BOK 이슈노트

뉴스 텍스트를 이용한 경기 예측: 경제 부문별 텍스트 지표의 작성과 활용

서범석
한국은행 경제통계국 통계연구반 과장
Tel. 02-759-5253
bsseo@bok.or.kr

2022년 5월 16일

최근 코로나19, 우크라이나 사태 등 경제불확실성이 높아짐에 따라 신속한 경기 판단을 위한 다양한 데이터의 활용이 더욱 중요해지고 있다. 특히 뉴스 텍스트 데이터는 다양(variety)하고 방대(volume)한 정보를 신속히(velocity) 전달한다는 점에서, 경기 예측을 위한 새로운 빅데이터로 주목받고 있다.

본 논고는 뉴스 텍스트를 정량화하여 경제지표로 작성하는 새로운 텍스트 마이닝 방법론을 제시하였다. 이를 위해 2005년 이후 연간 약 100만 건(문장기준 1800만 문장)의 경제뉴스를 분석하여, 생산, 물가, 고용, 주가, 주택가격 등 경제적으로 관심이 높은 15개 부문의 뉴스 텍스트 기반 경제지표를 작성하였다.

작성한 텍스트 지표는 대부분 관련 공식 통계와 높은 상관관계를 보이며, 공식 통계 대비 0~9개월 선행하는 것으로 나타났다. 이는 뉴스 텍스트 기반 경제지표가 경기 예측을 위한 중요한 정보를 내포하고 있음을 보여준다.

뉴스 텍스트 기반 경제지표를 이용하여 경기 예측모형을 구축한 결과, 텍스트 지표를 예측모형에 반영할 경우 분기 GDP에 대한 예측 정확도가 유의미하게 높아지는 것을 확인하였다. 본 논고는 텍스트 지표의 특성을 반영할 수 있도록, 동적인자모형(Dynamic Factor Model, DFM) 기반의 선형모형과 인공신경망모형(Convolutional Recurrent Neural Network, CRNN) 기반의 비선형모형을 비교하여, 텍스트 지표 활용에 적합한 새로운 예측모형을 도출하였다. 새롭게 구축한 DFM 기반 경기 예측모형은 공식 통계가 발표되지 않은 상황에서도 텍스트 지표를 이용하여 GDP, CPI 등 63개 경제변수를 월별로 동시에 예측할 수 있도록 설계되었다.

뉴스 텍스트에는 다양한 전문가의 견해 · 전망 등 정성적 정보가 포함되어 있으므로, 이를 종합하고 정량화하여 경기 예측에 활용할 필요가 있다. 이러한 뉴스 텍스트의 정량적 활용은 신속하고 정확한 경기동향 파악 및 경기 예측에 유용하며, 정성적 방법으로 뉴스를 이용하는 것에 비해 휴먼 에러를 줄이는 데도 기여할 것으로 기대된다.

■ 논고 작성에 많은 도움을 주신 문혜정 통계연구반장과 유익한 논평을 주신 경제연구원 이승철 과장, 조사국 정원석 과장께 감사를 표합니다. 본문에 남아있는 오류는 저자의 책임임을 밝힙니다.

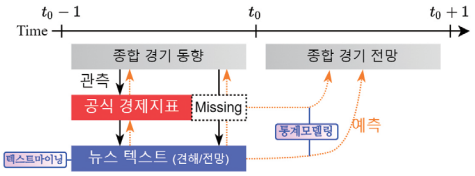
한국은행

Ⅰ. 검토배경

최근 코로나19, 우크라이나 사태 등 경제불확실성이 높아지면서, 신속한 경기 판단을 위한 빅데이터의 역할이 더욱 중요해지고 있다. 특히 최근 많은 연구들은 경기 판단을 위한 빅데이터로 뉴스 텍스트 데이터에 주목한다. 이는 뉴스가 다양(variety)하고 방대(volume)한 정보를 신속(velocity)하게 전달하기 때문이다. 본 논고는 뉴스 텍스트를 정량화하여 경제지표로 작성하는 새로운 방법을 제시하고, 작성한 뉴스 텍스트 기반 경제지표의 경기 예측효과를 실증분석을 통해 검증하고자 하였다.

각국 중앙은행 등 국내외 연구기관은 정책판단을 위한 신속한 경기 예측에 많은 노력을 기울이고 있다. 기존의 경기 예측은 주로 전문가들이 다양한 정보에 기반하여 종합적이고 정성적으로 평가하는 방식에 의해 이루어진다. 이에 반해, 경기 예측 관점에서 정량적 통계 모형은 제한적으로 이용되고 있다. 모형을 이용한 신속한 경기 예측이 어려운 이유는 기본적으로 통계 모형을 학습시키기 위해 필요한 충분한 정량적 정보를 장기 시계열로 확보하는 것이 어렵기 때문이다. 여기에는 구체적으로 다음의 세 가지 이유가 있다. 첫째, 금융부문을 제외한 실물부문의 경우 일, 주 단위의 고빈도 경제지표가 거의 전무하다. 둘째, 정량지표로 가장 중요하게 활용되는 대부분의 공식통계는 대상시점과 공표시점에 차이가 발생한다. 즉, 경제환경이 급변하는 상황에서는 공식통계의 활용이 어려워진다. 셋째, 정량 지표가 반영하지 못하는 정성적 정보는 통계 모형에 추가하는 것이 어렵다. 경기 상황에 따라 경제지표의 중요도가 달라지는데 이를 적절히 모형에 반영하는 것은 어려운 문제이다.

〈그림 1〉 경기 예측을 위한 뉴스 텍스트 데이터의 활용 예시



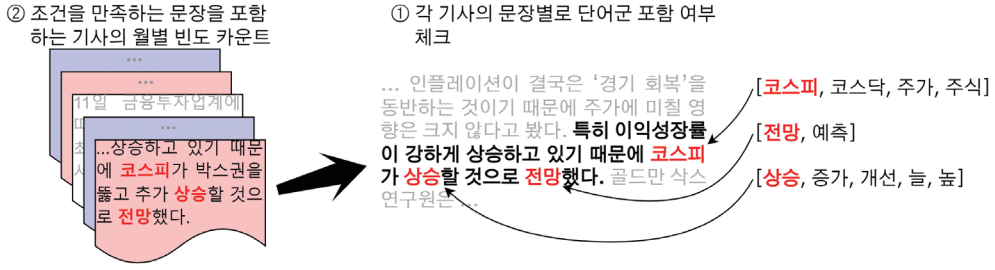
이러한 문제들로 인해 통계 모형에 의한 경기 예측은 주로 보조적 수단으로 활용되어 왔다. 그러나 최근 빅데이터 분석 기술의 발전은 경기 예측을 위한 다양하고 새로운 시도를 가능케 한다. 특히 Nowcasting이라고 불리는 실시간 경기 예측모형 개발을 위한 연구가 활발하게 진행되고 있고, 이러한 연구의 중심에서 뉴스 텍스트 등 새로운 빅데이터를 개발하는 것이 중요한 이슈로 부각되고 있다.

Thorsrud et al.(2020), Bybee et al.(2021) 등은 뉴스 텍스트에서 경제 정보를 추출하기 위해 토픽모델링 방식을 이용하였다. Thorsrud et al.(2020)은 비지도(unsupervised) 방식으로 뉴스 기사를 여러 개의 토픽으로 분류하고, 분류한 토픽의 시계열 변화를 예측모형에 추가하여 뉴스 정보를 모형에 반영하였다. 그러나 토픽모델링 방식은 추출한 토픽의 정의가 모호하고, 모형의 파라미터에 따라 토픽이 크게 달라지는 등 불안정한 모습을 보인다.

본 논고에서는 기존 연구와 달리, 경제적으로 관심이 높은 생산, 고용, 물가, 주가, 주택가격 등 15개 부문을 직접 선정하고, 각 부문

1) 편의를 위해 '뉴스 텍스트 기반 경제지표'에 대한 명칭으로 '텍스트 지표'를 혼용하여 사용하였다.

〈그림 2〉 뉴스 텍스트 기반 경제지표 작성 예시



별로 대상지표를 정하여, 이들 지표의 흐름을 예측하는 텍스트 지표를 각각 작성하였다. 이렇게 작성한 뉴스 텍스트 기반 경제지표[1]는 토픽모델링을 사용한 방식에 비해 대상 주제가 명확하여 검증이 가능하고, 예측모형에 반영할 경우 공식 통계가 발표되지 않은 상황에서 대상 부문의 대체 지표로 사용 가능하다는 장점이 있다.

본 논고는 텍스트 지표 작성 방법으로 해석이 용이하고 문어체 텍스트 분석에 우수한 사전접근법(lexical approach)[2] 방식을 문장별로 적용하는 새로운 방법을 제시하였다. 기존 연구는 문장이 아닌 기사를 기준으로 뉴스 텍스트를 분석하였으나, 보통 하나의 기사에 여러 의견이 나타나는 경우가 많으므로 뉴스 텍스트를 문장별로 검토하는 것이 정밀한 지표작성에 유리하다.

또한 본 논고는 노이즈가 많이 포함될 가능성이 있는 뉴스 텍스트 기반 경제지표의 특성을 고려하여, 동적인자모형(Dynamic Factor Model, DFM) 기반의 선형모형과 인공신경망모형(neural networks) 기반의 비선형모형을 비교 · 검토하였다. 이를 통해, 텍스트 지표의 특성을 잘 반영할 수 있는 새로운 경기예측모형을 제시하였다.

본 논고는 다음과 같이 구성하였다. 2장에서 뉴스 텍스트를 이용한 일반적인 지표 작성 방법과 경기 예측모형 구축 방법에 대해 살펴보았다. 이어지는 3장에서는 새로운 부분별 뉴스 텍스트 기반 경제지표 작성 방법과 작성 결과를 제시하였다. 4장에서는 텍스트 지표의 특성을 고려한 경기 예측모형 구축 방법을 제시하였고, 5장에서 텍스트 지표를 반영한 경기 예측모형의 추정 결과를 살펴보았다. 마지막으로 6장에서 본 논고의 시사점 및 향후 발전방향을 정리하였다.

II. 뉴스 텍스트 데이터의 활용

1. 텍스트를 이용한 지표 작성

뉴스 텍스트를 가공하여 텍스트 지표를 작성하기 위해서는 텍스트를 통하여 추출하고자 하는 주제와 그 추출 방법을 명확히 할 필요가 있다. 추출하고자 하는 주제가 너무 광범위할 경우 정보의 가치가 떨어질 수 있는 반면, 주제가 너무 좁은 범위라면 주제와 관련된 정보가

2) 미리 정해 놓은 단어의 포함 여부를 기준으로 텍스트를 분석하는 방식이다.

텍스트에 충분히 나타나지 않을 가능성이 있기 때문이다.

텍스트에서 얻고자 하는 주제를 결정하였다면 그에 맞는 추출 방법을 정하여야 한다. 텍스트에서 정보를 추출하는 방법은 크게 기계학습 등의 통계 모델링을 이용하는 방식(stochastic approach)과 사람이 직접 정의한 추출조건을 이용하는 방식(rule-based approach)이 있다. 통계 모델링 방식은 다시, 추출하고자 하는 텍스트의 예시 문장을 이용하여 모형을 학습시키는 지도학습 방법과, 예시 문장 없이 텍스트 간의 분포차이를 이용하여 모형을 학습시키는 비지도학습 방법으로 나뉜다. 어떤 방법을 사용할 것인지는 작성하고자 하는 텍스트 지표의 주제 및 텍스트 데이터의 성격에 따라 달라진다. 일반적으로 텍스트 데이터가 구어체에 가까운 비정형 텍스트일수록, 그리고 추출하고자 하는 주제가 복합적이고 추상적일수록 통계 모델링 방식이 우수한 성능을 보인다.

예를 들어 블로그의 글을 이용하여 경제 심리를 포착하는 지표를 작성하고자 한다면 문맥을 잘 학습하는 것으로 알려진 지도학습 모형을 사용하는 것이 바람직하다. 다만 지도학습 방식을 사용하기 위해서는 잘 정비된 예시 문장을 사람이 평가하여 모형에 학습시키는 과정이 필요하고, 이 과정에서 많은 비용이 발생한다. 한편, 비지도학습 방식은 인터넷 포탈 검색문장의 경제 관련 토픽 변화 등을 파악하는 데 사용될 수 있다. 사람들이 관심을 갖는 토픽이 어떻게 변화하는지를 파악하기 위해, 검색문장에 나타나는 텍스트간의 관계를 통계 분포로 근사하고, 이를 통해 텍스트를 주제별로 나누는(topic clustering) 것이 가능하다.

통계 모델링을 사용하는 대신 사람이 직접 추출하고자 하는 조건을 명시하여 텍스트 지표를 작성하는 방법(rule-based approach)도 가능하다. 이 경우, 추출하려는 주제와 관련된 단어들을 미리 정해 놓고, 이들 단어가 텍스트에 포함되는지 여부를 체크하여 지표를 작성할 수 있다. 미리 정해 놓은 단어를 이용하여 텍스트를 분류하는 방식을 특별히 사전접근법(lexical approach)이라고 한다. 사전접근법 방식은 문어체 텍스트 분석에 우수하며 해석이 직관적인 장점이 있다.

기존 연구들은 뉴스 텍스트를 이용하여 지표를 작성하는 경우 목적에 맞게 다양한 방식의 접근법을 사용하였다. 샌프란시스코 연준은 사전접근법 방식으로 뉴스 기사를 긍정기사와 부정기사로 분류한 뒤 이 둘의 상대비율을 지수화한 뉴스심리지수를 개발하여 공개하고 있다. Baker et al.(2016)은 경제, 정책, 불확실성 등 세 영역의 단어군을 이용하여 사전접근법 방식으로 경제불확실성 지수를 작성하는 방법을 제안하였다. Bybee et al.(2020)은 토픽모델링을 이용하여 분석한 경제 뉴스 토픽 변화가 경기 변동(business cycle)을 설명하는 데 도움이 된다고 주장하였다. IMF의 Caldara et al.(2022)은 지정학적 리스크를 평가하기 위해 사전접근법 방식으로 리스크 지표를 작성하여 유의성을 검증한 바 있다. 또한 한국은행은 지도학습 방식으로 뉴스심리지수를 개발하여 실험적 통계[3]로 공개하고 있다. 한국은행의 뉴스심리지수 작성 방법은 서범석 외(2022)가 자세히 소개하고 있다.

3) 새로운 유형의 데이터를 활용하거나 새로운 방식을 적용하여 실험적으로 작성한 통계로 국가통계는 아니다.

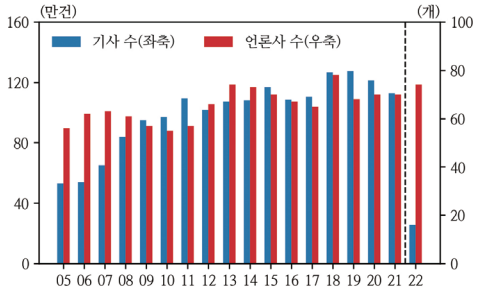
2. 텍스트를 이용한 경기 예측

최근 많은 연구들이 텍스트 지표에 주목하는 이유는 무수한 정보들이 텍스트 형태로 전달되기 때문이다. 특히 뉴스 텍스트는 다양(variety)하고 방대(volume)한 정보를 매우 신속(velocity)하게 전달한다는 특징이 있다. 따라서 텍스트 데이터를 예측모형에 적절히 반영할 경우 보다 빠른 경기 판단이 가능할 것이라 기대할 수 있다. Babii et al.(2021), Bybee et al.(2021), Thorsrud et al.(2020) 등은 경기 예측모형의 구축에 있어서 텍스트 데이터가 매우 중요함을 강조하였다.

또한 뉴스 텍스트는 통계치를 언급하는 등의 사실 정보 이외에도 전문가 인터뷰 등을 통하여 특정 정보에 대한 견해나 전망도 함께 전달한다. 따라서 텍스트에 나타나는 정보는 공식 통계와 달리 유연하게 변화하며, 사람들의 관심도 또한 반영한다. 예를 들어 2021년 상반기 코로나19 확진자수는 민간소비 예측을 위해 매우 중요한 지표로 활용되었는데, 2022년 들어서는 위드코로나 정책 시행과 함께 동 지표가 소비에 미치는 영향이 낮아졌다. 따라서 코로나19 확진자수를 정량지표로 활용하기 위해서는 그동안의 활용방식을 수정하는 것이 불가피해진다. 이에 반해 뉴스 텍스트에 나타난 코로나19에 대한 우려는 사람들의 관심도와 함께 2022년 들어 감소하였으며, 확진자수 통계보다 코로나19의 경제적 영향을 잘 보여주는 지표로 활용이 가능하였다. 다시 말해, 뉴스 텍스트 데이터는 견해, 전망 등 정성적 정보를 효과적으로 반영한다. 따라서 뉴스 텍스트를 정량화하여 예측모형에 반영할 경우 사람의 개입없이 정성적 정보를 모형에 고려할 수 있는 장점이 있다.

다만 텍스트 지표는 서베이를 통해 작성하는 공식 통계에 비해 변동성이 심하고 노이즈가 많이 포함될 가능성이 있으므로, 이 점에 주의하여야 한다. 공식 통계는 작성 주제가 명확하게 정의되고 표본 관리를 통하여 품질이 엄격히 통제되지만, 텍스트 지표는 텍스트에 나타나는 사람의 언어를 종합한 지표이므로 주제가 다소 불명확하고 작성 의도와 다른 언급 내용도 지표에 포함될 가능성이 있다.

〈그림 3〉 경제분야 뉴스기사 및 언론사 수



〈표 1〉 텍스트 지표와 공식 통계의 정보 비교

텍스트 지표	공식 통계
● 사실 정보 ● 전문가의 견해 및 전망 ● 언론사의 해석 ● 언론과 사회의 관심 ⋮	● 사실 정보

Ⅲ. 부문별 뉴스 텍스트 기반 경제지표의 작성

1. 한국어 뉴스 텍스트 데이터

뉴스 텍스트 기반 경제지표의 작성을 위해 2005년부터 2022년 3월까지 인터넷 포털

사이트에 게재된 경제분야 뉴스 기사를 수집하였다. 텍스트 지표는 노이즈를 포함할 가능성이 높기 때문에 최대한 많은 표본을 이용하는 것이 유리하다. 본 연구에서 사용한 기사는 약 70여개 언론사의 평일 기준 일평균 약 4000건(연평균 약 1백만건)의 기사이며, 문장 단위로 환산할 경우 연간 약 1천8백만 문장이다.

2. 뉴스 텍스트 기반 경제지표의 작성

본 논고에서는 경기 예측을 위한 텍스트 지표의 주제로 〈표 2〉와 같이 15개 부문의 주요 거시변수 및 일부 산업 관련 미시변수를 선정하였다. 여기서 텍스트 지표의 대상변수들은 국내총생산(GDP) 예측에 주요하게 영향을 미칠 것으로 판단되는 변수들로 구성하였다. 이는 텍스트 지표가 공식 통계에 선행하며 공식 통계의 대체 변수로 사용 가능하므로, 공식 통계가 발표되지 않은 상황에서 경기 예측모형이 텍스트 지표를 대신 이용하도록 하기 위함이다.

부문별 텍스트 기반 경제지표는 작성하고자 하는 주제가 명확하고, 이용하는 뉴스 데이터가 정형화된 문어체이기 때문에 사전접근법 방식으로 작성하는 것이 바람직하다. 지도학습 방식을 이용하면, 예시 문장 작성에 따른 시간과 비용이 소요되고, 통계 모델링 과정에서 불필요한 문장 요소를 학습함에 따라 오차가 증가하는 과적합(overfitting) 문제가 발생할 우려가 있다. 한편 토픽모델링 등의 비지도학습 방식은 추출하려는 주제가 명확하지 않을 때 사용하며, 구체적인 주제의 지표 작성 방법으로는 적절하지 않다.

본 연구에서는 15개 부문별 텍스트 지표에 더하여, 경제 전반에 대한 심리 및 불확실성을 나타내는 뉴스심리지수(News Sentiment Index, NSI)와 경제불확실성지수(Eonomic Policy Uncertainty, EPU)도 함께 이용하였다. 뉴스심리지수는 한국은행 경제통계시스템(ECOS)에 공개된 지수를 이용하였다. 뉴스심리지수는 지도학습 방법을 적용하여 개발되었으며, 자세한 작성 방법은 서범석 외(2022)에 기술되어 있다. 경제불확실성지수는 Baker et al. (2016), Lee et al.(2020)의 방법론을 바탕으로 국내 사정에 맞는 지수를 직접 개발하여 이용하였다. 경제불확실성지수는 주요 언론사의 뉴스 기사만 선별한 뒤, 선별된 뉴스 기사에 사전접근법 방식을 적용하여 개발하였다. 경제불확실성지수의 자세한 작성 방법은 〈부록 2〉에 기술하였다.

부문별 뉴스 텍스트 기반 경제지표는 특정 단어의 포함 여부를 문장 단위에 적용한 뒤, 해당 문장의 포함 여부를 기준으로 뉴스 기사를 분류하여 작성하였다. 이는 특정 단어의 포함 여부를 기사 단위에 적용하여 뉴스 기사를 분류한 기존 연구들과 다른 점이다. 텍스트를 기사 단위로 분석할 경우, 한 기사에 여러 의견이 혼재되어 나타나는 경우가 많기 때문에 텍스트 지표의 노이즈가 증가하고 공식 통계와의 상관관계가 하락하는 것으로 나타났다.

구체적으로 살펴보면, 부문별 뉴스 텍스트 기반 경제지표는 분야별 단어군을 사전에 정의하고, 이들 단어군을 포함하는 문장이 등장한 기사들의 기간중 상대빈도수를 계산하여 작성하였다. 즉, t시점에 발간된 N개의 뉴스 기사 집합을 $\Omega_t = \{A_1, \cdots, A_N\}$라고 하고, 각 뉴스 기사 A_i를 문장 S_{im}의 집합 $A_i = \{S_{i1}, \cdots, S_{iMi}\}$로 표현하자. 그러면 t시점의 뉴스 텍스트 기반 경제지표 R_t는 특정 분야의 단어군 $W^{(k)}$ =

$\{W_1^{(k)}, \cdots, W_{Lk}^{(k)}\}$, $(k = 1, \cdots, K)$와 그 단어군 k에 속하는 단어 $w_j^{(k)}(j = 1, \cdots, l_k)$에 대하여 다음과 같이 나타낼 수 있다.

$$R_t = \frac{\sum_{i=1}^{N} \hat{A}_i}{N},$$

$$\hat{A}_i = \bigvee_{m=1}^{M_i} C_{im},$$

$$C_{im} = \prod_{k=1}^{K} \bigvee_{l=1}^{L_k} I_{S_{im}}(w_l^{(k)}).$$

여기서 $I_S(w) = \begin{cases} 1, & \text{if } w \in S \\ 0, & o.w. \end{cases}$는 지시함수(indicator function),

$\bigvee_{m=1}^{M} C_m = \max(C_1, \cdots, C_M)$는 최대값 함수이다.

따라서 R_t는 0과 1사이의 실수값을 갖는다.

다만, 추출하고자 하는 텍스트 지표의 대상 주제를 긍정 및 부정 논조로 구분할 필요가 있는 경우에는, 긍정 단어군을 기준으로 추출한 긍정 지표 R_t^{pos}에서 부정 단어군을 기준으로 추출한 부정 지표 R_t^{neg}를 차감하여 뉴스 텍스트 기반 경제지표 R_t를 작성하였다.

$$R_t = R_t^{pos} - R_t^{neg}.$$

예를 들어 주가전망 텍스트 지표는 전체 표본기사를 대상으로 (코스피, 코스닥, 주가, 주식), (전망, 예측), (상승, 증가, 개선, 늘, 높) 등 3개 단어군 내에서 각각 하나 이상의 단어를 포함하는 문장이 기사 본문에 존재하는지 여부를 체크한 뒤, 이를 만족시키는 긍정기사의 월별 상대빈도수에서 (코스피, 코스닥, 주가,

〈표 2〉 부문별 텍스트 지표 작성을 위한 단어군[1)2)3)]

부문		작성 방법
산업	생 산	|(생산)&(상승,급등,증가,개선,가속,늘)| − |(생산)&(하락,급락,감소,악화,둔화,줄)|
	선 박	|(선박)&(수주)&(상승,급등,증가,개선,늘)| − |(선박)&(수주)&(하락,급락,감소,악화,줄)|
	자동차	|(자동차,승용차)&(상승,급등,증가,개선,가속)| − |(자동차,승용차)&(하락,급락,감소,악화,둔화)|
	반도체	|(반도체)&(상승,급등,증가,개선,가속)| − |(반도체)&(하락,급락,감소,악화,둔화)|
설비투자		|(설비투자,R&D)&(상승,급등,증가,개선,가속,늘,확대)| − |(설비투자,R&D)&(하락,급락,감소,악화,둔화,줄,감축)|
주택건설		|(주택,아파트)&(건설,건축,착공,시공)|
고용	실 업	|(실업)&(상승,증가,늘,악화)|
	채 용	|(채용,고용)&(상승,증가,개선,늘)| − |(채용,고용)&(하락,감소,악화,줄)|
	취 업	|(취업,구직)&(상승,증가,늘)| − |(취업,구직)&(하락,감소,줄)|
도 소 매		|(도매,소매,도소매)&(상승,급등,증가,개선,가속,늘)| − |(도매,소매,도소매)&(하락,급락,감소,악화,둔화,줄)|
정부지출		|(정부)&(지원,보조,지출)|
물가전망		|(물가)&(전망,예측,예상)&(상승,급등,올라,높)| − |(물가)&(전망,예측,예상)&(하락,급락,내려,낮)|
주가전망		|(코스피,코스닥,주가,주식)&(전망,예측)&(상승,증가,개선,늘,높)| − |(코스피,코스닥,주가,주식)&(전망,예측)&(하락,감소,악화,줄,낮)|
주택가격 전 망		|(주택,아파트)&(가격,매매가,전세가,분양가)&(전망,예측)&(상승,급등,확대,개선,가속,높)| − |(주택,아파트)&(가격,매매가,전세가,분양가)&(전망,예측)&(하락,급락,축소,악화,둔화,낮)|
세계교역		|(세계,글로벌)&(교역,무역,수출,수입)&(상승,급등,증가,개선,가속,늘,확대)| − |(세계,글로벌)&(교역,무역,수출,수입)&(하락,급락,감소,악화,둔화,줄,감축)|

주: 1) (a, b, …) 는 하나의 문장 안에서 나열한 원소(a, b, …)중 하나 이상의 원소가 포함되는 경우를 의미
2) a & b & … 는 하나의 문장 안에서 나열한 원소(a, b, …) 모두가 포함되는 경우를 의미
3) |A| 는 A의 조건을 만족하는 문장을 포함하는 기사의 수를 의미

주식), (전망, 예측), (하락, 감소, 악화, 줄, 낮) 등의 단어군으로 계산한 부정기사의 월별 상대빈도수를 차감하여 작성하였다.

텍스트 지표는 기간별 상대빈도수를 이용하여 작성하므로 일, 주, 월 등 대상기간을 달리하여 수시로 작성 가능하며, 고빈도 지표로 활용할 수 있다.

3. 텍스트 지표의 유용성 검토 결과

작성한 부문별 텍스트 지표를 관련 공식 통계와 비교한 결과, 텍스트 지표가 공식 통계에 선행하며 높은 상관관계를 보이는 것으로 나타났다. 〈표 3〉을 보면 텍스트 지표가 각 부문의 관련 공식 통계와 0~9개월 선행시점에서 0.35~0.73의 비교적 높은 상관관계를 보이는 것을 알 수 있다. 또한 Granger 인과성 검정 결과를 통해서도 텍스트 지표가 1~7개월 선행시점에서 대부분 유의한 인과성을 보이는 것으로 나타난다.

특히 '전망' 및 '예측' 단어를 포함하여 작성한 물가전망, 주가전망, 주택가격전망 지표의 경우 선행시점이 3~9개월로 높게 나타난다. 이는 텍스트 지표가 뉴스 본문에 나타나는 전문가의 견해 및 전망을 반영한 결과이다. 즉, 물가의 경우 약 반기 앞을 전망한 전문가 인터뷰가 텍스트 지표에 반영된 것으로 해석이 가능하며, 주가의 경우 약 3개월, 주택가격의 경우 약 9개월 앞에 대한 전망이 뉴스에 많이 나타났을 것으로 유추할 수 있다. 이들 선행시점은 각 텍스트 지표의 주제를 고려할 때 매우 합리적이다. 주가의 경우 관심이 높은 전망시점이 비교적 짧은

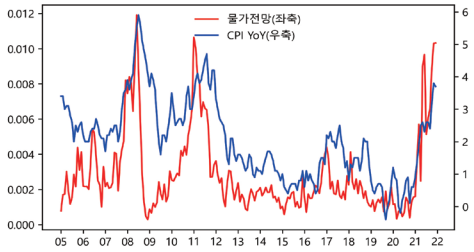
〈표 3〉 부문별 텍스트 지표와 관련 공식 통계의 비교

부 문		관련 공식 통계	공식통계 대비 상관시차/상관계수[1]	Granger 검정 선행시차/유의수준[2]
산업	생 산	선행종합지수 순환변동치 (통계청)	-1 / 0.69	-2~-5 / ***
	선 박	선박수주량 CGT (클락슨 리서치)	0 / 0.43	-1 / ***
	자동차	자동차판매대수 (한국자동차산업협회)	-1 / 0.50	-1~-3 / ***
	반도체	반도체 ICT수출액 (정보통신기획평가원)	-2 / 0.62	-1 / *
설비투자		설비투자지수 (통계청)	-1 / 0.47	-1~-4 / ***
주택건설		주택착공실적 (국토교통부)	0 / 0.48	유의하지 않음
고용	실 업	경제활동인구조사 실업률 (통계청)	-1 / 0.35	유의하지 않음
	채 용	경제활동인구조사 고용률 (통계청)	-1 / 0.61	-3~-4 / ***
	취 업	경제활동인구조사 취업자수 (통계청)	0 / 0.57	-1~-3 / ***
도 소 매		도소매 서비스업생산지수 (통계청)	-1 / 0.41	-1~-3 / ***
정부지출		통합재정수지 (기재부)	-3 / 0.66	-6~-7 / **
물가전망		소비자물가지수 (통계청)	-5 / 0.73	-2~-3 / **
주가전망		코스피지수 (한국거래소)	-3 / 0.65	-2 / **
주택가격전망		주택매매가격지수 (KB부동산)	-9 / 0.47	유의하지 않음
세계교역		International Merchandise Trade Volume (OECD)	-3 / 0.65	-2 / **
뉴스심리지수[3]		경제심리지수 (한국은행)	-2 / 0.61	-3 / ***
경제불확실성지수[3]		VKOSPI (한국거래소)	0 / 0.56	-1 / ***

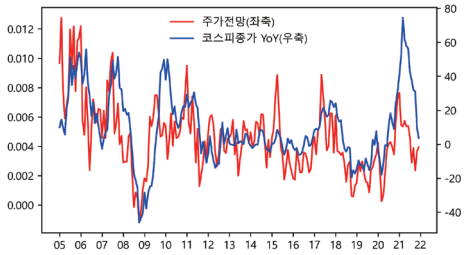
주: 1) 음수(-)는 텍스트 지표가 관련 공식 통계에 선행, 0은 동행함을 의미하며 동행 및 선행시차만을 고려했을 때의 최대 상관계수와 그 시차
2) *, **, ***는 각각 Granger 검정 통계량이 $\alpha = 0.10, 0.05, 0.01$ 유의수준에서 유의함을 의미
3) 뉴스심리지수와 경제불확실성지수는 본 연구에서 제시한 방법이 아닌 선행 연구를 통해 작성한 지표를 이용

〈그림 4〉 부문별 텍스트 지표와 관련 공식 통계 추이 예시

(물가전망 텍스트 지표와 소비자물가지수)



(주가전망 텍스트 지표와 코스피지수)



약 3개월 정도이며, 물가의 경우 약 3~6개월, 주택가격의 경우 약 반기에서 1년 정도인 것으로 보는 것이 자연스럽기 때문이다.

따라서 부문별 텍스트 지표는 선행지표로서 매우 유용한 것으로 판단되며, 각 부문의 동향 파악을 위해 개별적으로 활용될 수 있다. 〈그림 4〉는 물가전망 및 주가전망 텍스트 지표의 추이를 소비자물가지수 및 코스피지수의 전년동기대비 증가율과 함께 그려본 것이다. 〈그림 4〉를 보면 텍스트 지표가 관련 공식 통계의 흐름을 매우 잘 포착하는 것을 확인할 수 있다. 보다 상세한 부문별 텍스트 지표와 공식통계의 비교 결과는 〈부록 1〉에 기술되어 있다.

Ⅳ. 뉴스 텍스트를 이용한 경기 예측모형의 구축

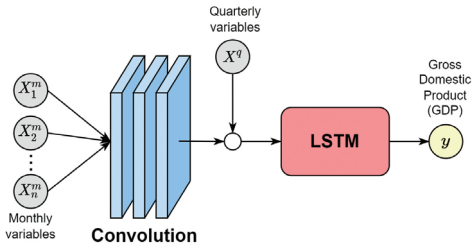
1. 경기 예측모형의 선택

텍스트 지표를 활용한 종합적인 경기 예측을 위해서는 경기 예측모형을 이용할 필요가 있다. 이때 텍스트 데이터의 특성을 고려한 적절한 예측모형의 선택이 중요하다. 텍스트 지표는 공식 통계에 비해 노이즈가 크게 나타나며 정상(stationary) 시계열로 변환하는 작업이 쉽지 않다[4]. 텍스트 지표는 언론에 나타나는 관심을 정량화한 지표이기 때문에 상승 추세 등의 명백한 비정상성(non-stationarity)은 발견되지 않으나, 변동성이 크게 나타나고 노이즈와 경제 시그널을 구분하는 것이 어려운 문제가 있다. 따라서 선형모형을 이용하면 정상성 가정이 완벽히 충족되지 않아 예측오차가 증가할 가능성이 있고, 비선형모형을 이용하면 노이즈를 적합하여 과적합(overfitting) 오차가 증가할 가능성이 있다. 이러한 점을 고려하여 본 연구에서는 선형모형과 비선형모형을 모두 이용하여 텍스트 지표의 예측 효과를 살펴보았다. 선형모형으로는 경기 예측모형으로 가장 널리 활용되는 동적인자모형(Dynamic Factor Model, DFM)을 이용하였고, 비선형모형으로는 인공신경망 기반의 Convolutional Recurrent Neural Network (CRNN)를 경제시계열 데이터에 알맞도록 구성하여 이용하였다.

먼저 비선형모형으로 고려한 CRNN은 64개의 Convolutional 필터를 3개월 단위로 적용한 뒤 이를 32개 유닛으로 구성한 LSTM

4) 통계 모델링을 이용하여 시계열 자료를 분석할 때 정상(stationary) 시계열이 아닌 경우 일반화(regularized 또는 out-of-sample) 예측오차가 증가하는 문제가 발생한다.

〈그림 5〉 CRNN 구조도



(Long Short-Term Memory) 레이어에 입력하여 구성하였다. Convolutional 레이어는 관측변수간 관계를 직접 추정하는 데 유용하고, 이어지는 LSTM 레이어가 이들 관계의 시계열 추세를 모형에 반영하게 된다. 특히 본 연구에서는 Convolutional 필터를 이용하여 월 및 분기 단위의 시계열 자료를 분기 단위로 집계함으로써, 혼합주기 시계열 자료를 CRNN 모형이 이용할 수 있도록 모형을 구성하였다.

다음으로 선형모형으로는 Nowcasting 모형으로 가장 널리 활용되는 DFM을 이용하였다. DFM은 다변량 관측변수들을 적은 수의 잠재요인(latent factor)으로 추정하며 고차원 시계열 변수들의 동조적 변동성을 잘 적합하는 것으로 알려져 있다. DFM 모형의 구성을 간략히 살펴보면, 다음과 같이 관측변수에 대한 선형모형과 잠재요인에 대한 선형모형으로 구성된다.

$$X_t' = \lambda(L) f_t + e_t$$
$$f_t = \psi(L) f_{t-1} + \eta_t$$

여기서 X_t'는 t시점의 p개 관측변수, f_t는 q개 잠재요인, e_t 및 η_t는 t시점의 idiosyncratic disturbance 및 factor innovation, L은 시차연산자, $\lambda(L)$ 및 $\psi(L)$은 각각 두 식의 시차 다항행렬식을 의미한다.

〈표 4〉 DFM의 요인별 관측변수[1)]

요인	관측변수
종합1	모든 변수
종합2	모든 변수
생산	제조업출하지수, 제조업재고지수, 서비스업생산지수, 광공업생산지수, 전산업매출BSI, 전산업업황실적BSI, 제조업수출BSI, 제조업가동률BSI, 제조업신규수주BSI, 제조업내수판매BSI, 제조업업황실적(SA)BSI, GDP(분기), **생산(뉴스)**, **선박(뉴스)**, **자동차(뉴스)**, **반도체(뉴스)**
대외	수출, 수입, GDP수출(분기), **세계교역(뉴스)**
투자	설비투자지수, GDP설비투자(분기), GDP건설투자(분기), **설비투자(뉴스)**, **주택건설(뉴스)**
고용	실업률, 고용률, 구인배수, 취업률, 취업자수, **실업(뉴스)**, **채용(뉴스)**, **취업(뉴스)**
소비	소매판매액지수, 현재경기판단CSI, 소비자심리지수, GDP민간소비(분기), **도소매(뉴스)**
물가	수출물가지수, 수입물가지수, 소비자물가지수, 생산자물가지수, 농산물및석유류제외지수, 식료품및에너지제외지수, WTI선물, 금선물, **물가전망(뉴스)**
주식	원달러환율, KOSPI, KOSDAQ, VKOSPI, **주가전망(뉴스)**
부동산	주택매매가격지수-서울, 주택매매가격지수-전국, 주택전세가격지수-서울, 주택전세가격지수-전국, **주택가격전망(뉴스)**
금리	콜금리, CD금리, 국고채3년금리
심리	경제심리지수, **뉴스심리지수(뉴스)**, **경제불확실성지수(뉴스)**
정부	**정부지출(뉴스)**

주: 1) 모든 관측변수는 추세적으로 상승하는 경우 전년동기대비 증가율을 사용하였으며, 진한 글씨는 텍스트 지표를 의미

본 연구에서는 DFM의 잠재요인으로 모든 변수를 고려한 종합 요인 2개와 63개 관측변수를 부문별로 나누어 구성한 11개 부문별 요인을 〈표 4〉와 같이 고려하였다. 이때 각 잠재요인의 시차변수는 선행연구를 참조하여, 2개 종합 요인의 경우 4기를, 11개 부문별 요인의 경우 1기를 고려하였다.

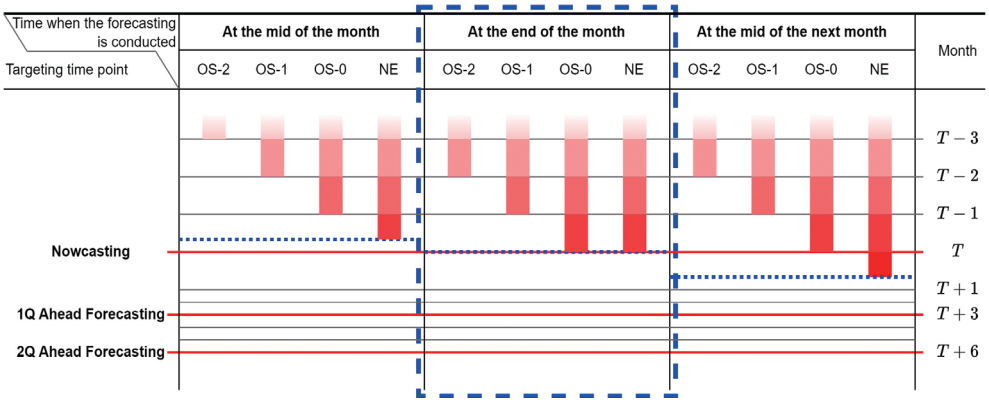
CRNN과 DFM은 모형의 구조에 있어서 매우 큰 차이가 있다. CRNN은 지도학습 방식으로 목적변수(y)를 가장 잘 예측하는 설명변수(X)의 비선형함수를 학습하는 반면, DFM은 비지도학습 방식으로 모든 관측변수($X' = [X:y]$)를 잘 적합하는 잠재요인(latent factor)을 학습하여 관측변수의 예측치를 계산한다. 따라서 DFM은 목적변수를 특정할 필요 없이 고려한 모든 관측변수를 동시에 월별로 예측할 수 있는 장점이 있다.

2. 학습데이터 구성

경기 예측모형을 학습시키기 위해서는 예측시점에 가용한 정보를 최대한 활용하여 학습데이터를 적절히 구성하는 것이 중요하다. 이렇게 구성한 학습데이터는 과거 시점 기준으로 작성하였다고 하여 vintage 데이터라고 부르며, vintage 데이터는 공표 주기 및 시차가 서로 다른 자료들로 구성된다. 따라서 vintage 데이터는 공표 주기가 서로 다른 자료를 데이터로 구성할 때 발생하는 혼합주기문제(mixed frequency problem)와 최종시점 자료가 누락되어 발생하는 최종시점차이문제(ragged-edge problem)를 해결해서 이용해야 한다.

보통 혼합주기문제와 최종시점차이문제는 변수 누락(missing variable) 문제와 동일하게 보고 보간(interpolation)을 이용하여 해결하거나 고차원 주기의 변수를 저차원 주기로 집계(aggregation)하여 해결한다. 본 연구

〈그림 6〉 예측시점별 가용 데이터[1]와 예측 시나리오



주: 1) 예측시점별 가용 데이터를 공표시차에 의해 다음과 같이 분류
- OS-2는 조사대상 익익월에 공표되는 공식 통계: 소매판매액지수
- OS-1은 조사대상 익월에 공표되는 공식 통계: 실업률, 고용률, 구인배수, 취업률, 취업자수, 수출물가지수, 소비자물가지수, 농산물및석유류제외지수, 식료품및에너지제외지수, 생산자물가지수, 설비투자지수, 제조업출하지수, 제조업재고지수, 서비스업생산지수, 광공업생산지수, GDP(국내총생산, 민간소비, 건설투자, 설비투자, 수출)
- OS-0은 조사대상 월말에 입수가능한 공식 통계: 전산업매출BSI, 전산업업황실적BSI, 제조업수출BSI, 제조업가동률BSI, 제조업신규수주BSI, 제조업내수판매BSI, 제조업업황실적(SA)BSI, 경제심리지수, 현재경기판단 CSI, 소비자심리지수, 주택매매가격지수(서울, 전국), 주택전세가격지수(서울, 전국), 콜금리, CD금리, 국고채3년금리, 원달러환율, KOSPI, KOSDAQ, VKOSPI, WTI선물, 금선물
- NE는 15개 부문별 텍스트 지표, 뉴스심리지수, 경제불확실성지수

〈표 5〉 GDP 전년동기대비 증가율 예측 평균오차(MAE)[1)]

모형 \ 데이터	공식 통계	텍스트 지표	공식 통계 & 텍스트 지표
DFM	0.743	0.872	0.681**
CRNN	0.995	0.971	0.816*

주 : 1) *, **는 공식 통계를 이용한 모형의 예측치 대비 해당 예측치의 예측력이 더 높음을 의미하는 Diebold-Mariano 통계량이 각각 α = 0.05 및 α = 0.025 유의수준에서 유의함을 의미

〈표 6〉 DFM 모형의 예측 평균오차(MAE) 비교[1)]

변수 \ 데이터		공식 통계	공식 통계& 텍스트 지표
분기	GDP	0.743	0.681
월	소비자물가지수	0.246	0.245
	(식료품및에너지제외)	0.120	0.118
	설비투자지수	5.846	5.728
	제조업출하지수	3.707	3.667
	서비스업생산지수	1.340	1.343
	광공업생산지수	3.211	3.203

주 : 1) 2016.1~2021.4분기 말월의 예측오차 평균

에서 DFM은 전자의 방법을, CRNN은 후자의 방법을 이용하여 혼합주기문제와 최종시점 차이문제를 해결하였다.

모형의 예측 시나리오는 〈그림 6〉과 같이 다양하게 구성하는 것이 가능하지만, 본 연구에서는 컴퓨팅 비용, 작업 편의 등을 고려하여 예측 수행시점은 분기 종료 당일로 하고, 예측 대상시점은 당분기로 한정하여 분석을 수행하였다. 모형 학습은 2005년부터 2021년까지의 46개 공식 통계자료(월 41개, 분기 5개)와 17개 뉴스 텍스트 기반 경제지표를 vintage 데이터로 구성하여 이용하였다. 이용한 변수의 구성은 〈표 4〉에 기술되어 있다.

모형의 평가를 위해서 2016년 1분기부터 2021년 4분기까지의 전년동기대비 국내총생산(GDP) 증가율의 일반화(regularized) 예측평균오차(mean absolute error, MAE)를 평가지표로 이용하였다. 즉, 매분기 말일까지 이용가능한 자료를 바탕으로 당분기 GDP 증가율을 예측하고, 이를 실제 공표수치와 비교하여 모형을 평가하였다. 또한 텍스트 데이터를 추가함으로써 얻는 정확도 개선효과와 모형 선택에 따른 정확도 개선효과를 구분하기 위하여, 예측평균오차를 데이터×모형의 크로스테이블로 작성하여 〈표 5〉와 같이 비교하였다.

V. 뉴스 텍스트를 이용한 경기 예측모형의 추정 결과

1. 경기 예측모형 추정 결과

텍스트 지표를 반영한 경기 예측모형의 GDP 예측 정확도를 분석한 결과, 선형 및 비선형 모형 모두에서 텍스트 지표를 추가한 경우 예측 정확도가 〈표 5〉와 같이 유의미하게 높아지는 것으로 나타났다. 여기서 예측 정확도는 학습오차

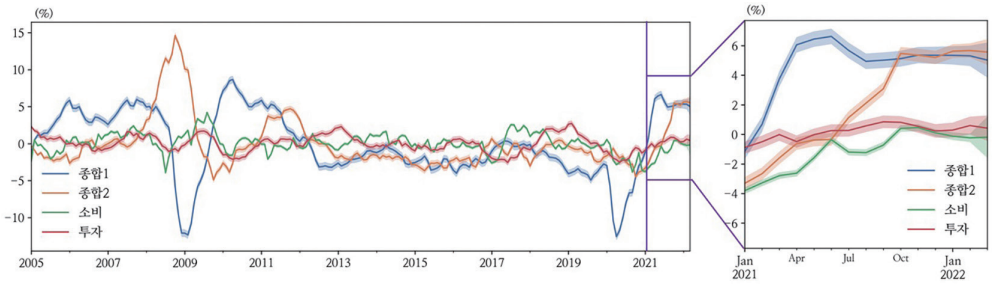
〈표 7〉 개별 요인(factor)의 관측변수 적합도[1)]

종합1 요인의 관측변수	R^2
제조업신규수주BSI	0.92
제조업내수판매BSI	0.91
제조업가동률BSI	0.91
전산업매출BSI	0.87
제조업업황실적(SA)BSI	0.87

종합2 요인의 관측변수	R^2
수입물가지수	0.71
수출물가지수	0.70
소비자물가지수	0.61
농산물 · 석유류제외지수	0.60
생산자물가지수	0.59

주: 1) 각 요인을 독립변수로 하여 관측변수를 적합할 경우의 적합도(coefficient of determination, R^2)

〈그림 7〉 종합1, 종합2, 소비, 투자 요인의 평균 추정값과 95% 신뢰구간

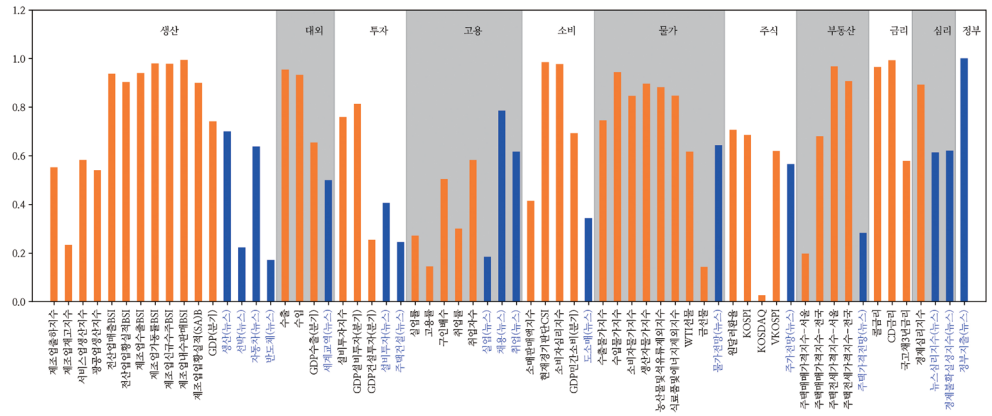


(training error) 기준이 아닌 예측시점 데이터를 학습하지 않은 일반화오차(regularized error) 기준으로 측정한 것이다. 즉, 예측 정확도의 향상은 텍스트 지표가 GDP 예측에 유의미한 정보를 추가함을 의미한다.

한편, 텍스트 지표를 반영한 경기 예측모형으로는 DFM이 CRNN보다 우수한 것으로 나타났다. DFM은 공식 통계를 이용할 수 없는 상황에서 텍스트 지표 및 요인별 추세(autoregression)를 이용하여 각 요인의 예측치를 추정하므로, 텍스트 지표가 각 요인의 대체 변수로 적절할 경우 모형의 적합도가 향상되고 예측치의 표준오차가 하락하게 된다. 또한 하위 부문의 합계로 구성되는 GDP의 특성을 고려할 때, 가용 데이터가 충분한 경우에는 과적합 오차가 발생할 수 있는 비선형모형보다 비교적 간단한 선형모형이 GDP 예측 모형으로 적합한 것으로 판단된다.

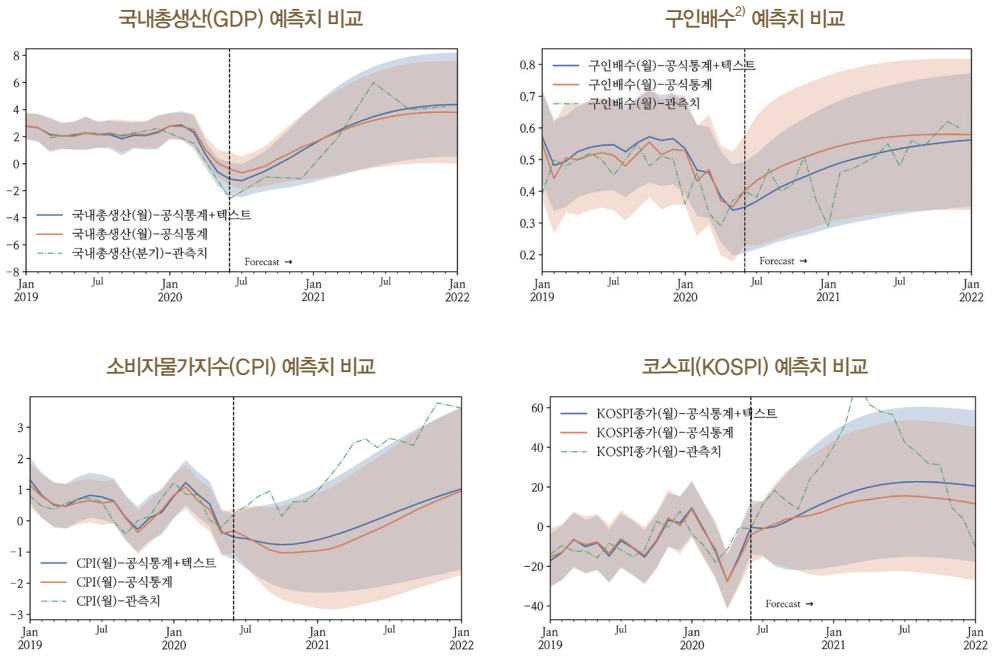
다만 DFM 모형을 보면 텍스트 지표만을 이용한 경우의 예측 정확도는 공식 통계만 이용한 경우보다 낮게 나타난다. 이는 텍스트 지표의 수가 17개로 제한적이고, 공식 통계 대비 노이즈가 크기 때문인 것으로 사료된다. 또한 텍스트 지표 반영에 따른 예측 정확도 향상 효과도 제한적인데, 이는 공식 통계를 이용할 수 있는 상황에서는 텍스트 지표의 추가적인 정보가 제한적이기 때문이다. 〈표 6〉에서 DFM 모형에 고려한 주요 경제변수들의 예측 정확도를

〈그림 8〉 DFM의 요인과 관측변수의 관계[1)2)]



주: 1) DFM에서 고려한 요인 전체를 독립변수로 하여 각각의 관측변수를 적합할 경우의 적합도(coefficient of determination, R^2)로 요인과 관측변수의 상관성을 보여줌
2) 파란색은 뉴스 텍스트 기반 경제지표를 나타냄

〈그림 9〉 2020년 6월 시점의 DFM 예측치 비교[1)]



주: 1) 음영은 예측치의 95% 신뢰구간을 의미
2) 구인 인원 대비 구직 건수 비율

비교해보면, 상대적으로 속보성이 높고 가용 정보가 많은 월지표의 경우, 텍스트 지표 반영에 따른 예측력 향상 효과가 분기 GDP에 비해 낮은 편으로 나타난다. 이는 월지표의 경우 텍스트 지표가 추가적으로 제공할 수 있는 정보가 많지 않기 때문인 것으로 보인다.

2. 텍스트 지표를 반영한 DFM 경기 예측 모형의 해석

DFM의 적합결과를 자세히 살펴보면, 〈표 7〉과 같이 2개의 종합 요인은 각각 생산 및 물가와 관련 있는 변수로 구성되었다. 따라서 〈그림 7〉에서 요인별 추정값 추이를 보면, 종합1 요인은 생산, 종합2 요인은 물가 요인을 나타내는 것으로 해석 가능하며, 여타 부문별 요인은 생산 및 물가의 장기 흐름을 제거한 단기적 변동으로 해석할 수 있다.

또한 DFM에 포함된 텍스트 지표의 예측 기여도를 〈그림 8〉에서 살펴볼 수 있다. 〈그림 8〉을 보면 생산(뉴스), 물가전망(뉴스), 채용(뉴스), 정부지출(뉴스), 뉴스심리지수(뉴스), 경제불확실성지수(뉴스) 등 거시적 지표의 성격을 갖는 텍스트 지표가 선박(뉴스), 반도체(뉴스) 등 미시적 지표 성격의 텍스트 지표에 비해 경기 예측모형에 유용한 것으로 나타난다.

마지막으로 〈그림 9〉에서 코로나19의 영향이 크게 나타났던 2020년 6월말 시점을 기준으로 DFM의 주요 경제변수 예측 결과를 살펴보았다. 〈그림 9〉에서 보는 바와 같이 공식 통계만 이용한 경우에 비해 텍스트 지표를 추가한 경우, GDP, CPI, KOSPI 등의 예측치가 관측치에 더 근접하는 것을 확인할 수 있다.

Ⅵ. 시사점

본 논고는 경제적으로 중요한 15개 부문의 주제를 선정한 뒤, 이들 부문의 경제적 흐름을 잘 반영하는 새로운 뉴스 텍스트 기반 경제지표의 작성 방법을 제시하였다. 뉴스 텍스트 기반 경제지표는 비교 대상 공식 통계에 비해 0~9개월 선행하며, 공식 통계와 높은 상관관계를 갖는 것으로 나타났다. 또한 뉴스 텍스트 기반 경제지표를 반영하여 경기 예측모형을 구축한 결과, 경기 예측 정확도가 유의미하게 향상되는 것을 확인하였다.

본 논고는 뉴스 텍스트를 경제지표로 작성하기 위해 뉴스 기사를 문장 단위로 분석하는 새로운 방법을 제시했다는 점과, 인공신경망 및 DFM 모형의 비교를 통하여 텍스트 지표의 특성을 반영하는 경기 예측모형을 도출하였다는 점에서 관련 연구에 기여할 것으로 평가된다.

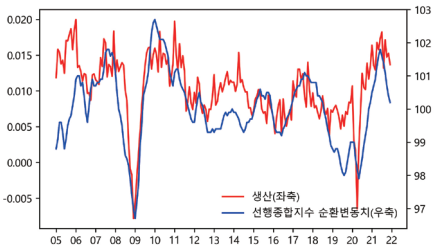
뉴스 텍스트 기반 경제지표는 실시간으로 작성 가능하므로 속보성 지표로 활용할 수 있고, 신속하고 정확한 경기동향 파악 및 경기 예측에도 유용하다. 또한 뉴스 기사를 정량화하고 통계 모형을 통해 체계적으로 뉴스 정보를 활용하는 방식은 연구자가 개별적으로 습득한 뉴스 정보를 정성적으로 활용하는 방식에 비해 휴먼 에러를 줄이는 데도 기여할 것으로 기대된다. 다만 이러한 경우 특정 의도가 담긴 뉴스가 많이 발간된다면, 경기 예측에 의도가 반영될 가능성이 있는 점에는 주의가 필요하다.

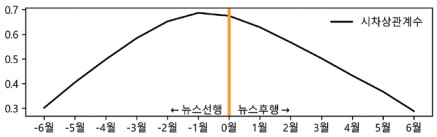
향후 뉴스 텍스트 기반 경제지표의 활용도를 높이기 위해서는 경기 예측에 유용한 텍스트 지표의 주제 발굴, 텍스트 지표의 정도 개선 등이 필요하다. 이를 위해 경기 예측에 중요한 새로운 부문의 텍스트 지표를 개발하는 한편, 필터링 기법 등을 통해 텍스트 지표의 노이즈를 감소시키는 연구를 발전시켜 나갈 필요가 있다.

〈부록 1〉

부문별 뉴스 텍스트 기반 경제지표 추이

① 생산

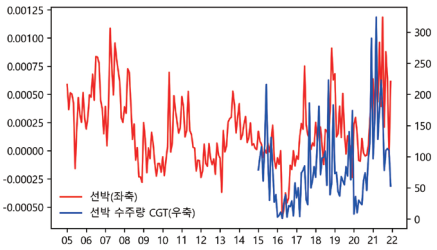


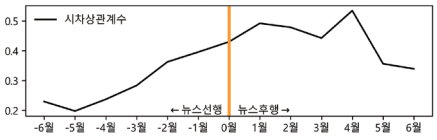


- 최대 선행 상관시차/계수 : −1 / 0.69
- 최대 후행 상관시차/계수 : 0 / 0.67
- Granger 인과성 검정 : −2~−5개월에서 유의 (α=0.01)

자료: 통계청 산업활동동향

② 선박

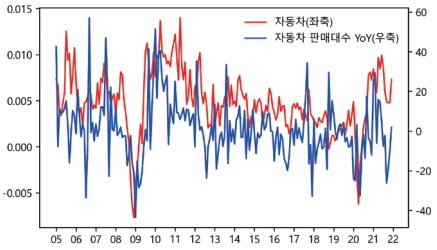


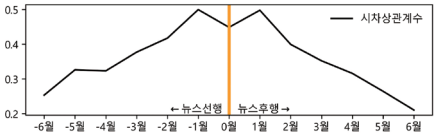


- 최대 선행 상관시차/계수 : −1 / 0.40
- 최대 후행 상관시차/계수 : 4 / 0.53
- Granger 인과성 검정 : −1개월에서 유의 (α=0.01)

자료: 클락슨 리서치

③ 자동차

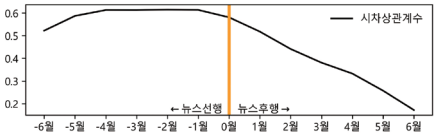




- 최대 선행 상관시차/계수 : −1 / 0.50
- 최대 후행 상관시차/계수 : 1 / 0.50
- Granger 인과성 검정 : −1~−3개월에서 유의 (α=0.01)

자료: 한국자동차산업협회

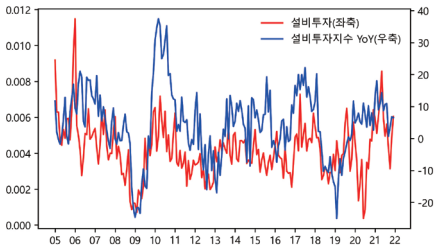
④ 반도체

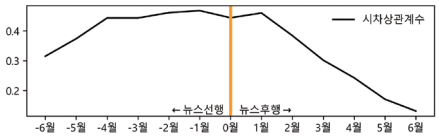


- 최대 선행 상관시차/계수 : −2 / 0.62
- 최대 후행 상관시차/계수 : 0 / 0.58
- Granger 인과성 검정 : −1개월에서 유의 (α=0.10)

자료: 정보통신기획평가원 ICT수출입통계

⑤ 설비투자

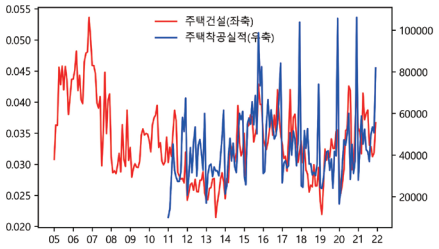


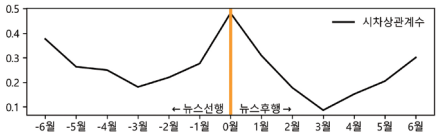


• 최대 선행 상관시차/계수 : −1 / 0.47
• 최대 후행 상관시차/계수 : 1 / 0.46
• Granger 인과성 검정 : −1~−4개월에서 유의 (α=0.01)

자료: 통계청 산업활동동향

⑥ 주택건설

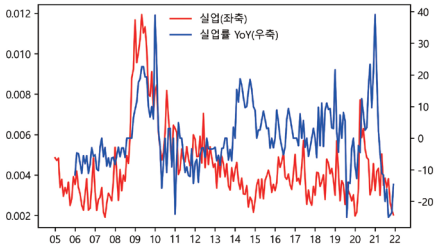


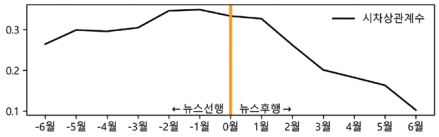


• 최대 선행 상관시차/계수 : −6 / 0.38
• 최대 후행 상관시차/계수 : 0 / 0.48
• Granger 인과성 검정 : 유의하지 않음

자료: 국토교통부

⑦ 실업

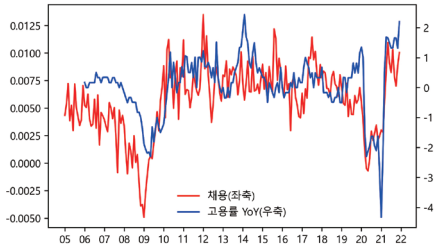


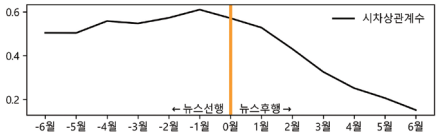


• 최대 선행 상관시차/계수 : −1 / 0.35
• 최대 후행 상관시차/계수 : 0 / 0.33
• Granger 인과성 검정 : 유의하지 않음

자료: 통계청 경제활동인구조사

⑧ 채용

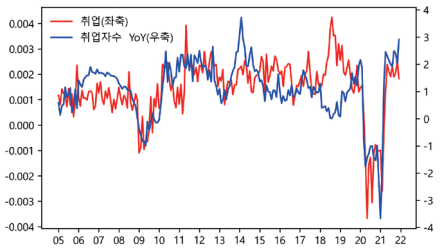


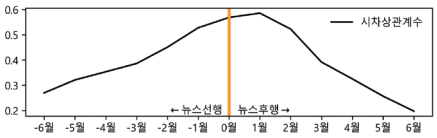


• 최대 선행 상관시차/계수 : −1 / 0.61
• 최대 후행 상관시차/계수 : 0 / 0.57
• Granger 인과성 검정 : −3~−4개월에서 유의 (α=0.01)

자료: 통계청 경제활동인구조사

⑨ 취업

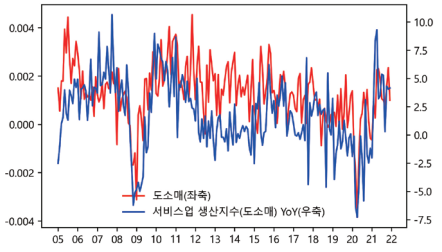


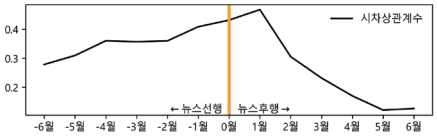


자료: 통계청 경제활동인구조사

- 최대 선행 상관시차/계수 : −1 / 0.53
- 최대 후행 상관시차/계수 : 1 / 0.59
- Granger 인과성 검정 : −1~−3개월에서 유의 (α=0.01)

⑩ 도소매

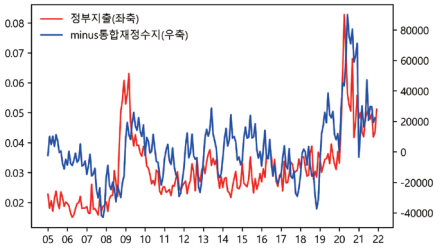


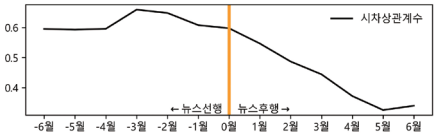


자료: 통계청 서비스업동향조사

- 최대 선행 상관시차/계수 : −1 / 0.41
- 최대 후행 상관시차/계수 : 0 / 0.47
- Granger 인과성 검정 : −1~−3개월에서 유의 (α=0.01)

⑪ 정부지출

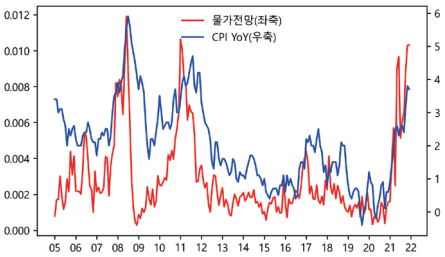


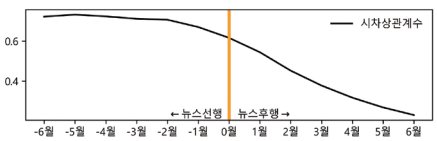


자료: 기획재정부

- 최대 선행 상관시차/계수 : −3 / 0.66
- 최대 후행 상관시차/계수 : 0 / 0.60
- Granger 인과성 검정 : −6~−7개월에서 유의 (α=0.05)

⑫ 물가전망

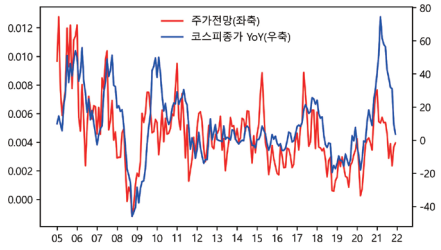


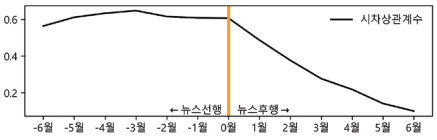


자료: 기획재정부

- 최대 선행 상관시차/계수 : −5 / 0.73
- 최대 후행 상관시차/계수 : 0 / 0.62
- Granger 인과성 검정 : −2~−3개월에서 유의 (α=0.05)

⑬ 주가전망

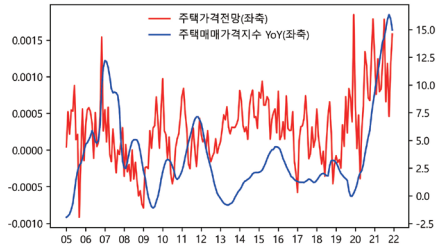


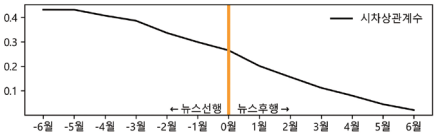


- 최대 선행 상관시차/계수 : −3 / 0.65
- 최대 후행 상관시차/계수 : 0 / 0.61
- Granger 인과성 검정 : −2개월에서 유의 (α=0.05)

자료: 한국거래소

⑭ 주택가격전망

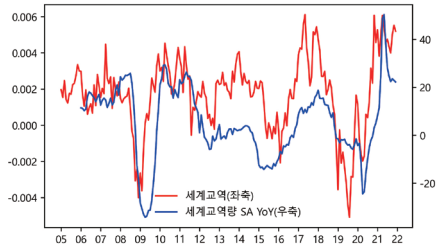


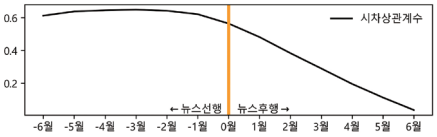


- 최대 선행 상관시차/계수 : −9 / 0.47
- 최대 후행 상관시차/계수 : 0 / 0.27
- Granger 인과성 검정 : 유의하지 않음

자료: KB부동산

⑮ 세계교역

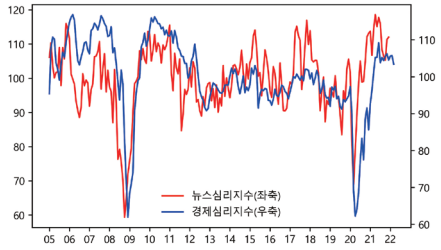


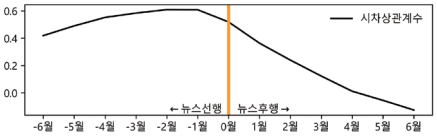


- 최대 선행 상관시차/계수 : −3 / 0.65
- 최대 후행 상관시차/계수 : 0 / 0.56
- Granger 인과성 검정 : −2개월에서 유의 (α=0.05)

자료: OECD

⑯ 뉴스심리지수

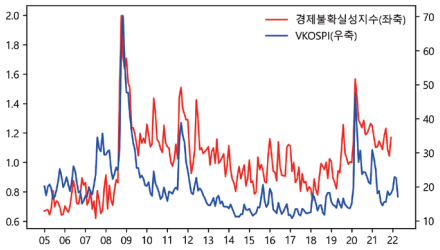


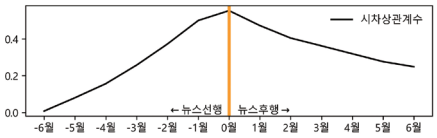


- 최대 선행 상관시차/계수 : −2 / 0.61
- 최대 후행 상관시차/계수 : 0 / 0.52
- Granger 인과성 검정 : −3개월에서 유의 (α=0.01)

자료: 한국은행

⑰ 경제불확실성지수





- 최대 선행 상관시차/계수 : -1 / 0.50
- 최대 후행 상관시차/계수 : 0 / 0.56
- Granger 인과성 검정 : -1개월에서 유의 (α=0.01)

자료: 한국거래소

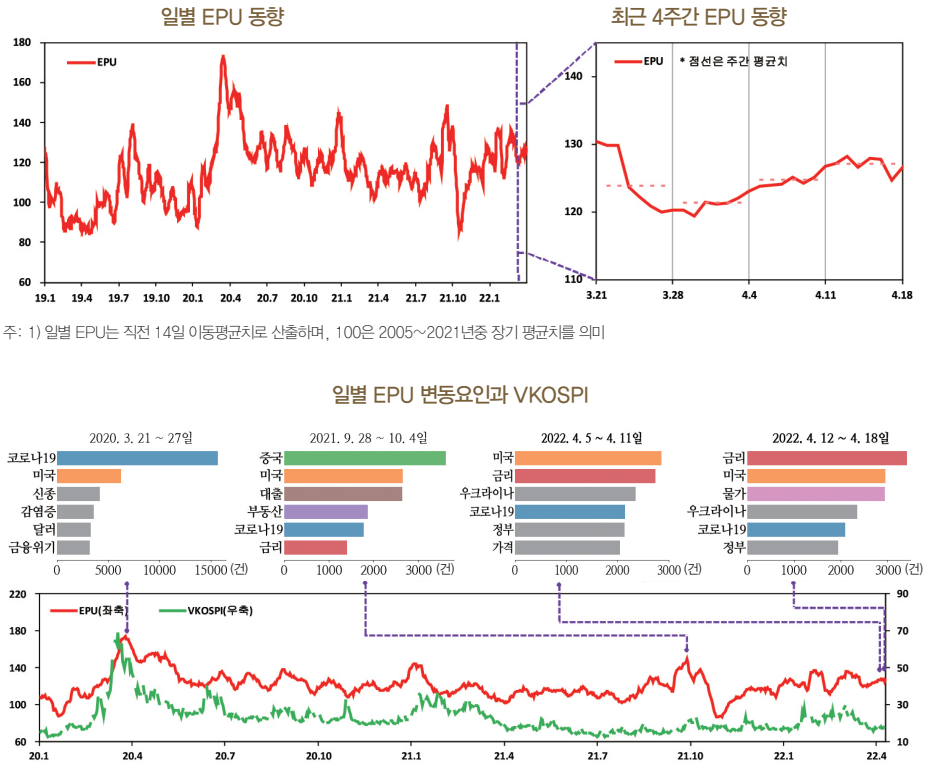
〈부록 2〉

경제불확실성지수(Economic Policy Uncertainty, EPU)

경제불확실성지수는 뉴스기사에 나타난 경제, 정책 관련 불확실성을 정량화한 지표로 가장 널리 연구된 뉴스 텍스트 기반 경제지표 중 하나이다. Baker et al.(2016) 연구팀이 처음으로 제시하였으며 사전접근법(lexical approach) 방식으로 작성되었다. 국내에서는 한국개발연구원(KDI)이 Lee et al.(2020)의 연구를 바탕으로 월별 EPU를 작성하여 공개하고 있다. 한국은행 경제통계국은 속보성 고빈도지표 개발 연구의 일환으로 일별 EPU를 내부적으로 작성하여 경기 동향 파악을 위해 활용하고 있다.

한국은행 경제통계국의 일별 EPU는 지수의 안정성을 고려하여 중앙지 10개사, 경제지 7개사, 인터넷 통신사 4개사, 방송사 4개사 등 25개 언론사를 대상으로 작성하였다. EPU 작성을 위한 단어군은 국내 사정에 맞도록 Lee et al.(2020)의 연구를 준용하여 Lee et al.(2020)의 EPU_KR_6를 기준으로 선정하였다. 〈그림 A1〉은 2022년 4월 셋째주의 EPU 동향을 살펴본 것이다. EPU는 다른 뉴스 텍스트 기반 경제지표와 마찬가지로 텍스트 분석을 통한 변동요인 파악이 매우 용이하다.

〈그림 A1〉 경제불확실성지수 동향



주: 1) 일별 EPU는 직전 14일 이동평균치로 산출하며, 100은 2005~2021년중 장기 평균치를 의미

〈참고문헌〉

김치호 · 김현정 (2016), "GDP 성장률의 Nowcasting에 관한 연구," 국민계정리뷰, 2016(2).

김수현 · 이영준 · 신진영 · 박기영 (2019), "경제분석을 위한 텍스트마이닝," BOK 경제연구, 2019(18).

김한준 · 조새롬 · 김동찬 (2021), "경제 텍스트 데이터를 활용한 키워드 분석방안 연구," 국민계정리뷰, 2021(1).

김현중 · 임종호 · 이혜영 · 이상호 (2019), "온라인 뉴스 기사를 활용한 경제심리보조지수 개발," 국민계정리뷰, 2019(2).

문혜정 · 이혜영 (2017), "빅데이터의 경제통계 활용 현황 및 시사점," 국민계정리뷰, 2017(3).

서범석 · 이영환 · 조형배 (2022), "기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용," 국민계정리뷰, 2022(1).

원중호 · 이한별 · 문혜정 · 손원 (2017), "텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류," 국민계정리뷰, 2017(4).

이긍희 · 황상필 (2016), "빅데이터를 이용한 경기판단지표 개발: 네이버 검색 경기지수 작성과 유용성 검토," BOK 경제분석, 2016(20-4).

이현창 · 최동규 · 김용건 · 허정 (2022), "디지털 신기술을 이용한 실시간 당분기 경제전망 (GDP Nowcasting) 시스템 개발," BOK 이슈노트, 2022(7).

전종준 · 안승환 · 이문희 · 황희진 (2020), "경제용어 감성사전 구축방안 연구," 국민계정리뷰, 2020(3).

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using Twitter?. Journal of Econometrics.

Armah, N. (2013), "Big Data Analysis: The Next Frontier," Bank of Canada Review, Summer, pp. 32-39.

Babii, A., Ghysels, E., & Striaukas, J. (2021). Machine learning time series regressions with an application to nowcasting. Journal of Business & Economic Statistics, 1-23.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4), 1593-1636.

Bańbura, M., & Michele M. (2014). Maximum likelihood estimation of factor

models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics 29(1), 133-160.

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2021). Business news and business cycles (No. w29344). National Bureau of Economic Research.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. Annual Review of Economics, 10, 615-643.

Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. American Economic Review, 112(4), 1194-1225.

Goshima, K., Ishijima, H., Shintani, M., & Yamamoto, H. (2021). Forecasting Japanese inflation with a news-based leading indicator of economic activities. Studies in Nonlinear Dynamics & Econometrics, 25(4), 111-133.

Huang, C., Simpson, S., Ulybina, D., & Roitman, A. (2019). News-based sentiment indicators. International Monetary Fund.

Jung, S., Lee, J., & Lee, S. (2021). The impact of geopolitical risk on stock returns: Evidence from inter-Korea geopolitics. IMF Working Papers, 2021(251).

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. Staff Working Paper in Bank of England.

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. Journal of Monetary Economics, 117, 507-520.

Lee, G. H., Cho, J. H., & Jo, J. G. (2020). New economic policy uncertainty indexes for South Korea. The Korean Journal of Applied Statistics, 33(5), 639-653.

Lee, Y., & Seo. B. (2022). Extracting Economic Sentiment from News Articles: The Case of Korea. Irving Fisher Committee on Central Bank Statistics (IFC) Conference in BIS 2022.

Mariano, R. S., & Murasawa, Y. (2010). A coincident index, common factors, and monthly real GDP. Oxford Bulletin of economics and statistics, 72(1), 27-46.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics, 34(4), 574-589.

Nguyen, K., & La Cava, G. (2020). Start Spreading the News: News Sentiment and Economic Activity in Australia. Sydney: Reserve Bank of Australia, 33.

Praptono1, N. H., & Zulen, Alvin Andhika (2021). Getting Insight of Employment Vulnerability from Online News: A Case Study in Indonesia. Irving Fisher Committee on Central Bank Statistics (IFC) Conference in BIS 2021.

Seki, K., Ikuta, Y., & Matsubayashi, Y. (2022). News-based business sentiment and its properties as an economic index. Information Processing & Management, 59(2), 102795.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. Journal of econometrics.

Shiller, R. J. (2017). Narrative economics. American economic review, 107(4), 967-1004.

Reichlin, L., Giannone, D., & Banbura, M. (2011). Nowcasting. Oxford Handbook on Economic Forecasting.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. Journal of Business & Economic Statistics, 38(2), 393-409.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics.