경제용어 감성사전 구축방안 연구

전종준*, 안승환**, 이문희***, 황희진****

뉴스 기사를 이용한 정보의 추출 및 가공은 분석의 적시성과 유연성, 경제적 효용성을 갖추면서 관련 연구가 활발히 진행되어 왔다. 기존의 연구에서 경제적 현상을 설명할 수 있는 감성단어를 선택하여 그 단어의 빈도를 이용해 지표를 만드는 방법을 제안하였고, 해당 모형의 해석력을 고려한 결과 경제현상을 추론할 수 있는 감성단어의 빈도를 이용하는 것이 장점이 크다는 결과가 있었다. 이에 본 연구는 경제현상을 나타내는 감성단어를 경제기사로부터 추출하는 방법에 대해 연구한다.

본 연구에서는 기본적으로 기계학습을 통해 기사 내에 있는 단어를 저차원 공간에 임베딩하는 방법을 적용한다. 이때 단순히 문맥의 정보만을 이용해 텍스트의 정보를 축약하는 것이 아니라, 감성분류의 방향을 고려하여 텍스트의 축약공간을 조정한다. 이렇게 감성이 고려된 텍스트의 축약공간 위에서 우리가 관심이 있는 주제단어를 대응시키고 그것에 가장 유사한 단어를 찾아내는 방법으로 감성사전을 구축한다.

Ⅰ. 서론

Ⅱ. 연구방법
1. 전처리 방법
2. 학습모형 및 계산
3. 감성사전의 구축

Ⅲ. 모형 적합 방법 및 비교
1. 전처리 방법의 비교
2. 학습 단위의 비교

Ⅳ. 분석결과
1. 심리 주체별 감성단어 사전의 구축
2. 기업과 소비자 간 감성단어 비교
3. 시사점

Ⅴ. 결론

부록

참고문헌

* 서울시립대학교 통계학과 (e-mail: jj.jeon@uos.ac.kr, phone: 02-6490-2637)
** 서울시립대학교 통계학과 학생연구원(e-mail: dpeltms79@gmail.com, phone: 02-6490-5693)
*** 한국은행 경제통계국 (e-mail: mhlee@bok.or.kr, phone: 02-759-5464)
**** 한국은행 경제통계국 (e-mail: hjhwang@bok.or.kr, phone: 02-759-4443)
※ 본 연구의 내용은 집필자들의 개인 의견으로 한국은행의 공식견해를 나타내는 것은 아님.

Ⅰ. 서론

뉴스 기사를 이용한 정보의 추출 및 가공에 대한 연구는 인터넷으로부터 적은 비용으로 많은 데이터를 수집할 수 있게 되면서 활발하게 연구되어 왔다. 예를 들어 경제 뉴스 기사는 사람들이 관심을 가지고 있는 경제 현상을 설명하거나 향후 경제전망을 포함하고 있어, 기존에 시간과 비용이 많이 들었던 경제조사 활동들을 뉴스기사 정보를 이용하여 보완하거나 혹은 대체할 수 있는 가능성에 대해서 연구가 이루어져 왔다. 이는 TV, 신문, 라디오를 통해 기사가 전달되었던 기존 미디어에 비해 인터넷을 통한 뉴스 기사의 생산, 유통, 소비가 활발해지면서 이러한 정보를 수집하는 비용이 급격하게 낮아진 것에 기인한다.

인터넷 뉴스 기사가 가지고 있는 정보가 경제 현상을 설명하는 효과적인 정보를 담고 있다는 가정이 성립할 경우 이 정보에 대한 효율적인 추출은 기존 조사방법이 가질 수 없었던 장점을 가진다. 먼저, 잘 갖추어진 시스템 하에서 적시성 있는 정보를 추출할 수 있다. 인터넷 기사는 실시간으로 수집 가능하므로 경제현상에 대한 민감한 변화를 즉각적으로 파악할 기회가 생긴다. 이는 잘 설계된 조사계획에 의해 경제지표를 도출해왔던 기존 조사방법과 구별되는 큰 차이점이다. 그리고 시스템이 정보를 추출하므로 정보의 처리 및 가공이 조사·분석을 수행하는 사람의 역량에 크게 의존하지 않는다는 장점이 있다. 뿐만 아니라, 축적된 데이터를 이용하여 여러 형태의 분석이 가능하다는 점에서 분석의 유연함을 줄 수 있다. 하지만 인터넷 뉴스 기사에 경제 현상을 설명할 수 있는 좋은 정보가 있는지 확증하거나, 혹은 문자로 이루어진 기사 자료에서 정보를 효율적으로 추출해 낼 수 있는 방법 및 그것이 가지는 적시성, 기존 조사방법과 비교한 효율성에 대한 문제는 잘 알려져 있지 않다.

김현중 외 3명(2019)은 경제 뉴스 기사를 이용하여 얻어진 정보를 이용해 경제심리지수(ESI), 소비자동향지수(CSI) 및 기업경기실사지수(BSI)를 근사하는 방법론을 연구하였다. 그 연구에서는 경제 뉴스 기사가 감성분류 라벨을 가지고 있는 경우, 분류된 감성라벨을 이용하여 기사의 텍스트 정보를 추출함으로써 경기지수와 상관성이 높은 지표를 생산할 수 있음을 보였다. 이 연구는 감성분류 라벨의 출현 확률이 기사의 특정한 단어와 연관이 있을 것이라 가정하였고, 이를 이용한 분류모형을 통해 설명력 있는 지표를 생산하였다.

김현중 외 3명(2019)의 결과에서 경제 뉴스 기사에 나온 단어의 특성값을 학습하여 직접 경제지표를 만드는 방법론과 경제적 현상을 설명할 수 있는 감성단어를 선택하여 그 단어의 빈도로 지표를 만드는 방법을 비교하였다. 연구 결과 제안한 두 방법론에 따라 산출된

생산 지표의 성능이 다르게 나오기는 했지만, 해석력을 고려한 결과 경제현상을 추론할 수 있는 감성단어의 빈도를 이용하는 것의 장점이 크다는 결론을 내렸다. 이에 본 연구는 경제현상을 나타내는 감성단어를 경제기사로부터 추출하는 방법을 연구한다.

주어진 데이터는 한국정보화진흥원에서 제공받은 2008.1월~2018.12월 인터넷 경제 뉴스기사 중에서 한국은행이 표본 14,000여 건을 추출하고 경제 관련 감성의 유무와 방향을 분류하여 라벨을 부여한 것이다. 감성라벨은 같은 데이터에 대해 기사별 혹은 문장별로 할당되어 있으며, 각 경제 주체인 기업, 소비자, 기타[1]의 세 종류의 감성 점수로 표시되어 있다. 라벨 점수는 부정에서 긍정에 이르는 리커트 척도를 사용하였다. 즉 감성라벨은 라벨링 대상이 기사인지 문장인지에 따라 두 가지로 구분되고 라벨 점수가 기업 측면인지 소비자 측면인지 혹은 기타인지에 따라 세 가지로 구분되어 총 여섯 가지 다른 방법으로 표시되어 있다.

본 연구에서는 감성라벨이 있는 경제 뉴스기사를 이용하여 기계학습을 통해 기사 내에 있는 단어를 저차원 공간에 임베딩[2]하는 방법을 적용한다. 단순히 문맥의 정보를 이용해 텍스트의 정보를 축약하는 것이 아니라, 감성분류의 방향을 고려하여 텍스트의 축약공간을 조정한다. 이렇게 감성이 고려된 텍스트의 축약공간 위에서 우리가 관심이 있는 주제단어를 대응시키고 그것에 가장 가까운 단어를 찾아내는 방법으로 감성사전을 구축한다. 2장에서는 텍스트를 전처리하는 두 가지 방법을 소개하고, 감성단어를 찾기 위한 방법을 소개한다. 3장에서는 전처리 방법과 학습의 단위에 따른 감성단어 추출결과를 간략하게 비교한다. 4장은 본 연구에서 선택한 전처리 방법과 학습의 단위, 자료에 따라 찾아낸 감성단어들을 경제주체별, 긍정-부정단어별로 소개하고 시사점을 제시한다. 마지막으로 5장에서 연구의 시사점을 검토하고 향후 연구방향을 고찰한다.

1) 기사 또는 문장 내에 명확한 경제심리 주체(기업 또는 소비자)가 없는 경우

2) 단어는 사전 내의 하나의 원소로 이산형 변수로 다룰 수 있다. 만약 사전이 10만개의 단어를 포함하고 있다면 이 변수는 10만 차원 내의 한 원소에 해당한다. 이 변수는 고차원인 반면 단 하나의 원소만 1이고 나머지는 0인 특징을 가지고 있다. 차원은 높지만 원소가 가진 정보의 단순함을 이용하여 정보의 손실을 최대한 줄이면서 낮은 차원의 원소(예: R^m 위의 실수)로 변환하는 것을 생각해볼 수 있다. 이를 단어의 임베딩이라고 한다.

Ⅱ. 연구방법

기계학습은 컴퓨터가 학습할 수 있도록 도와주는 알고리즘(모형)과 학습의 목표를 개발하는 인공지능 분야 중 하나이다. 설명변수와 반응변수로 이루어진 데이터가 주어졌을 때, 설명변수를 이용해 반응변수를 잘 설명할 수 있는 모형이 적합되었다면 이는 학습이 잘 진행되었다고 말할 수 있다. 본 연구에서는 설명변수인 감성라벨이 있는 경제 뉴스기사를 이용해 반응변수인 문맥 정보와 감성라벨 정보를 잘 설명할 수 있는 모형을 적합하는 것이 학습의 목표이다.

1. 전처리 방법

단어는 우리말에서 자립성이 있으며 일정한 뜻을 가지는 가장 작은 단위이다. 감성사전 구축에 관한 연구를 진행하기 위해서는 우선 전체 문서 데이터를 단어들로 분해하여 데이터를 재구성하는 과정이 필요하다. 일반적으로 텍스트 분석 연구에서 문장에서 단어를 분리해 나가는 방법에 따라 목표로 하는 텍스트 분석 모형의 성능이 크게 변화할 수 있어, 전처리 방법을 선택하는 것은 중요한 일이다. 뿐만 아니라 전처리 방법의 선택은 분석 목적, 데이터의 크기, 학습 알고리즘의 복잡도에 의존할 수 있으므로 본 연구에 맞는 전처리 방법을 어떻게 선택할지 먼저 조사하였다. 여기서는 한국어의 품사 특성을 반영할 수 있는 품사 기반 단어 전처리 방법과 품사에 의존하지 않고 단어를 추출하는 비지도 학습 방법을 소개한다.

가. 품사 기반 방법

품사 기반 단어 추출기는 기존에 학습된 사전과 규칙을 기반으로 문장으로부터 단어를 추출하는 방식의 추출기이다. 본 연구에서는 Python 프로그래밍 언어의 KoNLPy 패키지에서 제공하는 오픈소스 한국어 처리기(tokenizer)인 Okt class를 사용하였다. KoNLPy는 한국어 정보처리를 위한 Python 언어로 작성된 패키지이고 Okt class는 KoNLPy가 제공하는 여러 tag 패키지 중 하나다. Okt class는 한국어에 대한 텍스트 정규화(normalization), 토큰화(tokenization), 어근화(stemming), 어구 추출 등의 다양한 기능을 지원한다. Okt class를 이용해

분석 가능한 품사의 종류는 총 19가지로, KoNLPy가 제공하는 다른 tag 패키지들과 비교해 볼 때 적은 편이다. 하지만 본 연구에서는 세분화된 품사 분류가 필요하지 않아 상대적으로 속도가 빠른 Okt class를 사용하였다.

〈표 1〉 **품사 기반 단어 추출 결과 예시**

입력 문장
금융위기 이후 대기업과 중소기업 자금조달의 양극화가 갈수록 심화하고 있는 것으로 나타났다
출력 단어
('금융위기', 'Noun') ('이후', 'Noun') ('대기업', 'Noun') ('과', 'Josa') ('중소기업', 'Noun') ('자금', 'Noun') ('조달', 'Noun') ('의', 'Josa') ('양극화', 'Noun') ('가', 'Josa') ('갈수록', 'Noun') ('심화', 'Noun') ('하고', 'Josa') ('있다', 'Adjective') ('것', 'Noun') ('으로', 'Josa') ('나타나다', 'Verb')

나. 비지도 학습 기반 방법

비지도 학습 기반 단어 추출기는 단어의 품사 정보를 이용하지 않고, 한글 고유 어절의 구조적 특성을 이용하여 주어진 데이터만을 이용하여 단어를 추출하는 방식의 추출기이다. 여기서 사용한 한글 어절의 구조적 특성은 다음과 같다. 한글은 어절의 왼쪽 부분에는 의미를 지니는 단어, 즉 명사, 동사, 형용사, 부사, 감탄사 등이 사용된다. 그리고 어절의 오른쪽 부분에는 문법적 기능을 하는 단어인 조사가 등장한다. 이렇게 어절의 구성을 왼쪽과 오른쪽 부분으로 구분할 때, 단어는 어절, 어미 등도 포함하는 넓은 의미라고 생각할 수 있다. 단일 어절을 이루는 명사, 부사와 같은 단어는 어절의 오른쪽 부분이 없다고 생각할 수 있다. 그리고 복합명사를 하나의 단어로 생각한다면 동일한 방식으로 복합명사가 쓰인 어절도 왼쪽과 오른쪽 부분으로 구분할 수 있다.

이러한 한글 어절의 구조적 특성을 바탕으로 단어 가능 점수[3]를 정의하여 해당 점수를 기준으로 문장에서 유의미한 단어를 추출한다. 본 연구에서 사용한 단어 가능 점수는 두 가지로 Cohesion score와 Branching entropy를 이용하였다. 이 두 가지 단어 가능 점수에 대해서는 <부록 1>에서 자세히 설명한다. <표 2>는 두 방법을 이용한 비지도 학습 기반 단어 점수 계산 예시이다.

3) 주어진 데이터의 정보에 기반하여 해당 단어가 문서에서 유의미한 단어일 가능성을 나타내는 점수. 일정한 기준의 단어 가능 점수를 넘는 단어는 유의미한 단어로 분류한다.

〈표 2〉 비지도 학습 기반 단어 점수 계산 예시

단어: '뉴욕증시가'	Cohesion score	Branching entropy
'뉴'	0	0.66
'뉴욕'	**0.84**	**2.41**
'뉴욕증'	0.59	0.49
'뉴욕증시'	**0.65**	**2.01**
'뉴욕증시가'	0.53	**3.41**

2. 학습모형 및 계산

추출된 단어 사이의 유사도 등의 계산에 필요한 단어의 특성벡터(embedding vector)를 얻기 위한 학습모형으로는 Sentiment Specific Word Embedding Model(Fu, P. et al., 2018)을 이용하였다. 이 모형은 단어 주변의 문맥 단어들의 정보와 문서의 감성라벨 정보를 동시에 이용해 학습하여 단어를 표현하는 저차원 벡터를 구하는 방식이다. 여기서는 문맥 정보 모형과 감성 정보 모형에 대응되는 문맥 손실함수와 감성 손실함수를 각각 정의하고 두 개의 손실함수의 가중 평균을 전체 손실함수로 두었다. 학습 모형은 이 전체 손실함수를 최적화한다. 감성 손실함수의 손실을 줄여주는 방향으로 본 연구에서 사용된 모형을 적합하기 위해서는 문서 혹은 문장에 감성라벨이 포함되어 있어야 하므로, 감성라벨이 없는 경우에는 사용할 수 없음을 알려둔다[4].

문맥 정보는 문장 내에서 단어의 문법적 연결성을 의미하며 손실함수는 CBOW 모형(Mikolov, T. et al., 2013a)과 같이 단어의 출현에 대한 조건부 확률의 우도를 이용하여 모형화하였다. 문법적 연결성이 높다는 뜻은 문장을 이루는 연속된 단어의 집합들이 우리가 가지고 있는 데이터에서 흔히 관찰된다는 것을 뜻한다. 예를 들어 '인공지능을 이용한 예측모형'이라는 원문을 생각해보자. 여기서 이 예문은 단어 추출을 통한 전처리 과정을 거쳐 다음과 같이 '인공지능 이용 예측모형'으로 정제되었다고 가정하자. 정보기술 논문들을 모아둔 문서집합을 다루고 있다면, 예문의 첫 번째 단어인 '인공지능'과 세 번째 단어인 '예측모형'이 주어진 경우 두 번째 단어인 '이용'은 높은 확률로 관찰될 것이다. 즉, 이 문장은 높은 조건부 확률을 가질 것이다. 문서 집합 내에 첫 번째 단어인 '인공지능' 대신 '기계학

4) 감성라벨이 표시된 문서나 문장의 숫자가 적을 때에는 감성분류모형을 먼저 적합한 후 추정된 감성라벨을 이용하여 모형을 적합한다.

습', 'AI'와 같은 단어로 대치된 문장 역시 흔히 관찰할 수 있을 것이고, 이러한 문장을 문법적 연결성이 높은 문장이라 할 수 있다. 반면 '인공지능'이라는 단어 대신 '달리기'라는 단어가 온다고 생각하면 '달리기 이용 예측모형'은 빈도가 낮을 것이며 문법적 연결성이 낮은 문장이라 볼 수 있다.

문법적 연결성을 측정하는 관측빈도의 계산은 반응변수와 설명변수가 주어진 다항분류모형으로 생각할 수 있다. 위 예문에서 설명변수는 '인공지능', '예측모형'으로 반응변수는 '이용'으로 주어져 있다고 생각해보자. 만약 범주형 변수인 단어를 모두 실수 벡터로 변환하였다고 생각하면 이 문제는 두 개의 설명변수 '인공지능', '예측모형'이 주어진 경우 그 두 단어 사이에 올 단어를 예측하는 다항분류모형으로 적합할 수 있을 것이다. 만약 다항분류모형을 적합할 때 분류 정확도 혹은 모형 적합도를 다항로지스틱 모형의 우도함수를 사용하였다면 사용된 손실함수는 다항분포의 음우도함수로 주어질 것이다. CBOW 모형이 문장의 문법적 구조를 단어 간 분류모형으로 모형화하고 분류모형을 적용한 대표적인 예다. 여기서 특이한 점은 이 분류모형을 적합하는 가운데 이산형 변수인 단어를 연속형 변수인 특성벡터로 변환하는 과정을 포함시키면서 분류 정확도를 최대로 만들어 줄 수 있는 단어의 실수 표현형을 찾는다는 것이다.

예를 들면 잘 적합된 분류모형이라면 다음과 같이 확률 P('이용' | '인공지능', '예측모형'), P('이용' | '기계학습', '예측모형'), P('이용' | 'AI', '예측모형')에 대한 예측확률은 모두 높을 것으로 예상할 수 있다. 그렇다면, '인공지능', '기계학습', 'AI' 세 단어의 특성벡터가 서로 가까이 있어야 한다고 생각할 수 있고 CBOW 모형은 이 기준 하에서 범주형 변수인 단어를 실수벡터로 변환하는 좋은 대응을 찾을 것이다. 즉, 문법적 연결성을 최대로 만들어 줄 수 있는 단어의 변환 과정을 문법적 학습모형이라 부른다.

감성 정보는 감성분류 정확도를 의미하며 감성분류 라벨에 기반한 분류모형을 이용하여 손실함수를 정의하였다. 감성분류모형의 분석 단위는 감성라벨이 표기된 문장 혹은 문서가 된다. 감성분류모형은 다항분류모형을 이용하였으며, 설명변수로는 문장 혹은 문서를 요약한 특성벡터, 반응변수로는 감성라벨을 사용하였다. 감성분류모형 적합을 위한 문서, 문장의 변환은 해당 문서와 문장을 이루는 단어의 특성벡터의 평균으로 정의한다. 감성분류모형에서는 감성적 분류가 잘 되는지 혹은 그렇지 않은지를 기준으로 모형을 적합하기 때문에, 범주형 변수인 단어를 실수 값을 가지는 특성벡터로 변환하는 학습이 감성분류 정확도를 높여주는 방향으로 이루어진다. 이 모형을 감성 학습모형이라 부른다.

다음과 같은 두 문장을 생각해보자. '인공지능 이용 생산성', '인공지능 이용 실업'. 이 두 문장에 각각 긍정과 부정의 감성라벨이 달려있다고 생각해보자. 그러면 이 감성라벨의 차이는 '생산성'과 '실업'에서 발생하였음을 예상할 수 있다. 즉 감성분류모형은 이 두 단어가

분류에 중요한 역할을 하도록 모형을 예측해야 할 것이다. 그러기 위해서는 '생산성'과 '실업'이라는 단어는 숫자 공간에서 비슷한 위치가 아니라 멀리 떨어진 위치에 표시되어야 감성분류모형이 잘 작동할 것이다. 이는 문법적 연결성이 아니라 감성분류 라벨에 의해 단어의 특징이 어떻게 표현되어야 하는지를 학습하는 것이다.

우리가 사용한 모형은 문법 학습모형과 감성 학습모형에 사용한 단어의 특징 추출함수를 공유한다. 앞서 언급한 문법 학습모형과 감성 학습모형이 문법적 연결성을 최대로 하거나 감성적 분류 정확도를 최대화하는 방향으로 각각 단어의 특성벡터를 추출했다고 하면, 여기서는 두 개의 다른 분류 기준의 가중 평균으로 새로운 기준을 정하였고, 단어의 특징 추출이 그 분류 정확도를 최대화하는 방향으로 이루어지는 모형을 사용하였다. 결과적으로 감성적 분류기준과 문법적 연결성을 함께 고려한 단어의 표현형을 찾는다. 모형에 사용한 중요한 조절 모수로 α와 β가 있다. α는 감성분류모형에서 문장, 문서의 대표 벡터의 크기를 조절하고, β는 문법 학습모형과 감성 학습모형의 가중평균을 결정한다[5].

가. 학습 내용의 세부 설정

감성단어 추출을 위해 각 심리 주체(기업, 소비자, 기타)별로 감성이 있다고 표시된 기사와 문장만을 선택하여 사용하였다. 그리고 심리 주체별로 학습된 감성사전 결과를 비교할 수 있도록 심리 주체별로 모형을 별도로 학습하였다.

1) 문법 학습모형의 데이터 설정

문맥 손실함수는 조건부 확률을 예측하는 분류모형에 기반한다. 이 분류모형에 사용되는 반응변수는 품사 정보를 활용하여 정제하였다. 품사 기반 전처리 모형을 통해 도출된 단어에는 품사에 대한 정보를 저장해두어 이후 분석에 결과를 사용한다. 예를 들어 '경기'라는 단어가 긍정 감성라벨에 사용되었고 명사로 품사 할당이 되어 있는 경우 '경기_2_Noun'[6]으로 구분하여 분석하였다. 이는 감성과 연관된 단어가 문맥상 다른 의미를 가지게 될 수 있음을 반영하기 위해서다. '경기가 좋다', '경기가 나쁘다'라는 두 문장이 각각 긍정과 부정의 감성라벨을 가지고 있다고 가정하자. '좋다', '나쁘다'라는 단어는 감성라벨과 직접 연결

5) Ⅱ-2장 다. 문서 대표 벡터, 바. 모형 학습 참고
6) 부정 감성라벨에서 '경기'가 관찰된 경우 '경기_0_Noun', 중립 감성라벨에서 '경기'가 관찰된 경우 '경기_1_Noun'이라 표시하였다. 본고에는 따로 감성 정보와 품사 정보를 표기하지는 않았다.

이 되어 감성을 예측할 수 있지만 '경기'라는 단어를 통해 감성라벨을 분류할 수가 없을 것이다. 하지만 경기라는 단어는 감성사전을 구축하는 데 중요한 단어로 사용될 수 있는 연관어로 볼 수 있기 때문에 이 단어를 감성사전을 구축할 때 사용할 수 있는 단어로 찾아내고자 하였다. 따라서 감성라벨이 포함된 문장이나 문서 내에서 구분하여 단어를 사용하였다.

그리고 문맥 단어들이 문맥 정보 모형에서 특정 품사의 단어들만을 반응변수로 예측하도록 학습을 진행하였다. 예를 들면 문법 손실함수에서 반응변수로 명사, 동사, 형용사만을 고려하였고 그 외의 품사가 반응변수로 오는 확률함수를 학습모형에서 제외하였다. 설명변수는 모든 단어를 사용하였다.

2) 감성 학습모형의 데이터 설정

세부적으로 감성 정보 모형에서 사용하는 감성라벨은 다음과 같이 처리하였다. 본 연구에서 사용한 감성 문서 데이터는 각 기사별 혹은 문장별로 경제심리 없음, 긍정, 부정, 중립, 불확실한 긍정, 불확실한 부정 총 여섯 가지로 분류되어 있다. 이 중 경제심리가 없는 기사 또는 문장은 분석에서 제외하였고, 긍정과 불확실한 긍정을 하나의 긍정감성으로, 부정과 불확실한 부정을 하나의 부정감성으로 구분하여 경제심리가 있는 기사 또는 문장을 긍정, 부정, 중립의 3가지 감성 극성으로 축소하여 표현하였다. 기사별로 표시된 감성라벨은 기사 내에 표현된 모든 단어에 감성라벨을 할당하였고, 문장별로 표시된 감성라벨은 문장 내에 표현된 단어에만 감성라벨을 할당하였다.

3) 증감단어의 처리

경제기사의 특성을 반영하기 위해 세 가지 감성분류 기준 외에 감성이 포함된 경제기사의 특성을 추가적으로 이용하였다. 감성이 있는 경제기사가 같은 감성을 가지더라도 사용되는 단어에는 대부분 특정한 방향(증가 또는 감소)이 존재한다는 점을 반영한 것이다. 예를 들어 '실업률 감소'라는 내용의 문장과 '수출 증가'라는 내용의 문장은 모두 긍정 감성이지만, 사용되는 단어의 방향은 감소와 증가로 반대를 나타낸다. 증감을 나타내는 단어의 감성특성을 추출하면 경제기사의 정보를 효과적으로 요약할 수 있다는 가정 하에 감성이 있는 각 문장에 대하여 <표 3>과 같은 증감단어의 출현 빈도를 이용하여 증가, 감소, 증감 없음 이 세 가지의 추가적인 정보를 표시하였다. <표 3>의 증감단어는 김현중 외 3명

(2019)의 연구결과를 기반으로 선정하였다. 그 결과 (긍정, 증가), …, (부정, 감소) 등 총 9개의 감성라벨을 이용해 감성 데이터 문장들의 라벨을 새롭게 구성하였고, 감성 정보 모형의 학습을 진행하였다.

〈표 3〉 **증감 단어 목록**

증가 단어
'가중', '가속', '강화', '격화', '고조', '급증', '급등', '넘어', '넘다', '늘다', '늘어', '높아', '높이', '높게', '돌파', '상승', '상향', '상회', '솟', '심하', '심해', '심화', '오르', '올라', '올리', '인상', '증가', '증대', '증진', '증폭', '촉진', '최대', '최고', '커지', '커져', '폭증', '폭등', '확대', '확산', '확장'
감소 단어
'감소', '감축', '낮아', '낮춰', '낮추', '내려', '내리', '둔화', '떨어', '미미', '수축', '약화', '인하', '작아', '적어', '줄다', '줄어', '줄이', '지연', '지체', '축소', '최소', '최저', '하강', '하락', '하향', '하회'

나. 기호

다음에 설명할 모형과 그 손실함수들을 정의하기 위한 기호들은 다음과 같다.

- C : 전체 단어집합
- T_i : i번째 문서내의 단어 개수
- $w_i^t \in C$: i번째 문서의 t번째 단어
- x_i^t: 단어 w_i^t의 one-hot 벡터[7](∈$R^{|C|}$), 더미 변수(dummy variable) (단 |C|는 집합 C의 원소 개수)
- C_i^t : i번째 문서의 t번째 단어 주변의 주변 문맥 단어집합
- U : ($m \times |C|$ 행렬) 문맥 정보 모형에서 학습되는 모수 행렬로, 주어진 문맥 단어들의 one-hot 벡터(∈$R^{|C|}$)와 문서의 대표 벡터(∈$R^{|C|}$)를 m차원 임베딩(embedding) 공간으로 사상(mapping)시키는 역할을 한다.
- V : ($|C| \times m$ 행렬)문맥 정보 모형에서 학습되는 모수 행렬로, 임베딩 공간으로 사상된 벡터를 모형의 학습 목표인 목표 단어의 one-hot 벡터로 사상시키는 역할을 한다.
- W_1 : ($q \times m$ 행렬) 감성 정보 모형에서 학습되는 모수 행렬로, U행렬에 의해 임베딩

7) 하나의 단어를 표현할 때는 범주형 변수를 표현하는 더미변수와 동일하다. 문장을 one-hot 벡터로 표현할 때는 문장 내에서의 각 단어의 출현 빈도를 표시하거나 혹은 출현 유무를 표시하는 이산형 변수가 된다.

된 문서의 대표 벡터($\in R^m$)를 축소된 차원($\in R^q$)의 은닉층(hidden layer)으로 사상시키는 역할을 한다.

- W_2 : ($K \times q$ 행렬)감성 정보 모형에서 학습되는 모수 행렬로, 은닉층으로 사상된 벡터를 학습의 목표가 되는 문서의 감성라벨에 해당하는 출력층[8]의 벡터($\in R^K$)로 사상시키는 역할을 한다.
- $freq(w_i^t)$: w_i^t의 전체 데이터에서의 단어 출현 빈도수

예를 들어 첫 번째 문서에 대해 다음과 같은 전처리가 완료되었다고 가정하자.

〈문서〉

금융위기 이후 대기업 중소기업 자금 조달 양극화 갈수록 심화 나타나다.

그리고 문서가 단 하나만 있고 현재 사용가능한 단어집합이 다음과 같다고 가정하자.

〈단어집합〉

C = {갈수록, 금융위기, 나타나다, 대기업, 심화, 양극화, 이후, 자금, 조달, 중소기업}

이 때 전체 단어집합은

C={갈수록, 금융위기, 나타나다, 대기업, 심화, 양극화, 이후, 자금, 조달, 중소기업}

로 주어진다.

문서의 번호 i는 1이고, T_1은 문서에 10개의 단어가 있으므로 10이다. 전처리된 문서에서 w_1^3은 '대기업'이고 w_1^7은 '양극화'이다. 문맥 단어의 개수를 4개로 정한다면, C_1^4는 4번째 단어 '중소기업' 주변 문맥 단어로 '이후, 대기업, 자금, 조달' 4개의 단어이다. 한편 전체 단어집합 C에 표현된 단어의 순서대로 단어 벡터를 표현한다고 하면, 첫 번째 문서의 세 번째 단어(w_1^3)인 '대기업'은 C집합에서 네 번째에 있으므로

8) K개의 감성라벨이 있다.

$$x_1^3 = (0,0,0,1,0,0,0,0,0,0) \in R^{10}$$

로 표현한다. 그리고 전체 문서에서 모든 단어는 한 번 출현하였으므로 $freq(w_i^t) = 1$이다.

다. 문서 대표 벡터

감성학습 모형은 문서, 혹은 문장을 대표하는 특성벡터를 이용한다. 문서가 주어질 때 대표 벡터는 다음과 같은 방식으로 계산한다. 우선 w_i^t 단어의 전체 감성 데이터 내의 빈도수($freq(w_i^t)$)를 이용해 확률 p_i^t(corrupt probability)를 계산한다. 이때 α는 감성모형의 복잡도 혹은 민감도와 관련이 있는 조절 계수이다.

$$p_i^t = 1 - \left(\sqrt{\frac{\alpha}{freq(w_i^t)}} + \frac{\alpha}{freq(w_i^t)} \right)$$

다음 베르누이 확률을 이용해 one-hot 벡터 x_i^t를 변환[9]시킨다.

$$\tilde{x}_i^t = \begin{cases} 0, \; with \;\; probability \; p_i^t > 0 \\ \dfrac{1}{1-p_i^t} x_i^t, \; otherwise \end{cases}$$

해당 문서에 포함된 모든 단어들의 변환된 one-hot 벡터를 모두 더하면 다음과 같은 i번째 문서의 대표 벡터를 구할 수 있다.

$$\tilde{x}_i = \sum_{t=1}^{T_i} \tilde{x}_i^t \in R^{|C|}.$$

여기서 구한 $\tilde{x}_i$는 랜덤벡터이며 모형 학습단계에 사용되는 최적화 방법인 Stochastic Gradient의 하나의 반복(epoch)마다 새로 생성한다. p_i^t 확률을 이용하여 단어들을 추출하는 것은 문서 혹은 문장의 대표 벡터를 생성할 때 단어 포함 유무를 조절함으로써 문장의 특

9) corruption이라고 쓰지만 정규화(regularization) 기능을 하는 변환으로 사용하였다.

징을 효과적으로 나타낼 수 있도록 하는 정규화방법이다.

구체적으로 앞에서 사용한 모수 α는 감성이 분류된 문서의 대표 one-hot 벡터를 계산할 때, 문서에 출현하는 각 단어들의 one-hot 벡터를 변환하는 과정에서 사용되는 모수이다. 문서에 포함되는 각 단어들의 대표 벡터에의 포함 확률을 조절하며, α의 값을 작게 설정할수록 문서에 포함되어 있는 각 단어가 대표 one-hot 벡터 계산에 포함될 수 있는 확률이 감소한다. 따라서 감성라벨 정보 학습에 사용되는 단어의 개수, 즉 정보의 개수가 감소하게 된다. 이는 단어의 특성벡터 학습 과정에서 감성에 관한 정보를 거의 고려하지 않게 되는 문제를 발생시킨다. 또한, α를 지나치게 작게 설정하게 되면 문서에 포함된 모든 단어들이 문서의 대표 one-hot 벡터의 계산에서 제외되는 경우가 발생할 수 있다. 이 경우에는 감성 정보 손실함수의 계산이 불가능해 오류가 발생하여 학습을 정상적으로 진행하지 못하게 된다. 따라서 α값의 적절한 선택은 감성 정보 손실함수의 계산에 있어 매우 중요하다. 본 연구에서는 α의 값으로 0.0075를 선택하였다.

라. 문맥 손실함수

문맥 손실함수에서 문서의 대표 벡터 $\tilde{x_i}$와 목표 단어의 주변 문맥 단어들인 C_i^t이 주어졌을 때, 목표 단어 w_i^t가 출현할 확률을 구하는 식은 다음과 같다.

$$P(w_i^t|C_i^t,\tilde{x_i}) = \frac{\exp(V_{w_i^t}^T U(C_i^t + \frac{1}{T_i}\tilde{x_i}))}{\sum_{w'}\exp(V_{w'}^T U(C_i^t + \frac{1}{T_i}\tilde{x_i}))}$$

위의 목표 단어의 출현 확률을 이용한 문맥 손실함수는 다음과 같다.

$$L_1(V,U) = -\sum_{i=1}^{n}\sum_{t=1}^{T_i} f(w_i^t, C_i^t, \tilde{x}_i)$$

단, $f(w_i^t, C_i^t, \tilde{x}_i) = \log(P(w_i^t|C_i^t, \tilde{x}_i))$ 이다. 본 연구에서는 목표 단어를 중심으로 단어 여덟 개를 문맥 단어집합으로 사용하였다.

마. 감성 정보 손실함수

문서의 대표 벡터 $\widetilde{x_i}$가 주어졌을 때, 문서의 감성라벨을 예측하는 확률 계산식은 다음과 같다.

$$f(\widetilde{x_i}) = softmax(W_2(W_1 U(\frac{1}{T_i}\widetilde{x_i}) + b_1) + b_2) \in R^K,$$

(단, $softmax$는 softmax 함수, $b_1 \in R^q$, $b_2 \in R^K$인 bias 벡터)

위의 감성 정보 학습에 이용되는 손실함수는 다음과 같다.

$$L_2(U, W_1, W_2, b_1, b_2) = -\sum_{i=1}^{n} 1^T (y_i \bullet \log(f(\widetilde{x_i}))) \quad ,$$

(단, $y_i \in \{0,1\}^K$는 실제 문서의 감성라벨을 표시하는 벡터이고 1은 모든 원소가 1인 벡터)

바. 모형 학습

모형 학습 단계에서는 앞에서 정의한 두 가지 손실함수의 가중 평균된 손실함수를 최적화하는 방향으로 학습을 진행한다. 조절 모수 $\beta \in (0,1)$를 도입하여 최종 손실함수인 $\beta L_1(V, U) + (1-\beta) L_2(U, W_1, W_2, b_1, b_2)$ 를 최소화하는 방향으로 학습한다.

$$\min_{V, U, W_1, W_2, b_1, b_2} \beta L_1(V, U) + (1-\beta) L_2(U, W_1, W_2, b_1, b_2)$$

위 식은 $\min_{V, W_1, W_2, b_1, b_2} \min_U \left(\beta L_1(V, U) + (1-\beta) L_2(U, W_1, W_2, b_1, b_2)\right)$와 같으므로 단어의 인코딩 함수의 학습시 특성벡터의 학습방향에 감성 손실함수 값이 영향을 주게 된다. 즉, 문맥 정보의 학습에 필요한 단어의 특성공간을 감성 정보가 조정해주는 역할을 한다.

구체적으로 조절 모수 β는 단어의 임베딩 결과에 대한 문법적 학습 모형의 반영 크기를 나타낸다. β의 값이 증가하면, 최종 손실함수의 가중치가 문맥 손실함수에 치우치게 된다. 따라서 문맥의 정보를 더 많이 활용하게 되고 감성에 관한 정보는 거의 사용하지 않게 된

다. 따라서 $\beta=1$인 경우는 사용된 모형이 오직 문법적 학습모형에 의존하게 되며, 일반적으로 널리 알려져 있는 WORD2VEC 모형(Mikolov, T. et al., 2013b)과 같다. β의 값을 0에 가깝게 감소시키는 경우에는 문맥의 정보를 거의 고려하지 않으므로 문서의 감성 정보만을 이용하여 임베딩을 학습하게 된다. 본 연구에서 참고한 Fu, P. et al.(2018)의 논문에서는 β가 0.3일 때 대부분의 실험에서 좋은 결과를 냈음을 보였고, 본 연구에서도 β의 값으로 0.3을 선택하였다.

모형학습은 심리 주체별로 감성 극성 유무에 따라 사용되는 감성 데이터를 구분하고, 품사별로 반응변수를 설정하여 기업-명사 모형, 소비자-동사 모형, 기타-형용사 모형 등 9개의 모형을 적합하여 9개의 감성사전 결과를 얻었다.

3. 감성사전의 구축

감성사전의 구축을 위해 우선 증감단어와 함께 사용될 수 있는 주제단어를 선정한 다음, 단어 특성벡터 간의 유사도 계산을 통해 각각의 주제단어와 관련이 높은 단어들을 추출하는 방식으로 감성사전을 구축하였다.

감성사전을 구축하기 위한 검색 주제단어는 김현중 외 3명(2019)의 연구 결과를 기반으로 선정하였다. 특히 본 연구에서 주목하고 있는 경제 단어와 증감단어의 관계를 감안하여 증감단어와 같이 사용 가능한 경제용어를 선택하였다. 또한 복합명사는 추가적으로 더 작은 의미를 갖는 단어들로 분해하여 같이 사용하여 총 49개의 단어를 사용하였으며, 이는 <부록 2>에 나타나 있다.

감성라벨이 붙어 있는 문서 데이터를 이용하여 모형을 적합한 후, 추정된 모형계수 U를 이용하여 감성사전을 구축한다. U는 주어진 단어를 m차원 단어특성 공간으로 표현해주는 선형 사상이다. 어떤 단어 w가 one-hot 벡터 x로 표현되어 있다고 가정하자. Ux는 R^m의 원소가 된다. 주어진 단어 $x_i(i=1,\cdots,n)$에 대해서 $y_i = Ux_i$를 계산한다. 만약 키워드가 될 중심어 x가 주어진 경우

$$I=\{i : d(y_i, Ux) \le \gamma\} \text{ (단, } d(x,y)\text{는 } R^m \times R^m \text{에서 정의된 거리함수)}$$

를 계산하고 중심어 x 주변에 있는 단어들 x_i $(i \in I)$을 이용하여 감성사전을 구축한다. 예를 들면 '경기'라는 단어 w를 전체 사전 C 위에서 one-hot 벡터 x로 표현한다. 다음 학습된 모형계수 U를 이용하여 $y=Ux$를 계산하면 '경기'라는 단어는 R^m공간 위 실수 y로

표현된다. 미리 계산한 $y_i = Ux_i$ 중 y와 거리가 가까운 원소를 찾고 그 원소에 대응되는 x_i들을 모으면 그것이 바로 '경기' 단어와 가까운 단어가 되며, 같은 방식으로 다른 키워드에 대해서 단어를 모아 감성사전을 구축한다. 여기서 주제단어와의 거리, 즉 유사도는 코사인 측도를 사용하였다. 코사인 측도는 단어특성 공간의 두 벡터 y와 $\tilde{y}$에 대해 $y^T\tilde{y}/(\|y\|_2\|\tilde{y}\|_2)$로 정의한다.

모형을 통해서 U를 학습할 때, CBOW 모형은 문법적 학습만을 하지만, 여기서 사용하는 모형은 감성라벨을 이용하기 때문에 U행렬이 감성적 라벨 분류를 잘 할 수 있도록 추정된다. 앞서 언급한 것과 같이 감성적 라벨 분류 정보의 반영 정도를 β가 결정한다.

Ⅲ. 모형 적합 방법 및 비교

본 연구에서는 전처리 방법, 학습 데이터의 단위를 변화시키면서 모형 적합을 시도하였고 이 장에서는 적합 방법에 따른 결과들을 비교하였다. 첫 번째로 전처리 방법에 따른 단어 추출 결과를 비교하였다. 두 번째로 서로 다른 데이터 단위의 감성 정보를 이용한 분류모형에 따른 감성단어 추출 결과를 비교하였다. 본 연구에서는 기사별 감성분류 라벨과 문장별 감성분류 라벨 데이터를 모두 분석하였으며 그에 따른 기사 감성분류모형과 문장 감성분류모형에 대한 감성단어 추출 결과를 비교하였다.

1. 전처리 방법의 비교

본 연구에서는 비지도 학습 기반의 단어 추출이 아닌 품사 기반 단어 추출 방식을 이용해 진행하였다. 품사 기반 단어 추출 방식을 사용할 경우의 장점들은 다음과 같다. 우선 실질적으로 감성이 포함되어 있지 않을 것으로 예상되는 품사들을 전처리 과정에서 미리 제거할 수 있어 학습에 필요하지 않은 데이터가 포함되는 것을 제한할 수 있다. 따라서 학습에 사용되는 전체 감성 문서 데이터의 크기가 감소하여 적은 메모리를 사용한 빠른 학습이 가능하다. 본 연구에서는 명사, 동사, 형용사 품사가 유의미한 감성 정보를 가지고 있을 것으로 가정하였다. 따라서 이 품사들에 해당하는 단어들만 선별적으로 선택하여 학습을 진행하였다.

추가적으로, Okt class를 사용함으로써 동일한 어근을 가지는 단어들을 동일한 단어로 처리하여 학습을 진행하였다. 따라서 동일한 의미의 단어들이 전체 단어 사전 내에서 중복되는 것을 방지하여 전체 학습 데이터의 단어 사전의 크기가 커지지 않게 할 수 있었다. 마지막으로, 경제용어 감성사전의 주제단어들에 대하여 품사별 결과 해석이 가능하였다.

반면 품사 정보를 고려하지 않는 비지도 학습 기반의 단어 추출 방식은 품사별 단어 추출 방법의 장점을 이용할 수 없었다. 우선 주어진 문서 데이터만을 이용해 계산한 단어 가능 점수만을 기반으로 단어를 추출하기 때문에, 품사에 대한 제한 없이 일정 기준치의 점수보다 높은 단어들을 모두 의미 있는 단어로 추출하게 된다. 따라서 실질적으로 감성을 갖지 않을 것으로 예상되는 품사의 단어들도 모두 학습에 포함된다. 또한, 동일한 어근을 가져 유사한 의미를 담고 있지만 다른 어미와 쓰여 형태가 다른 단어에 대해서 별개의 단

어로 인식하여 단어를 추출하는 경우가 흔히 발생하였다. 이 결과, 학습할 전체 감성 데이터의 크기가 매우 증가하여 학습 시간이 매우 오래 걸리고, 감성 분석에 유의미하지 않은 단어들이 결과에 포함되어 그 해석이 어렵다는 단점이 있었다. <표 4>를 보면, 앞에서 언급한 것처럼 '다녀야', '직에'와 같은 감성적 의미가 없는 단어들이 최종 결과에 포함되어 해석이 어려운 문제가 발생한다.

〈표 4〉 **전처리 방법에 따른 기업 측면의 감성사전 분석 결과 비교 예시**

주제단어: 취업 (긍정, 명사)	
비지도 학습 기반 분석	품사 기반 분석
구직	실적
청년	투자
뽑는	시장
채용	경기
다녀야	고용
견줘	연구원
근무	달러
앞지르	지표
직에	기업
청의	증시
실직	코스피
경력	이후
자치단	지수
나누기	증가
컸기	전망

2. 학습 단위의 비교

본 연구에서 사용한 감성 문서 데이터는 기사별 그리고 문장별 감성분류 데이터이고, 각각 학습을 진행하여 감성사전을 추출한 결과를 비교하면 다음 <표 5>, <표 6>과 같다. 기업측면에서 기사별로 감성분류 라벨을 가진 데이터를 이용하여 학습한 단어 특성벡터 공간에서 긍정 감성 데이터에 사용된 '신용'이라는 주제단어 주변에 있는 단어를 가까운 순

서대로 모았고, 이는 <표 5>에 정리되어있다. '신용' 단어 주변에 '대출', '잔액'과 같이 의미상 관계가 높은 단어들이 탐지되지만 해당 단어에서 감성적 특성을 찾아내기 힘들었다. 반면 문장별로 감성분류 라벨을 가진 데이터를 이용하였을 경우 '긍정', '전망', '상승'과 같이 경기 변동과 관련한 감성적 특성을 가진 단어가 다수 포함되어 있었다.

〈표 5〉 **기업 측면의 기사별, 문장별 감성분류 데이터를 사용한 경우의 명사 품사의 감성사전 결과 비교**

주제단어: 신용 (긍정, 명사)	
기사별	문장별
대출	수출
보증	기록
잔액	기업
융자	긍정
부실	전망
만기	실적
정이	생산
상호	증권
채권	상승

소비자 측면에서 부정적인 감성을 가진 '인력' 단어는 기사별로 감성라벨이 분류된 자료와 문장별로 감성라벨이 분류된 자료에서 모두 문맥적, 감성적 관련성이 높은 단어가 탐지되었다. 하지만 관련 단어의 순서를 보면 문장별로 감성분류가 된 자료를 이용하여 학습한 경우 부정적 감성을 가진 '감축', '해고'와 같은 단어가 '인력'이라는 단어에 더 가까이 있음을 확인할 수 있었다. 다른 몇 가지 예시를 더 살펴본 결과 문장별로 정리된 감성라벨을 이용하여 단어의 특성벡터를 추출하고 감성사전을 구축하는 것이 더 좋을 것으로 생각된다. 이러한 결과는 다음의 <표 6>에 정리되어 있다.

〈표 6〉 소비자 측면의 기사별, 문장별 감성분류 데이터를 사용한 경우의 명사 품사의 감성사전 결과 비교

주제단어: 인력 (부정, 명사)	
기사별	문장별
점검	감축
효과	인원
기업	해고
구조조정	명의
원인	감원
계획	통폐합
전체	대규모
해고	여명
우선	직원

Ⅳ. 분석결과

1. 심리 주체별 감성단어 사전의 구축

문장 단위 감성분류 데이터로 학습한 단어 특성벡터를 이용하여 소비자, 기업, 기타 심리 주체별로 경제관련 주제단어와 유사도가 높은 감성단어 사전을 구축하였다. 각 주제단어에 대해 전체 단어집합의 단어 중 주제단어와 유사도가 높은 상위 20개의 단어를 소비자·기업·기타 심리 주체별, 긍정·부정·중립 감성별, 명사·동사·형용사 품사별로 각각 도출하였다. 다음의 표들은 각 주제단어와 유사도가 높게 나온 상위 15개 단어로 구성된 경제용어 감성사전의 예시들이다.

〈표 7〉 **소비자 측면 감성사전 결과 예시**

명사	동사	형용사
주제단어: 채용(긍정)	주제단어: 인력(부정)	주제단어: 경제(부정)
인턴	비다	심각하다
인원	줄어들다	불확실하다
공공기관	미루다	강하다
신입	옮기다	느리다
계획	시달리다	꾸준하다
상반기	하다	급속하다
규모	적다	아니다
대졸	밝히다	급격하다
공기업	받아들이다	성공하다
올해	따르다	기대하다
그룹	대다	힘들다
자의	잊다	낮다
정규직	머무르다	화하다
인력	오다	이다
모든	나타나다	미치다

소비자 심리 측면에서 분류된 감성 극성 정보에서의 감성사전의 예시는 <표 7>에 정리되어 있다. <표 7>을 보면 긍정적 문장에서 나온 '채용'은 명사로는 '인턴, 공공기관, 신입, 상반기'와 같은 단어들이 긍정 감성으로 가장 관련이 높게 분류되었다. 부정적 문장에서 나온 '인력'과 가까운 부정 감성을 가지는 동사로 '비다, 줄어들다, 적다'와 같은 단어들이 분류되었다. 그리고 부정 감성의 문장에서 등장한 '경제'는 가장 가까운 부정 형용사로 '심각하다, 불확실하다, 느리다'와 같은 단어들이 분류되었다.

〈표 8〉 **기업 측면 감성사전 결과 예시**

명사	동사	형용사
주제단어: 성장(긍정)	주제단어: 구매(부정)	주제단어: 수출(긍정)
경기	줄어들다	활발하다
달러	줄다	꾸준하다
증가	떨어지다	같다
시장	잦다	빠르다
증시	부딪히다	성장하다
지수	나타나다	톡톡하다
기업	치다	자리다
상승	그치다	있다
중국	치르다	유리하다
투자	쪼그라들다	멀다
회복	막다	좋다
미국	시달리다	빨르다
개선	늘다	회복하다
경제	추다	상당하다
기록	여기다	튼튼하다

기업 심리 측면에서 분류된 감성 극성 정보에서의 감성사전의 예시는 <표 8>에 정리되어 있다. <표 8>을 보면 긍정적 문장에서 나온 '성장'은 '시장, 증시, 기업' 등이 유사한 감성 극성 명사로 분류되었다. 부정적 문장에서 등장한 '구매'는 '줄어들다, 줄다, 떨어지다'와 같은 동사들이 관련된 부정 감성을 가지는 동사로 분류되었다. 그리고 긍정적 감성 문장의 '수출'이라는 단어에는 '활발하다, 꾸준하다, 성장하다'와 같은 단어가 관련이 높은 긍정 감성 형용사로 분류되었다.

〈표 9〉 **기타 측면 감성사전 결과 예시**

명사	동사	형용사
주제단어: 경기(긍정)	주제단어: 국가채무(부정)	주제단어: 주가지수(긍정)
회복	늘어나다	높다
둔화	없애다	충분하다
침체	늘다	가난하다
확장	줄어들다	다르다
실물	가다	장하다
지표	돌다	번영하다
대감	빼다	원활하다
국면	가리다	필요하다
위축	지나치다	마땅하다
부양책	가르다	만족하다
재정정책	드리워지다	흡족하다
주효	넘어서다	냉정하다
시그널	드러나다	약하다
두려움	들다	반하다
기술	기울이다	대담하다

기타 심리 측면에서 분류된 감성 극성 정보에서의 감성사전의 예시는 <표 9>에 정리되어 있다. <표 9>를 보면 긍정적 문장에서 나온 '경기'와 '회복, 둔화, 확장'이 관련이 높은 단어로 분류되었다. 부정 감성의 '국가채무'라는 단어와는 '늘어나다, 없애다, 늘다'와 같은 동사가 가장 유사한 동사로 분류되었다. 그리고 긍정적 감성 문장에서 나온 '주가지수'와 관련이 높은 형용사로는 '높다, 충분하다, 번영하다'와 같은 형용사가 분류되었다. 또한, '기타' 심리 주체는 소비자와 기업의 감성적 특성이 잘 구별되지 않고 중립적이며, 긍정과 부정 감성단어가 혼재되어 있는 결과를 보여주는 경향이 있다.

2. 기업과 소비자 간 감성단어 비교

아래 <표 10>~<표 12>에서는 같은 주제단어에 대해 소비자 심리 측면과 기업 심리 측면에서 가깝게 분류된 상위 20개 단어들을 비교하였다.

〈표 10〉 **명사 품사의 소비자와 기업 측면의 감성사전 결과 비교 예시**

주제단어: 인력(긍정, 명사)	
소비자	기업
채용	**수출**
규모	**심리**
확대	**재정**
공기업	**강세**
사정	**조정**
감축	전망
청년실업	가운데
신세계	이상
원금	우리나라
유망	제조업
한도	포인트
퇴직	지속
인턴	사상
교사	증가
서민	개선
공채	투자
다자	일본
세제	상승
정규직	기업
전체	이익

소비자 측면에서 긍정이라고 표시된 감성라벨을 이용하여 분석한 결과, 긍정적 감성을 가진 단어 '인력'은 단어의 특성벡터가 가까운 명사로서 '채용', '규모', '확대', '공기업' 등 취업과 관련이 있는 단어들이 탐지되었다. 반면 기업 측면에서 표시된 감성라벨을 이용하면,

인력과 가까운 특성벡터로 '수출', '심리', '재정', '강세' 등 기업활동과 관계가 있는 단어들이 탐지되었다.

〈표 11〉 **동사 품사의 소비자와 기업 측면의 감성사전 결과 비교 예시**

주제단어: 소비(긍정, 명사)	
소비자	기업
늘어나다	**늘어나다**
크다	기다리다
보이다	되어다
않다	벗어나다
받다	살아나다
부추기다	막다
이끌다	서다
비다	나타나다
따르다	**보이다**
높이다	하다
해오다	받다
감다	띄다
뽑다	바뀌다
넘어서다	찾다
하다	돼다
벌이다	걸다
되다	**크다**
입다	치르다
보다	즐기다
내리다	걸리다

같은 방법으로 긍정 감성 문장의 단어 '소비'와 가까운 동사를 소비자 측면과 기업 측면에서 비교해 보았다. 두 분야에서 모두 동사 '늘어나다'가 '소비'와 가까운 단어로 선택되었다. 하지만 상위 5개 단어 중 기업과 소비자 측면에서 동사는 하나를 제외하고 일치하지 않았다. 또한 '크다', '보이다' 등은 공통적으로 선택되었지만 '소비'와의 유사도 순위가 서로 다른 것을 볼 수 있다.

〈표 12〉 **형용사 품사의 소비자와 기업 측면의 감성사전 결과 비교 예시**

주제단어: 규제(부정, 명사)	
소비자	기업
면밀하다	불리하다
상당하다	미치다
강하다	**강화하다**
불편하다	새삼스럽다
새롭다	**불편하다**
빨르다	두드러지다
있다	별다르다
대단하다	높다
불확실하다	불가피하다
꾸준하다	빨르다
정교하다	있다
확실하다	마땅하다
강화하다	활발하다
많다	스럽다
선량하다	**강하다**
화하다	단단하다
실망하다	심각하다
불안정하다	없다
파산하다	느리다
유일하다	다르다

단어 '규제'에 대해서는 관련된 형용사 역시 기업과 소비자가 다른 양상을 보였다. 유사도가 높은 상위 5개 단어 중 '불편하다'만 일치하였고 '강하다', '강화하다' 등 공통적으로 선택된 단어도 유사도에 차이가 있는 것을 볼 수 있다.

이와 같은 예는 감성적 특성이 심리 주체에 대해서 따로 표시된 데이터를 활용하였을 경우 감성사전이 다르게 만들어질 수 있음을 보여준다. 다시 말해 감성사전의 활용 대상이 소비자심리지수인지 혹은 기업심리지수인지에 따라 다른 감성사전을 적용할 수 있음을 보여준다.

3. 시사점

본 연구에서는 같은 의미를 갖는 단어라도 감성별로 구별하여 사용하였다. 이는 같은 단어라도 긍정 문서에 쓰이는 경우와 부정 문서에서 쓰이는 경우 각각 실질적으로 내포하는 감성적 의미가 다를 것이고, 따라서 그에 따라 계산된 특성벡터가 달라야 할 것이기 때문에 이와 같은 연구방법을 선택하였다.

특히 경제기사의 증감단어에 대해서 이러한 경향은 더 뚜렷하게 나타난다. 예를 들어 '취업률 증가'라는 주제의 문서가 있다고 가정하면 이 문서는 긍정 감성을 지니고 있을 가능성이 크다. 또한 '실업률 증가'라는 주제의 문서는 부정 감성을 지니고 있을 것이라 생각해볼 수 있다. 두 문서에서 동일하게 등장한 '증가'라는 단어는 문서가 대표하는 감성 극성이 달리 보는 것이 합리적이며, 실질적으로 각 단어가 내포하는 감성적 의미는 긍정과 부정으로 구분하는 것이 효과적 단어분석을 위해 바람직할 것이다. 각 문서의 '증가' 단어에 해당하는 특성벡터는 별도로 학습되어야 할 것이고, 본 연구에서는 이에 따라 추출된 단어를 구분하여 학습하였다.

감성사전 분석과 같은 텍스트 마이닝에서는 좋은 결과를 위해 많은 양의 데이터를 필요로 한다는 점은 모형의 적합에 있어 난점이다. 왜냐하면 데이터의 양을 늘릴수록 그만큼 학습의 속도는 현저히 느려지기 때문이다. 이를 해결하는 방법으로 문맥 정보 모형의 학습에서 세부 목표 단어를 품사별로 나누어 진행하는 것을 고려해볼 수 있다. 이 방법은 사용되는 학습 데이터의 양이 감소하므로 보다 빠른 학습과 결과 확인이 가능하고, 데이터 도메인을 동질적 대상으로 한정시킨다는 점에서 모형학습에 유리한 측면이 있다. 형용사와 동사 같은 경우에는 전체 감성 문서 데이터에서 그 빈도가 높지 않아 계산상 수월한 학습이 가능하다. 하지만 품사의 제한은 그로 인한 분석모형의 편이를 발생시킬 수 있어, 계산의 속도와 모형의 확장성 사이에서 적절한 의사결정이 필요하다.

Ⅴ. 결론

본 연구에서는 경제기사에 표시된 감성 정보를 이용하여 단어의 감성적 특징 추출을 시도하였다. 우리가 사용한 단어 특징 추출 모형은 감성 정보에 대한 반영 크기에 따라 다른 형태의 단어 임베딩 결과를 주는 것을 확인하였으며, 이를 통해 주어진 키워드에 대한 감성적 특성을 가진 단어들을 단어 특징 공간 위에서 서로 가깝게 만들어 주는 것도 관찰하였다.

또한 단순히 주어진 모형을 적합만 한 것이 아니라, 모형적합을 위해 필요한 사전 작업과 여러 가지 모형 계수 선택에 따른 결과들을 비교, 검토하였다. 특별히 여기서는 단어를 추출하기 위한 두 가지 다른 방법의 전처리 과정을 검토하였고, 문장 및 기사 단위의 두 개의 다른 방법으로 만들어진 감성 데이터를 이용하여 그 성능을 비교하였다. 이번 연구에서는 감성사전을 구축함에 있어 단어를 품사 기반으로 분석하여 처리하는 것이 감성단어의 특징을 추출하고 결과를 해석하는 데 효과적임을 보였다. 특별히 품사 특성을 이용한 감성단어 분석에서 명사에도 감성적 특성이 있음을 데이터 결과를 통해 확인할 수 있었다. 세부적으로는 기사별 감성 분석보다는 문장별 감성 분석이 단어의 특징을 추출하는 데 효과적이었고, 감성 분석의 대상을 경제주체별로 나누어 따로 분석하였을 때 감성사전의 결과가 다르게 나올 수 있음을 확인하였다.

감성적 특성이 문서의 관찰 시점에 따라 달라진다는 것을 고려한다면, 본 연구에서 제시한 감성단어의 특징 추출방법을 기간별로 나누어진 데이터에 따로 적용하는 것이 도움이 될 것으로 보인다. 다만, 제안한 모형에는 감성라벨이 반드시 필요하므로 문서/문장의 감성라벨 예측 모형을 통해 감성라벨을 추정한 데이터를 생성한 후 단어의 특성벡터를 추출하는 것이 가능할 것으로 보인다.

부 록

1. 비지도 학습 기반 단어 가능 점수 : Cohesion score와 Branching entropy

- Cohesion score

$$Cohesion(w_{0:n}) = \left(\prod_{i=0}^{n-1} P(w_{0:i+1} | w_{0:i}) \right)^{\frac{1}{n-1}}$$

이 때 $w_{0:n}$는 음절이 n개인 단어를 의미한다. Cohesion score는 단어를 구성하는 글자들만의 정보를 이용하여 단어를 추출하기 위해 고안된 단어 점수 계산 방식이다. 특정한 음절 시퀀스(sequence)가 나올 확률은 어절들의 열이 마코프 체인(Markov chain) 가정을 만족하는 경우 Cohesion score로 주어진다.

음절 시퀀스가 주어졌을 때, 해당 음절 시퀀스 내부의 음절 시퀀스 각각에 대한 전체 문서 내의 출현 빈도를 이용해 확률을 계산한다. 이때 주의할 점은 내부의 음절 시퀀스를 정의할 때 전체 음절 시퀀스($w_{0:n}$)의 첫 번째 음절로 시작하지 않은 것은 무시해도 된다는 것이다. 이는 의미가 있는 단어는 반드시 어절의 첫 번째 음절부터 시작되는 한글의 구조적 특성 때문이다.

유의미한 단어의 경계에 가까워지면, 즉 단어의 경계의 직전에는 다음 음절을 예측하기 쉽다. 예를 들어 '컴퓨'라는 음절 시퀀스 다음에는 '터'라는 음절이 등장할 것이라고 쉽게 예측할 수 있다. 하지만 단어의 경계를 넘어서게 되면 이전까지 등장한 내부의 음절 시퀀스가 명확한 문맥적 의미를 주지 못한다. 예를 들어 '컴퓨터'라는 음절 시퀀스는 '컴퓨터는'이라는 음절 시퀀스의 문맥적 의미를 주지 못한다. 마지막으로 짧은 음절 개수의 단어만을 선호하는 결과를 방지하기 위해 기하평균으로 계산한다.

- Branching entropy

$$Entropy(c) = -\sum_{w' \in V} P(w'|c)\log P(w'|c)$$

이 때 w는 주어진 문서 데이터의 전체 단어 사전에 속하는 단어, c는 1개 이상의 음절로 이루어진 단어를 의미한다. (단 V는 문서 안에서의 음절집합이다.)

Branching entropy는 Jin and Tanaka-Ishii(2006)에서 제안된 방법으로, 단어의 앞, 뒤 경계에 출현하는 글자의 정보를 이용하여 단어의 경계를 찾아 이를 기준으로 단어를 추출하기 위해 고안된 단어 점수 계산 방식이다. Cohesion score와 다르게 단어 내부의 글자가 아닌, 단어를 구성하지 않는 외부 글자들의 정보를 이용한다. 주로 띄어쓰기가 없는 언어(중국어, 일본어 등)의 단어 세분화(word segmentation)에 많이 사용된다.

유의미한 단어의 경계에서는 다양한 단어가 등장할 수 있으므로, 그 경계의 주변 단어에 대해서는 불확실성, 즉 엔트로피가 증가하게 된다. 그리고 이 엔트로피를 글자 빈도의 분포를 이용해 계산한다. 예를 들어 어떤 음절 시퀀스의 경계에 등장할 음절이 명확하다면 불확실성이 낮아져 엔트로피가 작다. 하지만 명확하지 않다면 불확실성이 높아져 그 엔트로피가 높다. 따라서 엔트로피가 높은 부분은 단어의 경계일 가능성이 높으므로 이를 기준으로 유의미한 단어를 추출한다.

2. 감성사전의 주제단어

다음은 본 연구에서 사용한 주제단어들이다.

가계	구매	부채	소비	실업률	주가	취업
가계부채	구매력	비정규직	수출	실업자	주가지수	취업률
격차	국가부채	빈곤	시가	연체	증시	코스닥
경기	국가채무	빈곤층	시가총액	원자재	지수	코스피
경제	규제	생산	신용	유가	채무	코스피지수
경제성장	대출	성장	신용등급	인력	채용	퇴직
고용	변동성	소득	실업	일자리	출생	투자

참고문헌

· 김현중·임종호·이혜영·이상호(2019), "온라인 뉴스 기사를 활용한 경제심리보조지수 개발", 한국은행 국민계정리뷰 2019년 제2호, pp.1～33.

· Fu, P., Lin, Z., Yuan, F., Wang, W., & Meng, D. (2018, April). Learning Sentiment-Specific Word Embedding via Global Sentiment Representation. In Thirty-Second AAAI Conference on Artificial Intelligence.

· Jin, Z., & Tanaka-Ishii, K. (2006, July). Unsupervised segmentation of Chinese text by use of branching entropy. In Proceedings of the COLING/ACL on Main conference poster sessions . Association for Computational Linguistics. (pp. 428-435).

· Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

· Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).