기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용

서범석*, 이영환**, 조형배***

한국은행 경제통계국은 경제분야 뉴스기사에 나타나는 경제심리를 지수화한 뉴스심리지수(NSI, News Sentiment Index)를 개발하여 2022년 2월부터 매주 한국은행 경제통계시스템(ECOS)에 실험적 통계로 공개하고 있다. 뉴스심리지수는 인터넷 포털사이트의 경제기사 텍스트를 웹크롤링 기법으로 수집하고 이를 최신 자연어 처리(NLP, Natural Language Processing) 모형으로 분석하여 작성한다.

뉴스심리지수 작성 결과, 월별 뉴스심리지수는 주요 경제심리지표 및 실물경기지표에 1~2개월 선행하며 높은 상관관계를 보이는 것으로 나타났다. 또한 일별 뉴스심리지수는 월 단위의 공식통계가 공표되기 전에 경제심리 변화를 신속하게 포착할 뿐만 아니라 키워드 분석, 부문별 지수 산출 등을 통해 변동요인 파악이 용이하다. 따라서 뉴스심리지수는 신속한 경제심리 파악 및 경기판단을 위한 지표로서 유용하게 활용될 수 있다. 한편 뉴스심리지수 작성과정에 최신 텍스트 마이닝(text mining) 분석기법, 기계학습(machine learning) 방법론, 자동화 프로세스(automation process) 등을 선도적으로 도입한 점도 새로운 데이터 및 통계기법의 발전과 확대에 크게 기여한 것으로 평가된다.

* 한국은행 경제통계국 통계연구반 과장(email: bsseo@bok.or.kr, phone: 02-759-5253)
** 한국은행 경제통계국 통계연구반 과장(email: yhlee@bok.or.kr, phone: 02-759-4452)
*** 한국은행 경제통계국 통계연구반 조사역(email: hyungbae.cho@bok.or.kr, phone: 02-759-5252)

Ⅰ. 검토배경

한국은행 경제통계국은 매일 발간되는 뉴스기사의 텍스트 분석을 통하여 국민들이 느끼는 경제심리를 파악하고자 뉴스심리지수(NSI, News Sentiment Index)를 개발하여 실험적 통계[1]로 공개하고 있다. 뉴스심리지수는 단순히 뉴스기사에 나타난 긍정문장과 부정문장을 카운트한 뒤 지수화한 지표이다. 뉴스심리지수는 설문조사에 의존하는 다른 경제심리지수와 달리 수시로 입수 가능한 뉴스기사를 이용하여 작성하므로, 경제심리 변화를 신속하게 포착하고 변동요인을 쉽게 파악할 수 있는 장점이 있다. 그러나 뉴스기사를 분석하여 뉴스심리지수를 작성하는 방법은 기술적으로 간단(trivial)하지 않은데 이는 매일 발간되는 뉴스기사의 논조를 모두 분류하는 것이 현실적으로 불가능하기 때문이다. 따라서 뉴스심리지수 작성을 위해서는 최신 자연어처리(NLP, Natural Language Processing) 및 텍스트 마이닝(text mining) 기술들이 필요하고, 방대한 뉴스데이터 분석을 위한 빅데이터 분석 기법의 활용이 필수적이다.

최근의 자연어 처리 기술의 발전은 뉴스심리지수의 기반이 되는 비정형 텍스트 데이터에 대한 분석을 가능케 하였다. 이에 따라 텍스트 데이터에 대한 경제학적 관심이 크게 증가하였으며, 그 중에서도 뉴스기사 데이터에 대한 관심이 크게 높아졌다. 뉴스기사 데이터는 자료의 양(Volume), 속보성(Velocity), 그리고 정보의 다양성(Variety) 측면에서 매우 우수한 것으로 평가받으며, 이로부터 경제 정보를 추출하기 위한 연구가 학계 및 각국 연구기관에서 활발히 진행되고 있다. 미국 샌프란시스코 연준은 미리 정해 놓은 특정 단어들의 감성사전을 만드는 방법(lexical approach)으로 일별 뉴스심리지수를 산출하여 공개하고 있고, IMF, 호주, 노르웨이 중앙은행 등은 뉴스심리지수의 산출을 연구한 결과물을 최근 발표하였다. 여기서 더 나아가 학계에서는 뉴스기사에서 추출한 지표들을 경제전망에 활용하여 경기예측 정확도를 높이고자 하는 연구가 활발히 진행되고 있다.

한국은행 경제통계국은 이러한 시대적 흐름을 선도하여 한국 실정에 맞는 뉴스심리지수의 개발을 추진하여 왔다. 한국은행은 2018년 12월 뉴스심리지수 개발 작업을 착수한 이래 텍스트 분석 방법론 검토, 감성분류 모형 및 학습데이터 구축, 전문가와의 협의 등을 거쳐 2021년 4월 지수를 처음 시험공개하였으며, 이후 최신 인공신경망 모형을 도입하고 학습데

1) 새로운 유형의 데이터를 활용하거나 새로운 방식을 적용하여 실험적으로 작성하는 통계로 국가통계는 아니다. 통계청은 빅데이터 활용 통계 등 통계의 다양성을 확대하고 새로운 방식의 통계 작성을 활성하기 위해 2021년 9월 실험통계제도를 도입하고 “실험적 통계 작성을 위한 가이드라인”을 마련하였다.

이터를 추가하는 등 지수의 안정성 및 정확성을 개선하여왔다. 이러한 노력을 종합하여 2022년 2월부터는 뉴스심리지수를 통계청의 실험적 통계(2022-001호)로 등록하고, 매주 화요일마다 '일별' 및 '월별' 지수를 산출하여 한국은행 경제통계시스템(ecos.bok.or.kr)에 공개하고 있다.

본고에서는 뉴스데이터의 통계적 활용에 대한 지속적인 관심과 연구의 결과물인 뉴스심리지수의 전반적인 내용을 소개하고, 활용 방안 등에 대해 살펴보고자 한다. 이를 위해 먼저 Ⅱ장에서는 뉴스심리지수의 작성 방법을 구체적으로 살펴보고, Ⅲ장에서 뉴스심리지수의 특성과 유용성을 평가하였다. 이어서 Ⅳ장에서는 뉴스심리지수의 다양한 활용 방안을 살펴보았으며 마지막으로 Ⅴ장에서 시사점과 향후 과제를 정리하는 것으로 본고를 마무리하였다.

Ⅱ. 뉴스심리지수 작성 방법

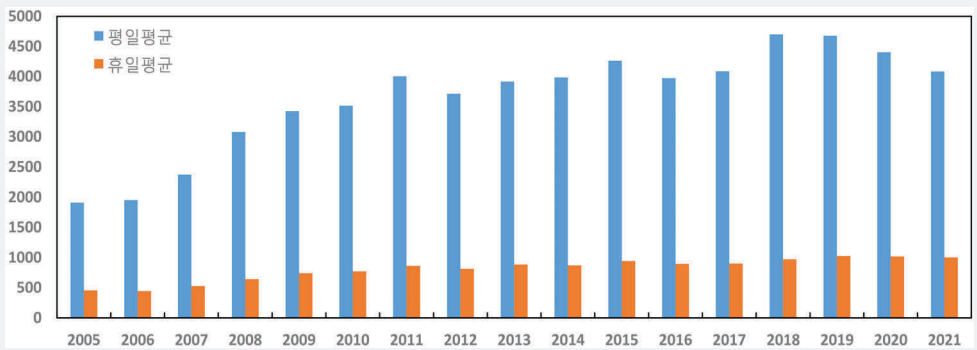
1. 데이터

뉴스심리지수 작성을 위한 뉴스기사 텍스트 데이터는 웹크롤링(web crawling) 방식으로 수집하였다. 웹크롤링은 인터넷에 노출된 데이터를 직접 다운로드하여 수집하는 방식이다. 인터넷 포털사이트에서는 기사 작성 주체인 언론사의 선택에 따라 뉴스기사를 정치, 경제, 사회, 생활/문화, IT/기술, 세계 등 6개의 분야로 분류하여 게재한다. 뉴스심리지수는 뉴스기사에 내포된 경제심리를 포착하기 위해 경제분야로 한정하여 인터넷 포털사이트에서 뉴스기사를 수집하였다. 수집 대상기간은 데이터의 크기, 통계 작성 및 분석에 필요한 시계열의 길이 등을 감안하여 2005년 이후로 정하였다. 이렇게 구축한 뉴스 데이터베이스는 현재 중앙지, 경제지, 방송사, 인터넷 통신사, 지역종합지 등을 포함한 약 50여개 언론사의 뉴스기사를 포함한다.

웹크롤링을 통해 데이터를 수집하는 경우 정형화된 데이터를 입수하는 것과 달리 '데이터 무결성(data integrity)' 문제가 발생할 가능성이 있다. 즉, 웹크롤링을 이용하는 경우 인터넷 포털의 운영상황이나 방침에 따라 중복된 뉴스기사 또는 광고성 기사가 함께 수집된다. 이를 처리하기 위해 뉴스심리지수 작성시에는 새로 수집한 뉴스기사 원문이 직전 30일 내에 한번이라도 동일하게 데이터베이스에 등장하는 경우, 이 뉴스기사를 데이터베이스에 추가하지 않고 배제하였다. 이는 광고성 기사의 경우 같은 내용의 기사가 반복되어 게재된다는 점과 중복된 기사의 경우 최초 게재 시점을 해당 뉴스가 나온 시점으로 보는 것이 타당하다는 측면에서 합리적인 원칙이라고 할 수 있다.

<그림 1>은 수집한 경제분야 뉴스기사 데이터베이스의 일평균 뉴스기사 수를 연도별 및 평일·휴일별로 집계한 것이다. 이를 살펴보면 일평균 뉴스기사 수(평일기준)는 2005년 약 2,000건에서 2021년 약 4,000건으로 크게 증가하였다. 이는 인터넷 포털에 기사를 게재하는 언론사의 수가 증가한 것과 함께 오래된 기사의 경우 언론사에서 게재를 취소하기 때문인 것으로 해석된다. 한편 평일과 휴일에 수집되는 뉴스기사 수의 차이는 일부 언론사 및 통신사를 제외한 주요 일간지의 경우 휴일에 뉴스기사를 생산하지 않기 때문이다.

〈그림 1〉 평일 · 휴일별 평균 경제 뉴스기사 수



2. 작성 방법

뉴스심리지수는 뉴스기사에 나타난 긍정문장과 부정문장을 카운트한 뒤 지수화하여 작성한다. 따라서 방대한 뉴스기사의 논조를 높은 정확도로 분류하는 것이 뉴스심리지수 작성의 핵심이 된다. 이를 위해 사람이 미리 분류해 놓은 학습 문장의 패턴을 통계 모형에 학습시킨 뒤, 동 모형을 새로운 문장에 적용하여 문장의 논조를 분류하는 기계학습(machine learning) 방법을 적용하였다. 기계학습 방법은 뉴스 문장($s \in S$)과 논조($l \in \{(p_0, p_1, p_2) \mid p_0 = P(Positive), p_1 = P(Negative), p_2 = P(Neutral)\}$)로 구성된 데이터쌍(data pair, $\{(s,l)_i\}_{i=1}^{N}$)을 이용하여 데이터에 가장 잘 부합하는 다음과 같은 감성분류 함수 f_θ 를 찾는 것과 같다.

$$f_\theta : S \rightarrow \{(p_0, p_1, p_2) \mid p_0 = P(Positive), p_1 = P(Negative), p_2 = P(Neutral)\}$$

이하에서는 뉴스심리지수 작성을 위해 뉴스문장의 논조를 기계학습 방법으로 분류하는 과정을 상세하게 살펴본다.

가. 기계학습 적용을 위한 전처리(Preprocessing) 과정

기계학습의 적용을 위해서는 감성분류모형(classifier)과 동 모형을 학습하기 위한 학습데이터(training data)의 구축이 필요하다. 이때 텍스트로 이루어진 뉴스데이터를 통계 모형이 인식할 수 있도록 하기 위한 전처리 과정을 먼저 거치게 된다. 전처리 과정은 텍스트 데이

터의 구조를 설정하고, 구조화된 텍스트 데이터를 형태소 단위로 분해(tokenizing)하며, 분해한 텍스트 데이터를 숫자로 전환(numerical encoding)하는 등의 과정을 포함한다. 이러한 과정은 일련의 텍스트 마이닝 기술을 순차 적용하는 것으로 이루어진다.

우선 감성분류모형에 사용할 입력 데이터의 구조를 정하는 것이 필요한데, 여기서 뉴스기사의 논조를 기사별로 분류할 것인지, 아니면 문장별로 분류할 것인지에 대한 판단이 필요하다. 일반적으로 뉴스기사는 긍정적인 내용과 부정적인 내용을 동시에 전달하는 경우가 많기 때문에 뉴스기사 한 편의 논조를 기사 기준으로 분류하는 것은 효과적이지 않다. 이 문제를 해결하기 위하여 뉴스심리지수는 기사가 아닌 문장 기준으로 논조를 분류하여 작성하였다. 따라서 입력 데이터는 뉴스기사를 구성하고 있는 각 문장($s \in S$)이 된다.

뉴스기사 문장을 모형에 입력하기 위해서는 문장을 의미가 있는 가장 작은 말의 단위인 형태소(POS, part-of-speech)로 분해하는 것이 필요하다. 이 과정을 형태소 분해(tokenizing)라고 하는데, 예를 들면 '뉴스심리지수를 작성하였다.'는 다음과 같이 분해하게 된다.

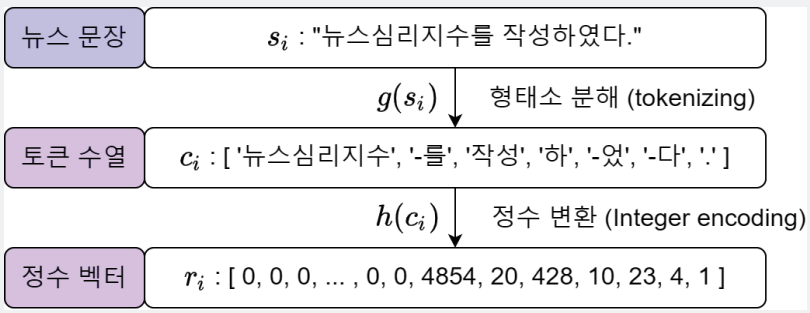
'뉴스심리지수(고유명사)' + '-를(목적격조사)' + '작성(일반명사)'
+ '하(동사파생접미사)' + '-었(선어말어미)' + '-다(종결어미)' + '.(마침표)'.

문장을 형태소 단위로 분해하면 문장의 다양한 변형에도 불구하고 동일한 의미의 말뭉치(corpus) 토큰(token)들을 구별해 내는 것이 가능해진다. 따라서 형태소 분해는 텍스트 데이터를 통계 모형에 적용하기 위한 필수적인 단계이다. 말뭉치 토큰의 수열 모임(a collection of sequences of tokens)을 C라고 하면, 형태소 분해 과정은 다음과 같이 나타낼 수 있다.

$$g : S \rightarrow C$$

형태소로 분해한 말뭉치 토큰들은 통계 모형에 입력할 수 있도록 정수 변환(integer encoding)을 이용하여 수치화한다. 즉, 유일한(distinct) 말뭉치 토큰별로 각각 하나의 정수를 매칭하여 문장을 하나의 수치형 벡터로 나타낸다. 이때 각 문장별로 수치형 벡터의 최대 길이(m)를 80으로 설정하고, 80보다 작은 길이를 갖는 문장의 경우 0 값을 채워 넣어(zero padding) 벡터의 길이를 동일하게 맞추었다. 여기서 최대 길이 80은 과거 뉴스기사 문장의 길이를 감안하여 학습데이터 기준 99%의 문장이 포함되는 길이로 임의 설정하였다. 즉, 정수 변환 과정은 다음과 같은 전단사(bijection) 함수로 표현할 수 있다.

〈그림 2〉 뉴스 텍스트 데이터의 전처리 과정



$$h : C \rightarrow \mathbb{R}^m$$

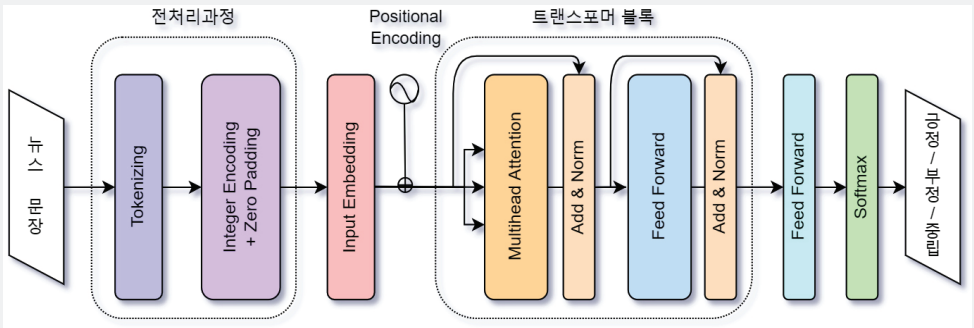
이상의 전처리 과정은 모형의 학습을 위한 학습데이터 뿐만 아니라 뉴스심리지수의 작성을 위해 입수하는 일별 뉴스데이터에 모두 동일하게 적용된다.

나. 감성분류모형(Classifier) 및 학습 데이터(Training Data)의 구축과 모형 학습(Model Training)

뉴스문장의 논조를 파악하기 위한 감성분류(sentiment classification) 모형으로는 최근 널리 이용되는 인공신경망 기반의 트랜스포머(transformer) 모형을 이용하였다. 트랜스포머 모형은 여러 개(multiple-head)의 Attention 구조를 갖는 인공신경망 모형으로 텍스트 분석을 위해 널리 이용되던 RNN(recurrent neural network) 및 LSTM(long short-term memory) 모형에 비해 입력 벡터 전체의 맥락(context)을 더 잘 파악하는 것으로 알려져 있다. 여기서 Attention 구조란 벡터를 입력데이터로 하는 예측 모형(sequential models)에서 입력 벡터의 값들 중 집중(attention)적으로 학습할 필요가 있는 값에 더 높은 가중치를 부여하도록 구성한 인공신경망 구조를 일컫는다. 여러 자연어처리(NLP) 모형은 Attention 구조를 통해 모형의 예측 정확도를 높일 수 있다.

<그림 3>은 뉴스심리지수의 작성을 위해 구축한 트랜스포머 기반의 감성분류모형을 도식화한 것이다. 감성분류모형은 뉴스문장을 입력변수로 받으며, 해당 문장의 감성이 긍정/부정/중립일 확률을 출력변수로 출력한다. 한편 트랜스포머 기반 인공신경망 모형(k_θ)의 파라미터 벡터를 θ라고 할 때 전처리 과정을 포함한 감성분류모형 f_θ는 다음과 같이 나타낼 수 있다.

〈그림 3〉 트랜스포머(Transformer) 기반 뉴스심리지수 감성분류모형의 도식



$$f_\theta : S \rightarrow \{(p_0, p_1, p_2)\},$$

$$f_\theta(s_i) = k_\theta \circ h \circ g(s_i) \equiv \hat{l}_i \ ,$$

여기서 $s_i \in S, \quad \hat{l}_i \in \{(p_0, p_1, p_2)\}$

구체적으로 설명하면, 뉴스문장은 먼저 전처리 과정을 거쳐 수치형 벡터로 치환된다. 수치형 벡터로 치환된 문장은 각 단어 토큰의 연관도를 학습하여 다시 다차원 공간으로 배치하는 Embedding 과정을 거치는데 뉴스심리지수는 Embedding을 위해 32차원을 고려하였다. 따라서 하나의 뉴스문장은 전처리와 Embedding 과정을 거치면서 80×32 차원의 수치행렬 값으로 치환된다. 이후 트랜스포머 블록과 이어지는 Feed Forward 네트워크는 입력 문장의 논조를 잘 예측하는 비선형 함수를 찾아 뉴스문장의 감성을 긍정/부정/중립 중 하나로 분류하게 된다. 마지막으로 Softmax 함수는 실수값의 출력변수를 0과 1사이의 확률값으로 변환하는 역할을 한다.

이와 같이 구성한 감성분류모형의 파라미터 θ값은 데이터쌍($\{(s,l)_i\}_{i=1}^N$)의 관측 레이블($l_i = (p_0^{(i)}, p_1^{(i)}, p_2^{(i)})$)과 모형의 예측치($\hat{l}_i = (\hat{p}_0^{(i)}, \hat{p}_1^{(i)}, \hat{p}_2^{(i)})$) 분포가 비슷해지도록 다음의 Cross-entropy 손실함수(L)를 최소화하는 과정(optimization)을 통해 추정(estimate or train)한다.

$$L(\theta \mid \{s_i, l_i\}_{i=1}^N) = -\sum_{i=1}^{N}\sum_{c=0}^{2} p_c^{(i)} \log \hat{p}_c^{(i)}$$

이제 감성분류모형의 파라미터(θ)를 학습하기 위하여 학습데이터를 구축한 방법에 대해서 살펴본다. 학습데이터는 2005년 1월부터 2021년 6월 중에 생성된 경제 뉴스기사에서 약 44만개 문장을 추출하여 구성하였다. 이 학습데이터 문장을 16명의 통계조사원이 직접 '긍정', '부정', '중립'으로 분류하여 각 문장별로 감성 레이블을 작성하였다. 감성 레이블은 사람이 작성함에 따라 발생할 수 있는 측정오류를 최소화하기 위해 최초 분류 이후 검토자의 검토를 거쳐 최종 레이블을 확정하였다. 이렇게 구성한 학습데이터는 총 문장 중 약 80%가 중립, 나머지 약 20% 중 절반이 긍정, 다른 절반이 부정으로 구성되었다. 이러한 학습데이터의 불균형한(imbalanced) 레이블 비중은 기계학습 모형의 최종 예측값에도 매우 큰 영향을 미친다. 이에 따라 모형의 예측 정확도(accuracy), 최종 산출한 뉴스심리지수의 적정성 및 안정성(stability) 등을 검토한 후 불균형 데이터의 조정을 위해 레이블 가중치를 모형 학습에 반영하였다.

한편 학습데이터는 문장의 논조 이외에도 내용의 지리적 범위를 기준으로 '국내', '국외', '국내·국외 모두 해당' 등 세 가지로 분류하였다. 뉴스문장 내용의 지리적 범위를 조사한 이유는 국외 뉴스기사의 경우 국내 경제심리와는 다른 양상을 보이는 경우가 많아 이를 뉴스심리지수 작성에서 제외하기 위함이다.

이와 같이 구성한 학습데이터를 이용하여 감성분류 모형과 국내외분류 모형 등 두 가지 모형을 트랜스포머 모형을 바탕으로 개별적으로 구성한 뒤 학습시켰다.

〈참고 1〉

감성분류모형의 성능 비교

감성분류모형의 성능 비교를 위해 Random Forest, Support Vector Machine (SVM), Single-head Attention 모형 등을 함께 고려하였다. <표 1>은 5000개 문장으로 구성한 검증 데이터(validation data)에 대하여 SVM과 트랜스포머 모형의 감성분류성능을 비교한 표이다. 분류성능 비교 결과 트랜스포머 모형이 여타 모형에 비해 분류성능이 우수한 것으로 나타나 이를 채택하였다. 한편 트랜스포머 모형의 가중치 조정 여부를 검토한 결과, 가중치를 조정하지 않은 경우 검증데이터의 감성분류 성능은 높게 나타나지만 뉴스심리지수 작성 시 변동성이 확대되는 문제가 발생하여 레이블 가중치를 모형 학습과정에 반영한 모형으로 최종 모형을 선택하였다.

〈참고 1〉 계속

〈표 1〉 감성분류모형의 성능 비교

	SVM		트랜스포머	
	가중치 조정 (balanced)	가중치 미조정 (imbalanced)	가중치 조정 (balanced)	가중치 미조정 (imbalanced)
정확도(accuracy)	0.86	0.86	0.96	0.98
민감도(sensitivity)	0.87	0.91	0.95	0.98
특이도(specificity)	0.84	0.79	0.97	0.99
정밀도(precision)	0.86	0.88	0.95	0.97
F1점수	0.86	0.89	0.96	0.98

다. 일별 뉴스데이터의 구축과 감성분류모형의 적용(Prediction)

뉴스심리지수는 앞 절에서 설명한 감성분류모형을 일별로 입수하는 뉴스데이터에 적용하여 산출한다. 이때 입수되는 뉴스기사의 수는 2021년 기준 일평균 약 4000개이며, 이를 문장으로 환산할 경우 약 7만 문장이 된다. 따라서 모든 문장에 대하여 전처리 과정 및 감성분류모형을 적용하는 것은 막대한 컴퓨팅 비용을 초래한다. 컴퓨팅 비용을 낮추기 위하여 뉴스심리지수는 입수한 뉴스기사 문장 중 일부를 표본추출(sampling)하여 작성하는 것으로 결정하였다. 즉, 입수한 뉴스기사를 문장 단위로 분해하여 일별 데이터 모집단을 먼저 구성한 다음 일별 모집단에서 1만 개의 표본문장을 임의로 추출하여 일별 데이터베이스를 구성하였다. 여기서 일별 표본문장의 수는 작업시간과 뉴스심리지수의 안정성 등을 고려하여 1만 개로 정하였다.

이와 같이 구성한 일별 뉴스데이터에 대하여 전처리 과정을 거친 후 감성분류 및 국내외분류 모형을 적용하였다.

라. 뉴스심리지수의 작성 및 표준화

뉴스심리지수는 일별로 입수한 뉴스기사의 표본문장들을 앞에서 설명한 과정을 거쳐 '국내·긍정문장'과 '국내·부정문장'으로 분류한 뒤, 두 분류(class)의 문장 개수를 카운트하여 작성한다. 이때 일별 뉴스심리지수의 경우 변동성을 고려하여 해당일 기준 직전 7일간 발간된 뉴스기사를 기준으로 작성하였으며, 월별 뉴스심리지수의 경우 해당월 중 발간된 뉴스기사를 기준으로 작성하였다. 또한 지수의 안정성 및 여타 지표와의 비교 등을 고려하여

뉴스심리지수는 2005년부터 해당일 기준 직전 연도까지를 표준화 구간으로 설정하고 이 기간중 지수의 평균과 분산을 이용하여 평균이 100, 표준편차가 10이 되도록 표준화하여 산출하였다.

구체적으로 특정일 t 에 대하여 국내외분류 모형에 의해 국내로 분류되면서 동시에 감성분류모형에 의해 긍정으로 분류된 문장의 수를 P_t 라고 하고 동일하게 국내로 분류되면서 동시에 부정으로 분류된 문장의 수를 N_t 라고 하면 일별 뉴스심리지수($NSI_t^{(daily)}$)는 다음의 산식에 의해 계산된다.

$$NSI_t^{(daily)} = \left(\frac{X_t - \overline{X}}{S} \right) \times 10 + 100,$$

$$\text{여기서 } X_t = \frac{\sum_{\tau=1}^{7} P_{t-\tau} - \sum_{\tau=1}^{7} N_{t-\tau}}{\sum_{\tau=1}^{7} P_{t-\tau} + \sum_{\tau=1}^{7} N_{t-\tau}},$$

$$\overline{X} = \frac{1}{|T|} \sum_{t \in T} X_t, \quad S = \sqrt{\frac{1}{|T|-1} \sum_{t \in T} (X_t - \overline{X})^2}.$$

표준화 구간을 나타내는 T 는 2005년 1월 1일부터 해당일 기준 직전 연도 말일까지의 모든 일자를 포함하는 집합에 대응하는 날짜 지표의 집합(index set)이다.

한편 특정월 u 에 대한 월별 뉴스심리지수($NSI_u^{(monthly)}$)는 위의 산식에서 T 를 2005년 1월부터 해당월 기준 직전 연도 말월까지의 모든 월을 포함하는 집합에 대응하는 월 지표의 집합 U 로 바꾸고, X_t 를 월별로 산출한

$$X_u = \frac{\sum_{\tau \in M_u} P_\tau - \sum_{\tau \in M_u} N_\tau}{\sum_{\tau \in M_u} P_\tau + \sum_{\tau \in M_u} N_\tau}$$

로 바꾸어 동일한 방식으로 계산한다. 여기서 M_u 는 특정월 u 에 포함되는 모든 일자 집합에 대응하는 날짜 지표의 집합이다.

위와 같이 표준화한 뉴스심리지수는 지수가 100보다 크면 뉴스기사에 나타난 경제심리가 과거 평균보다 낙관적, 100보다 작으면 비관적인 것으로 해석할 수 있다. 표준화 구간은 지수의 현실반영도 제고를 위해 매년 초에 전년말까지 연장하여 과거 시계열을 수정할 예정이다.

3. 전 과정의 자동화(Automation)

뉴스심리지수는 일별 뉴스기사 텍스트 데이터의 입수, 전처리, 예측 및 지수의 작성 등 전 과정이 파이썬 기반 자동화 프로그램을 통하여 이루어지도록 구축되었다. 자동화 프로그램은 매일 오전 5시에 데이터베이스의 마지막 시점부터 직전일까지의 뉴스기사를 수집하며, 이후 감성분류모형이 작동하여 해당일의 뉴스심리지수를 작성하여 출력한다. 이러한 뉴스심리지수 작성 전 과정의 자동화는 통계 작성의 효율을 크게 향상시킨 것은 물론이고 노동력 절감을 통하여 통계 작성 환경을 발전시킨 것으로 평가할 수 있다.

Ⅲ. 뉴스심리지수 유용성 평가

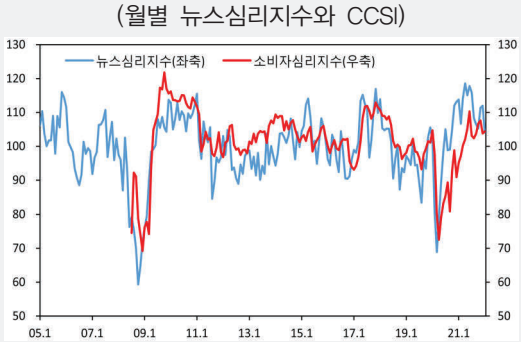
1. 공식 경제지표와의 상관성

뉴스심리지수가 뉴스기사에 나타난 경제심리를 반영하므로 기존의 공식 경제심리지표들과 비교해 볼 필요가 있다. 대표적인 경제심리지표로는 소비자의 경제에 대한 심리를 나타내는 소비자동향지수(CSI), 주요 CSI 항목을 합성한 소비자심리지수(CCSI), 기업가의 경제심리를 나타내는 기업경기실사지수(BSI), 소비자와 기업가를 합친 민간의 경제심리를 나타내기 위해 CSI와 BSI의 주요 항목을 합성한 경제심리지수(ESI) 등이 있다. 이들 경제심리지표는 매월 설문조사 결과를 바탕으로 작성되는 월별 지표이므로 뉴스심리지수도 월 단위로 작성하여 비교하였다. 한편 비교대상 경제심리지표의 이용가능한 시점을 고려하여 ESI와 BSI는 2005년 1월, CCSI와 CSI는 2008년 7월을 초기시점으로 하고 2021년 12월까지의 시계열을 대상으로 비교분석을 실시하였다.

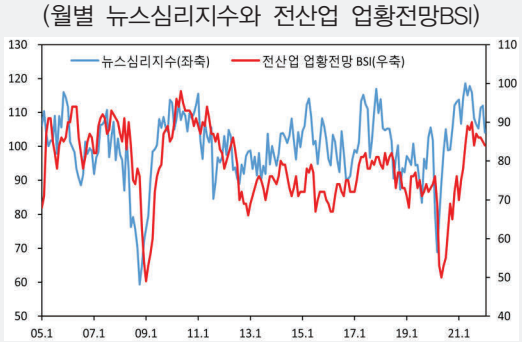
분석결과 월별 뉴스심리지수는 주요 경제심리지표에 1~2개월 선행하며 높은 상관관계를 보이는 것으로 나타났다. <그림 4>를 보면 파란색의 월별 뉴스심리지수가 CCSI 및 전산업 업황전망 BSI와 전반적으로 유사한 흐름을 보이고 있다. 또한 글로벌 금융위기, 코로나19 위기 시에 뉴스심리지수가 1~2개월 빨리 저점을 찍고 상승하는 모습을 볼 수 있다. 보다 구체적으로 <표 2>의 교차상관분석 결과를 보면, 뉴스심리지수는 CCSI 및 CSI 현재생활형편과 가장 밀접한 관계를 보이며 CSI에 대해서는 약 1개월 선행하고 BSI 전망지수에 대해서는 약 2개월 선행하는 것으로 나타났다.

〈그림 4〉 뉴스심리지수와 경제심리지표의 비교

(월별 뉴스심리지수와 CCSI)



(월별 뉴스심리지수와 전산업 업황전망BSI)



〈표 2〉 **뉴스심리지수와 주요 경제지표간 교차상관분석 결과**

비교대상 경제지표[1]		최대상관계수	최대상관시차[2]
CCSI		0.75	−1
ESI		0.61	−2
CSI	현재생활형편	0.74	−1
	현재경기판단	0.73	−1
	생활형편전망	0.73	−1
	향후경기전망	0.70	−1
	가계수입전망	0.68	−1
	소비지출전망	0.57	−1
	취업기회전망	0.72	−1
BSI	전산업 업황실적	0.64	−1
	전산업 채산성실적	0.68	−1
	전산업 자금사정실적	0.64	−1
	전산업 업황전망	0.61	−2
	전산업 채산성전망	0.65	−2
	전산업 자금사정전망	0.61	−2
실물지표	KOSPI (월별 전년동기대비 증가율)	0.68	−1
	선행종합지수 (순환변동치)	0.76	−2
	분기GDP (실질SA 전기대비 증가율)	0.53	0

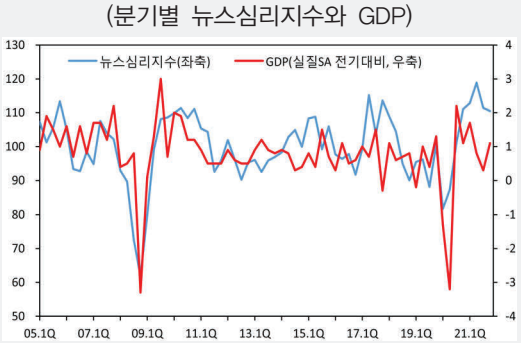
주: 1) 자료의 이용가능한 시점을 감안하여 CCSI와 CSI는 2008.7~2021.12월, 나머지 월별지표와는 2005.1~2021.12월, 분기GDP와는 2005.1~2021.4분기를 대상으로 비교
2) (−)는 뉴스심리지수가 비교대상 경제지표에 선행함을 의미

한편 뉴스심리지수는 실물경기지표와도 전반적으로 유사한 흐름을 보이며 높은 상관관계를 갖는다. 구체적으로 월별 뉴스심리지수는 선행종합지수 순환변동치(2005.1~2021.12월)에 2개월 선행하며 0.76의 상관관계를 보인다. 또한 분기별 뉴스심리지수는 GDP(실질 계절조정계열 전기대비 증가율, 2005.1~2021.4분기)와 0.53의 상관관계를 보이는 것으로 나타났다.

〈그림 5〉 **뉴스심리지수와 실물경제지표의 비교**

(월별 뉴스심리지수와 선행종합지수)

(분기별 뉴스심리지수와 GDP)



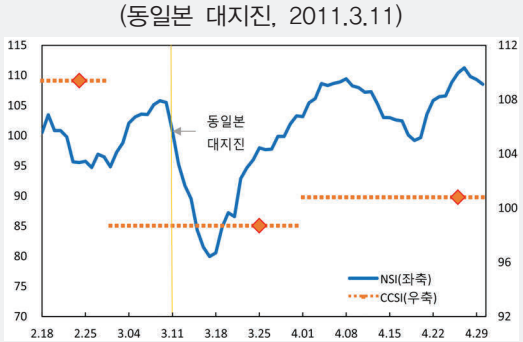
2. 속보성

뉴스심리지수는 일 단위로 작성 가능하므로 월 단위의 공식통계가 공표되기 이전에 경제심리 변화를 신속하게 파악할 수 있는 장점이 있다. <그림 6>은 주요 이슈 전후의 뉴스심리지수 추이를 살펴본 것이다. 이를 통해 2011년 동일본 대지진, 2015년 메르스 최초 사망자 발생 전후로 뉴스심리지수가 경제심리의 변화를 속보성 있게 포착하고 있음을 확인할 수 있다. 동일본 대지진이 발생한 2011년 3월 11일 직후 뉴스심리지수가 크게 하락한 뒤 반등하는데 3월 말에 공표된 소비자심리지수도 전월에 비해 크게 하락한 뒤 4월에는 소폭 상승하는 모습을 보인다. 또한 2015년 6월 1일 메르스 최초 사망자 발생시점에도 뉴스심리지수가 즉각적으로 반응하여 하락하고 이후 6월 말에 공표된 소비자심리지수가 비슷한 수준으로 하락하는 것을 확인할 수 있다. 이와 같이 뉴스심리지수와 소비자심리지수가 전반적으로 같은 방향으로 움직이므로 월말에 소비자심리지수가 공표되기 이전에 경제심리의 변화를 신속하게 파악하는 데 뉴스심리지수를 활용할 수 있다.

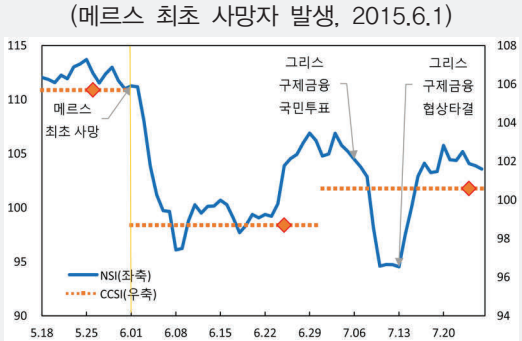
한편 <그림 7>은 2020년 1월 20일 코로나19 국내 첫 확진자가 발생한 이후부터 최근(2022년 1월 31일)까지 뉴스심리지수의 추이를 살펴본 것이다. 이를 통해 뉴스심리지수가 감염병 전개 양상, 주요 경제이슈 등에 따른 변화를 신속하게 포착하고 있음을 확인할 수 있다. 구체적으로 보면 뉴스심리지수가 코로나19 국내 첫 확진자 발생 및 1~4차 대유행 시점에 크게 하락한 반면, 재난지원금 지급, 거리두기 하향 조정 등과 같이 경제에 긍정적인 시그널이 되는 이슈 발생 다음에는 상승하는 모습을 보였다. 이와 같이 뉴스심리지수는 특정 이슈 발생에 따른 경제심리 변화를 즉각 포착할 수 있어 설문조사 기간 이후에 발생한 이슈가 누락될 수 있는 월 단위의 공식 경제지표들을 보완하는 역할을 할 수 있다.

〈그림 6〉 주요 이슈 전후의 뉴스심리지수 추이

(동일본 대지진, 2011.3.11)

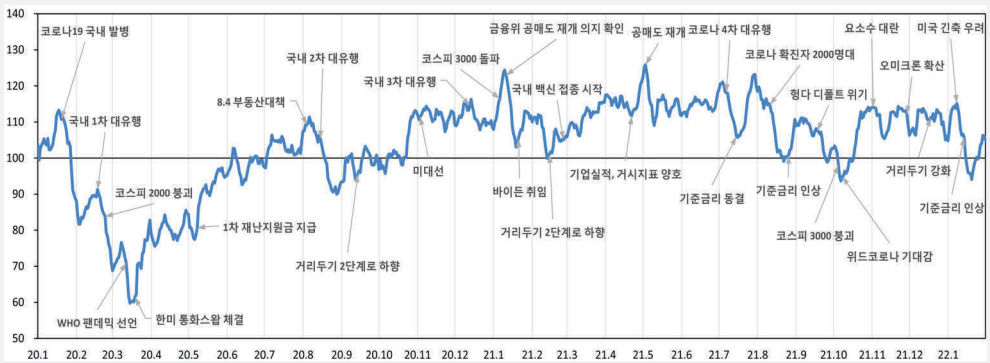


(메르스 최초 사망자 발생, 2015.6.1)



주 : 1) 붉은색 마름모는 월별 CCSI가 공표된 시점을 표시

〈그림 7〉 코로나 발생 이후 뉴스심리지수 추이

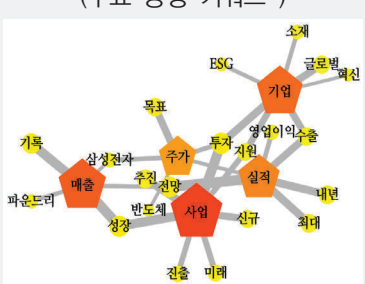


3. 키워드 분석을 통한 설명가능성

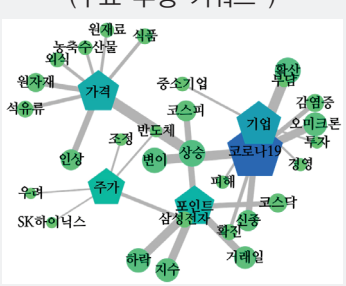
뉴스심리지수는 뉴스기사를 분석하여 작성하므로 키워드 분석을 통해 변동요인을 쉽게 파악할 수 있는 점도 큰 장점이다. <그림 8>은 코로나19 오미크론 변이가 확산하던 2021년 12월 첫째 주(11월 20일~12월 6일)를 대상으로 뉴스심리지수 작성에 활용된 뉴스문장의 키워드를 분석한 결과이다. 동 기간의 긍정 키워드에는 상장기업의 매출 등 실적 발표 내용이 주로 언급된 반면, 부정 키워드로는 코로나19 이슈와 함께 원자재 가격 등 물가 상승 요인이 많이 언급된 것을 확인할 수 있다. 또한 주요 급상승 키워드를 보면, 동 기간 중 슈퍼예산 통과에 따른 소상공인 지원 내용, 사상 첫 600억 달러를 돌파한 수출액 통계, 그리고 미국의 긴축 우려에 따른 국제 증시 및 유가 급락 등의 이슈를 함께 확인할 수 있다.

〈그림 8〉 뉴스심리지수 키워드 분석

(주요 긍정 키워드[1])



(주요 부정 키워드[1])



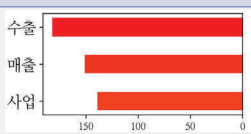
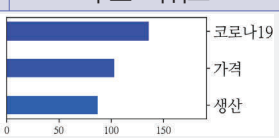
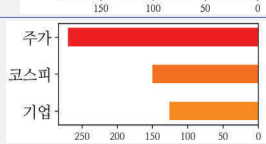
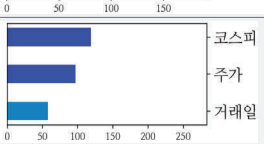
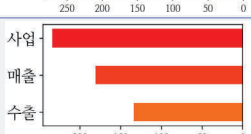
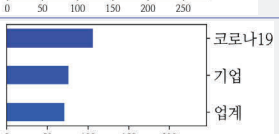
(주요 급상승 키워드[2])

키워드	금주 순위	최근 1개월 대비
오미크론	1	▲106
코스피	2	▲63
예산	17	▲227
돌파	23	▲113
급락	31	▲102
방역	36	▲369

주 : 1) 뉴스문장에 나타난 주요 키워드를 □, 주요 키워드의 연관단어를 ○로 표기한 도표. 교점(node)의 크기는 각 단어의 등장 횟수와 비례하고 연결선(edge)은 두 단어가 함께 등장하는 횟수에 비례
2) 이전 4주간 집계된 키워드 순위 대비 해당주 키워드 순위의 상승폭 기준으로 작성

뉴스심리지수 변동요인을 보다 자세히 살펴보기 위해서는 키워드를 하위 부문별로 나누어 살펴볼 필요가 있다. <표 3>은 위와 동일하게 2021년 12월 첫째 주를 대상으로 주요 긍·부정 키워드를 부문별로 나누어 살펴본 것이다. 금융부문 키워드를 보면 당시 혼조 양상을 보이던 주식시장의 영향이 뉴스심리지수에 반영된 것을 확인할 수 있다. 또한 거시부문에서는 뉴스심리지수 상승 요인으로 11월 수출액 발표가 크게 작용한 반면 산업활동동향의 광공업생산 악화 등은 하락요인으로 작용한 것으로 나타났다. 이와 같이 뉴스심리지수는 지수의 등락뿐 아니라 변동요인을 쉽게 파악할 수 있다는 점에서 설문조사에 의존하는 다른 심리지수에 비해 더 많은 정보를 내포한다고 할 수 있다.

〈표 3〉 **뉴스심리지수 부문별 긍정/부정 키워드**

부문	긍정뉴스		부정뉴스	
	연관단어	주요 키워드	주요 키워드	연관단어
거시	(증가, 11월, 최고, 실적) (분기, 증가, 기록, 성장) (투자, 지원, 글로벌, 성장)			(확산, 변이, 오미크론, 위기) (상승, 석유류, 외식, 원자재) (감소, 자동차, 광공업, 반도체)
금융	(상승, 전망, 실적, 삼성전자) (지수, 외국인, 상승, 코스닥) (성장, 투자, 확대, 미래)			(하락, 지수, 오미크론, 최저) (하락, 삼성전자, 업황, 우려) (포인트, 하락, 코스피, 지수)
산업	(투자, 추진, 소재, 친환경) (분기, 증가, 성장, 파운드리) (증가, 반도체, 11월, 실적)			(확산, 신종, 장기, 변이) (피해, 중소기업, 과징금) (항공, 반도체, 조선, 자동차)

주 : 1) 주요 키워드는 빈도수 순으로 막대 그래프로 배열하고, 각 키워드에 대한 연관단어는 양단에 괄호로 표시

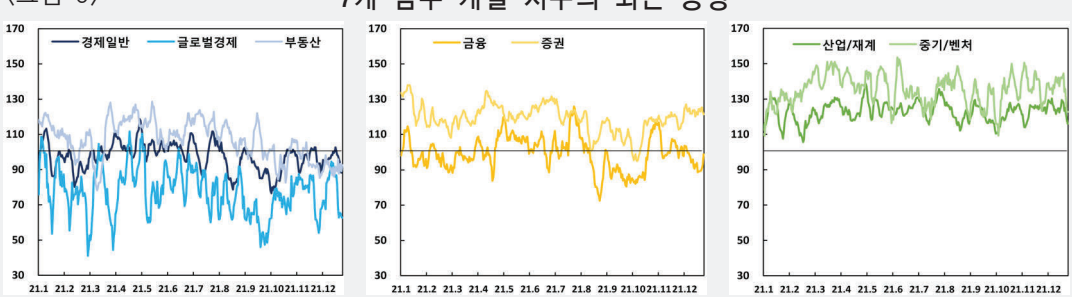
Ⅳ. 뉴스심리지수 활용 방안

현재 뉴스심리지수는 입수 가능한 경제뉴스를 대상으로 전체 부문 지수만 작성하여 공개하고 있다. 그러나 경제심리 변화를 보다 자세히 파악하기 위해서는 경제분야 뉴스기사를 부문별로 나누어 작성하는 것을 고려할 수 있다. 하위부문 지수는 전체 뉴스심리지수의 변동성을 설명하는 데 유용할 뿐만 아니라 경기변동 요인을 세밀하게 파악하는 데도 활용될 수 있다.

현재 한국은행에서 수집하는 뉴스기사는 인터넷 포털사이트 기준으로 경제일반, 글로벌경제, 부동산, 금융, 증권, 산업/재계, 중기/벤처 등 7개 범주로 분류된다. 7개 범주별로 산출한 개별 지수를 살펴보면 지수의 수준과 변동성, 전체 지수와의 상관성, 범주별 표본문장 수 등에서 서로 상이한 모습을 보인다. <그림 9>에서 증권, 중기/벤처 범주의 지수 수준이 다른 범주에 비해 대체로 높고 글로벌경제, 부동산 범주의 지수 변동성이 상대적으로 큰 것을 볼 수 있다.

한편 범주별로 표본문장 수의 비중을 보면 경제일반이 가장 크게 나타나고 증권, 산업/재계가 그 다음을 차지하며 글로벌경제, 중기/벤처[2]는 그 비중이 미미한 것으로 나타나 범주별 비중의 격차가 큰 것을 알 수 있다. 따라서 보다 안정적인 하위 부문 지수 산출을 위하여 인터넷 포털사이트 기준 7개 범주를 지수의 통계적 특성, 범주별 표본문장 비중 등을 감안하여 거시(일반경제, 글로벌경제, 부동산), 금융(금융, 증권), 산업(산업/재계, 중기/벤처) 등 총 3개로 재분류하였다. <그림 10>에서 재분류한 하위부문별 뉴스데이터 비중을 보면 거시부문이 약 절반을 차지하고 금융부문과 산업부문이 비슷한 비중으로 나머지 절반을

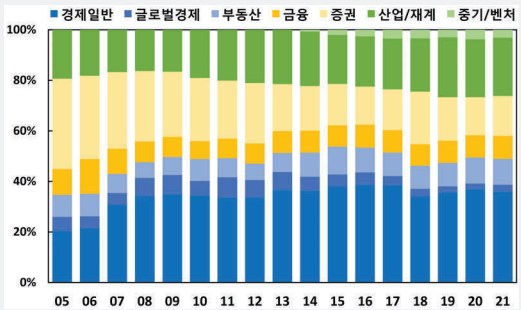
〈그림 9〉 7개 범주 개별 지수의 최근 동향



2) 중기/벤처 범주는 2014년 이후부터 존재한다.

〈그림 10〉 **뉴스데이터 분포**

(연도별)



(하위부문별)

부문	범주	비중(%)[1]	
		15년이후	18년이후
거시	경제일반, 글로벌경제, 부동산	50.0	47.8
금융	금융, 증권	25.2	25.9
산업	산업/재계, 중기/벤처	24.8	26.2

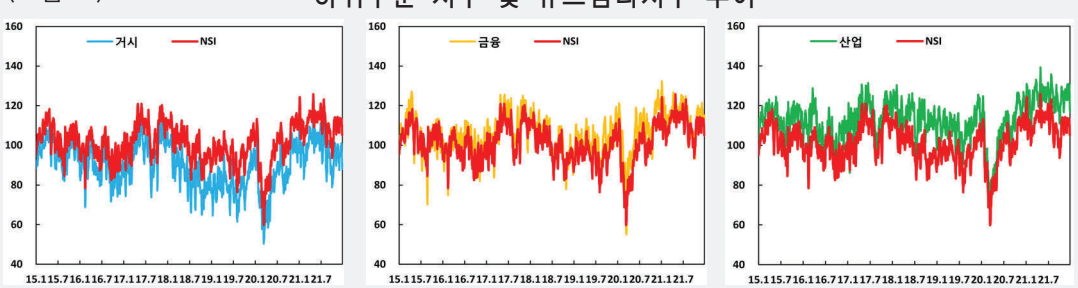
주: 1) 2018년 금융 부문의 비중이 일시적으로 상승

차지한다. 3개 하위부문의 연도별 비중은 2015년 이후 비교적 안정적으로 유지된 것으로 나타나 표본문장을 거시, 금융, 산업 부문으로 분류하고 이들 표본문장에 동일한 감성분류 모형을 적용하여 뉴스심리지수의 하위부문 지수를 작성해보았다.

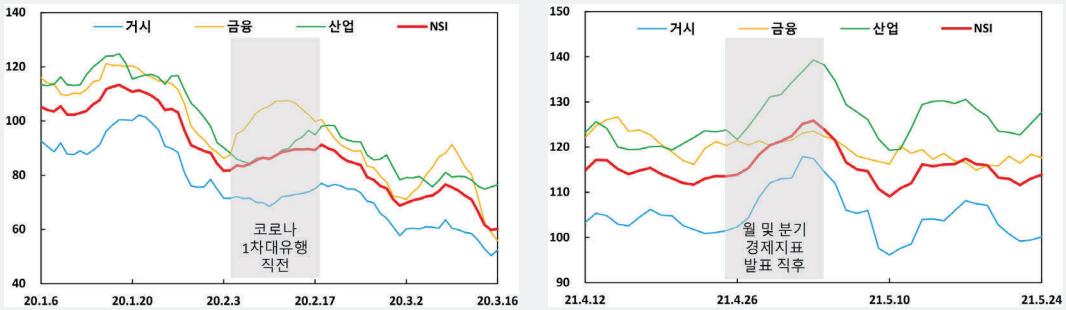
<그림 11>은 2015년 이후 3개 하위부문 지수의 추이를 보여준다. 하위부문 지수는 전체 뉴스심리지수를 기준으로 그 수준과 변동성 측면에서 차이를 보인다. 지수의 수준은 평균적으로 산업부문이 높고 거시부문이 낮은 편이며, 지수의 변동성은 거시부문이 가장 크게 나타난다. 또한 특정 이슈에 대한 영향도 부문별로 다소 상이한 모습을 보인다. <그림 12>를 보면 2020년 코로나 1차 대유행 직전에는 금융지주사 실적 발표 등의 영향으로 금융부문 지수가 크게 상승하였으며, 2021년 4월말에는 양호한 경제지표 발표 등으로 거시부문과 산업부문 지수가 크게 상승한 것을 확인할 수 있다.

한편 <표 4>에서 하위부문 지수와 뉴스심리지수 간의 상관관계를 살펴보면, 거시부문 지수가 뉴스심리지수와 0.9 이상, 여타 부문 지수와 0.7 이상의 높은 상관관계를 보인다. 이러한 지수 간 동조적 흐름이 글로벌 금융위기, 코로나19 등이 포함된 최근에 더욱 뚜렷해진 모습이 특징적이다.

〈그림 11〉 **하위부문 지수 및 뉴스심리지수 추이**



〈그림 12〉 하위부문 지수와 뉴스심리지수의 비교



〈표 4〉 하위부문 지수 및 NSI 간 상관계수[1)]

부문	2015년 이후			2018년 이후			2020년 이후		
	거시	금융	산업	거시	금융	산업	거시	금융	산업
거시	1.00	0.75	0.76	1.00	0.85	0.84	1.00	0.88	0.92
금융		1.00	0.67		1.00	0.71		1.00	0.79
산업			1.00			1.00			1.00
NSI	0.94	0.89	0.88	0.97	0.91	0.91	0.98	0.92	0.95

주: 1) 각 연도 1.1일부터 2021.12.31일까지의 일별 데이터 기준

앞에서 살펴본 바와 같이 3개 하위부문 지수가 뉴스심리지수의 전반적인 흐름을 설명하면서 부문별 특징도 포착하는 장점이 있으므로 뉴스심리지수의 동향을 파악하고 변동요인을 분석하는 데 유용하다. 다만 하위부문 지수를 공개하기 위해서는 부문별 충분한 데이터 확보, 지수의 안정성과 정합성 제고를 위한 추가적인 검토와 연구가 필요하다.

Ⅴ. 시사점 및 향후 과제

뉴스심리지수는 그동안 데이터로 활용하지 못하던 뉴스기사를 데이터로 정제하고, 이를 기계학습이라는 새로운 기법으로 분석하여 작성한 대표적인 실험적 통계이다. 뉴스데이터는 설문조사나 집계를 통하여 수집하는 정형데이터에 비해 입수에 소요되는 시간과 비용이 적으며 입수하는 정보의 양도 방대하여 다양한 분석에 적용할 수 있는 장점이 있다. 따라서 뉴스데이터를 기반으로 작성한 뉴스심리지수는 다음과 같은 의의를 갖는다.

첫째, 뉴스심리지수는 속보성을 갖는 고빈도 경제지표로서 기존의 경기지표를 보완한다. 금융지표를 제외한 여타 경제지표의 경우 고빈도 지표가 거의 전무하며, 비교적 고빈도로 발표되는 지표의 경우에도 공표 시점과 대상 시점이 한 달 이상 차이가 나는 것이 보통이다. 뉴스심리지수는 일별로 작성 가능하며 다른 심리지수 및 실물지표와 유사한 흐름을 보이는 데다 키워드 분석을 통한 변동요인 파악이 용이하므로, 이를 활용하여 경제상황을 보다 신속하게 파악하는 것이 가능하다.

둘째, 뉴스심리지수는 기계학습 방법 및 자동화 과정을 통계 작성에 적용하였다는 점에서 새로운 통계 작성기법의 안정성 및 효용성을 연구할 기반을 제시한다. 통계 작성에는 많은 시간과 비용이 투입되며, 특히 표본조사를 활용할 경우 표본의 대표성을 확보하기 위한 노력이 필요하다. 기계학습 방법을 통계 작성에 이용할 경우 대표본을 활용할 수 있고 자료수집 비용을 크게 절감할 수 있다. 또한 자동화를 통한 통계의 작성은 통계작성자의 의중이나 실수(human error)를 원천적으로 차단할 수 있다는 점에서 미래 통계가 나아갈 방향이라고 할 수 있다. 뉴스심리지수는 기계학습 방법 및 자동화(automation) 프로세스를 선도적으로 도입하여 통계 작성에 있어서 새로운 기법의 발전과 확대에 크게 기여한 것으로 평가된다.

다만 뉴스심리지수의 지속적 발전을 위해서는 하위부문별 지수를 세밀하게 작성하고 키워드 분석을 통하여 지수 변동요인을 구체적으로 파악하는 등, 지수의 적정성 및 안정성을 다양한 각도에서 검증하고 연구할 필요가 있다. 또한 뉴스심리지수를 Nowcasting 등 비교적 짧은 시계의 전망모형에 활용하는 등 활용 가능성을 확대해 나갈 필요가 있다.

디지털·정보화 시대에 다양한 형태의 데이터가 실시간으로 축적되고 있음을 고려할 때, 동 데이터로부터 의미 있는 정보를 추출하는 시도들이 매우 중요해졌음은 자명하다. 앞으로 뉴스심리지수와 같은 실험적 통계의 개발을 활성화하고 통계의 다양성을 확대해 가는 것이 향후 경기 파악 및 경기 전망을 위한 중요한 과정이 될 것으로 기대한다.

참고문헌

- 김한준·조새롬·김동찬 (2021), "경제 텍스트 데이터를 활용한 키워드 분석방안 연구," 국민계정리뷰, 2021년 제1호.
- 김현중·임종호·이혜영·이상호 (2019), "온라인 뉴스 기사를 활용한 경제심리보조지수 개발," 국민계정리뷰, 2019년 제2호.
- 문혜정·이혜영 (2017), "빅데이터의 경제통계 활용 현황 및 시사점," 국민계정리뷰, 2017년 제3호.
- 원중호·이한별·문혜정·손원 (2017), "텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류," 국민계정리뷰, 2017년 제4호.
- 전종준·안승환·이문희·황희진 (2020), "경제용어 감성사전 구축방안 연구," 국민계정리뷰, 2020년 제3호.
- 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.
- Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2021). Business news and business cycles (No. w29344). National Bureau of Economic Research.
- Huang, C., Simpson, S., Ulybina, D., & Roitman, A. (2019). News-based sentiment indicators. International Monetary Fund.
- 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.
- 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.

· 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.