미함

□ 특성추출어 집계 과정에서는 분석결과의 재현성을 유지하면서 동일하거나 분석상 불필요한 표현을 제한적으로 정리함

- ‘소비’, ‘소비자’와 같이 기사 선별기준에 의해 기계적으로 높은 빈도가 발생하는 표현과 단순 숫자·백분을 표현은 주요 특성추출어 집계에서 제외함
- 동일한 대상을 지칭하는 일부 표현은 일관된 명칭으로 통합함

- 예를 들어 ‘한은’은 ‘한국은행’, ‘소비자원’은 ‘한국소비자원’, ‘대한민국’·‘우리나라’는 ‘한국’, ‘코로나’는 ‘코로나19’로 통합함
- 특정 분석결과를 얻기 위해 개별 정책·기업·기관·경제용어를 사후적으로 제거하는 방식은 적용하지 않음

□ 소비 관련 뉴스에서 반복적으로 나타나는 주요 표현의 분포를 직관적으로 확인하기 위해 워드클라우드를 보조적으로 활용함

- 전체 분석기간의 특성추출어 기사 포함빈도를 기준으로 주요 표현의 크기를 다르게 표시함
- 워드클라우드는 주요 어휘의 분포를 탐색적으로 확인하기 위한 보조 시각화로 활용함
 - 단어의 크기는 기사 포함빈도를 의미하며 해당 표현의 경제적 중요성, 실제 소비에 대한 영향 또는 긍정·부정 방향을 의미하지 않음

다. BERTopic 기반 토픽분석

□ 소비 관련 뉴스의 주요 내용구조를 파악하기 위해 기사 제목을 대상으로 BERTopic(Grootendorst, 2022) 분석을 실시함

- 분석대상은 2022년 1월 1일부터 2026년 4월 30일까지의 소비 관련 기사 76,301건으로 설정함
- BERTopic은 사전학습된 Transformer 기반 언어모형을 활용하여 문서를 의미 벡터로 변환하고, 의미적으로 유사한 문서를 군집화한 뒤 class-based TF-IDF(c-TF-IDF)를 활용하여 각 군집의 대표적인 토픽 표현을 산출하는 임베딩 기반 토픽모형임
- 전통적인 단어 출현빈도 중심 분석과 달리 문서 전체의 의미적 유사성을 활용할 수 있으므로, 본 연구에서는 동일한 ‘소비’ 관련 기사집단에 포함된 서

로 다른 내용영역을 탐색적으로 구분하는 데 활용함

□ 기사 제목의 의미적 유사성을 반영하기 위해 다국어 문장 임베딩, 차원축소 및 밀도기반 군집화 방법을 순차적으로 활용함

- 기사 제목은 paraphrase-multilingual-MiniLM-L12-v2 모델을 이용하여 문장 수준의 의미정보를 반영하는 벡터로 변환함

- Sentence-BERT 계열의 문장 임베딩은 문장 전체를 의미적으로 비교할 수 있는 고정차원의 벡터표현으로 변환하도록 설계되었으며(Reimers & Gurevych, 2019), 다국어 문장 임베딩은 서로 다른 언어의 의미적으로 유사한 문장이 가까운 벡터 공간에 위치하도록 확장될 수 있음(Reimers & Gurevych, 2020)

- 변환된 문장 임베딩을 바탕으로 UMAP(McInnes et al., 2018)을 이용한 차원축소와 HDBSCAN(McInnes et al., 2017)을 이용한 밀도기반 군집화를 순차적으로 수행하여 의미적으로 유사한 기사들을 토픽별로 분류함¹⁵⁾

- 세부 모형 설정값은 분석의 재현성을 위해 기록하되, 본문에서는 토픽 도출과 선정 및 해석 절차를 중심으로 제시함

□ 초기 군집화에서 특정 토픽에 배정되지 않은 기사는 문장 임베딩 유사도를 이용하여 기존 토픽에 재배정함

- 밀도기반 군집화의 특성상 의미적 유사성이 충분하지 않은 일부 기사는 초기 단계에서 이상치(outlier)로 분류될 수 있으므로, BERTopic의 임베딩 기반 이상치 재배정 절차를 적용함¹⁶⁾

- 이상치 재배정은 새로운 토픽을 추가하는 것이 아니라 초기 군집에서 도출

15) 문장 임베딩은 paraphrase-multilingual-MiniLM-L12-v2 모델을 이용하여 384차원 벡터로 생성함. UMAP은 n_neighbors=15, n_components=5, min_dist=0, metric=cosine, random_state=42로 설정하였으며, HDBSCAN은 min_cluster_size=150, min_samples=20, metric=euclidean, cluster_selection_method=EOM을 적용함.

16) 초기 군집화에서는 총 78개 토픽이 도출되었으며, 전체 소비기사 76,301건 중 31,131건(40.8%)이 이상치로 분류됨. 이후 BERTopic의 임베딩 기반 이상치 재배정(reduce_outliers, threshold=0.20)을 적용하여 기존 토픽과의 의미적 유사성을 기준으로 재배정함.

된 토픽구조를 유지하면서, 기존 토픽과 일정 수준 이상의 의미적 유사성을 갖는 미배정 기사를 해당 토픽에 추가하는 방식으로 수행함

- 이상치 재배정 결과 초기 이상치로 분류된 기사들은 모두 기존 토픽에 재배정되어 최종적으로 별도의 미배정 기사(-1)는 남지 않음

□ 각 토픽의 내용은 초기 군집에 포함된 기사와 대표 기사 제목 및 BIGKinds 특성추출어를 종합하여 해석함

- 각 토픽의 대표 기사 제목과 주요 특성추출어를 검토하여 공통적으로 나타나는 내용적 특징을 파악하고 이를 바탕으로 토픽명을 부여함
 - 토픽명은 사전에 설정한 분류체계가 아니라 분석결과를 바탕으로 연구진이 부여한 해석적 명칭임
 - 자동 생성된 상위단어나 토픽명을 그대로 사용하지 않고 실제 기사 제목과 특성추출어를 함께 검토하여 토픽의 내용적 일관성을 확인함

□ 분석결과의 제시에서는 토픽의 규모와 초기 군집구조의 유지 정도를 고려하여 주요 토픽을 선별함

- 초기 군집에 직접 포함된 기사 수가 500건 이상이면서, 최종 토픽 기사 가운데 초기 군집 기사의 비율이 70% 이상인 토픽을 주요 토픽으로 선정함
- 이러한 기준을 모두 충족한 17개 토픽을 주요 토픽으로 선정하고, 토픽별 기사 규모와 연도별 비중 변화를 분석함
 - 초기 군집 기사 수 기준은 일정 규모 이상의 내용적 집단을 선별하기 위한 것이며, 초기 군집 기사 비율 기준은 이상치 재배정 이후에도 최초 군집구조가 상당 부분 유지되는 토픽을 선별하기 위한 것임
 - 이러한 기준은 통계적 유의수준이나 최적 임계값을 의미하는 것이 아니라, 전체 78개 토픽 가운데 일정한 규모와 초기 군집구조를 유지하여 내용적 해석이 상대적으로 용이한 토픽을 보고서에 제시하기 위한 실무적 선정기준임

- 선정된 17개 주요 토픽은 전체 소비 기사를 완전하게 분류하기 위한 체계가 아니므로 주요 토픽별 비중의 합계는 전체 소비 기사의 100%와 일치하지 않음

□ 주요 토픽의 시기별 변화는 각 연도의 전체 소비 기사 수를 기준으로 토픽별 기사 비중을 산출하여 비교함

- 각 연도의 토픽별 기사 수를 동일 연도의 전체 소비 기사 수로 나누어 토픽 비중을 산출함
- 2026년은 1월부터 4월까지의 자료만 포함하므로 연도 간 비교에서는 부분기간이라는 점을 고려함
 - 토픽 비중은 실제 소비 지출의 구성비가 아니라 해당 시기 소비 관련 뉴스 가운데 각 토픽이 차지하는 보도 비중을 의미함

□ 이상의 자료처리 및 분석과정에서 활용한 주요 정보와 분석방법을 정리하면 <표 3-3>과 같음

<표 3-3> BIGKinds 소비기사 분석에 활용한 주요 정보와 분석방법

분석정보	주요 내용	활용 목적
기사 일자·뉴스 식별자	기사 발행일과 기사별 고유 식별자	분석기간 설정, 중복기사 제거 및 연도·월별 집계
분석제외 여부	BIGKinds가 제공하는 분석 제외 표시	최종 분석대상 기사 확정
기사 제목	제목에 문자열 ‘소비’가 포함된 기사 선별	소비 관련 기사 분석집단 구축
기사 수·비중	기간별 소비기사 수와 동일 기간 전체 경제기사 대비 비중	소비 관련 보도의 규모와 시기별 상대적 변화 파악
특성추출어	BIGKinds가 기사별로 제공하는 가중치순 상위 특성추출어	소비 관련 뉴스의 주요 표현과 어휘구조 탐색
워드클라우드	특성추출어의 기사단위 포함빈도를 크기로 시각화	주요 특성추출어 분포의 탐색적 확인
BERTopic	기사 제목의 문장 임베딩과 군집화를 이용한 토픽 탐색	소비 관련 뉴스의 주요 내용구조와 시기별 토픽 변화 파악

주: 1) 소비기사는 기사 제목에 문자열 ‘소비’가 포함된 기사로 정의함. 2) 소비기사 비중은 동일 기간 전체 경제기사 수 대비 소비기사 수의 비율이며 실제 소비활동의 수준이나 증감률을 의미하지 않음. 3) 특성추출어는 BIGKinds가 기사별로 제공하는 가중치순 상위 표현을 활용하고, 기사 단위 중복을 제거하여 포함빈도를 산출함. 4) 워드클라우드는 특성추출어 분포를 확인하기 위한 탐색적 보조 시각화임. 5) BERTopic은 기사 제목의 의미적 유사성을 이용하여 주요 토픽을 탐색하고 시기별 변화를 분석하는 데 활용함.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

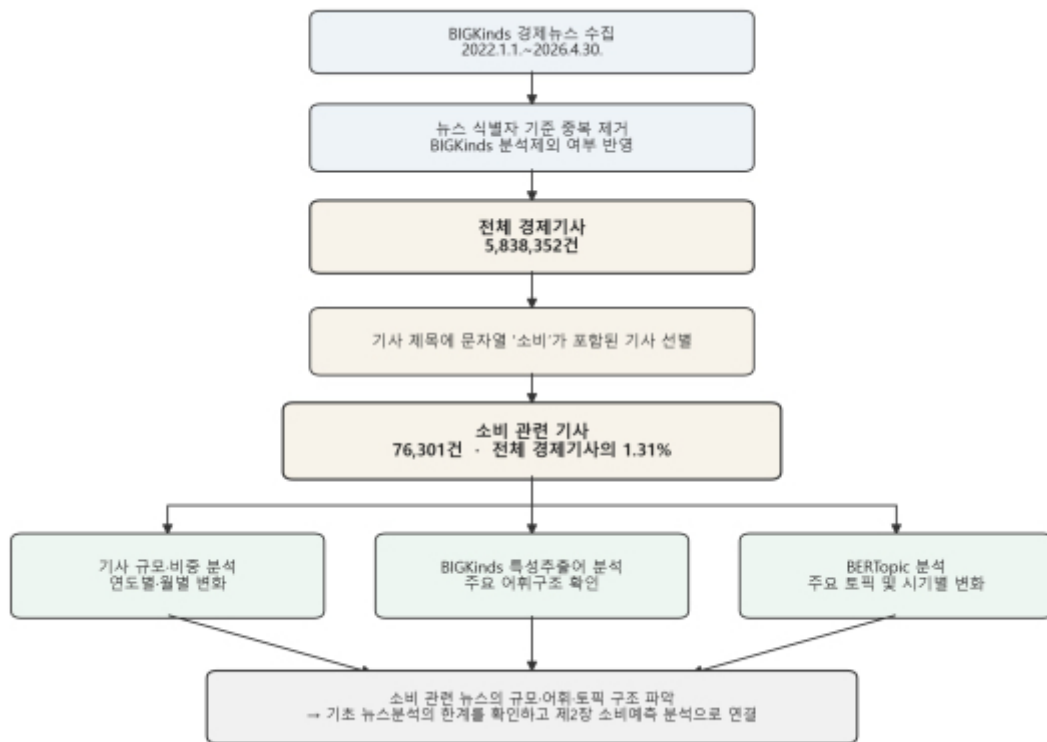
□ 본 분석은 기사 규모의 변화와 뉴스 내용의 구조를 단계적으로 파악하는 방식으로 구성함

- 먼저 연도별·월별 기사 수와 전체 경제기사 대비 소비기사 비중을 산출하여 소비 관련 보도의 규모와 시기별 변화를 확인함
- 다음으로 BIGKinds가 제공하는 특성추출어를 활용하여 소비 관련 뉴스에서 반복적으로 나타나는 주요 표현과 어휘구조를 파악함
 - 특성추출어의 기사단위 포함빈도를 기준으로 주요 표현을 정리하고, 워드클라우드는 이러한 분포를 직관적으로 확인하기 위한 보조적 시각화로 활용함
- 마지막으로 단어 단위 분석만으로는 파악하기 어려운 기사 간 의미적 유사성과 내용구조를 확인하기 위해 BERTopic 기반 토픽분석을 수행함
- 이를 통해 소비 관련 뉴스에서 반복적으로 나타나는 주요 토픽을 탐색하고, 각

토픽의 시기별 비중 변화를 비교함

- 이상의 자료 구축과 분석절차를 종합하면 <그림 3-1>과 같음
- 전체 경제뉴스에서 최종 분석대상을 확정한 뒤, 소비 관련 기사에 대해 기사 규모·비중, 특성추출어, BERTopic 분석을 순차적으로 적용함
- 각 분석은 소비 관련 뉴스의 규모, 주요 어휘, 내용구조를 서로 다른 수준에서 파악하기 위한 것으로, 개별 분석결과를 실제 소비수준이나 소비경기의 직접적인 지표로 해석하지 않음

<그림 3-1> BIGKinds 소비 관련 뉴스 분석 절차



주: 1) 전체 경제기사 중 기사 제목에 문자열 '소비'가 포함된 기사를 소비 관련 기사로 선별하였으며, 기사 규모·비중, BIGKinds 특성추출어 및 BERTopic을 활용하여 기초·탐색 분석을 수행함.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

- 이와 같은 기초·탐색 분석을 통해 뉴스 기사량과 단어 수준의 정보만으로는 소비동향을 충분히 설명하기 어렵다는 점을 확인하고, 이후 제4장에서는 기사 문맥과 고차원 표현을 활용한 소비예측 분석으로 확장함

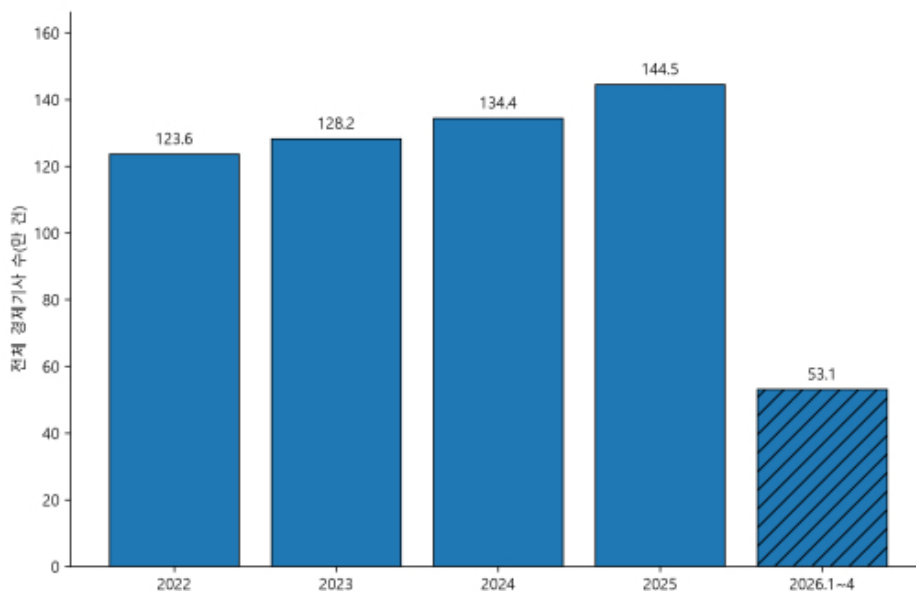
Ⅱ. 소비 관련 뉴스의 규모와 변화

1. 전체 경제기사 규모

□ 2022년 1월부터 2026년 4월까지 분석대상 전체 경제기사는 총 5,838,352건으로 나타남

- 연간 기사 수는 2022년 1,236,153건, 2023년 1,282,484건, 2024년 1,343,937건, 2025년 1,444,970건으로 증가함
 - 2022년 대비 2025년의 전체 경제기사 수는 약 20.9만 건 증가하여 분석기간 중 경제뉴스의 전체 생산규모가 확대되는 흐름을 보임
 - 다만 기사 수의 증가는 경제활동이나 특정 경제현상의 확대를 직접 의미하지 않으며, 경제현안과 언론의 보도량 및 기사 생산규모 등이 함께 반영된 결과로 해석할 필요가 있음

<그림 3-2> 연도별 전체 경제기사 규모



주: 1) 전체 경제기사는 BIGKinds 통합분류상 경제 분야 기사 가운데 뉴스 식별자 기준 중복기사와 BIGKinds 분석제외 기사를 처리한 최종 분석자료임. 2) 막대 위의 수치는 연도별 기사 수를 만 건 단위로 환산한 값임. 3) 2026년은 1월부터 4월까지의 부분기간이며 이전 연도의 연간 기사 수와 직접 비교하지 않음.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

- 2026년 1~4월에는 총 530,808건의 경제기사가 수집됨
 - 2026년 자료는 4개월의 부분기간이므로 이전 연도의 연간 기사 수와 직접 비교하지 않음
 - 하반기까지의 기사 생산규모가 동일하다고 가정한 단순 연환산도 실시하지 않으며, 이후 연도별 분석에서도 2026년은 1~4월의 부분기간으로 구분하여 해석함

- 전체 경제기사 규모는 소비 관련 보도의 상대적 규모를 비교하기 위한 기준 자료로 활용함
 - 전체 경제기사 수가 시기별로 달라질 경우 소비기사의 절대건수만으로 소비 관련 보도의 상대적 규모를 비교하기 어려움
 - 이에 따라 이후 분석에서는 소비기사 수와 함께 동일 기간의 전체 경제기사 대비 소비기사 비중을 제시함
 - 소비기사 비중은 전체 경제뉴스 생산량의 변화를 고려하여 소비 관련 보도가 차지하는 상대적 비중을 비교하기 위한 값임
 - 따라서 전체 경제기사 수와 소비기사 비중 모두 실제 경제활동이나 민간소비의 수준을 직접 나타내는 지표로 해석하지 않음

2. 소비기사 규모와 시계열 변화

- 2022년 1월부터 2026년 4월까지 전체 경제기사 가운데 소비기사는 총 1.31%를 차지함
 - 연도별 전체 경제기사와 소비기사 규모는 하단의 <표 3-4>과 같음
 - 연도별 소비기사 수는 2022년 14,491건, 2023년 15,846건, 2024년 16,564건, 2025년 23,423건으로 나타남
 - 같은 기간 전체 경제기사 대비 소비기사 비중은 각각 1.17%, 1.24%, 1.23%,

1.62%로 나타남

- 2022~2024년에는 소비기사 비중이 1.17~1.24% 수준에서 비교적 안정적으로 유지된 반면, 2025년에는 1.62%로 상승함
- 2026년 1~4월 소비기사는 5,977건으로 전체 경제기사의 1.13%를 차지함

<표 3-4> 연도별 전체 경제기사 및 소비기사 규모

연도	전체 경제기사 수	소비기사 수	소비기사 비중(%)
2022	1,236,153	14,491	1.17
2023	1,282,484	15,846	1.24
2024	1,343,937	16,564	1.23
2025	1,444,970	23,423	1.62
2026.1~4	530,808	5,977	1.13
합계	5,838,352	76,301	1.31

주: 1) 소비기사는 기사 제목에 문자열 '소비'가 포함된 기사임. 2) 소비기사 비중은 각 기간의 전체 경제기사 수 대비 소비기사 수의 비율임. 3) 2026년은 1월부터 4월까지의 자료임.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

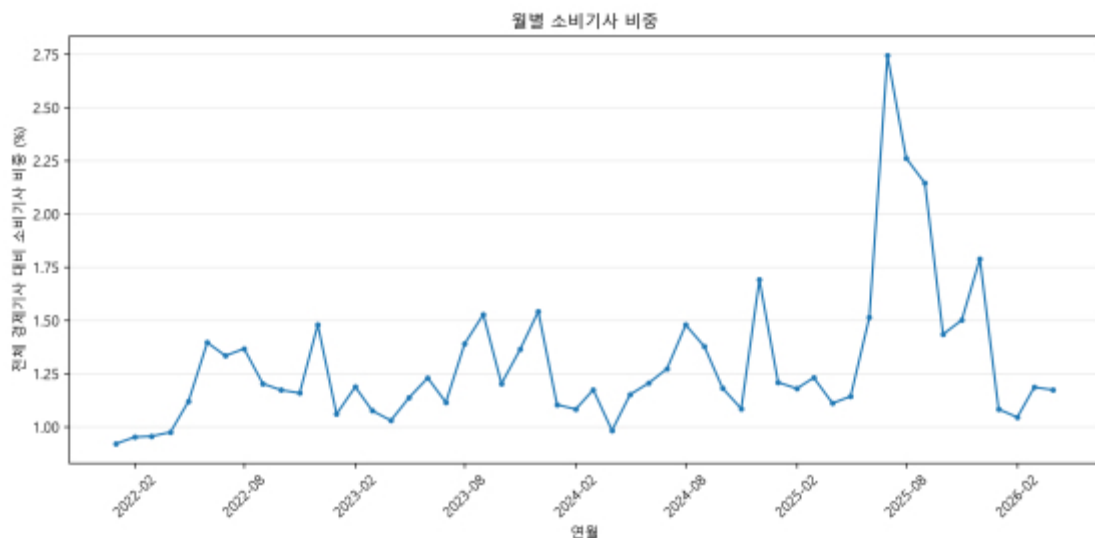
□ 소비기사의 절대건수와 전체 경제기사 대비 비중은 함께 해석할 필요가 있음

- 전체 경제기사 수 자체가 연도별로 변화하기 때문에 소비기사 수의 증감만으로 소비 관련 보도의 상대적 확대 여부를 판단하기 어려움
- 이에 따라 소비기사 수와 함께 전체 경제기사 대비 비중을 제시하여 시기별 전체 경제뉴스 생산규모의 차이를 고려함
- 소비기사 비중은 실제 민간소비의 증가율이나 소비경기의 수준을 의미하는 것이 아니라 전체 경제뉴스에서 소비 관련 보도가 차지하는 상대적 비중을 나타냄

□ 월별 소비기사 비중을 통해 소비 관련 보도의 단기적인 변화를 추가로 확인함

- 월별 소비기사 비중은 각 월의 소비기사 수를 동일 월의 전체 경제기사 수로 나누어 산출함
- 분석기간 전체에서 월별 비중은 시기별로 변동하며, 특히 2025년 중 일부 월에서 상대적으로 높은 수준이 나타남
 - 2025년 7월 소비기사 비중은 2.74%로 분석기간 중 가장 높았으며, 8월 2.26%, 9월 2.15%로 높은 수준이 이어짐
 - 이러한 변화는 실제 소비활동의 증가를 직접 의미하지 않으므로, 해당 시기의 소비 관련 뉴스 내용과 함께 해석할 필요가 있음

<그림 3-3> 월별 소비기사 비중



주: 1) 소비기사는 기사 제목에 문자열 '소비'가 포함된 기사임. 2) 월별 소비기사 비중은 각 월의 전체 경제기사 수 대비 소비기사 수의 비율임. 3) 해당 비중은 실제 소비활동의 수준이나 증감률을 의미하지 않음.
자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

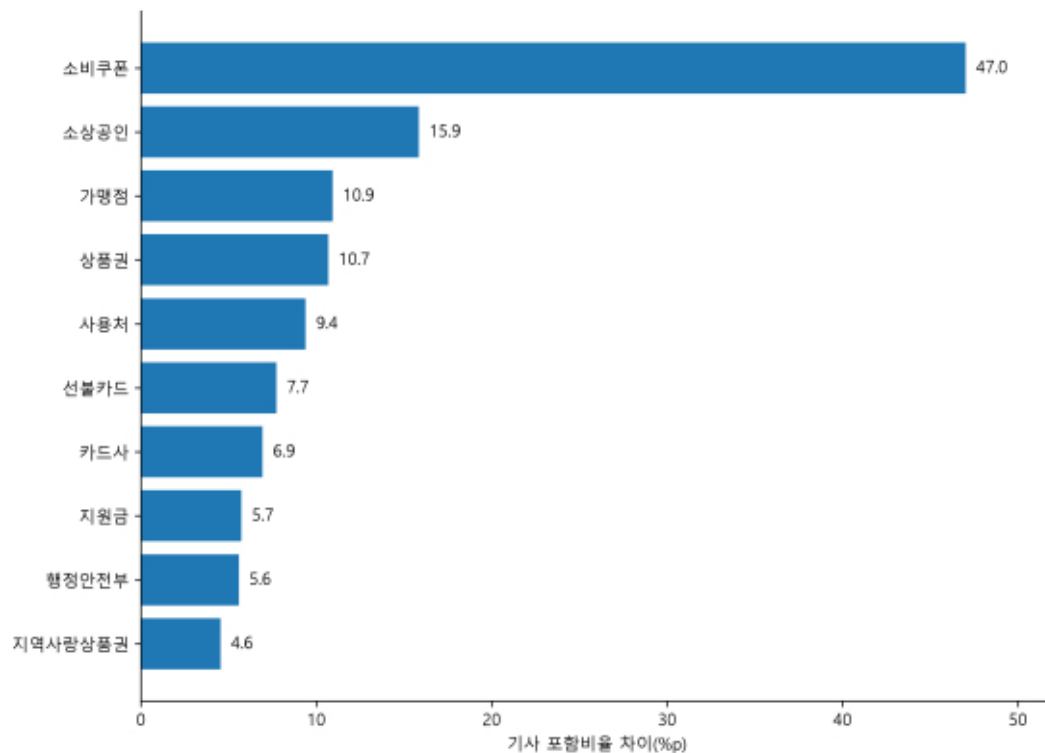
□ 월별 소비기사 비중은 시기별로 변동하며, 특히 2025년 7~9월에 상대적으로 높은 수준이 나타남

- 2025년 7월 소비기사 비중은 2.74%로 분석기간 중 가장 높았으며, 8월

2.26%, 9월 2.15%로 높은 수준이 이어짐

- 이러한 변화가 어떠한 뉴스 내용과 함께 나타났는지 확인하기 위해 2025년 7~9월 소비기사와 그 외 기간 소비기사의 주요 특성추출어 기사 포함비율을 비교함
 - 비교 결과 ‘소비쿠폰’의 기사 포함비율은 2025년 7~9월에 그 외 기간보다 47.0%p 높았으며, ‘소상공인’ 15.9%p, ‘가맹점’ 10.9%p, ‘상품권’ 10.7%p, ‘사용처’ 9.4%p 등의 순으로 차이가 크게 나타남
 - 이 밖에도 ‘선불카드’, ‘카드사’, ‘지원금’, ‘행정안전부’, ‘지역사랑상품권’ 등의 표현이 같은 기간 상대적으로 집중되어 나타남
- 이는 2025년 7~9월 소비기사 비중의 상승이 민생회복 소비쿠폰 및 관련 소비지원정책 보도의 집중과 함께 나타났음을 보여줌
 - 다만 이러한 결과는 특정 시기 뉴스 내용구성의 차이를 보여주는 기술적 비교이며, 실제 소비활동의 증가나 정책효과를 직접 의미하지 않음

<그림 3-4> 2025년 7~9월과 그 외 기간의 주요 특성추출어 기사 포함비율 차이



주: 1) 2025년 7~9월 소비기사 9,044건과 그 외 기간 소비기사 67,257건을 비교함. 2) 각 특성추출어의 기사 포함비율은 해당 기간 소비기사 중 해당 표현이 기사별 상위 특성추출어에 포함된 기사 비율임. 3) 그림의 값은 2025년 7~9월 기사 포함비율에서 그 외 기간의 기사 포함비율을 차감한 비율 차이(%p)임. 4) 이는 기사 내용구성의 차이를 나타내는 것으로 실제 소비증가 또는 정책효과를 의미하지 않음.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

Ⅲ. 소비 관련 뉴스의 주요 내용과 특징

1. 주요 특성추출어와 어휘구조

- 소비 관련 뉴스의 주요 내용을 파악하기 위해 전체 분석기간의 BIGKinds 특성추출어를 기사 단위로 집계함
 - 각 소비기사에서 BIGKinds가 제공하는 가중치순 상위 특성추출어를 활용하여, 전체 소비기사에서 반복적으로 나타나는 표현의 기사 포함빈도를 산출함
 - 동일한 표현이 하나의 기사에서 중복되는 경우 한 번만 집계하여 특정 표현이 몇 개의 소비기사에서 주요 특성추출어로 나타났는지를 기준으로 비교함
 - 따라서 특성추출어의 빈도는 단어 자체의 가중치 합계나 경제적 중요도가 아니라 해당 표현이 주요 특성추출어로 포함된 기사 수를 의미함

- 전체 분석기간에서 빈도가 높은 특성추출어로는 ‘미국’, ‘한국’, ‘온라인’, ‘서울’, ‘통계청’, ‘소비쿠폰’, ‘한국은행’, ‘코로나19’, ‘중국’, ‘상승률’ 등이 나타남
 - ‘미국’은 8,864건, ‘한국’은 7,583건, ‘온라인’은 7,329건, ‘서울’은 7,001건의 소비기사에서 주요 특성추출어로 나타남
 - 이와 함께 ‘통계청’ 6,130건, ‘소비쿠폰’ 5,503건, ‘한국은행’ 5,472건, ‘코로나19’ 4,803건, ‘중국’ 4,647건, ‘상승률’ 4,529건 등이 높은 빈도를 보임
 - 상위 특성추출어에는 소비 자체를 직접 설명하는 표현뿐 아니라 국가명, 지역명, 기관명, 거시경제 관련 표현도 다수 포함됨
 - 이는 단순한 상위 빈도만으로 소비 관련 뉴스의 핵심 내용을 곧바로 특정하기 어렵다는 점을 보여줌

<표 3-5> 소비 관련 뉴스의 주요 특성추출어 상위 15개

순위	특성추출어	기사 수	소비가사 대비 비중(%)
1	미국	8,864	11.62
2	한국	7,583	9.94
3	온라인	7,329	9.61
4	서울	7,001	9.18
5	통계청	6,130	8.03
6	소비쿠폰	5,503	7.21
7	한국은행	5,472	7.17
8	코로나19	4,803	6.29
9	중국	4,647	6.09
10	상승률	4,529	5.94
11	한국소비자원	4,470	5.86
12	소상공인	4,075	5.34
13	농산물	3,614	4.74
14	자동차	3,299	4.32
15	상승	3,233	4.24

주: 1) 특성추출어는 BIGKinds가 기사별로 제공하는 가중치순 상위 특성추출어를 활용함. 2) 동일한 특성추출어가 한 기사에서 중복되는 경우 한 번만 집계하였으며, 기사 수는 해당 표현이 주요 특성추출어로 포함된 소비기사 수를 의미함. 3) 소비기사 대비 비중은 전체 소비기사 76,301건 가운데 해당 특성추출어가 포함된 기사 비율임. 4) ‘소비’, ‘소비자’ 등 선별기준에 의해 기계적으로 빈도가 높아지는 표현과 단순 숫자·백분율 표현은 제외하고, 일부 동일 의미 표현은 통합함.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성

□ 주요 특성추출어에는 소비정책, 물가, 유통·결제, 산업 및 시장상황과 관련된 표현이 혼재되어 나타남

- ‘소비쿠폰’, ‘소상공인’, ‘상품권’, ‘가맹점’ 등은 소비지원 정책과 지역상권 관련 보도를 반영함
- ‘물가’, ‘소비자물가’, ‘인플레이션’, ‘상승률’ 등은 가격변동과 소비여건에 관한 보도에서 반복적으로 나타남
- ‘온라인’, ‘오프라인’, ‘자동차’, ‘반도체’ 등은 소비경로 또는 특정 산업·시장의 변화와 관련된 기사에서 나타남
 - 이처럼 동일한 ‘소비’ 기사집단 안에서도 정책, 물가, 산업, 시장, 제도 등 서로 다른 내용영역이 함께 포함되어 있음을 확인할 수 있음

- 워드클라우드는 주요 특성추출어의 분포를 직관적으로 확인하기 위한 보조적 시각화로 활용함
 - 특성추출어의 기사 포함빈도가 높을수록 크게 표시하여 전체 소비기사에서 반복적으로 나타나는 주요 표현의 분포를 시각적으로 확인함
 - 다만 워드클라우드의 단어 간 관계나 기사 문맥을 반영하지 못하므로 독립적인 분석결과라기보다 특성추출어 집계결과를 보완하는 탐색적 도구로 해석함

<그림 3-5> 소비 관련 뉴스의 주요 특성추출어 워드클라우드



주: 1) 워드클라우드의 BIGKinds가 기사별로 제공하는 주요 특성추출어의 기사 포함빈도를 기준으로 작성함.
 2) 동일한 특성추출어가 하나의 기사에서 중복되는 경우 한 번만 집계함. 3) 단어의 크기는 해당 표현이 주요 특성추출어로 포함된 기사 수를 의미하며 경제적 중요도나 실제 소비에 대한 영향의 크기를 의미하지 않음.
 자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

- 단순 특성추출어 분석만으로는 소비 관련 뉴스의 내용구조를 충분히 설명하는 데 한계가 있음
 - 상위 빈도어에는 ‘미국’, ‘한국’, ‘서울’과 같이 소비의 구체적 내용과 직접 연결하기 어려운 일반적 표현도 포함됨

- 또한 개별 단어의 빈도만으로는 동일한 표현이 어떠한 사건·정책·경제상황의 맥락에서 사용되었는지를 구분하기 어려움
 - 이에 따라 다음 절에서는 기사 제목 전체의 의미적 유사성을 반영한 BERTopic 분석을 통해 소비 관련 뉴스의 주요 토픽과 시기별 변화를 추가로 살펴봄

2. BERTopic 기반 주요 토픽과 시기별 변화

- BERTopic 분석을 통해 소비 관련 뉴스에 포함된 다양한 내용영역을 주요 토픽으로 구분함
 - 전체 소비기사에서 도출된 토픽 가운데 토픽의 규모와 이상치 재배정 이후 초기 군집구조의 유지 정도를 기준으로 17개 주요 토픽을 선정함
 - 주요 토픽에는 식품·농산물 소비, 금융소비자 보호, 쌀 소비촉진, 에너지·전력 소비, 소비자물가, 민생회복 소비쿠폰, 카드·결제 및 소비행태, 중국 소비시장·경기, 민간소비와 경제성장 등이 포함됨
 - 이 밖에도 부동산 소비심리, 지역 물가·소비심리, 생산·소비·투자 동향, 친환경·가치소비, 산업활동·소매판매, 자동차 소비시장, 보험·소비자보호, AI와 소비자·소비시장 등의 토픽이 확인됨
 - 주요 17개 토픽은 전체 소비기사의 25.69%를 차지하며, 전체 기사에 대한 완전한 분류체계가 아니라 규모와 해석 가능성이 상대적으로 높은 토픽을 중심으로 제시한 결과임

〈표 3-6〉 BERTopic 기반 주요 소비 관련 토픽

순위	주요 토픽	초기 군집 기사 수	최종 기사 수	초기 군집 비율(%)
1	식품·농산물 소비	2,404	2,832	84.9
2	금융소비자 보호	1,826	2,399	76.1
3	쌀 소비촉진	1,992	1,995	99.8
4	에너지·전력 소비	1,237	1,346	91.9
5	소비자물가	886	1,230	72

6	민생회복 소비쿠폰	960	1,208	79.5
7	카드·결제 및 소비행태	890	966	92.1
8	중국 소비시장·경기	838	966	86.7
9	민간소비와 경제성장	727	901	80.7
10	부동산 소비심리	796	844	94.3
11	지역 물가·소비심리	592	839	70.6
12	생산·소비·투자 동향	650	770	84.4
13	친환경·가치소비	565	734	77
14	산업활동·소매판 매	523	712	73.5
15	자동차 소비시장	524	659	79.5
16	보험·소비자보호	538	636	84.6
17	AI와 소비자·소비시장	505	561	90

주: 1) 주요 토픽은 초기 군집에 직접 포함된 기사 수가 500건 이상인 반면, 최종 토픽 기사 가운데 초기 군집 기사 비율이 70% 이상인 토픽을 대상으로 선정함. 2) 초기 군집 기사 수는 HDBSCAN 군집화 단계에서 해당 토픽에 직접 배정된 기사 수를 의미함. 3) 최종 기사 수는 이상치 재배정 이후 해당 토픽에 포함된 기사 수를 의미함. 4) 초기 군집 비율은 최종 기사 수 가운데 초기 군집에 직접 포함된 기사 수의 비율임. 이는 분류 정확도나 신뢰도를 의미하지 않음. 5) 토픽명은 대표 기사 제목과 주요 특성추출어를 검토하여 연구진이 부여한 해석적 명칭임.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

□ 주요 토픽의 규모는 토픽별로 차이를 보이며, 소비 관련 뉴스가 단일한 내용 영역으로 구성되지 않음을 확인함

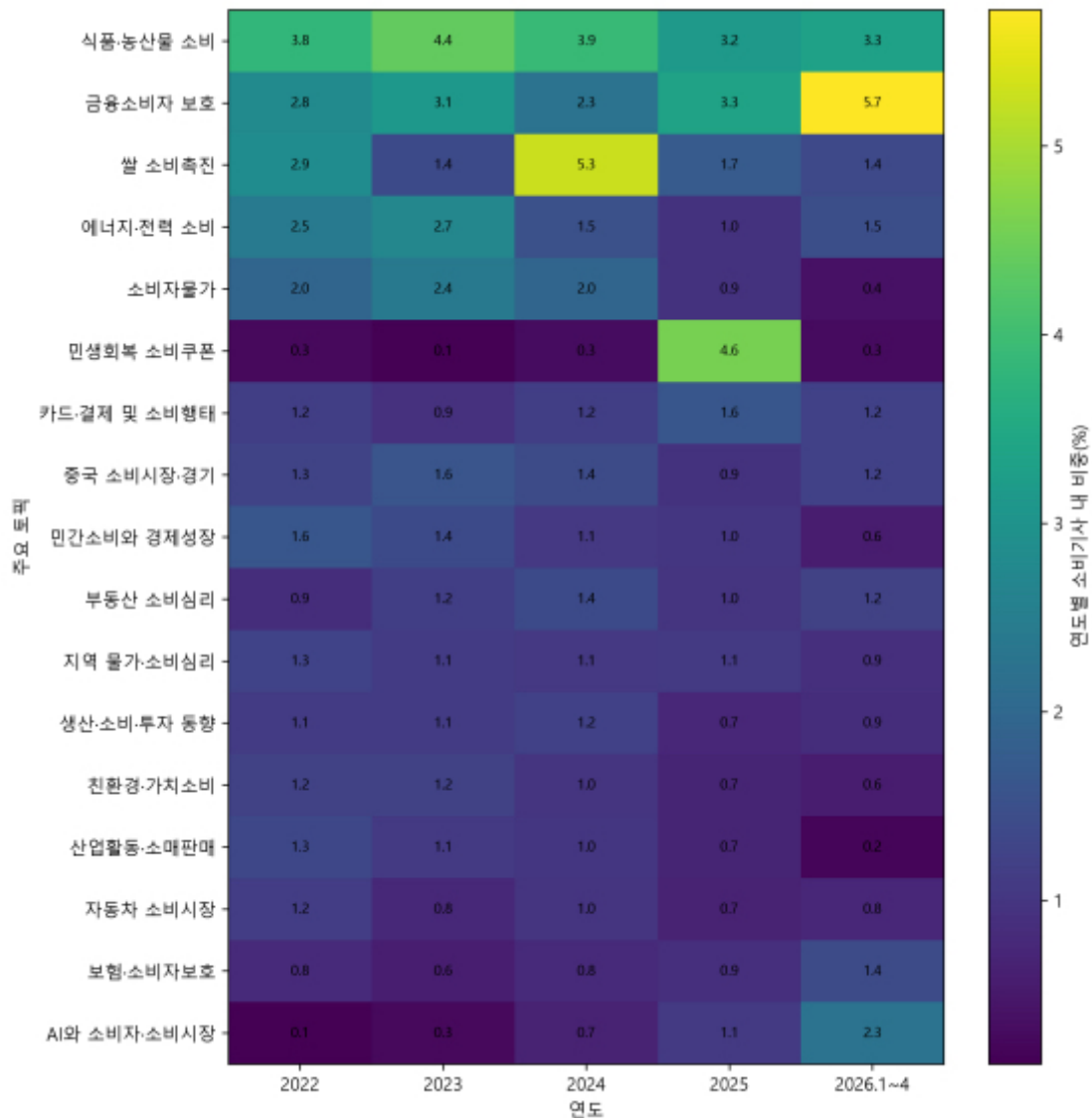
- 주요 토픽 가운데 식품·농산물 소비가 가장 큰 규모로 나타났으며, 금융소비자 보호, 쌀 소비촉진, 에너지·전력 소비, 소비자물가 등이 뒤를 이음
- 민생회복 소비쿠폰, 카드·결제 및 소비행태, 민간소비와 경제성장, 산업활동·소매판매 등은 소비정책이나 실제 경제활동과 상대적으로 직접 연결되는 내용영역으로 나타남
 - 반면 에너지·전력 소비처럼 ‘소비’라는 표현이 가계의 상품·서비스 구매와 다른 의미로 사용된 토픽도 확인됨

- 이는 기사 제목에 ‘소비’가 포함되었는지를 기준으로 분석대상을 선별할 경우 서로 다른 의미의 소비 관련 기사가 함께 포함될 수 있음을 보여줌

□ 주요 토픽의 연도별 비중을 비교한 결과 일부 토픽은 특정 시기에 집중되는 반면, 다른 토픽은 비교적 지속적으로 나타남

- 식품·농산물 소비, 금융소비자 보호 등은 분석기간 전반에 걸쳐 일정한 비중으로 나타나는 반면, 일부 토픽은 특정 연도에 상대적으로 높은 비중을 보임
- 쌀 소비촉진은 2024년에 상대적으로 높은 비중을 보였으며, 민생회복 소비쿠폰은 2025년에 뚜렷하게 확대됨
- AI와 소비자·소비시장 토픽은 2024년 이후 비중이 높아지는 흐름을 보이며 2026년 1~4월에는 2.26%까지 증가함
 - 이러한 결과는 소비 관련 뉴스의 내용구성이 고정되어 있기보다 정책환경과 산업·시장 변화에 따라 시기별로 달라질 수 있음을 보여줌

<그림 3-6> 주요 소비 관련 토픽의 연도별 비중 변화



주: 1) 각 셀의 값은 해당 연도의 전체 소비기사 수 대비 각 토픽 기사 수의 비율임. 2) 주요 토픽은 초기 군집 기사 수 500건 이상, 최종 토픽 기사 중 초기 군집 비율 70% 이상을 기준으로 선정한 17개 토픽임. 3) 전체 78개 토픽 가운데 주요 토픽만 제시하므로 각 연도의 토픽 비중 합계는 100%가 아님. 4) 2026년은 1월부터 4월까지의 자료임.

자료: 한국언론진흥재단 BIGKinds 경제뉴스 자료(2022. 1. 1.~2026. 4. 30.)를 이용하여 연구진 작성.

□ 특히 2025년 소비기사 비중의 상승은 민생회복 소비쿠폰 관련 토픽의 증가와 밀접하게 연결되어 나타남

- 민생회복 소비쿠폰 토픽의 연도별 비중은 2022년 0.28%, 2023년 0.15%, 2024년 0.33% 수준이었으나 2025년에는 4.58%로 크게 상승함

- 이는 앞서 확인한 2025년 7~9월 소비기사 비중 상승 및 ‘소비쿠폰’, ‘소상공인’, ‘가맹점’, ‘상품권’ 등의 특성추출어 집중과 일관된 결과임
 - 따라서 2025년의 소비기사 확대는 실제 민간소비의 전반적 확대라기보다 특정 소비지원 정책에 대한 집중보도가 상당 부분 반영된 것으로 해석할 수 있음

□ BERTopic 분석은 단순 어휘빈도 분석보다 소비 관련 뉴스의 내용적 맥락을 구분하는 데 유용하지만, 결과 해석에는 일정한 한계가 존재함

- 토픽은 기사 제목의 의미적 유사성을 중심으로 도출되므로 기사 본문의 전체 맥락이나 경제적 영향의 방향까지 직접 반영하지는 못함
- 또한 토픽명은 대표 기사와 주요 특성추출어를 바탕으로 연구진이 부여한 해석적 명칭이므로 토픽 자체를 고정된 경제분류체계로 해석해서는 안 됨
 - 따라서 BERTopic 결과는 소비 관련 뉴스의 주요 내용구조와 시기별 의제 변화를 파악하기 위한 탐색적 분석으로 활용함

IV. 기초 뉴스분석의 시사점 및 한계

1. 주요 시사점

- 전체 경제뉴스에서 소비 관련 보도가 차지하는 비중은 분석기간 동안 대체로 1%대 수준에서 나타났으나, 시기별로 일정한 변동이 확인됨
 - 2022~2024년 소비기사 비중은 1.17~1.24% 수준에서 비교적 안정적으로 나타났으나, 2025년에는 1.62%로 상승함
 - 월별로는 2025년 7~9월 소비기사 비중이 특히 높게 나타나 소비 관련 보도의 집중시기가 존재함을 확인함
 - 이는 소비 관련 뉴스의 규모가 일정하게 유지되는 것이 아니라 정책·사건 및 경제환경 변화에 따라 단기적으로 크게 변동할 수 있음을 보여줌
- 소비기사 비중의 상승은 실제 민간소비의 증가와 동일하게 해석하기 어려움
 - 2025년 소비기사 증가 시기에는 ‘소비쿠폰’, ‘소상공인’, ‘가맹점’, ‘상품권’ 등 특정 소비지원 정책과 관련된 표현이 집중적으로 나타남
 - BERTopic 분석에서도 민생회복 소비쿠폰 토픽의 비중이 2025년에 크게 확대되어 기사량 증가가 특정 정책 이슈와 밀접하게 관련되어 있음을 확인함
 - 따라서 소비기사 수나 비중의 상승은 실제 소비활동의 확대뿐 아니라 특정 정책·사건에 대한 언론보도 집중에 의해 발생할 수 있음
 - 뉴스 기반 소비동향을 해석할 때에는 기사량 자체뿐 아니라 어떤 내용의 보도가 증가하였는지를 함께 확인할 필요가 있음
- 소비 관련 뉴스는 단일한 경제현상보다는 다양한 내용영역으로 구성되어 있음
 - 주요 특성추출어에는 소비정책, 물가, 금융, 유통·결제, 산업 및 시장상황과

관련된 표현이 함께 나타남

- BERTopic 분석에서도 식품·농산물 소비, 금융소비자 보호, 소비자물가, 카드·결제 및 소비행태, 민간소비와 경제성장 등 서로 다른 내용영역이 구분됨
 - 이와 함께 에너지·전력 소비와 같이 민간소비와 직접 관련되지 않은 의미의 ‘소비’도 일부 포함됨
 - 따라서 소비 관련 뉴스의 변화는 하나의 경제적 의미로 단순화하기보다 복수의 의제와 사건이 결합된 결과로 해석할 필요가 있음

□ 단순 기사량과 단어빈도 분석만으로는 소비 관련 뉴스의 경제적 의미와 문맥을 충분히 식별하기 어려움

- 특성추출어 분석은 뉴스에서 반복되는 주요 표현을 확인하는 데 유용하지만, 국가명·기관명·일반 경제용어 등도 높은 빈도로 나타날 수 있음
- 워드클라우드 역시 주요 표현의 분포를 직관적으로 보여줄 수 있으나 단어 간 의미관계나 기사 문맥은 반영하지 못함
 - BERTopic 분석을 통해 기사 제목의 의미적 유사성을 반영함으로써 단순 어휘빈도보다 구체적인 내용구조를 확인할 수 있었으나, 기사 본문의 전체 맥락과 실제 경제적 영향까지 직접 식별하는 데에는 한계가 있음

2. 분석의 한계

□ 본 분석은 기사 제목에 ‘소비’가 포함된 뉴스를 일관된 기준으로 선별하여 소비 관련 보도의 규모와 내용구조를 탐색하였으나, 소비 관련 뉴스 전체를 포괄하는 분류체계는 아님

- 제목에 ‘소비’가 포함되지 않은 소비지출·소매판매 관련 기사는 제외될 수 있으며, 반대로 에너지 소비와 같이 민간소비와 직접 관련되지 않은 의미의 기사가 포함될 수 있음
- 따라서 본 분석의 소비기사 집단은 실제 민간소비 관련 뉴스 전체라기보다

제목에서 ‘소비’가 명시적으로 부각된 기사집단으로 해석할 필요가 있음

- 기사 수와 전체 경제기사 대비 소비기사 비중은 언론보도의 상대적 규모를 나타내는 지표로서 실제 소비활동의 수준이나 방향을 직접 측정하지 못함
 - 특정 정책이나 사건에 대한 집중보도가 발생할 경우 실제 소비동향의 변화와 무관하게 소비기사 수와 비중이 크게 상승할 수 있음
 - 실제로 2025년에는 민생회복 소비쿠폰 관련 보도가 집중되면서 소비기사 비중이 크게 높아진 것으로 나타남
 - 따라서 기사량 자체를 소비경기의 상승·하락이나 소비지출 증감으로 직접 해석하는 것은 적절하지 않음
- 특성추출어와 워드클라우드의 주요 표현의 분포를 파악하는 데 유용하지만 기사 문맥과 단어 간 관계를 충분히 반영하지 못함
 - 동일한 단어라도 기사 내에서 사용되는 경제적 의미와 맥락이 다를 수 있으며, 단순 빈도만으로는 이러한 차이를 구분하기 어려움
 - 상위 특성추출어에 국가명, 기관명, 일반 경제용어 등이 함께 나타나는 점도 단순 어휘기반 분석의 한계를 보여줌
- BERTopic 분석은 기사 제목의 의미적 유사성을 활용하여 단순 어휘빈도보다 구체적인 내용구조를 파악할 수 있으나, 기사 본문의 전체 문맥과 경제적 의미를 직접 식별하는 데에는 한계가 있음
 - 분석은 기사 제목을 중심으로 이루어지므로 동일한 제목 표현이라도 본문에서 제시되는 원인, 방향, 평가 및 경제적 영향까지 충분히 반영하지 못함
 - 토픽명 역시 대표 기사와 주요 특성추출어를 바탕으로 연구진이 부여한 해석적 명칭이므로 고정된 경제분류체계나 객관적 정답으로 해석해서는 안 됨

- 이상의 결과는 단순 기사량이나 어휘빈도만으로 소비동향을 식별하는 데 한계가 있으며, 뉴스의 문맥과 경제적 의미를 보다 정교하게 반영할 필요가 있음을 보여줌
 - 특히 동일한 ‘소비’ 관련 기사라도 정책지원, 물가, 금융소비자 보호, 산업활동, 민간소비 등 서로 다른 경제적 맥락을 포함하므로 기사 내용의 의미를 구분하는 과정이 필요함
 - 이에 제4장에서는 뉴스 기사의 문맥정보를 활용하여 소비 관련 정보를 보다 정교하게 식별하고, 이를 소비예측에 활용할 수 있는 지표 개발을 검토함
 - 제3장의 분석은 이러한 후속 분석의 필요성과 분석대상 뉴스의 기본적인 규모·내용구조를 확인하기 위한 기초분석으로 활용함

제4장 실시간 소비예측지표의 개발 및 검증

I. 사전기반 소비 방향성 분류

- 본 장은 뉴스자료를 활용한 민간소비 인덱스 구축을 위한 선행단계로서, 딥러닝 기반 소비 방향성 모형의 학습자료 설계와 분류체계 정립에 필요한 렉시콘 기반 사전분석 결과를 제시하는 데 목적이 있음
 - 본 연구의 최종목표는 대규모 뉴스자료로부터 내구재, 준내구재, 비내구재 등 주요 소비 품목군별 소비 방향성과 신호를 추출하고, 이를 고빈도 민간소비 지표로 확장하는 데 있음
- 딥러닝 모형이 소비관련 문맥을 효과적으로 학습하기 위해서는 먼저 대규모 뉴스자료에서 소비 품목군별 기사 후보군을 선별하고, 소비 상승, 하락, 보합 등 방향성 신호가 어떤 방식으로 뉴스에 반영되는 지를 확인할 필요가 있음
 - 뉴스 자료에는 소비와 직접 관련된 기사뿐 아니라 가격, 기업실적, 산업동향, 정책발표, 금융시장 등 다양한 문맥이 혼재되어 있음
 - 따라서 소비관련 기사 후보군을 구성하고, 방향성 표현의 기본 구조를 점검하는 선행작업이 필요함
 - 본 장의 렉시콘 기반 분석은 이러한 목적에서 수행한 딥러닝 모형 구축을 위한 선수 작업에 해당됨

- 전체 뉴스기사 집합을 다음과 같이 정의함

$$D = \{d_i\}_{i=1}^N$$

- 여기에서 d_i 는 개별뉴스 기사를 의미
- 본 연구에서는 기사 제목과 본문을 결합하여 각 뉴스 i 에서 분석에 활용되는 텍스트 x_i 를 다음과 같이 구성함

$$x_i = title_i \oplus content_i$$

- 여기에서 $title_i$ 은 기사제목, $content_i$ 기사본문을 의미하며, $\oplus$ 는 두 텍스트를 결합하는 연산을 의미함
- 소비방향성 신호는 기사 전체에서 단어가 단순히 함께 등장하는지만으로 판별하는 것이 아닌, 뉴스기사 내부의 문장 또는 근접 문맥 단위에서 품목군 표현, 소비활동 표현, 방향성 표현이 함께 나타나는지를 기준으로 식별함
- 따라서 뉴스 기사 i 의 분석 텍스트를 다음과 같은 하위 분석단위로 분해할 수 있음

$$x_i = \{s_{i1}, s_{i2}, \dots, s_{iH_i}\}$$

- 여기서 s_{ih} 는 뉴스기사 i 에 포함된 h 번째 문장 또는 근접 문맥 단위이며, H_i 는 뉴스기사 i 에서 추출된 분석 단위의 수를 의미함
- 실제 구현에서는 제목과 본문 도입 문단을 중심으로 분석하였으며, 소비활동 표현과 방향성 표현이 일정 거리 이내에서 결합되어 나타나는지를 기준으로 방향성 신호를 판별
- 다만 본 연구의 최종관심은 각 뉴스 텍스트 x_i 가 특정 소비품목군 c 와 관련되는지, 그리고 관련되는 경우 소비 방향성이 무엇인지를 판별하는 데 있음
- 소비 품목군 집합은 다음과 같이 정의할 수 있음

$$c \in C = \{durable, quasi-durable, nondurable\}$$

- 각 품목군 c 에 대응하는 품목사전은 다음과 같이 M_c 로 정의함

$$M_c = \{m_{c,1}, m_{c,2}, \dots, m_{c,J_c}\}$$

- 여기에서 $m_{c,j}$ 는 품목군 c 에 속하는 j 번째 품목 표현을 의미하며, J_c 는 해당 품목군 사전에 포함된 전체 표현의 수를 나타냄
- 품목군별 소비 특성과 뉴스 보도 문맥이 상이하므로, J_c 의 크기와 M_c 의 구성은 품목군별로 달라질 수 있음
- 예를 들어, 비내구재의 경우 식료품, 생활용품, 화장품 등 관련표현이 이에 해당

- 문장 또는 근접 문맥 단위 s_{ih} 가 품목군 c 와 관련되는지 여부는 다음의 지시 함수로 정의함

$$Q_{ih,c} = I\left(\sum_{j=1}^{J_c} 1(m_{c,j} \in s_{ih}) > 0\right)$$

- 즉, 분석단위 s_{ih} 안에 품목군 c 의 품목 사전에 포함된 표현이 하나 이상 등장하면 $Q_{ih,c} = 1$ 그렇지 않으면 $Q_{ih,c} = 0$ 으로 정의
- 뉴스 기사단위의 품목군 관련 여부는 기사 내부의 하위 분석 단위 중 하나이상에서 품목군 표현이 포착되는 경우로 함

$$Q_{i,c} = 1\left(\sum_{h=1}^{H_i} Q_{ih,c} > 0\right)$$

- 다음으로 소비 방향성 판별을 위해 소비활동 사전, 방향성 사전, 가격 문맥 사전을 정의
 - 소비활동 사전을 A , 상승방향 사전 U , 하락방향 사전 D^- , 포함방향 사전을 H , 가격 문맥 사전을 P 로 설정
 - $A = \{a_1, a_2, \dots, a_{J_A}\}$, $U = \{u_1, u_2, \dots, u_{J_U}\}$, $D^- = \{d_1, d_2, \dots, d_{J_D}\}$,
 $H = \{h_1, h_2, \dots, h_{J_H}\}$, $P = \{p_1, p_2, \dots, p_{J_P}\}$
 - J_A, J_U, J_D, J_H, J_P 는 각각 해당 사전에 포함된 표현의 수를 의미함
- 품목군 c 에 대한 전체 사전 기반 분류체계는 다음과 같이 표현할 수 있음

$$\Lambda_c = \{M_c, A, U, D^-, H, P\}$$

- 여기에서 Λ_c 는 품목군별 품목사전과 소비활동, 방향성, 가격문맥 사전을 결합한 규칙기반 분류체계임
- 소비방향성은 단순히 품목 사전과 소비활동, 방향성 단어가 동시에 등장하는지만을 판별하지 않고, 품목군 관련성, 소비활동 표현, 방향성 표현이 함께 충족되는지를 기준으로 함
- 본 연구는 소비활동 표현과 방향성 표현이 동일한 분석 단위 또는 제한된 근접 문맥 내에서 결합되어 나타나는 경우에만 소비방향성 신호로 간주함

- 이를 위해 뉴스 기사 x_i 안에서 표현 r 과 s 사이의 거리를 $dist(r, s|x_i)$ 로 정의하고, 두 표현이 하나의 의미 단위로 결합되어 있다고 판단할 수 있는 허용거리 δ 로 둬

- 품목군 c 의 소비 상승 신호가 포착되는지 여부는 다음과 같이 정의함

$$Z_{i,c}^{up} = Q_{ic} \cdot 1[\exists a \in A, \exists u \in U \text{ s.t. } dist(a, u|x_i) \leq \delta]$$

- 하락 및 보합 신호도 같은 방식으로 정의함

$$Z_{i,c}^{down} = Q_{ic} \cdot 1[\exists a \in A, \exists d \in D^- \text{ s.t. } dist(a, d|x_i) \leq \delta]$$

$$Z_{i,c}^{flat} = Q_{ic} \cdot 1[\exists a \in A, \exists h \in H \text{ s.t. } dist(a, h|x_i) \leq \delta]$$

- 여기에서 $Z_{i,c}^{up}, Z_{i,c}^{down}, Z_{i,c}^{flat}, Z_{i,c}^{price}$ 는 각각 기사 i 가 품목군 c 에 대해 소비상승, 소비하락, 소비보합, 가격 문맥 신호를 포함하는지를 나타내는 인디케이터임
- 예를 들어, “편의점 도시락 판매가 늘었다.” 라는 표현은 품목군 관련 표현, 소비활동 표현인 “판매”, 상승 방향 표현인 “늘었다”가 가까운 문맥에서 결합되어 있으므로 상승신호로 분류됨
- 반면 “식품업계는 신제품을 출시했다. 주가는 상승했다” 와 같이 품목군 표현과 상승 표현이 함께 등장하더라도 소비활동 표현과 의미상 직접 연결되지 않은 경우에는 소비상승 신호로 간주하지 않음
- 가격 문맥은 다음과 같이 별도로 정의함

$$Z_{i,c}^{price} = Q_{i,c} I\left(\sum_{s=1}^{J_p} 1(p_s \in x_i) > 0\right)$$

- 이때 가격 문맥은 소비 방향성과 동일하게 해석하지 않고, 별도로 구분하기로 함

□ 앞서 정의한 신호를 바탕으로 뉴스 기사 i 의 품목군 c 에 대한 사전기반 분류라벨을 분류함

- 여기서 사전기반 분류라벨은 향후 딥러닝 기반 문맥 분류모형의 학습자료를 구성하기 위한 기초라벨로 활용되며, 최종 학습자료 구축단계에서는 표본 검수와 라벨 정제 과정을 거쳐 보완될 수 있음

$$\tilde{y}_{i,c}^{(L)} = g_c(x_i; \Lambda_c)$$

- $\tilde{y}_{i,c}^{(L)}$ 는 뉴스기사 i 에 대해 품목군 c 기준으로 부여된 사전기반 1차 분류라벨이며 L 은 해당 라벨이 렉시콘 기반으로 산출되었음을 의미, $g_c(\cdot)$ 는 품목군별 사전체계 Λ_c 를 이용한 규칙기반 분류함수임
- 라벨은 소비상승, 하락, 소비보합, 혼재, 가격문맥, 무관 범주로 구분하도록 설정
- 품목군 관련 표현이 확인된 기사 중 소비활동 표현과 상승 방향 표현이 근접 문맥에서 결합되어 나타나는 경우에는 소비 상승 기사로 분류
 - 반대로 소비활동 표현과 하락방향이 결합되어 나타나는 경우에는 소비하락 기사로 분류하며, 소비활동 표현과 보합 또는 유지 관련 표현이 결합되는 경우에는 소비 보합 기사로 분류
 - 다만, 하나의 기사 안에서 상승신호와 하락신호가 동시에 포착되는 경우, 해당 기사를 특정 방향으로 단정하지 않고 혼재신호로 처리함
 - 예를 들어 한 뉴스 기사에서 “편의점 도시락 판매가 늘었다.” 라는 표현과 “화장품 판매는 둔화되었다” 라는 표현이 함께 등장하는 경우, 해당 기사는 비내구재 소비의 상승과 하락 신호를 모두 포함하므로 단일한 상승 또는 하락기사로 분류하지 않음
- 가격, 물가, 비용 관련 표현이 포함되어 있으나, 소비활동 증가·감소·보합 방향성이 명확하지 않은 경우에는 가격 문맥으로 구분함
 - 예컨대 “휘발유 가격이 하락했다” 는 해당 품목과 하락표현을 포함하지만, 하락한 대상은 소비나 판매가 아니라 가격임. 따라서 이러한 기사는 소비하락 기사로 직접 해석하지 않고 가격문맥으로 별도 처리하기로 함
- 품목군 관련 표현이 포함되지 않거나, 품목군 표현이 포함되어 있더라도 소비활동의 방향성을 판단할 수 있는 근접 문맥이 확인되지 않는 경우에는 미분류 범주로 처리
- 이와 같은 사전기반 1차 분류라벨은 최종 정답 라벨이라기보다는 대규모 뉴

스자료에서 소비방향성 신호를 체계적으로 선별하고 딥러닝 기반 문맥 분류 모형의 학습자료를 구성하기 위한 기초 분류값으로 활용

1. 비내구재

□ 본 분석에서는 비내구재 소비 흐름을 고빈도로 파악하기 위해 뉴스자료를 활용한 예비지표를 구축함

- 분석대상은 중복 제거가 완료된 2022.1월부터 2026.4월까지 뉴스자료로서, 이 중 경제 관련 영역에서 비내구재 소비와 관련성이 높은 세부 분류를 우선 선별함
- 이후 식료품, 가공식품, 신선식품, 음료, 생필품, 화장품, 의약품, 석유류, 대형마트·편의점 등 비내구재 관련 키워드를 적용하여 최종적으로 836,397건의 비내구재 관련 기사를 추출
- 추출된 비내구재 기사에 대해서는 소비 관련 표현을 중심으로 상승, 하락, 보합 및 혼재 방향성을 식별함
 - 본 단계에서는 딥러닝 기반 분류모형을 적용하기에 앞서, 규칙분류 방식을 활용함
 - 이는 최종적인 딥러닝 분류결과라기 보다는 향후 딥러닝 학습을 위한 1차 라벨 생성 및 오류 검토 자료로 활용하기 위한 절차임
- 규칙기반 분류결과, 뉴스 기사 제목과 본문을 포함할 경우 소비 상승 116,293건, 소비하락 13,479건, 혼재 20,058건, 보합 3,535건으로 식별됨

<표 4-1> 비내구재 뉴스의 소비 방향성 분포

(단위: 기사수, %)

소비 방향성	제목 기준	본문 기준	뉴스 기사전체 기준	비중
소비 상승	8,304	112,250	116,293	13.90
소비 하락	1,348	13,125	13,479	1.61
보합	75	3,392	3,535	0.42
전체	836,397			

주: 비중은 전체 비내구재 기사 대비 계산함. 뉴스 기사 본문 기준은 초반 400자를 대상으로 수행함. 이는 뉴스 기사의 핵심 내용이 제목 및 리드문 이후 초반부에 집중되는 특성을 고려하여 설정한 분석단위로, 기사 후반부의 배경 설명, 인용문, 부가 정보로 인한 분류 노이즈를 완화하기 위한 기준임

□ 본 장에서는 비내구재 관련 뉴스에서 소비흐름을 고빈도로 포착하기 위해 주간단위의 소비 방향성 지표를 구성

- 뉴스 텍스트는 공식 통계보다 빠르게 축적된다는 장점이 있으나, 특정기업이 슈, 계절적 이벤트, 가격인상 보도, 프로모션 기사 등에 따라 단기 변동성이 크게 나타날 수 있음
- 따라서 본 분석에서는 기사제목 및 본문에서 식별된 소비 상승 및 하락 방향성을 주간단위로 집계하고, 이를 바탕으로 소비여건의 변화 방향을 나타내는 후보지표를 산출
 - 단기 잡음을 완화한 평활화 지표도 병행 검토
 - 소비 방향성의 기초적 흐름을 파악하기 위해 이동평균과 지수가중이동평균을 함께 산출
- 주간 t 에서 뉴스기반 소비방향성 원지표를 R_t 라 하면, k 주 이동평균은 다음과 같이 정의됨

$$MA_k(t) = \frac{1}{k} \sum_{j=0}^{k-1} R_{t-j}$$

- $MA_k(t)$ 는 최근 k 주 동안의 뉴스지표를 동일한 가중치로 평균한 값
- 단순이동평균은 일시적인 기사 급증이나 특정 주의 이상치를 완화하는 데 유용하지만, 최근 변화와 과거 변화에 동일한 가중치를 부여하기 때문에 소비 흐름의 전환이 다소 늦게 반영될 수 있음

- 이를 보완하기 위해 4주 기준의 지수가중 이동평균도 함께 활용하기로 함

$$EWM_4(t) = \alpha R_t + (1 - \alpha)EWM_4(t-1)$$

- 여기에서 $EWM_4(t)$ 는 t 시점의 4주 기준 지수가중 평활화 지표, R_t 는 해당 주의 뉴스 기반 소비방향성 원지표, α 는 최근 관측치에 부여되는 가중치를 의미함

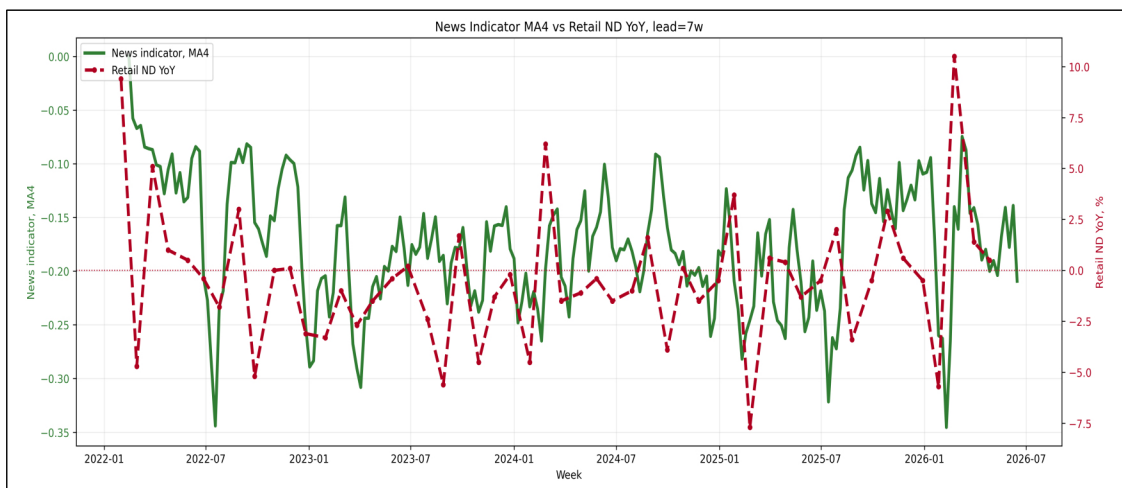
□ 본 분석에서는 뉴스기반 지표가 실제 비내구재 소매판매 전년동월대비 증가율보다 선행적으로 움직일 가능성을 고려, 후보 뉴스지표별 시차관계를 검토하였으며, 분석 결과 후보 지표 중 7주 선행 기준에서 실제 비내구재 소매판매 전년동월대비 증가율과의 관련성이 상대적으로 높은 것으로 나타남

□ 하단의 <그림 4-1>은 뉴스기반 비내구재 소비방향성 지표가 실제 비내구재 소매판매 흐름과 어느정도 유사한 움직임을 보이는지 확인하기 위해 소매판매 지표와의 패턴 비교를 수행한 것임

- 비내구재 관련 뉴스에서 소비여건의 변화를 포착하기 위해 주간단위로 산출하였으며, 단기 변동성을 완화하기 위해 4주 단순 이동평균과 4주 지수가중 이동평균을 함께 적용
- 하단의 <그림 4-1>은 4주 단순 이동평균을 적용한 뉴스기반 비내구재 소비방향성 지표와 실제 비내구재 소매판매 전년동월대비 증가율을 비교한 것임
 - 왼쪽 축은 뉴스기반 지표, 오른쪽 축은 비내구재 소매판매 전년동월대비 증가율을 의미함
 - 두 지표는 산출 주기와 단위가 서로 다르므로 절대적인 수준을 직접 비교하기보다는, 주요 국면에서 상승 및 하락 방향이 유사하게 나타나는지를 확인하는 데 초점을 둠
 - 뉴스지표와 실제 소매판매 증가율이 모든 시점에서 일치하지 않지만, 일부 구간에서는 뉴스지표의 변화가 이후 비내구재 소매판매 증가율의 둔화 또는 회복 흐름과 유사하게 나타나는 것으로 확인됨

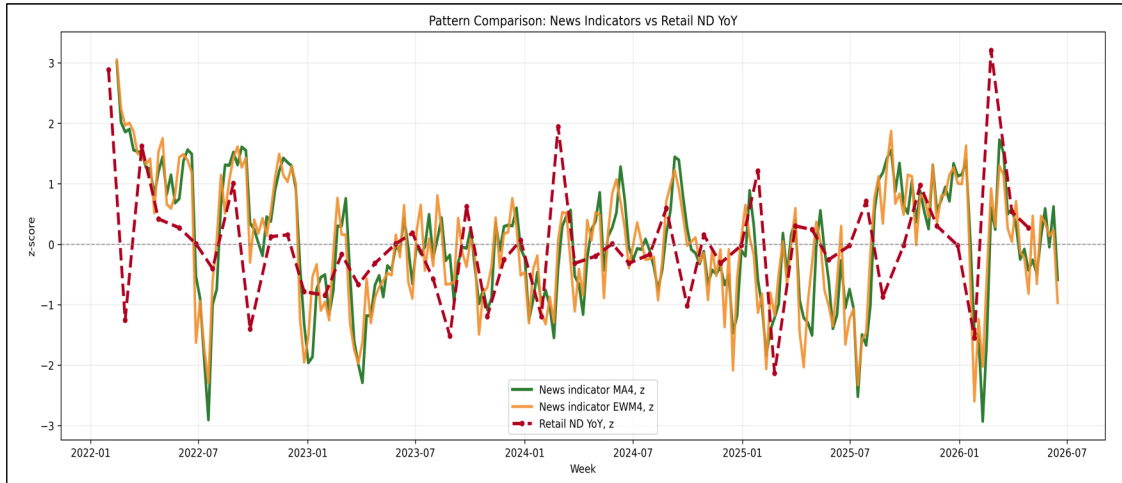
- 이는 주간 뉴스자료가 공식 월간 통계보다 빠르게 소비 관련 신호를 반영할 수 있음을 보여줌
- 특히, 뉴스기반 지표는 공식 통계발표 이전에 관측가능하다는 점에서, 비내구재 소비 흐름을 조기에 점검하기 위한 보조지표로 활용될 여지가 있음
- <그림 4-2>는 뉴스기반 비내구재 소비 방향성과 실제 비내구재 소매판매 전년동월대비 증가율을 모두 표준화하여 비교한 것임
 - <그림 4-1>에서는 뉴스지표와 공식통계가 서로 다른 단위와 축으로 표시되기 때문에, 두 지표의 변동폭이나 국면 전환을 동일한 기준에서 비교하는 데 한계가 있음
 - 7주 선행 정렬을 적용한 뉴스기반 지표 중 4주 단순 이동평균과 4주 EWM(지수가중이동평균)을 함께 제시함
 - <그림 4-2>에 따르면 두 지표가 대체로 유사한 방향으로 움직이는 것으로 나타났으며, 뉴스 기반 지표 흐름이 특정 평활화 방식에만 의존하지 않고 비교적 안정적으로 나타남을 보여줌

<그림 4-1> 뉴스기반 비내구재 소비 방향성 지표와 소매판매 비내구재 YOY 증가율



주: 뉴스지표는 비내구재 관련 뉴스자료를 활용하여 산출한 주간 소비방향성 보조지표, MA4는 단순이동 평균을 의미함. 본 그림은 다양한 선행시차를 비교한 결과를 바탕으로, 7주 선행정렬 기준에서 뉴스 지표와 비내구재 소매판매 전년동월대비 증가율을 비교한 것임

<그림 4-2> 뉴스기반 비내구재 소비방향성 지표와 소매판매 증가율 표준화 비교



주: MA4는 4주 단순이동평균, EWM4는 4주 기준 지수가중이동평균을 의미함. 모든 뉴스지표는 7주 선행정렬 후 비교. z-score는 단위가 다른 뉴스지표와 공식통계의 패턴 비교를 위한 것임.
자료: 국가데이터처, 저자 작성

□ 비내구재 뉴스기반 소비방향성 지표는 일부 구간에서 실제 비내구재 소매판매 증가율의 둔화 및 회복 흐름과 유사한 움직임을 보였으나, 전 기간에 걸쳐 안정적으로 일치한다고 보기는 어려움

- 이는 규칙기반 분류가 상승 및 하락 표현을 명시적으로 식별할 수 있다는 장점에도 불구하고, 기사 문맥에 따라 동일한 표현이 다른 의미를 가질 수 있다는 한계를 갖기 때문임

- 예를 들어 판매증가, 할인확대, 가격인상, 수요 둔화 등의 표현은 기사 맥락에 따라 소비개선 신호일 수 있으며, 비용부담이나 가격효과에 따른 신호일 수 있음

- 따라서 현재의 지표는 소비 방향성의 1차 신호를 포착하는 데 유용하지만, 소비 흐름과 직접 관련된 문맥과 단순 이벤트성, 홍보성 등의 보도를 구분하는 데에는 추가적인 보완이 필요함

□ 본 연구는 규칙 기반 분석결과를 1차 분류단계로 활용하고, 이를 바탕으로 KoBERT 기반 딥러닝 문맥분류모형으로 확장하는 방향을 고려함

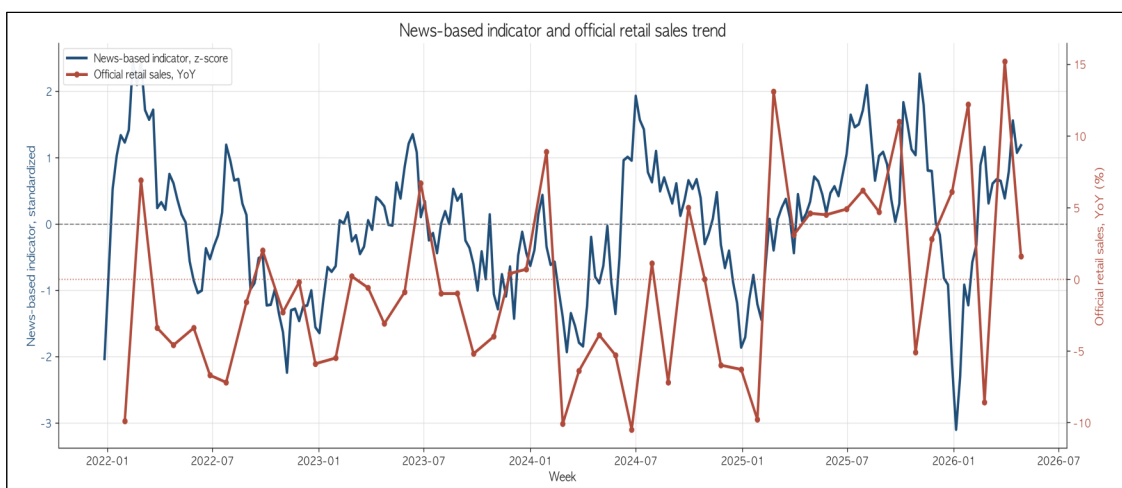
- 단순 키워드 출현 여부를 넘어 문장 내 소비 방향성의 의미를 보다 정교하게 판별할 수 있는지 검토하고, 이를 위해 본 연구의 1차 결과를 KoBERT 학습자료 구축의 기초 입력값으로 활용

2. 내구재

□ 하단의 그림(<그림 4-3>~<그림 4-5>)은 내구재 부문에 대해서 규칙기반 방식으로 산출한 뉴스기반 지표와 내구재 소매판매 전년동월대비 증가율 간의 관계를 나타낸 것으로, 이 그림들에서 보이는 바와 같이 뉴스기반 지표는 내구재 소매판매 지표(YOY)와 개선 및 둔화 흐름과 유사한 방향성을 보이는 것으로 파악됨

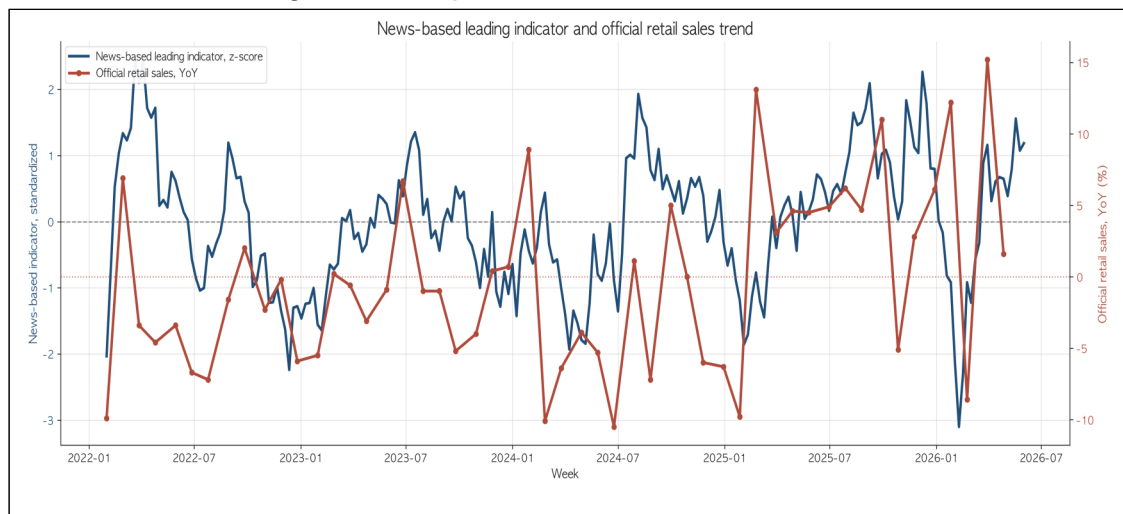
- 뉴스기반 지표는 주간단위로 산출되기 때문에 월간 공식 통계보다 빠르게 소비 관련 정보를 반영할 수 있으며, 선행시차를 조정한 이후에는 공식 통계의 변동 방향을 일정 부분 앞서 포착하는 모습이 확인됨
- 이는 규칙기반으로 구축한 내구재 뉴스기반 지표가 내구재 소비여건의 변화를 사전에 확인하기 위한 보조적 선행지표로서 활용가능성이 있음을 시사

<그림 4-3> 뉴스기반 내구재 지표와 내구재 소매판매 YOY 추이



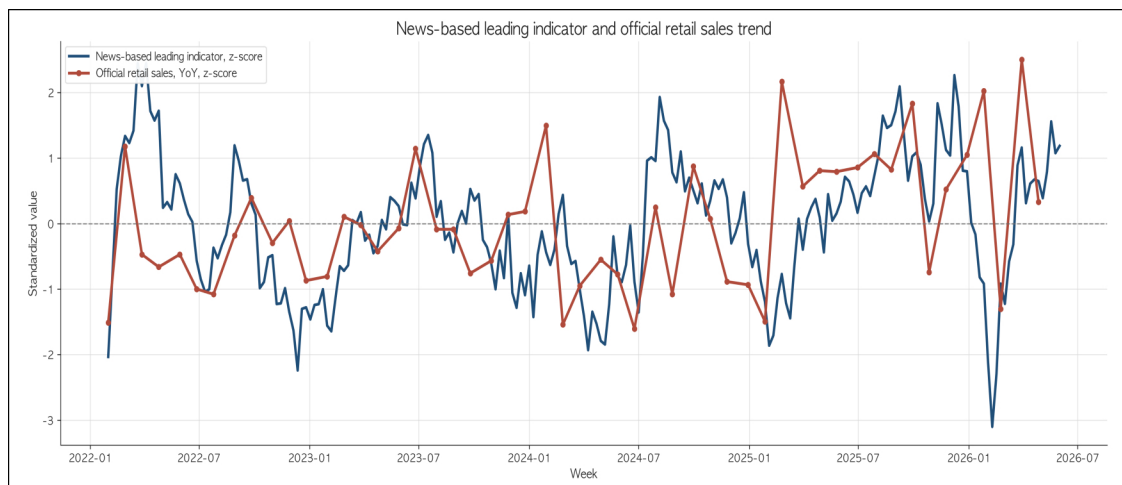
자료: 국가데이터처, 저자 작성

<그림 4-4> 선행시차(5주) 반영한 뉴스기반 내구재 지표와 내구재 소매판매 YOY



자료: 국가데이터처, 저자 작성

<그림 4-5> 선행시차(5주) 반영후 뉴스기반 지표와 내구재 소매판매 YOY 표준화 비교



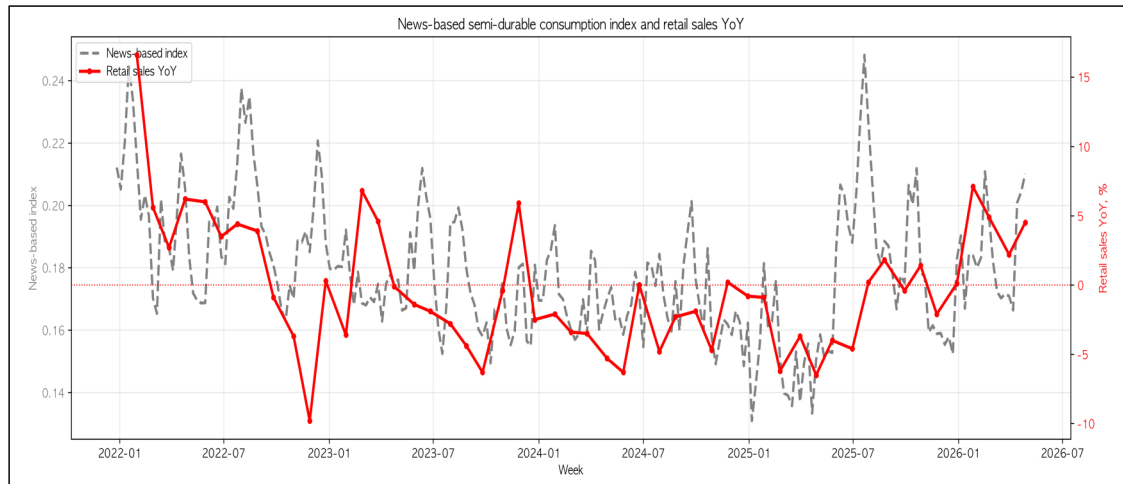
자료: 국가데이터처, 저자 작성

3. 준내구재

- 하단의 그림은 준내구재 부문에서 규칙기반 방식으로 산출한 뉴스 지표와 준내구재 소매판매액 지수 전년동월대비 증가율 간의 관계를 나타낸 것임

- 분석 결과, 주간단위로 산출된 뉴스기반 준내구재 지표는 월간 공식 통계에 비해 단기 변동성이 크게 나타났지만, 전반적인 흐름에서는 준내구재 소매판매 (전년동월대비) 증가율 국면변화를 일정 부분 반영하고 있는 것으로 확인됨
 - 2022년에는 준내구재 소매판매 증가율이 비교적 높은 수준을 유지하였고, 뉴스기반 지표 역시 상대적으로 높은 구간에서 움직인 것으로 나타남
 - 이후 2023년부터 2024년까지 공식통계의 증가율이 둔화되거나 마이너스권에 머무는 구간에서는 뉴스기반 지표 역시 대체로 낮아지는 모습을 보임
 - 2025년 이후에는 두 지표 모두 등락이 확대되는 양상을 보임
 - 뉴스기반 지표는 일부 시점에서 공식통계보다 빠르게 반응하거나 단기적으로 크게 움직이는 구간이 나타난 것으로 확인됨
 - 이는 준내구재 소비가 계절적 수요, 할인행사, 의류 및 신발 등 품목별 이슈, 소비심리 변화에 민감하게 반응하기 때문인 것으로 해석됨

<그림 4-6> 뉴스기반 준내구재 지표와 소매판매 증가율 추이



자료: 국가데이터처, 저자 작성

- 본 지표는 기사 본문 내 표현을 기반으로 산출된 예비적 지표로, 본문 문맥에서 실제 소비 및 판매 방향성을 정교하게 판별할 수 있도록 딥러닝 기반 문맥분류 모형을 적용, 계절적 요인과 이벤트성 기사에 따른 변동성을 추가적으로 조정할 필요가 있음

II. 딥러닝 기반 문맥분류 모형 활용

- 본 연구는 앞서 소비관련 품목어, 가격·수요·판매 관련 서술어 등을 활용한 렉시콘 기반 규칙분류를 통해 뉴스기사의 소비 관련 여부와 방향성을 일차적으로 식별함
 - 이 과정에서 기사 제목 및 본문 일부에 나타나는 품목군별 키워드와 소비 방향성 표현을 결합하여 예비분류하고, 이를 주간 단위의 고빈도 뉴스지표로 집계함
 - 이러한 규칙기반 접근은 분류 기준이 명시적이고 결과해석이 용이하다는 장점이 있어, 대규모 자료에서 소비관련 표본을 안정적으로 선별하기 위한 필터링 단계로 활용가능
- 다만, 렉시콘 기반 분류는 사전에 정의한 키워드와 방향성에 의존하므로, 문장 내 의미관계가 복합적으로 나타나는 경우에는 분류의 정밀도가 제한될 수 있음
 - 예를 들어 동일한 소비 품목이 포함된 기사라도 판매량, 가격, 수요, 재고 등 서로 다른 경제적 맥락이 함께 서술될수 있기 때문에, 하나의 기사 안에서도 긍정 신호와 부정 소비 신호가 동시에 나타날 수 있음
 - 또한, 규칙분류 방법은 연구자가 사전에 정의한 단어와 표현의 조합에 의존하기 때문에 문장 내 의미 관계가 복합적으로 나타나는 경우 분류의 정밀도가 제한될 수 있음
- 본 장에서는 딥러닝 기반 문맥 분류모형 (BERT 계열)을 활용하여 소비 방향성을 보다 정교하게 판별하고자 함
- BERT는 Bidirectional Encoder Representations from Transformers로, Transformer

의 Encoder 구조를 기반으로 하는 사전학습 언어모형임

- Transformer는 문장안의 단어들을 순서대로 처리하기 보다, 문장 전체에서 단어들 사이의 관계를 동시에 고려하는 신경망 구조임.
 - 즉, 입력된 토큰열을 문맥적 벡터 표현으로 변환하기 위한 신경망 구조
 - Vaswani et al. (2017)은 기존의 순환신경망이나 합성곱 구조에 의존하지 않고, attention mechanism 만으로 입력과 출력 간의 관계를 학습하는 Transformer 구조를 제안¹⁷⁾
- Transformer 는 기본적으로 Encoder와 decoder 구조로 구성되는데 , 이 중 encoder 는 입력 토큰별로 연속적인 벡터표현의 열로 변환하고, decoder 는 encoder가 산출한 표현을 바탕으로 출력 토큰열을 순차적으로 생성함
- 본 연구에서 중요한 것은 Transformer 의 전체 구조 중 encoder의 역할임
 - Encoder는 입력문장 내부의 단어들이 서로 어떤 관계를 갖는지를 반영하여 문장 또는 토큰의 문맥적 표현을 산출하는 장치임
- Vaswani et al. (2017)에 따르면 Encoder는 6개의 동일한 층으로 구성되며, 각 층은 두 개의 하위구조를 포함
- 첫 번째는 multi-head self-attention이고, 두 번째는 위치별 feed-forward network임
- multi-head self-attention 은 문장 안의 단어들이 서로 어느정도 관련되어 있는지를 여러관점에서 계산하는 부분이고, feed-forward network 는 이렇게 계산된 정보를 변환하여 다음 층으로 전달하는 부분임
 - 각 하위 구조에 residual connection과 layer normalization을 적용
 - 이는 Encoder가 입력 토큰을 한번에 단순 변환하는 것이 아니라, 여러층을 거치면서 토큰 간 관계의 위치별 비선형 변환을 반복적으로 반영하는 구조임을 의미
- self-attention 은 하나의 입력열 내부에 있는 서로 다른 위치의 토큰 간 관계를 계산하여 sequence 내부에서 서로 다른 위치들을 관련시켜 해당

17) We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

sequence의 표현을 계산하는 attention mechanism임

- 즉, 문장 안의 각 단어가 다른 단어들을 얼마나 고려해야 하는지를 계산하는 방식
- 예를 들어, 어떤 문장에서 “증가”라는 단어가 등장하더라도, 그 단어가 “판매”와 연결되는지, “가격”과 연결되는지에 따라 의미가 달라짐

□ Transformer 구조 상 Attention은 특정 위치의 표현을 계산할 때 다른 위치의 정보를 어느정도 반영할 지를 결정하는 연산으로, attention 함수는 query와 key-value 쌍을 입력으로 받아 output을 산출하는 식으로 정의함

- output 은 value들의 가중합으로 계산되며, 각 value에 부여되는 가중치는 query와 해당 key 사이의 적합도에 의해 결정됨
 - 이를 통해 attention은 입력된 정보들 가운데 현재 위치의 표현을 구성함에 있어 필요한 정보를 선택적으로 반영하는 가중결합 방식으로 이해할 수 있음

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$$

- 여기에서 Q 는 query 행렬, K 는 key 행렬, V 는 value 행렬, d_k 는 key 벡터의 차원임
- query 와 모든 key 간 내적을 계산하고, 이를 $\sqrt{d_k}$ 로 나눈 후, softmax 적용하여 value 에 대한 가중치를 얻음
- 이를 하나의 토큰 단위로 풀어쓰는 경우, 문장안의 i 번째 토큰을 기준으로 할 때, i 번째 토큰의 query 벡터를 q_i , j 번째 토큰의 key 벡터를 k_j , j 번째 토큰의 value 벡터를 v_j 라고 할 때, i 번째 토큰과 j 번째 토큰 사이의 관련성 점수는 다음과 같이 계산됨 $\text{score}_{i,j} = q_i k_j^T / \sqrt{d_k}$
- 이 점수는 softmax 함수를 통해 attention weight 로 변환됨

$$a_{i,j} = \frac{\exp(\text{score}_{i,j})}{\exp(\text{score}_{i,1}) + \exp(\text{score}_{i,2}) + \dots + \exp(\text{score}_{i,n})}$$

- 여기서 $\alpha_{i,j}$ 는 i 번째 토큰이 j 번째 토큰의 정보를 어느 정도 반영하는지를 나타내는 가중치를 의미
- $\alpha_{i,j}$ 가 클수록 i 번째 토큰의 문맥 표현을 만들 때 j 번째 토큰의 정보가 더 많이 반영되며, $\alpha_{i,j}$ 가 작을수록 j 번째 토큰의 정보는 적게 반영됨
- i 번째 토큰이 문장 내 여러 토큰의 정보를 $\alpha_{i,j}$ 에 따라 결합한 결과를 벡터 o_i 로 나타낼 수 있는데, o_i 는 문장 내 각 토큰의 value 벡터를 $\alpha_{i,j}$ 에 따라 결합하여 다음과 같이 산출됨

$$o_i = \sum_{j=1}^n \alpha_{i,j} v_j$$

- 여기서 o_i 는 i 번째 토큰의 문맥이 반영된 새 표현으로, 문장안의 1번째 토큰부터 n 번째 토큰까지의 value 정보와 $\alpha_{i,j}$ 가중치에 따라 결합된 결과임
- 예를들어, “가격이 상승했다” 라는 문장에서 i 번째 토큰을 “상승”이라고 가정함. 이 경우 모형이 “상승”을 해석시 “가격”을 중요하게 보아야 한다고 판단한다면, “가격”에 해당하는 $\alpha_{i,j}$ 가 크게 산출됨. 따라서 o_i 에는 “가격”의 value 정보가 크게 반영됨. 이때 o_i 는 “상승”이 “가격과 연결된 상승”의 문맥을 갖는 표현이 됨

□ Transformer 계열 모형은 문장을 문자 그대로가 아닌, 문장을 구성하는 각 토큰을 숫자 벡터로 변환하여 처리함.

- 문장에 n 개의 토큰이 있다고 할 때, i 번째 토큰이 특정 self-attention 층에 입력되기 직전에 갖고 있는 벡터 표현을 h_i 라고 가정함.
 - 여기에서 h_i 는 i 번째 단어 자체가 아닌, 해당 토큰이 모형 내부에서 계산 가능한 벡터로 표현된 값임
 - 문장 전체의 벡터표현은 각 토큰 표현을 순서대로 묶어 $H = (h_1, h_2, \dots, h_n)$ 으로 나타냄
- Transformer encoder 는 입력된 토큰열을 계산 가능한 벡터표현의 열로 바꾸는 장치로서, 동일한 encoder layer를 여러개 쌓은 구조로 되어 있음

- Vaswani et al. (2017)에 따르면 Transformer 모형에서 encoder가 6개의 동일한 층으로 구성되며, 각 층은 multi-head self attention 등을 포함
- h_i 는 특정 encoder layer의 self-attention에 입력되는 i 번째 토큰의 표현으로 볼 수 있으며, 이 h_i 로부터 query, key, value를 각각 산출. 즉, i 번째 토큰에 대해 다음과 같이 쓸 수 있음

$$q_i = h_i WQ$$

$$k_i = h_i WK$$

$$v_i = h_i WV$$

- 여기에서 WQ, WK, WV는 학습과정에서 추정되는 가중치 행렬임
 - 동일한 토큰 표현 h_i 에서 q_i, k_i, v_i 를 따로 만드는 이유는 attention 계산에서 세 벡터의 역할이 상이하기 때문임
 - q_i 는 i 번째 토큰이 어떤 정보를 필요로 하는지를 나타내는 표현, k_i 는 해당 토큰이 다른 토큰의 query와 얼마나 관련되는지를 비교하기 위한 표현, v_i 는 attention 가중치에 따라 실제로 결합되는 정보임
- 문장전체 기준으로, 개별 토큰의 표현을 묶은 H 로부터 모든 토큰의 query, key, value를 한꺼번에 다음과 같이 산출됨

$$Q = HWQ$$

$$K = HWK$$

$$V = HWV$$

- 즉, Q 는 문장 내 모든 토큰의 query 벡터, K 는 모든 토큰의 key 벡터, V 는 모든 토큰의 value 벡터를 모은 행렬을 의미함
- Transformer encoder에서 이러한 과정은 여러 encoder layer를 거치며 반복됨

- 각 층의 self-attention은 이전 층에서 전달된 토큰 표현을 입력으로 받아 새로운 문맥표현을 산출하고, 이후 feed-forward network 등을 거쳐 다음 층으로 전달
 - 이러한 구조를 통해 encoder는 입력 문장의 각 토큰을 단순한 개별단어가 아닌, 문장 내 다른 토큰들과의 관계가 반영된 문맥적 표현으로 변환됨
- 본 연구에서 활용하는 BERT 계열 모형은 Transformer 의 encoder 구조를 기반으로 하는 문맥표현 모형으로, BERT는 기계번역과 같이 새로운 문장을 생성하는 과제보다 입력문장 또는 문서가 어떤 의미 범주에 속하는지를 판별하는 자연어 이해과제에 적합한 구조를 가짐
- 본 연구의 목적도 뉴스기사를 새로이 생성하는 것이 아닌, 이미 수집된 기사나 문장이 소비확대, 소비둔화, 혼재 등 어느 방향의 신호를 갖는지를 분류하는 것임
 - 따라서 decoder 기반 생성 구조보다, 입력 문장의 문맥 표현을 산출하고 이를 분류층에 연결할 수 있는 encoder기반 BERT 구조가 연구목적에 부합
- 본 연구에서의 분석대상은 한국어로 되어 있는 뉴스기사이므로, 한국어 문장 구조와 표현 특성을 반영할 수 있어야 하므로, 한국어 BERT 사전학습 모형 중 하나인 KoBERT를 활용함
- KoBERT는 SK Telecom T-Brain 이 공개한 한국어에 특화된 BERT 사전학습 모형으로, 한국어 처리에서 BERT가 갖는 한계를 보완하기 위해 개발된 모형임
 - 해당 모형의 학습자료 규모는 약 500만 문장, 5400만 단어로 제시되어 있으며, KoBERT의 vocabulary 크기는 8002개이며, 한국어 위키 기반으로 학습한 SentencePiece 토큰라이저를 사용

1. 비내구재

□ 비내구재 부문에서는 기존 규칙기반 소비 분류 결과를 바탕으로 KoBERT 기반 문맥분류모형을 구축함

- 규칙기반 방식은 소비관련 키워드 표현을 활용해 기사를 신속하게 분류할 수 있다는 장점이 있으나, 표현의 맥락과 기사간 이질성을 충분히 반영하기에는 한계가 존재함
- 이에 본 장에서는 규칙기반 결과를 예비 학습자료로 활용하여 정제를 거쳐 KoBERT 모형을 미세조정함

□ 분석대상은 비내구재 관련 뉴스에서 추출한 기사 단위 분석자료이며, 각 기사 i 에 대해 분석 텍스트를 x_i , 소비 패턴 라벨을 y_i 로 정의

- 초기에는 상승, 하락, 중립 라벨을 균형있게 추출하여 학습 후보군으로 구성하였고, 이후 소비와의 직접적 관련성이 낮거나 라벨 해석이 불명확한 사례를 점검 후, 최종적으로 균형학습 자료를 구축함

□ 비내구재 뉴스의 특성을 반영하여 소비 패턴 여부와 방향성을 단계적으로 구분하는 방식으로 모형을 구성함

- 먼저 해당 분석 텍스트가 소비 패턴과 관련된 정보를 포함하는지 식별하고, 이후 소비 패턴이 확인된 관측치에 대해 상승 또는 하락 방향을 구분하도록 설계

$$q_i^{ND} = \text{소비 방향성 변화가 존재할 확률}$$

- 여기에서 q_i^{ND} 가 높을수록 해당 관측치가 비내구재 소비 증가 또는 감소와 관련된 정보를 포함할 가능성이 높다는 의미임
- 다음 단계에서는 소비 패턴 변화가 실제로 포착된 관측치에 한해 상승 또는 하락 방향을 구분함
 - 이는 모든 기사에 대해 곧바로 상승 또는 하락을 판단할 경우, 소비와 직접

관련이 없는 정보까지 방향성으로 해석되는 문제를 줄이기 위함임

- 따라서 소비패턴 변화 여부를 판별한 뒤, 변화가 있다고 판단된 경우에만 방향을 구분하는 조건부 구조를 다음과 같이 적용함

$$u_i^{ND} = P_i^{ND}(\text{상승} | \text{소비패턴 변화 있음})$$

$$d_i^{ND} = P_i^{ND}(\text{하락} | \text{소비패턴 변화 있음})$$

- 여기서 u_i^{ND} 는 비내구재 소비 증가 방향의 조건부 확률을 의미하며, d_i^{ND} 는 비내구재 소비 둔화 또는 감소 방향의 조건부 확률임
- 최종 소비 패턴 확률은 다음과 같이 산출됨

$$P_i^{ND}(\uparrow) = q_i^{ND} \times u_i^{ND}, \quad P_i^{ND}(\downarrow) = q_i^{ND} \times d_i^{ND} \text{18)}$$

- 이 구조는 비내구재 뉴스의 특성을 반영하도록 설계되어 있음. 즉, 소비 패턴이 약한 관측치는 q_i^{ND} 를 통해 지표 반영도가 낮아지고, 패턴이 뚜렷한 관측치는 상승 또는 하락 확률을 통해 주간지표에 더 크게 반영됨

□ 비내구재 주간지표 산출과정에서는 상승과 하락을 단순히 동일한 크기로 선택하지 않고, 하방신호를 상대적으로 더 민감하게 반영하는 비대칭 구조로 설계함

- 비내구재 소비가 일상적 소비 품목을 중심으로 구성되어 있어, 소비 둔화 또는 구매 위축과 관련된 뉴스신호가 단기 소비 흐름에 비교적 빠르게 반영될 가능성이 있기 때문
- 기사단위의 패턴 점수는 상승 방향 확률에서 하락 방향 확률을 단순 차감하는 것이 아닌 하락 확률에 조정 가중치 부여함

$$S_i^{ND}(\lambda) = u_i^{ND} - \lambda \times d_i^{ND}$$

- 여기에서 $S_i^{ND}(\lambda)$ 는 비내구재 관측치 i 의 상승 및 하락 방향 확률을 이용하여 구성한 방향성 점수이며, u_i^{ND} 는 상승 확률, d_i^{ND} 는 하락확률, λ 는

18) $P_i^{ND}(\text{중립}) = 1 - q_i^{ND}$

비내구재 부문 하방신호의 상대적 반영정도를 조정하는 계수임

- $\lambda=1$ 이면, 상승과 하락을 대칭적으로 반영하는 구조가 되고, 1보다 크면 하락 방향 신호가 지표에 상대적으로 더 크게 반영됨
- 본 분석에서는 λ 을 사전적으로 고정하지 않고, 다양한 지표 구성방식(기사수 등)을 검토함

□ 관측치 단위의 소비패턴 점수는 주간단위로 집계하였으며, 주간 비내구재 패턴 지수는 다음과 같이 산출함

$K_t^{ND}(\lambda)$ = 주 t 에 포함된 기사별 소비방향성 점수 $S_i^{ND}(\lambda)$ 를

각 기사의 소비방향성 관련성을 가중치로 하여 평균한 값

- 즉, $K_t^{ND}(\lambda)$ 는 해당 주의 비내구재 관련 뉴스 흐름이 전반적으로 소비증가 방향인지, 소비 둔화방향인지를 나타내는 값임
 - 값이 높을수록 상승 방향 신호가 상대적으로 강하고, 낮을수록 하락 방향 신호가 상대적으로 강하다고 할 수 있음
 - 각 기사에 따라 소비방향성과의 관련정도가 상이하므로, 소비 방향성 관련성이 높은 기사의 정보를 상대적으로 크게 반영하기 위해 가중평균 적용
- 이후 주간 뉴스량의 차이를 반영하기 위해 기사수 보정 ($K_t^{ND}(\lambda) \times \log(1+N_t^{ND})$)
 - 여기에서 N_t^{ND} 는 주 t 에 포함된 비내구재 관련 관측치 수이며, 로그 변환을 적용한 것은 뉴스량 정보를 반영하되, 기사수 증가가 지표를 과도하게 확대하지 않게끔 하기 위한 것임
 - 주간지표 단기 변동성 완화위해 이동평균을 적용하였으며, 비내구재 소매판매 YOY 와 비교할 수 있도록 평균과 표준편차를 이용하여 표준화함
- 선행 시차는 별도로 임의 지정하지 않고, t 주의 KoBERT 지표를 $t+L$ 주의 공식 비내구재 소매판매 YOY 와 매칭하여 상관계수를 산출

- 3주 이동평균, 4주 선행 정렬을 적용한 지표가 공식 비내구재 소매판매 YOY와 상대적으로 가장 양호한 관계를 보이는 것으로 나타남
 - 해당 지표는 월별 공식지표와 매칭 가능한 50개 관측치 기준으로 상관계수 약 0.3을 기록

□ 하단의 그림은 KoBERT 기반 주간지표와 비내구재 소매판매 YOY 지표를 나타낸 것임

- 첫 번째 그림은 KoBERT 모형에서 산출된 비내구재 주간지표 흐름을 나타낸 것이며, 두 번째 그림은 선행시차를 반영하지 않은 상태에서 뉴스 지표와 비내구재 소매판매 YOY를 비교한 것임
- 세 번째 그림은 뉴스지표를 4주 선행 정렬한 뒤 소매판매 YOY 와 비교한 것임

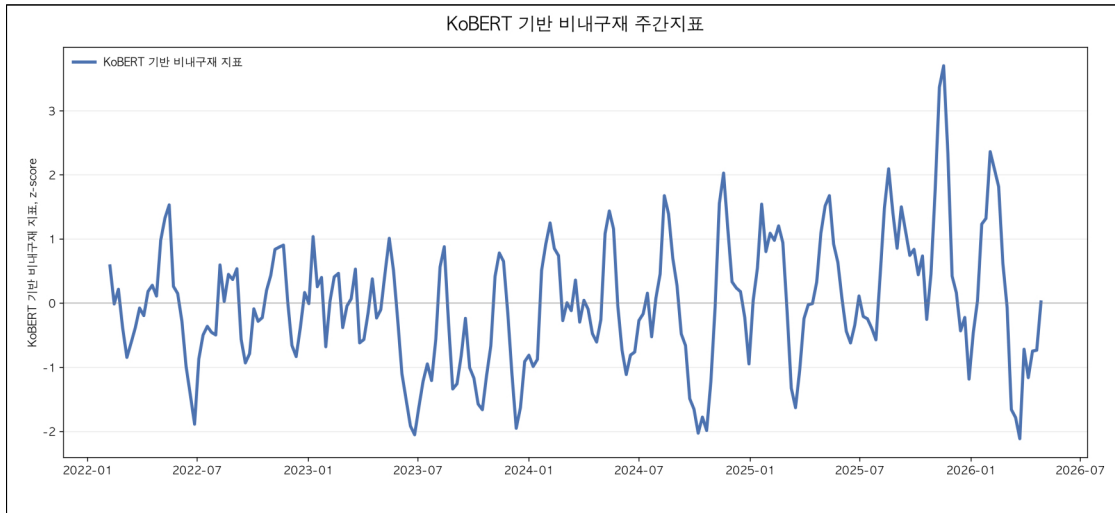
□ 비내구재 주간지표는 2022년 이후 비교적 높은 주간 변동성을 보이는 가운데, 일부 구간에서는 비내구재 소매판매 지수와 유사한 움직임이 있는 것으로 나타남

- 이는 비내구재 품목이 식료품, 생활용품, 화장품, 의약품 등 일상적 소비와 밀접한 항목을 포함하고 있어, 가격변화, 구매심리 등에 민감하게 반응하기 때문인 것으로 나타남
- 선행 시차를 반영하지 않은 경우에 뉴스지표와 비내구재 소매판매 YOY 간의 움직임이 일부 구간에서 어긋나는 모습이 관찰되는데, 이는 뉴스지표가 공식 통계와 동일한 시점에 반응하기 보다, 소비여건 변화에 대한 기대나 우려를 먼저 반영할 가능성이 있기 때문인 것으로 보임

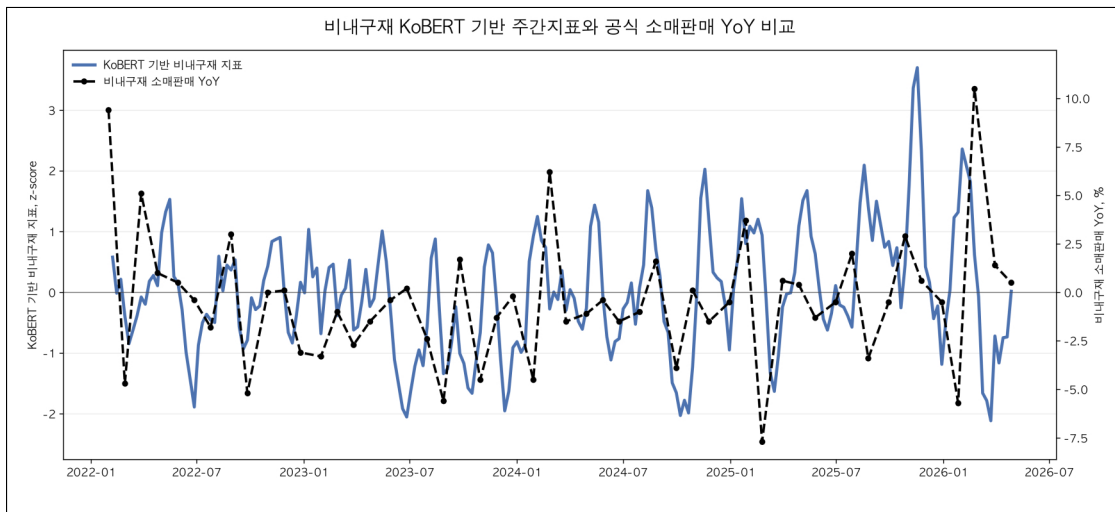
□ 4주 선행 정렬을 반영한 결과, 뉴스지표가 소매판매 YOY의 단기적인 상승 및 하락 방향을 일정 부분 선행적으로 반영하는 구간이 확인됨

- 월별 소매판매YOY 와 매칭한 결과, 관측치 기준 상관계수는 0.276으로 나타남
- 이는 강한 예측 관계라기보다, 비내구재 뉴스지표가 공식 통계 발표 이전에 소비 흐름의 방향성을 보조적으로 점검하는 데 활용될 가능성이 있음을 의미함
- 다만, 2024년 말부터 2025년 상반기까지 뉴스지표와 소매판매 YOY 간 괴리가 상대적으로 크게 나타나는데, 해당 구간은 정치적 불확실성, 사회적 혼란이 확대된 시기와 맞물려 있어, 뉴스기반 지표가 실제 소비활동 자체 뿐 아니라 불확실성, 소비심리 위축 등을 함께 반영했을 가능성이 있음
- 이 시기 괴리는 단순 모형오차라기 보다, 뉴스기반 고빈도 지표가 외생적 불확실성에 민감하게 반응하는 특성을 보여주는 것으로 해석됨
- 본 분석에서는 이러한 가능성을 확인하기 위해 2024년말 이후 정치적 불확실성 구간을 제외한 분석을 추가로 수행하였는데, 2024년 12월부터 2025년 5월까지를 제외하면 상관계수는 0.442로 크게 상승한 것으로 나타남
- 이는 2024년 말 이후의 외생적 불확실성 구간이 전체 상관관계를 낮추는 방향으로 작용했을 가능성과 추후 추가검토 필요성을 시사함

<그림 4-7> 비내구재 KoBERT 기반 주간지표 추이

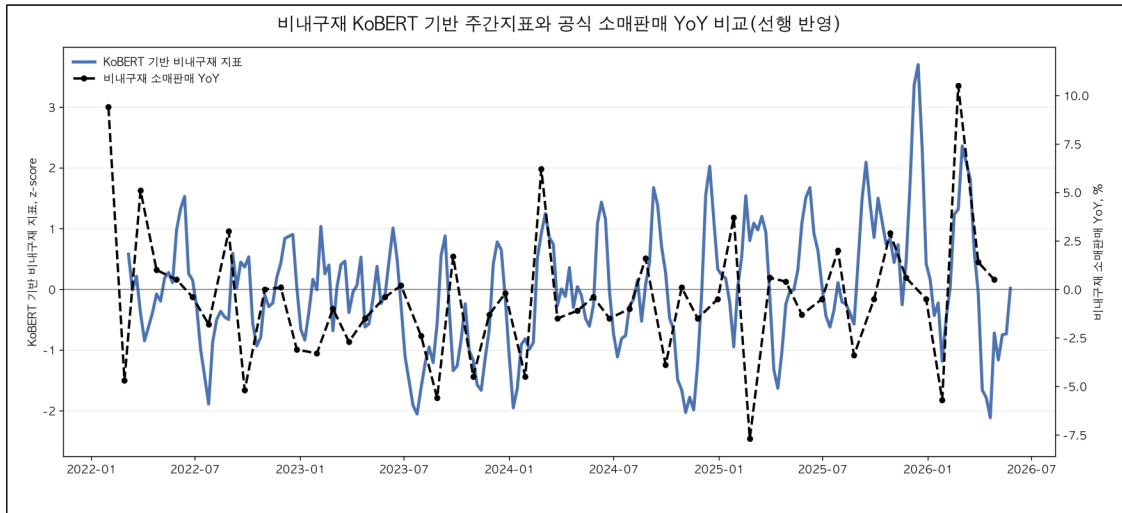


<그림 4-8> 비내구재 KoBERT 기반 주간지표와 비내구재 소매판매 증가율 추이



자료: 국가데이터처, 저자 작성

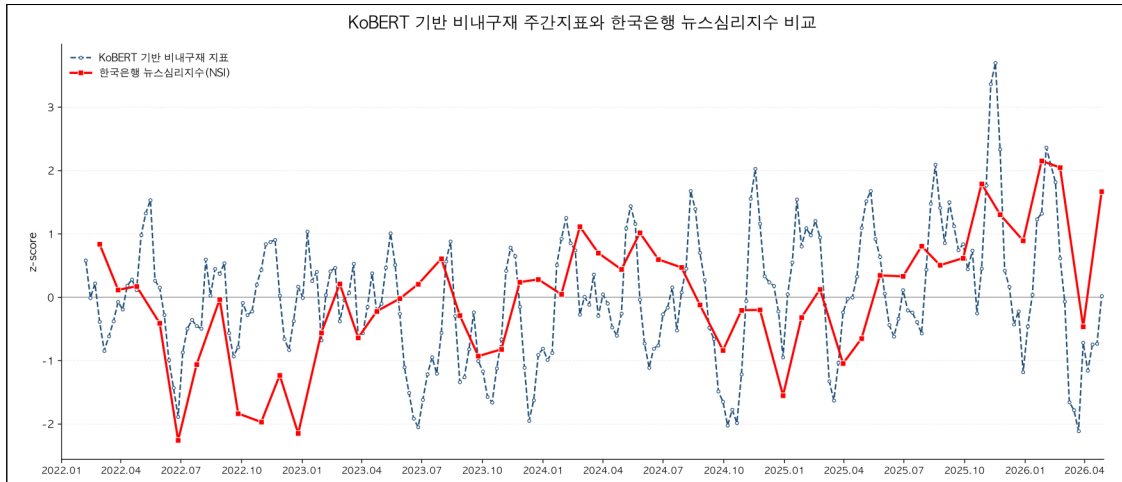
<그림 4-9> 비내구재 KoBERT 기반 주간지표와 비내구재 소매판매 증가율 추이(선행 반영)



자료: 국가데이터처, 저자 작성

- 비내구재 뉴스 지표 움직임을 경제 뉴스 전반의 심리흐름과 비교하기 위해 한국은행 뉴스심리지수(NSI)와의 관계를 살펴본 결과, 두 지표 간 상관관계수(분석 전체기간)는 0.351로 5% 수준에서 유의한 것으로 나타남. 두 지표가 일부 구간에서 유사한 방향으로 움직이는 경향이 관찰됨
 - 이러한 관계는 내구재 및 준내구재 지표에서는 상대적으로 뚜렷하게 나타나지 않음
 - 비내구재 소비는 식료품 등 일상적인 소비활동과 밀접하게 관련되어 있어, 전반적인 경제심리의 변화가 관련 뉴스에 일부 반영된 결과로 볼 수 있음

<그림 4-10> 비내구재 KoBERT 기반 주간지표와 新NSI 지수



주: 한국은행 新뉴스심리지수(NSI)는 월별 원 지수에 대해 분석기간의 평균을 차감하고 표준편차로 나누어 표준화하였으며, 각 월의 값은 해당 월에 포함된 주간 관측치 가운데 마지막 주에 표시.

출처: 한국은행 경제통계시스템

2. 준내구재

- 준내구재는 의류, 신발, 가방 및 관련 생활품목을 중심으로 구성되며, 구매시점의 선택 가능성이 비교적 높고 계절변화, 유행, 판촉행사 및 소비심리 등의 영향을 받는 특징이 있음
 - 준내구재 관련 뉴스에는 실제 판매와 수요의 증가, 감소 뿐 아니라 신제품 출시, 브랜드 소개, 패션동향 등 소비의 방향을 직접 나타내지 않는 정보도 상당부분 포함
 - 이에 따라 준내구재 품목이 기사에 등장했다는 사실만으로 해당 기사가 소비 확대 또는 위축을 나타낸다고 판단하기에는 한계가 있음
 - 특정 브랜드의 신제품 출시나 유행 전망을 다룬 기사는 준내구재 시장과 관련되어 있으나, 실제 소비의 증가 또는 감소를 직접적으로 의미하지 않을 수 있음

- 본 분석에서는 준내구재 뉴스지표는 전체 관련 기사를 일률적으로 상승과 하락으로 구분하기에 앞서, 준내구재 소비 패턴이 존재하는지를 판별한 이후 방향성이 확인된 기사에 한하여 상승, 하락 등을 구분하는 방식으로 구축
 - 소비 패턴이 불명확한 기사의 영향을 줄이고, 실제 판매 및 수요 변화와 연결되는 문맥을 보다 정확하게 식별하는 데 중점을 둠

- 준내구재 부문의 KoBERT 미세조정을 위해서 기존 규칙기반 분석에서 확보한 기사별 소비 방향 분류결과를 초기 학습정보로 활용하였으며, 이를 토대로 KoBERT가 준내구재 기사에 나타나는 소비 관련 표현과 문맥 구조를 학습하도록 구성함

- 모형의 입력자료는 준내구재 기사에서 핵심내용이 집중적으로 담긴 부분을 중심으로 구성함
 - 단순 기사제목이나 특정 키워드만을 활용하는 방식으로는 포착하기 어려운 문맥적 정보, 즉 준내구재 품목과 판매·수요 관련 표현이 실제 뉴스기사 내에서 어떠한 흐름과 맥락속에서 함께 등장하는지를 보다 정교하게 반영하기 위한 것임
 - 준내구재 관련 뉴스의 경우 제품 소개, 유행정보, 판촉 행사 안내 등 다양한 성격의 내용이 혼재되어 있어, 단편적인 단어 수준의 정보보다 문장 단위 이상의 문맥을 활용하는 것이 소비 패턴을 식별하는 데 유용함
 - 학습자료를 구성하는 과정에서 지나치게 짧아 의미있는 문맥을 확보하기 어려운 기사를 제외하여 입력 자료의 질을 확보하였으며, 동일한 본문이 학습자료와 검증자료 또는 평가 자료에 중복 포함되는 것을 방지하기 위해 본문 기준의 중복 제거 절차를 적용
 - 이러한 전처리 과정을 통해 모형이 특정 기사에 과도하게 적합되는 문제를 완화하였으며, 보다 일반화된 문맥 이해 능력을 학습할 수 있도록 함

- 준내구재 관련 뉴스에서 실제로 소비 증가나 감소와 같은 소비 패턴이 나타나는지를 확인하고, 이러한 패턴이 드러나는 경우에 한해 증가와 감소를 구분하도록 함

- 기사 i 의 본문을 x_i 라고 할 때, 다음의 조건부 확률을 추정하도록 함

$$p_i^D = P(D_i = 1 | x_i)$$

- 여기에서 $D_i = 1$ 은 해당 기사에 준내구재 소비의 증가 또는 감소 방향이 존재하는 경우를 의미하며, $D_i = 0$ 은 준내구재 품목이 언급되어 있더라도 소비 패턴이 식별하기에 어려운 경우를 의미함
 - 예를 들어 의류판매증가, 겨울 패딩 구매 증가, 의류 소비 부진과 같은 표현은 소비 흐름이 드러나는 기사로 판단될 수 있음
 - 신상품 의류 출시, 브랜드 협업 컬렉션 공개 및 패션행사 일정과 같이 준내구재와 관련은 있으나 실제 소비의 변화를 직접 나타내지 않는 기사는 소비 방향을 명확하게 파악하기 어려운 뉴스로 구분될 수 있음

- 하단의 그림은 KoBERT를 활용해 구축한 준내구재 뉴스기반 지표와 준내구재 소매판매 전년동월대비 움직임을 비교한 것임

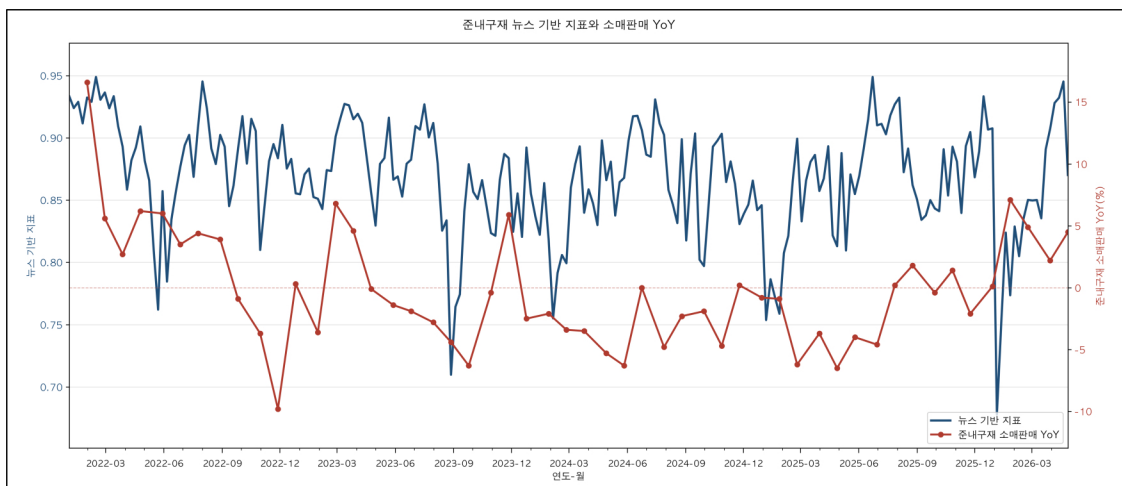
- 첫 번째 그림은 선행시차를 반영하지 않은 주간 뉴스지표와 월별 소매판매 전년동월비(YOY)를 각각의 축에 표시한 것이며, 두 번째 그림은 뉴스지표에 4주의 선행시차를 반영한 결과임
 - 세 번째 그림은 선행시차를 반영한 뉴스지표와 소매판매 전년동월비를 각각 표준화하여 동일한 척도로 비교한 것임

- 뉴스지표에 4주의 선행시차를 반영한 경우 소매판매지수의 상승 및 하락 전환에 앞서 뉴스지표가 동일한 방향으로 움직이는 구간이 보다 분명하게 관찰됨

- 4주 선행시차를 반영후 두 지표간 상관계수 0.426으로 나타남

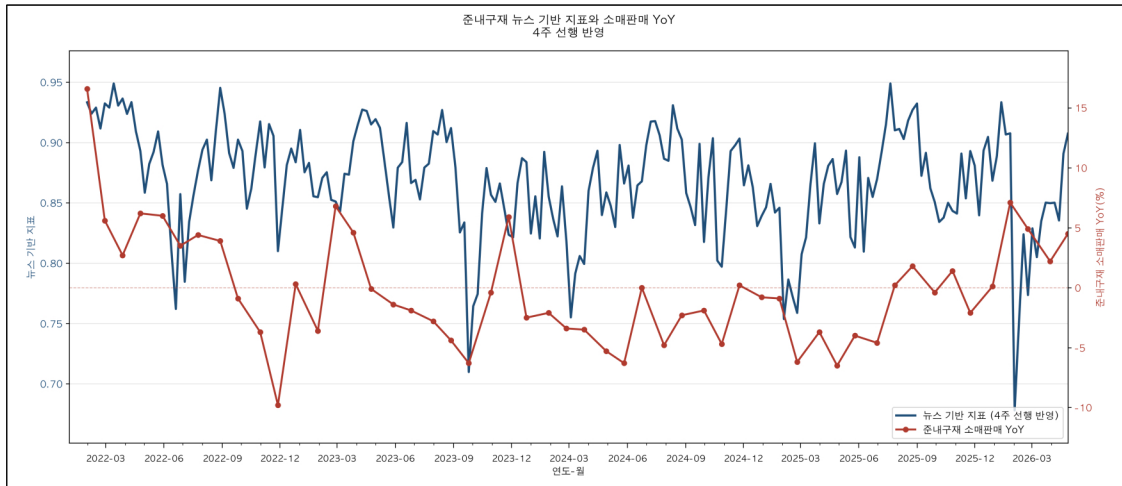
- 선행시차를 반영하지 않은 경우, 일부 구간에서 뉴스지표와 소매판매 YOY가 비교적 유사한 패턴을 보이고 있으나 변곡점의 시점과 단기 변동성에서 차이가 있는 것으로 확인됨
- 두 지표를 표준화한 그림에서도 4주 선행뉴스지표와 소매판매 YOY는 일부 주요 변동구간에서 같은 방향성을 보임
 - 2022년 연말 저점, 2023년 초 반등과 이후 둔화, 2023년 하반기 하락과 연말 반등, 2024년 상반기 부진과 하반기 중 일부 회복, 2025년 초 약화와 하반기 개선 등 큰 방향에서는 두 지표가 유사한 패턴을 보이고 있음

<그림 4-11> 준내구재 뉴스기반 주간지표와 준내구재 소매판매 증가율 추이



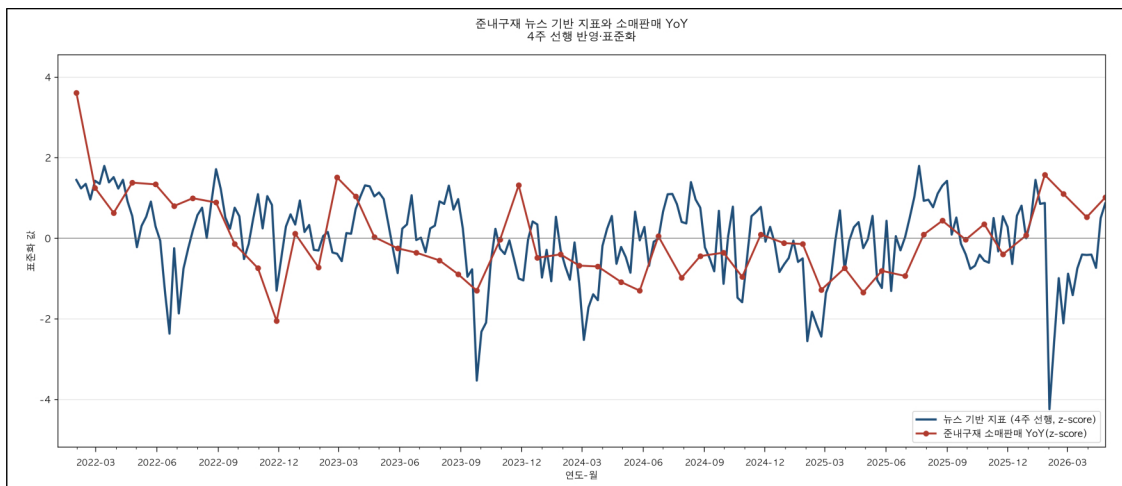
자료: 국가데이터처, 저자 작성

<그림 4-12> 준내구재 뉴스기반 주간지표와 준내구재 소매판매 증가율 추이(4주 선행 반영)



자료: 국가데이터처, 저자 작성

<그림 4-13> 준내구재 뉴스기반 주간지표와 준내구재 소매판매 증가율의 표준화 추이



자료: 국가데이터처, 저자 작성

3. 내구재

- 내구재 부문에서는 규칙기반 분석을 통해 기사별 소비 분석 결과를 학습자료로 활용하여 KoBERT 기반 모델을 구축함
- 내구재 관련 뉴스는 실제 판매 및 수요 변화를 다루는 기사 뿐 아니라 기업

인사, 기술개발, 신제품 소개, 산업정책 등 소비흐름과 직접 관련되지 않은 기사도 다수 포함되어 있음

- 모든 기사를 곧바로 분석에 활용하기보다, 먼저 기사내용이 소비 변화와 관련되어 있는지를 판별하고, 관련성이 확인된 기사에 대해서 소비 트렌드를 구분하는 순차적 분류 방식을 적용

□ 뉴스 기사를 활용한 뉴스지표를 구축하기 위해서 각 기사에 대해 산출된 분류 확률을 활용, 신뢰도가 충분히 높은 기사만을 선별함

- 이를 위해 본 연구에서는 소비 변화와 관련된 확률과 상승 및 하락 확률이 라는 두 가지 기준을 동시에 고려하였음
- 즉, 특정기사에 실제 소비 변화와 관련된 내용을 담고 있을 가능성을 나타내는 소비 변화관련성 확률이 일정 수준 이상이어야 하며, 동시에 해당 기사가 소비의 상승 또는 하락 방향을 명확하게 시사할 가능성을 나타내는 확률이 충분히 높아야 한다고 보았음
- 이에 따라 본 연구에서 상승 방향 확률이 높은 경우 상승기사, 하락쪽으로 높은 경우 하락기사로 분류함으로써 노이즈가 상대적으로 적고 방향성이 분명한 기사만을 기반으로 지표를 고려함
- 내구재 관련 뉴스에서는 소비 상승과 하락을 나타내는 신호가 뉴스지표에 동일한 정도로 반영되기는 어렵기 때문에, 소비의 상승 및 하락의 상대적인 반영정도를 조정할 수 있도록 조정계수(λ)를 적용
 - 조정계수 (λ)가 1이면 상승과 하락이 동일한 비중으로 반영되며 1보다 크면 하락의 반영정도를 확대하고 1보다 작으면 이를 축소함

□ 선별된 기사를 활용하여, 상승·하락 기사 비중의 차이, 로그 비율, 이동평균 기간을 조합한 후 여러 지표를 산출하여 검토한 결과, 4주 이동 평균을 적용한 지표가 내구재 소매판매 전년동월대비 증감률과 가장 높은 상관관계를 보임

- 상승 및 하락 기사 비중의 차이에 기사량을 반영하고, 4주 이동 평균을 적용할 경우, 12주 선행시차를 적용한 경우 두 지표의 움직임이 보다 유사하게 나타남

- 12주 선행을 반영한 확률기반 뉴스지표와 내구재 소매판매 전년동월대비 증감률 상관관계는 0.384로 나타남

□ 하단의 세 그림은 확률 기반 KoBERT 내구재 뉴스지수와 내구재 소매판매 전년동월대비 증감률 관계를 선행시차 여부와 표준화 방식에 따라 각각 비교한 것임

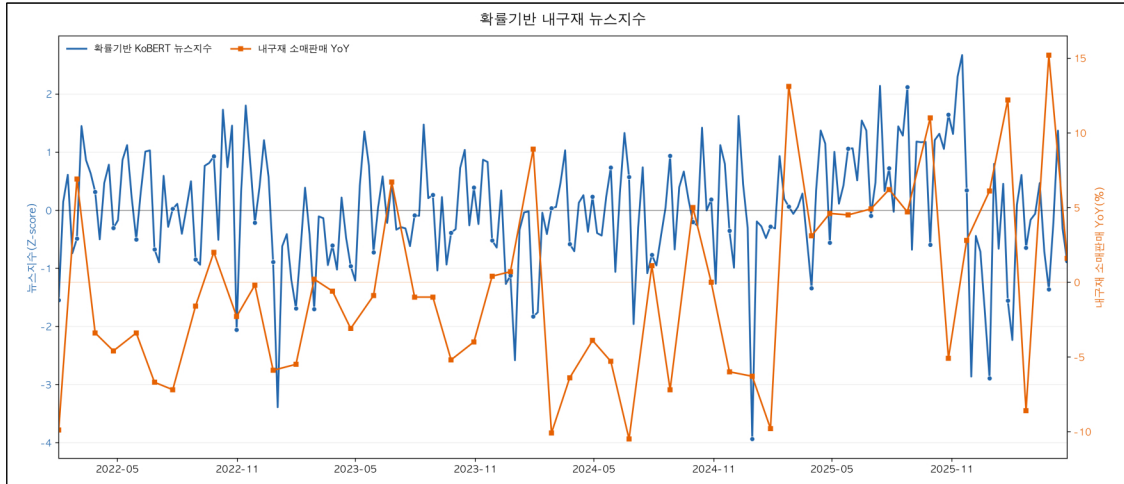
□ 뉴스지수에 12주 선행시차를 반영한 경우, 선행시차를 반영하지 않은 경우보다 뉴스 지수의 주요 상승 및 하락 구간과 소매판매 증가율의 변화가 상대적으로 유사하게 나타남

- 특히 뉴스지수가 상승한 이후 소매판매 증가율이 확대되고, 뉴스지수가 낮아진 이후 소매판매 증가율도 둔화되는 흐름이 여러 구간에서 관측됨
- 전체 분석기간을 기준으로 12주 선행 뉴스지수와 내구재 소매판매 전년동월대비 증감률 간의 상관계수는 0.384로 나타남

□ 표준화된 두 지표를 비교하면 전체기간 일부 단기적인 불일치가 존재하지만, 뉴스지수의 주요 상승 및 하락 흐름이 이후 소매판매 증감률의 변화와 일정 부분 대응하는 모습이 관측됨

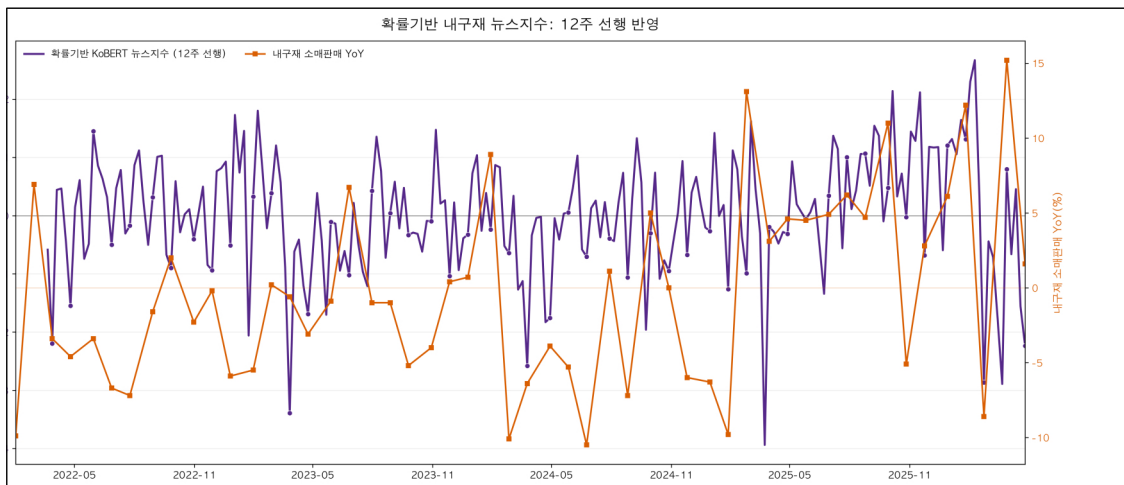
- 2025년 이후 구간만을 대상으로 계산한 12주 선행 뉴스지수와 내구재 소매판매 증가율간의 상관계수는 약 0.6으로, 전체 기간보다 높은 수준을 보임
- 이는 최근 구간에서 뉴스지수와 내구재 소매판매 간의 연계성이 보다 뚜렷하게 나타났음을 의미

<그림 4-14> 확률기반 내구재 뉴스지수와 내구재 소매판매 전년동월대비 증가율 추이



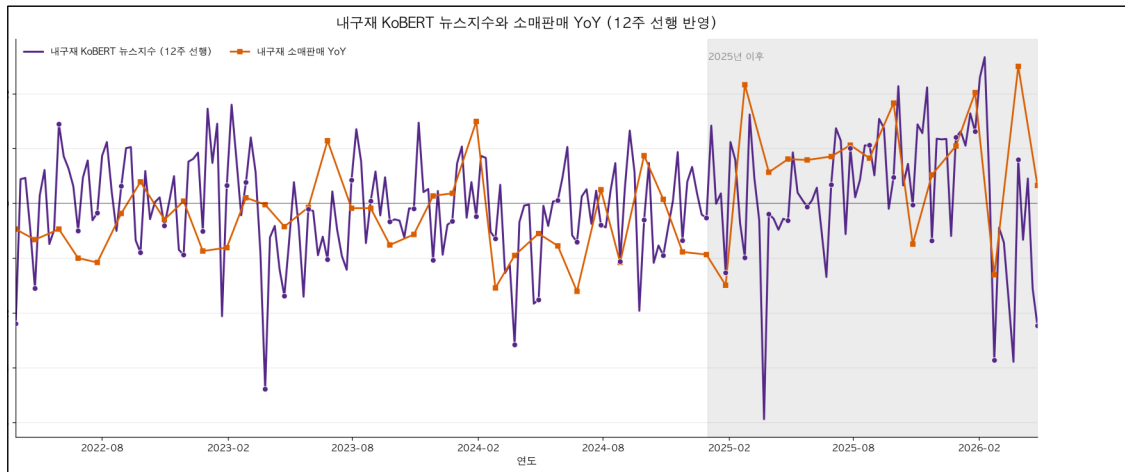
자료: 국가데이터처, 저자 작성

<그림 4-15> 확률기반 내구재 뉴스지수와 내구재 소매판매 전년동월대비 증가율 추이(12주 선행 반영)



자료: 국가데이터처, 저자 작성

<그림 4-16> 내구재 KoBERT 뉴스지수와 내구재 소매판매 전년동월대비 증가율의 표준화 추이(12주 선행 반영)



자료: 국가데이터처, 저자 작성

- 내구재 뉴스지수는 소매판매 증감률과 같은 시점에서 비교하는 것보다 12주 선행시차를 반영하였을 때 두 지표의 관계가 보다 분명하게 나타났음
 - 내구재 뉴스지수는 약 12주 이후의 판매 흐름을 살펴보기 위한 선행적 보조 지표로 활용할 수 있을 것으로 보임

Ⅲ. 주간 뉴스지수 기반 소비지표 예측

- 본 절에서는 앞에서 구축한 주간 뉴스지수의 소비지표에 대한 예측에 활용할 수 있는지 실제 예측을 수행하여 검토
 - 뉴스지수가 주별로 구축되어 뉴스데이터로부터의 정보를 빠르게 포착할 수 있으나 예측하고자 하는 소비지표는 월별 혹은 분기별로 발표되기 때문에 데이터 간 주기 불일치를 해결할 수 있는 혼합주기(mixed frequency) 모형으로 예측 수행
 - 이를 위해 본 절에서는 혼합주기 및 고차원 문제를 다룰 수 있는 모형을 검토하고 주간 뉴스지수 기반 월별 소매판매 나우캐스팅 가능성을 검토
- 소매판매액지수, GDP 구성요소 중 민간 소비 등 공식 소비통계는 월별 혹은 분기별로 공표되며, 실제 지표가 공표되기까지 상당한 시차가 존재하여 실시간으로 경기를 판단하기에는 한계가 존재
 - 월별 소매판매액지수는 익월말에 공표되며 분기별 지표인 GDP 중 민간 소비는 분기 종료 후 약 한달 반 가량 후에 속보치가 발표되어 공표시점까지 시차가 존재
- 반면 일별, 주별 뉴스지수 등 고빈도로 이용가능한 데이터는 확정된 소비 지표 공표 이전에 소비흐름을 포착할 가능성이 높기 때문에 실시간 경기 진단에 있어 활용도가 높을 가능성
 - 특히 뉴스지수는 공표까지 시차가 있는 거시경제적 지표와 다르게 새로운 뉴스가 수집될 때마다 실시간으로 진단에 필요한 정보를 수집
- 그러나 실제 실시간 경기 진단 수행에 있어 접하게 되는 문제는 예측변수(예

를 들어 주별 뉴스지수)와 예측타겟(예를 들어 월별 소비지표) 간 빈도가 달라 이를 분석하기 위한 계량경제학적 방법이 필요

- 혼합주기 모형은 설명변수와 종속변수 간 서로 주기가 다른 경우 모형 내에서 이를 통합하여 분석하는 방법으로 일반적인 실시간 경기 진단 상황과 같이 고빈도 예측변수를 이용한 저빈도 예측타겟 분석에 활용 가능
- 본 절에서는 고빈도 주별 뉴스지수를 이용하여 저빈도 월별 소비지표를 예측을 수행¹⁹⁾

□ 이에 따라 본 절에서는 전통적 시계열 모형과 함께 다양한 혼합주기 모형을 고려하여 예측을 수행하였으며 이 중 가장 예측력이 높은 모형을 선택

- y_t 는 시점 t 에서의 저빈도 예측 목표(예를 들어 t 월의 월별 소매판매액지수의 차분한 값), $x_{k,t}$ 는 시점 t 내 k 번째 고빈도 설명변수(예를 들어 t 월 k 번째 주 주별 뉴스지수의 차분한 값)
- 벤치마크 모형으로 전통적인 시계열 예측 모형인 Random Walk(RW), AR을 선택
 - RW: $Y_t = Y_{t-1} + u_t$
 - AR: $Y_t = \alpha + \beta Y_{t-1} + u_t$
 - 이와 같은 모형은 추가적인 설명변수 없이 예측 타겟의 전기 과거값만을 이용하여 예측하는 모형으로 예측 평가에 있어 실제 뉴스지수의 예측 기여가 유의미한지를 판정하는 기준선 역할을 수행
- 혼합주기 모형으로는 Bridge와 MIDAS(MIXed DATA Sampling) 계열 모형들을 고려
 - Bridge 모형은 단순히 고빈도 설명변수를 저빈도 예측변수와 동일한 주기로 집계(예를 들어 주별 뉴스지수를 집계하여 월별로 빈도 통일)하여 예측에 활용하는 방법

19) 분기별 GDP 민간소비에 대해서도 뉴스지수를 이용한 예측을 수행할 수 있으나 제4장에서 구축한 뉴스지수가 2022년 1월부터 2026년 4월까지로 최대 13분기만을 포함하고 있어 분기별 소비지표에 대한 예측 가능성 검증은 수행하기 어려운 한계가 있음.

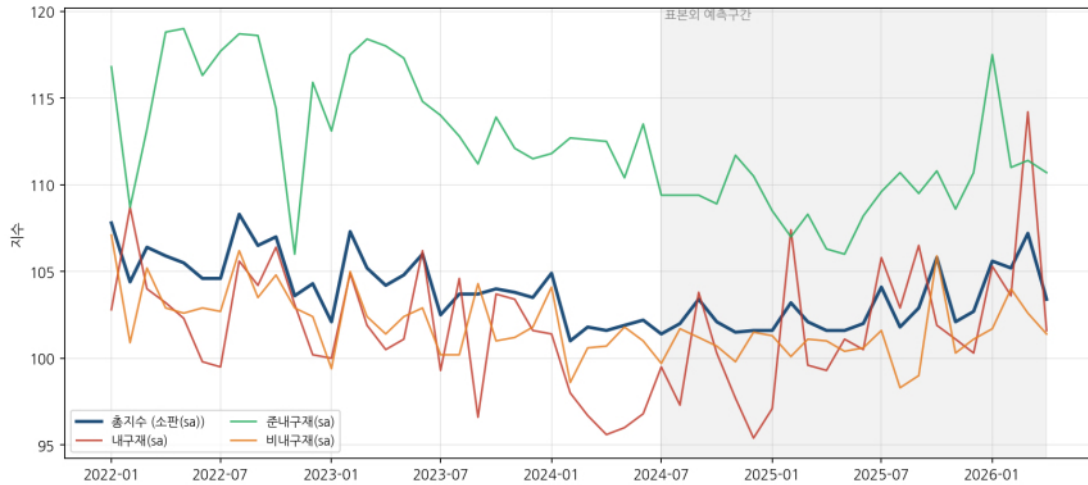
- Bridge: $Y_t = \alpha + \beta Y_{t-1} + \gamma x_t + u_t$, $x_t = \sum_k x_{k,t}$
- MIDAS(Ghysels et al., 2004)는 고빈도 변수를 시차가중 다항(예를 들어 Beta 함수 형태)을 바탕으로 저빈도로 변환(고빈도 변수의 가중평균으로 저빈도 변환)한 후 예측에 활용하는 방법
- MIDAS: $Y_t = \alpha + \beta Y_{t-1} + \sum_k B(k; \theta) x_{k,t} + u_t$, $B(k, \theta)$ =Beta 함수
- U-MIDAS(Foroni et al., 2015)는 무제약 MIDAS로 시차가중을 제약하지 않고 개별 계수를 추정하는 방법으로 저빈도와 고빈도 간 빈도차가 크지 않은 경우(예를 들어 월별과 주별 간 빈도차는 4 내외) 활용 가능
- U-MIDAS: $Y_t = \alpha + \beta Y_{t-1} + \sum_k \gamma_k x_{k,t} + u_t$
- 선행예측은 반복(iterated) 예측이 아닌 직접 예측(multi-step ahead)으로 수행
- 예측에 대한 평가는 평균제곱오차(RMSE)를 기준으로 수행하였고 RMSE가 작을수록 예측 오차가 낮은 수준임을 의미함.

$$\text{RMSE: } \sqrt{\frac{1}{T} \sum_t (Y_t - \hat{Y}_t)^2}$$

- 최종 예측 타겟은 계절조정 소매판매액지수로, 이를 설명하기 위한 뉴스지수는 주별 내구재, 준내구재, 비내구재 뉴스지수로 설정
 - 계절조정 소매판매액지수는 전월대비 증감률 등 소비의 단기적 흐름을 파악에 활용되고 있으며 속보성이 확보된 뉴스데이터를 활용한 예측 분석에 적합한 변수로 판단됨.
 - 예측에서 예측 타겟은 계절조정 소매판매액지수의 차분된 값을 이용
 - 이에 따라 예측 변수인 뉴스지수도 전주대비 차분한 값을 예측에 이용
 - 2022년 1월부터 2024년 6월까지를 모형 추정에 활용하고 예측 기간은 2024년 7월부터 2026년 4월로 설정하였으며 확장창 표본외 예측(expanding

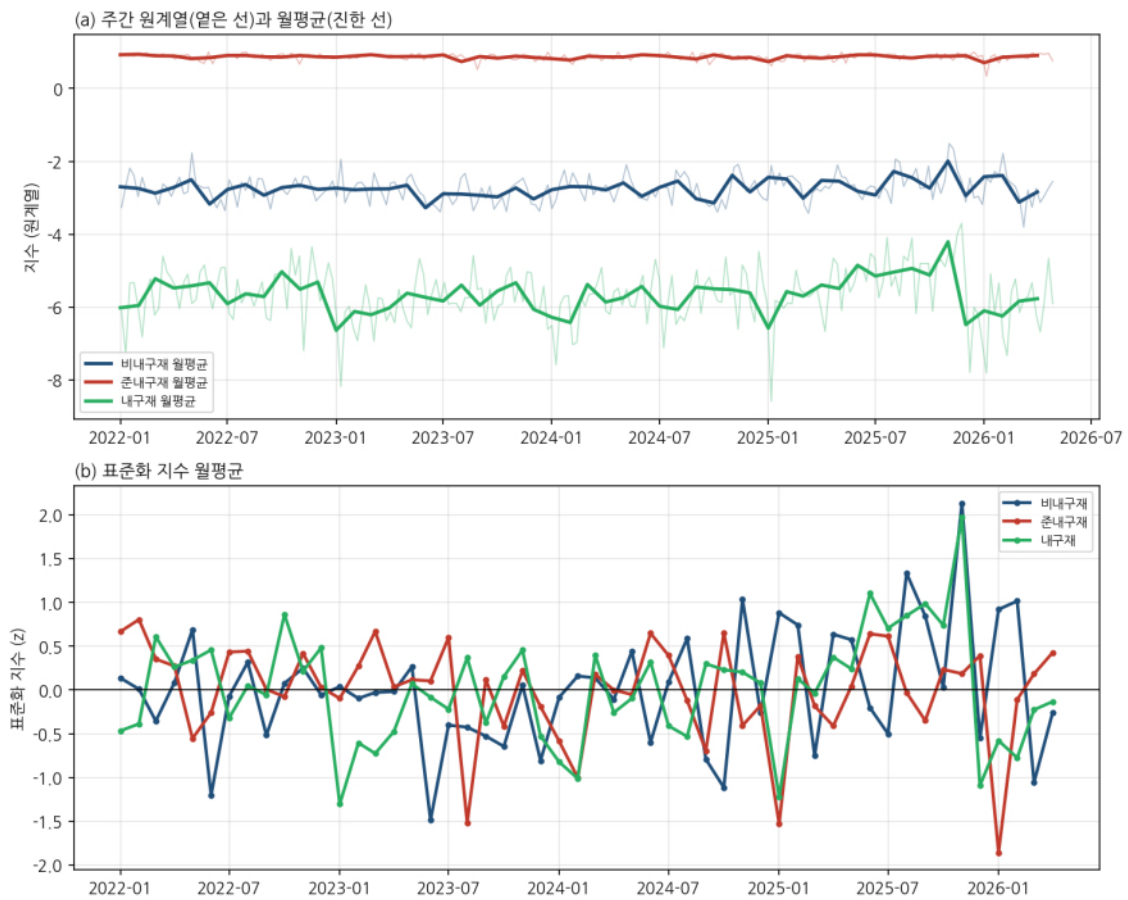
window out-of-sample forecasting)으로 평가

<그림 4-17> 월별 계절조정 소매판매액지수 (총지수 및 품목별 지수)



- 아울러 소비 기사 문장을 KoBERT를 이용하여 상승/하락 분류하여 주간 지수화하여 설명변수로 활용
 - 뉴스지수의 선행조정은 제4장에서의 분석 결과를 바탕으로 조정

<그림 4-18> 주간 뉴스 소비지수 (월평균 및 표준화 지수)



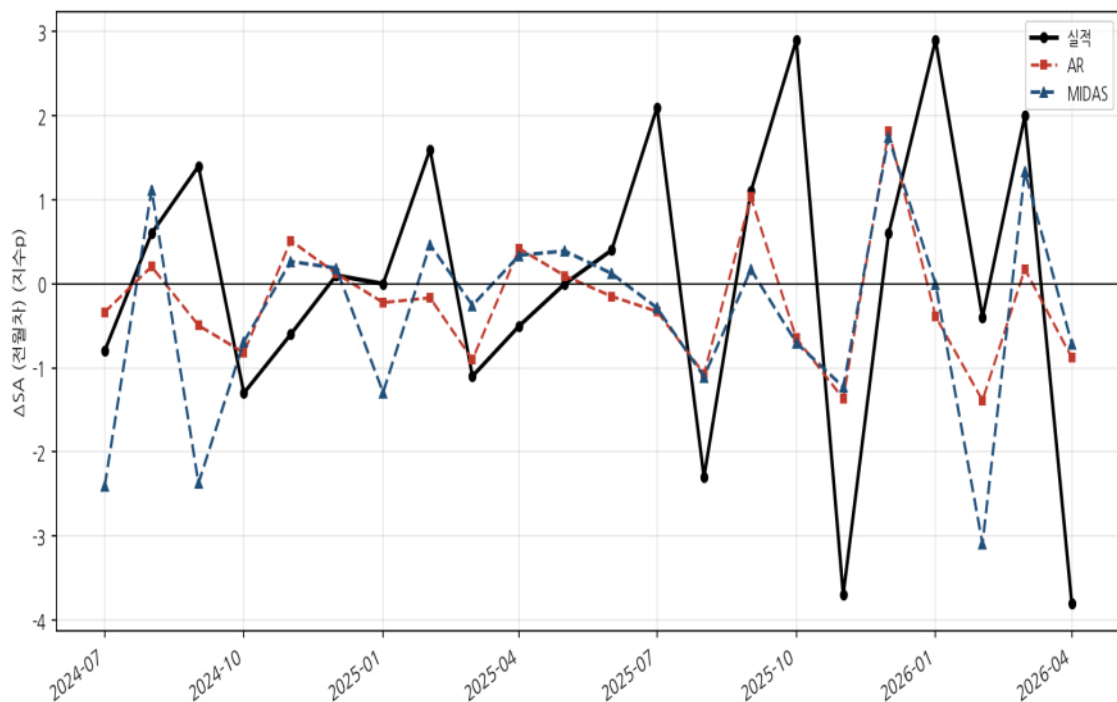
□ 당월 나우캐스트 수행 결과 뉴스지수를 활용한 혼합주기 예측 모형이 AR(1) 모형에 비해 예측력이 높지 않은 것으로 확인됨.

- 전반적으로 혼합주기 모형의 RMSE가 전통적인 시계열 모형 대비 높은 것으로 나타났으며 예측치도 실제 소비지표의 흐름을 제한적으로 포착하는 것으로 나타남.

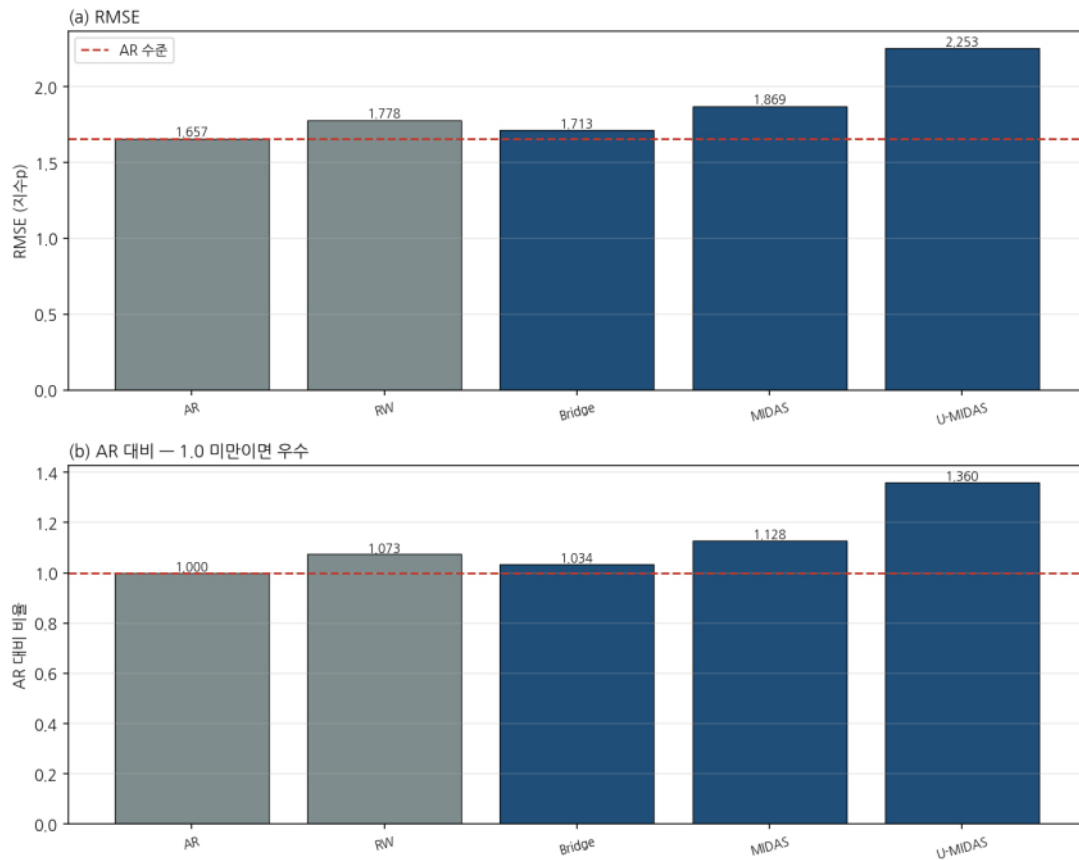
<표 4-2> 당월 나우캐스트 예측오차 비교

모형	RMSE	MAE	RMSE(AR 대비)
AR	1.66	1.27	1.00
RW	1.78	1.37	1.11
Bridge	1.71	1.33	1.03
MIDAS	1.87	1.51	1.13
U-MIDAS	2.25	1.79	1.36

<그림 4-19> 당월 나우캐스트 예측 결과 (AR 및 MIDAS)



<그림 4-20> 당월 나우캐스트 예측력 비교 (시계열 모형 및 혼합주기모형)



- 전통적인 시계열 예측 모형 중 AR의 RMSE는 1.66으로 RW(1.78)에 비해 낮은 수준
- 혼합주기 모형인 Bridge, MIDAS, U-MIDAS 모두 RMSE가 1.71~2.25로 AR에 비해 전반적으로 높게 나타났으며, 이와 같은 높은 예측오차는 AR에 비해 예측력이 높지 않음을 의미함.
 - 특히 MIDAS와 U-MIDAS를 비교했을 때 U-MIDAS의 예측오차가 MIDAS에 비해 매우 높은 수준을 보인 것은 소표본 문제와 함께 모수 제약의 부재로 인한 과적합 문제일 가능성이 높음
- 이와 같은 결과는 뉴스지수가 소비 예측에 유용하지 않음을 시사하는 것이 아니라 예측에서 ① 소표본 문제를 해결할 필요가 있고 ② 뉴스데이터의 신호 특성을 고려하여 예측에 활용하여야 함을 시사

- 특히 제4장에서의 분석에서 확인할 수 있듯이 뉴스데이터로부터의 신호는 동시성보다는 선행성에 있기 때문에 이러한 선행관계를 예측 분석에 활용하는 것이 필요

□ 이러한 분석 결과를 바탕으로 예측 모형을 다지수, 다시차 정보를 반영한 모형으로 확장하여 벌점화(penalized) U-MIDAS 모형을 고려

- 모형의 설명변수로 다양한 지수와 각 지수의 다양한 시차를 포함하게 되면 모형에서의 설명변수 숫자가 커지게 되어 표본과 유사하거나 혹은 표본보다 커지게 될 수 있음
 - 구체적으로 뉴스지수의 선행관계를 고려하여 각 뉴스지수(내구재, 준내구재, 비내구재)별 3개월까지의 시차항을 설명변수로 고려
- 이를 극복하기 위해 모형에 벌점을 도입하여 데이터를 바탕으로 중요한 변수만을 선별
- 모형 구성은 2단계(AR회귀, 뉴스지수 예측)로 구성
 - 1단계에서는 예측 타겟인 y_t 를 AR(1)으로 추정하여 전기값으로 설명할 수 있는 지속성을 가진 부분과 그 외 부분(잔차)로 구분
 - 2단계에서는 잔차를 다지수, 다시차 뉴스지수를 이용한 벌점화 U-MIDAS를 통해 예측
 - 최종 예측값은 $\hat{Y}_t = \hat{\alpha} + \hat{\beta} Y_{t-1} + \hat{u}_t$ 로 정의하였으며 이는 1단계 AR회귀에서 예측된 값($\hat{\alpha} + \hat{\beta} Y_{t-1}$)과 2단계 뉴스지수를 이용한 예측($\hat{u}_t$)로 구성
 - 일반적으로 설명변수 숫자에 비례하여 표본 내 예측(in-sample fit)은 좋아지나 표본 외 예측(out-of-sample forecasting)은 악화되나 뉴스지수가 소비 지표에 대한 예측력이 있다면 이러한 모형의 예측력이 우월할 수 있음.
- 뉴스지수에 대한 예측은 U-MIDAS 형태 모형에 벌점화 회귀를 결합한 형태로 설정하였으며 다양한 벌점화 방법을 고려
 - 이는 소비지표에서 전기 지표로 설명할 수 있는 부분을 제외한 후 고차원

고빈도 특성으로 남은 부분을 설명하며 벌점을 도입하여 과적합을 억제하고 설명력이 유의미한 지표만을 선별

- 벌점화 방법 중 하나로 고려한 LASSO는 L1 벌점을 도입하여 일부 계수를 정확히 0으로 만들어 희소 선택이 가능하나 상관이 높은 특성군에서는 임의로 하나만 선택하는 불안정성이 있고, 신호가 소수 특성에 집중될 때 유리
- ElasticNet은 L1과 L2의 혼합 벌점 형태로 L2 항이 상관 특성을 함께 수축(grouping effect)해 다수의 작은 계수를 허용
- Adaptive LASSO는 일반적으로 1단계 릿지 추정치를 바탕으로 특성별 가중치를 만들어, 중요한 특성은 약하게, 약한 특성은 더 강하게 벌점하는 방법으로 편의를 줄여 점근적 오라클 성질을 가지게 됨(Zou, 2006)
- 각 벌점화 모형에서 하이퍼파라미터에 대한 튜닝은 학습표본을 시간순서대로 나누어 적용한 3분할 교차검증(3-fold cross validation) 방법을 통해 수행

□ 벌점화 U-MIDAS 모형을 이용하여 다지수, 다시차 뉴스지수의 소비지표 예측을 수행한 결과 전반적으로 예측력이 개선되었으며 뉴스지수를 이용한 예측은 당기예측보다는 단기예측에 더 중요하게 작용

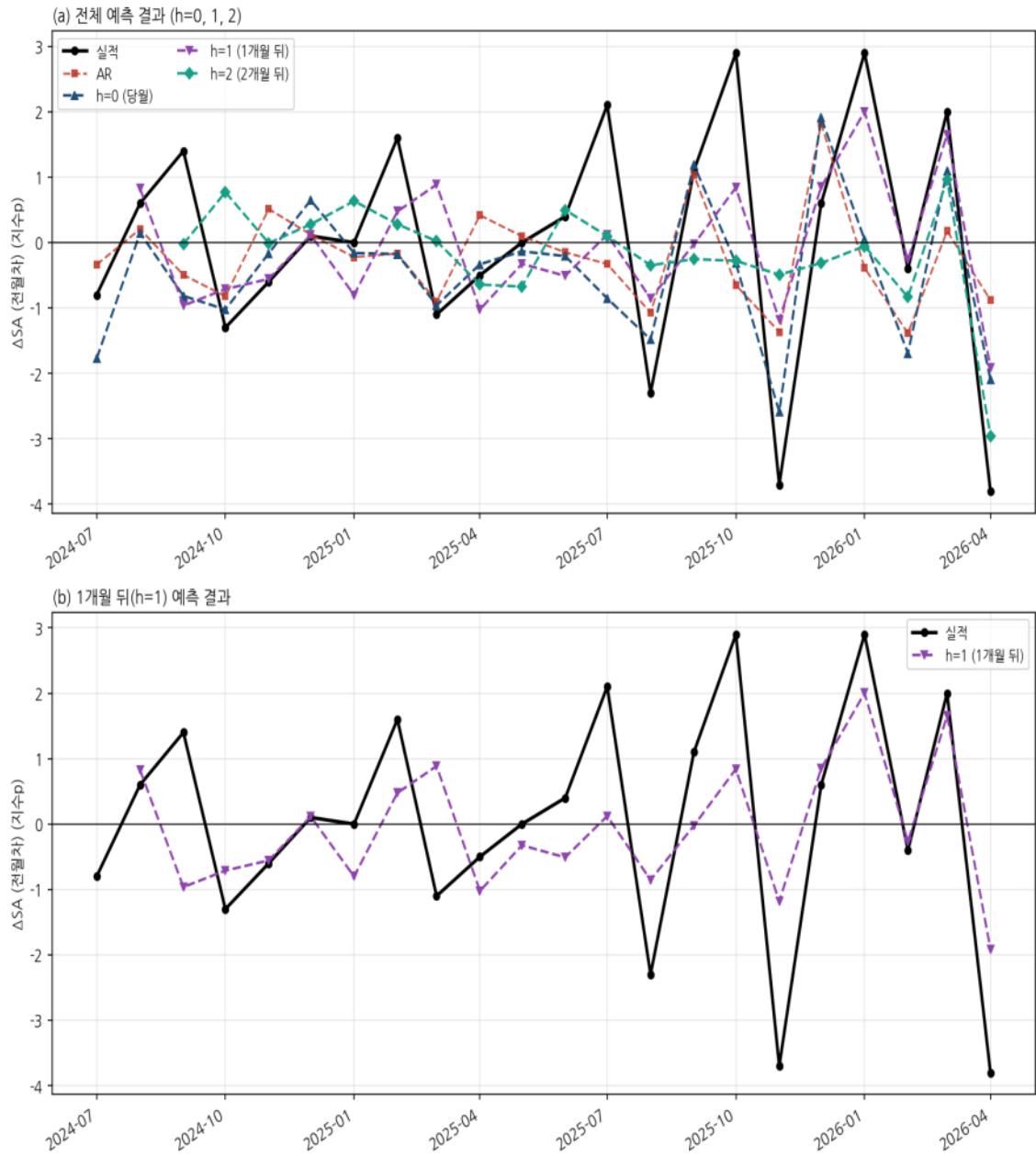
- 벌점화 U-MIDAS 모형 중 Ridge 회귀 추정치를 초기값으로 하여 가중치를 구성한 후 ELasticNet을 적용한 Adaptive Elastic-Net(Zou and Zhang, 2009)을 U-MIDAS에 적용(AEN U-MIDAS)
- AEN U-MIDAS로 예측을 수행한 결과 예측을 수행한 예측시계에서 전반적으로 예측력이 개선되었고 특히 단기예측에서 예측력이 제고

<표 4-4> AEN U-MIDAS 예측력 비교

	AR(RMSE)	AEN U-MIDAS(RMSE)	DMW Test
t 월 예측	1.66	1.45	1.60
$t+1$ 월 예측	1.89	1.30	2.33**
$t+2$ 월 예측	2.00	1.61	1.51

- 구체적으로 AEN U-MIDAS는 AR에 비해 전 예측 시계에서 더 작은 RMSE를 기록하였으며 특히 1개월 뒤 예측에서 RMSE의 감소폭이 가장 크게 나타남.
- 아울러 소표본의 한계가 있으나 모형별 예측력이 통계적으로 유의한지를 검정하는 Diebold-Mariano-West 검정(DMW Test)을 통해 살펴보았을 때 1개월 뒤 예측에서 AEN U-MIDAS와 AR 간 예측력 차이는 5% 수준에서 통계적으로 유의함
 - 다만, DMW 검정 결과로 보았을 때 당월 예측과 2개월 뒤 예측에서 AEN U-MIDAS의 예측력 개선은 통계적으로 유의하지는 않은 것으로 나타남.

<그림 4-21> AEN U-MIDAS 예측 결과



○ 이와 같이 구축된 예측치는 계절조정 소매판매액지수의 흐름을 일정 수준 이상 포착하고 있는 것으로 나타남.

- 특히 1개월 뒤 예측에서 AR에 비해 예측력 개선이 크게 나타났고 실제 1개월 전에 예측된 예측치(보라색)의 흐름이 소매판매지수(검정색)의

흐름과 유사하게 나타남.

□ 이러한 예측 결과는 경제학적 해석과도 부합함.

- 내구재는 상대적으로 고가 품목이며 계획적인 소비를 하게 되어 소비심리가 실제 구매를 선행할 수 있으며 뉴스지수 분석에서는 2개월 선행신호로 나타남.
- 의류 등 준내구재도 유사하게 심리가 실제 구매를 선행할 수 있으며 뉴스지수 분석에서는 1개월 선행신호로 나타났으며 내구재에 비해서는 선행신호가 보다 단기적
- 반면 생필품 등 비내구재는 습관적, 비탄력적 소비로 지속성이 낮을 수 있으며 이에 따라 뉴스데이터에서 포착되는 소비 심리로는 예측이 어려울 수 있음.

□ 따라서 뉴스지수는 당월 나우캐스팅보다는 단기 예측을 위한 지표로 활용하는 것이 뉴스데이터로부터의 정보 활용을 극대화하는 방안으로 평가됨.

- 실제 각 지표가 1~2개월 선행지표로 작용하고 있으며 뉴스데이터를 활용한 예측 결과 역시 단기 예측에서 예측력이 개선

□ 다만 이러한 분석은 가용한 뉴스 데이터의 범위가 제한적인 상황에서 발생하는 소표본의 한계로 인해 해석에 유의할 필요가 있음

- 소표본으로 인해 모형 간 예측력 차이가 통계적으로 유의미한지 확인하는 검정(DMW 검정 등)의 검정력(power)이 크게 저하될 수밖에 없음.
- AEN U-MIDAS에서 초기 추정치인 Ridge 회귀 추정치가 소표본에서는 불안정할 수밖에 없으며 이로 인해 추정량의 오라클 성질이 훼손될 우려가 있음
- 아울러 모형 내 하이퍼파라미터의 튜닝 과정에서 활용한 교차검증 또한 소표본으로 인해 모형 학습에 활용되는 데이터가 매우 제한적이게 되며 따라서 모형 안정성이 저하될 우려가 있음.

- 또한 예측에 활용한 변수들이 타겟 변수의 경제적 관계를 반영하지 않고 최근 단기간 흐름에 과적합되었을 가능성이 있어 향후 예측력이 저하될 위험도 상존

□ 따라서 본 절에서의 분석은 뉴스지수가 소비지표 예측력 제고에 도움이 될 수 있다는 제한적인 수준의 증거를 제공하는 것으로 볼 수 있음.

- 이에 따라 향후 충분한 수준의 시계열 데이터가 축적된 후 엄밀한 분석과 추가 검증이 필요한 상황으로 판단됨.

제5장 결론

I. 주요 연구결과

- 본 연구는 국내 경제뉴스 빅데이터를 활용하여 소비 관련 뉴스의 특성과 내용구조를 파악하고, 뉴스기사에 포함된 소비 관련 문맥과 방향성을 식별하여 고빈도 소비예측지표를 구축·검증하고자 하였음
 - 이를 위해 먼저 BIGKinds 경제뉴스를 활용한 기초분석을 통해 소비 관련 뉴스의 규모와 내용적 특성을 파악하고, 이후 사전기반 분류와 KoBERT 기반 문맥분류를 통해 내구재·준내구재·비내구재의 품목군별 뉴스지수를 구축함
 - 마지막으로 주간 뉴스지수와 월별 소매판매액지수의 서로 다른 자료주기를 고려한 혼합주기 예측모형을 적용하여 뉴스 기반 소비지표의 단기 예측 가능성을 검토함

- 뉴스 빅데이터 기초분석 결과, 소비 관련 보도의 규모는 시기별로 변화하였으며 기사량의 변화만으로 실제 민간소비의 증감을 직접 판단하기 어렵다는 점을 확인함
 - 2022년 1월부터 2026년 4월까지 분석대상 경제기사는 총 5,838,352건이었으며, 이 가운데 제목에 ‘소비’가 포함된 소비 관련 기사는 76,301건으로 전체 경제기사의 약 1.31%를 차지함
 - 소비기사 비중은 2022~2024년 1.17~1.24% 수준에서 비교적 안정적으로 나타났으나 2025년에는 1.62%로 상승하였으며, 월별로는 2025년 7월 2.74%, 8월 2.26%, 9월 2.15%로 상대적으로 높은 수준이 나타남
 - 그러나 해당 시기에는 ‘소비쿠폰’, ‘소상공인’, ‘가맹점’, ‘상품권’ 등 민생회복 소비쿠폰과 소비지원정책 관련 표현이 집중적으로 나타났으며, 이는 소비기사의 증감이 실제 소비활동뿐 아니라 특정 정책이나 사건에 대한 언론보도의 집중에 의해 크게 영향을 받을 수 있음을 보여줌

- 소비 관련 뉴스의 내용분석 결과에서도 동일한 ‘소비’ 관련 기사집단 안에서 다른 경제적 의미와 의제가 혼재되어 있음을 확인함
 - 주요 특성추출어에는 소비정책, 물가, 금융, 유통·결제, 산업 및 시장상황과 관련된 표현이 함께 나타나 단순한 단어빈도만으로 소비뉴스의 경제적 의미를 충분히 파악하기 어려운 것으로 나타남
 - BERTopic 분석에서는 식품·농산물 소비, 금융소비자 보호, 쌀 소비촉진, 에너지·전력 소비, 소비자물가, 민생회복 소비쿠폰, 카드·결제 및 소비행태, 민간소비와 경제성장 등 서로 다른 내용의 주요 토픽이 확인됨
 - 특히 민생회복 소비쿠폰 토픽은 2025년 전체 소비기사의 4.58%까지 확대되어 같은 시기의 소비기사 비중 상승과 일관된 모습을 보인 반면, 에너지·전력 소비와 같이 가계의 상품·서비스 구매와 직접 관련되지 않는 의미의 ‘소비’도 별도의 토픽으로 나타남
 - 이는 뉴스자료를 소비동향 분석에 활용하기 위해서는 기사량이나 특정 키워드뿐 아니라 개별 기사가 어떠한 소비 관련 문맥을 포함하고 있는지를 구분할 필요가 있음을 보여줌

- 이에 본 연구에서는 소비 품목군별 관련 표현과 소비활동·방향성 표현을 결합한 사전기반 분류를 실시하고, 그 결과를 바탕으로 KoBERT 기반 문맥분류모형으로 분석을 확장함
 - 사전기반 접근은 대규모 뉴스자료에서 내구재·준내구재·비내구재 관련 기사를 선별하고 소비 상승·하락 등의 방향성을 일차적으로 식별하는 데 활용하였으나, 동일한 표현이라도 가격·판매·수요 등 기사 문맥에 따라 의미가 달라질 수 있다는 한계가 확인됨
 - 이에 KoBERT 기반 문맥분류에서는 품목의 단순 출현 여부보다 해당 기사에 실제 소비 패턴이 존재하는지를 먼저 식별하고, 소비 관련성이 확인된 기사에 대해 상승·하락의 방향성을 구분하는 방식으로 뉴스정보의 문맥을 보다 정교하게 반영함

□ KoBERT 기반 품목군별 뉴스지수와 공식 소매판매지표를 비교한 결과, 소비 품목의 특성에 따라 뉴스정보와 실제 소비지표 간 관계 및 선행시차에 차이가 나타남

- 비내구재 뉴스지수는 4주 선행 정렬한 경우 소매판매 전년동월대비 증가율과 0.276의 상관관계를 보였으나, 일부 기간에서는 두 지표 간 괴리가 크게 나타나 전 기간에 걸쳐 안정적인 관계가 있다고 보기는 어려웠음
- 다만 비내구재 뉴스지수는 한국은행 뉴스심리지수(NSI)와 0.351의 상관관계를 보였으며 5% 수준에서 통계적으로 유의한 것으로 나타나, 식료품 등 일상적 소비와 관련된 뉴스정보가 전반적인 경제심리의 변화를 일정 부분 함께 반영할 가능성이 확인됨
- 준내구재의 경우 4주 선행 뉴스지수와 소매판매 전년동월대비 증가율 간 상관관계수가 0.426으로 나타났으며, 일부 주요 변동구간에서 뉴스지수가 공식통계의 변화보다 앞서 동일한 방향으로 움직이는 모습이 확인됨
- 내구재의 경우 12주 선행 뉴스지수와 소매판매 전년동월대비 증가율 간 상관관계수가 0.384로 나타났으며, 특히 2025년 이후 구간에서는 약 0.6 수준으로 높아져 내구재 뉴스정보에서 상대적으로 긴 선행관계가 나타날 가능성을 확인함
 - 이러한 결과는 소비 품목군별 구매행태와 정보반영 속도가 상이하므로 뉴스 기반 소비지표를 하나의 동일한 구조로 구성하기보다 품목별 특성과 선행시차를 고려하여 활용할 필요가 있음을 보여줌

□ 주간 뉴스지수를 활용하여 월별 소매판매액지수를 예측한 결과, 기본적인 혼합주기 모형에서는 뉴스지수의 추가가 당월 예측력을 개선하지 못한 반면, 다지수·다시차 정보를 반영한 벌점화 모형에서는 예측오차가 상대적으로 낮게 나타남

- 당월 나우캐스팅에서는 AR 모형의 RMSE가 1.66으로 나타난 반면, Bridge

1.71, MIDAS 1.87, U-MIDAS 2.25로 나타나 기본적인 혼합주기 모형이 AR 모형보다 낮은 예측오차를 보이지 않음

- 이에 내구재·준내구재·비내구재 뉴스지수의 다지수·다시차 정보를 함께 활용하고 과적합을 완화하기 위해 별점화 U-MIDAS 모형으로 분석을 확장함
- Ridge 회귀 추정치를 초기값으로 활용하여 Adaptive Elastic-Net을 적용한 AEN U-MIDAS 모형의 경우, RMSE는 당월 예측 1.45, 1개월 뒤 예측 1.30, 2개월 뒤 예측 1.61로 각각의 AR 모형보다 낮게 나타남
- Diebold-Mariano-West 검정을 통해 모형 간 예측력 차이를 추가로 검토한 결과, 1개월 뒤 예측에서는 AEN U-MIDAS와 AR 모형 간 예측력 차이가 5% 수준에서 통계적으로 유의하게 나타났으나, 당월과 2개월 뒤 예측에서는 통계적으로 유의한 차이가 확인되지 않음
- 이러한 결과는 뉴스지수를 활용한 소비예측의 가능성이 모든 예측시계에서 동일하게 나타나는 것은 아니며, 본 분석에서는 특히 1개월 뒤 단기예측에서 상대적으로 뚜렷하게 나타났음을 보여줌

□ 뉴스 빅데이터는 공식 소비통계를 직접 대체하는 지표라기보다 소비와 관련된 변화와 기대를 고빈도로 파악하여 기존 소비통계를 보완하는 자료로 활용할 가능성이 있음

- 특히 단순한 기사량이나 키워드 빈도보다는 기사 문맥을 반영한 분류, 소비 품목군별 이질성, 뉴스신호의 선행시차 및 복수의 뉴스지수를 함께 고려할 때 뉴스정보의 활용 가능성이 보다 구체적으로 나타남
- 따라서 뉴스 기반 소비지표의 유용성은 하나의 단일 지수 자체보다는 고빈도 뉴스정보를 품목별·문맥별로 구조화하고 이를 기존 공식통계와 결합하여 활용하는 데서 찾을 수 있음

II. 연구의 시사점

- 첫째, 뉴스 빅데이터는 공식 소비통계를 대체하기보다 공식 통계의 공표 시차를 보완하는 고빈도 정보원으로 활용할 가능성이 있음
 - 소매판매액지수와 같은 공식 소비통계는 소비동향을 파악하기 위한 핵심 자료이지만 월별로 작성·공표되므로 급변하는 소비여건을 실시간에 가깝게 확인하는 데 일정한 시차가 존재함
 - 반면 뉴스자료는 일별·주별로 지속적으로 축적되므로 소비 관련 정책, 시장 여건, 소비심리 및 품목별 수요 변화에 관한 정보를 상대적으로 신속하게 포착할 수 있음
 - 본 연구에서도 뉴스 기반 소비지표가 당월의 소비수준을 직접 추정하는 지표라기보다 일부 품목에서 향후 소비 흐름과 선행적으로 관련되는 것으로 나타나, 공식 소비통계와 병행하여 활용할 수 있는 보완지표로서의 가능성을 확인함
- 둘째, 뉴스자료를 경제지표로 활용하기 위해서는 단순한 기사량이나 키워드 빈도보다 기사의 내용과 문맥을 식별하는 과정이 중요함
 - 기초 뉴스분석에서는 소비 관련 기사량이 크게 증가한 시기라 하더라도 실제 민간소비의 증가가 아니라 특정 소비지원정책에 대한 보도 집중이 상당 부분 반영될 수 있음을 확인함
 - 또한 BERTopic 분석 결과 동일한 ‘소비’ 관련 기사집단 안에서도 소비정책, 물가, 금융소비자 보호, 산업활동, 에너지 소비 등 서로 다른 내용영역이 함께 나타나, 기사량이나 특정 단어의 출현만으로 소비동향을 일관되게 해석하기 어렵다는 점을 보여줌
 - 이에 따라 뉴스 기반 경제지표를 구축할 때에는 단순한 단어 출현 여부를 넘어 어떤 품목의 어떤 소비활동이 어떠한 방향으로 변화하는지를 문맥적으

로 구분하는 과정이 필요하며, 본 연구의 사전기반 분류와 KoBERT 기반 문맥분류는 이러한 방향으로의 확장 가능성을 보여줌

□ 셋째, 뉴스 기반 소비지표는 하나의 통합지표로 접근하기보다 소비 품목군별 특성과 정보반영 시차를 고려하여 구성할 필요가 있음

- 분석 결과 비내구재, 준내구재 및 내구재 뉴스지수는 공식 소매판매지표와의 관계와 선행시차에서 서로 다른 특성을 보였음
- 준내구재에서는 약 4주의 선행관계가 상대적으로 뚜렷하게 나타난 반면, 내구재에서는 약 12주의 선행관계가 관찰되어 계획적 구매의 성격이 강한 품목에서 뉴스정보가 실제 구매보다 비교적 먼저 반영될 가능성이 확인됨
- 반면 비내구재는 공식 소매판매지표와의 관계가 상대적으로 불안정한 반면 한국은행 뉴스심리지수(NSI)와 일정한 관련성을 보여, 일상적 소비와 관련된 뉴스정보가 실제 판매뿐 아니라 전반적인 경제심리와의 함께 움직일 가능성이 있음
- 이는 소비의 성격과 구매주기가 품목군마다 상이한 만큼 뉴스지수의 구성방식과 선행시차 역시 획일적으로 적용하기보다 품목별 특성에 맞게 설정할 필요가 있음을 시사함

□ 넷째, 뉴스정보의 예측 활용에서는 단순히 고빈도 변수를 추가하는 것보다 복수의 뉴스지수와 시차정보를 구조적으로 결합하는 방식이 중요함

- 당월 나우캐스팅에서는 Bridge, MIDAS, U-MIDAS 등 기본적인 혼합주기 모형이 AR(1) 기준모형보다 우수한 예측성과를 보이지 않아, 고빈도 뉴스정보를 추가하는 것만으로 예측력이 자동적으로 향상되는 것은 아닌 것으로 나타남
- 반면 내구재·준내구재·비내구재의 복수 뉴스지수와 다양한 시차항을 함께 고려하고 벌점화 회귀를 적용한 AEN U-MIDAS 모형에서는 AR 모형보다 전반적으로 낮은 예측오차가 나타났으며, 특히 1개월 뒤 예측에서는 모형 간

예측력 차이가 5% 수준에서 통계적으로 유의하게 나타남

- 따라서 뉴스 기반 소비지표를 예측에 활용할 때에는 개별 지수의 동시적 움직임만을 이용하기보다 품목별 뉴스정보와 다양한 시차를 함께 고려하는 방식을 검토할 필요가 있음

□ 마지막으로 본 연구는 뉴스 기반 소비지표를 공식 소비통계를 대체하는 지표라기보다 소비동향 판단과 단기 전망을 보완하는 지표로 활용할 가능성을 검토함

- 공식 소비통계는 실제 소비활동을 측정하는 기준자료로서의 역할을 유지하되, 뉴스 기반 지표를 함께 활용할 경우 통계 공표 이전에 나타나는 소비 관련 관심, 심리 및 품목별 변화신호를 추가적으로 확인할 수 있음
- 특히 뉴스지표의 절대적인 수준 자체보다 평상시와 다른 변화, 품목별 방향 전환 및 일정한 선행시차에서 나타나는 신호를 공식 통계와 함께 점검하는 방식으로 보완적으로 활용할 수 있을 것으로 판단됨

Ⅲ. 연구의 한계 및 향후 과제

- 첫째, 뉴스자료는 실제 소비활동뿐 아니라 정책발표, 사회적 사건, 언론의 관심 및 경제심리 등 다양한 요인의 영향을 받을 수 있으므로 뉴스지표의 변화를 실제 소비의 변화로 직접 해석하는 데에는 한계가 있음
 - 특히 소비 관련 기사 선별과 텍스트 분석은 뉴스의 내용과 문맥을 최대한 반영하도록 구성하였으나, 모든 민간소비 관련 기사를 완전하게 포괄하거나 기사 본문의 경제적 의미를 모두 식별하는 데에는 제약이 있음
- 둘째, 품목군별 뉴스지수와 공식 소비지표 간의 관계 및 선행시차는 분석기간과 시점에 따라 달라질 수 있으므로 결과의 안정성을 지속적으로 검증할 필요가 있음
- 향후 뉴스자료와 공식 소비통계가 추가적으로 축적되면 품목군별 선행관계와 지표의 분류·예측성능이 새로운 시기에도 유지되는지를 점검할 필요가 있음
- 셋째, 본 연구의 예측평가는 비교적 짧은 표본외 예측기간과 제한된 월별 관측치에 기초하고 있어 예측결과의 안정성과 통계적 검정력에 일정한 제약이 있음
 - Diebold-Mariano-West 검정 결과 1개월 뒤 예측에서는 AEN U-MIDAS와 AR 모형 간 예측력 차이가 5% 수준에서 유의하게 나타났으나, 당월과 2개월 뒤 예측에서는 통계적으로 유의한 차이가 확인되지 않음
 - 따라서 현재의 결과는 뉴스 기반 소비지표의 활용 가능성을 검토한 결과로 해석하고, 향후 표본이 충분히 축적되면 모형의 안정성과 예측성능에 대한

추가 검증을 수행할 필요가 있음

- 향후에는 뉴스자료와 공식 소비통계가 추가적으로 축적됨에 따라 뉴스 기반 소비지표를 기존 소비통계를 보완하는 지표로 활용할 수 있는 범위와 조건을 지속적으로 검토할 필요가 있음

참고문헌

서범석·이영환·조형배. (2022). 「기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용」. 『국민경제리뷰』, 2022년 제1호, 68-90.

- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593-1636. <https://doi.org/10.1093/qje/qjw024>
- Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment (Federal Reserve Bank of San Francisco Working Paper 2017-01). <https://doi.org/10.24148/wp2017-01>
- Buckman, S. R., Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). News sentiment in the time of COVID-19. *FRBSF Economic Letter*, 2020-08.
- Huang, C., Simpson, S., Ulybina, D., & Roitman, A. (2019). News-based sentiment indicators (IMF Working Paper No. 2019/273). International Monetary Fund. <https://doi.org/10.5089/9781513518374.001>
- Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*, 79(5), 3105-3147. <https://doi.org/10.1111/jofi.13377>
- Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1), 137-162. <https://doi.org/10.1016/j.jfineco.2016.01.032>
- Babii, A., Ghysels, E., & Striaukas, J. (2021). Machine learning time series regressions with an application to nowcasting (LIDAM Discussion Paper LFIN 2021004). Université catholique de Louvain.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models (CIRANO Working Paper 2004s-20). CIRANO.
- Faroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A*, 178(1), 57-82. <https://doi.org/10.1111/rssa.12043>
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429. <https://doi.org/10.1198/016214506000000735>
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253-263. <https://doi.org/10.1080/07350015.1995.10524599>
- West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067-1084.

- <https://doi.org/10.2307/2171956>
- Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578. <https://doi.org/10.1111/j.1468-0262.2006.00718.x>
- Federal Reserve Bank of San Francisco. (2026). Daily News Sentiment Index. <https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/>
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. <https://doi.org/10.48550/arXiv.2203.05794>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP 2020* (pp. 4512–4525). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
- McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 861. <https://doi.org/10.21105/joss.00861>
- McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205. <https://doi.org/10.21105/joss.00205>
- Sentence Transformers. (n.d.). sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [Model card]. Hugging Face. <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>