

## ICT 산업 경기 파악을 위한 ICT 뉴스경기지수 작성\*

손녕선 정보통신정책연구원 연구위원 | nsson@kisdi.re.kr

### 개요

비정형 텍스트 데이터에 대한 접근성이 높아지고 자연어 처리(NLP) 기술이 발달하면서, 텍스트 데이터를 활용한 분석에 대한 관심이 꾸준히 높아지고 있다. 그중에서도 뉴스 기사를 사용하는 경우가 많은데, 이는 뉴스 기사가 자료의 양, 속도성, 다양성에서 강점을 가지고 있기 때문이다(서범석 외, 2022). 이러한 뉴스 기사의 장점을 활용해 경제적 환경과 산업 동향을 예측하려는 연구들이 활발하게 진행되고 있다. 해외 연구를 보면, 「Baker et al.」(2016)은 경제정책 관련 기사를 수집하여 정책의 불확실성을 확인하는 작업을 수행하였고, 「Scott and Varian」(2014)은 뉴스 기사를 통해 실업률 및 물가의 초단기 변동을 측정하고자 하였다. 「Shapiro et al.」(2022)은 감정 분석을 위한 개선된 모델을 논의하고 있으며 뉴스를 통한 감정 분석의 경제적 함의를 제공하고 있다. 각국의 중앙은행과 국제통화기금(IMF) 등에서도 뉴스를 통한 심리지수 관련 연구를 수행하고 있으며, 한국은행에서도 2021년부터 뉴스심리지수를 작성하여 공개하고 있다.

이처럼 경제 전망과 현황 파악을 목적으로 뉴스 기사 분석이 활발하게 이루어지고 있는 가운데, ICT 산업의 경기 파악을 위해 뉴스 기사를 활용한 지표를 작성하는 것 또한 의미가 클 것으로 생각된다. ICT 산업은 디지털전환의 중심에 위치하고 있으며, 혁신이 빠르게 이루어지는 것만큼 변동성이 큰 분야이기도

\* 본고는 KISDI에서 수행한 「ICT 주요이슈 분석체계 구축」(정책연구 23-28) 보고서의 '제3장 ICT 뉴스 경기지수 구축 및 고도화' 내용을 요약 및 정리한 것이다. 분석 결과는 최신 자료를 사용하여 업데이트를 진행하였다.

하다. 이러한 산업의 특성상, 산업의 경기를 신속하게 파악하는 일이 중요하다. 시의성 있는 뉴스 기사를 활용하여 산업 경기를 반영하는 지수를 작성하는 것이 이러한 목적을 달성하는 하나의 방법이 될 수 있다. 이를 배경으로 본고에서는 뉴스 기사를 통해 ICT 산업의 경기를 측정하는 뉴스경기지수를 작성한 방법을 소개하고 주요 결과를 소개하고자 한다.

## 작성방법

### 1. 키워드 선정 및 뉴스 기사 수집

ICT 뉴스경기지수 작성을 위해서는 먼저 ICT와 관련된 뉴스들을 수집할 필요가 있다. 이를 위해 ICT 관련 뉴스를 대표할 수 있는 키워드를 선정하는 작업을 수행하였다. 먼저 ICT 산업, ICT 기업, ICT 수출, ICT 품목, IT 산업, IT 기업, IT 수출, IT 품목 이상 8개 키워드를 우선으로 선정하였다. 8개 키워드는 ICT 분야를 큰 틀에서 포괄할 수 있는 단어들이다. 이후 보다 세부적인 영역에서 키워드를 추가하기 위해 과학기술정보통신부, 가트너(Gartner) 등 주요 기관과 기업에서 발표하는 최신 자료들을 참고하여 ICT 트렌드를 조사하였다. 이렇게 조사한 결과를 바탕으로 키워드 후보를 선정하고, 이들 중

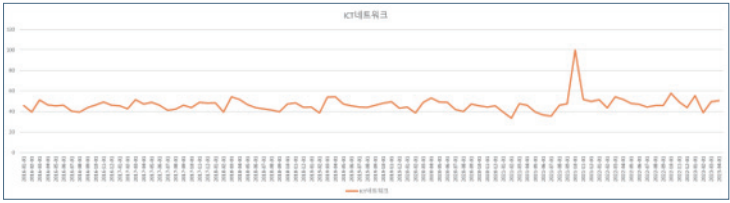
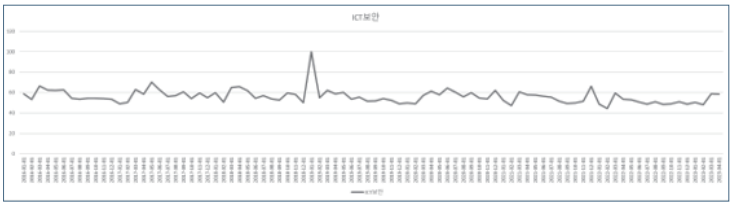

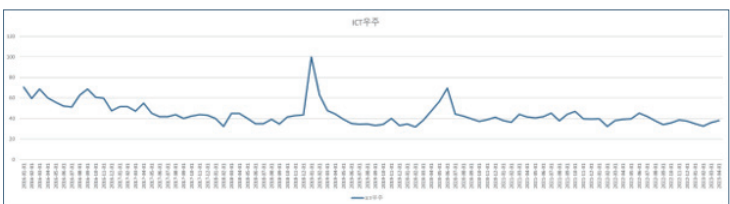



적절한 키워드를 선별하는 작업을 진행하였다. 키워드 선정 기준은 구글 트렌드(Google Trend)와 네이버 데이터랩(DataLab)에서 제공하는 각 키워드의 검색량을 활용하였다<sup>1</sup>. 검색량 결과에 따라 “과거부터 현재까지 꾸준히 언급되는 ICT 키워드”와 “최근에 나타난 ICT 키워드” 2개의 그룹으로 키워드를 분류하였다. 먼저 “과거부터 현재까지 꾸준히 언급되는 ICT 키워드”는 2010년대 이후부터 2020년대까지 일정 수준 이상으로 검색량이 유지된 키워드들을 의미한다. 판단의 기준으로는 ‘ICT’ 키워드의 평균 검색량과 분산을 사용하였다. ‘ICT’ 키워드가 ICT 전반을 포괄하는 키워드이기 때문에 벤치마크로 사용하기에 적절하다고 판단하였다. ‘ICT’ 키워드와 평균 검색량이 비슷한 수준이거나 기간 내 평이한 수준을 유지하는 키워드들, 그리고 ‘ICT’ 키워드의 평균 분산과 비슷하거나 더 낮은 키워드들을 선정하였다. 이는 해당 키워드가 꾸준히 일정 수준의 검색량을 유지하였는지를 판단하기 위한 방법이다. 이렇게 “과거부터 현재까지 꾸준히 언급되는 ICT 키워드”에는 ICT 네트워크, ICT 보안, ICT 로봇, ICT 우주, ICT 플랫폼 이상 5개를 선정하였다.

<sup>1</sup> 네이버 데이터랩은 2016년 이후부터 검색량을 제공하기에 이전 기간은 구글트렌드를 활용하였다.

[그림 1] 키워드 검색 동향: 과거부터 현재까지 꾸준히 언급되는 ICT 키워드

(단위: 개)

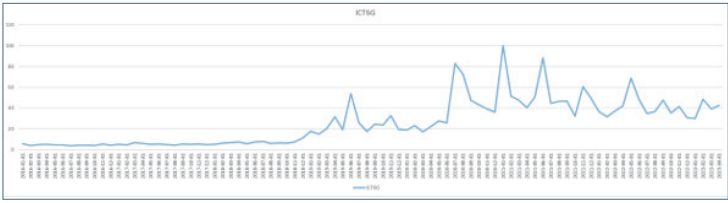
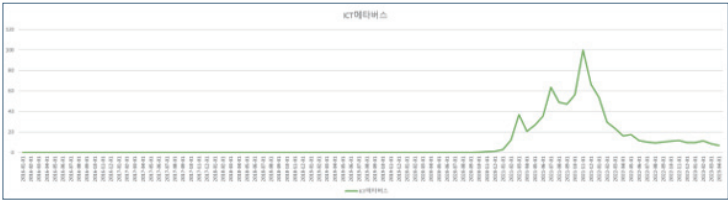

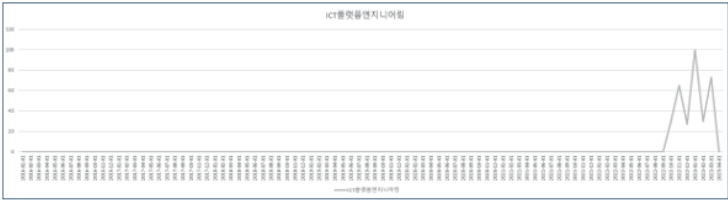
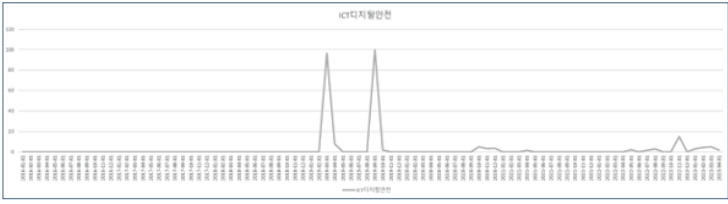
키워드명	검색 트렌드	평균	분산
ICT 네트워크		46.6	54.8
ICT 보안		56.1	49.9
ICT 로봇		44.6	126.1
ICT 우주		44.3	122.2
ICT 플랫폼		39.5	135.7

다음으로 “최근에 나타난 키워드”에는 2010년대에 검색량이 없었으나 2020년대에 검색량이 증가한 키워드들을 포함하였다. ‘ICT’ 키워드의 검색량보다 낮지만 분산은 높은 키워드들을 선정하였는데, 이는 신규 등장한 키워드의 특성상 평균 검색량이 낮아도 변동 가능성이 높기 때문이다<sup>2</sup>. “최근에 나타난 ICT 키워드”에는 ICT 6G, ICT 메타버스, ICT 모빌리티, ICT 플랫폼 엔지니어링, ICT 디지털 안전 이상 5개를 포함하였다.

<sup>2</sup> 검색 트렌드의 평균과 분산은 2016년 1월~2023년 4월 기간의 결과이다.

[그림 2] 키워드 검색 동향: 최근에 나타난 ICT 키워드

(단위: 개)

키워드명	검색 트렌드	평균	분산
ICT 6G		25.7	484.8
ICT 메타버스		8.8	338.7
ICT 모빌리티		9.4	123.4
ICT 플랫폼 엔지니어링		3.7	239.5
ICT 디지털 안전		2.9	218.0

최종적으로 18개 키워드를 사용하여 ICT 관련 뉴스를 웹 크롤링(Web Crawling) 방법으로 수집하였다. 수집 기간은 2008년 1월 1일~2024년 2월 29일까지이며, 중복을 제외한 약 86만 개의 기사를 활용하였다.

[표 1] 선정 키워드 리스트

ICT 산업 관련	과거부터 꾸준히 언급된 키워드	최근에 나타난 키워드
ICT 산업	ICT 네트워크	ICT 6G
ICT 기업	ICT 보안	ICT 메타버스
ICT 수출	ICT 로봇	ICT 모빌리티
ICT 품목	ICT 우주	ICT 플랫폼 엔지니어링
IT 산업	ICT 플랫폼	ICT 디지털 안전
IT 기업	-	-
IT 수출		
IT 품목		

## 2. 모델

ICT 뉴스경기지수는 기사에 대한 감성분석을 기반으로 작성하였다. 감성분석은 기사 내에서 긍정 및 부정 단어의 등장 빈도를 측정하고, 각 단어의 상대 빈도에 따라 기사를 긍정 혹은 부정으로 분류하는 것을 의미한다. 이렇게 각 기사의 긍·부정을 측정한 후 특정 기간 내 긍정, 부정 기사의 숫자를 측정하여 해당 기간의 뉴스경기지수를 계산하는 것이다.

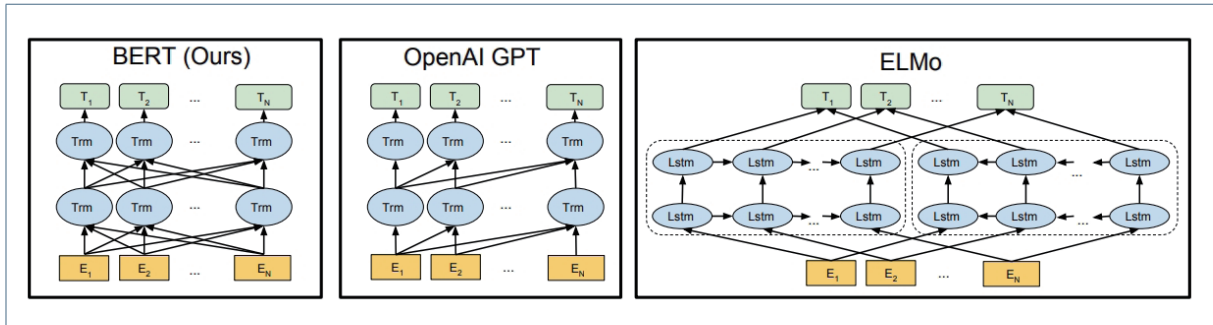
구체적으로 절차를 정리하면 다음과 같다. 먼저 기사를 문장 단위로 나눈 후, 각 문장별로 긍·부정 라벨링 작업을 진행하였다<sup>3</sup>. 문장의 긍·부정 라벨링에는 KoBERT(Korean BERT) 모델을 사용하였다. KoBERT는 SK Telecom에서 개발한 모델로 기반은 BERT(Bidirectional Encoder Representations from Transformers)에 두고 있다. BERT는 2018년에 구글이 발표한 NLP 방법의 일종으로, 대규모의 자연어 데이터 학습을 통해 문맥을 이해하며 단어 의미를 파악할 수 있도록 만든 사전 학습 모델이다. 특히 기존 단방향 모델과 달리 단어 확률 예측 시에 양방향을 사용하여 전체 문장을 파악할 수 있도록 만들었다는 점이 특징이다. 다만 BERT는 한국어 작업을 하는 경우 성능이 떨어지는 단점이 있었는데, KoBERT는 BERT의 한국어 성능 제약을 극복하고자 수백만 개의 한국어 문장을 학습시켜 만든 모델이다.

KoBERT를 통한 뉴스 기사 라벨링을 위해 먼저 사전 기반 라벨링 작업을 진행하였다. 사전 기반 라벨링은 단어별로 긍정, 중립, 부정으로 분류해 둔 단어 사전(Lexicon)을 이용하여 문장을 긍·부정으로 분류하는 것을 의미한다. 이를 위해 한국은행 뉴스심리지수에서 사용한 영역별 긍·부정 키워드(서범석 외, 2022), KNU 감성사전, 씬트렌드에서 ICT 기업, ICT 산업, ICT 수출과 관련한 기사 긍·부정 분석에서 등장한 단어들, 한국은행의 사전((김현중 외, 2019), (전종준 외, 2020)), KOSELF를 엮어서 사용하였다. 이렇게 구축한 단어사전을 통해 뉴스 기사를 문장별로 긍·부정으로 분류한 후, 이 데이터를 KoBERT로 학습시켜 결과를

<sup>3</sup> 하나의 기사에 긍·부정 라벨링을 측정할 수도 있고 기사를 여러 문장으로 나누어 긍·부정 라벨링을 측정할 수도 있다. 본고에서는 문장 단위에서 라벨링을 측정하는 방식을 사용하였는데, 이는 문장 단위 라벨링이 기사 단위 라벨링보다 안정적인 결과를 보였기 때문이다.

살펴보았다. 학습 결과 훈련 세트(Training set)에서는 98.6%, 테스트 세트(Test set)에서는 98.2%의 정확도를 기록하였고 예측 정확도(Accuracy Score)는 98.3%를 보였다.

[그림 3] BERT와 타 자연어 처리 모델 비교



자료: 인공지능 신문, ‘인공지능(AI) 언어모델 BERT(버트)는 무엇인가’ (<https://www.aitimes.kr/news/articleView.html?idxno=13117>)

### 3. 산출방법

이렇게 KoBERT를 통해 문장의 긍·부정을 분류한 후, 이 결과를 바탕으로 ICT 뉴스경기지수를 산출하였다. ICT 뉴스경기지수 산출은 한국은행의 뉴스심리지수 산출 방식(서범석 외, 2022)을 참고하였다.

$$\text{뉴스경기지수} = \left( \frac{X - \bar{X}}{S} \right) \times 10 + 100$$

$$X = \frac{\text{기간내 긍정문장수평균} - \text{기간내 부정문장수평균}}{\text{기간내 긍정문장수평균} + \text{기간내 부정문장수평균}}$$

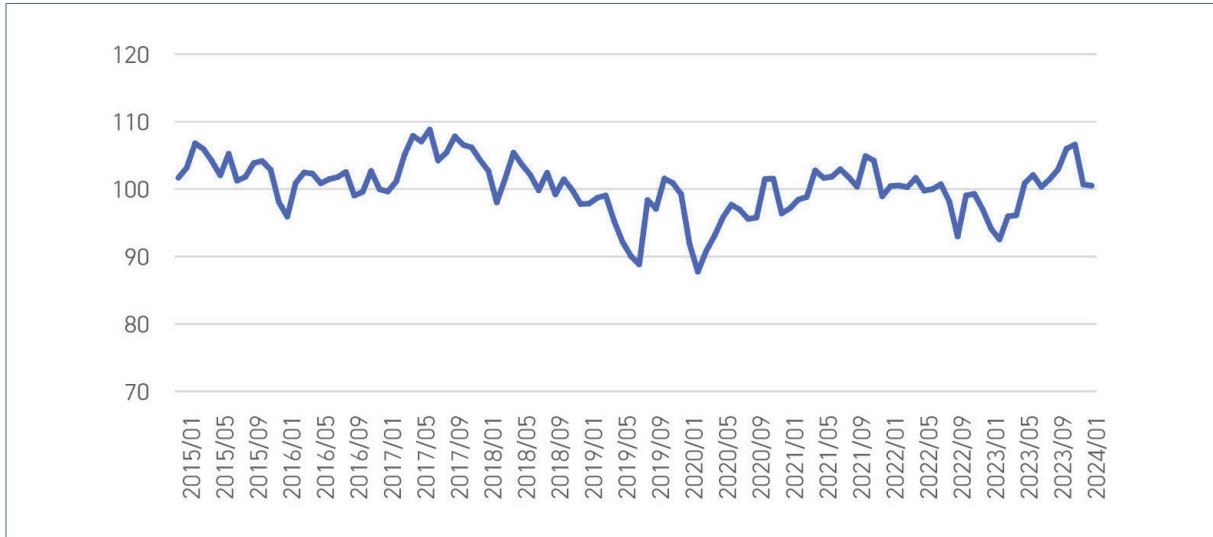
$$\bar{X} = \text{표준화 구간평균}, S = \text{표준화 구간 표준편차}$$

일별 ICT 뉴스경기지수는 특정 기간 내 긍정 문장과 부정 문장의 숫자를 비교하여 측정하게 된다. 기간은 7일을 설정하였다. 즉 특정 일자의 뉴스경기지수 값은 해당 일자 기준으로 최근 7일 동안 발표된 뉴스들을 기반으로 계산하는 것이다. 표준화 구간은 2008년 1월 1일~2023년 4월 30일로 하였다. ICT 뉴스경기지수는 100을 기준으로 하여, 100보다 클 경우 표준화 구간보다 긍정적인 기사 내용이 많았음을 의미한다. 반대로 100보다 작을 경우 부정적인 기사 내용이 많았다고 볼 수 있다.

이렇게 계산한 일별 ICT 뉴스경기지수를 월 단위로 평균치를 계산하여 월별 ICT 뉴스경기지수를 산출하였다. 일별 단위의 경우 변동성이 크고 기사 수가 부족한 과거 시점에서는 값을 측정하기 어려운 한계가 존재한다. 이러한 한계를 보완하고 지수의 안정성을 높이기 위해 월별 단위 결과를 활용하였다.

## 작성결과

[그림 4] ICT 뉴스경기지수 결과



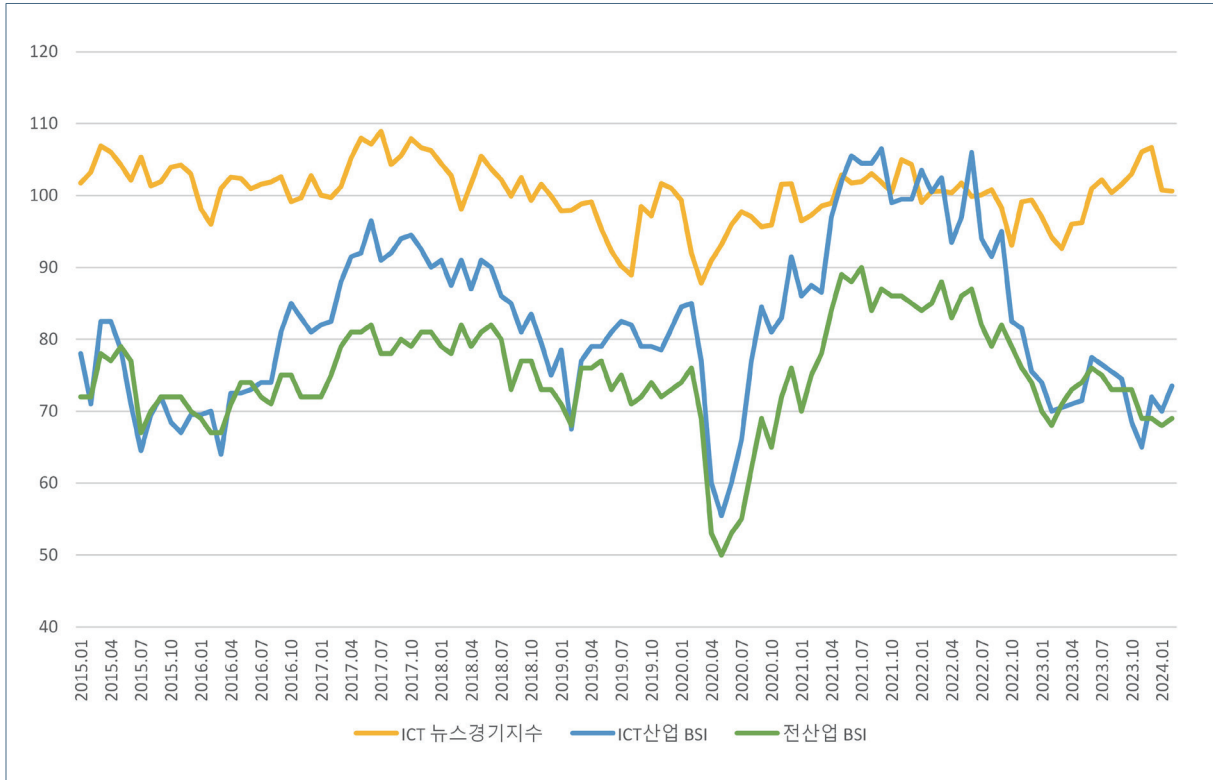
주: 2015년 1월~2024년 2월 기간의 ICT 뉴스경기지수 결과이다.

[그림 4]는 월별 ICT 뉴스경기지수 결과이다. 전체 결과 중 수집된 기사수가 충분하고 결과의 안정성이 높은 2015년부터의 결과를 제시하였다. 2015년부터 2018년까지는 전반적으로 ICT 산업에 긍정적인 뉴스가 많았음을 알 수 있다. 반대로 2019년부터는 ICT 산업 경기에 부정적인 내용들이 많았으며, 2020년 코로나19가 발생한 시점에 최저점을 찍고 있음을 알 수 있다. 이후 지수 값은 회복세를 보이다가 2022년 후반부터 다시 하락 추세에 직면했으나, 2023년 후반에는 다시 상승하고 있음을 볼 수 있다. 실제로 ICT 산업의 수출과 매출은 2015년~2018년에 꾸준히 증가하는 등 긍정적인 신호가 있었던 반면<sup>4</sup>, 2019년과 2020년에는 역성장을 겪는 모습을 보였다.

이 외에도 ICT 뉴스경기지수의 경제적 의미를 보다 다양하게 살펴보기 위해, 한국은행에서 발표하는 기업경기실사지수(BSI)와 ICT 뉴스경기지수의 상관관계를 분석하였다. BSI는 기업 대상으로 경기에 대한 판단, 전망 등을 설문조사한 결과이며 월별로 발표되고 있다. 기업들의 경기에 대한 주관적 판단을 대표하는 지표로서, ICT 뉴스경기지수가 제공하고자 하는 정보와 유사한 의미를 지닌다. 만약 ICT 뉴스경기지수가 BSI와 의미 있는 상관관계를 보인다면, ICT 뉴스경기지수를 통한 산업 현황 측정이 실제로 기업가들이 체감하는 경기 상황을 반영하는 측면이 있다고 볼 수 있을 것이다. ICT 뉴스경기지수는 ICT 산업에 대한 경기 측정이 목적임을 고려하여, 전산업 BSI에 더해 ICT 산업 BSI와도 비교하였다.

<sup>4</sup> 분석기간 ICT 산업의 생산(매출)액 정보는 과학기술정보통신부의 ICT실태조사 결과를 참고하였고, 동기간 수출금액은 과학기술정보통신부·정보통신기획평가원의 ICT 수출입통계를 참고하였다.

[그림 5] ICT 뉴스경기지수와 BSI



[표 2] ICT 뉴스경기지수와 BSI 간 상관관계

	ICT BSI	전산업 BSI
동기	27.2%	40.7%
1개월 선행	33.5%	45.5%
2개월 선행	30.4%	41.3%
3개월 선행	23.7%	35.9%

주: 2015년 1월~2024년 2월 기간의 ICT 뉴스경기지수와 각 BSI와의 상관관계를 비교하였다.

[그림 5]와 [표 2]는 ICT 뉴스경기지수와 전산업 BSI, ICT 산업 BSI와의 비교 결과이다. 결과를 살펴보면 동기 기준으로 ICT 뉴스경기지수와 ICT BSI는 27.2%, 전산업 BSI는 40.7%의 상관관계를 보이고 있다. ICT 뉴스경기지수가 실시간 뉴스를 사용하기에 상대적으로 속보성 있는 결과를 제공해준다는 점에 착안하여, ICT 뉴스경기지수의 선행 결과와 각 BSI와의 상관관계를 확인하였다. 분석 결과 1개월 선행에서 ICT BSI와 33.5%, 전산업 BSI와 45.5%의 상관관계를 보여 동기보다 더 높은 상관관계를 갖는 것으로 나타났다. 2개월 선행의 경우 ICT BSI와 30.4%, 전산업 BSI와 41.3%의 상관관계를 보였으며, 3개월 선행에서는 ICT BSI와 23.7%, 전산업 BSI와 35.9%의 상관관계를 보였다. 종합하면, 동기보다는 1개월 선행과 2개월 선행 지수에서



ICT BSI 및 전산업 BSI와의 상관관계가 높은 것을 확인할 수 있다. 이는 ICT 뉴스경기지수가 BSI를 선행하는 경향이 있어 선행지표로서의 가능성을 가지고 있음을 암시한다고 볼 수 있다.

## 결론

디지털화가 심화되면서 디지털 인프라에 기반한 다양한 빅데이터가 생성되고 있으며, 이러한 데이터를 활용하는 방법에 대한 연구도 다방면으로 진행되고 있다. 생성되는 데이터의 양과 정보, 신속성을 고려할 때 향후 디지털 기반의 빅데이터는 다양한 방면에서 활용될 가능성이 높다. 이러한 관점에서 ICT 뉴스경기지수의 작성은 뉴스 기반의 데이터 활용 방법을 모색하고 제시한다는 점에서 의미가 있다. ICT 뉴스경기지수는 ICT 산업의 경기를 신속하게 가늠할 수 있는 지표로 활용될 수 있으며, 산업 전망을 위한 주요 설명 변수로서의 역할을 할 수 있을 것으로 기대된다.



### ◎ 참고문헌

- 김현중, 임종호, 이혜영, 이상호(2019), 「온라인 뉴스 기사를 활용한 경제심리보조지수 개발」, 국민계정리뷰, 2019년 제2호, 한국은행.
- 서범석, 이영환, 조형배(2022), 「기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용」, 국민계정리뷰, 2022년 제1호, 한국은행.
- 손녕선, 진정민, 이은영, 김재민, 박재홍(2023), 「ICT 주요이슈 분석체계 구축」, 정책연구 23-28.
- 전종준, 안승환, 이문희, 황희진(2020), 「경제용어 감성 사전 구축방안 연구」, 국민계정리뷰, 2020년 제3호, 한국은행.
- 조수지, 김흥규, 양철원(2021), 「기업 재무분석을 위한 한국어 감성사전 구축」, 한국증권학회지 50.2, 135-170.
- Baker, S. R., Bloom, N., and Davis, S. J. (2016), 「Measuring economic policy uncertainty」, The Quarterly Journal of Economics, 131(4), 1593-1636.
- Scott, S. L. and Varian, H. R. (2014), 「Predicting the present with bayesian structural time series」 International Journal of Mathematical Modelling and Numerical Optimisation, 5(1-2), 4-23.
- Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022), 「Measuring news sentiment. Journal of econometrics」, 228(2), 221-243.
- 인공지능신문(2019. 1. 3.), '인공지능(AI) 언어모델 BERT(버트)는 무엇인가'

### ◎ 웹사이트

네이버 Datalab(<https://datalab.naver.com/keyword/trendSearch.naver>)

썸트렌드(<https://some.co.kr>)

전산업 BSI([https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT\\_512Y008&vw\\_cd=MT\\_ZTITLE&list\\_id=J1\\_301018\\_003&scrlId=&seqNo=&lang\\_mode=ko&obj\\_var\\_id=&itm\\_id=&conn\\_path=MT\\_ZTITLE&path=%252FstatisticsList%252FstatisticsListIndex.do](https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT_512Y008&vw_cd=MT_ZTITLE&list_id=J1_301018_003&scrlId=&seqNo=&lang_mode=ko&obj_var_id=&itm_id=&conn_path=MT_ZTITLE&path=%252FstatisticsList%252FstatisticsListIndex.do))

ICT 산업 BSI([https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT\\_512Y008&vw\\_cd=MT\\_ZTITLE&list\\_id=J1\\_301018\\_003&scrlId=&seqNo=&lang\\_mode=ko&obj\\_var\\_id=&itm\\_id=&conn\\_path=MT\\_ZTITLE&path=%252FstatisticsList%252FstatisticsListIndex.do](https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT_512Y008&vw_cd=MT_ZTITLE&list_id=J1_301018_003&scrlId=&seqNo=&lang_mode=ko&obj_var_id=&itm_id=&conn_path=MT_ZTITLE&path=%252FstatisticsList%252FstatisticsListIndex.do))

Google Trend(<https://trends.google.co.kr/trends>)