Journal of KIIT. Vol. 22, No. 6, pp. 83-93, Jun. 30, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.6.83

사전학습모델 기반 소비자심리보조지수(S-CCSI) 개발

정건영*[1], 박성호*[2], 이채원*[3], 함승훈*[4], 이강배**

Development of Supplementary Composite Consumer Sentiment Index based on a Pretrained Language Model

Geonyeong Jung*[1], Sungho Park*[2], Chaewon Lee*[3], Seunghoon Ham*[4], and Kangbae Lee**

이 논문은 동아대학교 대학원의 연구비 지원을 받아 수행된 연구임

요 약

경기판단을 위해 실물지표와 더불어 소비자심리지수를 비롯한 심리지표가 도구로 활용되고 있다. 경기판단을 위해 실물지표와 소비자심리지수와 같은 심리지표가 주로 활용된다. 그러나 심리지표는 설문조사 방식을 통해 작성되어 커버리지, 이용 가능 시점, 시간적·경제적 비용에 대한 한계점이 존재한다. 이러한 한계점을 보완하기 위해 본 논문은 KB-BERT 임베딩과 DNN을 활용하여 소비자심리지수를 보조할 수 있는 소비자심리보조지수를 개발한다. 소비자심리보조지수는 현행 심리지표와의 비교 분석과 GDP 경제성장률 예측력 비교를 통해 보조지수로서의 유용성을 검증하였다. 검증 결과 소비자심리보조지수는 기업경기실사지수, 경제심리지수에 약 1~2개월 선행하며, 두 지수와 높은 상관관계가 나타났다. 또한, 소비자심리보조지수를 활용하였을 때 GDP 경제성장률 예측 성능이 전반적으로 높게 나타나는 것 확인했다. 이는 소비자심리보조지수 신뢰할 수 있는 지수이며, 경제 예측에 활용될 수 있음을 시사한다.

Abstract

In economic assessment, sentiment indices such as the Composite Consumer Sentiment Index(CCSI), along with real economic indicators, are utilized as tools. However, sentiment indices, being based on survey methods, have limitations in terms of coverage, availability timing, and temporal and economic costs. To complement these limitations, this paper develops the Supplementary Composite Consumer Sentiment Index(S-CCSI) using KB-BERT embeddings and DNN. The utility of the S-CCSI as a supplementary index is verified through comparative analysis with sentiment indices and comparison of GDP economic growth rate forecasting capabilities. The results show that the S-CCSI leads the BSI and ESI by approximately 1-2 months, and exhibits a high correlation with these indices. Furthermore, when utilizing the S-CCSI, it was confirmed that the performance of GDP economic growth rate forecasting is generally higher. This suggests that the S-CCSI is a reliable index and can be utilized for economic forecasting.

Keywords

KB-BERT, deep learning, CCSI, GDP economic growth rate, DNN

* 동아대학교 경영정보학과
- ORCID[1]: https://orcid.org/0009-0002-6301-9916
- ORCID[2]: https://orcid.org/0000-0001-5419-3774
- ORCID[3]: https://orcid.org/0009-0004-0306-2554
- ORCID[4]: https://orcid.org/0009-0002-5950-2155
** 동아대학교 경영정보학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-6384-4268

· Received: May 07, 2024, Revised: Jun. 07, 2024, Accepted: Jun. 10, 2024
· Corresponding Author: Kangbae Lee
Dept. of Dong-A University, 225, Gudeok-ro, Seo-gu, Busan, Republic of Korea
Tel.: +82-51-200-7448, Email: kanglee@dau.ac.kr

Ⅰ. 서 론

본 연구는 온라인에서 생성되는 방대한 양의 뉴스 기사를 활용하여 한국은행 경제통계시스템에서 매월 발표하는 소비자심리지수(CSI, Composite Consumer Sentiment Index)를 보조할 수 있는 지수를 개발하고자 한다. 소비자심리지수는 경제주체인 소비자의 주관적인 경기판단을 종합한 심리지표로, 국내총생산(GDP, Gross Domestic Production) 및 수출액 등 실물경제에 근거한 실물지표와 함께 경제 상황을 이해하는 데 중요한 도구이다[1]. 소비자심리지수를 포함한 심리지표는 경제주체인 가계와 기업이 현재와 단기적인 미래의 경제 상태를 판단하고, 그에 따라 경제주체들의 경제활동을 조정하는 데 중요한 역할을 한다.

심리지표는 주로 설문조사(Survey)를 통해 매월 작성되어 발표되기 때문에 분기별로 발표되는 GDP와 같은 실물지표보다 빠르게 이용자가 경제 상황과 전망을 파악할 수 있다. 또한, 다른 경기 관련 자료와 달리 경제주체의 주관적이고 심리적인 요소까지 반영해 경제정책을 입안하는 데 중요한 자료로 활용된다. 소비자심리지수는 소비자동향지수 항목 중 현재생활형편, 생활형편전망, 가계수입전망, 소비지출전망, 현재경기판단, 향후경기전망 등 총 6개의 항목의 지수를 합성한 지수로 일반인들이 경제 상황에 대해 전반적으로 어떻게 생각하고 있는지를 알려주는 지표이다. 소비자심리지수는 한국은행 경제통계시스템을 통해 매월 발표되며, 소비자심리지수를 비롯해 한국은행 경제통계시스템을 통해 같이 발표되는 심리지표에는 기업경기실사지수(BSI, Business Sentiment Index), 경제심리지수(ESI, Economic Sentiment Index)가 있다.

그러나 설문조사를 통해 작성되는 심리지표는 몇 가지 한계점을 지고 있다. 첫째, 커버리지(Coverage)에 관한 것이다. 설문조사 기반의 심리지표는 제시된 항목에 한정된 정보만을 수집하여 지수에 반영하기 때문에, 이외의 다른 이벤트가 발생했을 때 그에 대한 경제주체의 인식을 지수에 반영하는 데 한계가 있다. 둘째, 자료의 이용 가능 시점에 있다. 소비자동향조사 기준 설문조사부터 집계, 작성 과정을 거쳐 발표까지 약 15일 소요되므로, 정보이용자가 이용하기까지의 지연이 불가피하다. 마지막으로 설문조사의 시행에 막대한 시간적, 경제적 비용이 발생한다는 것이다.

최근 자연어 처리(NLP, Natural Language Processing) 기술의 발전으로 비정형 데이터(Unstructured data)인 텍스트 데이터에 대한 처리 및 분석이 가능해졌으며, 이에 따라 위와 같은 한계점을 지닌 심리지표의 작성 방식을 보완하기 위해 텍스트 데이터를 활용한 심리지표 개발에 관한 연구가 활발하게 이루어지고 있다. 대표적으로 한국은행 경제통계국은 매일 발간되는 뉴스 기사의 텍스트 분석을 통하여 국민들이 느끼는 경제심리를 파악하고자 뉴스심리지수(NSI, News Sentiment Index)를 개발하여 실험적 통계로 2022년 2월부터 공개하고 있다. 특히 온라인에서 실시간으로 생성되는 뉴스 기사를 활용한 연구가 활발하게 이루어지고 있다. 뉴스 기사는 트위터, 블로그 등 다른 매체보다 키워드, 헤드라인 등이 포함된 표준화된 형식을 갖추고 있으며, 문법적 완성도가 높기 때문에 다른 매체의 텍스트 데이터보다 데이터 분석에 유용하다.

본 연구는 경제 뉴스 데이터를 활용하였다. 경제 뉴스는 경제통계, 경제정책, 경제전망 등 경제상황에 대한 내용을 포함하고 있다. 가계는 이러한 경제 관련 뉴스를 통해 소비, 저축과 같은 경제활동과 관련된 의사결정 하게 되므로 가계의 심리조사 결과는 뉴스 기사와 관련성이 높을 것으로 판단된다.

이에 본 연구는 경제 관련 뉴스 기사를 활용하여 경제활동을 수행하는 소비자의 심리지표인 소비자심리지수가 가진 한계점을 보완하기 위한 보조지수를 개발하고 보조심리지표로서의 유용성을 검증하는 것을 목적으로 한다. 본 연구에서 제시한 소비자심리보조지수(S-CCSI, Supplementary Composite Consumer Sentiment Index) 개발을 위해 소비자의 경제 상황과 관련 키워드를 바탕으로 온라인 뉴스 기사를 수집하였다. 이를 대표적인 사전학습 언어모델인 BERT(Bidirectional Encoder Representations from Transformers)를 기반으로 개발된 KB-BERT와 대표적인 인공신경망 알고리즘인 심층신경망(DNN, Deep Neural Network)을 활용하는 새로운 접근법을 제안한다.

보조지수로서의 유용성을 검증하기 위하여, 상관관계 분석, 추세 분석, 예측력 검증을 수행하였다. 검증결과 기존 심리지표와 상관관계가 높음을 확인하였고, 주요 경제이슈에 대한 추세를 보다 신속하게 파악하였다. 또한, 거시경제지표인 GDP 경제성장률 예측을 통해 본 연구에서 제시한 소비자심리 보조지수의 우수성을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 텍스트를 활용한 심리지표 개발과 경제예측에 대한 선행연구를 중심으로 살펴본다. 3장에서는 소비자심리 보조지수 개발에 관한 전반적인 내용을 다루며, 데이터 수집 및 전처리부터 KB-BERT 임베딩, 소비자심리보조지수 모델에 대한 설명을 한다. 4장에서는 본 연구에서 작성한 소비자심리보조지수에 대한 유용성을 검토하며 분석 결과를 다룬다. 마지막으로 5장에서 연구내용을 요약하고, 시사점과 한계점 그리고 향후 연구과제를 제시하며 마무리한다.

II. 관련 연구

2.1 비정형 텍스트를 활용한 심리지표 개발

비정형 텍스트를 활용한 심리지표 개발에 대한 선행연구로서 M. C. Song and K. S. Shin(2017)[2]은 온라인 뉴스 기사 비정형 텍스트 데이터를 활용하여 '소비자의 경기심리지수'를 생성하였다. 소비자의 경기심리지수 생성을 위해 감성사전(Sentiment lexicon)를 구축하고 이를 기반으로 뉴스 기사에서 추출된 어휘의 감성을 긍정, 부정, 중립으로 분류해 각 감성의 어휘 수를 산출식에 대입하여 소비자의 경기심리지수를 생성하였다. 또한, 경제분석에서의 활용 가능성을 검토하였다. Y. S. Kim et al.(2017)[3]는 소셜 미디어에 내재된 소비자의 직접적이고 즉시성 있는 의견을 경제적 측면에서 활용할 수 있는 온라인 소비자감성지수(e-CCSI) 모형을 제시하고 구현하였다. 온라인 소비자감성지수 생성을 위해 어휘 분류체계(온톨로지)와 감성사전을 구축하고 감성분석을 실시하여 생활형편, 경제상황, 소비와 수입 4가지 영역의 소셜 감성지수를 도출하였다. K. Seki et al.(2021)[4]은 경기에 대한 기업의 심리를 정량화한 기업경기실사지수인 'S-APIR'을 개발하였다. 이를 위해 일간 신문 기사로부터 기업 심리가 반영된 비정형 텍스트 데이터 각 기사의 내용을 문장 단위로 감성 분류하였다. 또한, 특정 사건이 예측된 S-APIR에 얼마나 영향을 미쳤는지를 시계열 분석을 통해 살펴보았다. H. J. Kim et al.(2019)[5]은 온라인 경제뉴스 기사에 내포된 심리를 반영한 '뉴스 경제심리지수'를 개발하여 현행 경제심리지수(ESI, Economic Sentiment Index)를 보조하는 지수로서의 유용성을 검증하였다. 뉴스 경제심리지수 작성을 위해 각 기사를 BSI/CSI에 맞게 분류하고 어휘사전을 기반으로 극성을 긍정/부정으로 분류하여 BSI/CSI 보조지수를 각각 산출하고, 이를 결합한 ESI 보조지수를 생성하였다. 추가적으로 생성된 ESI 보조지수를 이용하여 실질 GDP와 같은 거시 경제활동 지표와의 연관성을 추이 분석(Trend analysis)을 통해 살펴보았다.

2.2 경제예측

경제예측에 관한 선행연구로서 S. G. Jung et al.(2022)[6]은 국민총생산(GDP)이 가진 시차라는 제약을 해소하기 위해 GRU(Gated Recurrent Unit)와 LSTM(Long-Short Term Memory) 기반의 각 경제성장률 예측모형을 구축하여 경제성장률을 예측하였다. S. S. Lim and H. S. Choi(2021)[7]은 온라인 뉴스의 비정형 텍스트 데이터를 활용하여 DNN 알고리즘을 적용해 기업경기실사지수, 소비자동향지수를 예측하고 예측된 기업경기실사지수, 소비자동향지수를 설명변수로 설정하여 실질 GDP(RGDP, Real Gross Domestic Product)를 실시간으로 예측하는 모형을 구축하였다.

본 연구는 설문조사 기반 소비자심리지수의 한계점을 보완하기 위해 소비자 관련 경제 뉴스 기사의 텍스트를 지수화한 소비자심리보조지수를 개발하며, 생성된 소비자심리보조지수의 유용성을 검토하고자 한다. 비정형 텍스트를 활용한 심리지표 개발에 관한 선행연구는 대부분 사전접근방법(Sentiment lexicon-based approach)을 활용함으로써 감성사전을 구축하고 이를 기반으로 감성 어휘 단어의 수를 계산하여 심리지표를 산출하였다. 이러한 사전접근법의 경우 감성사전 구축 과정에서 많은 비용이 발생

과 연구자의 주관이 반영 그리고 감성 단어의 수를 계산하는 데 중점을 두어, 문맥상의 의미나 어휘 간 복잡한 관계를 포착하지 못하는 한계점을 가진다. 본 연구는 선행연구와 달리, 사전접근법이 아닌 KB-BERT를 사용하여 텍스트 데이터를 문맥적 특성과 어휘 간 복잡한 관계가 고려된 임베딩 데이터로 변환하여 이를 딥러닝 알고리즘 활용해 기존 소비자심리지수를 보조할 수 있는 심리지표의 개발 모형을 제시하였다.

III. 연구 방법

3.1 연구 프로세스

그림 1은 연구 프로세스를 도식화한 그림이다. 본 연구는 먼저, 소비자의 경제 상황과 관련 키워드를 바탕으로 온라인 뉴스 기사를 웹 크롤링(Web crawling)을 사용해 수집하고 중복 및 광고성 기사를 배제하는 데이터 필터링(Data filtering) 작업과 텍스트 데이터를 인공지능 모형에 활용할 수 있도록 변화하는 데이터 전처리(Data preprocessing) 과정을 수행한다.

다음으로 각 기사 제목의 텍스트를 경제·금융 도메인에 특화된 사전학습모델인 KB-BERT를 거쳐 학습 모델이 이해할 수 있는 수치화된 형태로 임베딩한다. 이를 통해 수치화된 데이터는 문장 내 복잡한 특성과 의미 구조를 담고 있다. 이후 임베딩된 데이터를 Input Layer의 Input Node로 구성하여 DNN 알고리즘 기반의 소비자심리보조지수 모델을 구축하고 월별 소비자심리보조지수를 산출한다.

마지막으로 월 단위로 산출된 소비자심리보조지수를 기존 기업경기실사지수(BSI), 경제심리지수(ESI)와 비교 분석하고 GDP 경제성장률 예측 모형에 소비자심리보조지수 활용유무에 따른 예측력 차이를 통해 보조지수로서의 유용성을 검토하고자 한다.

3.2 데이터

소비자심리보조지수 작성을 위한 뉴스 기사는 웹 크롤링 기법을 사용하여 네이버 온라인 뉴스에서 수집하였다. 웹 크롤링은 인터넷에 노출된 데이터를 직접 다운로드하여 수집하여 방식이다. 키워드 '가계 경기'를 설정하여 웹 크롤링을 통해 관련 기사를 수집하였다. 데이터 수집기간은 2013년 1월부터 2022년 12월로 설정하여 약 10년간의 데이터 56,307건을 수집하였다.

뉴스 기사의 본문이 아닌 뉴스 기사의 제목을 분석 대상으로 설정하였다. 뉴스 기사의 본문은 복합적인 주제가 포함된 내용으로 인해 주제에서 벗어난 정보가 포함될 가능성이 있기 때문이다. 반면 제목은 기사의 전체적인 논조를 반영하여 본 연구는 뉴스 기사의 제목을 분석대상으로 설정하였다.

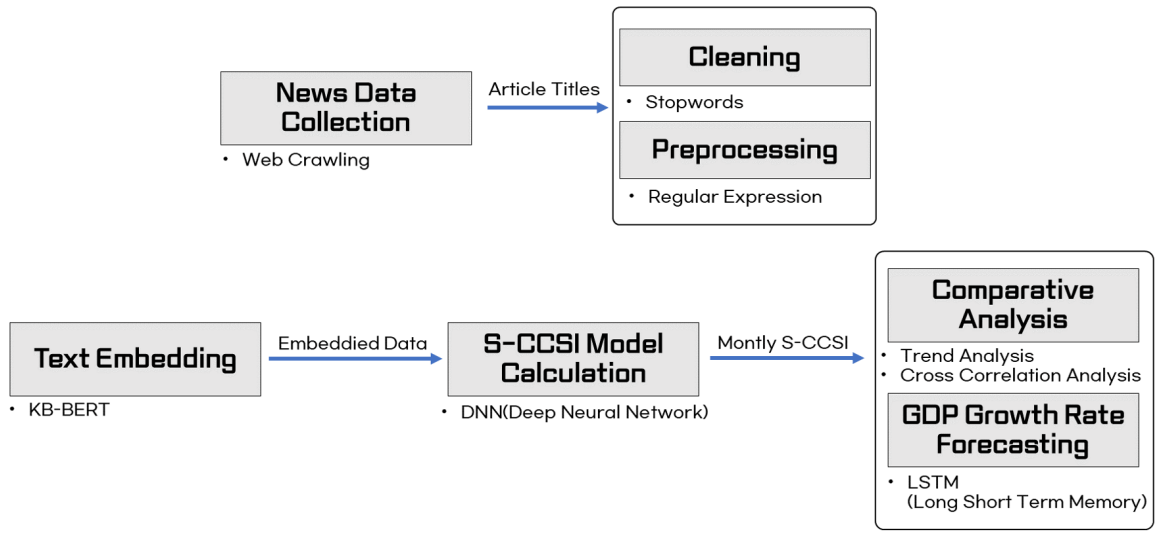


그림 1. 연구 프로세스
Fig. 1. Research process

3.2.1 데이터 필터링 & 전처리

웹 크롤링을 통해 데이터를 수집하는 경우 정형화된 데이터를 입수하는 것과 달리 인터넷 포털의 운영 상황이나 방침에 따라 중복된 뉴스 기사 또는 광고성 기사가 함께 수집된다. 이는 데이터의 정확성, 유효성이 보장되지 않아 데이터 무결성을 해친다는 것을 의미한다. 이를 처리하기 위해 중복된 기사를 배제하고 기사의 본문 내용이 비어있는 기사를 배제하는 데이터 필터링 작업을 수행하였다.

또한, 웹 크롤링을 통해 수집된 뉴스 기사는 원본 형태의 원천 데이터로, 이러한 텍스트 데이터는 모델 학습에 적합하지 않다. 따라서, 수집된 뉴스 기사를 모델 학습에 활용할 수 있도록 변환하는 전처리 과정이 필요하다.

본 연구에서 수집된 뉴스 기사는 자연어 처리에 방해가 되는 특수기호나 특수문자 등이 포함되어 있다. 이러한 문제를 해결하기 위해 특수기호나 특수문자를 제거하는 전처리 과정을 수행하였다. 이 과정에서 활용되는 도구로 정규표현식(Regular expression)을 사용하였다. 이 도구는 특정한 규칙을 가진 문자, 문자열을 추출, 제외, 변환하는 데 유용하다. 본 연구에서는 파이썬(Python)에서 제공하는 ‘re’ 라이브러리를 활용하여 구현하였다. 표 1은 전처리 전후 데이터 수를 비교한 표이다.

표 1. 전처리 전후 데이터 수 비교
Table 1. Comparison of data quantities before and after preprocessing

Year	Num of raw data	Num of pre-processed data
2013	5,374	5,354
2014	5,977	5,946
2015	6,333	6,288
2016	6,175	6,137
2017	5,922	5,898
2018	4,846	4,824
2019	4,564	4,549
2020	5,009	4,996
2021	5,795	5,778
2022	6,312	6,211
Sum	56,307	55,981

3.3 KB-BERT 임베딩

본 연구에서는 인간이 사용하는 자연어를 인공지능 모형이 이해할 수 있는 형태로 변환하기 위해 경제·금융 도메인에 특화된 KB-BERT를 사용하여 임베딩 작업을 수행하였다.

이러한 과정은 기계 번역, 감성분석, 문서 분류 등 다양한 자연어 처리(NLP) 태스크의 성능에 많은 영향을 미치기 때문에 중요하다. 본 연구에서 사용한 KB-BERT는 2018년 공개된 Google BERT를 Base Model로 하는 사전학습 언어모델로, 문맥을 반영한 임베딩(Contextual embedding)을 사용하고 있다. 문맥을 반영한 임베딩 방식은 원-핫 인코딩(One-hot encoding)과 같은 전통적인 언어모델부터 시작하여 Word2Vec, GloVe 등의 의미기반 언어모델의 한계를 극복하기 위해 나온 언어모델이다[8].

원-핫 인코딩은 단어 집합 내 표현하고자 단어의 인덱스만 1로 표시하고, 이외 다른 단어의 인덱스는 0으로 표시하여 단어를 벡터로 나타낸다. 예를 들어, ‘가계 대출’, ‘소비자물가’라는 2개의 단어만 존재한다고 했을 때, ‘가계 대출’은 [1, 0], ‘소비자 물가’는 [0, 1]로 표현된다. 이러한 전통적인 언어모델은 단어 간의 의미적 관계를 전혀 표현하지 못하며, 단어 집합의 크기에 따라 차원 수가 증가하여 자연어 처리 과정에서 많은 컴퓨팅 자원 소모와 시간적 비용을 발생시킨다는 단점이 있다. Word2Vec, GloVe 등의 의미기반 언어모델은 단어 간의 의미적인 연관성을 반영하여 언어를 표현한다. 의미를 고려한 언어표현을 위해 언어모델은 중심 단어를 주변 단어들로부터 예측하거나 주변 단어들을 중심 단어로 예측하도록 학습함으로써, 유사한 의미를 지닌 단어들이 벡터 공간상 가까운 거리에 분포된다. 이는 원-핫 인코딩의 한계를 극복했지만, 문맥 정보를 반영하지 못한다는 한계를 가진다. 예를 들어, ‘길거리에 많은 은행들이 땅에 떨어져 있다’와 ‘은행에서 돈을 인출했다’와 같이 같은 문자임에도 사용되는 문맥에 따라 다른 의미 지니는 데 이러한 차이를 구분하지 못한다[9]. 반면 문맥기반 언어모델인 BERT는 주어진 문장 내에서 단어의 위치에 따라 단어의 의미가 어떻게 변화하는지 문맥을 학습하여, 결과적으로 문맥에 따른 단어의 의미 차이를 구분이 가능하다.

그러나 BERT는 사전학습에 사용된 말뭉치의 특성에 따라 하위 태스크의 성능에 많은 영향을 미치기 때문에 영어 위키백과와 Book-Corpus로 학습된 Google BERT는 본 연구에 필요한 한국어 경제 및 금융 관련 텍스트 처리에는 적합하지 않다고 판단하였다[10]. 이러한 문제점을 근거로 한국어 위키백과, 경제 및 금융 뉴스 및 문서로 구성된 말뭉치로 사전학습된 KB-BERT가 본 연구에 필요한 도구로 적합하다고 판단하여 임베딩 작업에 활용하였다. S. H. Park et al.(2023)[11]는 온라인 뉴스 감성 분류에 KB-BERT를 사용함으로써 온라인 뉴스를 활용할 수 있는 방안을 확인하였다.

본 연구에서는 뉴스 기사 제목의 텍스트를 변환하기 위해 정보 손실 최소화와 컴퓨팅 자원을 고려하여 주어진 텍스트 데이터를 모델이 처리할 수 있는 최대 토큰 수를 45개로 설정하여, 자연어 형태인 각 기사의 텍스트 데이터를 768차원의 수치형 데이터로 변환하였다.

3.4 소비자심리보조지수(S-CCSI) 모델

본 연구는 대표적인 딥러닝 알고리즘인 DNN 활용하여 소비자심리보조지수 모델을 구축하였다. 소비자심리보조지수 생성을 위해 KB-BERT를 거쳐 임베딩된 데이터를 DNN 모델의 Input Layer로 구성하고 Hidden Layer와 Output Layer를 구성하였다. 텍스트 기반 심리지수 개발에 관한 선행연구는 대부분 감성 어휘의 수를 분류하여 지수를 산출하는 방식을 사용하였다. 이러한 방식은 특정 감성 단어가 포함된 양을 계산하는 데 중점을 두지만, 문맥상의 의미나 어휘 사이의 복잡한 관계를 완전히 포착하지 못하는 경우가 존재한다. 이러한 문제점을 인식하고 보완하기 위해 본 연구는 단어의 문맥적 관계와 의미 변화를 포괄적으로 이해하고자 소비자심리보조지수 개발에 임베딩된 데이터를 입력변수로 활용하였다. J. Devlin et al.(2019)[12]는 문맥기반 언어모델인 BERT를 다양한 NLP 태스크에 적용하여 각 태스크의 성능을 기존 언어모델과 비교했을 때 우수한 성능을 입증하였다. 이는 문맥기반 언어모델이 전통적인 언어모델, 의미기반 언어모델보다 단어의 문맥적 관계와 의미 변화를 더 잘 포착할 수 있음을 시사한다. 본 연구는 대표적인 딥러닝 알고리즘인 DNN 활용하여 소비자심리보조지수 모델을 구축하였다. 소비자심리보조지수 생성을 위해 KB-BERT를 거쳐 임베딩된 데이터를 DNN 모델의 Input Layer로 구성하고 Hidden Layer와 Output Layer를 구성하였다. 텍스트 기반 심리지수 개발에 관한 선행연구는 대부분 감성 어휘의 수를 분류하여 지수를 산출하는 방식을 사용하였다. 이러한 방식은 특정 감성 단어가 포함된 양을 계산하는 데 중점을 두지만, 문맥상의 의미나 어휘 사이의 복잡한 관계를 완전히 포착하지 못하는 경우가 존재한다. 이러한 문제점을 인식하고 보완하기 위해 본 연구는 단어의 문맥적 관계와 의미 변화를 포괄적으로 이해하고자 소비자심리보조지수 개발에 임베딩된 데이터를 입력변수로 활용하였다. J. Devlin et al.(2019)[12]는 문맥기반 언어모델인 BERT를 다양한 NLP 태스크에 적용하여 각 태스크의 성능을 기존 언어모델과 비교했을 때 우수한 성능을 입증하였다. 이는 문맥기반 언어모델이 전통적인 언어모델, 의미기반 언어모델보다 단어의 문맥적 관계와 의미 변화를 더 잘 포착할 수 있음을 시사한다.

DNN 모델은 입력층 512개의 노드, 중간층 256개의 노드로 2층, 출력층 1개의 노드로 구성하였다. 과적합 방지를 위해 드롭아웃(Dropout) 비율을 0.3로 설정하여 적용하였다. 출력값은 sigmoid 함수를 사용하여 0~1범위의 값으로 출력하고 설문조사 기반의 심리지표와 동일하게 0~200범위를 가지는 지수값으로 조정하여 초기 월별 소비자심리보조지수를 산출하였다. 이후, 초기 소비자심리보조지수 고도화를 소비자심리지수 데이터를 목적변수로 설정하여 초기 소비자심리보조지수를 피팅(Fitting)하였다. 뉴스 기사는 시장의 중요한 이슈나 경제 동향을 반영하므로, 이를 통해 얻은 정보를 기존 소비자심리지수에 피팅함으로써 시장 변화를 반영할 수 있다. 이러한 과정은 보조지수의 타당성을 강화하고, 기존 지수의 한계를 보완할 수 있다.

Ⅳ. 유용성 검토

4.1 소비자심리보조지수 산출

소비자심리보조지수는 0~200범위를 가지는 지수값으로 조정하여 2013년 1월부터 2022년 12월까지의 월별 소비자심리보조지수를 산출하였다. 표 2는 현행 소비자심리지수와 비교한 표이다. 소비자심리지수와 소비자심리보조지수와의 상관계수는 0.84으로 강한 상관관계가 나타났으며, 이는 소비자심리보조지수가 기존 소비자심리지수를 대체할 수 있음을 시사한다.

표 2. 월별 소비자심리지수와 소비자심리보조지수
Table 2. Monthly CCSI & S-CCSI

Date	CCSI	S-CCSI
2013.01	101.8	101.6
2013.02	101.1	100.5
2013.03	104.2	103.3
2013.04	101.6	100.8
2013.05	104.1	103.4
…	…	…
2022.10	89	91.7
2022.11	86.6	89.7
2022.12	90.1	91.9

4.2 기업경기실사지수(BSI) 및 경제심리지수(ESI) 비교

산출된 소비자심리보조지수를 기존의 공식 심리지표인 기업경기실사지수(BSI)와 경제심리지수(ESI) 월 단위로 작성하여 비교 분석하였다.

기존의 두 심리지표와의 비교 분석한 결과, 그림 2와 같이 소비자심리보조지수는 두 심리지표와 전반적으로 유사한 흐름을 보이는 것을 확인할 수 있다. 특히, 흐름이 전환되는 변곡점을 기준으로 살펴보았을 때 소비자심리보조지수가 이러한 변곡점을 두 심리지표보다 신속하게 포착하는 것을 확인할 수 있다. 예를 들어, 그림 2를 보면 소비자심리보조지수는 국내 코로나 19 첫 확진자가 발생한 2020년 1월에 포착하여 특정 이슈가 기업경기실사지수보다 빠르게 심리지수에 반영되는 것을 확인할 수 있다.

다음으로 표 3은 소비자심리보조지수를 교차상관분석 결과이다. 분석 결과 소비자심리보조지수는 소비자심리지수에 약 1개월 선행하고 0.68의 상관관계를 보인다. 또한, 경제심리지수에는 약 2개월 선행하며 0.81의 강한 상관관계를 보인다.

본 연구에서 제시하는 소비자심리보조지수와 기업경시실사지수 및 경제심리지수를 비교 분석한 결과, 소비자심리보조지수는 두 지수와 강한 상관관계를 보이며, 약 1~2개월 선행하는 것을 확인했다. 이는 소비자심리보조지수가 두 심리지표의 선행지표로서 유용하다는 것을 시사한다.

따라서 소비자심리보조지수를 주 단위로 작성한다면 월별로 발표되는 기존의 소비자심리지수가 가진 정보 이용 시점에 대한 한계점을 보완할 수 있을 것이며, 이는 정보이용자가 신속하게 정보에 접근하여 신속한 경기판단에 도움이 될 것으로 기대된다.

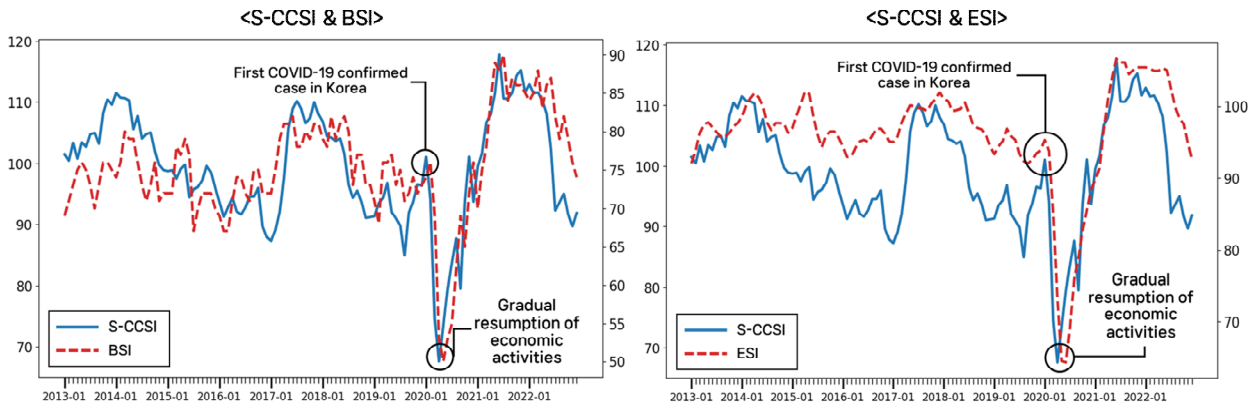


그림 2. 기존 심리지표와 추이 비교
Fig. 2. Trend comparison with original sentiment index

표 3. 교차상관분석 결과
Table 3. Cross correlation analysis results

Sentiment index	Maximum correlation coefficient	Maximum cross-correlation lag
CCSI	0.84	0
BSI	0.68	-1
ESI	0.81	-2

4.3 GDP 경제성장률 예측

본 연구는 보조지수로서 또 다른 기능인 시너지 효과(Synergy effect)를 검증하기 위해 거시경제지표인 GDP 경제성장률 예측모형을 구축하여 소비자심리보조지수의 활용 유무에 따른 예측력 차이를 비교한다.

GDP 경제성장률은 특정 기간 동안 각 경제활동 부문이 생성한 부가가치의 연간 증가율을 측정하여 경제의 성과를 평가하는 주요 지표로 사용된다. 그러나 GDP 경제성장률은 분기별로 발표되기 때문에, 데이터의 적시성에 제약을 받는다는 중요한 한계를 갖고 있다. 따라서 이러한 문제를 해결하고자 분기별 GDP 경제성장률 데이터를 월별 데이터로 세분화하여 추정하였다. 또한, 기존 관련 연구(e.g, 오현희, 2021; 김덕파·강성진, 2020)를 바탕으로 국내총생산과 연관성이 높을 것으로 기대되는 월별 자료로 예측모형의 설명변수를 설정하였다.

또한, 최근 경제예측모형에 관한 연구에서 많이 활용되고 있는 LSTM(Long-Short Term Memory)을 기반으로 예측모형을 구축하였다. LSTM은 전체 체인을 관통하는 셀 스테이트(Cell state)를 통해 과거 학습결과를 큰 변함없이 전달하는 구조로 과거의 RNN(Recurrent Neural Network)이 가지고 있는 단기 기억(Short memory) 한계점을 극복하여, 장기간 정보를 유지할 수 있어 시계열 데이터 분석에 적합하다[13]. 뿐만 아니라, 비선형적이고 동적인 관계를 더 잘 파악할 수 있다는 특징을 가진다.

본 연구는 먼저, 소비자심리보조지수 활용 유무에 따른 GDP 경제성장률 예측력을 비교하고 다음으로, 현재 시점부터 미래의 여러 예측 시점의 GDP 성장률을 예측하여 선행연구 결과와 비교하였다. 연구에서 사용된 학습 데이터(Training data)는 2013년 1월부터 2020년 10월까지이며, 테스트 데이터(Test data)는 2021년 1월부터 2022년 12월까지이다.

아래의 표 4는 소비자심리보조지수의 활용 유무와 과거 자료 활용 정도에 따른 GDP 성장률 예측 결과를 보여준다. 예측의 정확성은 평균 절대 오차(MAE, Mean Absolute Error)를 사용하여 측정하였다. 결과적으로, 소비자심리보조지수를 함께 사용했을 때 예측 성능이 전반적으로 높게 나타나는 것을 확인할 수 있다. 이는 소비자심리보조지수가 기존 소비자심리지수와의 함께 활용될 수 있음을 시사한다. 또한, 설명변수의 과거 자료를 많이 활용할 수 더 많이 활용했을 때 예측력이 높게 나타났다.

표 4. GDP 경제성장률 예측 결과(MAE)
Table 4. Results of GDP growth rate prediction(MAE)

Extent of past data utilization	Inclusion of S-CCSI	Exclusion of S-CCSI
5	**0.3407**	0.4453
4	**0.3561**	0.3848
3	**0.4921**	0.5042
2	**0.4162**	0.5073
1	0.5950	**0.4775**

다음으로 5개월의 과거 자료를 이용하여 현재 시점부터 여러 미래 시점의 GDP 성장률을 예측하고 이를 기존 선행연구(e.g, 정수관·표동진, 2022)와 비교했을 때 소비자심리보조지수를 포함한 본 연구의 예측모형의 예측력이 높게 나타나는 것을 확인하였다[6]. 표 5는 선행연구와 비교한 결과이다. 이러한 결과는 소비자심리보조지수가 경제예측에 있어 중요한 도구로 활용될 수 있음을 시사한다.

표 5. GDP 경제성장률 시점별 예측 결과(MAE)
Table 5. Results of GDP growth rate prediction per time point(MAE)

Forecasting point	Inclusion of S-CCSI	Prior research
y	**0.2974**	0.3745
y+1	**0.3407**	0.4587
y+2	**0.4028**	0.6081
y+3	**0.5415**	0.7053

V. 결 론

가계, 기업, 정부 등 경제주체는 소비자심리지수를 비롯한 다양한 심리지표를 경기에 대한 판단을 하기 위한 근거자료로 활용하고 있다. 그러나 현재 한국은행에서 매월 발표하는 심리지표는 주로 설문조사 방식을 통해 작성되므로 커버리지, 자료의 이용 가능 시점, 시간적·경제적 비용에 대한 한계점을 지닌다. 이러한 한계점을 극복하기 위해 최근 텍스트 데이터를 활용한 심리지표 개발에 관한 연구가 활발히 진행되고 있다.

이에 본 논문에서는 온라인 뉴스 기사를 활용한 심층신경망(DNN) 모델을 구축하여 소비자심리보조지수를 개발하였다. 소비자심리보조지수 개발을 위해 사전접근법을 사용한 대부분의 선행연구와 달리, 본 연구에서는 KB-BERT와 DNN을 활용한 새로운 방법론을 제시하였다. 금융·경제 도메인에 특화된 사전학습 모델인 KB-BERT를 통해 각 기사 제목의 텍스트를 문장 내 복잡한 특성과 의미 구조를 담고 있는 수치화된 형태로 변환하였다. 이후, 수치화된 데이터를 심층신경망 알고리즘의 Input Layer로 입력하여 기존 소비자심리지수를 보조할 수 있는 소비자심리보조지수를 월 단위로 산출하였다.

산출된 소비자심리보조지수의 유용성을 설문조사 기반 심리지표와 비교와 GDP 경제성장률 예측을 통해 검토하였다. 먼저, 설문조사 기반 기업경기실사지수, 경제심리지수와 비교 분석한 결과 소비자심리보조지수는 두 심리지표와 강한 상관관계가 나타났으며, 두 심리지표에 1~2개월 선행하는 것으로 나타났다. 이러한 분석 결과는 소비자심리보조지수가 두 심리지표의 선행지수로서의 유용성을 시사하며, 소비자심리보조지수 작성 이점상 최소한 주 단위로 발표 주기를 단축하여 월별로 발표되는 소비자심리지수가 가진 적시성에 대한 한계점을 보완할 수 있을 것이다. 다음으로, 소비자심리보조지수 활용 유무에 따른 GDP 경제성장률 예측력 평가를 통해 소비자심리보조지수가 소비자심리지수와 함께 활용되었을 때 더 나은 성능이 나타난 것을 확인했고 기존 연구결과와 비교를 통해 더 우수한 예측 성능을 나타난 것을 확인하였다.

그러나 본 연구는 몇 가지 한계점을 지닌다. 첫째, 데이터의 선별성(Representativeness)이다. 본고에서 소비자심리보조지수 개발을 위해 키워드 '가계 경기'만을 설정하여 관련 데이터를 수집하여 활용하였다는 점이다. 둘째, GDP 경제성장률 예측 단계에서 분기별 GDP 경제성장률을 월별 GDP로 세분화하기 위해 선형보간법을 사용하였는데, 이는 선형적인 변화를 가정하여 추정하므로 시계열 데이터의 비선형적 특성을 잘 반영하지 못할 수도 있다는 점이다.

본 연구 결과와 한계점을 토대로 몇 가지 향후 연구과제를 제시하면 다음과 같다. 첫째, '가계 경기' 외 가계와 관련된 주요 경제 키워드를 설정하여 충분한 데이터를 확보하여 소비자심리보조지수 산출에 활용하는 것이다. 둘째, 주기가 더 짧은 소비자심리보조지수를 작성하는 것이다. 본 연구에서는 소비자심리보조지수를 월 단위로 산출하여 대체가능성을 검증하였다. 따라서 최소한 주 단위의 소비자심리보조지수를 작성하여 이를 검증할 필요가 있다. 마지막으로 주기 불일치, 결측치(Missing value) 문제를 해결하고 다양한 경제지표를 입력변수로 활용이 가능한 동적요인모형(DFM, Dynamic Factor Model)을 분기별 GDP 경제성장률 데이터를 월별 데이터로 추정하여 모형을 구축하고 예측하는 것이다[13].

References

[1] C. H. Kim, T. Y. Kim, I. H. Kim, and J. J. Ahn, "A study on the improvement of the economic sentiment index for the Korean economy", Journal of the Korean Data and Information Science Society, Vol. 26, No. 6, pp. 1335-1351, Nov. 2015. http://dx.doi.org/10.7465/jkdi.2015.26.6.1335.

[2] M. C. Song and K. S. Shin, "Construction of Consumer Confidence index based on Sentiment analysis using News articles", Journal of Intelligent Information Systems, Vol. 23, No. 3, pp. 1-27, Sep. 2017. http://dx.doi.org/10.13088/jiis.2017.23.3.001.

[3] Y. S. Kim, S. G Hong, H. J. Kang, and S. R Jeong, "Electronic-Composit Consumer Sentiment Index(CCSI) development by Social Bigdata

Analysis", Journal of Internet Computing and Services, Vol. 18, No 4, pp. 121-131, Aug. 2017. http://dx.doi.org/10.7472/jksii.2017.18.4.121.

[4] K. Seki, Y. Ikuta, and Y. Matsubayashi, "News-based business sentiment and its properties as an economic index", Journal of Information Processing and Management, Vol. 59, No. 2, pp. 102795, Mar. 2022. https://doi.org/10.1016/j.ipm.2021.102795.

[5] H. J. Kim, J. H. Lim, H. Y. Lee, and S. H. LEE, "Development of News Economic Sentiment Index using Online news articles", Bank of Korea, Jun. 2019.

[6] S. G. Jung, D. J. Pyo, and D. Y. Heo, "A Study on Real-time Forecasting GDP and GRDP Using Deep Learning", Ordo Economics Journal, Vol. 26, No. 2, pp. 17, Jun. 2023.

[7] S. S. Lim and H. S. Choi, "Real-time Forecasting of Real GDP using Text Mining", Journal of Corporation and Innovation, Vol. 44, No. 4, pp. 91-106, Dec. 2021. http://dx.doi.org/10.22778/jci.2021.44.4.91.

[8] C. J. Park, W. S. Lee, Y. K. Kim, J. H. Kim, and H. S. Lee, "Survey on Large Language Models", Communications of KIISE, Vol. 41, No. 11, pp. 8-24, Nov. 2023.

[9] S. H. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings", Journal of Expert Systems With Applications, Vol. 117, pp. 139-147, Mar. 2019. https://doi.org/10.1016/j.eswa.2018.08.044.

[10] D. G. Kim, D. Y. Lee, J. W. Pa, and S. W. Oh, "KB-BERT: Korean pre-trained language model specialized Finance and Utilization", Journal of Intelligent Information System, Vol. 28, No. 2, pp. 191-206, Jun. 2022. http://dx.doi.org/10.13088/jiis.2022.28.2.191.

[11] S. H. Park, H. W. Lee, Y. B. Jo, and K. B. Lee, "An Investigation into the Impact of Pretrained Language Models on Online News Sentiment and its Influence on Investor-Specific Trading Intensity", Asia-pacific Journal of Convergent Research Interchange, Vol. 9, No. 10, pp. 225-235, Oct. 2023. http://dx.doi.org/10.47116/apjcri.2023.10.19.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 4171-4186, Jun. 2019. https://doi.org/10.48550/arXiv.1810.04805.

[13] D. H. Shin, G. H. Choi, and C. B. Kim, "Deep Learning Model for Prediction Rate Improvement of StockPrice Using RNN and LSTM", Journal of KIIT, Vol. 15, No. 10, pp. 9-16, Oct. 2017. http://dx.doi.org/10.14801/jkiit.2017.15.10.9.

[14] H. C. Lee, D. G. Choi, and Y. G. Kim, "Dynamic Factor Model and Deep Learning Algorithm for GDP Nowcasting", Bank of Korea, Vol. 28, No. 2, Jun. 2022.

저자소개

정 건 영 (Geonyeong Jung)

2021년 2월 : 부경대학교 재무회계학과(경영학사)
2022년 9월 ~ 현재 : 동아대학교 경영정보학과 석사과정
관심분야 : 자연어 처리, 딥러닝, 빅 데이터, 경제예측

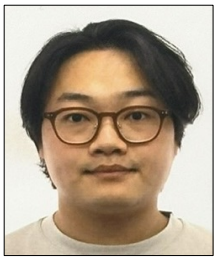
박 성 호 (Sungho Park)

2017년 2월 : 동아대학교 경영정보학과(경영학사)
2019년 2월 : 동아대학교 경영정보학과(경영학석사)
2023년 2월 : 동아대학교 경영정보학과(경영학박사)
관심분야 : 자연어 처리, 딥러닝, 경제예측, 시계열 분석

이 채 원 (Chaewon Lee)

2023년 2월 : 동아대학교
경영정보학과(경영학사)
2023년 3월 ~ 현재 : 동아대학교
경영정보학과 석사과정
관심분야 : 자연어 처리, 딥러닝,
경제예측, 시계열 분석

함 승 훈 (Seunghoon Ham)



2024년 2월 : 동아대학교
경영정보학과(경영학사)
2024년 3월 ~ 현재 : 동아대학교
경영정보학과 석사과정
관심분야 : 거대언어모델, 생성AI,
자연어처리

이 강 배 (Kangbae Lee)

1989년 2월 : 고려대학교
산업공학과(공학사)
1991년 2월 : 한국과학기술원
산업공학과(공학석사)
1995년 8월 : 한국과학기술원
산업공학과(공학박사)
2004년 9월 ~ 현재 : 동아대학교
경영정보학과 교수
관심분야 : 인공지능, MIS, 경제예측, 시계열 분석,
딥러닝