텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류*

본 연구에서는 다양한 주제를 가진 텍스트 문서 중 경제심리와 관련된 문서를 분류하는 방법에 대해 살펴본다. 텍스트 문서는 수치형 자료와 달리 비정형적(unstructured)인 특성을 가지고 있으므로 데이터 마이닝 방법론을 적용하기 위해서는 이들 자료를 벡터, 행렬 등의 수치형 자료로 변환하는 기법이 필요하다. 이러한 변환 방법 중 「Bag-of-words」 방식과 word embedding 방식의 하나인 「word2vec」 기법을 이용하여 분석 비교하였다. 이들 방법을 이용하여 실제 언론기사 자료**를 벡터로 표현하고 로지스틱 회귀모형을 적용하여 경제심리 관련 문서 분류 성능을 평가해 보았다. 분석결과 「Bag-of-words」 방법에 비해 word embedding 방식을 사용하는 「word2vec」 기법이 전반적으로 경제심리 관련 문서 식별 성능이 우수한 것으로 나타났다.

* 본고는 서울대학교 통계학과 원중호 교수, 이한별 학생연구원, 한국은행 문혜정, 손원 과장이 작성한 것으로, 본고의 내용은 집필자의 개인의견으로서 한국은행의 공식견해를 나타내는 것은 아님.
** 본 연구에 사용된 언론기사 자료는 한국정보화진흥원(NIA)으로부터 제공받았음.

Ⅰ. 서 론

정보통신기술의 발달로 인해 상거래, 소셜미디어(social media), 인터넷 블로그, 사물인터넷(internet of things) 등을 통해 생성된 수많은 빅데이터(big data)가 전자정보 형태로 실시간으로 기록·축적되고 있다. 이 중 텍스트 데이터(text data), 사진, 동영상 등과 같은 비정형적(unstructured) 데이터가 빅데이터의 상당수를 차지하고 있는데 특히 텍스트 데이터에 포함된 정보를 분석· 활용하기 위한 시도가 다양한 분야에서 활발히 이루어지고 있다. 이처럼 텍스트 데이터로부터 유용한 정보를 추출해내는 과정을 텍스트 마이닝(text mining)이라 한다. 텍스트 마이닝 기법은 최근 다양한 분석 방법론의 개발, 정보처리 속도의 향상, 저장장치의 발달 등을 기반으로 급속히 발전하고 있으며 이에 대한 관심도 높아지고 있다.

텍스트 데이터로부터 의미 있는 정보를 찾아낼 수 있을 것이라는 기대는 "통계적 의미론 가설(statistical semantics hypothesis)"에 근거한다. 텍스트 마이닝의 기틀이 되는 이 가설은 "사람들의 글, 말 등에서 드러나는 단어 사용의 통계적 규칙성으로부터 사람들이 말하고자 하는 바를 찾아낼 수 있다"(Turney and Pantel, 2010)는 전제를 바탕으로 한다.

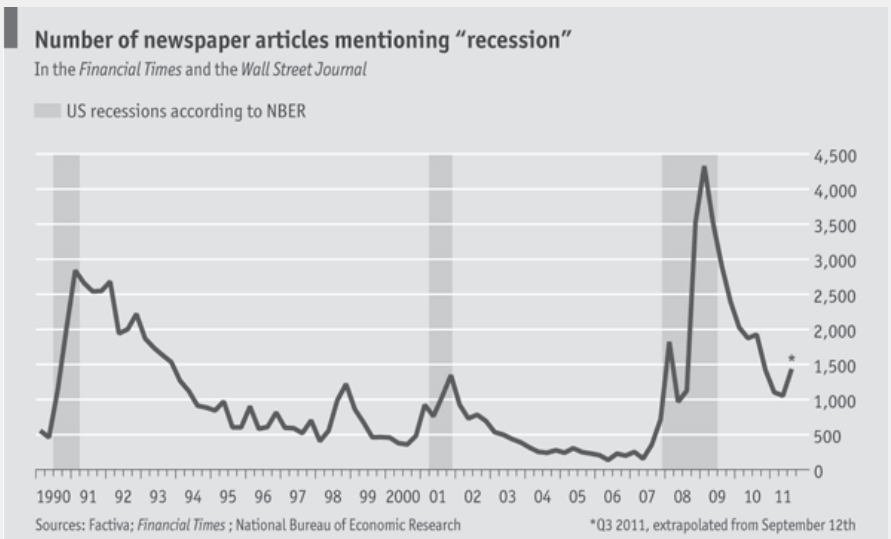
하지만 텍스트 데이터는 벡터, 행렬 등의 수치형 자료와 같은 정형적인 형태를 가지고 있지 않아 기존 수치형 데이터(numerical data)에 비해 분석이 어렵다. 또한 텍스트 데이터에서는 다양한 단어들이 복잡하고 미묘한 문법적 구조 안에서 표현되고, 동일한 문자로 기록된 단어도 표현 방식에 따라 다른 의미를 지닌다. 이러한 문법적인 구조와 의미의 다양성을 모두 수치형 자료로 표현하기 위해서는 매우 큰 차원의 벡터 또는 행렬이 필요하고, 이는 비효율적인 저장공간 활용, 계산시간의 기하급수적 증가 등의 문제를 야기하게 된다. 따라서 의미의 손실이 다소 발생하더라도 일정한 가정 아래에서 텍스트 데이터를 가능한 간략한 수치형 자료로 표현하는 방법을 선택하는 것이 일반적이다.

한편 텍스트 마이닝은 다음과 같은 세부 분야로 구분할 수 있다. 주제에 따라 문서를 나누는 문서 분류(document classification)와 군집화(clustering), 원하는 정보를 가진 문서를 찾는 정보 검색(information retrieval), 문서로부터 필요한 정보를 찾는 정보 추출(information extraction) 등이 있으며 이외에도 작성자의 심리, 감정 등을 파악하는 감성 분석(sentiment analysis)도 많이 활용되고 있다(Weiss et al., 2010).

경제·금융 분야에서 텍스트 마이닝 기법을 활용한 가장 단순한 형태 중 하나는 Economist (2011.9.16일)에 소개된 R-word 인덱스이다. R-word 인덱스는 『파이낸셜 타임즈』와 『월스트리트 저널』에 게재된 기사 중 경기침체를 뜻하는 "recession"이라는 단어를 포함한

기사의 수를 합산하여 지수화한 것으로 NBER(national bureau of economic research)에서 공표하고 있는 미국의 공식 경기 주기를 잘 반영하는 것으로 알려져 있다(<그림 1-1> 참조).

〈그림 1-1〉 Economist에 소개된 R-word 인덱스



최근에는 텍스트 데이터 분석기법이 더욱 정교해지고 있으며 텍스트 데이터의 활용을 위한 시도가 다양한 분야로 확대되고 있다. 거시경제 분야에서는 Baker et al.(2016)이 "uncertainty" 또는 "uncertain" 등의 단어가 포함된 경제정책 관련 기사 자료를 분석하여 경제정책 불확실성(Economic Policy Uncertainty) 지표를 제시한 바 있다. 그리고 Scott and Varian(2014)은 구글(Google) 검색 자료를 이용하여 물가와 실업률 등의 초단기 변동을 예측하여 저빈도 자료(low-frequency data)를 고빈도 자료(high-frequency data)로 전환하는 방안을 제안하였다. 또한 금융 분야에서도 Gentzkow et al.(2017)이 언론기사, 소셜미디어, 기업의 공시자료 등을 통해 수집한 정보를 자산가격 예측에 활용한 연구 결과를 발표하기도 하였다.

이러한 연구들은 대부분 특정 키워드에 의존하는 방식을 택하고 있다. 그러나 키워드 방식은 경우에 따라서는 정확성 등에 한계가 발생할 수 있다는 주장이 제기되고 있다. Butler(2013)는 독감(flu)에 대한 구글 검색빈도로 독감 발병률을 예측하는 모형이 일정 시점 이후에는 상당한 과추정(overestimation) 현상을 보인다는 주장을 제기한 바 있다.

특히 문서의 분류 문제에 있어서는 문서의 주제를 특정 키워드의 포함 여부에만 의존할 경우 부정확한 결과로 이어질 수 있다. 아직까지 동의어와 유의어, 동음이의어 등을 완벽하게 처리하는 것은 쉽지 않은 상황이기 때문이다. 예를 들어 '경기'라는 단어는 경제 상

황이라는 의미로 사용되기도 하지만 운동시합, 지역명 등으로 사용되기도 하므로 '경기'라는 단어의 포함 여부를 기반으로 한다면 정확한 문서 분류가 이루어지지 못할 가능성이 있다. 또한 세월이 흐름에 따라 단어의 의미가 변화(semantic change)하기도 하고 평소에 잘 사용되지 않는 특정 단어가 일시적으로 널리 사용되는 경우도 있다. 이러한 점들을 고려하여 본고에서는 특정 키워드에 의존하지 않고 전체 단어를 활용한 문서 분류 기법을 제안하고자 한다.

본고의 구성은 다음과 같다. 먼저 Ⅱ장에서는 문자열로 이루어진 비정형적인 텍스트 데이터를 벡터, 행렬 등의 정형적인 수치형 자료로 표현하는 방법에 대해 소개하고 각 방법의 특성을 살펴본다. Ⅲ장에서는 Ⅱ장에서 소개한 방법으로 표현된 수치형 자료를 이용하여 경제 상황에 대한 판단, 즉 경제심리를 내포하고 있는 문서(언론기사)를 분류하고 그 결과를 비교해 본다. 마지막으로 Ⅳ장에서는 본 연구의 결과를 정리하고 텍스트 데이터를 이용한 경제심리 분석을 위한 앞으로의 연구 방향을 간단히 언급한다.

Ⅱ. 텍스트 데이터의 수치화

다량의 문서를 분석하여 유의미한 정보를 찾아내기 위해서는 이들 텍스트 데이터를 전산기기를 이용하여 분석하기 쉬운 형태로 바꾸어 주는 과정이 필수적이다. 텍스트 데이터를 벡터, 행렬 등의 형태로 표현(representation)하면 이미 잘 구축되어 있는 데이터 마이닝 기법들을 활용하여 보다 쉽게 결과를 도출할 수 있는 장점이 있다. 따라서 이 장에서는 텍스트 데이터를 수치형 자료, 특히 벡터, 행렬 등의 형태로 표현하는 대표적인 방법들을 살펴보고 이 방법들을 평가해 본다.

1. 전처리(pre-processing) 과정

텍스트 데이터를 수치형 자료로 표현하기 위해서는 대부분 전처리 과정을 거치게 된다. 많은 경우 본격적인 데이터 마이닝 과정보다 전처리 과정에 더 많은 시간이 소요되는 것으로 알려져 있으며 텍스트 마이닝 결과를 결정하는 중요한 요인이 되기도 하는 것으로 평가되고 있다. 텍스트 마이닝의 목적이나 방법에 따라 다양한 전처리 방법이 있지만 본고에서는 전처리 과정 중 중요성이 높은 토큰 분리, 정규화 및 품사 태깅 과정에 대하여 간략히 소개하고자 한다.

가. 토큰 분리 (tokenization)

앞서 언급한 바와 같이 텍스트 마이닝은 단어의 사용 패턴을 통해 글쓴이가 전달하고자 하는 의미를 파악하는 과정이라 할 수 있다. 여기에서 '단어'의 개념을 보다 엄밀히 정의할 필요가 있다. 대체적으로 단어가 의미를 전달하는 가장 중요한 요소이지만 한글의 경우 때로는 단어보다 더 잘게 쪼개진 형태소(morpheme)나 두 단어 이상의 단어로 이루어진 구절이 의미를 표현하는 데에 더 적합한 요소로 사용되기도 한다. 이렇게 의미를 나타내는 데 적합한 가장 기본적인 단위를 텍스트 마이닝에서는 '토큰(token)'이라 한다. 형태소나 단어가 토큰이 되기도 하지만 때로는 둘 이상의 단어가 토큰으로 사용될 수도 있다는 점에서 토큰, 형태소, 단어는 서로 구별된다. 예를 들어 "New York"의 경우 두 단어로 이루어져 있지만 각각의 단어 "New"와 "York"가 단순히 결합된 것과는 전혀 다른 의미를 지니고 있으므로 정확한 의미 파악을 위해서는 하나의 독립적인 토큰으로 사용하는 것이 바람직하다.

텍스트 마이닝을 위해서는 연속된 문자열로 표현된 문서를 의미 표현의 기본 단위인 토큰으로 나누어주는 작업이 필요한데 이러한 작업을 '토큰 분리' 또는 '토큰화(tokenization)'라고 한다. 한편 한글의 경우 영어에 비해 토큰을 분리하는 작업이 더 까다로운 것으로 알려져 있다. 영어의 경우 "doesn't" 등과 같은 축약된 표현, "New York" 등과 같이 둘 이상의 단어가 새로운 의미를 만드는 경우 등 몇몇 특수한 경우를 제외하면 공백을 기준으로 토큰들이 구분되므로 토큰화 과정이 비교적 용이하다. 반면 한글의 경우 조사가 공백 없이 결합되어 있고 한 글자 안에서도 어근과 어미의 일부가 결합된 형태로 나타나기도 하므로 어미와 조사를 구분하는 과정이 쉽지 않다. 또한 여러 단어로 이루어진 합성어의 경우에도 사람에 따라 띄어쓰기를 하기도 하고 띄어쓰기 없이 표현하기도 하므로 영어에 비해 정확한 토큰 분리 작업에 많은 어려움이 따른다.

나. 정규화 (normalization)

토큰은 문장 속에서 여러 형태로 다양하게 변형되어 표현되기도 한다. 대표적인 예가 영어 동사의 시제, 성, 수 등에 따른 변화이다. 정규화는 이렇게 변형되어 나타난 토큰들을 하나의 기본 형태로 묶어주는 과정이다. 정규화를 통해 변형된 표현 형태들을 하나의 토큰으로 나타냄으로써 변수의 개수를 줄이고 텍스트 마이닝 작업의 효율을 높일 수 있다. 하지만, 한편으로는 일정 부분 의미의 손실을 감수해야 하는 경우도 있다. 예를 들어 '한다'와 '하였다'는 정규화를 통해 '하다'로 표현될 수 있지만 '하다'로 이 두 단어를 정규화하여 표현하는 경우 각 단어가 가지고 있던 현재와 과거의 의미는 사라지게 된다.

영어, 독일어 등의 경우 성, 수, 시제 등에 따라 명사와 동사의 변화가 심하고 문장 첫머리에 나오는 글자는 대문자로 표현되는 등 정규화 작업에 고려할 요소가 많다. 반면, 한글의 경우 상대적으로 성, 수 등에 따른 변화는 크지 않기 때문에 정규화는 상대적으로 용이한 편이라 할 수 있다.

다. 품사 태깅 (part-of-speech tagging)

단어의 의미는 문장 구조 또는 다른 단어들과의 관계를 통해 드러나기도 한다. 같은 문자로 표현된 단어도 문맥에 따라 전혀 다른 의미로 해석될 수 있다. 예를 들어 영어 단어 "saw"는 동사로 사용되었을 때는 "보았다"라는 뜻이 되지만 명사로 사용되었을 때에는 "톱"을 뜻한다. 우리말에서도 "다"는 부사로 사용되었을 때에는 "모두"라는 뜻으로 쓰이지만 어미로 사용되거나 조사로 쓰이기도 하므로 의미를 정확히 나타내기 위해서는 품사의

구분이 중요하다. 이와 같이 단어의 의미를 정확히 포착하기 위해서는 문장 안에서 단어가 어떤 품사로 사용되고 있는지 파악하는 작업이 필요한 경우도 있는데 이러한 작업을 품사 태깅 과정에서 수행하게 된다.

라. 전처리 과정의 예시

지금까지 소개한 텍스트 데이터의 전처리 과정을 예를 통해 살펴본다. 전처리를 위해서는 Python 패키지인 KoNLPy를 사용하였으며 KoNLPy에 내장된 형태소 분석기 중 트위터에서 개발한 한글 형태소 분석기를 활용하였다.

아래는 2017년 1월 13일 발표된 한국은행 통화정책방향 결정문에서 발췌한 문장이다.

"금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를 현 수준(1.25%)에서 유지하여 통화정책을 운용하기로 하였다."

이 문장에 대한 토큰 분리, 정규화, 품사 태깅 과정은 다음과 같이 진행된다. 먼저 토큰 분리 과정을 거쳐 이 문장에서 토큰들을 추출해 낸 결과는 아래와 같다. 조사와 어미가 결합되어 있어 까다로운 측면이 있음에도 불구하고 아래와 같이 토큰 분리 과정이 비교적 잘 수행되었음을 확인할 수 있다.

"금융통화위원회/는 다음 통화정책/방향 결정/시/까지 한국은행 기준금리/를 현 수준/ (/1/./25/%) /에서 유지/하여 통화정책/을 운용/하기로 하였다/."

다음으로 변형되어 나타난 토큰들을 정규화를 통해 하나의 표준적인 토큰으로 재분류하는 과정을 살펴본다. 위의 문장에서는 "하여", "하기로" 및 "하였다"라는 토큰들이 나타나는데 이 토큰들은 모두 "하다"라는 기본형에서 파생되어 나온 것이다. 따라서 이 토큰들을 표준 형태인 "하다"로 묶어줌으로써 이들 토큰들이 가지는 의미의 유사성을 살리면서 변수의 개수를 줄일 수 있다. 다만, 앞에서 언급했듯이 정규화를 통해 세 가지 표현이 갖는 차이점들은 사라지게 되므로 다소간의 의미의 손실이 발생하게 된다.

마지막으로 품사 태깅을 통해 각 토큰들의 품사를 구분해 주면 기본적인 전처리 과정이 완료되어 아래와 같이 표현된다.

"금융통화위원회/Noun 는/Josa 다음/Noun 통화정책/Noun 방향/Noun 결정/Noun 시/Noun 까지/Josa 한국은행/Noun 기준금리/Noun 를/Josa ..."

2. 토큰의 수치형 자료로의 표현 방법

전처리 과정이 완료되면 텍스트 데이터를 수치형 자료로 표현할 수 있게 된다. 이 절에서는 텍스트 데이터를 수치형 자료로 표현하는 방법들과 그 특성에 대해 살펴본다.

가. Bag[1]-of-words 표현방식

Bag-of-words 표현방식은 "문서에서 특정 단어의 출현빈도(frequency)를 통해 질의(query)와 문서의 관계를 파악할 수 있다"는 "Bag of words 가설"에 기반한다(Turney and Pantel, 2010). 이 정의에서 알 수 있듯이 "Bag of words 가설" 하에서는 단어 사용 빈도는 고려하지만 문장의 문법적 구조나 단어의 선후 관계 등은 고려하지 않는다.

Bag-of-words 표현방식에서는 각 토큰을 벡터의 고유한 인덱스로 간주하여 각 토큰이 문서에 몇 번이나 포함되어 있는지를 해당 인덱스에 기록한다. 예를 들어 위에서 소개한 통화정책방향 결정문의 문장은 문장부호를 제외하면 <표 2-1>과 같은 벡터로 표현할 수 있다.

〈표 2-1〉 Bag-of-words 표현 방식의 예

인덱스	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
토큰	결정	금융통화위원회	기준금리	까지	는	다음	를	방향	수준	시	에서	운용	유지	을	통화정책	하다	한국은행	현
빈도	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	1	1

이렇게 표현된 문서의 벡터는 one-hot 벡터[2]로 표현된 단어 벡터들의 합으로 볼 수 있다. one-hot 벡터란 하나의 인덱스에서만 1의 값을 가지고 나머지 인덱스에서는 모두 0의 값을 갖는 벡터를 말한다. 즉, 위의 보기에서 '하다'라는 단어의 경우 16번째 인덱스에 1을, 나머지 다른 인덱스에는 0을 기록한 단어 벡터로 표현할 수 있다. 주어진 문장에서 '하다'는 세 번 나타났으므로 16번째 인덱스에 3을 기록하게 된다. 이렇게 각 단어 벡터를 문서

1) "Bag"이란 원소의 중복을 허용하는 중복집합(multiset)을 뜻한다.
2) Bag-of-words 방식은 문서, 문장 등을 표현하는 방식이고 동 방식의 기본이 되는 단어 표현 방식이 one-hot encoding 방식이다.

(문장) 전체에서 합산하면 예시와 같은 문서 벡터가 주어진다.

Bag-of-words 표현방식은 각 문서 또는 문맥 내에서의 단어 출현횟수를 기록한 단어-문서 행렬(term-document matrix), 단어-문맥 행렬(term-context matrix) 등에서 많이 사용되어 왔다. 예를 들어 만약 단어 w가 특정한 문맥 c에서 나타나는 빈도가 다른 문맥들에 비해 높다면 w는 문맥 c를 표현하는 데 중요한 역할을 하고 있다는 것을 알 수 있다. 즉, 단어 w가 문맥 c에서 나타난 빈도가 단어 w의 일반적인 출현확률 $\mathrm{P}(w)$와 문맥 c의 비율 $\mathrm{P}(c)$에 비해 높다면 단어 w와 문맥 c의 연관성이 높음을 알 수 있는데 이러한 관계를 표현한 지표가 PMI (pointwise mutual information)이다.

$$\mathrm{PMI}(w,c) = \log \frac{\mathrm{P}(w,c)}{\mathrm{P}(w)\mathrm{P}(c)}$$

다만, Bag-of-words 표현방식은 다음과 같은 한계를 지닌다. 먼저 방대한 양의 문서를 사용하여 텍스트 마이닝 작업을 하는 경우 토큰의 수가 매우 많아지므로 벡터의 차원도 크게 늘어나게 된다. 하지만 각각의 문서에서는 한정된 수의 토큰만 사용되는 경우가 많으므로 각 문서 벡터는 많은 수의 0으로 채워지게 된다(Jurafsky and Martin, 2014). 따라서 Bag-of-words 방식으로 표현된 단어-문서 행렬 등은 고차원(high-dimensional)의 성긴(sparse) 구조를 지닌 벡터이며 통계이론 측면에서는 성질이 좋지 않은 경우가 많다.

그리고 Bag-of-words 방식으로 문서를 표현하면 비슷한 의미를 가진 문서도 의미의 유사성을 포착해 내기 어려운 경우가 있다. 또한 반대로 동음이의어가 사용된 경우 문서 벡터의 유사성은 높지만 실제로는 의미의 유사성이 없는 경우도 있을 수 있다. Bholat et al.(2015)은 <그림 2-1>과 같은 예를 통해 Bag-of-words 표현방식에서 의미의 유사성을 정확히 드러내기 어려운 경우를 제시하였다.

〈그림 2-1〉 **Bag-of-words 방식으로 표현된 문서의 유사성**

	문서1	문서2		문서3	문서4
school	0	10	tank	5	5
university	5	0	marine	5	5
college	5	0	frog	3	0
teacher	1	4	animal	2	0
professor	2	0	navy	0	4

벡터 사이의 유사성을 표현하는 방법은 다양하다(Bholat et al., 2015). 그 중 코사인 유사도(cosine similarity)는 두 벡터 사이의 각도를 기준으로 두 벡터의 유사성을 측정한다. 구체적으로 단어 w와 단어 v 사이의 코사인 유사도는 다음과 같이 정의된다.

$$\text{코사인유사도} = \frac{w^t v}{|w||v|}$$

여기서 $w^t v$는 단어 w와 단어 v의 내적(inner product)을, $|w|$와 $|v|$는 두 단어의 길이(length)를 의미한다. 이렇게 정의된 코사인 유사도를 기준으로 보면 Bag-of-words 방식으로 표현하는 경우 각 단어는 서로 다른 인덱스가 1인 one-hot 벡터로 표현되므로 두 단어의 내적은 0이 되며 코사인 유사도도 0이 되어 두 단어 사이의 의미의 유사성이 없는 것으로 나타난다. 따라서 위의 예에서 school, university, college는 비슷한 의미를 지닌 단어이지만 이러한 의미의 유사성을 Bag-of-words 방식으로는 표현하기 어렵게 된다.

나. Word Embedding 표현방식

한편 Bag-of-words 표현방식에 의한 고차원의 성긴(sparse) 구조를 지닌 벡터의 한계점을 보완하기 위해 저차원의 조밀한(dense) 벡터 공간에 표현하는 방법들이 제안되어 왔다(Jurafsky and Martin, 2014). 대표적인 방법으로 Deerwester et al.(1990)이 제안한 LSA(Latent semantic analysis)가 있다. LSA는 단어-문서 행렬의 특이값 분해(singular value decomposition)를 통해 특이값이 큰 성분만을 이용하여 단어-문서 행렬을 근사적으로 표현하는 방법으로 단어 벡터의 차원을 크게 줄일 수 있는 장점이 있다. 그러나 의미의 유사성을 단어 벡터에 반영하기 어려우므로 벡터 공간 표현에 있어서의 최적화가 이루어지지 않았음을 의미하는 것으로 인식되고 있다(Pennington et al., 2014).

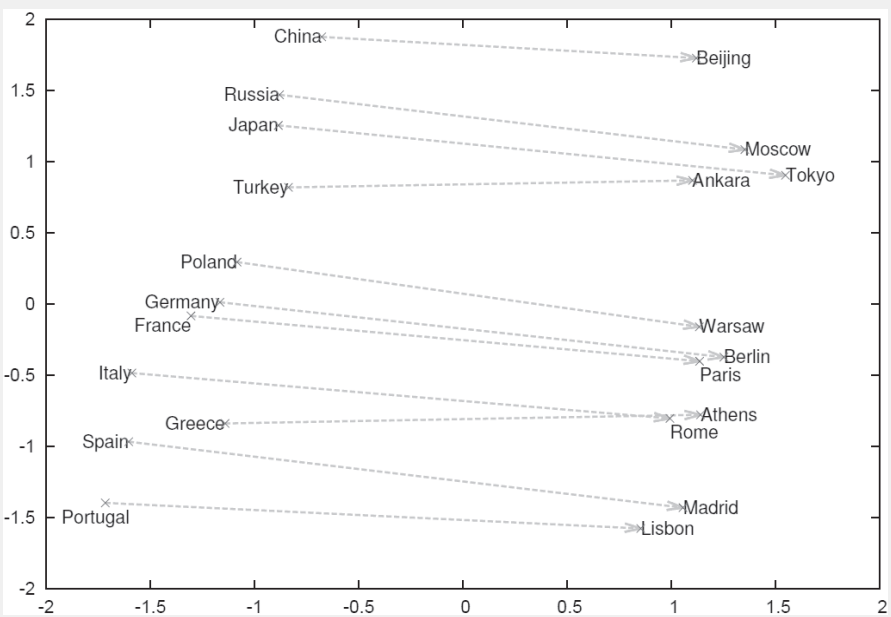
또 다른 방법으로 Mikolov et al.(2013a)이 제안한 word2vec 방법을 소개한다. word2vec 표현방식에서는 인공신경망(neural network) 모형을 사용하여 아래와 같은 목적함수를 최적화하는 벡터 공간을 구한다.

$$Q_{ij} = \frac{\exp(w_i^t \tilde{w}_j)}{\sum_{k=1}^{V} \exp(w_i^t \tilde{w}_k)}$$

여기서 단어 $\widetilde{w}_j$는 단어 문맥 속에서 자주 단어 w_i와 함께 나타나는 단어로, V는 문서에 나타난 단어의 수로 정의된다. 이렇게 정의할 때 Q_{ij}는 같은 문맥 상에서 자주 함께 등장하는 단어 벡터 $\widetilde{w}_j$가 다른 단어 벡터들에 비해 상대적으로 얼마나 단어 w_i 벡터에 가깝게 표현되는지 측정하는 지표로 해석할 수 있다. 이 목적함수 Q_{ij}는 유사성이 높은 단어의 내적은 크고 유사성이 낮은 단어의 내적은 0에 가까울 때 최대가 된다. 따라서 word2vec으로 표현된 벡터는 단어 사이의 의미의 유사성이 높은 경우, 조금 더 정확하게 표현하면 단어들이 비슷한 문맥에서 자주 표현되는 경우, 코사인 유사도도 높게 나타나게 된다.

이렇게 word2vec 방식으로 단어를 표현하는 경우 비슷한 의미의 단어들이 비슷한 형태의 벡터로 표현될 수 있으므로 텍스트 마이닝 과정에서는 단어 벡터들의 덧셈, 뺄셈 등의 연산이 가능하게 된다는 장점이 있다. <그림 2-2>에서 보는 바와 같이 개별 국가명과 수도가 대체로 평행한 관계를 가지고 있으며 이러한 관계를 활용하면 'A국가명-B국가명+B국 수도'라는 연산을 통해 A국가의 수도를 구할 수 있게 된다.

〈그림 2-2〉 **word2vec 방식으로 표현된 국가와 수도 벡터의 관계**



이와 같은 word2vec 표현 방법의 특성은 문서를 주제별로 분류하는 데 큰 장점이 될 수 있다. 예를 들어 경제와 관련된 문서의 경우 경제용어들이 많이 나타나게 되고 이들 단어들은 서로 유사한 형태의 벡터로 표현될 것으로 추정된다. 따라서 문서 또는 문장 내의 단

어들의 평균이나 합을 구하는 것만으로도 해당 문서 또는 문장의 주제가 경제와 관련되어 있는지 쉽게 파악할 수 있게 된다.

word2vec 표현 방식은 마치 어린 아이들이 어른들의 대화를 들으며 언어를 배우는 것과 같이 문맥을 통해 단어의 의미를 학습하게 되므로 많은 양의 문서를 학습할수록 좋은 결과를 얻을 수 있는 것으로 알려져 있다. 따라서 초기 학습과정에서 많은 양의 자료와 시간이 필요하다. 반면에 단어들이 저차원의 벡터로 표현되므로 일단 학습이 완료되면 이후의 분석 과정이 빠른 속도로 진행될 수 있다는 장점을 가진다.

한편 word2vec 방법은 짧은 문맥 안에서 함께 나타나는 단어들만으로 벡터 표현방식을 찾기 때문에 더 넓은 맥락에서의 의미의 유사성을 표현하는 데에는 한계가 있을 수도 있다. 이러한 점에 착안하여 Pennington et al.(2014)은 전체 문서의 맥락(context)을 고려하는 'GloVe' [3]라는 벡터 표현방법을 제안한 바 있다.

본 연구에서는 Bag-of-words 방식과 word2vec 방식으로 단어를 표현하고 이들 방식으로 문서를 분류한 결과를 비교해 보기로 한다.

3) 단어들의 동시출현 빈도수를 가중 최소제곱법(weighted least squares)으로 최적화하여 단어 사이의 의미 유사도를 구하는 방법 (Global Vectors for Word Representation)이다.

Ⅲ. 실제 텍스트 자료의 분석

지금까지 살펴본 텍스트 데이터의 수치형 자료로의 표현 방식을 실제 텍스트 자료에 적용하여 보았다. 분석을 위한 자료로는 한국정보화진흥원(NIA)으로부터 제공받은 2007년 1월부터 2017년 6월까지의 한글 언론기사 데이터를 사용하였다. 전처리를 위해서는 파이썬(Python) KoNLPy 패키지(Park and Cho, 2014)와 소셜네트워크 서비스 트위터(Twitter)에서 개발한 형태소 분석기를 사용하였다. 트위터의 형태소 분석기는 KKMA와 MeCab-ko 등의 다른 형태소 분석기에 비해 상대적으로 처리 속도가 빨라 대량의 문서를 처리하는 데 유용한 것으로 알려져 있다.

word2vec 방법을 이용한 단어 임베딩(embedding) 작업에는 파이썬 gensim 패키지(Řehůřek and Sojka, 2010)를 사용하였다. gensim 패키지는 LSA(Latent Semantic Analysis), LDA(Latent Dirichlet Allocation) 등의 토픽 모형(topic model)을 구현하는 데에도 유용하게 활용되고 있다. word2vec에서 벡터의 차원 선택과 관련해서는 아직 많은 연구가 이루어지지 않았는데 Levy and Goldberg(2014)는 50~300 차원의 벡터를 사용한 바 있다. 본 연구에서는 100차원 word2vec를 사용하였다.

1. 단어의 의미 유사성

전술한 바와 같이 word2vec 알고리즘은 비슷한 뜻을 가진 단어들을 비슷한 형태의 벡터로 표현하는 장점을 가지고 있다. 영어 단어에 대한 학습 결과, 단어-문맥(word-context)을 통한 학습 방법에 비해 word2vec 학습 방법으로 작성된 단어 벡터들이 단어의 유사성을 더 잘 표현하는 것으로 알려져 있다(Levy and Goldberg, 2014).

위에서 서술한 한글 언론기사 자료들을 학습용 데이터(training data)로 사용하여 각 단어들을 벡터로 표현하고 비슷한 뜻을 가진 단어의 벡터들이 실제로 코사인 유사도가 높은 것으로 나타나는지 확인해 보았다. Levy and Goldberg(2014)는 영어 문서들에 등장한 단어 벡터들의 유사성이 잘 표현되는지 확인하기 위해 MSR, GOOGLE 등[4]에 사용되었던 유추(analogy) 문제 데이터를 사용하였다. 한글 단어들의 경우 이런 대규모의 검증용 데이터 셋

4) MSR(Microsoft Research) 데이터 셋은 Mikolov et al.(2013b)이 사용한 약 8,000개의 검증용 유추 문제를, GOOGLE 데이터 셋은 Mikolov et al.(2013c)이 사용한 19,544개의 유추 문제를 포함한다.

이 존재하지 않기 때문에 출현 빈도가 높은 단어를 중심으로 유사성을 점검할 수 있는 소규모의 검증용 데이터 셋을 작성하여 결과를 확인하였다.

먼저 언론기사 자료에서 출현 빈도가 높은 단어들을 기반으로 공통점이 있는 단어 3개와 이들 단어와 개념에 다소 차이가 있는 단어 1개로 4개의 단어를 구성하고 이들 단어 중 가장 연관성이 낮은 단어를 판별하는 문제를 통해 word2vec 학습 방법의 성능을 점검하였다.

<표 3-1>에서 처음 3개의 단어는 유사한 의미를 지닌 단어이고 다음 열에 기록된 정답은 3개 단어들과 의미에 차이가 있는 단어를 의미한다. 표에서 볼 수 있듯이 word2vec 기법은 전체 64개의 문제 중에서 56개 문제의 정답을 맞춰 87.5%의 높은 정확도로 개념상 차이가 있는 단어를 식별해 내었으며 경제적인 의미를 가진 단어도 잘 분별하고 있음을 알 수 있다. 다만, 일부 단어 조합에서는 의미의 유사성이 없는 단어를 선택하여 문맥을 통한 단어의 의미 학습 절차가 완벽하지는 않음을 보여준다.

다음으로 단어 쌍의 관계를 이용한 유추(analogy) 문제를 이용하여 단어 벡터의 학습이 잘 이루어졌는지 확인해 보았다. 유추 문제는 '단어 쌍 A와 a의 관계는 B와 b의 관계와 같다'라는 문장에서 단어 b가 주어져 있지 않을 때 b를 추론하는 문제로 다양한 답이 존재할 수 있다.

앞서 언급한 바와 같이 word2vec으로 학습된 단어 벡터들은 의미 관계를 벡터 연산으로 표현할 수 있다는 장점이 있다. 단어 A, B, a가 주어져 있을 때 단어 b를 구하는 유추 문제에서 A-a = B-b 임을 이용하여 미지의 단어 b를 b = B-A+a로 구할 수 있다. 이 연산에 대한 답은 파이썬 gensim 패키지의 most_similar 명령어를 사용하여 구하였다.

<표 3-2>는 30개의 유추 문제에 대해 가장 유사도가 높은 단어 3개를 보여주고 있다. 매칭 문제에서와 마찬가지로 word2vec 기법으로 학습된 단어 벡터들은 유추 문제를 잘 풀고 있음을 확인할 수 있다. 다만, 유추 문제에서의 정확도는 80%로 매칭 문제에 비해 떨어지는 것을 확인할 수 있는데 이는 단어 집합에서 어울리지 않는 단어를 찾는 문제에 비해 유추 문제에 훨씬 많은 경우의 수가 존재하기 때문인 것으로 판단된다.

〈표 3-1〉 word2vec 학습 결과 : 단어 벡터의 유사성 검증

유사단어			정답	응답결과		유사단어			정답	응답결과	
한국	미국	중국	정부	정부	o	금액	액수	돈	부피	부피	o
기사	신문	뉴스	영화	영화	o	교과서	참고서	교재	필기구	필기구	o
정부	국회	법원	협회	협회	o	불법	위반	위법	준수	준수	o
서울	부산	대전	섬	섬	o	원인	이유	요인	결과	이유	x
사람	인류	인간	사물	사람	x	백화점	마트	편의점	미용실	미용실	o
기업	회사	업체	개인	개인	o	수도권	지방	서울	고향	고향	o
사업	영업	장사	봉사	봉사	o	매각	판매	매출	도입	도입	o
시장	거래	매매	기부	기부	o	교회	사찰	성당	고궁	고궁	o
경제	생산	소비	여가	여가	o	질환	질병	병	회복	회복	o
세계	글로벌	해외	국내	세계	x	변호사	판사	검사	의사	의사	o
투자	수출	소비	하락	하락	o	겨울	봄	여름	휴가	휴가	o
학교	교육	대학	정치	정치	o	공항	항구	선착장	휴게소	휴게소	o
회사	기업	사업	교육	교육	o	졸업	퇴직	수료	이민	이민	o
운영	관리	경영	폐업	폐업	o	판결	항소	항고	검거	검거	o
주택	부동산	토지	주가	주가	o	근로자	노동자	노무자	고용주	고용주	o
거래	매매	교환	보유	교환	x	야구	축구	농구	무용	무용	o
생활	인생	삶	가족	가족	o	버스	기차	택시	잠수함	잠수함	o
영화	공연	연극	출판	출판	o	빚	부채	채무	이익	이익	o
우려	위기	위험	단체	단체	o	남편	배우자	아내	가족	배우자	x
협력	협상	합의	연금	연금	o	코	눈	귀	허리	코	x
주식	채권	펀드	환율	환율	o	목사	승려	수녀	회장	회장	o
교사	교수	학생	소방관	교수	x	자본	자산	부채	임금	임금	o
보도	기사	뉴스	회의	회의	o	업종	산업	업체	정부	정부	o
금리	주가	수익률	임금	임금	o	학부모	학생	교사	배우	배우	o
선거	투표	표결	퇴임	퇴임	o	화학	물리	생물	영어	영어	o
과거	현재	미래	장소	장소	o	새벽	오후	오전	개월	개월	o
상승	하락	보합	현실	현실	o	크기	면적	부피	합	합	o
수익	손실	이익	채권	채권	o	선고	판결	공판	기소	공판	x
고용	채용	실업	물가	물가	o	경제	환율	주가	정치	정치	o
아시아	유럽	남미	도시	도시	o	투자	소비	수출	주가	주가	o
일자리	직장	직업	확산	확산	o	물가	유가	환율	투자	투자	o
아버지	어머니	부모	친척	친척	o	협의	논의	토의	결론	결론	o

〈표 3-2〉 word2vec 학습 결과 : 단어 벡터의 유추 성능 검증

단어A	단어a	단어B	단어b	결과	응답1	응답2	응답3
한국	서울	미국	워싱턴	x	용산구	광진구	사당동
기자	신문	배우	영화	x	신문기자	콩트	이서진
의원	국회	판사	법원	o	서울중앙지법	법무부	법원행정처
국가	국민	학교	학생	o	학부모	학생	교사
병원	의사	대학	교수	o	대학교수	학자	재학생
대출	이자	임대	임대료	x	덕정역	동백동	동탄신도시
여왕	왕	여자	남자	o	남자	오랑캐	왕후
왕비	왕	여자	남자	o	남자	쇼트트랙	스피드스케이팅
소비	침체	주가	하한가	o	하한가	악재	어닝쇼크
호황	호조	불황	침체	o	강세	부진	약세
불황	침체	호황	활황	o	활황	둔화	오일쇼크
수출	수입	매출	매입	x	판매량	판매	마진
학교	교사	대학	교수	x	졸업생	대학원생	재학생
행정부	대통령	국회	국회의장	o	새누리당	원내대표	국회의장
사장	직원	교장	학생	o	원생	교사	학생
채권	금리	주식	수익률	o	수익률	투자수익률	이자율
영화	배우	콘서트	가수	o	가수	인디밴드	디너쇼
미국	달러	일본	엔화	o	만엔	엔	호주달러
삼성	갤럭시	애플	아이폰	o	아이폰	안드로이드	아이패드
공급	증가	가격	상승	x	감소	늘어나다	줄다
아버지	어머니	아빠	엄마	o	엄마	친정엄마	아이
유가	원유	환율	원화	o	원화	유로화	엔화
영화	상영	연극	공연	o	공연	대학로	공연예술
토지	면적	휘발유	리터	o	ℓ	리터	소비전력
아침	오전	저녁	오후	o	오후	밤	새벽
근로자	고용주	학생	교사	o	학부모	교사	재학생
화재	소방관	사건	경찰관	o	경찰관	공범	범인
학교	졸업	직장	퇴직	o	입사	취직	퇴직
기사	기자	소설	작가	o	작가	원작	나희덕

마지막으로 경제와 관련된 단어들의 유사성이 정확히 포착되는지 확인해 보기 위해 파이썬 gensim 패키지의 most_similar 명령문을 사용하여 유사도가 높은 단어들을 추출해 보았다. <표 3-3>에서는 앞선 유추 문제에서와 마찬가지로 대체로 유사한 의미의 단어를 잘 파악해 내고 있는 것을 확인할 수 있다.

하지만, '상승'과 같은 단어의 경우에는 비슷한 의미를 가진 '급등' 대신 유사한 문맥에서 사용되는 단어이기는 하지만 정반대되는 의미를 가진 '하락'을 유사도가 가장 높은 단어로 분류한 것으로 나타났다. 이렇게 동의어와 반의어를 정확하게 포착해 내지 못하는 것은 word2vec 학습방법의 특성 때문인 것으로 판단된다. word2vec 학습방법의 경우 문맥상에서 단어의 의미를 포착해 내기 때문에 비슷한 문맥에서 함께 사용될 수 있는 단어의 경우 의미를 정확히 구분해 내기 어렵다. '상승'과 '하락'의 경우 "시장 금리가 ...%p 상승/하락하였다"라는 문구에서 알 수 있는 바와 같이 비슷한 문맥에서 사용될 수 있는 경우가 많으므로 word2vec 학습방법으로는 유사도가 매우 높은 단어로 판별되지만 실제로는 정반대되는 의미를 내포하고 있다.

또한 '경기'와 같이 여러 가지 의미로 사용되는 동음이의어의 경우에도 정확한 식별이 쉽지 않다는 점을 확인할 수 있다. 언론기사 자료에서 '경기'는 경제적인 의미 보다는 스포츠와 관련된 단어들과 의미의 유사성이 높은 것으로 식별되었다.

한편, 동의어와 반의어의 처리 문제의 경우 전체적인 문서의 주제를 분류하는 데에는 큰 영향을 미치지 않지만 동음이의어의 경우 문서 주제를 분류하는 데 교란요인으로 작용할 가능성이 있다. 예를 들어 '경기'라는 단어가 자주 사용된 기사의 경우 원래 주제가 경제 상황과 연관되어 있다고 하더라도 word2vec을 통한 단어 벡터 표현상으로는 경제 상황이 아닌 스포츠와 관련된 것으로 분류될 수 있다.

〈표 3-3〉 word2vec 학습 결과 : 유사성이 높은 단어 벡터의 추출

	유사도 1순위	유사도	유사도 2순위	유사도	유사도 3순위	유사도
시장	틈새시장	0.681	신흥시장	0.625	업계	0.606
경제	경제성장	0.740	거시경제	0.739	경제정책	0.705
금융	금융기관	0.748	상호금융	0.714	은행	0.713
회사	업체	0.719	자회사	0.714	거래처	0.683
달러	호주달러	0.832	파운드	0.784	유로	0.756
주택	임대주택	0.788	전월세	0.762	전세	0.752
가격	값	0.841	출고	0.662	우윳값	0.647
은행	금융기관	0.841	은행권	0.721	금융	0.713
거래	매매	0.683	외환거래	0.651	선물거래	0.634
소득	소득세	0.656	가처분소득	0.655	중산층	0.640
판매	구매	0.762	직수입	0.674	출시	0.662
부동산	주택	0.709	집값	0.654	전세	0.651
주식	배당	0.685	채권	0.684	펀드	0.683
소비자	고객	0.745	이용자	0.638	사용자	0.637
자산	주식	0.664	투자수익률	0.644	채권	0.627
대출	연체	0.769	채무자	0.705	보증	0.686
금리	기준금리	0.796	이자율	0.792	시중금리	0.753
상승	하락	0.957	급등	0.871	급락	0.804
수익	수익률	0.665	이익	0.661	투자수익률	0.656
재정	국가부채	0.705	예산	0.681	추경	0.670
이익	순이익	0.714	이윤	0.703	손실	0.676
고용	일자리	0.746	시간제	0.712	근로	0.696
수출	수출량	0.695	수출입	0.691	제조업	0.671
세금	소득세	0.727	법인세	0.684	월세	0.675
채권	회사채	0.720	차입	0.698	유동성	0.688
수입	수입품	0.728	외국산	0.678	수출량	0.662
주가	증시	0.692	우선주	0.675	뉴욕증시	0.674
실적	업황	0.726	어닝	0.632	순익	0.631
일자리	고용	0.746	양질	0.658	시간제	0.658
수요	공급물량	0.646	소비	0.607	호재	0.592
소비	총수요	0.632	구매력	0.619	물가	0.612
임금	근로시간	0.795	상여금	0.776	최저임금	0.728
취업	구직	0.806	대졸	0.705	고졸	0.697
매각	인수	0.835	증자	0.763	매입	0.737
부채	국가부채	0.795	국가채무	0.759	나랏빚	0.735
적자	흑자	0.821	부채	0.653	손실	0.631
전세	월세	0.840	전셋값	0.794	집값	0.780
지출	재정	0.640	가처분소득	0.597	가계	0.596
증시	주식시장	0.845	미국증시	0.759	다우지수	0.738
재산	증여	0.682	증여세	0.668	명의신탁	0.659
채용	공채	0.738	고졸	0.722	신입	0.706
손실	환차손	0.710	손해	0.690	이익	0.676
인하	인상	0.736	금리인하	0.647	동결	0.639
급여	퇴직금	0.767	월급	0.749	수당	0.741
환율	원화	0.838	엔화	0.751	절상	0.719
물가	물가상승률	0.762	인플레이션	0.687	금리	0.681
매입	매각	0.737	현금화	0.709	발행	0.648
경기	이경기	0.736	문학구장	0.602	김온아	0.598
자금	뭉칫돈	0.693	현금	0.685	유동성	0.678

2. 경제심리 관련 문서의 분류

분류 문제는 대표적인 지도학습(supervised learning) 문제로 사전에 분류가 이루어져 있는 훈련용 자료들(training data)을 기반으로 분류 모형을 작성하고 검증용 자료(test data)에 이 분류 모형을 적용하여 분류 모형의 성능을 확인한다. 분류 문제는 가장 일반적인 데이터 마이닝 문제 중 하나로 Fisher(1936)의 선형판별분석(linear discriminant analysis) 이래 많은 통계적 방법론들이 개발되어 왔다. 대표적으로 많이 사용되는 방법들로는 Cox(1958)의 로지스틱 회귀(logistic regression) 모형, Cortes(1995)의 SVM(support vector machine)을 비롯하여 의사결정나무(decision tree) 모형, 랜덤 포레스트(random forest), 부스팅(boosting), 인공신경망(neural network) 모형 등이 있다. 본 연구에서는 일반적으로 가장 많이 활용되는 로지스틱 회귀모형(logistic regression model)을 사용하였다.

로지스틱 회귀모형에서는 관측값 x가 주어졌을 때 모집단 1이 참일 확률 $P(Y=1|x)=\pi(x)$와 모집단 1이 참이 아닐 확률 $1-\pi(x)$의 비(odds ratio)에 로그를 취한 로짓(logit)이 아래와 같은 선형회귀모형을 따른다는 가정하에 모형을 적합시킨다.

$$\log\frac{\pi(x)}{1-\pi(x)}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$

본 연구에서는 분류 모형을 추정하고 검증하기 위해 5개 언론사[5]의 2013년 1월부터 2014년 3월까지의 기사 중 총 9,771건의 기사를 사용하였다. 설명변수 x로는 word2vec 방법으로 작성된 100차원 벡터와 Bag-of-words 방식으로 기록된 벡터 중 출현빈도가 높은 100개, 500개, 1000개의 단어로 이루어진 100, 500, 1000차원 벡터를 사용하였다. 이들 단어 벡터를 설명변수로 하여 전체의 75%에 해당하는 훈련용 데이터로 로지스틱 회귀모형을 추정하고 나머지 25%의 검증용 데이터로 경제심리와 관련된 기사를 잘 분류하는지 확인해 보았다.

가. 분류 성능의 평가

분류 대상 집단의 수가 두 개인 이진 분류(binary classification)의 경우 가능한 분류 결과는 모두 네 가지이다. 참값이 양(positive)인 자료를 양(positive)으로 식별(true positive, TP)하는 경우와 참값이 음(negative)인 자료를 음(negative)으로 식별(true negative, TN)하는 경우는

5) 조선, 중앙, 경향, 한겨레, 매일경제 등 주요 일간지 기사를 사용하였다.

정확한 분류가 이루어진 경우들이다. 반대로 잘못된 분류 결과로는 참값이 양(positive)인 자료를 음(negative)으로 식별(false negative, FP)하는 경우와 참값이 음(negative)인 자료를 양(positive)으로 식별(false positive, FP)하는 경우가 있다. 이들 분류 결과는 아래 <표 3-4>와 같이 2-by-2 분할표(contingency table, confusion matrix)를 이용하여 정리할 수 있다. 본 연구에서 사용하고 있는 기사 자료에서는 경제상황과 관련된 기사를 양(positive)으로, 그렇지 않은 기사를 음(negative)으로 간주하였다.

〈표 3-4〉 **이진 분류에서의 가능한 분류 결과**

		True value	
		Positive	Negative
Classification result	Positive	**TP** (true positive)	**FP** (false positive)
	Negative	**FN** (false negative)	**TN** (true negative)

분류 결과의 평가를 위해서 많이 사용되는 지표로는 정확도(accuracy), 민감도(sensitivity, recall, TPR), 특이도(specificity), 정밀도(precision) 등이 있다(<표 3-5> 참조). 정확도(accuracy = [TP + TN] / [TP + TN + FP + FN])는 전체 자료 중 정확히 분류된 자료의 비중을 뜻한다. 정확도가 분류 성능 평가의 한 기준이 되지만 분류 성능을 정확히 평가하기 위해서는 다른 지표들도 확인해 볼 필요가 있다(Fawcett, 2006).

민감도(sensitivity, recall = TP / [TP + FN])는 참값이 양인 자료들 중 분류 결과 양으로 식별된 자료의 비중을, 특이도(specificity = TN / [TN + FP])는 참값이 음인 자료들 중 음으로 분류된 자료의 비중을 의미한다. 정밀도(precision = TP / [TP + FP])는 양으로 분류된 자료들 중 참값이 실제로 양인 자료의 비중을 의미한다. 참값이 양인 자료를 정확히 분류하는 것이 중요할 때에는 이들 지표 중에서 민감도와 정밀도가 상대적으로 중요한 지표가 된다.

〈표 3-5〉 **분류 성능 검증에 사용되는 지표**

정확도(accuracy) = [TP + TN]/[TP + TN + FP + FN]
민감도(sensitivity, recall, TPR) = TP/[TP + FN]
특이도(specificity) = TN/[TN + FP]
정밀도(precision) = TP/[TP + FP]

먼저 9,771건의 전체 기사 자료에 대해 word2vec과 Bag-of-words 방법으로 만든 설명변수를 이용하여 로지스틱 모형을 만들고 네 가지 분류 성능 검증 지표를 비교하여 보았다. <표 3-6>에 나타난 바와 같이 모든 검증 지표에서 word2vec을 사용한 모형이 Bag-of-words 방식에 비해 더 좋은 분류 성능을 나타내었다. 다만, 정확도와 특이도는 대체적으로 높게 나타나 경제심리와 관련되지 않은 기사는 대체로 정확하게 분류하고 있는 반면, 민감도와 정밀도는 이에 크게 미치지 못해 경제심리와 관련된 기사의 분류에는 문제가 있음을 확인할 수 있다.

〈표 3-6〉 **로지스틱 회귀모형의 분류 성능(과소표본추출 미적용)**

	word2vec (dim=100)	Bag-of-words		
		dim=100	dim=500	dim=1000
정확도(accuracy)	0.969	0.961	0.954	0.955
민감도(sensitivity, recall, TPR)	0.604	0.321	0.538	0.481
특이도(specificity)	0.985	0.990	0.973	0.976
정밀도(precision)	0.646	0.586	0.467	0.472

위의 결과에서 알 수 있듯이 경제심리 관련 문서 분류의 문제에 있어서 고려해야 할 요소 중 하나는 경제상황에 관한 판단을 포함하고 있는 기사와 그렇지 않은 기사로 구성된 두 집단 간의 불균형에 관한 것이다. 전체 기사 중에서 경제상황과 관련된 문서의 비중은 상대적으로 낮은 편이며, 이렇게 불균형이 심한 자료의 경우 비중이 큰 집단은 정확히 추정하지만 비중이 작은 집단에 대한 분류 정확성은 떨어지는 편향된 모형이 추정될 가능성이 높다.

구체적으로 분류 모형을 적용할 기사 자료의 경우 총 9,771건의 기사 중 4.46%에 해당하는 436건의 기사만 경제상황에 관한 판단을 포함하고 있으며 나머지 대부분의 기사는 경제심리와는 무관한 기사로 비대칭적인 분포를 보인다. 이런 경우 예를 들어 기사 내용에 관계없이 무조건 경제상황에 관한 기사가 아니라고 판단하는 모형이라 하더라도 436건의 경제심리 관련 기사는 잘못 분류하지만 나머지 9,535건의 기사는 정확히 분류할 수 있으므로 분류 정확도가 95.54%가 된다. 하지만, 이런 모형은 관심의 대상이 되는 경제상황에 대한 기사를 전혀 식별해 낼 수 없으므로 좋은 분류 모형으로 받아들이기 어렵다.

이러한 모형 추정의 편향성을 완화하기 위해 다양한 방법들이 제안되어 왔다. 이들 방법은 과소표본추출(undersampling 또는 downsampling), 과표본추출(oversampling) 등을 통해 분류

모형 구축을 위해 사용하는 자료의 집단간 비중을 대등하게 만들어 주는 방법과 비중이 작은 집단에서 분류를 잘못했을 때 보다 더 큰 벌점을 부여하는 비용함수 이용법으로 크게 나누어 볼 수 있다(Ganganwar, 2012).

본고에서는 상대적으로 비중이 높은 경제심리와 무관한 자료의 비중을 축소하는 과소표본추출을 통해 집단간 비중을 조정하여 새로운 모형을 추정해 보았다. 이를 위해 경제심리와 관련이 없는 9,535건의 기사 중 2,000건을 랜덤추출하여 경제심리를 포함하고 있는 436건의 기사와 합하여 총 2,436건의 기사로 새로운 데이터 셋을 구성하였다. 이 중 75%를 사용하여 로지스틱 모형을 추정하고 나머지 25%의 기사로 추정된 모형의 분류 성능을 평가해 본 결과가 <표 3-7>에 기록되어 있다.

〈표 3-7〉 **로지스틱 회귀모형의 분류 성능(과소표본추출 적용)**

	word2vec (dim=100)	Bag-of-words		
		dim=100	dim=500	dim=1000
정확도(accuracy)	0.946	0.910	0.913	0.936
민감도(sensitivity, recall, TPR)	0.790	0.680	0.760	0.800
특이도(specificity)	0.976	0.955	0.943	0.963
정밀도(precision)	0.868	0.747	0.724	0.808

과소표본추출을 적용한 모형의 분류 성능을 살펴보면 과소표본추출 미적용시에 비해 정확도와 특이도가 소폭 떨어진 반면, 민감도와 정밀도는 상대적으로 크게 상승한 것을 확인할 수 있다. 과소표본추출에서도 전반적으로 word2vec 방식으로 만들어진 벡터를 설명변수로 사용한 로지스틱 모형이 더 좋은 분류 결과를 보였음을 확인할 수 있다.

한편 이들 네 가지 분류 성능 지표의 값은 분류함수의 임계치(threshold)를 어떻게 선택하느냐에 따라 달라질 수 있다. 임계치를 조정함으로써 양(positive)으로 분류되는 자료의 비중을 선택할 수도 있다. 분류함수는 참값이 양인 자료를 잘 분류(TP)하기 위해 임계치를 조정하면 참값이 음인 자료도 양으로 분류(FP)될 수 있고 반대로 음인 자료를 잘 분류(TN)하기 위해 임계치를 조정하면 참값이 양인 자료도 음으로 분류(FN)될 가능성이 있다. 예를 들어 모든 자료를 양으로 분류하면 FPR(= FP / [TN + FP] = 1-특이도) = 1이 되고 반대로 모든 자료를 음으로 분류하면 민감도(= TP / [TP + FN])는 0이 된다.

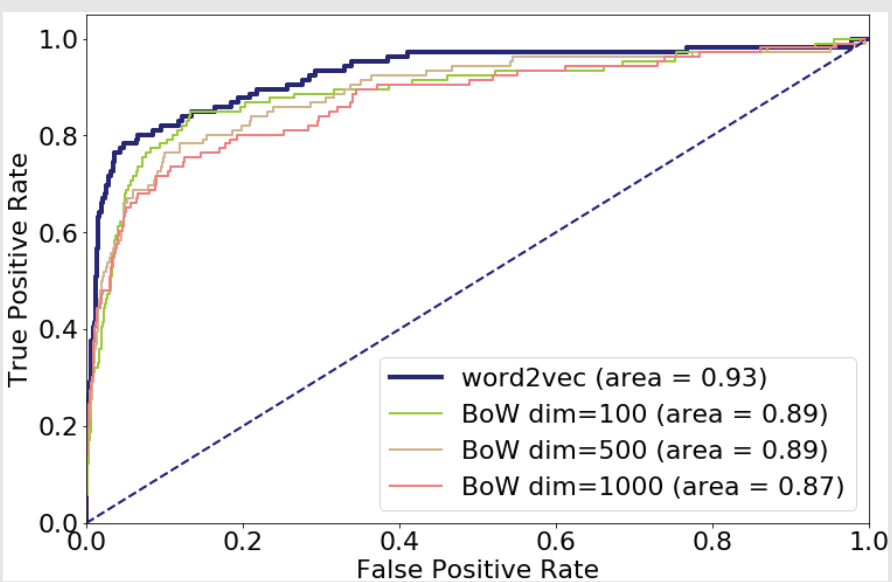
이러한 임계치 조정에 따른 민감도와 FPR의 변화를 ROC 곡선(receiver operating characteristic curve)을 통해 관찰할 수 있다. ROC 곡선에서는 곡선이 높은 쪽에 위치해 있을

수록 또 곡선으로 둘러싸인 부분의 면적인 AUC(area under curve)가 클수록 분류 성능이 상대적으로 좋음을 의미한다. ROC 곡선은 이 연구에서 분석하고 있는 기사 자료에서와 같이 두 집단의 분포 비중에 큰 차이가 있는 자료에도 유용하게 활용될 수 있음이 알려져 있다 (Fawcett, 2006).

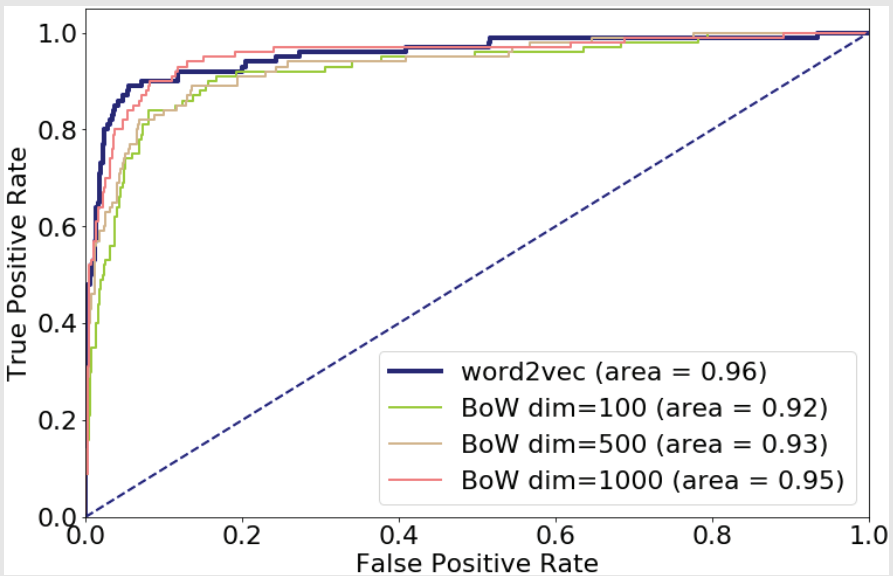
기사자료 분류를 위해 사용한 로지스틱 모형의 ROC 곡선은 <그림 3-1> 및 <그림 3-2>와 같다. 그림에서 볼 수 있는 바와 같이 과소표본추출이 적용된 경우와 그렇지 않은 경우 모두 word2vec 방식으로 만들어진 벡터를 설명변수로 사용한 로지스틱 회귀 모형이 Bag-of-words 방식으로 추정된 로지스틱 회귀 모형에 비해 전반적으로 높은 쪽에 위치해 있어 word2vec 방식의 벡터로 작성된 분류 모형의 성능이 더 좋음을 알 수 있다. 다만, 과소표본추출을 적용할 경우 1,000차원 Bag-of-words 방식 벡터를 이용한 로지스틱 회귀모형이 FPR이 0.1에서 0.3 사이일 때 조금 더 좋은 성능을 나타내었다.

AUC를 이용한 비교에 있어서도 word2vec 방식으로 표현된 벡터를 사용한 모형이 더 나은 성능을 보이는 것을 <표 3-8>에서 확인할 수 있다. 과소표본추출을 사용한 경우와 사용하지 않은 경우 모두 word2vec 방식 벡터를 사용한 로지스틱 회귀모형이 Bag-of-words 방식 벡터를 사용한 모형에 비해 AUC의 값이 크게 측정되었다.

〈그림 3-1〉 로지스틱 회귀모형의 ROC 곡선(과소표본추출 미적용)



〈그림 3-2〉 로지스틱 회귀모형의 ROC 곡선(과소표본추출 적용)



〈표 3-8〉 로지스틱 회귀모형의 분류 성능(AUC)

	word2vec	Bag-of-words		
		dim=100	dim=500	dim=1000
과소표본추출 미적용	0.93	0.89	0.89	0.87
과소표본추출 적용	0.96	0.92	0.93	0.95

Ⅳ. 결론 및 시사점

경제주체들의 심리는 실제 경제상황을 반영하고 있으며 앞으로의 경제활동에도 큰 영향을 미칠 수 있다. 이러한 경제 주체들의 심리를 파악하기 위해 전통적인 표본조사 방식 이외에 소셜미디어 등의 빅데이터(big data)를 활용하려는 시도들이 최근 많이 이루어지고 있다. 한편, 다양한 주제를 가진 텍스트 데이터(text data)로부터 경제심리를 분석하기 위해서는 해당 문서가 경제심리와 관련된 문서인지 확인하는 과정이 선행되어야 한다. 이에 따라 본고에서는 경제심리와 관련된 내용을 포함하고 있는 언론기사를 분류하는 방법을 모색해 보았다.

텍스트 데이터를 분석하기 위해서는 우선 텍스트 데이터를 수치형 자료로 표현해야 한다. 본 연구에서는 수치형 자료 표현 방식의 성능을 확인하기 위해 각각의 단어를 하나의 인덱스로 간주해 문서를 표현하는 Bag-of-words 방식과 각 단어를 보다 작은 차원의 축약된 벡터공간에 표현하는 word2vec 방식을 비교해 보았다. 이들 두 방식으로 만들어진 벡터를 설명변수로 하는 로지스틱 회귀모형을 추정하여 분류 성능을 비교해 본 결과, word2vec 방식이 Bag-of-words 방식에 비해 경제심리와 관련된 문서를 더 잘 분류해 내는 것으로 나타났다.

다만, 본 연구는 몇 가지 제한적인 상황에서 검증된 것이므로 향후 더 엄밀한 확인과정을 거칠 필요가 있는 것으로 판단된다. 먼저 word2vec 방식을 이용하여 생성한 단어 벡터들이 의미의 유사성을 잘 반영하고 있는지에 대해서는 비교적 적은 개수의 문제에 한해서 평가가 이루어졌으므로 보다 다양한 문제들을 이용하여 결과를 재확인할 필요가 있다. 또한 본 연구에서는 시간 제약 등으로 한정된 수의 훈련 및 검증용 자료만 사용하였으므로 향후 훈련 및 검증용 자료의 확충을 통해 분류 성능에 대한 보다 정확한 검증을 시도해 볼 필요가 있다.

아울러 이러한 연구 결과를 경제심리 분석 방법으로 발전시키기 위해서는 문서에 포함된 경제심리의 긍정/중립/부정 등의 극성(polarity)을 파악하는 방법론을 개발할 필요가 있다. 본문에서도 살펴본 바와 같이 word2vec 방식을 사용하는 경우 문맥을 통해 단어의 의미를 벡터로 표현하기 때문에 상승/하락, 개선/악화 등의 반대되는 심리를 가진 단어들도 유사성이 높은 단어로 판별하는 단점이 있을 수 있다. 이런 점을 감안할 때, 현재의 방식을 경제심리 분석에 바로 적용하는 것에는 무리라고 판단되며 향후 이를 보완하기 위한 연구가 더 진행되어야 할 것이다.

참고문헌

· Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131 (4), 1593-1636.

· Bholat, D. M., Hansen, S., Santos, P. M., and Schonhardt-Bailey, C. (2015). Text mining for central banks.

· Butler, D. (2013). When google got flu wrong. Nature, 494 (7436), 155.

· Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine learning, 20 (3), 273-297.

· Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), 215-242.

· Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41 (6), 391.

· Fawcett, T. (2006). An introduction to roc analysis. Pattern recognition letters, 27 (8), 861-874.

· Feldman, R. and Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

· Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of human genetics, 7 (2), 179-188.

· Ganganwar, V. (2012). An overview of classication algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2 (4), 42-47.

· Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. Technical report, National Bureau of Economic Research.

· Jurafsky, D. and Martin, J. H. (2014). Speech and language processing, volume 3. Pearson London.

· Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In Proceedings of the eighteenth conference on computational natural language learning, 171-180.

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

· Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In hlt-Naacl, volume 13, 746-751.

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

· Park, E. L. and Cho, S. (2014). Konlpy: Korean natural language processing in python. In Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology, 133-136.

· Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

· Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50. Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en

· Scott, S. L. and Varian, H. R. (2014). Predicting the present with bayesian structural time series. International Journal of Mathematical Modelling and Numerical Optimisation, 5 (1-2), 4-23.

· Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of articial intelligence research, 37, 141-188.

· Weiss, S. M., Indurkhya, N., and Zhang, T. (2010). Fundamentals of predictive text mining, volume 41. Springer.