온라인 뉴스 기사를 활용한 경제심리보조지수 개발

김현중[a], 임종호[b], 이혜영[c], 이상호[d]

본 연구의 목적은 경제 상황에 대한 시장 참여자들의 인식과 관련하여 온라인 경제뉴스 기사에 내포된 심리를 나타내는 '뉴스 경제심리지수'를 개발하는 것이다. 이를 위하여 속성 기반 오피니언 마이닝(ABOM: Aspect-based Opinion Mining) 기법이 적용되었다. 동 방법은 온라인 리뷰나 뉴스 기사와 같은 비정형 텍스트 빅데이터의 정보를 요약하는 데 자주 사용된다. 본고에서는 ABOM 방법 중 속성 탐지와 감성 분석을 함께 수행하는 결합 방법이 고려되었다. 구체적으로, (1) 온라인 뉴스 기사를 특정 주제에 매핑(mapping)하여 속성을 탐지하고 (2) 각 주제에 할당된 기사들을 대상으로 감성점수를 계산하였다. 인터넷 포털에서 수집한 실제 뉴스 기사에 대하여 기계학습(machine-learning) 접근법과 룰기반(rule-based) 접근법을 모두 적용하여 보았다. 비정형 빅데이터인 뉴스 기사를 활용하여 이번에 개발한 새로운 경제심리지수 NESI (News ESI)는 현행 ESI와 유사한 움직임을 보일 뿐만 아니라 거시 경제지표와도 연관성을 가지는 것으로 나타남에 따라 현행 ESI의 보조지수로서 의미 있는 역할을 할 수 있을 것으로 기대된다.

a. 연세대학교 응용통계학과 교수(e-mail: hkim@yonsei.ac.kr, phone: 02-2123-2545)
b. 연세대학교 응용통계학과 교수(e-mail: ijh38@yonsei.ac.kr, phone: 02-2123-2539)
c. 한국은행 경제통계국 금융통계팀 과장(e-mail: hy.lee@bok.or.kr, phone: 02-759-4411)
d. 한국은행 경제통계국 통계기획팀장(e-mail: shyi@bok.or.kr, phone: 02-759-4384)
* 본고의 내용은 집필자들의 개인의견으로 한국은행의 공식견해를 나타내는 것은 아님.
** 본 연구에 사용된 온라인 뉴스 기사 데이터는 한국정보화진흥원(NIA)으로부터 입수하였음.

Ⅰ. 서 론

미래의 경기 상황은 정책당국 뿐만 아니라 기업, 가계 등 다른 경제 주체 모두에게 매우 중요한 이슈이다. 현재와 미래의 경기 상황을 정확히 파악하고 예측하기 위해 지금까지 많은 모형들이 개발되었는데, 보통 이러한 모형들은 'Hard information'과 'Soft information'을 결합하는 형태로 제안되었다. 여기서, Hard information은 객관적이고 정량적인 변수, 예를 들면 국민소득, 물가, 실업률 등과 같은 정보를 의미하고 Soft information은 설문조사 방식을 통해 얻은 현재와 미래의 경기 상황에 대한 시장 참여자들의 인식과 전망 등에 관한 정보를 의미한다(Liberti and Peterson, 2018). Soft information과 같이 가계와 기업의 경기에 대한 판단 및 전망은 생산, 소비, 투자의 변화를 통해 단기적 경기변동을 발생시키므로 경제주체들이 느끼는 체감 경기와 거시 경제지표들 간에 밀접한 관련성을 보이는 것으로 알려져 있다(Bram and Ludvigson, 1998; Ludvigson, 2004; Gelper *et al.*, 2007; Shapiro *et al.*, 2017).

국내에서 경제 주체들을 대상으로 그들의 경기 인식과 전망에 관하여 조사하는 대표적인 예로, 한국은행에서 매달 설문조사를 통해 작성·발표하는 소비자동향지수(CSI: Consumer Survey Index), 소비자심리지수(CCSI: Composite Consumer Sentiment Index), 기업경기실사지수(BSI: Business Survey Index), 경제심리지수(ESI: Economic Sentiment Index) 등이 있다(한국은행, 2014). 먼저 소비자동향지수는 경제 상황에 대한 소비자의 인식 및 향후 전망 등을 설문조사한 결과이다. 다음으로 소비자심리지수는 우리나라 가계부문의 현재생활형편, 생활형편전망, 가계수입전망, 소비지출전망, 현재경기판단, 향후경기전망 등 총 6개의 소비자동향지수를 합성한 지수로 일반인들이 우리나라 경제 상황에 대해 전반적으로 어떻게 생각하고 있는지를 알려주는 지표이다. 이와 대조적으로 기업경기실사지수는 뉴스 기사에서 '기업 체감 경기'를 나타내는 지표로 다루어지며 기업가의 주관적이고 심리적인 요소를 조사한 지수로 객관적인 수치로 표현되는 다른 경제지표와는 다르다. 한국은행은 크게 제조업과 비제조업으로 나누어 현재와 다음 달의 업황, 매출, 생산설비수준 및 설비투자실행, 경영 애로사항 등을 설문 항목으로 놓고 조사를 한다. 마지막으로 경제심리지수는 가계와 기업의 생산, 소비, 투자, 고용 등 총체적 경제활동에 대한 심리상태로 소비자동향지수와 기업경기실사지수에 가중치를 주고 합성하여 산출한다.

경제 주체들의 경기에 대한 심리, 전망, 태도 등과 경제 활동 사이에 어떤 연관이 있는지에 대해서 여전히 많은 논쟁이 있지만, 연구자들 사이에서는 경기에 대한 심리와 미래의 경제 활동 사이에는 연관성이 존재하고, 경기 심리에 대한 연구가 예측 목적으로 유용하다

는 공통된 의견이 있다. 그러나 설문조사를 통해 발표되는 지수들은 설문조사 항목에 관한 정보 위주로 반영되므로 특정 이슈가 발생할 경우 그것이 경제 주체들의 경기인식에 미치는 영향을 알 수 없다(송민채·신경식, 2017). 또한, 조사 및 수집, 집계까지 상당한 시간이 소요되어 자료의 이용 가능 시점에 제약이 있고, 많은 비용이 든다는 문제점도 있다.

따라서 본 연구는 이러한 기존 심리지수들의 단점들을 보완하기 위해 온라인 뉴스 기사를 활용하여 새로운 경제심리지수인 News ESI (NESI)를 개발하고자 하였다. 본고에서는 NESI 측정 방법론과 개발된 NESI의 유용성과 한계를 확인하는 내용에 대하여 소개한다. 온라인에서 생성되는 뉴스 기사는 속보성이 뛰어나고 커버리지가 넓어 경제와 관련하여 어떤 이슈가 발생할 경우 이것이 경제에 미치는 영향을 빠르게 파악할 수 있다. 또한, 조사에 필요한 시간과 비용을 줄일 수 있어 온라인 뉴스 기사를 통해 얻은 텍스트 빅데이터의 유용성이 확인되면 그 활용도는 크게 높아질 것으로 기대된다.

텍스트 데이터와 같은 비정형(unstructured) 빅데이터를 활용한 다양한 연구방법 중, 본 연구에서는 속성 기반 오피니언 마이닝(ABOM: Aspect-based Opinion Mining) 방법을 응용하였다. 그 이유는 온라인 뉴스 기사를 활용하여 현행 ESI의 보조지수를 만드는 작업이 속성 탐지(Aspect Detection)와 감성 분석(Sentiment Analysis) 과정의 결합으로 이해될 수 있기 때문이다. 즉, 주어진 온라인 기사가 기술하는 주된 내용이 무엇인지를 파악하는 문제는 속성 탐지에 대응되고, 그 뉴스 기사의 톤을 평가하는 작업은 감성 분석의 일종이기 때문이다. 최근 ABOM의 경우 문장 단위(Schouten et al., 2018) 혹은 절 단위(Im et al., 2018)까지 속성을 찾고 감성점수(sentiment score)를 부여하는 방법론이 개발되었지만, 본 연구에서는 뉴스 기사 단위(document level)에 한정하여 속성을 찾고 그 속성에 대한 감성점수를 부여하였다.

온라인 뉴스 기사에서 추출된 정보로 만들어진 NESI가 실제로 활용될 수 있기 위해서는 다른 경제지표와의 연관성이 매우 중요하다. 본 연구에서 생성한 NESI의 유용성을 점검하기 위하여 현행 ESI 및 거시 경제의 대표지표인 실질 GDP와 그 연관성을 비교·분석하였다.

본고의 구성은 다음과 같다. Ⅱ장에서는 ABOM의 간략한 아이디어 및 연계 방법을 정리하였다. Ⅲ장에서는 통계모형 기반 방법과 어휘사전 기반 방법에 대하여 설명하고, 최종적으로 본 연구에 사용된 NESI 산출 공식을 간략하게 소개하였다. Ⅳ장에서는 실제 온라인 뉴스 기사(텍스트 빅데이터)에 두 방법론을 적용하여 News BSI (NBSI)와 News CSI (NCSI) 점수를 도출하고 이를 결합하여 NESI 점수를 계산하였다. 그리고 생성된 NESI를 GDP 및 ESI와 간략하게 비교·분석하였다. 마지막으로 Ⅴ장 결론에서는 연구 내용을 요약하고 향후 추가적인 연구 과제를 제안하였다.

Ⅱ. 선행 방법론 연구

1. 텍스트 데이터 처리 방법

텍스트 데이터를 처리하는 대표적인 방법으로는 Bag-of-words[1] 방법이 있다. Bag-of-Words 표현방식은 텍스트 데이터를 개별적인 단어들의 집합으로 이해하고 이를 단어의 출현여부 혹은 빈도로 수치화하여 나타내는 대표적인 텍스트 처리 방법이다. 하지만 Bag-of-Words 방식은 유니크한 단어가 많은 경우(예를 들면, 고차원 데이터의 sparsity 문제)에 잘 작동하지 않는 단점이 있다(Im *et al.*, 2018). 이러한 Bag-of-Words 방식의 한계점을 보완하기 위해서, 단어 임베딩(Word Embedding)과 토픽 모델링(Topic Modeling)이 주요한 대안으로 제안되었다.

단어 임베딩은 단어들이나 구(phrase)를 실수 공간으로 맵핑(mapping)하는 일련의 자연어 처리(NLP: Natural Language Process)의 일종인데, 대표적인 방법으로 Mikolov *et al.*(2013)가 제안한 word2Vec이 있다. 신경망(Neural Network)을 사용하여 텍스트 데이터를 변환하는 것이 주요한 내용으로 단어 하나가 주어지면 그 단어와 주변 단어가 같이 일어날 확률을 구하여 단어의 의미를 수치화한다. 가령 Bag-of-Words 방식에서는 '고양이'가 들어간 문장 혹은 구가 고양이 = [0, 0, 0, ..., 1, 0, 0]의 형태로 디지털 숫자로 표현되지만, word2Vec에서는 '귀엽다', '야옹거린다'와 같은 주변 단어가 의미에 영향을 끼쳐서 고양이 = [1.281, -2.321, ... 3.212]와 같이 실수 형태로 표현된다.[2] 한편, Pennington *et al.*(2014)은 문장이나 구 단위가 아니라 문서 단위에서 의미를 고려하여 저차원 벡터(Low-dimensional Vector)로 표현하는 방법을 제안하였다.

토픽 모델링은 문서 내용을 단순히 단어가 아니라 토픽(주제)라는 큰 의미 단위로 파악하는 기법이다. 토픽은 단어의 분포이며, 토픽마다 단어의 분포가 다르다. 토픽 모델링은 어떤 문서가 토픽들로부터 생성된 단어로 이루어지고, 이 토픽들이 문서에서 차지하는 비율은 서로 다르다고 가정한다. 예를 들어 토픽 모델링을 통해 어떤 신문기사는 스포츠 토픽과 연예인 토픽이 7:3의 비율로 섞여 있다고 해석할 수 있다. Hofmann(1999)이 제안한 probabilistic Latent Semantic Indexing(pLSI)과 Blei *et al.*(2003)이 제안한 LDA(Latent Dirichlet

1) '단어들의 가방'이라는 의미로, 텍스트에서 단어가 출현한 만큼 단어가방(설명변수 집합)에 1씩 더해 준다.
2) 보다 자세한 설명은 한국은행 발간 「국민계정리뷰」 2017년 제4호를 참조.

Allocation; 잠재 디리클레 할당) 기법이 토픽 모형에서 많이 쓰이는 대표적인 방법들이다. 문서가 어떤 토픽을 가질 확률, 각 단어가 어떤 토픽에 해당할 확률, 그 토픽에 따라 단어가 어떤 확률로 생성될지를 정의해서 문서를 확률 모델로 설명한다. 문서 안의 토픽 하나하나는 Dirichlet 분포를 따르며, 각 문서를 이 토픽들에 할당한다. 엄밀하게 군집화(clustering)와 토픽 모델링은 접근 관점이 다르지만, 유사도를 기반으로 데이터를 묶는다는 면에서 공통점이 있다.

2. ABOM

속성 기반 오피니언 마이닝(ABOM: Aspect-Based Opinion Mining)은 텍스트 데이터의 속성(aspect) 혹은 감성(sentiment)을 파악하여 주어진 텍스트를 속성에 맞게 요약하는 일련의 알고리즘 혹은 통계적 방법론을 통칭한다. ABOM은 크게 세 가지 방법으로 분류된다(Schouten and Frasincar, 2016). 첫 번째는 속성 탐지(aspect detection) 방법이고 두 번째는 감성 분석(sentiment analysis) 방법이며 세 번째는 속성 탐지와 감성 분석을 함께 수행하는 결합 방법(JADSA: Joint of Aspect Detection and Sentiment Analysis)이다. 이 중에서 세 번째 방법이 중요한데, 이는 속성 탐지와 감성 분석 과정을 결합하여 시너지를 창출할 수 있기 때문이다.

JADSA 방법은 보통 세 단계로 구분 되는데, 이는 추출, 분류 그리고 감성 분석이다(Hu and Liu, 2004; Schouten and Frasincar, 2016). 첫 번째 단계는 속성-감성 단어로 이루어진 단어집합을 주어진 텍스트 데이터에서 추출하는 것이다. 다음 단계인 분류에서는 속성 단어는 몇 가지 속성으로 할당되고, 감성 단어는 미리 결정된 감성 카테고리, 예를 들면 긍정, 부정, 중립 등으로 분류된다. 마지막 단계인 감성 분석에서는 각 속성으로 분류된 감성 단어들의 점수를 모두 합하여 그 속성의 감성 점수가 결정된다.

한편, ABOM은 구현 방법에 따라서 네 가지 방식으로도 구분할 수 있는데, 이는 통사 기반 방식(syntax-based methods), 지도 학습(supervised machine learning), 비지도 학습(unsupervised machine learning), 하이브리드 학습(hybrid machine learning)이다. 통사 기반 방식이란 감성 단어들이 먼저 식별된 다음 속성 단어들과 문법적으로 연결되는 것이다. 지도 학습은 라벨링된 데이터가 존재하는 경우인데, 이 방식은 보통 CRF(Conditional Random Field)를 선택한다(Li *et al.*, 2010; Zhong and Wang, 2010). 비지도 학습에서는 지도 학습과 달리 속성 탐지와 감성 분석을 라벨링이 되지 않은 데이터로 진행한다. 앞서 언급한 pLSI 기반과 LDA 기반 토픽 모델링이 비지도 학습의 대표적인 예이다(Hofmann, 1999; Lu *et al.*,

2011; Titov and McDonald, 2008). 마지막으로 하이브리드 학습은 두 가지 이상의 방법을 동시에 적용하는 경우를 지칭하는 것으로, Popescu and Etzioni(2005)와 Church and Hanks(1990)에 기본 아이디어가 소개되어 있다.

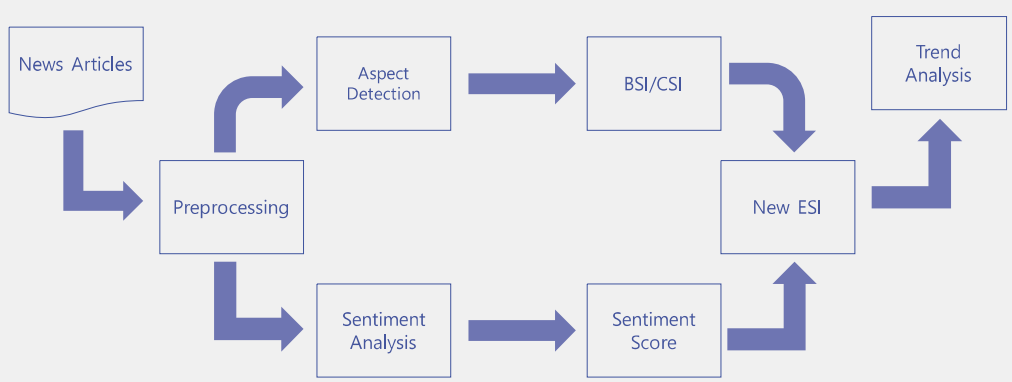
ABOM의 속성 탐지에 관한 초기 연구에서는 단어의 출연 빈도 기반으로 토픽들을 요약했다. 따라서 의미적으로 전혀 상관이 없지만 높은 빈도로 출연한 단어들이 대표 속성으로 선택되었다. 이는 속성 단어들을 전혀 다른 속성들로 분류하는 것과 같은 문제들을 발생시켰다. 이런 이유로 많은 문장 압축(sentence summarization) 알고리즘들이 개발되었고, 입력 문장 내의 토큰 (token)들의 활용 여부가 결정되었다. 문장 압축 알고리즘의 결과로 정보가 적게 담긴 단어들이 삭제된 단순한 문장이 남게 되었다. 이때, 의미론적인 유사성을 측정하는 두 가지 기준이 특별히 주목을 받게 되었는데, 이는 글뭉치 기반(corpus-based)과 지식 기반(knowledge-based)이다. 글뭉치 기반은 글뭉치에서 추출된 단어들 사이의 유사성을 측정하는 것이다. Point-wise Mutual Information, Latent Semantic Analysis(LSA), 그리고 LDA와 같은 토픽 모델링이 글뭉치 기반 방식이다. 글뭉치 기반과 대조적으로 지식 기반 방식은 의미론적 네트워크에서 얻은 추가 정보를 토대로 단어들이 어느 정도로 정량적으로 연결되어 있는지를 보는 방식이다. 이러한 방식에서는 주로 WordNet과 같은 어휘 관련 데이터베이스가 활용된다. 이와 유사한 어휘사전 기반 접근 방법(Lexicon-based Approach)은 단어 사전을 활용한다. 미리 작성된 단어 사전과 부여된 점수를 기준으로 주어진 텍스트를 분류하고 점수를 계산하는 방식이다.

Ⅲ. 제안 방법론

본 연구는 통계모형 기반 방법(Statistics Model-based Method)과 어휘사전 기반 방법(Lexicon-based Method)의 두 가지 다른 접근방식으로 진행하였다. 텍스트에 전통적으로 많이 사용되는 두 가지 방법론을 각각 독립적으로 진행한 다음 그 결과를 비교하고 발전된 방향으로의 새로운 방법론을 모색하는 것을 목표로 하였다.

실제 BSI/CSI 분류와 스코어링(scoring) 작업은 다르게 구현이 되었지만, 일련의 처리 과정은 <그림 3-1>과 같이 동일하다. 먼저 전처리(pre-processing) 과정을 통해 생성된 데이터는 속성 탐지 및 감성 분석의 두 방향으로 분석이 동시에 진행된다. 그리고 속성 탐지를 통하여 해당 기사가 BSI와 CSI 가운데 어떠한 것과 주로 관련이 있는지를 파악하고, 감성 분석 단계에서 해당 기사의 전체적인 톤, 즉 긍정/부정 여부를 분류한다. 마지막으로 속성 탐지와 감성 분석을 결합하여 NBSI와 NCSI를 각각 산출하고, 이를 통하여 최종적으로 NESI 값을 생성하게 되는 것이다. 추가적으로 이렇게 생성된 NESI를 이용하여 실질 GDP와 같은 거시 경제활동 지표와의 연관성을 추이 분석(trend analysis)을 통해 살펴보았다.

〈그림 3-1〉 **온라인 뉴스 기사 경제 지수화 작업 순서도**



1. 통계모형 기반 방법

가. 텍스트 데이터 전처리

한글 자연어 처리(NLP)에서 가장 선행되어야 하는 작업은 입력 문자열을 형태소 열로 바꾸는 형태소 분석 내지는 품사 태그(Part-of-speech tags) 작업이다. 형태소 분석이란 띄어쓰기 단위의 어절인 단어(word)를 더 이상 분석 불가능한 최소 의미 단위인 형태소들로 분리하고, 용언의 활용, 불규칙 활용이나 축약, 탈락 현상이 일어난 형태소는 원형을 복원하는 과정을 의미한다(강승식, 2002). 하지만 한글 텍스트 데이터를 각각의 단어 혹은 형태소로 분리하는 토큰화[3](tokenization) 작업이나 품사를 태깅하는 작업은 다른 언어(예, 영어)에 비하여 쉬운 작업이 아니다. 아카데믹에서 많이 연구된 방법론이나 알고리즘들이 제한적으로 사용될 수밖에 없기 때문이다. 이러한 한계는 한글 고유의 특징, 경제 관련 뉴스 기사의 특징, 그리고 제반 관련 연구 부족 등에서 기인하는데 몇 가지 주요한 한계점을 <부록 1>에 정리하였다.

본 연구에서는 경제기사 데이터에서 우선 특수문자, 알파벳, 구두점 등 불용어[4](stopwords)를 삭제하고 토큰화를 수행한 뒤, 주어진 텍스트 데이터를 어간으로 변형시켰다. 그리고 분석 대상으로 명사, 동사, 형용사, 부사 등 주요 품사만을 추출하여 사용하였다.

나. 데이터로 변환 및 차원 축소

토큰화가 완료된 텍스트 데이터는 Ⅱ장에서 소개된 방법들로 수치형 데이터로 전환될 수 있다. word2Vec과 LDA를 적용하여 통계적으로 차원 축소를 수행하기 위해서는 우선 텍스트 데이터를 수치형 행렬로 변환해야 한다. 대표적인 방법을 적용하면, 경제기사 글뭉치(corpus)를 빈도 기반 Bag-of -Words 방법으로 표현할 수 있다. 행은 각 경제기사를 순서대로 나열한 것이며, 열은 전체 글뭉치에 존재하는 모든 단어들이 변수의 형태로 구성되어 있다. <표 3-1>을 보면 이렇게 얻어진 데이터들이 예시로 표기되어 있는데, 값이 0이라는 것은 해당 단어(열)가 해당 뉴스 기사(행)에 포함되어 있지 않다는 것을 의미한다. 성긴 정

3) 토큰들로 문자열을 분리하는 작업으로, 형태소 분석 시 품사 태그를 달지 않고 형태소만 남겨놓는 것을 의미한다.

4) 불용어란 정보를 전달하지 않고 기능적 역할을 하는 단어들이다. 영어의 the, a, in, about…, 한국어의 있, 않, 없… 등을 말한다. 텍스트 분석 시 더 나은 결과를 내기 위해 불용어를 삭제한다.

도(sparsity)를 확인해 보면 99% 이상의 원소 값이 0을 차지한다.

〈표 3-1〉 문서-단어(document-term) 행렬의 예시

	가격	가계	가구	가장	가치	...	흑자	흔들	힘입
0	0	0	0	0	0	...	0	0	0
1	0	0	0	0	1	...	0	0	0
2	0	0	0	0	0	...	0	0	0
3	0	0	0	0	0	...	0	0	0

이러한 고차원(high-dimensional)의 성긴(sparse) 데이터에 word2Vec 방법을 적용하면 각 단어가 주변 단어와의 유사도로 의미가 부여되면서 새로운 벡터공간에 표상될 수 있다. 아래의 <표 3-2>는 100차원의 단어 임베딩으로 표현된 각 단어 벡터들의 예시를 보여준다. 따라서 각 경제기사는 아래 단어 벡터들의 선형결합으로 표현된다.

〈표 3-2〉 word2Vec으로 표현한 단어 벡터 예시 (100차원)

	0	1	2	3	...
성장률	0.017934	−0.044546	−0.077594	0.131177	...
목표	0.045437	−0.028163	−0.077815	0.117374	...
가치	0.099601	−0.069812	−0.091431	0.107722	...
급락	0.094426	−0.036966	−0.187226	0.037322	...

LDA와 같은 토픽모델링이나 LSA(Latent Semantic Analysis)와 같은 행렬 축소 방법을 적용하면, 글뭉치 데이터는 위와 같은 분산 표현 형태의 행렬로 차원이 축소될 수 있다. 각 경제기사는 토픽의 개수만큼 차원을 갖고 토픽의 비중만큼 원소 값을 갖는 벡터로 표현된다. <표 3-3>은 주어진 경제 기사를 20개의 토픽 차원에 적합한 결과를 나타낸다. 각 토픽에서 높은 확률로 매칭된 세 개의 단어들이 표기되어 있다.

〈표 3-3〉 토픽 20개에서 매칭 확률이 높은 상위 3개 단어들

Topic 1	마감, 정부, 확대	Topic 11	사상, 산업, 식품
Topic 2	외국인, 약세, 코스피	Topic 12	뉴욕, 금융시장, 건설
Topic 3	벤처, 돌파, 중소	Topic 13	소비, 지수, 제조업
Topic 4	금리, 인상, 세계	Topic 14	고용, 가계부채, 中企
Topic 5	부진, 금융, 지원	Topic 15	분기, 성장, 서울
Topic 6	한국, 판매, 전망	Topic 16	강세, 불황, 앞두
Topic 7	경기, 회복, 연속	Topic 17	지표, 펀드, 악재
Topic 8	대출, 주택, 가격	Topic 18	기업, 은행, 매출
Topic 9	위기, 유럽, 미국	Topic 19	가전, 중국, 올해
Topic 10	글로벌, 추진, 종합	Topic 20	시장, 유가, 영향

다. 분류 예측

전처리와 차원 축소 과정이 끝나면, 처리된 데이터에 BSI/CSI 분류와 긍정/부정 분류 모형을 추정할 수 있게 된다. 본 연구에서는 대표적인 통계 분류 모형으로 단순 베이즈(Naive Bayes), 로지스틱 회귀(Logistic Regression) 및 신경망(Neural Network) 기법을 사용하였다.

첫째, 단순 베이즈 기법은 변수들 간 조건부 독립 가정을 사용하여 모형을 적합시킨다. 가령, 경제기사가 부정적 심리를 담고 있고 기사에 '경기 불황'이라는 단어가 포함되어 있다는 사실과, 기사 내에 곧 '주가 하락'이라는 단어가 등장할 것이라고 생각하는 기대가 전혀 관계가 없어야 한다는 뜻이다. 단순 베이즈는 빠르게 계산할 수 있고 구현이 간단하여 폭넓게 활용된다. 특히 데이터가 성긴 경우에도 다른 판별보다 더 좋은 성능을 보인다고 알려져 있다.

둘째, 로지스틱 회귀 기법은 반응변수가 연속형이 아니라 긍정 또는 부정의 이진(binary) 값을 가질 때 효과적으로 분류 예측을 수행할 수 있는 모형이다. 연속형 반응변수는 통상적인 선형 회귀에서 정규 분포를 따른다고 가정하지만, 값이 두 개만 존재하는 반응변수는 이 가정이 유효하지 않다. 따라서 0과 1 사이 산출 값을 갖는 로지스틱 연결 함수(logistic link function)를 사용해 설명변수의 선형 조합을 반응변수에 연계시킨다. 그 결과 반응변수는 값이 긍정일 확률을 나타내며, 임계치와 같거나 크면 긍정으로 예측하고 그렇지 않으면 부정으로 예측한다. 로지스틱 회귀는 일반화된 선형모형의 한 종류이다.

셋째, 본 연구에서 사용된 신경망(Neural Network) 기법은 최근의 복잡한 딥러닝(deep

learning) 알고리즘들이 발전하게 된 시초인 다층 퍼셉트론(Multi-layer Perceptron)이다. 신경망은 입력층(input layer), 은닉층(hidden layer) 및 출력층(output layer)으로 구성되어 있으며, 은닉층이 많아질수록 깊은 신경망이 된다. 신경망에서는 입력층과 은닉층 내, 은닉층과 출력층 내 유닛(unit)들이 서로 가중치 합과 비선형 함수(활성화 함수)로 연결되어 있다. 따라서 설명변수 값에 해당하는 데이터가 입력층에 유입되면 출력층에 나타나는 예측변수 값이 실제 값과 가까워지도록 가중치(계수)를 학습하게 된다. 신경망은 은닉층의 개수, 각 은닉층의 유닛수, 활성화 함수의 종류(ReLU, 하이퍼볼릭 탄젠트 등) 등 매우 다양하게 모수 값을 통제할 수 있지만, 본 연구에서는 가장 단순한 버전의 신경망을 사용하였다.

2. 어휘사전 기반 방법

어휘사전 기반 방법(Lexicon-based Method)은 분류 사전(classification lexicon)을 이용하여 주어진 텍스트 데이터를 분류하는 방법이다. 본 연구 과제는 결국 주어진 온라인 기사를 BSI/CSI에 맞게 분류하고 난 뒤에 각 기사의 톤(긍정/부정)을 분류하는 것으로 치환될 수 있기 때문에, 분류 사전 기법이 적용될 수 있다. 이러한 어휘사전 기법의 예로 감성 분석을 생각해 볼 수 있다. 영어 감성 분석의 경우에는 일반사전인 'WordNet'에 감성 정보를 추가한 'SentiWordNet'과 같이 공개된 감성 사전이 많이 사용된다. 어휘사전 기법의 성과는 주어진 분류 어휘사전의 품질에 매우 의존적이다. 그러나 한글은 공개적으로 많이 사용되는 어휘사전이 없어 본 연구에서는 극성이 미리 정의된 온라인 뉴스 기사를 활용하여 자체적으로 BSI/CSI 분류 사전과 긍정/부정 분류 감성 사전을 자체적으로 구축하였다. 이에 대한 자세한 알고리즘은 <부록 2>에 소개되어 있다.

NESI를 산출하기 위해서는 각 기사가 BSI/CSI 중 어느 것에 해당하는지 분류해야 하고(속성 탐지), 또한 그 기사의 톤이 긍정적인지, 아니면 부정적인지를 예측해야 한다(감성 분석). 이러한 과정은 앞서 구축한 BSI/CSI 어휘사전 및 긍정/부정 어휘사전을 이용하여 진행할 수 있다.

(Step 1) 주어진 뉴스 기사에 속해 있는 단어들을 토큰화한 뒤 벡터(V)로 표현한다. 이 때, 불용어는 삭제한다.

(Step 2) BSI/CSI 분류

단어 벡터 V와 BSI/CSI 어휘사전을 비교하여 매칭이 되는 단어의 수를 결정한다. 이 때,

V_{BSI}와 V_{CSI}를 어휘사전에 매칭된 단어들의 집합으로 정의하고, $N_{V,BSI}$와 $N_{V,CSI}$를 이 집합들의 크기로 정의하자. 각 집합의 크기를 BSI 어휘사전 및 CSI 어휘사전의 크기로 정규화시키면 상대 빈도 수가 얻어지는데, 이를 $N_{\tilde{V},BSI}$와 $N_{\tilde{V},CSI}$로 표기한다.

$$N_{\tilde{V},BSI} = \frac{N_{V,BSI}}{N_{BSI}}, \quad N_{\tilde{V},CSI} = \frac{N_{V,CSI}}{N_{CSI}}.$$

여기에서 N_{BSI}와 N_{CSI}는 각각의 어휘목록의 크기이다. 최종 BSI/CSI 분류점수는 다음과 같이 계산한다.

$$S_{BSI/CSI} = \log(N_{\tilde{V},BSI}) - \log(N_{\tilde{V},CSI}) .$$

만약 분류점수 $S_{BSI/CSI}$가 0보다 크면 해당 기사는 BSI에 할당하고 0보다 작으면 CSI에 할당한다[5].

(Step 3) 긍정/부정 분류

BSI/CSI 분류에 사용된 알고리즘을 그대로 사용하여 긍정/부정 기사를 분류할 수 있다. BSI/CSI 분류에 사용된 방법을 그대로 적용하면 다음과 같은 분류점수를 얻을 수 있다.

$$S_{P/N} = \log(N_{\tilde{V},P}) - \log(N_{\tilde{V},N}).$$

여기에서 $N_{\tilde{V},P}$와 $N_{\tilde{V},N}$는 긍정 어휘사전 및 부정 어휘사전 목록에 매칭된 단어들의 상대적 크기를 의미한다. 만약 분류점수 $S_{P/N}$이 0보다 크면 긍정으로 분류하고 0보다 작으면 부정으로 분류한다.

5) 분류점수 0 부근 위아래로 오차를 두어 중립영역을 만들 수 있으나, 본 연구에서는 편의를 위하여 이러한 내용을 고려하지 않았다.

3. 새로운 지수 계산

통계모형 기반 방법 혹은 어휘사전 기반 방법을 적용하고 나면 각 기사에 대하여 BSI/CSI 분류 결과와 긍정/부정 분류 결과를 얻게 된다. 이러한 결과들을 이용하여, 목표 기간에 해당하는 NBSI 및 NCSI를 다음과 같이 계산할 수 있다.

$$NBSI/NCSI = \frac{\text{긍정으로 분류된 기사의수} - \text{부정으로 분류된 기사의수}}{\text{기간내 긍정} + \text{부정 기사 수}} \times 100 + 100$$

즉, BSI 분류 기사 중에서 긍정으로 분류된 기사와 부정으로 분류된 기사의 수를 이용하여 NBSI에 해당하는 점수를 계산할 수 있고, 마찬가지로 NCSI에 해당하는 점수를 계산할 수 있다. NESI는 이렇게 얻어진 NBSI, NCSI의 표준화지수를 아래와 같이 가중평균한 후 지수의 평균이 100, 표준편차는 10이 되도록 변환하였다.

$$NESI = (0.75 \times NBSI) + (0.25 \times NCSI)$$

새로운 NESI 공식에 사용되는 가중치는 조사로 얻어지는 ESI 공식에서 차용하였다.

Ⅳ. 실제 데이터 분석

1. 포털 뉴스 데이터

통계모형 기반 접근 방법과 어휘사전 기반 접근 방법의 성능을 확인하기 위하여 10년치(2008~2017) 온라인 뉴스 기사를 사용하였다. 이 데이터는 한국정보화진흥원이 수집하여 한국은행에 제공한 포털 뉴스 '경제' 카테고리 기사들의 일부이다. 이 데이터는 서브 카테고리, 최초 게재일, 최종 수정일, 언론사, URL, 제목, 내용 등 크롤링(crawling) 기법을 통해 얻을 수 있는 기본 정보를 포함한다. 2017년 데이터에는 기본 정보 외에 경제 심리 유무, 국내/해외 기사 여부, BSI/CSI 해당 여부, 긍정/부정 5점 척도 등이 포함되어 있다[6]. <표 4-1>에 2017년 데이터의 일부가 예시로 표기되었다.

〈표 4-1〉 **2017년 포털 뉴스 기사 샘플 일부**

날짜	기사	BSI	CSI
2017-01-01	새해 경제 3대 복병…정치 포퓰리즘 · 보호무역주의 · 미국 금리인상	2	2
2017-01-02	[RUN to YOU] 반갑다 트럼프 기대 더 커진 엔저 효과	1	4
2017-01-08	한은 글로벌 경제 불확실성 지속…정책적 대응 시급	3	2
2017-01-09	은퇴족부터 고등학생까지 … '부동산 매매업' 열풍	3	2
2017-02-01	나바로 美무역위원장 "獨 저평가된 유로화로 미 · EU 착취"	1	1

2017년도 포털 뉴스 샘플 기사 2,947개 중에서 가장 먼저 BSI/CSI 점수가 빈칸인, 즉 경제 심리가 없는 1,040개는 모형 적합에서 우선적으로 제외하였다. 또한 BSI/CSI 점수가 동시에 존재하는 619개와 감성 점수가 중립인 기사 106개 역시 분류의 정확성을 위하여 제외하고 최종적으로 1,182개 기사를 모형 적합에 사용하였다. 1,182개 기사를 70:30 비율로 트레이닝(training)/테스트(test) 데이터로 나눈 뒤, 트레이닝 데이터를 이용하여 모형을 적합시키고 테스트 데이터를 이용하여 모형 적합 결과를 평가하였다. 트레이닝 데이터와 테스트 데이터를 분리할 때는 무작위로 분리하는 방법과 시간 순에 따라서 분리하는 두 가지 방

6) 한국은행 경제통계국에서 별도로 작업하여 제공하였다.

법을 모두 고려하였다. 각 데이터는 통계모형 기반 방법론과 어휘사전 기반 방법론에 공통적으로 사용되었다. 트레이닝 데이터에 관한 기본 정보[7]는 <표 4-2>에 정리되어 있다.

〈표 4-2〉 **트레이닝 데이터 기본 정보**

(개수)

BSI/CSI	BSI	541
	CSI	286
감성점수	긍정	248
	부정	579

2017년 데이터를 활용하여 모형 적합된 방법론들은 2013~2016년 뉴스 기사 데이터에 적용이 되어 NBSI, NCSI 및 NESI를 계산하는데 사용되었다. 이 때,전체 대상 기사 중에서 경제 심리가 있는 1,861개만을 선택[8]하여 각 지수를 산출하는 데 사용하였다.

2. 방법론 분류 성능 평가

본고에서는 경제기사가 긍정적 심리인지 부정적 심리인지, 그리고 기업심리(BSI)에 영향을 미치는지 소비자심리(CSI)에 영향을 미치는지 등을 예측할 때 이진(binary) 분류 모형에서 사용되는 분할표(contingency table)를 바탕으로 성능을 평가하였다.

〈표 4-3〉 **분할표**

긍정심리	예측: 긍정	예측: 부정
실제: 긍정	True Positive (TP)	False Negative (FN)
실제: 부정	False Positive (FP)	True Negative (TN)

7) 실제 데이터에서 감성에 대한 극성은 5점 척도였지만 분석과 지수화의 편의상 긍정/부정으로 재분류하여 사용하였다.

8) 경제심리 여부에 대한 분류 방법은 한국은행 발간 2017년 국민계정리뷰 제4호에 소개되어 있으며, 이를 적용한 데이터를 이용하여 지수를 산출하는 것을 암묵적으로 가정하였다.

각 방법의 성능 평가 척도로는 정밀도(Precision), 재현율(Recall) 및 F1 점수 (F1 score)를 사용하였다. 정밀도는 모형이 긍정적 심리로 예측한 경제기사 중에 실제로 정확하게 예측된 경제기사의 비율이다. 재현율은 실제 긍정적 심리의 경제기사 중에 모형이 제대로 예측한 경제기사의 비율이다. 그리고 F1 점수는 정밀도와 재현율의 조화 평균이다. 정밀도와 재현율이 불균형일수록 그 값이 낮아지며, 정밀도와 재현율이 모두 완벽하면 1의 값을 가진다.

- 정밀도 = TP / (TP + FP)
- 재현율 = TP / (TP + FN)
- F1 점수 = 2×정밀도×재현율 / (정밀도 + 재현율)

가. 통계모형 기반 방법 성능 평가

한국은행에서 제공한 2017년 온라인 샘플 뉴스기사 데이터에 대하여 기계학습(machine learning)을 수행하였다. 파이썬(Python) KoNLPy 패키지(Park and Cho, 2014)에서 제공하는 트위터 형태소 분석기를 바탕으로 전처리를 완료하였다. 텍스트 전처리한 데이터를 TF-IDF 방식의 Bag-of-Words 행렬로 수치화한 후, 분류 기계학습을 수행하였다. 앙상블 기법(랜덤 포레스트, 그라디언트 부스팅), 서포트벡터머신(Support Vector Machine) 등 여러 방법론을 수행하였으나, 계산 효율 대비 비교적 예측 성능이 좋은 3가지 방법(로지스틱 모형, 단순 베이즈 모형 및 신경망 모형)에 대해서만 평가 결과를 표기하였다. 첫째, 로지스틱(Logistic) 모형에서는 L2 패널티를 부과하여 과적합(overfitting)을 줄였고, 둘째, 범주형 설명변수에 따라 다항(multinomial) 단순 베이즈(NB)를 수행하였고, 셋째, 신경망(NN, 다층 퍼셉트론)은 10개의 노드를 지닌 은닉층 1개를 포함하고 있다. 예측하고자 하는 반응변수 값은 BSI/CSI, 또는 긍정/부정으로 나누어 수행하였다.

〈표 4-4〉 BSI/CSI 예측 - 무작위로 분리된 데이터 사용

모형	분류	정밀도	재현율	F1	빈도
Logistic	BSI	0.91	0.9	0.91	236
	CSI	0.81	0.82	0.82	119
	AVG/Total	0.88	0.88	0.88	355
NB[1)]	BSI	0.92	0.88	0.9	236
	CSI	0.78	0.86	0.82	119
	AVG/Total	0.88	0.87	0.87	355
NN[2)]	BSI	0.92	0.86	0.89	236
	CSI	0.76	0.86	0.81	119
	AVG/Total	0.87	0.86	0.86	355

주: 1) Naive Bayes (이하 같음)
2) Neutral Network (이하 같음)

〈표 4-5〉 BSI/CSI 예측 - 시간 순으로 분리된 데이터 사용

모형	분류	정밀도	재현율	F1	빈도
Logistic	BSI	0.89	0.93	0.91	234
	CSI	0.85	0.77	0.81	121
	AVG/Total	0.87	0.87	0.87	355
NB	BSI	0.96	0.86	0.91	234
	CSI	0.77	0.93	0.84	121
	AVG/Total	0.89	0.88	0.88	355
NN	BSI	0.87	0.93	0.9	234
	CSI	0.84	0.74	0.78	121
	AVG/Total	0.86	0.86	0.86	355

BSI와 CSI를 예측하는 문제에서 로지스틱 분류 모형의 결과가 정밀도, 재현율, F1 점수 등에서 단순 베이즈와 다층신경망보다 모두 높았다. 로지스틱 모형이 BSI로 예측한 경제기사 중에 정확하게 예측한 경제기사의 비율(정밀도)과 실제 BSI인 경제기사 중에 모형이 제대로 예측한 경제기사의 비율 모두 90%에 가깝다. 또한 BSI의 경우 재현율이 높고, CSI의 경우 정밀도가 높은 경향이 있다. 이는 기업 심리를 담고 있는 경제기사가 소비자심리를

내포한 기사보다 약 두 배 많은데, 이에 따라 수량 기준에서 얼마나 제대로 맞추었는지 측정하는 민감도는 BSI가 더 높고, 품질 기준에서 얼마나 정밀한지를 측정하는 정밀도는 CSI가 높은 것으로 평가된다. 한편 임의로 기사의 날짜 순서를 섞어 트레이닝 데이터와 테스트 데이터를 분리 후 예측한 결과와, 2017년의 1~9월 기사로 2017년 9~12월 기사의 라벨을 예측하는 결과는 둘 다 예측력이 유사하였다. 즉 시간 변화에 따른 예측력 변화는 크지 않은 것으로 판단된다.

긍정 및 부정적 심리를 예측하는 경우는 BSI 및 CSI 예측보다 전반적으로 F1 점수가 낮았다. BSI는 '기업', '수출' 등, CSI는 '가계', '저축' 등 단어 하나가 라벨값을 명확하게 결정짓는 특성을 갖지만, 긍정 및 부정적 심리는 맥락에 따라 같은 단어도 심리가 다를 수 있는 특징 때문에 예측 성능이 좀더 낮은 것으로 판단된다. 가령 '상승'이라는 단어는 '주가 상승'에서 긍정적, '유가 상승'에서 부정적으로 분류된다. 또한, 무작위로 추출된 트레이닝/테스트 데이터와 시간 순서대로 추출된 데이터의 분류 성능이 다르게 나왔다. 시간 순으로 구분하였을 때의 성능이 조금 더 안 좋게 나오는데, 이는 긍정과 부정의 의미가 시간에 따라서 다르게 해석 될 수 있음을 의미한다. 분석 결과는 <표 4-6>와 <표 4-7>에 소개되어 있다.

〈표 4-6〉 **긍정/부정 예측 – 무작위로 분리된 데이터 사용**

모형	분류	정밀도	재현율	F1	빈도
Logistic	Negative	0.84	0.87	0.85	254
	Positive	0.63	0.58	0.61	101
	AVG/Total	0.78	0.79	0.78	355
NB	Negative	0.85	0.76	0.8	254
	Positive	0.53	0.66	0.59	101
	AVG/Total	0.76	0.74	0.74	355
NN	Negative	0.82	0.89	0.86	254
	Positive	0.65	0.52	0.58	101
	AVG/Total	0.78	0.79	0.78	355

〈표 4-7〉 긍정/부정 예측 - 시간 순으로 분리된 데이터 사용

모형	분류	정밀도	재현율	F1	빈도
Logistic	Negative	0.79	0.81	0.8	237
	Positive	0.6	0.57	0.58	118
	AVG/Total	0.73	0.73	0.73	355
NB	Negative	0.81	0.8	0.8	237
	Positive	0.6	0.62	0.61	118
	AVG/Total	0.74	0.74	0.74	355
NN	Negative	0.77	0.76	0.77	237
	Positive	0.53	0.53	0.53	118
	AVG/Total	0.69	0.69	0.69	355

나. 어휘사전 기반 방법 성능 평가

1) BSI/CSI 분류 결과

Ⅲ장에서 소개된 BSI/CSI 어휘사전 생성 알고리즘을 2017년 데이터에 적용하여 BSI 어휘사전 목록과 CSI 어휘사전 목록을 우선적으로 생성하였다. BSI 어휘사전에는 4,620개 단어가 CSI 어휘사전에는 2,372개가 단어가 최종적으로 포함되었다. 이는 BSI 관련 기사가 CSI 관련 기사보다 약 2배 정도 많은데, 이러한 데이터 특성이 반영된 것으로 해석할 수 있다. <부록 3>에 어휘사전 목록의 일부가 소개되어 있다.

생성된 어휘사전에 형태소로 분리된 각 기사들을 매칭하면 기사마다 BSI 어휘사전과 CSI 어휘사전에 겹치는 단어들의 수를 계산할 수 있다. <표 4-8>은 이러한 매칭 과정을 단순화하여 예시하고 있다. 이렇게 계산된 단어들 수의 차이를 어휘사전 목록 크기에 대비하여 상대화한 다음 로그 차이(log difference)로 분류 점수를 만들고 이에 따라 각 기사를 BSI 혹은 CSI에 할당하게 된다.

〈표 4-8〉 BSI/CSI 매칭 예시

기사	형태소로 표현된 기사 내용	BSI	CSI
1	한겨레 / 회계 / 연도 / 국가채무 / ...	0	2
2	디자이너 / 부동산 / 악재 /겹치 ...	1	0
3	통계청 / 년 / 연간 / 분기 / 가계 동향 ...	0	2
4	투입 / 후 / 채권 / 단 / 정부 / 추가 ...	1	0

아래의 <표 4-9>와 <표 4-10>은 무작위 및 시간 순으로 트레이닝/테스트로 분리된 데이터에 방법론을 적용하여 얻은 분류 결과를 보여준다. 시간 순으로 나누어서 모형을 적합할 경우 전체적인 성능이 약간 떨어지는 것을 확인할 수 있다. 뉴스 기사에 주로 쓰이게 되는 단어나 표현 등이 시간에 따라 달라짐을 고려할 때 예측 가능한 결과이다. 따라서 어휘사전 기반 방법론을 실제로 적용할 때는 이점을 고려하여 주기적으로 어휘사전을 업데이트할 필요가 있다. 또한, CSI의 재현율이 상대적으로 떨어지는 것을 확인할 수 있는데 이 또한 어휘사전의 크기와 연관이 되어 있다. 어휘사전의 양적·질적 수준이 모형의 성능에 중요한 영향을 미친다는 것을 확인할 수 있다.

〈표 4-9〉 BSI/CSI 예측 – 무작위로 나눈 경우

분류	정밀도	재현율	F1	빈도
BSI	0.86	0.83	0.84	244
CSI	0.66	0.70	0.68	111
AVG/Total	0.80	0.79	0.79	355

〈표 4-10〉 BSI/CSI 예측 – 시간 순으로 나눈 경우

분류	정밀도	재현율	F1	빈도
BSI	0.80	0.93	0.86	234
CSI	0.85	0.44	0.58	121
AVG/Total	0.82	0.76	0.77	355

2) 긍정/부정 분류 결과

BSI/CSI 분류처럼 우선적으로 긍정/부정 어휘사전을 2017년 데이터를 활용하여 구축하였다. 최종적으로 긍정어가 1,258개 포함된 긍정 어휘사전을 생성하였고 부정어가 5,894개 포함된 부정 어휘사전을 생성하였다.[9] 상당수의 뉴스 기사가 부정적인 내용을 담고 있는 현상이 반영되었다. <부록 3>에 어휘사전 목록의 일부가 소개되어 있다.

어휘사전을 이용하여 각 기사에 포함된 긍정/부정 단어의 수를 계산할 수 있고 이의 상대적 차이를 계산하여 긍정과 부정 점수를 계산하였다(Ⅲ장 참고). <표 4-11>와 <표 4-12>는 이러한 방법으로 계산된 긍정/부정 예측 성능을 나타내고 있다.

무작위로 나눈 경우 긍정의 정밀도, 민감도, F1 점수가 부정에 비해 낮은 결과를 보인다. 이는 실제 부정 기사의 개수가 실제 긍정 기사에 비해 약 2.6 배가 많고 이로 인해 부정 사전의 단어 개수가 긍정 사전의 단어 개수보다 약 4.6배 정도 많기 때문에 발생한 현상이다. 이러한 현상은 어휘사전의 상대적 크기를 보정하면 해결할 수 있는데 해당 내용은 본 연구에서는 심도 있게 고려하지 않았다. 시간 순으로 나눈 경우에도 역시 정밀도, 민감도 및 F1 점수가 긍정보다는 부정 예측에서 더 낮았는데 같은 이유로 해석 가능하다. 또한, BSI/CSI 분류와 마찬가지로 시간 순으로 나눠서 평가했을 경우의 성능이 상대적으로 더 떨어지는 것을 확인할 수 있다. 어휘사전 목록은 단어들의 사용 빈도에 따라서 영향을 받기 때문에 시간 변화에 따른 단어 사용 변화에 민감한 편임을 재확인할 수 있다.

〈표 4-11〉 **긍정/부정 예측 – 무작위로 나눈 경우**

분류	정밀도	재현율	F1	빈도
Positive	0.65	0.57	0.60	109
Negative	0.82	0.86	0.84	246
AVG/Total	0.77	0.77	0.77	355

9) 본 연구에서는 각 사전의 단어 수에서 발생하는 차이를 컨트롤 하지 않았지만 추후 분석 및 연구에서는 이 부분을 보정하여 진행하는 것이 Imbalanced 데이터에서 발생하는 편차를 줄일 수 있다.

〈표 4-12〉 긍정/부정 예측 - 시간 순으로 나눈 경우

분류	정밀도	재현율	F1	빈도
Positive	0.60	0.45	0.51	118
Negative	0.76	0.85	0.80	237
AVG/Total	0.71	0.72	0.71	355

3. NBSI, NCSI, NESI 계산 및 거시 경제지표 비교

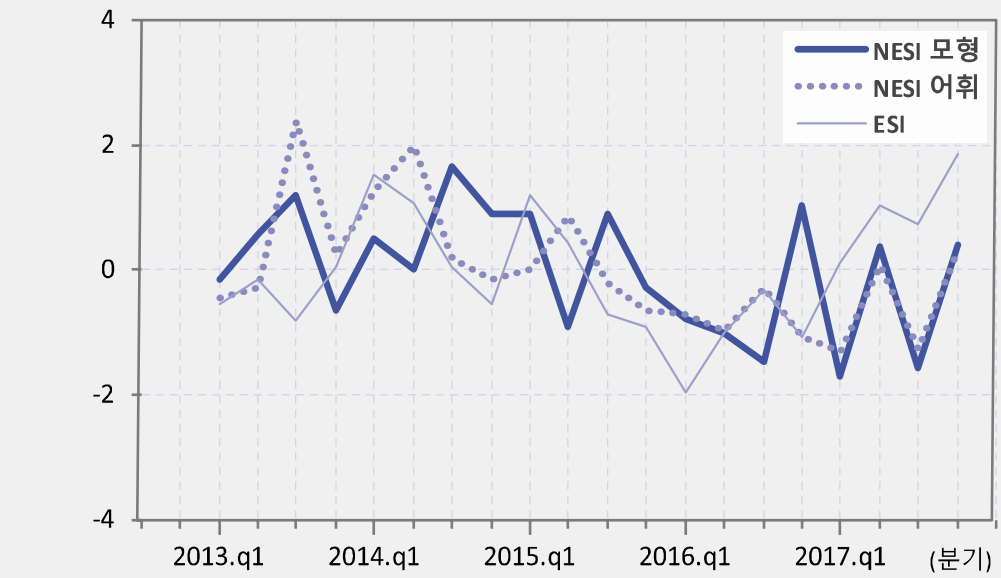
2013년부터 2017년까지 5개년도[10]의 경제 심리 뉴스 기사의 샘플에 통계모형 방법과 어휘사전 방법을 적용하여 각각의 기사를 BSI/CSI 및 긍정/부정으로 분류하였다. 이렇게 분류된 기사를 분기 단위로 묶은 뒤, Ⅲ장에서 소개한 지수공식을 사용하여 해당 분기의 NBSI, NCSI 및 NESI를 각각 계산하였다. 분기별 지수 값은 <부록 4>에 정리되어 있다. 각 지수의 변동성은 다소 큰 편인데, 이는 해당 분기 내 샘플 기사의 숫자가 크지 않기 때문이다. 더 많은 기사를 활용한다면 현재의 점수보다는 조금 더 변동성이 완화될 것으로 여겨진다.

<그림 4-1>은 한국은행에서 현재 공표하고 있는 ESI의 움직임과 본 연구에서 새롭게 개발한 방법론을 이용하여 생성한 NESI의 움직임을 보여주고 있다. 각 지수의 공정한 비교를 위해 5년간 지수 값의 평균과 표준편차를 이용하여 정규화[11](Normalization)하였다. 전반적으로 공표 ESI와 두 NESI는 추세적으로 동일하게 움직이는 것을 볼 수 있다. 특히 2015년 2분기 이후 세 지수의 움직임이 이전보다 더 유사한 것을 확인할 수 있다. 지수 작성의 고유한 이점상 일별, 주별 및 월별로 손쉽게 계산될 수 있다는 점과 현행 ESI와 유사한 움직임을 보인다는 점을 고려하면 NESI가 현행 공표 ESI의 보조지수 역할을 할 수 있을 것으로 기대된다. 나아가서 측정 오차 모형 등을 통하여 현행 ESI를 업데이트 하는 형식으로도 사용될 수도 있을 것이다.

10) 한국은행에서 실제로 제공한 데이터는 10년 치이지만, 실제 분석은 5년 치만 사용하였다. 앞서 살펴본 것처럼 모형의 성능은 시간과 밀접한 관련이 있기 때문에, 상대적으로 시간 갭이 큰 2008~2012년 기사는 분석에서 제외하였다. 이 문제는 보다 많은 데이터를 가지고 트레이닝을 시키면 자연스럽게 완화될 수 있다.

11) 일반적인 정규화 공식을 사용하였다. 즉, (관측값-평균)/표준편차.

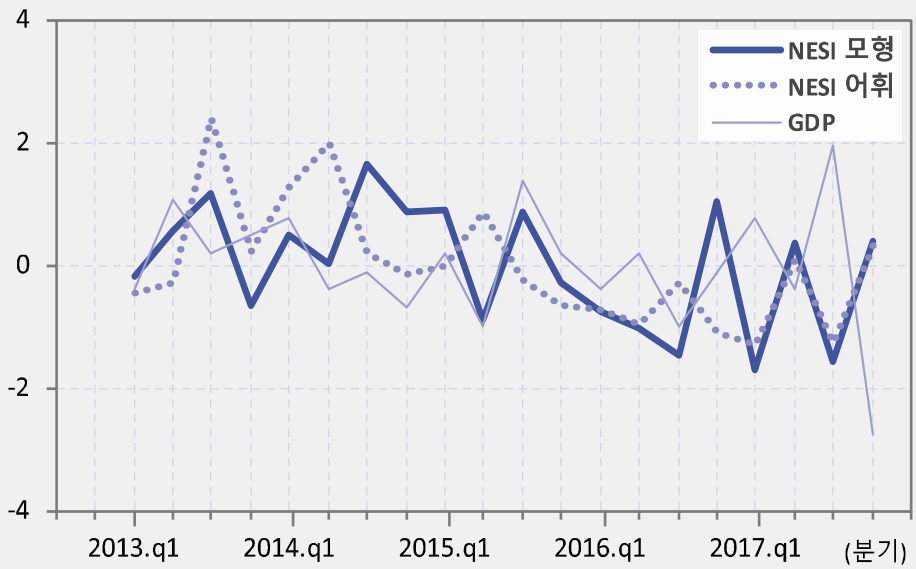
〈그림 4-1〉 현행 ESI와 NESI 추이



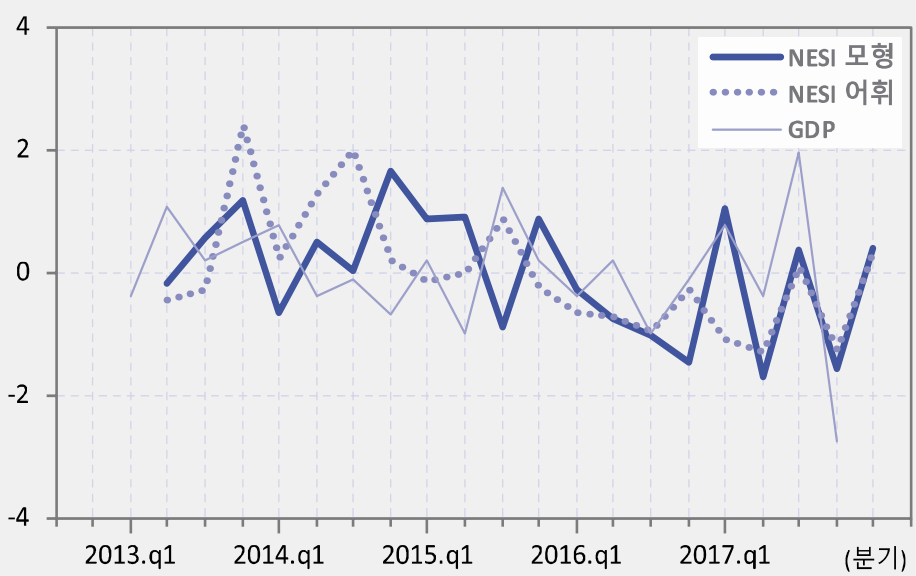
다음으로 분기별로 계산된 NESI가 실제 거시 경제지표와 밀접한 연관이 있는지 확인하기 위해, 거시 경제지표 중 하나인 실질 GDP (전기비)와 비교하였다. 실질 GDP는 기준연도 대비 변화율로 표현되기 때문에 NESI와는 계산 방식과 표기 방식이 다르다. 따라서 기준연도를 고정해서 각 지수를 산출하지 않는 한 정확한 비교는 어렵다. 차선책으로, 본 연구에서는 실질 GDP 지표를 앞선 방법으로 정규화하여 5년간의 변화 정도를 비교하였다.

<그림 4-2>를 살펴보면 온라인 뉴스 기사 데이터를 통해 얻어진 NESI의 움직임이 실질 GDP의 움직임과 유사함을 확인할 수 있다. 대체적으로 추정된 NESI가 실질 GDP보다 한 분기 정도 빠르게 움직이거나 거의 동일하게 움직이고 있음을 유추할 수 있다. 즉, 지표들 간 움직임의 유사성은 시점에 따라 그 형태가 조금 다른데, 특정 시점에서는 실질 GDP와 NESI가 약간의 갭을 두고 움직이기도 하며(2013년 2분기-2014년, 2016년 이후), 또 다른 특정시점에서는 거의 동일한 시점에서 비슷하게 움직이기도 한다(2015년 3분기-2016년). 시점에 따른 상관관계(time varying correlation)는 <그림 4-3>에서 조금 더 쉽게 확인할 수 있다. <그림 4-3>은 실질 GDP는 현재 분기를 기준으로 NESI는 전분기를 기준으로 도식화한 것인데, 같은 분기를 기준으로 그린 <그림 4-2>보다 지표들 간 움직임의 유사성을 조금 더 명확하게 파악할 수 있다.

〈그림 4-2〉 실질 GDP와 NESI 추이



〈그림 4-3〉 실질 GDP (현재 분기)와 NESI (전분기) 추이



Ⅴ. 결 론

본 연구는 온라인 뉴스 기사 빅데이터를 활용하여 새로운 경제심리지수(NESI)를 개발하는 것이 목적이었다. NESI 개발을 위해 두 가지 다른 접근법을 고려하였는데, 하나는 통계모형을 활용한 방법이고 다른 하나는 어휘사전을 활용한 방법이다. 통계모형 기반 방법론은 전처리 과정을 거친 텍스트 빅데이터의 차원을 축소하여 Bag-of-Words 행렬로 수치화한 후 로지스틱 회귀 분석 기법, 단순 베이즈 기법 및 신경망 기법 등 몇 가지 모형을 적용하여, 각각의 기사를 BSI/CSI 및 긍정/부정으로 분류하였다. 어휘사전 기반 방법론은 전처리 및 라벨링이 된 데이터로부터 각각 BSI/CSI 사전 및 긍정/부정 사전을 생성하였고 이를 활용하여 각 기사를 BSI/CSI, 긍정/부정으로 나누는 알고리즘을 개발하였다.

두 가지 방법론은 2017년 샘플 데이터를 활용하여 평가를 진행하였는데 전반적인 정확도는 대략 0.7~0.8 사이였다. 이는 기사 10개 중에 7~8개는 BSI/CSI 분류와 긍정/부정 분류가 제대로 이루어졌다는 뜻이다. BSI/CSI 분류에서는 통계모형 기반 방법론이 어휘사전 기반보다 정확하게 분류하였고, 감성 점수의 경우엔 상대적으로 어휘사전이 보다 정확하게 예측하였다. 다만, 통계모형에 사용된 설명변수들이 예측하고자 하는 뉴스 기사들에 사용된 단어까지 포함하여 Bag-of-Words로 구현해야 하는 것을 감안하면, 어휘사전 기반 방법이 상대적으로 강점이 더 많은 것으로 판단된다. 보다 정확한 비교를 위해서는 더 많은 트레이닝 및 테스트 데이터를 활용해야 할 것으로 생각된다.

본 연구결과를 토대로 몇 가지 추후 연구 과제를 정리하면 다음과 같다. 첫째, 본고에서 제안한 방법론은 뉴스 기사 단위(document-level)의 속성 탐지와 감성 분석이다. 최근에는 기사(문서) 단위(document level) 또는 기사의 문장 단위(sentence level)를 넘어 구(phrase)나 절 단위(clause level)까지 속성을 찾고 감성 점수를 부여하는 방법론이 개발(Im et al., 2018) 되었는데, 이는 뉴스 기사의 내용별로 BSI/CSI를 할당하고 감성 점수를 부여하는 것이 가능하게 되었음을 의미한다. 앞으로 한글 감성 사전과 분석기 등에 대한 많은 연구가 이루어지고, 어휘사전 업데이트를 위한 알고리즘이 개발된다면 보다 더 정확한 결과를 얻을 수 있을 것으로 기대된다.

둘째, 주기가 더 짧은 News ESI (NESI)를 작성하는 것이다. 본 연구에서는 경제심리를 내포한 온라인 뉴스 기사의 샘플 수가 충분하지 않았기 때문에 부득이하게 분기별로 지수를 산출하였다. 앞으로 전체 경제심리 기사를 활용한다면 월 단위뿐만 아니라 주 단위로도 지수를 작성할 수 있을 것이다. 이렇게 되면 새로운 경제심리지수의 활용도가 더욱 높아질

것이다.

셋째, NESI와 거시 경제지표의 관계를 제대로 살펴보기 위해서는 다변량 시계열 모형에 적합하는 과정을 거쳐 관계를 비교하는 등의 추가 분석이 필요하다. 연구 과제의 범위상 NESI와 실물 경제지표와의 연관성 분석이 심도 있게 진행되지 못하였다. 앞으로 제안된 방법론을 더 많은 뉴스 기사에 적용하여 보다 안정적인 NESI를 만들고 이를 여러 거시 경제지표와 비교하는 작업이 필요한 것으로 생각된다.

마지막으로, 새로 개발한 NESI를 어떻게 활용할 것인가에 대한 문제가 남아 있다. 설문조사를 통해 얻은 현행 ESI와 이번에 산출한 NESI를 결합한 지수 형태로 개발한다면 유용성이 높아질 것으로 판단된다. 측정오차모형을 이용하면 현행 공표 ESI와 NESI를 결합할 수 있을 것이다.

참고문헌

· 강승식 (2002).『한국어 형태소 분석과 정보 검색』, 홍릉 과학 출판사

· 송민채·신경식 (2017). "뉴스기사를 이용한 소비자의 경기심리지수", 「지능정보연구」 23, 1-27.

· 원중호·문혜정·손원 (2017). "텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류", 한국은행 「국민계정리뷰」 28 제4호, 1-27.

· 한국은행 (2014).『알기쉬운 경제지표 해설』, 경성문화사, 256-264.

· Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). "Latent Dirichlet Allocation", *Journal of Machine Learning Research 3*, 993-1022.

· Bram, J. and Ludvigson, S.C. (1998). "Does consumer confidence forecast household expenditure? A sentiment index horse race", *Economic Policy Review 4*, 59-78.

· Church, K.W. and Hanks, P. (1990). "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics*, 16, 22-29.

· Gelper, S., Lemmens, A. and Croux, C. (2007). "Consumer sentiment and consumer spending: decomposing the Granger causal relationship in the time domain", *Journal of Applied Economics 39*, 1-11.

· Hofmann, T. (1999). "Probabilistic latent semantic indexing", *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 50-57.

· Hu, M. and Liu, B. (2004). "Mining and summarizing customer reviews", *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168-177.

· Im, J., Song, T., Lee, Y. and Kim, J. (2018). "Confirmatory Aspect-based Opinion Mining Processes". Working Paper.

· Li, M., Bai, M., Wang, C. and Xiao, B. (2010). "Conditional random field for text segmentation from images with complex background", *Pattern Recognition Letters 31*, 2295-2308

· Liberti, J.M. and Petersen, M.A. (2018). "Information: Hard and Soft", *National Bureau of Economic Research Working Paper.*

· Lu, Y., Mei, Q. and Zhai, C. (2011). "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA", *Information Retrieval 14*, 178-203.

· Ludvigson, S.C. (2004). "Consumer confidence and consumer spending", *The Journal of Economic Perspectives 18*, 29-50.

· Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space", *ICLR workshop*.

· Park, E. and Cho, S. (2014). "KoNLPy: Korean natural language processing in Python", *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*.

· Pennington, J., Socher, R. and Manning, C. D. (2014). "GloVe: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 1532-1543.

· Popescu, A.M. and Etzioni, O. (2005). "Extracting Product Features and Opinions from Reviews", *Proceedings of the Conference on Human Language Technology and Conference on Empirical Methods in Natural Language Processing 2005 (HLT/EMNLP 2005). ACL*, 339 –346.

· Schouten, K. and Frasincar, F. (2016). "Survey on aspect-level sentiment analysis", *IEEE Transactions on Knowledge and Data Engineering 28*, 813- 830.

· Schouten, K., Weijde, O., Fransincar, F. and Dekker, R. (2018). "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data", *IEEE Transactions on Cybernetics 48*, 1263-1275.

· Shapiro, A.H., Sudhof, M., and Wilson, D. (2017). "Measuring News Sentiment", *Federal Reserve Bank of San Francisco Working Paper*.

· Titov, R. and McDonald, R. (2008). "A joint model of text and aspect ratings for sentiment summarization", *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 308-316.

· Zhong, P. and Wang, R. (2010). "Learning Conditional Random Fields for Classification of Hyperspectral Images", *IEEE Transactions on Image Processing 19*, 1890-1907

〈부록 1〉

한글 데이터 처리 과정 한계점

영어 등 서양 언어는 불용어를 제거(stopwords removal)하고 어간을 추출(stemming)하고, 원형을 복원(lemmatization)하는 단어 단위의 분석이 적합하고 보편적이지만, 한글은 의미 단위의 개념을 가지기 때문에 이 방법을 적용하기 까다롭다. 우리말은 교착어(agglutinative language)의 일종으로 어간(명사/동사)과 기능어(조사/어미)가 조합하여 어절을 구성하며, 상황에 맞게 문장을 활용한다.

한글과 영어의 언어적 차이
영어: I / go / to / school. (단어 단위)
한글: 나는 / 학교에 / 갑니다. (어절 단위)

한글 형태소 분석과정의 난제는 첫째, 분석이 모호하고[12], 둘째, 아직 표준화나 통일된 기준이 없으며[13], 셋째, '가계부채' 등 복합명사로 처리되어야 하는 단어가 '가계', '부채' 등 독립된 명사로 품사 태깅되어 의미구조가 약화된다는 어려움이 있다.

Shineware에서 제작한 오픈소스 형태의 코모란 형태소 분석기와 SNS 텍스트에 특화된 트위터 형태소 분석기로 수행한 텍스트 전처리 결과를 비교해 보면 범용적으로 활용할 수 있는 형태소 분석기가 없다는 것을 아래에서 확인할 수 있다.

원 경제기사 제목: '공영홈쇼핑 중소벤처 가전판매 연 500억 돌파'
– 코모란: 공영, 홈쇼핑, 중소, 벤처, 가전, 판매, 열, 돌파 (Good)
– 트위터: 공영, 홈쇼핑, 중소, 베다, 처, 가전, 판매, 연, 억, 돌파 (Bad)

원 경제기사 제목: '달러 가치 급락.. 유로에 1개월래 최저'
– 코모란: 가치, 급락, 유로, 월래, 최저 (Not good)
– 트위터: 달러, 가치, 급락, 유로, 개월, 최저 (Good)

12) '감기는'을 형태소 분석하면 아래와 같이 다양한 결과가 나타날 수 있다.
① (N 감기) + (J 는): 질병 '감기는'
② (V 감) + (E 기) + (J 는): "줄을 감다"의 '감기는'
③ (V 감기) + (E 는): "머리를 감기다"의 '감기는'

13) "아버지가방에들어가신다"는 형태소 분석기마다 다양하게 품사 태깅이 된다.
① 꼬꼬마(Kkma): 아버지 / NNG, 가방 / NNG, 에 / JKM, 들어가 / VV, 시 / EPH, ㄴ다 / EFN
② 코모란(Komoran): 아버지가방에들어가신다 / NNP
③ 트위터(Twitter): 아버지 / Noun, 가방 / Noun, 에 / Josa, 들어가신 / Verb, 다 / Eomi

본 연구의 텍스트 전처리 과정에서 해외의 선행연구를 바로 적용하기 어려운 또 다른 특징은 한글에서 명사의 중요성이다. 영어 텍스트 환경에서 감성분석은 동사와 형용사 품사가 중요한 반면에, 한글은 명사를 반드시 포함해야 하기 때문에 품사 태깅 단계에서 단어 차원을 줄이기 어렵다. 가령 '악화되다'라는 뜻의 'worse'의 경우 영어는 형용사이지만, 한글은 '악화', '되다'로 분리되기 때문에 '악화'라는 부정적 심리의 어휘를 포함하기 위해서는 반드시 명사를 포함해야 한다.

또한, 경제심리 판단은 아래와 같이 분야 자체의 특수성에 따른 어려움도 있다.

① '숨통 텄다' 코스피 장중 1360선 **돌파**…환율 1200원선 **급락 (긍정)**
② 국제유가 115弗 **돌파** 또 사상최고 (부정)
③ 집값 **급락** 가능성 **제한적** (긍정)

위의 예제에서, ①의 경우 '급락'은 보통 부정적 의미의 단어이지만, 이 경제기사에서는 글로벌 투자심리가 회복됨에 따라 우리나라 주식가격도 상승하고 원/달러 환율이 하락하였다는 긍정적인 의미로 사용되었다. 반면 ②에서의 '돌파'는 유가가 상승하여 국내 기업 및 가계에 비용부담으로 작용한다는 부정적인 의미를 전달하면서, ①의 긍정적인 기사와 상반되는 심리로 사용되었다. 한편 ③에서의 '급락'은 부정적 의미로 사용되었지만, 뒤의 '제한적'이라는 단어가 이를 이중부정하면서 최종적으로 긍정적 심리를 나타낸다. 따라서 경제심리의 긍정적인 정도를 예측하는 문제는 경제 심리의 유무보다 복잡하다. '농구', '손흥민' 등의 단어가 많으면 스포츠로, '국회', '선거' 등의 단어가 많으면 정치로 신문기사의 종류를 예측하는 문제보다 더 기준이 까다롭다.

〈부록 2〉

어휘사전 생성 알고리즘

아래 소개된 알고리즘은 Im *et al.* (2018)에 소개된 어휘사전 생성 알고리즘의 한국어 버전이다.

Input: 극성(속성)이 부여된 온라인 기사 집합

(Step 1) 극성에 따라 각각의 기사 집합을 단어(혹은 구 단위) 집합의 형태로 재구성한다.

BSI 분류 기사들의 단어 집합 (A) VS CSI 분류 기사들의 단어 집합 (B)
긍정 분류 기사들의 단어 집합 (C) vs 부정 분류 기사들의 단어 집합 (D)

(Step 2) 불용어와 공통 단어들을 각 단어 집합에서 제외한다.

(Step 3) 차분하여 각각의 속성에 맞는 어휘사전을 정의한다.

BSI 분류 사전: A＼B
CSI 분류 사전: B＼A
긍정 분류 사전: C＼D
부정 분류 사전: D＼C

본 부록에 소개된 어휘사전 생성 알고리즘은 지도학습 방법론으로 이해될 수 있다. 분류 속성이 부여된 기사들을 활용하여 각각의 분류에 적용될 수 있는 사전을 생성하기 때문이다. 이렇게 생성된 사전은 추후에 라벨링이 되어 있지 않은 기사들을 이용하여 업데이트할 수 있는데, 단어들의 co-occurrence 구조를 파악하여 Schouten *et al.* (2018)이 제안한 알고리즘을 적용하면 된다.

〈부록 3〉

BSI/CSI 및 긍정/부정 어휘사전 예시

〈3-1〉 BSI/CSI 사전 예시

구 분	단 어
BSI 사전	국제무역, 계획경제, 신재생에너지, 자금세탁, 상업시설, 무역흑자, 코스닥시장, 시장점유율, 자본잠식, 가공무역, 상장회사, 무역흑자, 무역특화지수, 이머징마켓, 보복조치, 통신업, 농작물, 동맹국, 주문량, 선물옵션, 기간산업, 우대금리, 회계기준, 생산요소, 다국적 기업, 셰일오일, 통상협정, 신용거래, 차익거래, 전자상거래
CSI 사전	소비자 보호법, 가스요금, 파트타임, 생활비, 궁핍, 노년, 운전자, 소시민, 청춘, 최저생계비, 출근시간, 기대수명, 가스요금, 임용고시, 양도소득, 잔여재산, 가난, 집세, 아르바이트생, 알뜰, 담뱃값, 인간관계, 노동시간, 교육비, 노점상, 무기계약직, 디딤돌대출, 궁핍, 누진세, 월셋값

〈3-2〉 긍정/부정 사전 단어 예시

구 분	단 어
긍정 사전	메리트, 상쇄, 흥행, 점증, 훈훈하다, 순조롭다, 성공사례, 활약, 대단하다, 순항, 고취, 깔끔하다, 도와주다, 드디어, 고밸류, 업그레이드, 대등, 점령, 밝아지다, 뚜렷하다, 준법, 풀어지다, 선명하다, 안전지대, 호감, 귀재, 첨가, 붐, 건강하다, 꿰찬
부정 사전	단절, 다급하다, 위험수위, 변질, 쪼그라들, 과소, 무분별, 슬프다, 자살, 답보, 불가항력, 구멍, 족쇄, 진퇴양난, 버겁다, 불리, 분주하다, 위태롭다, 불공평, 불충분, 가난, 고비, 당하다, 후유증, 문책, 멸렬, 아슬아슬, 내몰렸, 농락, 앓다

〈부록 4〉

분기별 NBSI, NCSI 및 NESI 점수

〈표〉 분기별 NBSI, NCSI, NESI 점수, 공표 ESI 및 실질 GDP

분기	NBSI_M	NCSI_M	NESI_M	NBSI_L	NCSI_L	NESI_L	ESI	Real_GDP[1)]
2013.1	77.4	63.4	98.4	71.4	33.3	95.4	94.1	0.6
2013.2	89.3	57.1	105.8	74.2	34.0	97.1	95.1	1.1
2013.3	92.1	82.8	111.9	112.8	64.9	123.9	93.5	0.8
2013.4	65.7	84.2	93.5	77.6	53.3	102.4	95.6	0.9
2014.1	88.4	56.6	105.1	103.1	33.3	112.8	99.3	1.0
2014.2	82.4	52.9	100.3	115.1	35.9	119.9	98.2	0.6
2014.3	104.2	59.3	116.7	72.1	66.7	101.8	95.6	0.7
2014.4	89.6	75.0	108.9	69.2	57.1	98.5	94.1	0.5
2015.1	84.4	100.0	109.2	71.6	58.1	100.0	98.5	0.8
2015.2	70.0	50.0	91.1	97.3	27.8	108.7	96.6	0.4
2015.3	91.1	68.2	108.9	81.0	16.7	97.7	93.7	1.2
2015.4	78.8	50.0	97.3	68.8	30.3	93.5	93.2	0.8
2016.1	67.2	70.6	92.4	63.3	43.6	92.8	90.6	0.6
2016.2	63.2	72.0	89.8	63.8	28.6	90.4	93.0	0.8
2016.3	59.2	61.5	85.3	75.8	30.0	97.2	94.7	0.4
2016.4	90.9	77.8	110.3	71.0	0.0	89.2	92.8	0.7
2017.1	62.8	31.2	83.0	58.2	26.7	87.0	95.8	1.0
2017.2	92.5	29.9	103.8	82.9	28.6	100.9	98.1	0.6
2017.3	65.7	27.4	84.4	60.6	21.9	87.4	97.3	1.4
2017.4	90.3	41.5	104.0	85.3	34.6	103.3	100.1	−0.2

주 : 1) 실질GDP 전분기대비 성장률(%)

※ _M은 통계모형 기반, _L은 어휘사전 기반에 의한 지수를 의미함.

※ 해당 점수는 분기별 샘플 뉴스 기사로만 작성된 것이기 때문에, 분기 내 기사숫자가 적은 경우 점수의 변동성이 더 커질 수 있음. 예를 들어, 2016년 4분기 CSI 점수의 경우 0인데 이는 샘플 기사 중에서 긍정으로 분류된 기사가 해당 분기에 없다는 것을 의미한다. 경제심리가 포함된 전체 기사로 확장할 경우 해당 점수는 0보다는 큰 값을 가질 것으로 예측된다.