거시경제 분석을 위한 텍스트 마이닝*

김 수 현** · 이 영 준*** · 신 진 영**** · 박 기 영*****

요 약

본고는 텍스트 마이닝의 주요 방법론 및 경제 분석의 활용 사례를 소개하는 것을 목적으로 한다. 언어 특유의 다의성과 비정형성에도 불구하고 텍스트로부터 수치화된 정보를 추출해 내는 텍스트 마이닝을 활용할 경우 새로운 데이터를 만들거나 기존의 데이터를 보완할 수 있고, 기존 방법론으로 분석이 어려웠던 주제들에 대한 연구도 가능하다. 빅데이터와 전산 기술의 발전으로 텍스트 마이닝 방법론은 빠르게 발전하고 있으며 이미 학계, 산업계, 정부 부문에서 활발하게 연구되고 이용되고 있다. 기존 데이터 및 방법론을 보완함과 동시에 새로운 정보를 제공하는 방법론으로 텍스트 마이닝의 수요는 더욱 증대될 것으로 기대한다.

핵심 주제어 : 텍스트 마이닝, 머신러닝, 경제 분석
JEL Classification Numbers : A12, B41, C80

Ⅰ. 서 론

Nate Silver[1]는 트위터(Twitter) 게시물을 분석하여 2008년 미국 대통령 선거 결과, 2010 총선, 2012 대선 결과 등을 그 어느 예측가보다 정확히 예측할 수 있

* 투고일(2019년 6월 8일), 수정일(2019년 10월 10일), 게재확정일(2019년 10월 28일)
** 제1저자, 한국은행 경제연구원 국제경제연구실 과장, E-mail: soohyonkim@bok.or.kr
*** Precourt Institute for Energy, Stanford University, Visiting Scholar, E-mail: yj.lee@yonsei.ac.kr
**** 연세대학교 경영대학 교수, E-mail: jshin@yonsei.ac.kr
***** 교신저자, 연세대학교 경제학부 교수, E-mail: kypark@yonsei.ac.kr
1) 미국의 정치평론가, 언론인으로 선거와 정치에 관련된 웹사이트 FiveThirtyEight을 운영한다. 저서로는 “The Signal and the Noise” 등이 있다.

었다. 검색엔진에서는 찾고자 하는 논문의 제목 또는 논문의 저자와 발간 연도를 검색하면 해당 논문뿐만 아니라 해당논문이 인용되었거나 유사한 주제를 다룬 논문까지 함께 찾아준다. 미국에서는 검색엔진에 독감 등 특정 질병 검색이 급증하면 해당 지역으로 백신을 미리 보내기도 한다. 또한 번역기에서는 어떤 단어든지 다양한 언어로 번역해준다. 이 모든 것이 가능하게 된 배경에는 텍스트 마이닝(text mining)이라는 빅데이터 처리 기법이 있다.

IBM 보고서(2015)[2]에 따르면 전 세계 데이터의 80%가 텍스트와 같은 비정형 데이터(unstructured data)로 이루어져있다. 우리가 매일 접하는 언어와 텍스트에는 풍부한 정보가 내재되어있으나, 하루에도 수백만 건의 문서가 생산되는 오늘날에는 인지능력만으로 방대한 텍스트 자료를 처리하는 데는 한계가 있다. 이제는 텍스트 마이닝으로 컴퓨터로 하여금 방대한 텍스트 데이터를 자동으로 분석하게 함으로써 유익한 정보를 효율적으로 얻을 수 있게 되었다.

텍스트 마이닝이란 비정형 데이터인 텍스트를 컴퓨터가 읽고 분석할 수 있도록 고안된 알고리즘을 다루는 분야이다. 텍스트 마이닝은 우리가 인지하지 못한 사이에도 이미 생활과 밀접한 관계에 있다. 특정 주제어로 논문을 검색하거나 책을 찾을 때에도 텍스트 마이닝으로 문서들이 요약된 데이터 베이스를 사용한다. 제품이 처음 출시되었을 때 신상품을 접한 고객들의 리뷰의 감성분석을 통해 제품의 초기 반응과 품질을 점검하기도 한다. 전화나 인터넷으로 서비스에 대한 불만 등이 제기되는 경우, 이를 분석하여 향후 서비스질 개선에 활용한다. 소셜 미디어에서는 게시글 및 댓글 데이터를 분석하여 잠재적 사이버 범죄를 사전에 예방하기도 한다. 또한 여러 게시글 데이터를 분석하여 마케팅에 사용할 수도 있다. 일부 소셜 미디어에서는 과거에 검색했던 주제와 관련된 광고가 선별되어 나타나는 것을 볼 수 있는데, 이것 또한 텍스트 데이터를 활용하여 검색어 혹은 검색어 묶음과 관련된 상품 혹은 서비스를 찾아주는 알고리즘에 기반한 것이라 볼 수 있다.

텍스트 데이터는 여러 연구에도 널리 활용되고 있는데 대표적으로 문헌정보학에서 텍스트 데이터는 오래전부터 연구의 대상이었다. 또한 정치학, 사회학에서

2) https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

도 공학이나 의학계열에서도 활용되고 있으며, 특히 공학 분야에서는 토픽 모델링(topic modelling)으로 과거 및 최신 연구동향을 점검하여 향후 진행할 연구 주제를 탐구하기도 한다.

이 외에도 다양한 분야에서 텍스트 마이닝을 활용한 예를 찾는 것은 어렵지 않다. 박진수 외(2018)는 112 신고내용에 대한 분석으로 주요 불법행위의 대상과 장소, 그 유형을 탐지하여 방법활동의 범위와 방향을 설정하고 소속 지구대 인원과 구성을 위한 기초자료를 제공하기도 하였다. 권충훈(2018)은 중등교사임용 시험 내용을 분석하여 교원양성기관 교육과정과의 정합성을 확인하기도 하고 권순보, 유진은(2018)은 수능 절대평가 관련 뉴스 기사를 수집하여 절대평가 관련 여론의 변화 추이를 분석하기도 하였다. 텍스트 마이닝은 의학분야에서는 이미 많이 알려진 연구방법론으로 빅데이터를 활용한 원격진료시스템 구축 사업의 일환으로 연구가 널리 진행되고 있다. Park *et al.*(2016)은 TF-IDF를 가중치로 환자가 구술한 진료기록을 분석하여 병명을 진단할 수 있는 연관규칙을 구성하였으며, 배효진 외(2018)는 한의학 원전(황제내경)을 텍스트 분석하여 한의학 교육에 사용할 수 있음을 밝혔다.

감미아, 송민(2012)은 언론사가 잘 알려진 보수 · 진보 성향에 따라 동일한 토픽에 대해 상이한 논조로 기사를 서술하는지 여부를 나이브 베이즈 분류모형[3](NBC: Naïve Bayes Classifier)으로 점검한 결과 정치 등 토픽에 대해 논조의 차이가 발생함을 발견하였다. 또한 김남원, 박진수(2012)는 나이브 베이즈 분류모형을 활용하여 SNS에 게시된 글의 개인정보 포함 여부를 판단하기도 하였다. 김유신 외(2012)는 뉴스의 감성을 분류하여 긍정 또는 부정적 뉴스에 따라 투자하는 의사결정모형을 구축하였다.

위의 사례 외에도 텍스트 데이터는 이미 재무, 마케팅, 심리학 등 다양한 사회과학 분야에서 널리 활용되고 있으나, 근래까지 경제학에서는 생소하게 인식되고 있으며 특히 국내의 경우 텍스트 데이터를 활용한 경제학적 분석 사례를 손에 꼽을 정도이다. 그러나 경제학적 분석에 활용할 수 있는 구조적 데이터의 한계를 보완하는데 텍스트 데이터가 지닌 장점이 분명히 존재하므로 향후 경제학

3) 나이베 베이즈 분류모형은 머신러닝의 한 방법으로 베이지안 정리를 응용한 분류기법이다.

연구에도 텍스트 마이닝 기법에 대한 수요가 있을 전망이다. 따라서 본고에서는 텍스트 마이닝이 경제학에 활용될 수 있는 방법론과 사례에 대해 상세하게 논하여 관련 연구에 도움이 되고자 하였다.

본고는 다음과 같이 구성되어있다. 2절에서는 텍스트 마이닝의 배경과 정의에 대해 알아보고 3절에서는 텍스트 마이닝의 방법론에 대해 논한다. 4절에는 텍스트 마이닝을 활용한 경제학적 연구와 활용 사례들을 소개하며 5절에서는 현 시점 텍스트 마이닝의 한계를 간략하게 서술한다.

Ⅱ. 텍스트 마이닝과 텍스트 데이터

1. 텍스트 마이닝의 정의

> "The individual words in a language name objects - sentences are a combination of such names. Every word has a meaning, it stands for something."
>
> - Saint Augustine

> "The meaning of words is best understood as their use within a given language-game"
>
> - Ludwig Wittgenstein, *Philosophische Untersuchungen*, 1953

텍스트 마이닝은 무엇인가? 이에 대해 우선 언어관에 대한 이해와 철학적 배경을 살피는 것이 텍스트 마이닝 근저에 자리한 아이디어를 이해하는 데도 도움이 될 것이다. 아우구스투스는 단어는 사전적으로 정의되어있고 문장이란 사전정의된 단어로 구성된다고 하였으며, 이렇게 구성된 언어는 만물과 일대일 대응구조를 갖는다고 주장한다. 가까운 사례로 우리가 흔히 접하는 어학사전에는 단

어(표제어)에 대해 '고유한 뜻'이 정의되어 있으며, 이러한 사전은 언어학습에 있어 아우구스투스의 언어관이 잘 반영된 결과라 할 수 있다. 이에 반해 비트겐슈타인의 철학에 따르면 삶의 형식이 변화하면 그에 따라 언어의 의미도 변화하며, 이에 따라 단어의 뜻은 언어의 사용례에 의해서만 이해될 수 있다고 한다.[4] 일상적인 예로 누군가가 내 뒤에서 "와 죽인다!"라고 감탄할 때 우리는 누구도 "죽임"을 사전에 정의된 부정적 의미로 해석하지 않는다. 이는 생활 속에서 사용되는 언어의 의미를 우리는 이미 알고 있기 때문이다. 이는 언어가 지닌 다의성 때문이며, 다의성으로 인하여 특정 단어가 사용된 문장 내에서 어떠한 의미를 나타낼 것인지에 대한 불확실성이 있다. 다른 한편으로 이러한 다의성과 불확실성은 구조화된 기호 및 숫자와 구분되는 언어의 고유한 특성이다(김규선, 2000).

단어에는 고유한 의미가 정의되어 문법에 맞게 단어를 배열하는 것이 인간의 언어라는 기존의 통념과 달리 사용된 단어의 관계 속에서 단어의 의미를 찾아낸다는 비트겐슈타인의 시각은 인간이 언어를 배우는 과정을 잘 묘사해준다.[5] 고유한 의미를 지닌 단어를 문법에 맞게 배열하는 것으로 언어사용을 정의하던 통념과 달리 단어의 관계 속에서 의미를 찾을 수 있다는 비트겐슈타인의 시각이 언어학습을 더욱 적절히 설명한다. 쉬운 예로 어린아이가 말을 배울 때 과정도 부모의 언어사용을 모방하는 것으로부터 시작하여 새로운 단어에 대한 뜻과 문맥적 활용법을 익히는 것이다. 비트겐슈타인의 언어관을 반영하면 기계(컴퓨터)로 하여금 사람의 언어를 학습시킬 수도 있을 것이다. 실제로 인공신경망 등에 의해 언어를 학습시키면 컴퓨터도 사람의 언어를 이해하고 분석할 수 있다. 이렇게 학습된 인공지능을 활용하면 자동화된 언어분석으로 우리에게 유익한 정보를 추출할 수 있다. 이러한 이론적 배경과 전산기술 발달, 인공지능(AI) 출현 등 여

4) 이를 철학에서는 언어게임(language-game)이라고 한다. 비트겐슈타인의 언어게임이란 패러다임의 변화가 언어 등의 변화를 수반한다는 점에서 토마스 쿤(T. Kuhn)이 제창한 과학혁명의 패러다임과도 일맥상통한다. 아우구스투스가 당위적(sollen) 언어관을 주장하였다면 비트겐슈타인은 언어의 실존(sein) 측면을 중시하였다고 볼 수 있다.

5) 언어에 관한 비트겐슈타인의 관찰은 논리실증주의에 기반하며, 향후 일상철학학파라는 철학의 분파 및 언어학(linguistics)에도 많은 영향을 주었다. Furnas *et al.*,(1983)의 "사용되는 언어의 통계적 유형으로 문서가 의미하고자 하는 바를 알 수 있다"라는 통계적 의미론 가설(statistical semantic hypothesis)은 비트겐슈타인의 철학적 배경을 언어학적 시각으로 해석한 것이라 볼 수 있다(Turney and Pantel, 2010).

러 학문분야가 복합적으로 구성되어 텍스트 마이닝 방법론으로 발전하였다.

오늘날 텍스트 마이닝은 빅데이터 분석 또는 데이터 과학의 대표적인 한 분야[6]로 자리매김하였으며, 머신러닝 등 전산기술을 활용하여 언어에 담긴 정보를 정형적 데이터로 추출해냄으로써 여러 학문분야에 기여하고 있다. 또한 텍스트 마이닝은 언어학, 전산학, 통계학이 관여하는 대표적인 다학제(multi-disciplinary) 연구 분야이다.[7] 텍스트 마이닝은 필연적으로 전산처리기술의 발달과도 밀접한 관련이 있는데, 텍스트를 전산처리 할 경우 초당 수백만 페이지의 텍스트를 읽고 자동으로 분석이 가능하기 때문이다(Cambria and White, 2014). 이뿐만 아니라 텍스트 마이닝은 검색엔진 등 인터넷 정보검색(information retrieval) 등 일상생활에서도 우리가 늘 사용하는 편리한 도구로써 널리 활용되고 있다.

2. 텍스트 데이터의 특징

텍스트 마이닝 방법론을 논하기 앞서 수치자료와 다른 텍스트 데이터가 지닌 특징 등을 우선 논할 필요가 있다. 텍스트 데이터의 특징은 언어의 특징이기도 하다. 텍스트가 지닌 비구조성, 모호성, 다의성 등은 텍스트로부터 정보를 추출하는데 상당한 장애가 된다. 이러한 특징은 후술하는 전처리(preprocessing) 등을 통해 텍스트를 정제하는 과정을 거쳐야하는 이유이기도 하다.[8]

---

6) 소위 빅데이터 시대에 텍스트 마이닝과 데이터 마이닝(data mining)은 흔히 접할 수 있는 용어이다. 데이터 마이닝의 경우 분석 대상이 정형화된 형태의 수치자료인 반면 텍스트 마이닝은 비정형 텍스트라는 점에서 차이가 있다. 또한 데이터 마이닝에서는 다량의 데이터에서 추세나 유형을 찾는 것이 주된 목적(Fayyad and Uthrusamy, 1999)이라면 텍스트 마이닝의 경우 텍스트 속에서 새로운 사실을 찾아내는 것이 목적(Hearst, 1999)이라 할 수 있으므로 미묘한 차이가 있다. 그러나 텍스트 마이닝도 빅데이터 분석에 해당하는 만큼, 데이터 마이닝에 활용되는 다양한 머신러닝 분류 방법이 함께 활용된다. 머신 러닝을 활용한 문서분류 방법은 본고에 기술되어 있다.

7) 과거 텍스트 마이닝이라 하면 텍스트 데이터 마이닝(Text Data Mining)이라는 매우 협소한 의미를 지칭하는 말이었으나, 현재는 품사 태깅(part-of-speech tagging), 사전구축 등의 기법을 활용하는 자연어 처리(natural language processing)와 전산언어학(computational linguistics)을 포괄하는 다학제 연구분야를 지칭하는 넓은 의미로 사용된다(Hearst, 1999).

8) 비정형화된 텍스트를 수치화된 데이터로 변환하여 얻는 결과물의 객관성 또는 자의성을 문제로 삼는 반대논리도 있을 수 있으나, 빈도나 코사인 유사성 등의 명확한 기준에 의해 변환하는 것이므로 오히려 연구자의 자의성은 문제될 소지가 적다. 그리고 분석 대상이 되는 텍스

우선 텍스트 데이터는 비구조적 자료이다. 컴퓨터의 발달과 문서 작성 프로그램의 등장으로 과거 수십여 년부터 현재까지 문서는 전자 방식으로 작성되며 저장되고 있다. 컴퓨터에서 작성되는 문서는 컴퓨터가 인식할 수 있도록 일정 범위의 정수값으로 변환하게 되는데 이를 문자 인코딩(character encoding)[9][10]이라 한다. 우리가 문서를 작성할 때 인코딩을 거쳐 컴퓨터 내에서 처리되고 저장되며 컴퓨터는 이를 다시 우리가 인식할 수 있는 문자로 화면이나 종이에 출력해준다. 텍스트 데이터는 컴퓨터 내부에 정수값으로 처리됨과 동시에 우리가 읽을 수 있는 형태로 출력될 수 있는 데이터 형태로 정의할 수 있다.

텍스트가 아닌 구조화된 통계자료의 경우 수학적 연산을 통해 우리가 원하는 분석결과를 비교적 쉽게 얻을 수 있다. 통계자료의 경우 분석에 사용될 목적으로 일정한 약속된 기준에 맞춰 편제되었기 때문이다. 이에 반해 텍스트를 인코딩한 정수 값은 비록 연산이 가능하다 하더라도 정량적 의미를 내포하지 않으므로, 그 자체로 분석을 위한 목적으로 사용할 수는 없다. 전통적으로 텍스트는 우리가 읽고 정보를 얻기 위함으로 사용되었을 뿐, 텍스트를 정보의 원천으로 간주하지는 않았다. 텍스트 마이닝은 연산이 어려운 비구조적 텍스트 데이터를 수치화 또는 벡터화하여 구조화된 자료로 변환시킴으로써 정보의 원천으로 활용하게하는 기술이다.

텍스트 데이터의 또 다른 특징은 모호함과 다의성이다. 일반적으로 언어에서는 한 단어로 표현할 수 있는 개념의 범위에는 한계가 존재한다. 통계자료의 경우 숫자의 크기만으로도 모호함 없이 정보 전달이 가능하나 텍스트의 경우 전체 문맥을 다 읽고 나서야 정확한 의미가 전달되는 경우가 비일비재하다. 언어로 표현하는 개념에는 미묘하고 추상적인 관계가 존재하고 이를 표현하기 위해 수많은 단어 조합이 가능하기 때문이다. 따라서 언어로 표현된 수많은 개념간 관계를

트 샘플이 충분하지 못하고 편향될 경우 계량분석에서 논하는 표본편의(sample bias)는 발생할 우려가 있으나 교차 검증(cross-validation)을 통해 부분적으로 해결할 수 있고, Hamilton *et al.* (2016)이 제안한 SentProp 기법을 이용해서 핵심어(seed words) 선정의 자의성도 해결할 수 있다.

9) 문자 인코딩의 시초는 Samuel Morse가 숫자를 전자적 신호로 조합한 모스 부호(Morse code)이다. 이 후 Alfred Veil이 모스 부호로 알파벳을 조합하였으며, 종이 테이프에 찍힌 전자적 신호를 텍스트로 해석할 수 있게 되었다.

10) 영문 알파벳과 숫자를 표현하는 7비트 이진수 표기법인 아스키(ASCII) 인코딩에서 다중 언어를 지원하는 8비트 표기법인 UTF-8 등이 대표적이다.

규정하고 규정된 관계 하에 단어 사용 유형 등을 파악하기는 쉽지 않다. 이뿐 아니라 동의어, 동음이의어가 존재하므로 이를 사람이 읽을 경우 큰 문제가 없지만 컴퓨터가 읽을 경우 동의어를 다른 의미로 인식하거나 동음이의어를 같은 의미로 인식할 수도 있다.

또한 텍스트 데이터는 차원(dimension)이 매우 높다. 각각의 단어가 하나의 자질(feature)만 나타낸다고 가정하더라고 문서에 사용된 단어의 수를 감안하면 텍스트 데이터의 차원은 수백만 또는 수천만을 크게 상회한다. 소프트웨어와 하드웨어 기술의 발전으로 다차원 자료 처리 속도가 매우 빨라졌다고는 해도, 차원의 저주(curse of dimensionality)가 분명 존재하여 텍스트 데이터의 분석을 어렵게 한다. 또한 차원이 높은 텍스트 데이터는 변수간 관계를 그래프 등으로 시각화하기 어렵다는 문제도 있다.

Ⅲ. 텍스트 마이닝 방법론

여기에서는 텍스트 분석을 위한 방법론에 대해 상세히 기술한다. 자료를 정리하는 전처리와 전처리된 텍스트를 수학적 표현으로 변환하는 단어표상은 텍스트 분석을 위해 거쳐야 하는 순차적 작업이다. 이후 이에 여러 분석기법을 적용하여 텍스트를 분석할 수 있다. 머신러닝의 로지스틱 회귀분석, SVM(support vector machine), kNN(k-th nearest neighbor) 등을 적용하여 문서를 특성별로 분류할 수 있고, 토픽모형을 적용하여 문서들의 특징적인 주제를 추출할 수 있다. 감성사전(sentiment lexicon)이 마련되어있다면 유사한 주제를 다룬 문서에 내재된 어조를 추출하여 유용한 정보를 포함하는 시계열을 추출해낼 수도 있다.

1. 전처리(Preprocessing)

1.1 토큰화(tokenization) 및 불용어(stop words)

많은 자연어 처리 알고리즘이 단어 수준의 텍스트 분석에 근거한다. 단어를

식별해내는 것을 토큰화(tokenization)라고 한다. 이는 텍스트가 주어졌을 때 이를 하나의 유용한 의미 단위들로 분해하는 과정이다.[11] 영문에서는 조사가 없으므로 토큰화는 단어를 기준으로 이루어진다. 문서에서 사용된 띄어쓰기, 탭, 줄바꿈, 쉼표, 마침표, 콜론, 세미콜론, 물음표, 느낌표, 괄호 등이 나타나면 그 기호 또는 문자를 중심으로 텍스트를 토큰화 할 수 있다.

그러나 토큰화된 모든 문자열을 분석에 사용할 수 있는 것은 아니다. 언어에는 자주 사용되지만 별 의미가 없이 관용적으로 사용되는 단어도 상당수 존재한다. 영문에서는 관사 'the'가 대표적인데, 브라운대의 현대 표준 말뭉치(Brown University Standard Corpus of Present-Day American English)에 제시된 500개 샘플(총 1만여 개 단어)에서 'the'는 비중이 7%에 달한다고 한다. 관사뿐만 아니라 전치사도 문법에 의해 항상 자주 등장하나 그 자체로 뜻을 나타낸다고 볼 수는 없다. 이러한 단어는 토큰화 이후에도 분석에 사용될 수 없으며, 앞서 언급한 차원의 부담을 줄이기 위해서도 제거하는 것이 분석의 효율을 높일 수 있는 방법이다. 이런 단어를 불용어[12](stop words)라고 하며 보통 프로그램 라이브러리나 패키지에 포함되어 전처리 과정에서 제거할 수 있도록 되어 있다.

1.2 어간추출(stemming)과 표제어추출(lemmatization)

어간추출은 단어의 접사 등을 제거하고 어간을 분리[13]해 내는 작업으로 관련

---

11) 하나의 의미를 나타내는 일련의 문자열은 영문에서는 단어가 되겠지만 한글에서는 조사가 붙게 되므로 이를 제외한 어근 등의 형태소가 된다. 형태소는 일정한 의미가 있는 가장 작은 말의 단위로 한글 텍스트를 분석할 때에는 형태소를 기준으로 분석하므로 영문에 비해 훨씬 복잡한 알고리즘으로 토큰화 작업이 필요하다.

12) 지프의 법칙(Zipf's law)에 따르면 텍스트에 나타난 출현 빈도에 따라 단어를 순서대로 정렬하면, 해당 단어의 사용빈도가 순위와 역의 상관관계를 갖는다.

$$f(k;s,N)=\frac{\frac{1}{k^n}}{\sum_{n=1}^{N}\frac{1}{n^s}}$$

여기에서 N은 총단어의 수, k는 특정 단어의 순위, s는 분포의 특징을 나타내는 지수값이다. 따라서 빈번하게 등장하는 단어일수록 내재된 정보의 가치는 희박하다고 볼 수 있다.

13) 어근(語根)은 단어를 분석할 때 실질적 의미를 나타내는 중심이 되는 부분으로 단어의 가장 중심이 되는 형태소를 의미한다. 어간(語幹)은 어미(語尾)에 대응되는 말로 활용어가 활용할

단어들이 일정하게 어간으로 매핑(mapping)하는 과정이다. 어간추출은 형태론과 정보 검색 등 분야에서 시작되었으며 1960년대부터 컴퓨터를 활용한 어간추출 알고리즘[14]으로 연구되었다. 영문의 경우 동사와 형용사에 접두사(prefix)나 접미사(suffix)가 붙어 수많은 단어가 파생될 수 있다.[15] 특히 한글에서는 동사, 형용사, 서술격 조사와 같은 활용어의 경우 어미에 따라 동일 어간이 다른 단어로 인식될 수 있다. 텍스트 분석의 효율성 및 정확도 등을 제고하기 위해 이러한 변형된 단어는 하나의 문자열로 간주하는 것이 바람직하며 어간추출은 텍스트의 전처리에서 일반적으로 거치는 과정이다.

표제어추출(lemmatization)은 어간 추출과 같이 변형된 단어를 기준이 되는 단어의 원형(representing lexemes)으로 되돌리는 기법이다.[16] 원형(lemma)이라고 함은 사전에 등장하는 표제어(headword)를 의미한다. 예를 들어 ‘is’, ‘was’, ‘are’, ‘were’의 경우 사전에서는 검색하는 표제어는 ‘be’이며 이 경우 ‘be’가 원형이 된다. 어간추출이 단어의 의미적 단위(semantic unit)를 고려하여 접사 등을 제거하는 기계적 방법이라면 표제어추출은 형태소(morpheme) 분석으로 머신러닝 등을 통해 단어 분석의 정확도를 높이고 있다.[17] 따라서 어간추출에 비해 과정이 좀 더 복잡하며, 표제어추출을 위해서는 다음에 서술하는 품사 태깅 과정이 선행되어야 한다.

때 변하지 않는 부분이며 어근 자체가 어간이 되기도 하고 다른 말과 합쳐져서 어간이 되기도 한다.

14) 우리가 흔히 접하는 검색 엔진의 경우 어간추출을 통해 동일한 어간을 지닌 단어는 동의어로 취급한다.

15) 현재까지 널리 사용하는 Porter stemmer의 경우 접미사를 제거하는 방식으로 어간을 추출하는 알고리즘이다(Porter, 1980).

16) lemmatization을 표제어추출이라고 번역하였지만, 원래 의미는 언어학(linguistics)에서 사전을 편찬할 때 표제어를 선정하는 과정을 의미한다. 텍스트 데이터의 전처리 과정으로서 lemmatization은 변형된 단어를 기존 사전의 표제어로 되돌리는 작업을 의미하므로 원래의 의미와는 다르게 사용되고 있다. 자세한 내용은 https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/lemmatiza tion.html을 참조하기 바란다.

17) 다양한 언어에 대해 어간추출 기법과 표제어추출 기법의 성능을 비교 분석하는 연구가 많이 진행되어 있다(Kettunen *et al.*, 2005; Tala, 2003 등).

1.3 품사태깅(POS tagging) 및 형태소 분석(morpheme analyzing)

품사(part-of-speech, POS)란 단어를 문법적 기능에 따라 구분한 것이며[18] 품사 태깅은 텍스트에 나타난 단어를 해당되는 품사로 꼬리표를 달아주는 과정을 뜻한다. 품사는 단어의 기능이나 형태에 따라 몇 가지 품사를 가질 수 있으므로 품사태깅은 이런 모호함을 해소하는 과정이기도 하다(Kroeger, 2005). 영문의 경우 하나의 문장은 단어의 배열이므로 띄어쓰기 등을 기준으로 토큰화한 단어에 직접 품사를 부여하게 되지만, 한국어의 경우 어절 또는 형태소의 조합으로써 한 단어에 다수의 형태소(morpheme)가 포함된 경우가 많다. 또한 결합된 형태소간 의존성을 지니고 있어 통계적 또는 머신러닝(machine learning based) 방법을 사용함으로써 품사태깅의 정밀도를 향상시킬 수 있다(신중호 외, 1994). 따라서 한글 텍스트 분석에서는 형태소 분석이 품사태깅에 우선하여 이루어져야 한다.

형태소 분석은 단어를 구성하는 각각의 형태소들을 인식하고 용언의 활용, 불규칙 활용이나 축약, 탈락현상이 일어난 형태소를 원형으로 복원하는 과정을 의미한다(강승식, 2002). 형태소 분석은 언어학적으로 언어의 생성과정을 설명하는 용도로 사용되나 전산언어학에서는 정보검색이나 텍스트 데이터의 전처리 용도로 사용될 수 있다(송민, 2017). 한글 텍스트 분석에는 형태소 분석기라 하여 텍스트를 형태소 단위로 분석하고 품사를 함께 출력해주거나 특정 품사에 해당하는 형태소만 선별해주는 패키지들이 있다.[19] 아래는 형태소 분석기로 “한국은행이 12일 금융통화위원회(금통위) 회의를 열고 기준금리를 현행 연 1.50%로 동결했다.”라는 문장을 형태소 분석하고 품사태깅을 마친 결과이다.

> 한국은행/NNG, 이/JKS, 12/SN, 일/NNG, 금융통화위원회/NNG, 금통위/NNG, 회의/NNG, 를/JKO, 열/VV, 고/EC, 기준금리/NNG, 를/JKO, 현행

---

18) 한글의 경우 명사, 대명사, 수사, 조사, 동사, 형용사, 관형사, 부사, 감탄사의 아홉가지 품사가 있다.

19) 서울대학교 IDS(Intelligent Data Systems) 연구실에서 자연어 처리를 위해 개발한 KKMA(Kind Korean Morpheme Analyzer)를 비롯하여 Mecab, Komoran, Hannanum(한나눔) 등이 있다. Python을 이용하여 텍스트 데이터를 분석할 경우 KoNLPy(박은정, 조성준, 2014) 패키지를 이용하면 이들 형태소 분석기를 모두 사용할 수 있다.

/NNG, 연/NNG, 1/SN, ./SW, 50/SN, %/SW, 로/JKB, 동결/NNG, 했/XSV, 다/EC (Lee, Kim and Park, 2019a)

결과물은 형태소에 품사태그가 딸린 형태로 출력된다. 품사태그는 <부록>의 세종품사태그에 나온 예시와 같이 정의되어 있는데, 형태소 분석기마다 조금 다르게 정의하는 경우도 있으나 세종품사태그와 유사하게 정의하고 있으므로 큰 차이는 없다.

전처리시 품사태깅을 하면 분석에 필요한 품사만 한정하여 데이터의 차원을 줄일 수 있으므로 더욱 유용하다. 기본적으로 텍스트 데이터는 우리가 매일 사용하는 자연어(natural language)를 분석대상으로 하며 말의 어휘만큼 어느 빅데이터보다도 더욱 큰 차원을 지닐 수 있으므로 이들을 모두 분석대상으로 삼을 경우 매우 비효율적이다. 따라서 명사만 분석대상을 삼는 경우도 있으며, 감성분석에서는 형용사, 부사, 동사 등을 함께 분석하는 등 데이터의 차원을 낮추고 분석의 효율성을 높일 수 있다.

1.4 N-gram

하나의 단어만으로 형태소를 분석하기에는 모호함이 있을뿐더러 그 뜻도 사용하기에 따라 다르다. 예를 들면 '감기는'이라는 단어는 질병 감기를 의미할 수도 있고 줄이 감기는 현상을 설명하는 말이 될 수도 있다(송민, 2017). 그 뜻도 모호하지만, 어떤 뜻으로 사용되었는지에 따라 형태소도 다르게 분석되고 품사도 달리 태깅(tagging)될 수 있다. 또한 여러 단어가 함께 나타나는 경우 긍정과 부정의 어조가 다르게 인식되기도 한다. 예를 들면 'recovery'는 경제가 회복한다는 긍정적인 뜻을 담고 있으나, 'sluggish recovery'라고 쓰면 회복이 더딘 경기 침체를 뜻하는 부정적 뜻을 담게 된다(Lee, Kim and Park, 2018). 따라서 하나의 단어 단위(uni-gram)로 분석하는 것보다 여러 단어(또는 형태소)를 묶어 하나의 단위(n-gram)로 분석하는 것이 분석의 정도를 제고할 수 있다.[20]

---

20) 하나의 단위만 사용하면 단어 단위 분석과 같으며 이를 uni-gram, 두 단위의 경우 bi-gram이라 하며, 단위의 수를 'n'으로 표기하여 몇 단위까지 묶어서 분석하였는지를 나타낸다. Lee,

일반적으로 n-gram 모형은 n개의 단어나 형태소 단위로 말의 집합(코퍼스)을 구성한다는 것인데, 이 경우 분석 모형의 정도를 제고할 수 있는 반면 단어조합(window)에 벗어나는 경우 분석 대상으로 인식이 되지 않을 수 있다는 단점이 있다. 따라서 n-gram으로 출현빈도를 계산하거나 신경망 등을 학습시킬 때 n-gram 모형이란 unigram과 n-gram 사이에 모든 단어조합을 고려하는 것이 타당하다. 한편 n의 크기를 증대시킬 경우 더 많은 조합으로 학습 및 분석이 가능할 것이나, 차원이 기하급수적으로 증대되는 문제가 있다.[21] 따라서 이 경우에도 품사를 제한하여 n-gram을 구성하는 것이 차원을 줄이는 합리적인 방안이 될 수 있다.

2. 단어표상(Word Representation)

전처리를 통해 필요한 텍스트 데이터만 수집한 후 얻게 되는 단어집합(corpus)으로 본격적인 텍스트 분석을 시작하게 된다. 전처리에서까지는 텍스트가 유니코드와 같은 문자열(string)이나 단순 텍스트(plain text) 형태로 처리되고 있었다. 그러나 컴퓨터 알고리즘으로 단어의 의미, 관계 등을 분석하기 위해 우리가 읽을 수 있는 텍스트가 아닌, 컴퓨터가 이해할 수 있는 형식으로 표현할 필요가 있다. 전처리된 단어의 집합을 확률적 또는 비확률적 방법으로 벡터공간에 표현하는 과정을 단어표상(word representation)이라 하며, 단어표상으로 텍스트를 수치적으로 표현한 후에야 연산을 통한 정보 분석이 가능하다.

2.1 빈도를 기준으로한 단어의 표현

빈도기준으로 단어를 표현할 경우 사람의 눈으로도 셀 수 있는 출현 빈도 등

Kim, and Park(2019a)은 금융통화위원회 의사록을 5-gram 분석을 하였으며, Picault and Renault(2017)는 ECB의 introductory statements를 10-gram으로 분석하였다. 일반적으로 n의 크기가 클수록 분석의 정확도는 향상될 수 있으나, 동시에 차원이 높아지므로 복잡한 모형을 활용한 분석에는 적합하지 않을 수도 있다.

21) 이러한 trade-off 문제로 최대한 n=5로 설정하는 것이 권장된다(유원준, 2019). Lee, Kim, and Park(2019a)는 이에 대한 구체적 사례를 제시하여주고 있다.

단어의 쓰임새를 통해 텍스트의 정보를 추출하는 방법이다. 빈도기준 단어표현은 비확률적(deterministic) 과정이며 표현 결과 나타나는 행렬에서 유용한 정보를 추출하기 위해 주성분 분석(principal component analysis) 등 추가적 연산을 필요로 한다.

2.1.1 빈도 행렬(TDM: Term Document Matrix)

빈도 행렬[22]은 전처리로 거친 텍스트의 토큰이 각 문서에 등장하는 빈도로 구성된 행렬이다. 예를 들어 간단한 두 문장이 있다고 가정해보자.

D_1: “그는 은행 직원이다. 그녀도 은행 직원이다.”
D_2: “나는 한국은행 직원이다.”

위 두 문장을 전처리하면 [그/NP, 그녀/NP, 나/NP, 는/JKS, 도/JKS, 은행/NNG, 한국은행/NNP, 직원/NNG, 이/XR, 다/EF]를 얻을 수 있다. 이를 빈도 행렬도 나타내면 아래 <표 1>와 같다.[23]

〈표 1〉 빈도행렬 예시

	그	그녀	나	는	도	은행	한국은행	직원	이	다
D_1	1	1	0	1	1	2	0	2	2	2
D_2	0	0	1	1	0	0	1	1	1	1

<표 2>에 예시된 행렬의 각각의 열은 특정 단어의 출현 빈도를 나타내는 벡터

22) TDM(Term-Document Matrix) 또는 DTM(Document-Term Matrix)라고도 한다. 행렬의 행이 term을 나타내는 경우 TDM이라고 하고 행이 문서를 나타내는 경우 TDM이라고 하여 구분한다.

23) 일반적으로 행렬의 차원을 최소화하고 분석의 효율성을 제고하기 위해 의미를 내포하지 않는 조사, 어미, 접사 등은 제외하게 되나, 여기에서는 예시를 위해 조사를 포함하여 빈도행렬을 구성하였다.

이다. 예를 들어 '직원'에 대한 빈도 벡터(count vector)는 [2, 1]이라 할 수 있다.

빈도 행렬은 수백만 건에 달하는 문서에서 추출한 단어 혹은 토큰을 기준으로 구성되므로 행렬이 매우 크고 원소중 '0'이 많이 나타나는 희소행렬(sparse matrix)이 된다. 따라서 의미가 있는 명사, 동사, 형용사 등 특정 품사만 사용하는 등 분석의 효율을 제고할 필요가 있다. 빈도행렬의 일반적 구조는 <그림 1>과 같은 희소행렬(sparse matrix)로 이를 활용하여 분석할 때에는 차원축소 등의 방법을 사용하여야 한다.

〈그림 1〉 빈도 행렬(TDM)의 구조

	문서1	문서2	문서3	문서4	문서5	문서6	문서7	문서8
term(s) 1	0	3	0	0	0	0	0	0
term(s) 2	2	1	3	0	5	1	0	0
term(s) 3	0	2	0	0	0	0	0	2
term(s) 4	0	0	0	0	0	1	0	0
term(s) 5	1	0	4	0	0	0	1	0
term(s) 6	0	0	0	0	7	0	0	2
term(s) 7	0	3	0	2	0	0	2	0
term(s) 8	0	0	0	2	0	2	0	0

2.1.2 TF-IDF(Term Frequency-Inverse Document Frequency)

빈도 행렬이 문서에 등장하는 단순 빈도를 측정하는 반면 TF-IDF는 빈도에 상대적 중요도에 따라 가중치를 부여하는 방식이다(Salton and McGill, 1983). 어떤 문서가 특정 주제를 다룰 경우 해당 주제와 관련된 단어[24] 또는 단어의 조합이 자주 등장할 것이다. 이를 TF(Term Frequency)라 한다. 또한 빈도가 높은 단어라도 모든 문서에 흔히 등장하는 경우에는 낮은 가중치를 부여(penalizing)할 수 있다. 이를 IDF(Inverse Term Frequency)라 한다. 예를 들어 경제와 관련된 문헌들

24) 이해를 돕기 위해 단어라고 지칭하고 있으나 특정 형태소가 될 수도 있고 n-gram이 될 수 있다.

이라면 '경기'또는 '금리' 등 경제 관련 단어가 자주 등장할 것이다. 한편 경제 관련 문헌을 분석할 때 '경제'라는 단어는 공통적으로 흔히 등장할 수 있으므로 이에 대한 단순 빈도만으로 얻을 수 있는 정보는 제한되어있다. 반면 '경기 한파'와 같은 bigram은 흔히 등장하는 단어 조합이 아니므로, 해당 조합이 등장하는 문서의 성격은 부정적 어조를 내포하고 있을 가능성이 매우 높다. 따라서 '경기 한파'라는 bigram이 등장할 경우 '경제'라는 단어에 비해 얻을 수 있는 정보가 더욱 많으며 이에 따라 '경기 한파'의 가중치를 높게 책정하는 것이 타당하다.

TF-IDF를 계산하는 방식은 다양하다. 총 N건의 문서 중 어떠한 문서(D)에서 특정 단어(t)가 f번 출현하였고 N건의 문서 중 t가 출현한 문서의 수가 n이라 하면 TF-IDF는 아래와 같이 산출한다.

$$TF-IDF = \frac{f}{T}\log\left(\frac{N}{n}\right) \tag{1}$$

여기에서 T는 표준화 인자(normalizing factor)로 한 단어가 습관적으로 자주 등장하는 경우 이를 표준화하는 역할을 한다. $\log(N/n)$이 가중치이며 t가 출현한 문서의 수가 적을수록 가중치가 증대됨을 알 수 있다. 위 빈도 행렬의 예에서 T '가 문서의 총 단어수라고 하면, 직원'은 D_1과 D_2빈도가 각각 2, 1로 높은 편이라 할 수 있다. 그러나 D_1에서 '직원'의 TF-IDF는 아래와 같다.

$$TF-IDF_{D_1,\,직원} = \frac{2}{12}\log\left(\frac{2}{2}\right) = 0.167 \times 0 = 0 \tag{2}$$

모든 문서에 등장하는 '직원'이라는 단어의 출현에서 얻을 수 있는 정보가 매우 제한적이므로 TF-IDF가 0이 됨을 알 수 있다. 반면 '한국은행'의 TF-IDF는 아래와 같으며, '직원'과 '한국은행'이 동일 문서내 빈도는 0.167로 동일하지만 '직원'은 두 문서에 모두 등장하므로 TF-IDF 기준에 따르면 '한국은행'이 내포한 정보가 더 유의하다는 것이다.

$$TF-IDF_{D_2,\text{한국은행}} = \frac{1}{6}\log\left(\frac{2}{1}\right) = 0.167 \times 0.301 = 0.05 \quad (3)$$

2.1.3 동반출현행렬(Co-occurrence Matrix)

'딸기'와 '바나나'는 대표적인 과일이며 음료를 만들 때에도 종종 함께 사용되기도 한다. 따라서 '딸기'라는 단어가 사용된 한 문서에서는 어딘가 '바나나'도 함께 등장할 것으로 기대할 수 있다. 둘의 관계는 과일이라는 범주에 속할 뿐만 아니라 아이들이 좋아하며, 음료를 만들 때에도 함께 사용되는 등 공통점이 많다. 따라서 '딸기'와 '바나나'를 표현하는 벡터는 서로 매우 유사한 형태를 지닐 것으로 상상할 수 있다. 이렇게 유사성에서 밀접한 두 단어는 동반출현행렬을 분석함으로써 벡터의 형태로 표현할 수 있다.

동반출현행렬은 일정한 크기 m의 윈도우(window)에 따라 규정되는 텍스트 혹은 말뭉치(corpus)의 부분집합이 있을 때, 두 단어가 함께 출현하는 부분집합의 수(동반출현빈도)를 행렬에 기록하는 것이다. 여기에서 윈도우는 계량기법에서 rolling windows와 같이 지정된 계산 범위를 의미한다. $m = 10$인 경우 한 단어를 기준으로 앞 다섯 단어와 뒤 다섯 단어를 포함한 윈도우 내에서 동반출현빈도를 계산하는 것이 된다. 이 부분집합에 포함되는 단어를 context words라고 한다. window의 크기 $m = 10$인 부분집합에서 context words는 임의의 단어를 기준으로 앞 다섯 단어와 뒤 다섯 단어가 된다. 앞의 예에서는 행렬의 (딸기, 바나나) 혹은 (바나나, 딸기) 원소가 m크기의 윈도우에서 동시에 출현하는 빈도를 나타낸다. 동반출현행렬은 그 자체로는 유용한 정보를 얻거나 분석하기에 용이하지도 않을뿐더러 행렬의 차원이 매우 높은데다 희소행렬(sparse matrix)이므로 계산이 용이하지 않다. 따라서 주성분 분해와 같은 차원축소 방법으로 행렬을 변환하여야 필요한 정보를 분석해 낼 수 있다.

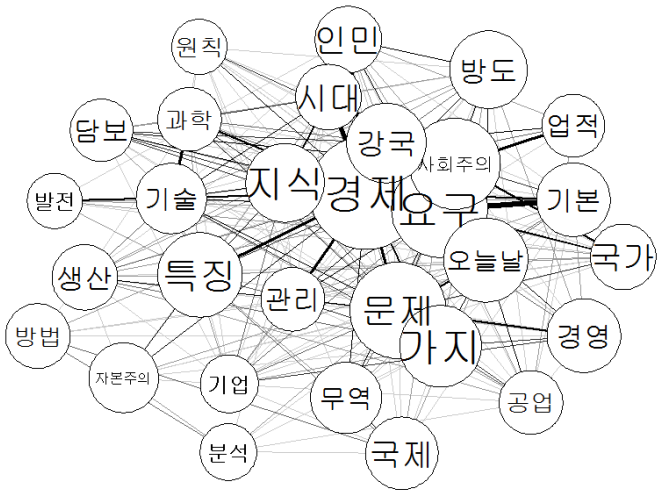
2.1.4 단어빈도의 시각화

문서의 전처리 과정을 거쳐 단어의 품사별로 빈도를 계산하게 되면, 이를 활용하여 다양한 시각화를 할 수 있다. 이러한 시각화 과정은 문서의 주제, 특징

등을 보다 직관적으로 표현할 수 있어 보편적으로 사용되는 방법이다. 빈도를 기준으로 그래프를 그릴 수도 있겠으나, 워드 클라우드(word clouds)는 빈도에 비례하여 단어의 크기를 나타내어 말뭉치를 시각화하는 방법으로 문서의 특징을

〈그림 2〉 금통위 의사록 워드 클라우드(Lee, Kim, and Park, 2019a)

〈그림 3〉 단어들의 동시 출현 관계 다이어그램 예시



보다 직관적으로 나타내준다. <그림 2>는 2005~2017년 중 한국은행 금융통화위원회 의사록에 나타난 단어중 명사만을 추출하여 그린 워드 클라우드이다.

시각화 방법은 워드 클라우드와 같이 단어의 빈도에 비례하는 시각화 외에도 단어간 동시출현빈도를 그림으로 나타내는 방법도 있다. 빈도행렬(TDM)을 D라고 할 때 DD^T는 공분산행렬을 구하는 것과 유사하게 단어간 출현 관계를 나타내준다. 이를 그림으로 도식하면 <그림 3>과 같이 단어의 동시출현관계를 하나의 다이어그램으로 나타낼 수 있다.

2.2 확률적 단어표상(Distributional Representation)

확률적 단어표상 또는 단어의 확률적 벡터화 표현(word embedding as a vector)은 "유사한 의미를 지닌 단어는 유사한 분포를 가진다"는 언어학의 확률적 분포 가정(distributional hypothesis)에 기반한다. 즉 '사과'와 '배'는 껍질을 벗겨 먹기도 하는 과일인데, "저녁은 먹었는데, 아직 허전하네. ____깎아 먹을까?"라는 문장 빈칸에 '사과', '배'가 모두 들어갈 수 있다. 여기에서 언어의 확률적 가정에 따르면 '사과'와 '배' 등은 보다 유사한 단어로 판단한다.

구체적으로 확률적 단어표상은 신경망(neural network)을 사용하여 단어를 힐베르트 공간[25]에 위치한 벡터로 변환한다. 초기 신경망 언어모형[26]은 연산량이 매우 많은 모형이었다. NNLM(neural net language model)[27]에서 시작하여 RNNLM (recurrent neural net language model) 등으로 개선이 이루어지며 연산시간을 축소시켰지만 여전히 효율성은 낮은 수준이었다. 이후 구글의 Mikolov *et al.*(2013)가 연산량을 획기적으로 줄이는 반면 정확도도 높일 수 있는 word2vec을 개발하여 확률적 단어표상 기법의 표준이 되었다.

---

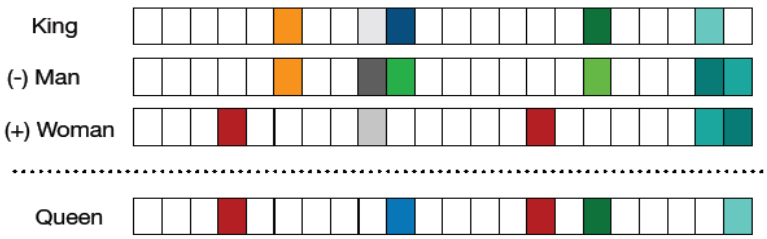
25) 유클리드 공간을 일반화한 추상적 벡터 공간으로 벡터의 내적과 norm 등을 정의할 수 있음.

26) 신경망 언어모형은 Rumelhart *et al.*(1986)에서 시작되었다고 볼 수 있다. Rumelhart *et al.*(1986)에서는 input layer와 output layer만 존재하는 간단한 신경망이다. 오늘날 사용되는 신경망은 input과 output 사이에 다층의 hidden layer가 존재하며 이들 신경망을 추정하는 과정을 학습(learning)이라 한다.

27) Bengio *et al.*(2003)은 n-1개의 단어들로 n번째 단어를 예측하는 n-gram 모델을 제시하였다. 이때 단어의 예측은 출력 가능한 각 단어들에 대한 조건부 확률로 표현된다.

단어의 의미를 힐베르트 공간의 정규화된 벡터로 표현할 경우 내적 또는 cosine 값으로 단어간 유사도[28]를 측정할 수도 있다. 또한 단순 벡터간 단순 가감할 경우 나타나는 벡터는 유사어 또는 반의어가 됨을 확인할 수 있다. 예를 들어 '여왕(Queen)', '왕(King)', '남자(Man)', '여자(Woman)'를 나타내는 벡터를 각각 Q, K, M, W라 하자. 확률적 단어표상에서는 <그림 4>와 같은 $K-M+W=Q$ 관계가 성립한다. 아래에서는 확률적 단어표상의 대표적 방법인 word2vec의 Skip-gram과 CBOW에 대해 알아본다.

<그림 4> 확률적 단어표상의 예시(Young *et al.*, 2017)



2.2.1 Skip-gram과 CBOW

CBOW(Continuous Bag-of-Words)와 Skip-gram은 word2vec에서 신경망을 학습시키는 방법이다. CBOW는 단어 집합이 주어졌을 때 특정 단어(target words)를 예측할 수 있는 확률 과정을 신경망을 통해 학습하는 방법이다. 반면 Skip-gram은 특정 단어가 주어질 때 주변 단어를 예측하는 방법으로 단어표상을 학습한다. CBOW와 Skip-gram이 유사한 방법으로 단어 벡터를 학습하게 되나, CBOW의 경우 한 단어에 대해 한 번의 학습만 가능한 반면, Skip-gram의 경우 윈도우 크

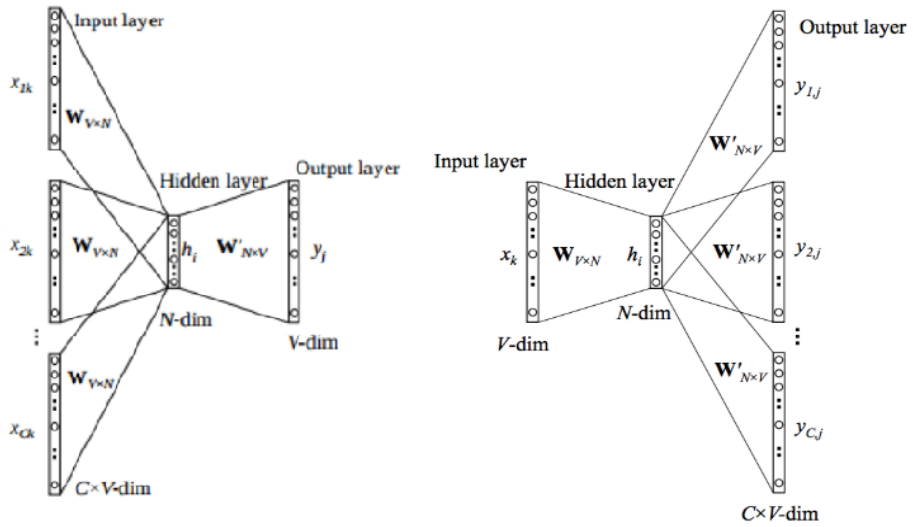
28) cosine 유사도는 벡터간 방향의 일치성을 측정함으로써 단어의미의 일치성을 보여준다. 두 벡터의 cosine 유사도는 아래와 같이 측정할 수 있다. 즉 두 벡터의 내적을 두 벡터 크기의 곱으로 나눈 것이다. (‖는 벡터의 norm) 벡터가 정규화되어 있을 경우 cosine 유사도는 두 벡터의 내적으로 표현된다.

$$\cos = \frac{u^T v}{\|u\| \, \|v\|}$$

기만큼 반복적 학습이 가능하므로 Skip-gram의 효율성이 더 높은 것으로 알려져 있다. <그림 5>은 CBOW와 Skip-gram의 흐름의 차이를 시각화하고 있다. <그림 5>의 왼쪽은 CBOW의 flow를 나타내며 주변단어 벡터를 원-핫 벡터(one-hot vector)[29] 형태로 입력해주면 신경망을 거쳐 중심단어(center word)를 예측하는 모형이다. <그림 5>의 우측은 Skip-gram를 나타내며 중심단어가 주어질 때 그 주변에 쓰일 수 있는 단어를 예측하는 flow이다.

CBOW와 Skip-gram은 신경망(neural network) 학습 과정에서 단어벡터를 추정한다는 점에서 동일하다. 이들은 중층구조(multiple layers)로 이루어진 신경망을 학습하게 되나 은닉층(hidden layer)에 비선형함수를 사용하지 않으므로 실질적으로는 선형모형으로 구성되어 있다. 학습된 결과 얻게 되는 단어벡터는 코퍼스 내에 존재하는 V개의 모든 단어들과의 관계 속에서 해당 단어의 특징 및 의미를 내포하는 N차원의 벡터이다. (<그림 5> 참조)

<그림 5> CBOW와 Skip-gram (Mikolov *et al.*, 2013)



29) 원-핫 벡터 또는 원-핫 인코딩(one-hot encoding)이라고도 하며 텍스트에서 전처리한 단어 집합을 순차적으로 두고 해당되는 단어에만 “1”을 표시하는 방식으로 특정 단어를 나타낸다. 원-핫 벡터는 예측 기반 단어표현에서 입력 벡터 혹은 출력 벡터와 비교하기 위해 사용된다.

학습은 각각의 입력단어에 대해 아래의 softmax 함수[30]를 극대화하는 과정이며 입력단어 벡터(input vector)와 출력단어 벡터(output vector)가 유사할 경우 그 내적값은 극대화하고 유사도가 낮은 단어 벡터간 유사도는 최소화하는 과정이다. 아래 softmax 함수에서 o는 출력단어, c는 입력단어(context vector), u_0와 v_c는 각각 출력 및 입력단어의 확률적 표상이다. 분모는 모든 출력단어와 특정 입력단어 내적 지수값의 합이 된다. 내적은 두 벡터간 거리에 반비례하므로 신경망 예측값의 정확도가 높을 경우 softmax 값은 1에 근사한다.

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \tag{4}$$

2.2.2 회선신경망(Convolutional Neural Network)과 N-gram

우리는 언어사용에서 단어의 결합과 순서의 변화가 매우 중요해지는 경우를 종종 목격한다. 어감이 중립적인 단어가 결합할 경우 긍정 내지 부정적 어감이 되기도 한다. “recovery”와 “staggered”는 “회복”과 “주저하다”라는 중립적인 어감임에도 “staggered recovery”는 부정적 어감을 갖는다. 한편 말의 쓰임에 따라 어떤 단어를 앞에 두느냐에 따라 어감의 차이가 발생하기도 하고 비문이 되기도 하는데, 예를 들어 “가벼운 발걸음”과 “발걸음이 가볍다”는 순서에 관계없이 어감이 비슷하다. 그러나 “가벼운 농담”과 “농담이 가볍다”에서 전자는 분위기를 밝게 하는 농담[31]이라는 긍정적으로 해석될 수 있는 반면, 후자에서처럼 두 단어는 주술관계로는 사용하지 않는 비문이다. 또한 “포용적 제도”와 “제도적 포용”에서 전자는 제도의 성격이 포용적이라는 의미를 담고 있는 반면 후자는 포용이 제도적, 정책적임을 의미한다고 볼 수 있다. 영문에서도 한글의 사례와 유

30) softmax는 정규화 지수함수(normalized exponential function)라고 하며 softmax 함수는 각 단어간 내적의 결과를 확률의 형태로 표현해준다. 즉 학습의 결과인 출력단어 벡터간 내적으로 구성된 V차원의 벡터는 각 원소의 합은 1로 표준화된다.

31) 이와 같이 언어사용 습관으로 하나의 숙어처럼 사용되는 말을 연어(collocation)라 한다(김규훈 외, 2013).

사하게 단어나 표현의 순서는 의미의 차이를 나타나게 하거나 문장성립 여부를 결정하는 중요한 정보를 담고 있다.

Word2vec이 확률적 단어표상에서 매우 널리 사용되기는 하지만 이러한 단어의 순서를 고려하지 않는 단점이 있다. 신경망학습으로 단어의 확률적 벡터로 나타내는 동시에 단어의 순서를 고려하는 방법을 회선신경망(CNN)이라 한다. 앞서 기술한 n-gram 모형과 회선신경망은 이러한 단어의 순서에 따라 어조와 의미가 변화하는 경우에도 이를 구분하여 분석할 수 있는 방법이다.

2.3 특성공학(Feature Engineering)

전처리에서 단어표상으로 넘어가는 과정은 텍스트를 벡터로 변환하는 과정이라 할 수 있다. 텍스트가 벡터로 변환이 되면 이후의 분석은 일반적인 머신러닝이나 딥러닝 모형을 사용하여 진행할 수 있다. 전처리의 결과인 단어 혹은 토큰은 머신러닝의 용어로는 특성(feature)이라고 할 수 있다. 머신러닝과 같은 모형에 사용할 특성을 선정하기에 앞서 성능 향상을 위해 특성공학(feature engineering)이라는 작업을 할 수 있다. 앞서 설명한 빈도에 기반한 TF-IDF, 동반출현 확율기반의 비지도학습의 결과인 word2vec 등도 특성공학의 일종이라 할 수 있다. 이러한 방식은 분석대상 텍스트의 정보만 사용함으로써 이미 알고 있는 도메인 지식을 반영하지 못하는 단점이 존재한다. 사용하는 언어가 일반적인 경우와 차별화되는 바이오의학이나 경제 분야와 같은 경우 그 분야에 특화된 지식을 활용하게 되면 해당 도메인에 최적화된 분석을 할 수 있게 된다. 도메인 지식을 반영하기 위한 특성공학의 방법으로는 온톨로지, 지식그래프 등을 활용하는 방법이 있다. 온톨로지(ontology)는 일관성 있고 명확한 지식공유와 지식통합의 틀을 제공하기 위한 개념모형이다. 온톨로지는 개념에 대해 정의하고 다른 개념과의 관계를 연결하는 역할을 한다. 따라서, 특정 도메인의 온톨로지는 다양한 용어의 해석을 도메인에 적합한 것으로 제한하고 도메인의 구조를 반영함으로써 텍스트 마이닝의 성능을 향상시킬 수 있다(Spasic *et al.*, 2005). Wang *et al.* (2014)는 지식그래프에서 객체(entity)사이의 관계와 텍스트에서 동반출현 정보를 동시에 반영하여 벡터화(embedding)시키는 방법을 제안하였는데, 이들의 방법은 개별적

벡터화 방식에 비해 사실예측능력의 정확성을 향상시키고 특히 지식그래프에 존재하지 않는 객체에 대한 예측능력을 높이는 결과를 보였다. 최근에는 BERT, ELMO, GPT-2나 ULMfit과 같이 대규모 텍스트와 컴퓨팅 능력에 기반한 비지도 학습으로 문맥의 정보를 반영하는 언어모형이 개발되어 사용되고 있다.[32] 이들 모형이 특성공학 없이도 다양한 NLP 작업에서 우수한 성능을 보여주면서 특성공학의 종말이라는 표현도 나오고 있으나 이들 모형은 막대한 자원을 필요로 하기 때문에 여전히 특성공학은 소규모 작업에서는 머신러닝의 성능향상에 필요한 작업이라고 할 수 있다.

3. 머신러닝을 활용한 문서 분류

전처리와 단어표상 과정을 거친 이후에는 텍스트는 정량적, 구조적 자료로 변환된다. 따라서 빅데이터 분석에 활용되는 여러 분석기법을 적용하여 텍스트 자료를 분류하고 분석할 수 있다. 일반적으로 텍스트 데이터는 차원이 높아 머신러닝의 분류기법을 활용한다. 로지스틱 회귀분석, 서포트 벡터 머신(SVM, support vector machine), 나이브 베이즈 분류기(naive Bayesian classifier) 등 지도학습(supervised learning)과 k-mean 군집화(clustering), 비모수 베이지안 군집화(nonparametric Bayesian clustering) 등 비지도학습(unsupervised learning)을 적용하면 문서를 특성별로 분류할 수 있다. 모든 지도학습은 이미 분류된 학습자료(training data)와 검정자료(test data)를 필요로 하는 반면 비지도학습은 기 분류된 학습자료를 필요로 하지않는다. 학습자료를 활용하여 합습(추정)한 뒤 검정자료를 통해 분류(예측)가 올바르게 이루어졌는지 여부를 판단할 수 있다. 아래에서는 문서분류에 활용될 수 있는 지도학습과 비지도학습 기법 대해 알아본다.

---

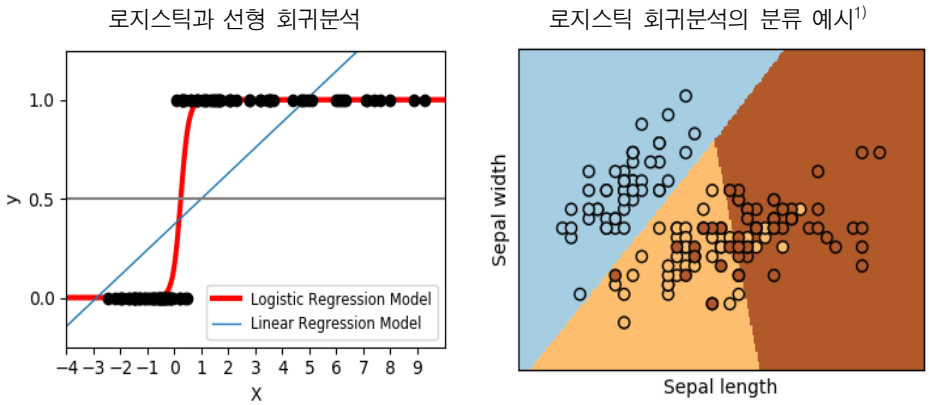
32) 구글의 BERT는 Bidirectional Encoder Representation의 약자, Ai2의 ELMo는 Embeddings from Language Models의 약자, OpenAI의 GPT-2는 Generative Pre-trained Transformer 2의 약자, fast.ai의 ULMfit은 Universal Language Model Fine-tuning의 약자이다.

3.1 로지스틱 회귀분석(Logistic Regression)

로지스틱 회귀(Cox, 1958)는 독립 변수의 선형 결합이 특정 사건 발생 가능성을 설명하거나 예측하는 확률적 모형이다. 텍스트 마이닝에서는 문서에 내재된 임의 특성의 선형결합으로 문서를 특정 범주로 분류[33]하는데 사용될 수 있다. 로지스틱 회귀는 일반 선형회귀 분석과 유사하지만, 종속변수가 [0, 1]의 범위 내에 있다는 점이 다르다. 따라서 값의 범위가 [-∞, ∞]인 독립변수의 변화에 대응시키기 위해 로지스틱 회귀분석은 아래와 같이 로짓 변화(l)된 종속변수와 독립변수 벡터(X), 회귀계수 벡터(β)의 모형으로 표현한다.

$$l(p) = \log \frac{p(y=1|x)}{1-p(y=1|x)} = \beta^T X \tag{5}$$

〈그림 6〉 로지스틱 회귀분석 분류



주 : 1) 자료는 붓꽃은 꽃받침(sepal)의 길이(length)와 넓이(width)를 2차원 평면에 나열한 것이며 로지스틱 회귀 분석에 의해 꽃받침의 길이와 넓이에 따라 세 종류(Virginica, Versicolor, Setosa)로 분류할 수 있다.

자료 : scikit-learn.org

33) 로지스틱 회귀분석은 후술하는 서포트 벡터 분류기법과 함께 타 분류 방법에 비해 분류 정확도가 매우 높은 기법이다.

여기에서 $p(y=1|x)=1/(1+e^{-\beta^T X})$로 표현할 수 있으며 이를 로지스틱 함수[34](logistic function)라 한다. 회귀모형은 실질적인 종속변수(y)가 0 또는 1에 속할 확률에 대한 추정식이며, 추정의 목적은 분류의 기준이 되는 의사결정 경계(decision frontier) 또는 초평면(hyperplane) $\beta^T X$를 구하는 과정이다. 일반적 회귀분석 모형과 달리 결정경계에 멀리 떨어진 샘플이 모수에 과도한 영향을 미치는 것을 방지한다는 점에서 장점이 있다.

3.2 서포트 벡터 머신(SVM, Support Vector Machine)

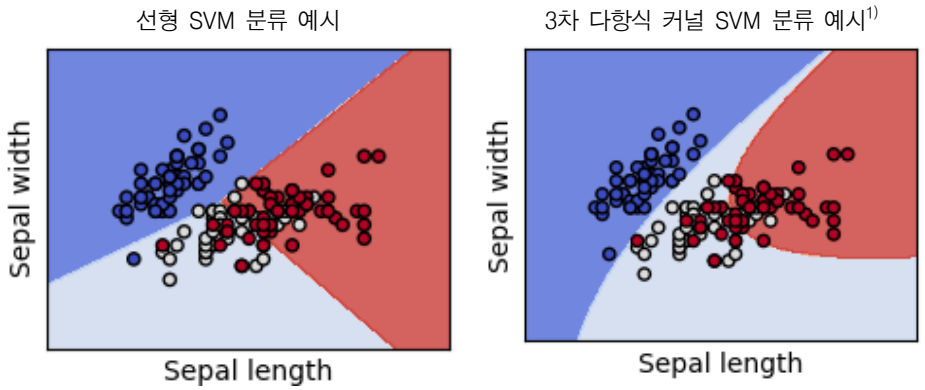
서포트 벡터 머신(Vapnik and Chervonenkis, 1964)의 기본 아이디어는 로지스틱 회귀분석과 유사하게 의사결정 영역(decision boundary)을 찾는 과정이다. 서포트 벡터 머신의 추정은 각 카테고리에 속한 자료가 나열되어있는 벡터 공간에서 가장 넓은 폭의 경계벡터(support vectors)를 지나는 초평면 또는 MHH(Maximum Marginal Hyperplane)를 구하여 데이터를 분류한다. 그러나 데이터를 정확히 분류할 수 있는 초평면은 데이터가 속한 차원에서는 쉽게 구해지지 않는 경우가 많다. 따라서 SVM에서는 데이터를 한 차원 높은 공간에 매핑(mapping)하여 초평면을 구할 수 있다. 그러나 차원과 연산 효율성은 반비례하므로 커널(kernel)함수를 활용하여 현 차원에서 계산한 내적을 더 높은 차원 계산한 내적으로 갈음하고 그것을 통해 초평면을 찾는 커널 트릭(kernel trick)을 사용한다.[35](Grus, 2015)

---

34) 머신러닝 또는 딥러닝(deep learning)에서는 이를 sigmoid 함수라고도 하며 쌍곡 탄젠트 함수(tanh, hyperbolic tangent), 경사함수(ReLU, Rectified Linear Unit) 등과 함께 신경망의 활성함수(activation function)로 활용된다.

35) 대표적인 커널함수 $K(X_i, X_j)$는 다항식 커널: $(X_i X_j+1)^h$, 가우스 방사기저함수(Gaussian radial basis function) 커널: $\exp(-\|X_i X_j\|^2/2\sigma^2)$, 시그모이드(sigmoid) 커널: $\tanh(\kappa X_i X_j-\delta)$ 등이 있다.

〈그림 7〉 서포트 벡터 머신(SVM) 분류



주 : 1) 자료는 붓꽃은 꽃받침(sepal)의 길이(length)와 넓이(width)를 2차원 평면에 나열한 것이며 SVM 분석에 의해 꽃받침의 길이와 넓이에 따라 세 종류(Virginica, Versicolor, Setosa)로 분류할 수 있다.
자료 : scikit-learn.org

MHH를 찾는 방식은 하드 마진(hard margin)과 소프트 마진(soft margin)이 있다. 하드 마진은 분류 오차를 허용하지 않는 엄격한 방식으로 초평면을 찾는 방식이며 데이터에 따라 초평면을 구하지 못하는 경우도 발생한다. 소프트 마진은 미리 정해진 범위 내 오차를 허용하는 범위 내에서 초평면을 구하므로 낮은 차원에서도 분류가 용이하다. 예를 들어 각각의 범주($y = \{-1, 1\}$)로 분류할 데이터의 속성 벡터(X)가 주어졌을 때, 가중치 벡터(W)와 상수(a, b)에 대해 MHH는 아래와 같이 정의된다.

$$y_i(b + WX_i) \geq a, \text{for} \forall i, y_i \in \{-1, 1\} \quad (6)$$

만약 b 값을 적정히 구하게 된다면 초평면의 a와 $-a$와의 거리가 동일하게 되고 y_1는 {-1, 1}의 값을 갖게 되므로, 위 초평면은 데이터를 두 범주로 구분함과 동시에 초평면이 지나는 벡터[36]간 거리가 극대화된다. 만약 소프트 마진을 활용하여 초평면을 찾는다면 아래와 같은 초평면이 구해질 수 있다.

36) 이를 서포트 벡터라 한다

$$y_i(b+WX_i) \geq a-\varepsilon,\ \varepsilon \geq 0 \text{ for} \forall i\ y_i \in \{-1,1\} \tag{7}$$

3.3 나이브 베이즈 분류모형(Naive Bayes Classifier)

나이브 베이즈 분류모형은 베이즈 법칙과 특성간 독립 조건부확률이라는 가정하에 스팸메일분류 등에 널리 사용되는 분류기법이다. 문서 i의 n가지 특성을 나타내는 특성벡터(feature vectors) $x_i = (x_{i1}, ..., x_{in})$와 범주 y가 주어졌을 때 베이즈 법칙에 의해 사전확률 $p(y_i)$와 사후 $p(y_i | x_{i1} \cdots x_{in})$은 아래와 같이 표현된다.

$$p(y_i | x_{i1} \cdots x_{in}) = \frac{p(y_i)\, p(x_{i1} \cdots x_{in} | y)}{p(x_{i1} \cdots x_{\in})} \tag{8}$$

만약 모든 특성이 나타날 조건부 확률이 독립[37]이라면 위 식은 아래와 같이 단순화하여 표현할 수 있다.

$$p(y_i | x_{i1} \cdots x_{in}) = \frac{p(y_i) \prod_{j=1}^{n} p(x_{ij} | y)}{p(x_{i1} \cdots x_{\in})} \tag{9}$$

여기에서 $p(x_{i1} \cdots x_{in})$는 상수이므로

$$p(y_i | x_{i1} \cdots x_{i2}) \propto p(y_i) \prod_{j=1}^{n} p(x_{ij} | y) \tag{10}$$

따라서 위 조건부 사후확률(posterior probability)를 극대화하는 값으로 범주 $\hat{y_i}$를 추정하고 문서를 분류할 수 있다. 나이브 베이즈 분류모형은 문서에 대한 분류 뿐만 아니라 단어 또는 n-gram의 극성 분류에도 응용할 수 있다(Lee, Kim,

37) 범주에 속할 사전확률(prior)이 동일하거나, 각 단어들이 등장할 조건부 확률이 독립적인 전제 하에 분류하므로 “순진함(Naive)”이라는 수식어가 붙는다고 한다.

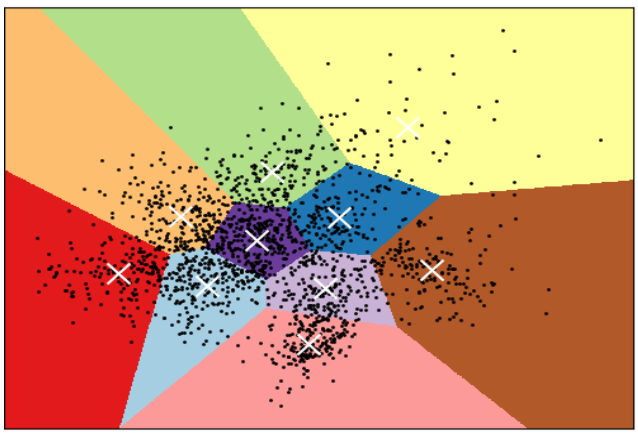
and Park, 2019a).

3.4 k-평균 군집화(K-means Clustering)

k-평균 군집화(McQueen, 1967)는 사전 분류되지 않은 데이터를 k개의 범주로 분류하는 비지도학습기법이다. 주어진 임의의 중심점(centroid)과의 거리에 기반하여 비용함수를 최소화하는 방식으로 데이터를 군집(clustering)하는 알고리즘이며, 추정과정에서 동일 그룹 내 객체 간 유사도(similarity)가 증대되는 반면 타 그룹내 객체와의 유사도는 점차 감소한다. 예를 들면 d건의 문서를 나타내는 벡터집합 $x = \{x_1, \dots, x_d\}$이 주어졌을 때, 아래의 식을 만족하는 $k(\leq d)$개의 집합 $c = \{c_1, \dots, c_k\}$과 중심점 $\mu = \{\mu_1, \dots, \mu_k\}$으로 문서를 분류한다.

$$c = \arg\min \sum_{j=1}^{k} \sum_{x_i \in c_j} \| x_i - \mu_i \| \tag{11}$$

〈그림 8〉 k-평균 군집화 분류



주 : 1) 자료는 숫자를 적은 손글씨의 특성을 나타내는 벡터를 주성분분석(PCA) 기법으로 2채원으로 축소한 후 2차원 평면에 나열한 것이며 k-평균 군집화 분류에 의해 1~9까지의 숫자로 분류해 내고 있다.

자료 : scikit-learn.org

k-평균 군집화 추정은 각 문서벡터를 임의의 중심점에 할당하는 것으로 시작한다. 할당된 범주 내에서 중심점을 다시 계산하여 중심점이 수렴할 때까지 이 과정을 반복하며 문서의 군집화가 이루어지는 방식이다. k-평균 군집화는 직관적으로 이해하기 용이한 알고리즘일 뿐만 아니라 연산부담도 낮은 편이다. 그러나 k-평균 군집화는 비지도학습임에도 불구하고 범주의 수 k를 사전 지정해야하며, 극단치(outlier)에 민감하다는 단점이 있다.

3.5 비모수 베이지안 군집화(Nonparametric Bayesian Clustering)

비모수 베이지안 군집화는 문서의 분포를 임의의 유동적 분포로 가정하고 이 분포에 사전분포를 설정하여 사후추론 하는 과정이다(Noh *et al.*, 2014). 나이브 베이즈 분류모형은 베이즈 법칙과 단순 가정으로 문서를 분류할 수 있는 지도학습이라면 비모수 베이지안 군집화는 계층적 베이지안 군집화 모형을 사용하는 비지도학습 기법이다. 앞서 설명한 분류기법이 문서 분포에 대한 가정 없이 사전 지정된 범주의 수에 따라 거리 등의 측도(metric)에 기반하여 분류하는 기법이라면, 비모수 베이지언 군집화에서는 문서가 임의 분포에 따른다는 가정 하에 통계적으로 추론하여 최적 범주의 수를 유동적으로 추정할 수 있다는 점이 다르다.

비모수 베이지언 군집화에서는 계산의 단순화 등의 이유로 k차 심플렉스 위에 정의되는 이산분포인 디리클레 확률과정(Dirichlet process, DP)이 가장 널리 사용된다. i번째 문서(y_i)가 따르는 임의의 연속 분포 $f(y_i|\theta_i)$의 모수 θ_i는 분포 $G(\theta_i)$에서 추출될 수 있으며 모수의 분포 $G(\theta_i)$는 디리클레 확률과정을 따르는 계층적 모형을 가정한다.[38]

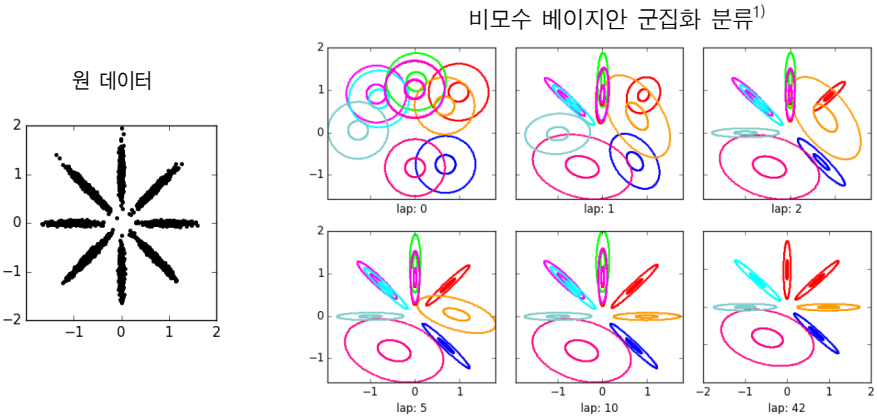
$$y_i|\theta_i \sim F(\theta_i),\ \theta_i \sim G(\theta),\ G(\theta) \sim DP(\alpha, G_0) \tag{12}$$

이를 디리클레 확률과정 혼합모형(Dirichlet process mixture model; DPM)이라고도 부르며 문서 y_i가 따르는 분포 $f(y_i|\theta_i)$ 모수 θ_i의 확률적 동일성 여부로 군

38) 비모수 베이지안 군집화의 추정과 관련된 자세한 도출과정 등은 Noh *et al.*(2014)과 Peter Orbanz의 강의 노트(http://stat.columbia.edu/~porbanz/npb-tutorial.html) 등을 참조

집화가 이루어진다. 따라서 서로 다른 문서 y_i와 y_j가 동일한 모수에 기반한 분포 $f(y|\theta)$에서 추출될 수 있다고 추론될 경우 하나의 범주로 묶이게 된다.

〈그림 9〉 비모수 베이지안 군집화 분류

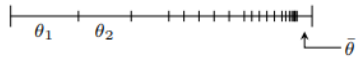


주 : 1) Astericks(*) 형태로 분포하도록 임의 생성된 자료를 비모수 베이지안 군집화 기법으로 사전분포(prior distribution)에서 시작하여 점차 사후분포(posterior distribution)로 추론해가는 과정으로 서로 다른 분포에 속한 점들은 다른 범주로 분류됨

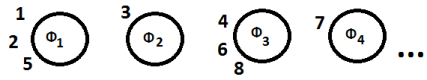
자료 : bnpy documentation

한편 범주의 수가 정해져 있지 않으므로 특성이 매우 다른 문서 y_k가 존재할 경우 새로운 범주를 대표하는 분포 $f(y_k|\theta_k)$가 생성되어 군집화 과정에 유연성을 높일 수 있다.[39] 비모수 베이지언 군집화에서는 이산형 분포인 디리클레 분

39) 이러한 추정과정을 막대 쪼개기 과정(stick-breaking process), 중국음식점 과정(CRP, Chinese restaurant process) 등으로 부른다. 막대 쪼개기 과정은 범주 구분시 데이터를 아래와 같이 하나의 긴 막대기를 쪼개듯 모수 θ_i의 범주에 할당하는 방식이다.



중국음식점 과정은 데이터를 범주에 할당하는 방식을 중국음식점에서 손님들이 테이블에 앉는 과정으로 표현한 것이다. 두 방식 모두 새로운 범주가 생성되는데 패널티가 부여되므로 과적합(overfitting)문제를 최소화하기 위한 군집화 과정이다.



포를 활용하여 연속 분포의 혼합 분포를 구성함으로써 자료가 지닌 특성을 최대한 분포에 반영할 수 있다는 장점이 있는 반면 직관적인 이해가 어렵고 계산과정이 복잡해질 수 있다는 단점이 있다.

3.6 SO-PMI(Semantic Orientation from Pointwise Mutual Information)

SO-PMI는 두 단어의 동반출현확률(PMI)로 단어의 극성을 분류하는 기법이다.(Turney and Littman, 2002) 우선 두 단어 w_1, w_2가 주어졌을 때 w_1, w_2가 한 문서내 동시 출현할 확률을 $p(w_1,w_2)$, w_1, w_2가 출현하는 확률을 각각 $p(w_1)$, $p(w_1)$라고 하면 PMI는 아래와 같이 측정한다.

$$PMI(w_1,w_2)=\log\frac{p(w_1,w_2)}{p(w_1)p(w_2)} \tag{13}$$

만약 두 단어 w_1, w_2가 독립이라면 $p(w_1,w_2)=p(w_1)p(w_2)$가 되어 PMI는 0이 된다. SO-PMI는 긍정, 부정과 같은 극성을 띄는 대표적 단어와의 PMI 크기를 비교하여, 임의의 단어의 극성을 분류한다. 예를 들어 긍정의 극성을 띄는 대표 단어 집합인 {좋다, 멋지다, 훌륭하다, 긍정적이다, 다행이다, 맞다, 우월하다}과 부정적 극성을 띄는 대표 단어 집합 {나쁘다, 볼품없다, 불쌍하다, 부정적이다, 불행하다, 틀렸다, 열등하다}를 w_2로 하여 임의의 w_1과의 PMI를 측정하고 이를 아래와 같이 비교하여 w_1의 극성을 분류한다.

$$SO-PMI=PMI(w_1,\{긍정어\})-PMI(w_1,\{부정어\})$$

4. 토픽 모형(Topic Models)

토픽 모형도 머신러닝 분류기법과 유사하게 데이터의 차원을 축소하거나 이를 바탕으로 데이터를 분류하는 기법이다. 그러나 머신러닝 분류에서는 문서에 내재된 임의의 특징을 기준으로 분류한 반면 토픽 모형에서는 문서에 내재된 주제

를 기준으로 문서의 차원이 축소되고 분류된다. 또한 분류 결과 문서가 지닌 특징적인 주제들에 대한 정보를 구할 수 있다는 장점이 있다.

4.1 잠재의미분석(LSA, Latent Semantic Analysis)

Deerwester *et al.*(1990), Papadimitriou *et al.*(2000) 등은 빈도행렬(TDM) 또는 tf-idf 행렬의 스펙트럼 분석을 통해 문서-의미(document-semantic) 행렬과 의미-단어(semantic-term) 행렬로 분해하는 잠재의미분석[40](LSA)을 제안하였다. 이는 빈도행렬 등을 구성하는 선형공간에서 본 행렬의 특징을 가장 잘 나타내는 부분공간(linear subspace)을 식별하는 과정이다(Blei *et al.*, 2003). 잠재의미분석은 정보검색(information retrieval) 분야에서 정보검색엔진의 효율을 높이는 방법으로 널리 사용되었다. 문서의 특징을 나타내는 빈도행렬 등은 기본적으로 차원의 수가 높아, 정보검색을 위한 데이터베이스에 수 많은 문서정보를 담아두는 것은 비효율을 유발하기 때문이다(Hoffman, 1999). 잠재의미분석은 이후에도 문서의 유사도 등을 측정하는 방법으로도 활용되고 있다.

단어수가 n, 문서 수가 d, 계수가 r인 $n \times d$ 빈도행렬인 A의 특잇값($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$)은 AA^T 또는 A^TA의 고윳값의 제곱근으로 구할 수 있다. 특잇값을 대각원소로 하는 정방행렬을 $D_{r\times r} = diag(\sigma_1,\dots,\sigma_r)$과 AA^T와 A^TA를 각각 고유분해(eigenvalue decomposition)하여 얻은 고유벡터로 구성된 정규직교행렬(column-orthonormal matrix) $U_{n\times r} = (u_1,\dots,u_r)$와 $V_{d\times r} = (v_1,\dots,v_r)$을 이용[41]하면 A는 아래와 같이 특잇값 분해[42]가 가능하다.

---

40) 원문에는 잠재의미지수(latent semantic indexing)로 표기되어 있으나, 오늘날에는 잠재의미분석으로 더 널리 알려져 있다.

41) 여기에서 U와 V의 열벡터들을 각각 left singular vector, right singular vector라 한다.

42) 주성분 분석과 특잇값 분해는 차원축소방법이다. 주성분 분석(PCA, Principal Component Analysis)은 텍스트 마이닝뿐만 아니라 공학과 심리학, 경제학 분야에서 차원축소 방법으로 널리 활용된다. 고차원의 행렬을 하나의 축에 사상(projection)시켰을 때 가장 분산이 커지는 축을 첫 번째 주성분, 두 번째 축을 두 번째 주성분이 되는 방식으로 새로운 좌표계로 행렬을 직교 변환(orthogonalized linear transformation)한다. 직교변환 과정에서 공분산 행렬의 고윳값(eigenvalue)의 크기를 기준으로 고유벡터를 기저로 선형변환할 수도 있고 특잇값을 기준으로 사용하는 경우 이를 특잇값 분해(SVD, Singular Vector Decomposition)라고도 한다.

$$A = UDV^T \tag{14}$$

잠재의미분석은 아래와 같이 빈도행렬의 특징을 가장 잘 나타내는 k개의 특잇값만 선택하여 A를 근사함을 의미한다.

$$A_k = U_k D_k V_k^T \tag{15}$$

잠재의미분석의 수학적 의미는 A의 열벡터가 U_k의 열벡터를 기저로한 벡터공간으로 사영(projection)된 것이다.[43] 따라서 위 잠재의미분석에서 $V_k D_k$의 행이 각각 문서의 특징을 나타낸다고 볼 수 있다. 잠재의미분석의 수학적 의의는 Eckart and Young(1936)의 정리에 따르면 빈도행렬 A를 가장 잘 근사하는 행렬 A_k를 구하는 것이다. 잠재의미분석은 비확률적 과정으로 단어와 문맥 간의 잠재적인 의미(latent/hidden meaning)를 효과적으로 보존할 수 있고, 텍스트 데이터가 지닐 수 있는 노이즈를 제거(Rapp, 2003)하거나 sparsity를 감소(Vozalis and Margaritis, 2003)시켜 문서 간 유사도 측정 모형의 성능향상에 기여한다고 알려져 있다(Deerwester *et al.*, 1990; Landauer and Dumais, 1997).

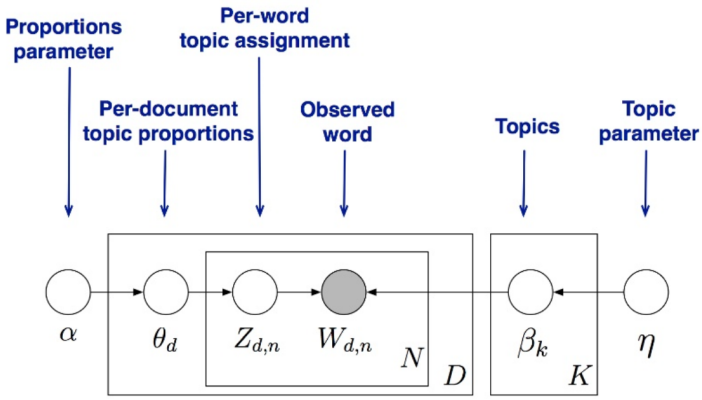
4.2 잠재디리클레할당(LDA, Latent Dirichlet al.location)

문서의 내재적 의미 보존 외에도 유사어 식별 등에도 널리 활용되는 잠재의미분석은 pLSI(probabilistic Latent Semantic Indexing; Hoffman, 1999)로 확률과정을 도입한 분석 방법으로 발전하였다. 이후 Blei *et al.*(2003)은 pLSI에 대한 과적합(overfitting) 문제 등을 제기하며 잠재디리클레할당(이하 LDA)을 통한 문서 주제 분류기법을 제시하였다. 단어의 표상에 확률 과정을 도입한 확률적 의미론(distributional semantics)에 기반한 word2vec과 달리, LDA는 주어진 주제 하에 특정 문서와 그에 사용된 단어가 추출될 수 있는 확률과정을 바탕으로 문서의

43) 특잇값분해는 원행렬의 성질을 가장 잘 반영하는 저차원의 행렬을 구하는 과정이다. 일상적으로 컴퓨터의 이미지 압축(compression) 기술로도 잘 활용되고 있는데, 특잇값 분해를 통해 이미지를 나타내는 행렬 A에서 A_k를 구함으로써 그림의 해상도를 낮출 수 있다.

주제를 분류하는 확률모형이다.

<그림 10> LDA 추론(Blei *et al.*, 2003)



LDA는 D를 말뭉치 전체 문서의 수, K는 전체 토픽의 수, N은 문서d의 단어 수라 할 때 우리가 문서d에 등장한 n번째 단어 $w_{d,n}$를 관찰하고 <그림 5>의 과정으로 토픽의 분포를 추정하는 과정이다. 여기에서 α, β는 연구자가 임의로 지정하거나 중층적 비모수 베이지안(hierarchical non-parametric Bayesian) 모형으로 추정할 hyper parameter이다. <그림 5>의 추정과정은 디리클레 분포[44]에서 임의 추출된 토픽 분포에서 해당 단어 및 문서에서 관찰된 말뭉치가 확률적으로 관찰될 가능성(likelihood)을 가장 높이는 토픽 분포를 찾는 과정이라 할 수 있다. LDA는 문서별 토픽분포와 토픽별 단어분포의 결합분포에 의해 문서에 사용된 단어들이 생성된다고 가정한다. 그러나 우리가 알고자 하는 바는 문서별 토픽분포와 토픽별 단어분포이므로, 관찰된 단어를 바탕으로 이들 결합분포의 확률을 극대화시키는 방향으로 추론하게 된다. <그림 6>은 Blei *et al.*,(2003)이 제시한 LDA 추론과정이다. <표 2>는 2005~2017년 금통위 의사록, 뉴스, 채권 보고서 등을 LDA로 분석한 토픽 비중을 나타낸다.

---

44) 디리클레 분포(Dirichlet distribution)는 베타분포의 일반화된 분포이다. 베이지안 추정에서 다항분포(multinomial distribution)의 사전분포(prior distribution)로 사용된다. 자세한 내용은 Andrew *et al.*(2014) 참조

〈표 2〉 금통위 관련 문서의 토픽 비중 예시(Lee *et al.*, 2019a)

No.	Topic name	Total	Minutes	News	Report
1	Foreign Currency	5.24	11.20	5.94	3.75
2	Financial Policy	2.69	2.24	3.15	1.99
3	Bond Issue Market 1	3.35	0.73	1.29	6.79
4	Monetary Policy	3.81	12.56	4.47	2.20
5	Bond Issue Market 2	2.79	1.32	2.67	3.08
6	Financial Crisis	1.79	1.03	2.09	1.36
7	Swap Market	4.30	3.05	4.02	4.82
8	Inflation	3.32	10.56	2.68	3.89
9	Credit Ratings	1.42	0.38	0.93	2.26
10	Real Estate	1.20	0.15	1.71	0.46

5. 감성사전을 활용한 논조분석

감성분석은 의견분석(opinion mining), 논조분석(tone analysis) 또는 자동감성분석(sentiment AI)이라고도 하며 단어 또는 문서의 극성(polarity) 등을 측정하여 텍스트의 논조를 읽는 방법이다. 흔한 예시로는 대량의 영화감상평이나 상품평에서 긍·부정 논조를 측정하여 하나의 대상에 대한 점수를 계산할 수 있다. 주로 특정 대상이나 주제에 대한 전반적인 감성 도는 논조(tone)를 측정하는데 사용되며 마케팅, 사회관계망(SNS) 서비스, 서베이 응답, 소비자 불만사항 등 다양한 범위에 폭넓게 활용될 수 있는 방법이다.

5.1 감성사전(sentiment lexicons)

Stone *et al.*(1966)은 오늘날 Harvard-IV 감성사전으로 알려져 있는 General Inquirer를 개발함으로써 감성분석의 시초가 되었다. Harvard-IV 감성사전은 심리사전으로도 알려져 있는데 이는 언어를 통한 심리학적 분석에 널리 활용되었기 때문이다. Harvard-IV 감성사전은 현재에도 감성분석의 벤치마크 사전으로 쓰이

고 있다. 그러나 일반적 용도로 개발된 Harvard-IV의 단점은 특정 분야에서 사용되는 언어에 대한 감성 분석에는 적절하지 못하다는 점이다. 예를 들어 debt, tax 등은 중립적인 감성으로 분류되어야 하나 일반적인 경우 부정적 감성으로 인식될 수 있다. Loughran and McDonald (2011)는 경제 · 금융 분야에 특화된 감성사전을 개발하였는데, 이후 FOMC 발표문, 의사록의 감성분석 등에 널리 활용되고 있다.

한글의 경우 여러 종류의 감성사전이 개발되어있다. 김문형 외(2013)는 한국어 감성사전 구축방법을 제안하고 KOSAC(Korean Sentiment Analysis Corpus)를 개발하였다. 안정국, 김희웅(2015)은 각 단어 감성에 대한 투표라는 집단지성방법으로 한국어 감성사전을 구축하였다. 김건영, 이창기(2016)는 회선신경망(CNN)을 활용하여 영화감상평의 감성을 분석하기도 하였다. 그러나 한글의 경우에도 Harvard-IV 감성사전의 경우에서와 같이 일반적으로 널리 활용될 수 있는 감성사전의 한계가 존재한다. 예를 들어 소비자의 심리를 나타내는 논조를 측정할 때와 금통위 의사록에서 경제현황에 대한 금통위원들의 평가를 나타내는 논조를 측정할 때 동일한 감성사전을 사용하는 것은 분석의 효율성을 떨어뜨릴 수 있다. 쉬운 예로 '불확실성'은 일반적으로 중립적인 감성을 지니는 것으로 분류되나, 경제 · 금융 분야에서는 부정적인 감성으로 해석되어야 한다.

특정 분야의 감성 사전을 구축할 때 자의성을 최대한 배제해야 하는데, 이러한 시도의 예로 경제 · 금융 분야에서 활용될 수 있는 감성사전을 구축한 Lee, Kim, and Park(2019a)을 들 수 있다. 이들은 약 20만개 이상의 문서에서 읽은 n-gram을 가지고 각각 시장 접근법과 어휘 접근법을 활용하여 감성사전을 구축하였다. 먼저 하나의 단어 단위(uni-gram)은 문맥을 제대로 반영하지 못 할 수 있기 때문에 n-gram을 사용하였다. 예를 들어, 'recovery'는 경제가 회복한다는 긍정적인 의미이나, 'sluggish recovery'라고 쓰면 회복이 더딘 경기 침체를 의미한다. 이들은 n = 5인 n-gram을 사용하였는데 감정 표현에 반드시 필요한 정보를 반영하면서 차원의 급격한 증가를 방지하기 위해 n-gram을 구성하는 단어는 품사정보(POS)를 바탕으로 명사, 형용사, 부사, 동사 및 부정어로 제한하였다. 그리고 n-gram의 극성을 시장 접근법, 어휘 접근법 두 가지를 사용하였다. 시장접근법은 문서가 발생한 일자의 금리반응을 기반으로 문서를 n-gram을 특성으로 하는 나이브 베이즈 분류모형[45]으로 분류하여 n-gram이 문서 분류에 기여하는 조

건부 확률로 극성을 구분하였다. 쉽게 표현하자면 금리가 일정 수준 이상 상승한 날과 하락한 날에 주로 출현하는 단어들(여기에서는 n-gram)을 구분하여 극성(polarity)을 부여하는 것이다. 시장접근법은 금융 시장의 반응과 연계시키기 때문에 연구자의 주관적인 판단이 배제되어 있다는 것이 장점이나 필연적으로 금리 변수와 높은 상관관계를 보이게 되어 인과관계 판단에 제한적일 수밖에 없는 단점이 존재한다. 이를 보완하기 위해 어휘 접근법도 시도하였다. 어휘접근법은 사전적으로 정의된 핵심어(seed word)와의 유사도를 기반으로 극성을 분류하는 방식이다. 어휘접근법은 문서에 존재하는 어휘간의 관계만 이용하기 때문에 시장접근법이 가지는 내재적 문제점이 존재하지 않으나 연구자의 핵심어 선정에 따른 자의성 문제가 존재한다. 이 문제를 해결하기 위해 Hamilton *et al.* (2016)이 제안한 SentProp 기법을 사용했다. 이는 후보군에서 임의로 선정한 핵심어를 기준으로 극성을 분석하는 절차를 반복적으로 실행하고 그 결과의 평균을 사용하는 부트스트래핑 방법론을 적용함으로써 핵심어 선정의 자의성을 줄이는 방법이다. 이들 연구는 출발점이 전혀 다른 두 방식으로 분류한 사전이 유사하다는 것을 확인함으로써 분석결과의 신뢰도를 제고하였다.

5.2 논조분석(tone analysis)

수작업 또는 기계를 이용하여 감성사전을 구축한 후 이를 활용하여 새로운 문서샘플의 감성을 측정하여 논조지수를 작성할 수 있다. 문서를 기간별 또는 이벤트 별로 분류하면 해당 문서샘플에서 얻을 수 있는 논조의 변화에 내재된 경제학적 의미를 추출하는 것이 논조분석의 목적이다. 이렇게 측정된 논조는 개개의 경제 시계열이 전달해주는 정보보다 훨씬 종합적이며 직관적일 수 있다. 거시경제

45) 앞서 기술한 바와 같이 나이브 베이즈 분류모형(Naive Bayes Classifier)은 주어진 문장 또는 단어를 사전 정의한 범주로 분류할 수 있는 기법이다. 나이브 베이즈 분류모형으로 단어 또는 n-gram의 극성을 측정할 수 있다. 주어진 사전확률(p(hawkish) 또는 p(dovish))이 동일하다는 가정 하에 단어가 나타날 조건부 확률의 비율로 단어 또는 n-gram의 극성을 측정하는 것이다(Lee, Kim, and Park, 2019a).

$$polarity\,score = \frac{p(w|hawkish)}{p(w|dovish)}$$

현황을 나타내는 지수는 시계열로부터 얻을 수 있는 구체적인 정보를 종합하여 편제해야 할 것이다. 거시경제를 한 눈에 살펴볼 수 있는 지수의 편제는 계속 이루어지고 있으며, 시카고 연준에서 발표하는 CFNAI[46]의 경우가 대표적이다. 금리를 결정하는 금융통화위원회의 경우 일반적으로 시장에 비해 정보가 우월하다고 알려져 있다(Kawamura *et al.*, 2019). 이들 정보를 종합하여 경제현황을 논의하고 전망하는 토론 과정을 기록한 금통위 의사록의 논조 변화는 이들 정보가 종합적으로 내재되어있다고 할 것이다(Lee, Kim, and Park, 2019a). 달리 말하면, 경제현황을 나타내는 데이터와 금통위원들의 성향 등 모든 가능한 정보를 종합하여 하나의 지수로 나타내는 CFNAI와 같은 지수를 감성분석으로 편제한 것과 같다.

6. 평가(evaluation)

머신러닝 모형의 학습 과정이 끝나면 모형에 대한 평가가 이루어져야 한다. 다양한 평가 기법이 제안되고 사용되고 있으며 아래 표는 지도학습 여부, 분류와 회귀 여부에 따른 평가 기법들을 보여 주고 있다.

〈표 3〉 모형과 분석 목적에 따른 평가 기법들

지도 학습 모형 (supervised learning models)	
분류 (classification)	회귀분석 (regression)
Accuracy, Precision, Recall, F1 score	MSE (mean squared error)
ROC-AUC	MAE (mean absolute error)
Log-Loss	R^2, $\overline{R}^2$
비지도 학습 (unsupervised learning models)	
Rand index	
Mutual information	
기타	
CV error, BLEU score	

46) 시카고 연준에서 작성하는 CFNAI(Chicago Fed National Activity Index)는 80여개 시계열을 주성분분석을 활용하여 하나의 지수로 편제한 것이다.

본고에서는 텍스트 마이닝 분야에서 많이 쓰이는 지도 학습 모형 중 분류(classification) 관련 평가 지표(evaluation metrics)로 가장 널리 쓰이는 Accuracy, Precision, Recall, F1 Score와 ROC-AUC를 설명한다.[47]

아래 표는 분류 문제에서 실제 값과 모형의 예측에 따라 네 가지 조합을 보여주는데 오차행렬(confusion matrix) 또는 혼동행렬이라 불리운다. TP는 True Positive(s)의 약자이며 실제 값이 참(True)인데 모형의 예측도 일치한 경우를 말한다. 스팸 메일의 예를 든다면 실제로 어떤 이메일이 스팸 메일인데 모형의 예측도 스팸 메일로 나온 경우이다. 실제 값이 참이나 모형의 예측은 반대인 경우 FN에 해당하며 실제 값이 거짓(False)이나 모형의 예측은 참인 경우 FP에 해당한다. TN은 실제 값과 모형의 예측 모두 거짓인 경우이다.

〈표 4〉 오차행렬 (confusion matrix)

		머신러닝 모형의 예측	
		Positive	Negative
실제 값	True	TP	FN
	False	FP	TN

오차행렬에 나오는 TP, TN, FP, FN의 값에 기반하여 평가 지표를 계산하다. 먼저 정확도(accuracy)는 다음 식으로 정의되는데 전체 값들 중에서 실제 값과 모형의 예측이 일치한 비중을 보여주는 지표이다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

정밀도(precision)는 모형이 참이라고 예측한 것 중에 실제 값도 참인 것의 비중이다.

$$\mathrm{P}recision = \frac{TP}{TP + FP} \tag{17}$$

47) 지도학습 중 회귀모형의 경우 계량경제학에서 널리 쓰이는 지표들을 사용하므로 별도의 설명을 하지 않는다.

재현율(recall)은 실제 값이 참인 것들 중에서 모형이 참라 맞춘 것의 비율이다.

$$Recall = \frac{TP}{TP+FN} \tag{18}$$

F1 score는 정밀도와 재현율의 조화 평균으로 계산한다.

$$F1score = 2\frac{\mathrm{P}recision^{*}Recall}{\mathrm{P}recision+Recall} \tag{19}$$

이들 지표들은 분석하는 데이터의 특성에 따라 조심스럽게 선택되어 사용해야 한다. 예를 들어 1개의 스팸 메일과 999개의 정상 메일이 있다고 하자. 모든 메일을 정상 메일로 분류하는 형편 없는 모형이 있다고 하자. 이런 경우에도 정확도를 계산하며 그 값은 99.9%(=999/1,000)나 된다. 데이터가 참, 거짓 중 한 쪽으로 치우져 있을 경우 정확도는 좋은 평가 지표가 될 수 없다.

정밀도는 FP가 중요할 때 유용한 지표가 될 수 있다. 스팸 메일의 예로 설명하자면 FP는 정상 메일인데도 불구하고 스팸 메일로 분류될 경우이다. 즉 중요한 업무상 메일이 스팸 메일로 분류되어 메일을 볼 수 없는 상황이 일어날 수 있는데 이때 정밀도로 모형을 평가할 수 있다. 비슷한 논리로 재현율은 FN이 중요할 때 유용한 지표가 된다. 예를 들어, 신용카드 사기의 패턴이 나타남에도 불구하고 사기가 아니라고 분류한다든가, 어떤 사람이 특정 병에 걸렸음에도 불구하고 환자로 분류하지 않는 상황을 들 수 있다. F1 score는 Precision과 Recall의 조화평균으로 계산되기 때문에 FP와 TN을 모두 고려하는 장점이 있으며 한쪽으로 치우친 데이터의 경우에도 유용하게 사용될 수 있다.

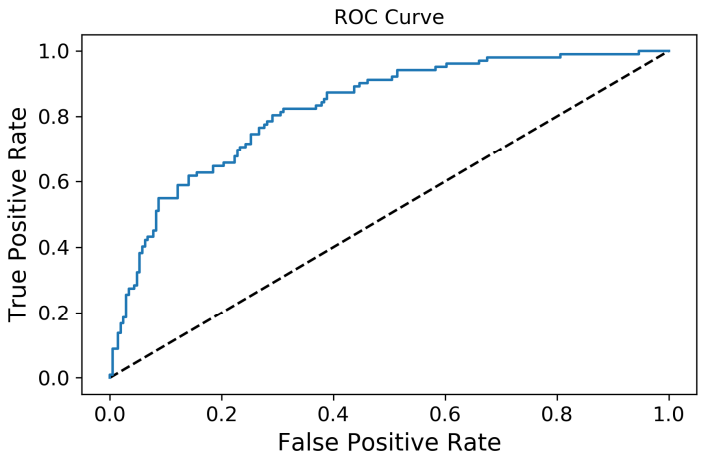
ROC 곡선은 가로축에 FP/(FP+TN)로 정의하는 FP 비율(False Positive rate)을, 세로축에 TP/(TP+FN)로 정의하는 TP 비율(True Positive rate)을 놓고 그린 곡선을 말하며 이 곡선의 아래 부분 면적을 AUC(Area Under Curve)라 한다.[48] ROC

48) FP 비율은 실제 값이 참인데도 모형의 예측이 거짓인 비중이며 일종의 오경보율(false alarm rate)로 해석할 수 있다. TP 비율은 실제 값이 참일 때 모형이 얼마나 정확하게 참으로 분류하는지를 보여 주며 민감도(sensitivity)라고도 불린다.

곡선은 여러 모형들의 상대적 성능을 시각적으로 비교할 수 있는 장점이 있으며 AUC는 ROC 곡선의 정보를 하나의 수치로 보여줄 수 있다.

<그림 11>은 ROC 곡선의 형태를 보여 준다.[49] 파란 색 곡선이 ROC 곡선의 전형적인 형태이며 ROC 곡선의 아래 면적이 AUC가 된다. True와 False를 완벽하게 분리해서 분류할 수 있는 이상적인 상황에서는 AUC가 1이 된다. True와 False를 전혀 구별해 내지 못 하는 경우 ROC 곡선은 점선으로 표시된 대각선이 되며 AUC는 0.5이다. True와 False를 정반대로 모형이 예측하는 경우 AUC는 0이다.

<그림 11> ROC 곡선의 예



Ⅳ. 텍스트 마이닝을 활용한 경제 분석

이 장에서는 경제학적 분석에 텍스트 마이닝을 활용한 사례를 소개하고자 한다. 텍스트 마이닝은 여러 맥락에서 활용될 수 있는데 첫째로 텍스트 마이닝 기법 없이는 분석 자체가 어려운 사례를 들 수 있다. 예를 들어 Gentzkow and

49) 다음 사이트의 파이썬 코드를 이용하였다: https://github.com/ageron/handson-ml

Shapiro(2010)는 언론의 논조가 사주의 정치적 입장에 의해 결정되는 것이 아니라 이윤 극대화의 결과라는 것을 보였다. Stock and Trebbi(2003)는 텍스트 마이닝을 활용해서 문체의 특성을 분석하는 방식으로 도구변수(instrument variable)를 Philip Wright가 처음 제안했는지 밝혀 냈다.[50]

둘째, 데이터의 주기(frequency)를 높이거나 아예 실시간 데이터 구축이 가능하다. 앞에서 예로 든 나우캐스팅 이외에도 Abraham *et al.*(2018)은 구글 트렌드(Google Trend) 데이터와 트위터 게시글의 논조 등으로 암호화폐 가격예측 모형을 구축하고, 암호화폐의 가격이 트위터 게시글의 논조보다 게시글의 수와 더 밀접한 관계가 있음을 밝히고 있다.

셋째, 기존의 데이터와 결합시켜 더 나은 경제 분석을 할 수 있다. Kelly *et al.*(2018)은 기술 혁신(technological innovation)을 텍스트 마이닝으로 측정하고 해당 지표가 산업별, 기업별 수준에서 생산성을 예측할 수 있음을 보였다. 기술 혁신를 수치화하기 쉽지 않은데 이들은 특허 문서에서 사용되는 유사도를 이용해서 기술 혁신을 지표화했다. 추가적인 장점으로는 특허 문서라는 텍스트가 19세기 초중반부터 존재하기 때문에 해당 지표도 1840년부터 구축이 가능하였다.

아래에서는 검색어 및 소셜미디어를 활용한 연구, 뉴스에 반영된 심리를 측정하는 연구, 중앙은행 관련 연구, 우리나라 현안에 관련된 연구 사례, 그리고 정부 및 기업 부문에 활용된 사례로 나누어서 관련 연구를 살펴 본다.

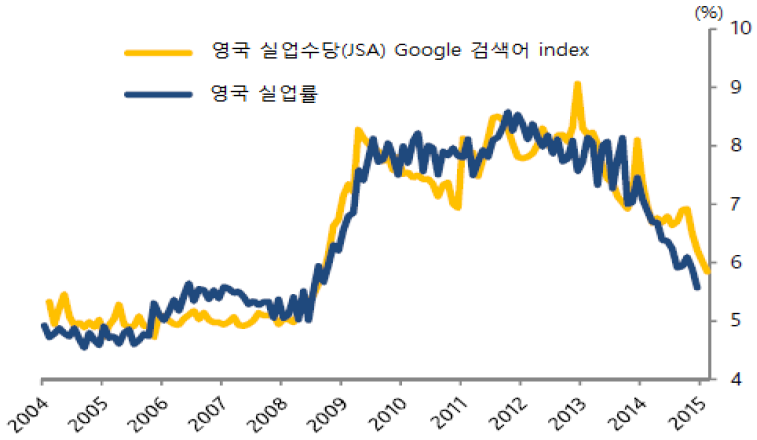
1. 검색어 및 소셜 미디어를 활용

구글(Google) 등 검색엔진이 보유한 검색어 데이터베이스를 활용하면 경기상황 및 불확실성 관련 정보를 추출하고 경제현황을 실시간 예측하는 나우캐스팅(nowcasting)을 할 수 있다. 노동시장과 주택시장을 설명하는 모형의 설명변수로

50) 도구변수는 경제학에서 널리 사용되어 왔지만, 최초로 누가 제안했는지 알려지지 않았다. Philip Wright의 저서 The Tariff on Animal and Vegetable Oils 부록에 도구변수의 개념이 설명되어 있었으나 부록을 Philip Wright가 썼는지, 당시 유명한 통계학자였던 아들인 Sewall Wright이 썼는지 논쟁이 이어져 왔다. Stock and Trebbi(2003)는 부록이 아버지 Philip Wright에 의해 집필되었음을 보였다.

구글 검색어 빈도를 지수화하여 추가하였을 때 설명력과 예측력이 유의하게 향상되었다(McLauren and Shanbogue, 2011). 영국의 경우 실업수당(JSA: Job Seeker's Allowance) 구글 검색어 빈도가 실업률을 잘 예측하는 것으로 나타났다.

〈그림 12〉 영국 실업수당 검색빈도와 실업률
(McLauren and Shanbogue, 2011)



주 : Google에서 실업수당(JSA: Jobseeker's Allowance) 검색어 빈도를 지수화

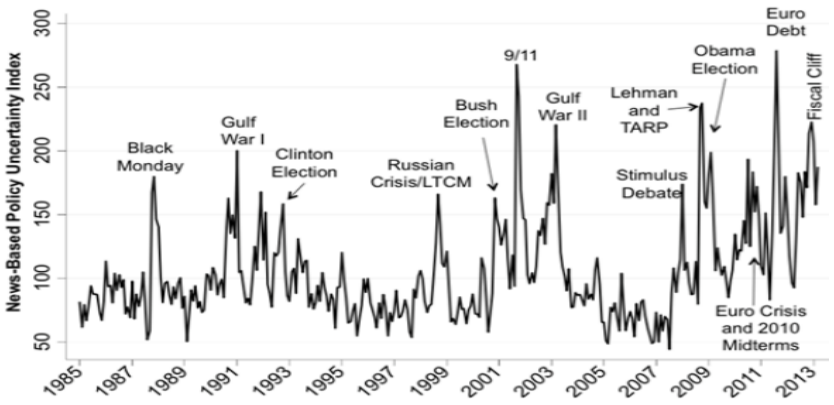
미국의 경우에도 경기선행지수로 잘 알려진 신규 실업수당 청구건수(Initial Job Claims)를 피설명변수로 하는 모형에 일자리(jobs), 복지(welfare) 관련 구글 검색어를 설명변수로 포함시키는 경우 설명력 및 예측력이 크게 향상된 것으로 나타났다(Choi and Varian, 2012). 유사한 방법으로 평면TV, 냉장고 등 구글 검색어 빈도수가 높아질수록 내구재 수요가 증대되는 것이 확인되었다(Choi and Varian, 2009).

Baker, Bloom, and Davis(2013)는 미국과 유럽의 주요 언론에서 불확실성 관련 단어를 포함하는 기사의 수를 불린 검색[51](Boolean methods)으로 측정하여 정책

51) 불린 검색이란 특정 검색어 포함여부로 문서를 나타내는 방법이다. 앞서 기술한 빈도행렬에서 어떤 문서에 특정 단어가 등장할 경우 원소를 "1"로 두고 AND, OR, NOT의 연산을 수행할 경우 검색조건에 따라 해당 문서만 추출할 수 있다. 구굴 검색에서 검색어 해당 건수를 반

불확실성지수[52](EPU: Economic Policy Uncertainty)를 편제하였다. 경제정책 불확실성을 나타내는 정책불확실성지수(EPU)가 상승할 경우 주가변동성이 확대되고 투자와 고용이 위축되는 것으로 나타났다.

〈그림 13〉 정책불확실성지수(Baker Bloom and Davis, 2013)



트위터(Twitter) 등 소셜 미디어나 로이터(Reuter) 뉴스 기사 데이터베이스에서 특정 이벤트 또는 심리를 나타내는 단어로 검색하였을 때, 그 결과로 추출되는 게시물 또는 기사의 수를 계산하여 시장의 심리를 측정하기도 한다. 영란은행은 과거 스코틀랜드의 분리독립 투표 당시 스코틀랜드 소재 은행의 뱅크런 발생 위험을 진단하기 위해 트위터에서 관련 검색어로 추출한 게시물 등을 통해 금융리스크를 실시간 모니터링하기도 하였다. 미국의 경우 미연준의 출구전략(tapering) 시행시기 관련 기대변화를 나타내는 트위터 게시물의 수로 출구전략 시기에 대한 기대변화를 측정하였다. 트위터에 나타난 기대변화는 자산가격에 유의한 영향이 있음이 밝혀졌는데, 출구전략 조기 시행을 예상하는 트위터 게시물이 10% 증가할 시 10년물 정부채 금리가 3bp(basis points)상승하며 미달러화는 유로화 대비 0.2% 절상하는 것으로 확인되었다(Meinusch and Tillmann, 2017).

환하는 연산도 이러한 불린 검색방법에 속한다.

52) 현재 우리나라를 포함한 20개 주요국에 대해 매월 EPU를 산출하여 발표하고 있다. www.policyuncertainty.com 참조

2. 뉴스에 반영된 심리

시장의 심리가 뉴스 기사의 어조에 반영된다면 뉴스 기사를 텍스트 분석하여 금융시장 움직임에 대한 예측이 가능하다. Nyman *et al.*(2015)은 영란은행 일간 보고서와 뉴스에 사용된 단어를 흥분(excitement)과 우려(anxiety)라는 두 요소(emotional factor)로 구분하여 내재된 시장 심리(sentiment)를 추출하였다. 추출된 지수는 VIX 등 금융시장 지표와 비교하였는데, 특히 영란은행 일간보고서(MCDAILY: Market Comment Daily)에서 추출한 심리는 VIX와 유사한 추세를 갖고 있으며 글로벌 금융위기 기간과 그 이후에는 VIX에 선행하는 모습도 관측되었다.[53] Tetlock(2007)은 Wall Street Journal의 “Abreast of the Market” 칼럼을 Harvard-IV 감성사전으로 분석하여 각 칼럼 당 77개 감성에 해당하는 단어의 수를 측정하였다. 주성분분석으로 분석한 결과 해당 감성의 단어 수중에 비관(pessimism)에 해당되는 단어의 수 변화가 주가 움직임을 가장 잘 나타낸다는 점을 발견하였다. 한편 우리나라의 경우에도 뉴스에 사용된 단어들의 긍정 또는 부정 논조를 가중평균하여 얻은 심리지수를 활용한 연구가 있다. Pyo and Kim (2018)은 긍정적 혹은 부정적 논조를 내포한 핵심어(seed lexicon)를 사전 정의하고, 이들과의 동반출현확률[54](PMI: Point-wide Mutual Information)로 단어의 극성[55]을 측정하였다. 기사의 단어의 극성을 가중 평균하여 시기별 심리지수를 편

53) Bholat *et al.*(2015)은 다음과 같은 이유로 텍스트 분석을 통한 불확실성 지수가 VIX에 비해 우월하다고 주장한다: ① 금융시장에 국한된 불확실성보다 넓고 다양한 분야의 불확실성 측정이 가능 ② 자산가격의 움직임으로 간접적으로 측정한 불확실성보다 텍스트에 나타난 불확실성 표현이 더욱 명확 ③ 오래전 자료를 활용할 수 있어 1900년대부터 편제 가능(VIX는 1980년대 후반부터 가능) ④ 국가별로 편제되고 국가간 비교도 가능(VIX는 미국 주식시장 가격으로 측정)

54) 동반출현확률(PMI: Point-wise Mutual Information)은 문헌내 특정 어휘의 등장 여부를 확률과정에 따른다고 가정하고 두 단어가 함께 나타날 수 있는 조건부확률을 계산한다. 이 조건부확률이 높을수록 다시 말하면 동반출현확률이 높을수록 두 단어가 지니는 논조 혹은 감성이 일치한다고 할 수 있다. 두 단어($w1, w2$)의 동반출현확률은 아래와 같이 계산한다.

$$PMI(w1, w2) = \log \frac{p(w1, w2)}{p(w1)p(w2)}$$

55) 단어의 극성(polarity)은 긍정적 단어와 동반출현빈도가 높을수록 긍정(+1)에 가깝게 나타나며, 반대의 경우 부정(-1)에 가까운 값을 부여함으로써 측정할 수 있다.

제한 결과 동지수는 주가, 금리, 환율 등에도 관계가 있다는 것을 보였다.

3. 중앙은행 커뮤니케이션

3.1 통화정책 관련 커뮤니케이션 분석

중앙은행 커뮤니케이션인 의결문, 기자간담회 등을 분석하면 통화정책 방향, 경제현황에 대한 평가와 중앙은행이 평가하는 불확실성 정도 등을 추출할 수 있다. FOMC 의결문의 논조변화는 최소 일년 이상 FOMC의 통화정책 방향 및 장기 국채금리에 영향을 미치며, 이로써 FOMC 의결문이 기존 테일러 준칙에 사용되는 거시경제변수에 담긴 정보 외에 추가적인 정보를 내포하는 증거가 된다. 특히 글로벌 금융위기 이후 초저금리 제약(zero lower bound) 하에서 선제적 지침(forward guidance)와 같은 중앙은행의 커뮤니케이션 정책은 텍스트 마이닝이 중요한 역할을 할 수 있다.

Lucca and Trebbi(2011)는 핵심어 동반출현확률[56](SO-PMI) 기법으로 FOMC 의결문에 사용된 단어의 극성을 분석하고 이를 심리지수로 편제하여 의결문이 향후 통화정책 방향에 대한 상당한 정보를 내포하고 있음을 확인하였다. ECB의 경우 기자회견문에 담긴 어조(tone)가 향후 ECB 통화정책 방향에 유의한 수준의 설명력을 지니고 있으며, 특히 긍정적 어조(positive tone)가 금융시장의 변동성을 감소시키는 효과가 있는 것으로 나타났다. Picault and Renault(2017)가 ECB 커뮤니케이션에 특화된 사전을 구축하고 n-gram 단위의 조건부 확률 언어모형으로 통화정책방향과 경제현황에 대한 평가와 관련된 논조를 추출하고, 이를 확장된 테일러준칙[57](augmented Taylor rule)의 설명변수로 하여 정책금리와 통화정책

56) 미리 정의된 긍정적 핵심어(seed lexicons)와 부정핵심어를 각각 pw, nw라 할 때 이들과의 동반출현확률로 단어 w를 아래와 같이 계산할 수 있는데, 이 경우를 핵심어 동반출현확률(SO-PMI: Semantic Orientation from Point-wise Mutual Information)이라 한다. 핵심어 동반출현확률은 아래와 같이 계산할 수 있다.

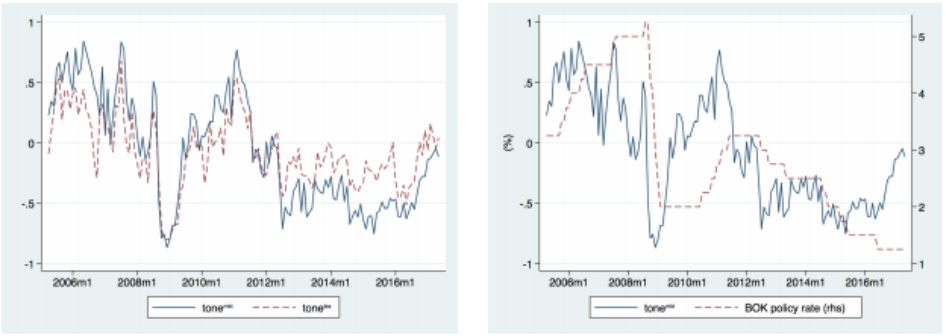
$$SO-PMI(w) = \sum_{pw \in PW} PMI(w, pw) - \sum_{nw \in NW} PMI(w, nw)$$

57) 인플레이션갭과 GDP갭이 포함된 기존 테일러 준칙에 ECB 기자회견문에서 추출한 긍정, 부정 논조 지수를 설명변수로 부가하였다.

스탠스에 대한 예측력과 설명력을 검정한 결과 상당한 수준의 설명력과 예측력이 있음이 확인되었다.

Lee, Kim, and Park(2019a)은 2005~2017년중 한국은행 금융통화위원회 의사록 논조를 측정하여 수치화된 논조지수가 기준금리에 대한 설명력과 예측력을 지니고 있음을 발견했다. 의사록의 논조는 해당 연구 과정에서 머신러닝 방법론으로 구축한 시장 접근법과 어휘 접근법 감성사전을 활용하였으며, 두 종류의 사전으로 측정한 의사록의 논조의 설명력과 예측력이 유사한 수준으로 밝혀졌다. 아래 <그림 14>는 두 가지 방식으로 구축한 금통위 의사록의 논조가 유사하면 실제 기준금리와 비슷하게 움직이는 것을 보여준다.[58]

<그림 14> 의사록 논조와 기준금리(Lee *et al.*, 2019a)



(a) $tone^{mkt}$ and $tone^{lex}$ (b) $tone^{mkt}$ and BOK policy rate

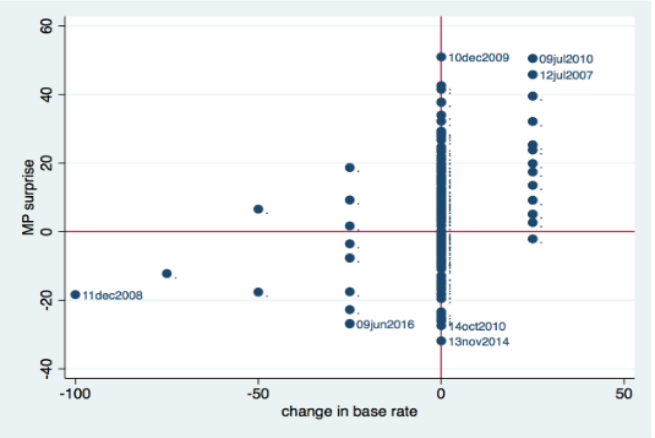
중앙은행의 통화정책관련 커뮤니케이션은 금융시장뿐만 아니라 실물변수에도 영향을 미칠 수 있으며, 커뮤니케이션의 내용과 논조에 따라 장단기 금리가 영향을 받거나 가격 변동성이 완화되는 등 특히 금융시장에 더 큰 영향력을 지니고 있는 것으로 분석되었다. FOMC 의결문의 사전적 정책방향 제시(forward guidance)에 해당하는 내용은 실물변수에도 영향이 있으나 특히 금융시장에 파급력이 더욱 큰 것으로 나타났다. Hansen and McMahon(2016)은 잠재디리클레할당(LDA:

58) 2015년 중반부터 기준금리와 금통위 의사록의 논조가 괴리를 보이는데 이는 아래에서 설명한다.

Latent Dirichlet al.location)을 활용하여 FOMC의 의결문을 경제현황(economic outlook)과 사전적 정책방향 제시로 구분하고 감성사전을 활용하여 어조를 측정하였다. Hendry and Madeley(2010)는 통화정책방향 의결문에 포함된 지정학적 위험, 주요 대내외 충격요인, 사전적 정책방향 제시 관련 발언 등의 내용이 단기 금리 변동성에 상당한 영향이 있음을 확인하였다.

Chague, De-Losso, Giovannetti and Manoel(2013)은 캐나다 중앙은행(Bank of Canada) 통화정책방향 의결문에 대해 잠재의미분석(LSA)을 적용하여 상위 10개 주제를 추출한 후, 해당 주제 관련 논의가 금리변화에 미치는 영향에 대해 회귀분석을 실시하였다. 그 결과 통화정책회의 의사록에 내포된 낙관적 어조는 장기 선물금리에도 큰 영향을 미치고 불확실성을 감소시켜 금리 변동성을 줄이는 효과가 있음이 확인되었다.

〈그림 15〉 기준금리 변경과 통화정책 서프라이즈(Lee, Kim and Park, 2019b)



Lee, Kim, and Park(2019b)는 금융통화위원회 전후 금통위 관련 신문기사 논조의 차이를 통화정책으로 인한 서프라이즈로 정의하고, 측정된 통화정책 서프라이즈가 금리, 주가 등 자산가격에 미치는 영향을 연구하였다. 1년 이하 단기 금리에 영향을 미치는 기준금리 변경과 달리, 통화정책 서프라이즈는 1년 이상 중장기 금리에 유의하게 영향을 미치며 Gurkaynak *et al.*(2005)의 경로요인(path

factor)의 성질을 갖고 있는 것으로 확인되었다. 한편 기준금리 변경이 없는 시기에도 서프라이즈는 더욱 큰 폭으로 나타날 수 있어, 사전적 정책방향 제시(forward guidance) 등 중앙은행의 커뮤니케이션이 효율적 통화정책 수단이 될 수 있는 증거를 제시하였다. <그림 15>는 가로축에 한국은행 기준금리 변화, 세로축에 통화정책 서프라이즈를 표시하는데 기준금리 변화가 전혀 없었던 금통위 회의 일자 전후로도 서프라이즈의 크기는 매우 상이하다는 것을 보여 준다.

3.2 거시건전성 정책 관련 커뮤니케이션 분석

중앙은행의 거시건전성 정책관련 커뮤니케이션도 금융시장 움직임에 유의한 영향을 미치며, 시중 은행에서 작성된 문헌에도 시장의 심리가 반영되어 있어 금융안정과 밀접한 관련이 있음이 확인되었다. 우리나라를 포함한 37개 중앙은행의 1,000여 건의 금융안정 보고서와 연설문 및 언론 인터뷰 등을 사전방식으로 분석하여 긍정(+1)과 부정(-1) 문장으로 구분하여 분석한 사례도 있다. Born *et al.*(2014)은 금융안정보고서에 담긴 논조가 낙관적인 경우 주식시장의 추가 수익이 장기에 걸쳐 관찰되고 시장 변동성도 감소하나, 연설문에 담긴 논조가 낙관적인 경우 가격에는 영향을 미치는 반면 변동성에는 영향이 없는 것을 확인하였다.

Bholat *et al.*(2017)은 머신 러닝을 활용하여 영란은행 건전성 감독기구의 비밀서한을 분석한 결과 서한의 어조가 수신 기관의 위험도에 따라 상이함을 발견하였다. 또한 유로지역 소재 개별 은행의 연차보고서와 은행장이 발송한 서한 등에서 추출된 불확실성과 부정적 감성(sentiment)은 은행권 전체의 자기자본비율과 높은 수준의 상관관계가 존재함이 드러났다. Nopp and Hanbury(2015)은 유럽 소재 민간은행 은행장의 공개된 서한과 은행 연차보고서 등의 어조와 내포된 감성을 사전(lexicon)기법으로 분류하였다.

3.3 커뮤니케이션 투명성 제고의 영향

통화정책의 투명성을 제고하기 위해 의사록을 공개하였는데 통화정책 의사결정자들에게는 적극적인 토론으로 회의에 참여할 유인이 증대되고 의사결정구조에 긍정적 효과를 가져왔으나, 본인의 발언에 대한 리스크도 동시에 부담하게 되

므로 소수의견을 피력하는 경우가 점차 감소하는 부작용을 초래한다는 것이 밝혀졌다. Hansen, McMahon and Prat(2018)은 FOMC 회의의 투명성 제고 이후 임명된지 얼마 되지 않은 신참(rookie) 멤버일수록 자료 인용 횟수가 많아지고 토의 발언을 더 길게 끌어가며, 다양한 주제에 관하여 적극 논의하는 등 회의 준비를 위해 노력하는 모습을 보이는 경향이 있는 점을 밝혔다.

FOMC 투명성 제고가 멤버들의 적극적인 토론 참여로 이어진 것을 확인한 또 다른 연구로는 Acosta(2014)가 있다. FOMC 의사록 공개 이후 멤버들 발언 시간이 평균적으로 길어졌으며, 지난 의사록을 참조하고 자료들을 사전에 준비하여 발언시간 동안 적극 인용하는 모습을 보였다. 그러나 FOMC 의사록 공개 결정 이후 멤버들은 여타 멤버들의 의견에 적극적 반대 의견을 개진하는 횟수는 감소하였다. 또한 Acosta and Meade(2015)는 잠재의미분석(LSA)과 벡터공간내 코사인(cosine) 유사도 분석으로 FOMC 멤버들 발언이 점차 비슷(conformity)해지는 경향을 발견하였다. FOMC는 1999년부터 매 회의 이후 통화정책방향 의결문을 발표하였으며 2002년는 금리결정 찬성과 반대를 표시한 멤버를 공개한 이후부터 전기 대비 의결문의 유사도가 점차 상승한 것으로 나타났다. 전기의 FOMC 의결문과 유사도는 육안으로 확인하였을 때보다 벡터공간에서 측정한 경우 더 높게 나타나며 유사도가 점차 상승하는 추세가 관찰되었다.[59]

4. 텍스트 마이닝을 활용한 현안 분석

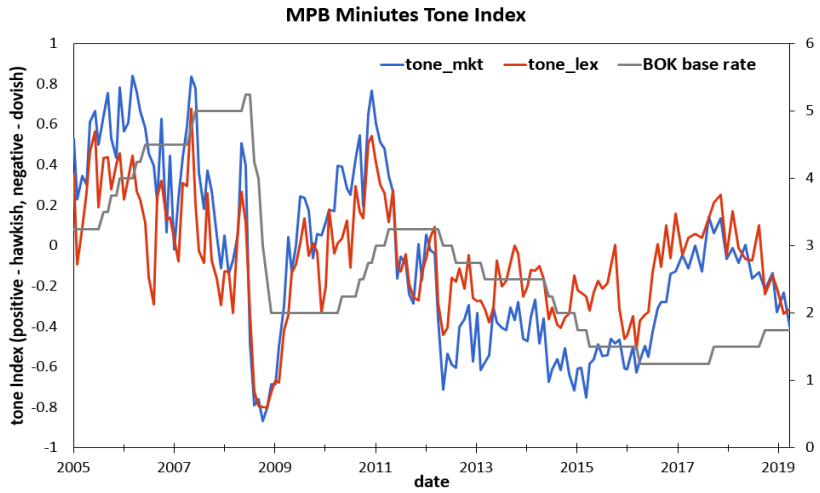
앞서 나열한 분석 사례는 연구자들이 텍스트 마이닝을 활용하여 다양한 거시경제 분석을 할 수 있는 가능성을 보여준다. 아래에는 텍스트 마이닝으로 현재 관심을 받을 수 있는 이슈에 적용한 사례로 우리 경제의 현안 분석에 어떻게 활용할 수 있는지에 대한 예시이다.

<그림 16>은 Lee, Kim and Park(2019a)에서 구축한 금통위 의사록의 논조를 2019.5월까지 연장한 것을 보여 준다. 최근 두 번의 금리인상의 배경을 이해하는

59) 126개 FOMC 의결문의 1,302개 단어의 빈도수를 원소로 하는 1302×126 행렬을 구성하고 벡터공간모형에서 두 의결문(a_i,b_i)의 코사인 유사성(cosine similarity)을 측정

데 도움을 준다. 한국은행 기준금리는 그림과 같이 2017.11월 1.50%로 한 차례 인상되었고, 1년 후 2018.11월에 1.75%로 다시 한 번 인상되었다. 그러나 두 경우의 논조를 비교하자면 같은 금리인상에도 불구하고 인상의 성격과 그 배경에 차이가 있음을 알 수 있다. 2017.11월 인상시에는 논조지수가 상당히 매파(hawkish) 쪽으로 상승한 시점이었고 이는 미국 연준의 금리 인상 시기와 맞닿아 있다. 2018년 이후 논조는 보다 dovish 쪽으로 하락하는 추세로 전환하였고, 2018.11 기준금리 인상은 경기회복 국면에 대응한 정책적 판단보다는 금리 정상화 또는 금융 불균형 완화가 주된 배경이었음을 짐작할 수 있다. 아울러 논조 지수의 추세에 따르면 오히려 금리인하에 대한 시그널을 내포하고 있다고 볼 수 있으며 실제로 2019년 7월 기준 금리는 1.50%로 인하되었다.

〈그림 16〉 금융통화위원회 의사록 논조와 금리



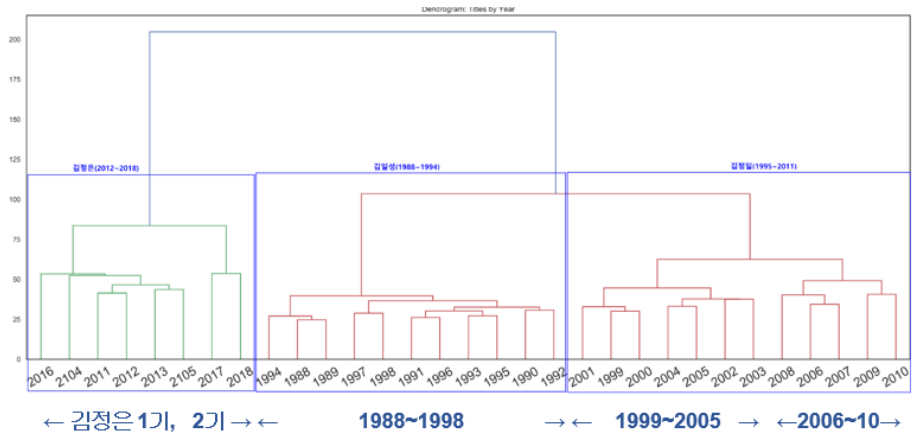
주 : tone_mkt와 tone_lex는 각각 Lee, Kim, and Park(2019a)에서 시장접근법(market approach)와 사전접근법(lexicon approach)으로 구축된 감성사전을 활용하여 측정한 의사록 논조 지수를 의미하며 BOK base rate은 한국은행 기준금리

<그림 17>은 북한의 경제학술지 「경제연구」 2575편의 제목을 텍스트 마이닝을 활용하여 빅데이터로 변환하고 이를 머신 러닝[60]으로 분류한 사례이다. 북한에서 「경제연구」와 같은 공식문건에 표현되는 제목의 경우 저자는 주제와 어휘

선택에 있어 권력자의 정책 방향과 맥을 같이 할 수 있도록 심사숙고하는 경우가 많다. 따라서 「경제연구」의 제목만으로도 북한 지도부의 정책 방향과 그 이면에 담긴 대내외 경제적 배경에 대해 분석해볼 수 있다.

본 사례에서는 「경제연구」 연도별 제목을 전처리하여 벡터로 변환한 후에 사전 정보 없이 벡터의 코사인 값(cosine similarity)만으로 제목을 분류하였다. 분류결과 김일성, 김정일, 김정은 시대에 따라 비교적 잘 분류되는 것으로 나타났다.[61] 그 외에도 대내외 주요 경제적 환경변화에 따라 논문제목의 유사도가 구분되고 있는데 이러한 시대 구분은 북한의 수출입의 추세적 변화와도 잘 일치한다. 예를 들어 북한의 수출입은 1998년까지 감소세를 보여주고 있으며 이 시기는 또한 북한체제의 몰락 위험이 대두되던 시기이다. 1999년 이후 2005년까지는 북한 체제가 아래로부터의 시장화가 진행되었으며 북한 당국은 이를 수용하는 입장이었다. 그러나 2006년부터 2010년까지 다시 시장화를 억제하기 시작하였고

〈그림 17〉 북한 「경제연구」 논문제목의 연도별 군집화



주 : 2757편의 논문제목 샘플 기간(1988~2018)을 김일성 시대(1988~94), 김정일 시대 (1995~2011), 김정은 시대(2012~18)로 구분함

60) 여기에서는 비지도 학습 기법중 하나인 계층적 군집화 방법을 사용하였다. 이를 도식적으로 표현한 그림을 dendrogram이라 한다.

61) dendrogram에서 가지의 높이가 낮을수록 유사도가 높다. 따라서 가지의 높이에 따라 연도별 제목의 유사도가 다르게 나타남을 볼 수 있다.

김정은이 본격적으로 권력을 갖게 된 2011년 이후는 시장의 활용기로 수출입이 크게 증대된 것을 볼 수 있다. 그러나 2017년 이후 대북제재가 본격화되며 수출입 또한 대폭 감소하였다. 이러한 일련의 경제 환경 변화는 <그림 17>와 같이 사전 정보 없이 텍스트를 분석하였을 경우에도 잘 나타나고 있음을 알 수 있다.

5. 정부 및 기업 부문의 활용

텍스트 마이닝은 기업 부문에서는 이미 널리 쓰이고 있으며 공공정책 분야에서도 활발하게 활용되기 시작하고 있다. <표 5>는 정부와 기업 부문 활용 사례를 정리해서 보여준다. 특히 전자의 경우 앞에서 강조한 바와 같이 수치화하기 힘든 데이터를 수치화하거나, 실시간 데이터를 얻기 위해 텍스트 마이닝을 활용하고 있다.

〈표 5〉 정부와 기업 부문 텍스트 마이닝 활용 사례

정부 부분 활용 사례
국방, 치안[62]
— 소셜 미디어를 분석해서 EU 지역 테러리즘에 대한 조기경보 수단으로 사용 (Stevenson, 2017) — 영국에서는 공격적이거나 폭력적 언어 사용을 탐지해서 극단주의자들을 식별하는데 활용 (Davey *et al.*, 2018) — 경찰 보고서(police report)는 급하게 작성되거나 피해자, 목격자의 진술이 부정확하게 기록되는 경우가 있기 때문에 수기 인식 등 자연어 처리 기법을 이용해서 주요 정보의 정확성을 제고 (Ku *et al.*, 2008) — 미국 시카고, 산타 바바라, 더햄 등에서는 텍스트 마이닝을 이용한 예측 기법으로 경찰 병력을 효율적으로 배치하고 범죄를 예방 (Ebi, 2014) — 미국 버지니아와 뉴 올리언즈에서는 온라인 광고를 분석해서 인신매매를 방지 (Quinn, 2016) — 텍스트 분류 기법을 이용해서 백만 건에 달하는 미국의 비분류된 외교 문서를 분석하고 기밀로 분류해야 할 문서를 식별 (Souza *et al.*, 2018)
금융 감독
— 미국 증권거래위원회(SEC)는 LDA 기법을 이용해서 공시 관련 부정행위를 감시 (Bauguess, 2017) — 호주 증권투자위원회(ASIC)는 불공정 금융상품 판매 및 광고를 식별 (Hendry, 2018)

공공정책 개선
— 영국 정부 정책을 안내하는 사이트(www.gov.uk)의 민원 및 사용자 의견을 LDA 기법으로 분석해서 정부 정책을 개선 (Heron, 2016) — 텍스트와 정형화된 데이터를 결합해서 청정 에너지 관련 혁신 지표를 제작, 공표해서 투자자, 연구자, 기업들의 관련 의사 결정을 보조 (NASDAQ, 2018) — grade.dc.gov를 통해 수집한 미국 워싱턴 주 정부 행정에 대한 의견을 분석해서 정책 개선에 사용 (McClellan and Goldsmith, 2018) — 싱가포르는 공공기관 서비스를 안내하는 'Ask Jamie'라는 챗봇을 활용 (Government Technology Assistance, 2018) — 세계은행에서는 LDA기법으로 남미와 스페인의 대통령 연설문을 분석해서 정책 우선순위 결정에 반영 (Calvo-González *et al.*, 2018)
보건의료
— 미국의 경우 문자로 된 진단서에서 개인 정보를 비식별화(de-identification)하면서 중요한 의료 정보를 저장 (US National Library of Medicine, 2018) — 미국 FDA의 의약품 부작용 보고서를 토픽 모델링으로 분석해서 약품과 부작용 사이 관계를 식별하고 향후 예측에 활용 (Rocca, 2017) — 감성 분석을 이용해서 4년 동안 27백만 개의 트위터를 분석해서 미국 환자들의 병원 경험 평가가 지역마다 어떻게 달라지는지 분석 (지역간 어떻게 달라지는지 분석 (Sewalk *et al.*, 2018))

기업 부문 활용 사례[63]
고객 응대
— 음성 인식(speech recognition), 챗봇을 활용해서 고객 응대를 하고 관련 인력을 줄일 수 있음
기업 평판 관리
— 소셜 미디어, 사용 후기 등에 감성 분석을 적용해서 소비자의 평가를 수치화할 수 있음
광고
— 기존 방식에서는 인구학적 특성(인종, 성별, 나이 등), 경제적 지위 등에 기반해서 광고를 했다면 이메일, 소셜 미디어, 웹 검색 기록 등을 분석해서 광고를 효율적으로 배치할 수 있음
시장 정보 획득
— NLP를 이용해서 너무 많은 데이터로부터 관심 있는 분야(예를 들어, 인수 및 합병, 산업 전

62) 미국 국방부는 2017년 기준 74억 달러를 인공지능 관련 예산으로 책정했는데 자연어 처리 관련 예산은 83백만 달러로 2012년 예산보다 17% 증가했으며 비중이 가장 크게 증가한 분야 중 하나이다. 2018년 미국 국방부 산하 국방고등연구계획국(DARPA, Defense Advanced Research Projects Agency)은 향후 새로운 인공지능 기술 개발을 위한 예산을 200억 달러로 늘리겠다고 발표했다. 관련 기사는 다음을 참고 https://www.c4isrnet.com/it-networks/2017/12/06/dods-leaning-in-on-artificial-intelligence-will-it-be-enough,https://www.forbes.com/sites/samshead/2018/09/07/darpa-plans-to-spend-2-billion-developing-new-ai-technologies/#76b04b443ae1

망 등)에 특화된 정보만을 추려낼 수 있음
규제 준수
— 규제 준수 여부를 확인하고 관련 조치를 선제적으로 마련할 수 있음. 예를 들어, 소셜 미디어, 이메일, 웹 검색 데이터에 NLP의 개체명 인식(named entity recognition), 관계 인식(relation detection) 기술을 적용해서 특정 의약품의 부작용 여부를 파악할 수 있음

Ⅴ. 결 론

위에서 살펴 본 텍스트 마이닝의 폭넓은 활용 가능성이 있지만 몇 가지 제약을 염두에 둘 필요가 있다. 텍스트 데이터는 일부 서베이 자료에 비해 비용이나 효율성 측면에서 우월한 정보의 원천도 될 수 있다. 그럼에도 불구하고 텍스트 데이터와 전통적인 구조화 데이터(structured data)와의 완전 대체관계는 존재하지 않는다. 텍스트 데이터가 근본적으로 사람이 알고 있는 지식과 의견의 발현이라는 점에서, 사람의 생각을 거칠 필요가 없는 객관적 자료의 경우 기존대로 측정하는 방식의 통계자료들이 더 적합한 정보를 제공하기 때문이다. 일례로 GDP와 같은 통계를 텍스트 데이터를 집계하여 편제할 수는 없다. 따라서 텍스트 데이터는 구조화된 자료를 보완하거나 구조화된 자료와 함께 분석에 활용될 때 효용이 높을 수 있다. 즉 대체 가능성과 보완성을 고려하여 적절한 방법론을 선택할 필요가 있다.

보다 큰 범위의 문제로는 최근 활발하게 이용되고 있는 인공신경망 또는 딥러닝(deep learning)을 포함해서 머신러닝 방법론에 공통적으로 적용될 수 있는 한계로검증을 위한 통계적 방법론이 부족하다는 점이다.[64] 텍스트 데이터의 확률

63) 기업 부문의 경우 활용 사례가 너무 많으므로 활용의 목적 또는 범주 별로 간략하게 소개한다. 범주의 분류는 다음을 참고하였다: https://emerj.com/ai-sector-overviews/natural-language-processing-business-applications/

64) 2019년 Turing Award 수상자 Yann LeCun, Yoshua Bengio, Geoffresy Hinton 모두 인공지능 분야 연구의 선두주자들과 동시에 텍스트 마이닝 분야 전문가이며, Hinton과 LeCun은 각각 텍스트 데이터를 적극 활용하는 구글과 페이스북 부사장이라는 점에서 텍스트 분석에서 인공지능이 차지하는 중요도를 짐작할 수 있다.

적 표상만 하더라도 학습 결과의 강건성 등에 대한 통계적 유의성 등을 검증할 대안이 마련될 필요가 있으며, 머신 러닝으로 구축한 감성사전의 경우에도 학습된 사전을 검증할 기준을 엄밀히 설정할 필요가 있다. 추가적으로 분야에 따라 텍스트 데이터를 구축하는데 많은 시간이 소요될 수 있다는 점은 연구자들에게 부담이 될 수 있다.

이런 제약에도 불구하고 텍스트 마이닝은 기존 경제학의 분석범위를 더욱 넓혀줄 뿐만 아니라 절대적으로 부족한 자료의 한계를 극복하는데 도움이 될 수 있다. 게다가 빅데이터의 출현과 인공지능 기술 발전에 힘입어 최근 들어 새로운 정보의 원천으로 인식되고 있다. 경제학은 경제 활동을 연구하는 사회과학 분야로 자연과학과 달리 자연실험(natural experiment)에 의해 데이터를 구하기 어렵다. 게다가 데이터의 이면에 있는 경제주체의 기대변화, 심리변화는 수치화된 통계자료로 식별하는 것은 경제학의 오랜 과제가 되어왔으며 상대적으로 강한 가정에 기반한 이론을 통해 식별을 시도해 오고 있다. 이런 맥락에서 볼 때 텍스트 마이닝은 새로운 데이터, 또는 기존의 데이터를 보완할 수 있기 때문에 앞으로도 경제학 분야에서 그 활용도가 점점 커지리라 기대한다.

참 고 문 헌

감미아 · 송민,「텍스트 마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석」,『지능정보연구』, 제18권 제3호, 2012, pp.53-77.

권순보 · 유진은,「텍스트 마이닝 기법을 통한 수능 절대평가」,『열린교육연구』, 제26권 제2호, 2018, pp.57-79.

권충훈,「텍스트 마이닝과 언어네트워크 분석을 활용한 중등교사임용 교육학시험 내용분석」,『교육혁신연구』, 제28권 제3호, 2018, pp.1-25.

김규선,『국어과 교육의 원리』, 학문사, 2000.

김규훈 · 이준영 · 김혜숙,「국어학: 연어(collocation)의 어휘 교육적 활용 방안 탐색」,『새국어교육』, 제95권, 2016, pp.521-555.

김건영 · 이창기,「Convolutional Neural Network를 이용한 한국어 영화평 감성분석」, 한국정보과학회 학술발표논문집, 2016, pp.747-749.

김남원 · 박진수,「Naive Bayes 방법론을 이용한 개인정보 분류」,『지능정보연구』, 제18권 제1호, 2012, pp.91-107.

김문형 · 장하연 · 조유미 · 신효필,「KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스」, 한국정보과학회 학술발표논문집, 2013, pp.650-652.

김유신 · 김남규 · 정승렬,「뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형」,『지능정보연구』, 제18권 제2호, 2012, pp.143-156.

박은정 · 조성준,「KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지」, 제26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.

박진수 · 박현호 · 조준택,「112신고내용에 대한 텍스트 마이닝 분석과 시사점」,『한국치안행정논집』, 제15권 제3호, 2018, pp.159-186.

박호식 · 이민수 · 황성진 · 오상윤,「의료 정보 추출을 위한 TF-IDF 기반의 연관규칙 분석 시스템」,『소프트웨어 및 데이터 공학』, 제5권 제3호, 2016, pp.145-154.

배효진 · 김창업 · 이충렬 · 신상원 · 김종현,「텍스트 마이닝(Text mining)을 활용

한 한의학 원전 연구의 가능성 모색-'황제내경(黃帝內徑)'에 대한 적용례를 중심으로」, 『대한한의학원전학회지』, 제31권 제4호, 2018, pp.27-46.

손욱 · 성병묵 · 권효성, 「통화정책 발언과 금융시장의 반응」, 『경제분석』, 제11권 제4호, 2005, 한국은행, pp.1-44.

신중호 · 한영석 · 박영찬 · 최기선, 「어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사태깅」, 한국정보과학회 언어공학연구회 학술발표 논문집, 1994, pp.389-394.

심광섭 · 양재형, 「인접 조건 검사에 의한 초고속 한국어 형태소 분석」, 정보과학회논문지: 소프트웨어 및 응용, 제31권 제1호, 2004, pp.89-99.

유원준, 『딥 러닝을 이용한 자연어 처리 입문』, 위키북스, 2019.

안정국 · 김희웅, 「집단지성을 이용한 한글 감성어 사전 구축」, 『지능정보연구』, 제21권 제2호, 2015, pp.49-66.

Abraham, Jethin, Higdon, Daniel, Nelson, John, and Ibarra, Juan. "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis", *SMU Data Science Review*, Vol.1, No.3, 2018, Article 1.

Acosta, M., "FOMC Responses to Calls for Transparency", Finance and Economics Discussion Series 2015-60, Board of Governors of the Federal Reserve System (U.S.), 2014.

Acosta, M. and Meade, E. E., "Hanging on Every Word: Semantic Analysis of the FOMC's Postmeeting Statement", FEDS Notes, Washington: Board of Governors of the Federal Reserve System, 2015.

Almatarneh, S. and Gamallo, P., "Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification", Conference Paper in Advances in Intelligent Systems and Computing, 2018.

Baker, S. R., Bloom, N. and Davis S. J., "Measuring Economic Policy Uncertainty", *The Quarterly Journal of Economics,* Vol.131, No.4, 2016, pp.1593-1636.

Bang, H. C. and Ha, J., "Monetary Policy and Communication: How Bank of Korea's Decision Making Affects Media's Attention to Interest Rate Policy", *Journal of Money and Finance* (in Korean), 2013.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C., "A Neural Probabilistic Language Model", *Journal of Machine Learning Research*, Vol.3, 2003, pp.1137-1155.

Bholat, D., Hansen, S., Santos, P. and Schonhardt-Bailey, C., *Text Mining for Central Banks: Handbook*, Centre for Central Banking Studies (33), 2015, pp.1-19, ISSN1756-7270.

Bholat, D., Brookes, J., Cai, C., Grundy, K. and Lund, J., "Sending Firm Messages: Text Mining Letters from PRA Supervisors to Banks and Building Societies They Regulate", Staff Working Paper No. 688, Bank of England, 2017.

Born, B., Ehrmann, M. and Frtzscher, M., "Central Bank Communication on Financial Stability", *Economic Journal*, Vol.124, No.577, 2014, pp.701-734.

Bouma, G., "Normalized (Pointwise) Mutual Information in Collocation Extraction", Conference Proceedings, 2009.

Bauguess, Scott W., "The role of big data, machine learning, and AI in assessing risks: A regulatory perspective", Champagne keynote address, US Securities and Exchange Commission, June 21, 2017.

Calvo-González, Oscar., Axel Eizmendi, and Germán Reyes, "Winners never quit, quitters never grow: Using text mining to measure policy volatility and its link with long-term growth in Latin America", policy research working paper 8310, World Bank Group, 2018.

Cambria, E. and White, B., "Jumping NLP Curves: A Review of Natural Language Processing Research", *IEEE Computational Intelligence Magazine*, Vol.9, No.2, 2014, pp.48-57.

Carlin, Gelman A., J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B., *Bayesian Data Analysis*, Chapman and Hall/CRC, 3rd edition, 2014.

Chague, F., De-Losso R., Giovanetti, C. B. and Manoel, B., "Central Bank Communication Affects Long-Term Interest Rates", Working Papers, No. 2013-07, Department of Economics, University of Sao Paulo, 2013.

Choi, H. and Varian, H., “Predicting Initial Claims for Unemployment Benefits”, SSRN: https://ssrn.com/abstract=1659307, 2009

_____, “Predicting the Present with Google Trends”, Special Issue: Selected Papers from the 40th Australian Conference of Economists, Vol.88, No.1, 2012, pp.2-9.

Cox, D. R., “The Regression Analysis of Binary Sequences”, *Journal of the Royal Statistical Society*, Series B (Methodological), Vol.20, No.2, 1958, pp.215-242.

Davey, Jacob., Jonathan Birdwell, and Rebecca Skellett, *Counter conversations: A model for direct engagement with individuals showing signs of radicalisation online*, Institute for Strategic Dialogue, 2018, p.6.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R., “Indexing by Latent Semantic Analysis”, *Journal of the Association for Information Sceience and Technology*, Vol.41, No.6, 1990, pp.391-407.

Ebi, Kevin, “How Durham, N.C. fights crime with data—and wins”, Smart Cities Council, 2014.

Eckart, C. and Young, G., “The Approximation of a Matrix by Another of Lower Rank”, *Psychometrika*, Vol.1, 1936, pp.211-218.

Fausch, J. and Sigonius, M., “The Impact of ECB Monetary Policy Surprises on the German Stock Market”, *Journal of Macroeconomics*, Vol.55, 2018, pp.46-63.

Fawley, B. W. and Neely, C. J., “The Evolution of Federal Reserve Policy and the Impact of Monetary Policy Surprises on Asset Prices”, Review, Federal Reserve Bank of St. Louis, Vol.96, 2014, pp.73-109.

Friedman, M. and Schwartz, A. J., *A Monetary History of the United States, 1867-1960*, Princeton University Press, 1963.

Furnas, G. W., Landauer, T. K., Gomez, L. M., Dumais, S. T., “Human Factors and Behavioral Science: Statistical Semantics: Analysis of the Potential Performance of Key-word Information System”, *The Bell System Technical*

Jounal, Vol.62, No.2, 1983.

Gentzkow, M. and Shapiro, J. M., "What Drives Media Slant? Evidence From U.S. Daily Newspapers", *Econometrica*, Vol.78, No.1, 2010, pp.35-71.

Gentzkow, M., Kelly, B. and Taddy, M., "Text as Data", NBER Working Paper, No. 23276, 2017.

Gertler, M. and Karadi, P., "Monetary Policy Surprises, Credit Costs, and Economic Activity", *American Economic Journal: Macroeconomics*, Vol.7, 2015, pp.44-76.

Giannone, D., Reichlin, L. and Sala, L., "Tracking Greenspan: Systematic and Unsystematic Monetary Policy Revisited", CEPR Discussion Papers 3550, 2002.

Government Technology Agency, "'Ask Jamie' virtual assistant", accessed December 19, 2018.

Grus, J., Data Science from Scratch, O'Reilly Media, 2015.

Gurkaynak, R., Sack, B. and Swanson, E., "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements", *International Journal of Central Banking*, Vol.1, 2005.

Hamilton, W. L., Clark, K., Leskovec, J. and Jurafsky D., "Inducing Domain Specific Sentiment Lexicons from Unlabeled Corpora", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Conference on Empirical Methods in Natural Language Processing, 2016, pp.595-605.

Hansen, S. and McMahon, M., "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communications", *Journal of International Economics*, Vol.99, No.1, 2016, pp.114-133.

_____, "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication", *Journal of International Economics*, Vol.99, S114-S133, 38th Annual NBER International Seminar on Macroeconomics, 2016.

Hansen, S., McMahon, M. and Prat, A., “Transparency and Deliberation within the FOMC: A Computational Linguistics Approach”, *The Quarterly Journal of Economics*, Vol.133, No.2, 2018, pp.801-870.

Hendry, S. and Madeley, A., “Text Mining and the Information Content of Bank of Canada Communications”, Staff Working Papers, Bank of Canada, 2010.

Hendry, Justin, “ASIC eyes AI to crack down on dodgy financial practices”, *itnews,* February 22, 2018.

Heron, Dan, “Understanding more from user feedback”, Data in Government blog, Gov.UK, November 9, 2016.

IBM, “Big Data and Analytics”, http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html/, 2015.

Kawamura, K., Kobashi, Y., Shizume, M. and Ueda, K., “Strategic Central Bank Communication: Discourse Analysis of the Bank of Japan’s Monthly Report”, *Journal of Economic Dynamics and Control*, Vol.10, 2019, pp.230-250.

Kelly, Bryan., Dimitris Papanikolaou, Amit Seru and Matt Taddy, “Measuring Techonological Innovation over the Long Run”, NBER Working Paper No. 25266, 2018.

Kettunen, K., Kunttu, T., Järvelin, K., “To Stem or Lemmatize a Highly Inflectional Language in a Probabilistic IR Environment?”, *Journal of Documentation*, Vol.61, No.4, 2005, pp.476-496.

Kim, Y., “Convolutional Neural Networks for Sentence Classification”, Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp.1746-1751.

Kroeger, P., *Analyzing Grammar: An introduction, Cambridge*, Cambridge University Press, 2005.

Ku, Justin., Alicia Iriberri and Gondy Leroy, “Natural language processing and e-government: Crime information extraction from heterogeneous data sources”,

conference paper at the 9th Annual International Digital Government Research Conference, 2008.

Kuttner, K. N., "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market", *Journal of Monetary Economics*, Vol.47, 2001, pp.523-544.

Landauer, T. K. and Dumais, S. T., "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge", *Psychological Review*, Vol.104, No.2, 1997, pp.211-240.

Lapowski, Issie, "How bots broke the FCC's public comment system", November 28, 2017.

Lee, Y., "Introduction to eKoNLPy: A Korean NLP Python Library for Economic Analysis", https://github.com/entelecheia/eKoNLPy, 2018.

Lee, Y., Kim, S. and Park, K. Y., "Deciphering Monetary Policy Committee Minutes with Text Mining Approach: A Case of Korea", *Korean Economic Review*, forthcoming, 2019a.

_____, "Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea", BOK Working Paper, No.2019-11, 2019b.

Lin, Jessica *et al.*, "Ecosystem discovery: Measuring clean energy innovation ecosystems through knowledge discovery and mapping techniques", Oak Ridge National Laboratory, study abstract, accessed December 19, 2018.

Liu, B., *Sentiment Analysis and Subjectivity*, in Handbook of Natural Language Processing, New York, NY, USA: Marcel Dekker, Inc, 2009.

_____, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.

Loughran, T. and McDonald, B., "When is a Liability not a Liability? Textual analysis, dictionaries, and 10-Ks", *The Journal of Finance*, 2011.

Lucas, R. E., "Econometric Policy Evaluation: A Critique", Carnegie-Rochester Conference Series on Public Policy, Vol.1, 1976, pp.19-46.

Lucca, D. O. and Trebbi, F., “Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements”, NBER Working Paper Series, 2011.

MacQueen, J. B., “Some Methods for Classification and Analysis of Multivariate Observations”, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp.281-297.

Manning, C. D., Raghavan, P. and Schutze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

McLaren, N. and Shanbhogue, R., “Using Internet Search Data as Economic Indicators”, Bank of England Quarterly Bulletin, No. 2011Q2, 2011.

McClellan, Matthew and Stephen Goldsmith, “From comment cards to sentiment mining: The future of government service rating”, Data-Smart City Solutions, September 18, 2013.

Meinusch, A. and Tillmann, P., “Quantitative Easing and Tapering Uncertainty: Evidence from Twitter”, *International Journal of Central Banking*, Vol.13, No.4, 2017, pp.227-258

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

Nasdaq, “How artificial intelligence is taking over oil and gas”, August 10, 2018.

Noh, H., Park, J., Sim, G,, Yu, J. and Chung, Y., “Nonparametric Bayesian Models in Biomedical Research”, *The Korean Journal of Applied Statistics*, Vol.27, No.6, 2014, pp.867-889.

Nopp, C. and Hanbury, A., “Detecting Risks in the Banking System by Sentimental Analysis”, Conference on Empirical Methods in Natural Language Processing, 2015.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P. and Smith, R., “News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment”, Bank of England Staff Working Paper, No.704, 2018.

Oshima, Y. and Matsubayashi, Y., "Monetary Policy Communication of the Bank of Japan: Computational Text Analysis", Discussion Papers, No.1816, Graduate School of Economics, Kobe University, 2018.

Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S., "Latent SEmantic INdexing: A Probabilistic Analysis", *Journal of Computer and System Science,* Vol.61, No.2, 2000, pp.217-315.

Pescatori, A., "Central Bank Communication and Monetary Policy Surprises in Chile", IMF Working Paper, No. 18156, 2018.

Picault, M. and Renault, T., "Words Are Not All Created Equal: A New Measure of ECB Communication", *Journal of International Money and Finance*, No.79, 2017, pp.136-156.

Porter, M. F., "An Algorithm for Suffix Stripping", Program, Vol.14, No.3, 1983, pp.130-137.

Pyo, D. and Kim, J., "News Media Sentiment and Asset Prices: Text-mining Approach", KIF Working Paper, 2017.

Quinn, Kristin, "Modern slavery: Cognitive computing and geospatial technology help law enforcement track, locate, and rescue human trafficking victims", *Trajectory*, 2016.

Rapp, R., "Word Sense Discovery Based on Sense Descriptor Dissimilarity", In Proc. of the Machine Translation Summit IX, New Orleans, 2003.

Rocca, Mitra, *Lessons learned from NLP implementations at FDA*, US Food and Drug Administration, 2017.

Romer, C. D. and Romer, D. H., "A New Measure of Monetary Shocks: Derivation and Implications", *American Economic Review*, Vol.94, No.4, 2004, pp.1055-1084.

Rumelhart, D. E., Hintont, G. E. and Williams, R. J., "Learning Representations by Backpropagating Errors", *Nature*, Vol.323, No.6088, 1986, pp.533-536.

Salton, G. and McGill, M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

Sewalk, Kara C. *et al.*, "Using Twitter to examine Web-based patient experience sentiments in the United States: Longitudinal study", *Journal of Medical Internet Research*, Vol.20, No.10, 2018.

Shibamoto, M., "Empirical Assessment of the Impact of Monetary Policy Communication on the Financial Market", Discussion Paper Series, DP2016-19, Research Institute for Economics Business Administration, Kobe University, 2016.

Sohn, W., Sung, B. and Kwon, H., "The Financial Markets' Responses to Monetary Policy Announcements", Bank of Korea Economic Analysis (in Korean), 2005.

Souza, Renato Rocha *et al.*, "Using artificial intelligence to identify state secrets", Semanticsscholar.org, accessed December 19, 2018.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A., "Text mining and ontologies in biomedicine: making sense of raw text", *Briefings in bioinformatics*, Vol.6, No.3, 2005, pp.239-251.

Stevenson, John, "EU-funded project uses artificial intelligence to tackle terrorist cyber-propaganda", City University of London, 2017.

Stone, P. J., Dunphy, D. C. and Smith, M. S., *The general inquirer: A computer approach to content analysis*, MIT Press, 1966.

Tala, F. Z., "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia", mimeo, 2003.

Tetlock, P. C., "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *The Journal of finance*, Vol.62, 2007, pp.1139–1168.

Turney, P. D. and Litterman, M. L., "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus", CoRR, vol.cs.LG/ 0212012, 2012.

Turney, P. D. and Pantel, P., "From Frequency to Meaning: Vector Space Models of Semantics", *Journal of Artificial Intelligence Research*, Vol.37, 2010, pp.141-188.

US National Library of Medicine, "De-identification tools", accessed December 19,

2018.

Vapnik, V. N., Chervonenkis, A. Y., “A Class of Algorithms for Pattern Recognition Learning”, Avtomat. i Telemekh., Vol.25, No.6, 1964, pp.937-945.

Vozalis, M. and Margaritis, K. G., “Analysis of Recommender Systems’ Algorithms”, In Proc. of the 6th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA), Athens, Greece, 2003.

Wang, Z., Zhang, J., Feng, J., & Chen, Z., “Knowledge graph and text jointly embedding”, Working Paper, 2014.

Young, T., Hazarika, D., Poria, S. and Cambria, E., “Recent Trends in Deep Learning Based Natural Language Processing”, arXiv preprint arXiv:1708. 02709, 2017.

〈부록〉 세종품사태그

대분류	태그	설명	대분류	태그	설명
체언	NNG	일반 명사	선어말 어미	EP	선어말 어미
	NNP	고유 명사	어말 어미	EF	종결 어미
	NNB	의존 명사		EC	연결 어미
	NR	수사		ETN	명사형 전성 어미
	NP	대명사		ETM	관형형 전성 어미
용언	VV	동사	접두사	XPN	체언 접두사
	VA	형용사	접미사	XSN	명사 파생 접미사
	VX	보조 용언		XSV	동사 파생 접미사
	VCP	긍정 지정사		XSA	형용사 파생 접미사
	VCN	부정 지정사	어근	XR	어근
관형사	MM	관형사	부호	SF	마침표, 물음표, 느낌표
부사	MAG	일반 부사		SP	쉼표, 가운뎃점, 콜론, 빗금
	MAJ	접속 부사		SS	따옴표, 괄호표, 줄표
감탄사	IC	감탄사		SE	줄임표
조사	JKS	주격 조사		SO	붙임표(물결, 숨김, 빠짐)
	JKC	보격 조사		SW	기타기호 (논리수학기호, 화폐기호)
	JKG	관형격 조사			
	JKO	목적격 조사	분석 불능	NF	명사추정범주
	JKB	부사격 조사		NV	용언추정범주
	JKV	호격 조사		NA	분석불능범주
	JKQ	인용격 조사	한글 이외	SL	외국어
	JX	보조사		SH	한자
	JC	접속 조사		SN	숫자

Text Mining for Economic Analysis

Soohyon Kim* · Youngjoon Lee** · Jhinyoung Shin*** · Ki Young Park****

Abstract

Text mining or natural language processing is a multi-discipline technique that can distill and obtain useful information from text. With the development of AI and increasing abundance of big data, text mining is becoming one of the essential technologies in the fields of business, governement and academia. In this regard, the purpose of this paper is to provide detailed description of how text data analysis can be used for economic analysis by reviewing the key methodologies and related literature. We expect that text mining will complement the limitation of current data and methodologies by serving as a new source of information.

Keywords: Text Mining, Economics, Machine Learning
JEL Classification: A12, B41, C80

* Economist, Economic Research Institute, Bank of Korea, E-mail: soohyonkim@bok.or.kr
** Precourt Institute for Energy, Stanford University, E-mail: yj.lee@yonsei.ac.kr
*** Professor, Yonsei Business School, E-mail: jshin@yonsei.ac.kr
**** Professor, Yonsei School of Economics, E-mail: kypark@yonsei.ac.kr

지 정 토 론

주 제 : 『거시경제 분석을 위한 텍스트 마이닝』에 대한 논평
논평자 : 김태경 (수원대)

이 글은 텍스트 마이닝에 관련된 여러 내용을 소개하고 이를 경제 분석에 활용하는 사례들을 제시한다. 저자는 텍스트 마이닝의 주요 기법을 비교적 체계적으로 다루고 있으며 워드 임베딩(Word Embedding) 기술을 포함하여 여러 데이터 마이닝 모형에 적용될 수 있는 방법들을 언급한다.

텍스트 사례를 분석하여 결과를 활용하는 일은 상당히 의미 있어 보이며 특히 비정형 데이터 분석의 대상으로 속기록이나 회의록 등을 분석함으로써 경제 현상을 설명할 필요가 있다는 의견은 타당성이 있다. 경제 정책을 담당하는 다양한 주체들이 생산한 비정형 기록물을 분석해서 그것과 관련된 어떤 다양한 결과들과 결부하는 연구는 그 필요성에 비해 축적된 연구성과는 적은 편이다. 이와 같은 측면에서 볼 때, 경제 분석에 텍스트 마이닝을 도입하고자 하는 주장과 취지를 공감한다.

텍스트 분석이 의미가 있으려면 말뭉치 혹은 코퍼스(Corpus)를 수치화하기 위한 기초 자료가 마련되어야 할 필요가 있다. 말뭉치에서 추출한 텍스트 토큰(Token)을 어떻게 평가할 것인가? 평가된 결과는 어떻게 요약할 것인가? 요약된 결과는 어떤 방식으로 벡터로 변환하며, 그 결과를 활용할 수 있는 예측 모형은 어떻게 만들어야 하는가?

담화를 분석하여 주가를 예측하려는 연구 사례는 텍스트 마이닝의 실용적인 사례들 가운데 하나다(Hagenau *et al.*, 2013; Nassirtoussi *et al.*, 2014). 상품이나 서비스에 대한 대중의 선호를 판단하거나(e.g., Mostafa, 2013) 시장의 수요를 예측하는 일에도 텍스트 마이닝 방법은 사용될 수 있다(e.g., Yu *et al.*, 2005). 본 논문은 경제 이슈에 관련된 회의자료 역시 분석이 가능한 데이터로 보고 있으며 상당히 의미가 있는 결과가 기대된다. 앞으로 어떠한 알고리즘을 선택하거나 개발할 것이며, 어떤 문제에 적합할 것인지를 살펴보는 후속 연구가 이어져야 할

것으로 본다.

참 고 문 헌

Hagenau, M., Liebmann, M. and Neumann, D., “Automated news reading: Stock price prediction based on financial news using context-capturing features”, Decision Support Systems, Vol.55, No.3, 2013, pp.685-697.

Mostafa, M. M., “More than words: Social networks’ text mining for consumer brand sentiments”, *Expert Systems with Applications*, Vol.40, No.10, 2013, pp.4241-4251.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. and Ngo, D. C. L., “Text mining for market prediction: A systematic review”, *Expert Systems with Applications*, Vol.41, No.16, 2014, pp.7653-7670.

Yu, L., Wang, S. and Lai, K. K., “A rough-set-refined text mining approach for crude oil market tendency forecasting”, *International Journal of Knowledge and Systems Sciences*, Vol.2, No.1, 2005, pp.33-46.

지 정 토 론

주 제 : 『거시경제분석을 위한 텍스트 마이닝』에 대한 논평
논평자 : 최동욱 (상명대학교)

텍스트데이터는 소위 빅데이터의 일종이다. 빅데이터의 정의에 대해 논할 때 다양한 소스로부터 얻은 대규모의 데이터라는 직관적인 개념도 중요하지만 조사자료와 같이 연구자가 계획적으로 만들어낸(made data) 자료가 아닌 인간의 선호와 감정 등이 자연스럽게 행태로 드러난(found data)자료라는 점이 매우 중요하다(Connelly *et al.*, 2016). 이는 현시선호(revealed preference)와 같이 보수적인 기준을 선호하는 경제학 연구자들이 빅데이터에 관심을 갖게 되는 중요한 이유라고 할 수 있다. 또한 빅데이터 분석방법은 알고리즘을 통해 인간의 개입을 최소화한 객관적인 근거를 만들어 낼 수 있다는 점, 컴퓨터를 이용하여 대량의 자료를 처리가능하다는 점 등이 기존의 사회과학 분석 방법에 대비해서 가질 수 있는 중요한 장점이라고 할 수 있다. 텍스트데이터도 이러한 성격을 모두 만족하고 있으며 경제학 연구에 활용될 수 있는 잠재력에 대해 많은 연구자들이 관심을 가지고 있다. 이러한 관점에서 텍스트마이닝의 활용에 대한 논의가 더 발전하기를 기대하면서 논문에서 몇 가지 고려해야 할 사항들에 대해 첨언을 하고자 한다.

저자들이 아우구스티누스와 비트겐슈타인을 인용하여 텍스트마이닝을 소개한 부분은 매우 인상적이다. 저자들은 단어와 의미의 관계 - 소쉬르에 따르면 기표(signifiant)와 기의(signifie)에 해당하는 - 에서 일물일어(一物一語)의 법칙이 아닌 언어게임(language game)의 성격을 강조한다. 언어의 의미 파악에 있어서 단순히 사전(dictionary)을 기반으로 하는 것이 아니라 상황과 문맥을 고려해야 한다는 뜻이다. 이러한 논리로 저자들은 머신러닝과 같은 기술을 활용한 텍스트 분석의 필요성을 강조한다. 이러한 주장이 인상적인 이유는 텍스트 분석을 경제학에서 활용할 때 단지 공학적인 이해뿐만 아니라 언어학에 대한 기초적인 이해도 필요하다는 점을 보여주고 있기 때문이다. 텍스트분석이 학제간 연구로서 다루어져

야 하는 이유이기도 하다.

경제학자의 입장에서 볼 때 언어학에서 배울 수 있는 흥미로운 법칙은 언어학자 폴 그라이스(Paul Grice, 1913-1988)의 '양의 격률maxim of quantity'과 '관련성의 격률maxim of relevance'이다. 이는 각각 인간이 말을 하거나 글을 쓸 때 합리적으로 정보의 양을 선택한다는 가정, 그리고 항상 주제나 문맥과 관련된 내용을 선택한다는 가정을 뜻한다. 즉 이는 언어학에서 다루는 인간의 합리성 가정으로서 경제학의 합리적 선택 가정에 상응한다고 볼 수 있다. 사실 우리가 텍스트 데이터를 활용할 때는 이미 이러한 가정을 전제한다고 볼 수 있다. 관심 대상인 경제주체가 어떤 표현을 과도하게 한다거나 혹은 과도하게 언급을 피하는 것은 모두 합리적인 이유가 있으며 해당 주체의 선호나 심리, 성향을 반영하고 있다고 해석할 수 있다는 뜻이다. Gentzkow and Shapiro(2010)는 이러한 점을 뉴스기사의 정치성향을 판별하는데 활용하였으며 최동욱(2017)에서는 Gentzkow and Shapiro(2010)의 방법론을 한국어에 적용한 사례를 제시하였다. 정치적 성향을 드러내는 구문을 판단하기 위해 아래와 같은 χ^2통계량을 활용했는데 여기서 분자의 $f_{pr}f_{\neg pd}-f_{pd}f_{\neg pr}$값은 주어진 구문p의 보수성향을 의미한다. 이 값은 해당 표현을 보수적 성향의 인물이 언급한 빈도(f_{pr})가 높을수록, 진보 성향의 인물이 언급한 빈도(f_{pd})가 낮을수록 증가한다.

$$\chi_p^2=\frac{(f_{pr}f_{\neg pd}-f_{pd}f_{\neg pr})^2}{(f_{pr}+f_{pd})(f_{pr}+f_{\neg pr})(f_{pd}+f_{\neg pd})(f_{\neg pr}+f_{\neg pd})}$$

특기할 만한 점은 진보성향의 인물이 언급하지 않는 빈도($f_{\neg pd}$)가 높아져도 이 값이 증가한다는 점이다. 예컨대 Gentzkow and Shapiro(2010)가 소개한 2005년 당시 미국 의회의 사례를 보면 공화당 의원들은 의도적으로 상속세(estate tax)를 "사망세(death tax)"라고 표현했으며 민주당 의원들은 의도적으로 그러한 표현을 사용하지 않았다. 그리고 실제 데이터에서도도 "death tax"가 포함된 표현은 당시 정파를 구분할 수 있는 지표 중 하나로 나타났다. 이러한 사례는 발화에 있어서 합리적으로 정보량을 선택한다는 폴 그라이스의 법칙이 반영된 것이라고 해석할 수 있다. 다만 주의해야할 점은 "언급하지 않은 빈도"를 계산하기 위해서

는 전체 표현(단어, 어구 등)을 포함한 집합이 닫힌집합(closed set)이어야 하며 이 집합의 구성 및 크기를 결정하는 과정에는 연구자의 자의적 판단이 개입되어야 한다는 점이다.

저자들은 본 논문을 통해 텍스트 분석을 위한 다양한 방법들을 소개했지만 연구자들이 실제 구현할 때 유의해야 할 문제들에 대해서는 앞으로 더 많은 논의가 필요하다. 특히 텍스트마이닝 기법에 쓰이는 알고리즘들이 완벽하게 객관적인 결과를 생성해주는 것이 아니라 항상 일정정도는 연구자의 자의적 선택을 요구한다는 점에 유의해야 한다. 결과를 해석하는 입장에서는 이러한 자의성이 분석 결과에 영향을 끼칠 가능성에 대해서 인식할 필요가 있고, 따라서 방법론을 논의할 때 이에 대한 언급이 필요하다. 예컨대 불용어사전을 구성할 때 특정 단어의 포함여부와 관련해서 연구자의 자의적인 선택이 개입될 수 있으며, 감성/성향/긍부정 등의 구분 역시 그렇다. 저자들의 분석 사례에서 시장적 접근방법을 사용하여 단어의 논조를 구분할 때 seed 단어의 선정에도 일정부분 자의적 판단이 필요했을 것으로 보인다. 최동욱(2017)의 사례에서도 정치적 표현들의 전체 집합을 구성할 때 그 범위의 설정도 어떤 객관적 기준이 존재하는 것이 아니라 연구자의 자의적 선택에 따른 것이다. 머신러닝의 지도학습에서도 훈련데이터셋의 설정에 따라 결과가 바뀐다는 사실은 잘 알려져 있다. 이 때 이런 기준의 변화에 따라 분류결과가 민감하게 변동한다면 기법들의 신뢰성에 대해 의문을 가지게 될 것이다.

비지도학습처럼 인풋에서 자의성이 개입되지 않는 경우라 하더라도 여전히 문제는 존재한다. 예컨대 토픽모형의 경우 분류된 결과에 대해 각 그룹을 명명하고 해석하는 단계에서 연구진의 자의성이 개입된다. 본문 <표 2>의 LDA 분류결과에서 각 토픽그룹을 명명하는 것이 그러한 사례다. 첫 번째 그룹의 이름을 Foreign Currency라고 설정했을 때 그 이름이 포함된 단어들의 성격을 얼마나 잘 반영하는지 객관성에 대한 의문이 제기될 수 있다. 이는 데이터마이닝 기법 중 하나인 군집분석과 마찬가지의 문제라고 할 수 있다. 이처럼 텍스트마이닝 기법에 대해 논의할 때 자의성이 개입될 수 있는 지점을 명확하게 설명해준다면 앞으로 발전적인 논의를 이끌어 가는 데 더 많은 도움이 될 것이다.

데이터는 주어진 연구문제를 해결하기 위한 사용하는 도구의 하나일 뿐이다.

하지만 많은 경우 데이터의 사용가능성(availability)이 어떤 문제를 정의할 때 실질적인 제약조건으로 작용한다. 따라서 데이터의 범위가 확장되었다는 것은 곧 해결가능한 문제의 범위가 확장되었다는 의미이기도 하다. 빅데이터가 경제학에 기여하는 점은 이렇듯 상상력의 범위를 확대했다는 것이라고 할 수 있다 (Stephens-Davidowitz, 2017). 이러한 관점에서 볼 때 경제학에서 텍스트데이터의 활용가능성을 소개한 본 논문의 기여는 명확하다고 할 수 있으며 이를 바탕으로 경제학계에서 더 발전된 논의가 이루어질 수 있기를 기대한다.

참 고 문 헌

최동욱, 「인터넷 포털의 경쟁과 뉴스 편향도의 선택」, 『산업조직연구』, 제 25권 2호, 2017, pp.1-40.

Connelly, R. *et al.*, “The role of administrative data in the big data revolution in social science research”, Social Science Research. 2016.

Gentzkow and Shapiro, “What drives media slant? Evidence from U.S. daily newspapers”, *Econometrica*, Vol.78, No.1, 2010, pp.35-71.

Stephens-Davidowitz, Seth, Everybody lies: Big data, new data and what the internet can tell us about who we really are, Harper Collins, New York. 2017.

일 반 토 론

주 제 : 거시경제 분석을 위한 텍스트 마이닝

사회자(김동헌) : 저도 한 가지 질문을 드리고 싶은데요. 지금 사실 비슷한 맥락인 것 같습니다. 비용 편익과 비슷한 맥락인데 사실 중요한 것은 우리가 정형적인 데이터를 많이 가지고 경제학에서 분석을 많이 시도하고 있는데 정형적인 데이터에서 어떤 부분들을 비정형데이터로 개선시켜 나가고, 예를 들어 저희가 한 번 구글트렌드 데이터를 가지고 조기경보지수를 보완하는 연구를 했었을 때 조금 개선되는 결과를 내본 경험이 있습니다. 여기에서도 지금 앞에 기본적인 설명들을 아주 잘 해주신 것 같아요. 방법론을 그런데 정말 우리가 더 중요한 것 중의 하나는 어떤 부분에 응용했을 때에 그 응용된 부분에 좀 더 자세히 설명을 해주셔서 정형적인 부분이 이렇게 돼 있는데 여기에 비정형이 어떻게 들어와서 개선돼서 이런 쪽에 많은 보완을 하고 있다. 조금 더 이쪽에 관심을 갖는 분들한테 실제 예를 자세히 제시해 주시면 많은 도움이 될 것이라는 생각이 들었습니다. 그러면 플로어에서 질문을 받도록 하겠습니다.

전현배(서강대) : 저도 상당히 재밌게 들었는데요. 경제학에서 제일 걱정하는 것이 사실은 규범적인 부분이나 실증적인 부분이잖아요. 연구자의 시각이 들어가는데, 이게 실증분석을 한 것이라고 생각한다면 보통 과학성을 담보하기 위해서 저희가 복제(replication)를 할 수가 있어야 되거든요. 다른 시각을 가진 사람이라도 데이터가 프로그램이 있으면 똑같은 결과가 나와야 되거든요. 근데 지금 이런 류의 분석에서는 과연 시각이 다른 사람들이 똑같은 작업을 했다고 해도 결과가 다시 재현될 수 있느냐 이것이 과학성의 근본인데 그 부분이 어떻게 가능할 수 있을지에 대해서 말씀해 주셨으면 좋겠습니다. 그래서 규범적인 결과가 실증적인 것으로 포장이 되면 경제학에서는 그것을 가장 두려워하고 있기 때문에 과학적으로 재현하고 그 다음에 모든 사람들에 의해서 객관성이 담보될 수 있는 그런 것들에 대해서 어느 정도 진행이 되었는지 좀 알려주시면 감사하겠습니다.

김동헌(고려대) : 또 혹시 플로어에서 질문 있으십니까.

이명헌(인천대) : 전혀 모르는 분야인데 흥미 있게 들었고요. 이게 지금 보여주신 것처럼 한국은행의 커뮤니케이션 이런 것이 어떤 영향을 주냐는 이런게 굉장히 현실과 밀착된 연구인데 경제학이나 경영을 넘어서는 분야에서 완전히 문외한 입장에서 생각해보면 경제사나 경제사상사, 경제학설사 같은 데서 이용될 수 있는지 궁금합니다. 예를 들어 더 올라가면 가장 연구가 많이 된 텍스트는 성경일 텐데 성경 신학에서 가장 중요한 논쟁거리 중 하나가 이 글을 쓴 사람이 저 글도 썼느냐 하는 질문들이거든요. 이것이 경제학설사 같은 데서도 그럴 수 있는데 지금 말씀하신 이런 방법론들이 두 개의 서로 다른 저작이 같은 사람이 썼는지 또는 두개의 서로 다른 저작이 같은 영향권에서 상호작용을 주고받은 사람끼리 쓴 텍스트인지 그런 것을 식별 가능한지 궁금합니다. 왜냐하면 그것이 경제학설사 같은 것으로 내려오게 되면 근대 것들은 다수의 동의(correspondence)가 있지만 거슬러 올라가면 그런 것이 없기 때문에 서로 다른 학자들 사이에 영향력이 있었는지 이 사람이 저 사람 텍스트를 읽은 적이 있는지 이런 것들이 굉장히 중요한 주제가 되기 때문에 혹시 그런 연구된 사례가 있는지 궁금합니다.

사회자(김동헌) : 네, 한두 분 더 질문

연태훈(금융연구원) : 질문이라기보다 현실적인 요청을 좀 드리겠는데요. 어쨌거나 분석패널에 기고한 논문으로서 굉장히 많은 분들이 관심을 가지고 있고, 정말 도움이 되는 페이퍼이긴 한데 뭔가 좀 응용적인(innovative) 부분을 조금 추가해 주실 수 있으면 좋겠어요. 지금 토론자분들이 말씀하셨듯이 대한민국 언어로 된 경제분석에서 열거하시고 정리해 놓으신 부분들에서 어떤 부분에 추가의 노력이 필요하고 어떤 부분에서 이런 고민들이 필요하다는 정리들이 정리될 수 있으면 대부분의 서베이 페이퍼가 그렇듯이 논문으로의 구성을 갖추게 되는데 아직까진 그부분이 조금 부족하다는 느낌을 제가 피할 수가 없어서 그 부분의 답변을 지금 당장 주시는 것보다는 토론자 분들하고 좀 더 말씀을 나누셔서 그런 내용이 좀 추가될 수 있으면 좋겠다는 희망사항을 말씀드리겠습니다.

박정수(서강대) : 저도 연 박사님하고 거의 같은 의견이고 한국경제 분석패널이기 때문에 저도 약간 걱정을 하는 부분이 있습니다. 사실 이 논문이 현재로썬 정말 훌륭한 가치가 있다고 보지만 문헌정리가 상당히 되어 있는 것이고 소개가 주된 내용입니다. 여기 한국 사례라도 조금이나마 분석을 해주셨으면 하는 바람에서 혹시 2019년A 2019년B를 봤더니 2019년A는 이미 발간된 것이고 그 두 페이퍼의 연장(extension)이나 해서 한 예시를 보여주는게 어떨까 합니다. 그 분석을 통해서 한국경제에 이런 것을 적용해서 분석했다 이런 것을 추가를 해주시면 좋을 것 같습니다. 한 가지 아까 그 이명헌 교수님께서도 말씀하셨지만 동일한 사람이 발언한 것을 확인할 수 있는지도 궁금하거든요. 인터넷상에서 댓글이 많은데 이런 것도 분석을 할 수 있지 않을까 해요. 또 혹시 매크로로 생성된 댓글인지 아닌지를 판별하실 수 있는지 궁금합니다.

김용진(아주대) : 제가 하나만 물어볼까요? 제가 얼마전에 이런 연구를 발표하는 것을 봤거든요. 그리고 최근에 제가 심리학 페이퍼를 봤어요. 그 페이퍼에 예를 들면 심리가 어떻게 전달이 되나 이런 건데 난 사실 내가 아는 연구에 버블에 관련돼서 우리가 신고전학파적인(neo classical) 매커니즘으로 설명하는 것이 있고 또 우리가 설명을 못하면 심리학 쪽에서 심리 측면에서 설명하는 것이 있거든요. 내가 생각하는 것 중에 하나가 한국은행도 통화정책을 DSGE 갖고 멋지게 이야기 하지만 많은 경우를 보면 사실 이 버블현상에 있어서 시장의 감정(market sentimental)을 조절하는 것이 많거든요. 제가 하나 제안하고 싶은 것은 이런 빅데이터를 사용하고 대중들의 심리를 잡아내는 키워드를 추출하고 이런 것을 가지고 시간별 빈도(frequency)가 어떻게 변하는가를 가지고 심리를 측정을 한다면 그 심리에 따른 통화정책이라든지 또는 시장(market)의 자산 가격(asset price)이 어떻게 반응하는지 이런 것을 보면 재밌겠다는 생각이 듭니다.

사회자(김동헌) : 네. 추가적으로 또 질문이 있으신지요. 네 질문이 없으시면 발표하신 김수현 박사님께서 답변을 해주시고요. 마무리하도록 하겠습니다.

신진영(연세대) : 제가 이 연구를 어떻게 시작했고 그 백그라운드를 말씀드리겠

습니다. 연구는 제 지도학생이었던 이영준 박사가 원래 컴퓨터공학 전공이에요. 그래서 금융 쪽을 하다가 연구주제를 찾다가 마침 텍스트마이닝 관계된 강의를 듣고 이게 적용이 되겠다고 했고, 금융 쪽의 주제를 찾다 보니까 박기영 교수님이 거시를 하시고 김수현 박사랑 해서 결국 한국은행 쪽 의사록 분석, 그리고 이게 해외에서도 보면 FOMC분석과 ECB분석이 돼 있어서 이쪽 주제로 맞다고 되는데 아까 말씀하신 그 문제가 그대로 있어요. 이게 처음에 저희가 할 때는 우리말로 금융쪽에서 분석을 한 것이 없어서 구글 번역기로 의사록을 번역하고 그러면 영어는 다 돼있어요. 사전도 많고 워낙 있기 때문에 그걸로 해서 한건데 이렇게 되면 우리 말의 맥락(context)이 사라져서 사실은 아까 중간에 사전 이야기를 했는데 이영준 박사하고 김수현 박사가 사실은 사전을 만들었어요. 그럴 수밖에 없는게 이게 우리말은 굉장히 어려운 말이고, 한국은행의 금융통화위원 분들이 말을 굉장히 우리 말을 알아들을 수 없게 하시잖아요. 아닌게 아닌거고 뭐 이렇게 되니까, 이게 분석이 어렵고, 아까 n이 5까지 가고 그 다음에 이제 여러 사전을 했지만 서울대학에도 있고 한데 이게 금융쪽을 한게 없어요. 금융은 우리끼리는 쓰는 용어지만 아까 왜 금통위 뭐 이런거 없어요. 그러니까 그걸 결국은 사전을 편찬하는(build up)하는 작업을 하는데 그걸 그렇다고 그걸로 학위논문을 받는 것도 아니고, 그걸로 퍼블리쉬가 되는 것도 아니고 그래서 우리가 어떻게 할까 하다가 방법을 알리고 그래서 뒷부분의 용역 말씀은 저희가 보완을 해드리지만 사실은 이게 비정형 데이터이고 중요한게 뭐냐면 모든 사람이 합리적으로 얘기하지는 않아서 그리고 경우에 따라선 이게 뭔가 뒤에 있지만 말을 할 때는 그게 나오지 않지만 간접적으로(circuitous) 드러나는 것을 어떻게 잡아낼지가 이런 부분에서 가능하기 때문에 사실은 활용은 될 데가 많아요. 지금 김수현 박사도 그렇고 이영준 박사도 그렇고 공동 또 각자하는 것 중에는 기업 실증에 대한 애널리스트 리포트를 가지고 기업 실증을 해요. 왜냐면 저희가 그 재무 쪽에서 보면 애널리스트 리포트들이 대부분 바이 레코멘데이션(buy recommendation)이에요. 아니면 중립이에요. 그런데 물어보면 왜 셀 리포트(sell report)가 없냐 하는데 중립을 셀(sell)로 보시면 됩니다 해요. 결국은 맥락을 보면 나오거든요. 그런 것들이라든지 경기 예측에 관계된 것들 왔다 갔다하는 신문 기사라든지 이런 것들을 보면 나오는 것들에 대해서 활용범위는 많고, 말씀하셨다시피 이게 비정형 데

이터고 계속해서 이게 만들어지는(build up) 과정에서 일정기간에 AI에서 하는 트레이닝 시키고 그걸 다시 하는건데 이게 계속 무빙이 되기 때문에 그 부분은 이게 한 두 학자가 하기보다는 사전을 결국 퍼블릭 도메인(public domain)에 올려놨고, 누군가 그걸 가지고 업데이트하면서 되는 과정에서 말씀드렸다시피 누가 이걸 보느냐에 따라서 주관적으로 왜곡되는 것은 거기서 나름 완화될 수 있는 여지가 있는 것 같아요. 그래서 아까 말씀하셨지만 이것을 공짜로 올려놓는다고 하지만 그 과정에서 검증되고 정화되는 부분이 있기 때문에 이 부분은 연구영역이 어렵고. 그리고 우리 말이 굉장히 어려워요. 이걸 먼저 처음에 하신 분은 문헌정보학과의 이쪽 일 많이 하신 분인데 이분은 영어로만 논문을 쓰세요. 우리말이 노력이 너무 많이 들고 결과도 잘 안나오고 그래서. 테크니컬한 부분은 김수현 박사가 말씀해 주시겠습니다.

발표자(김수현) : 신진영 교수님 많은 부분을 대신 답변해 주셔서 감사합니다. 저는 그럼 구체적인 답변을 드리겠습니다. 최동욱 교수님 전체적으로 많은 관심을 가져 주시고 코멘트 주셔서 감사합니다. 첫 번째 최동욱 교수님 말씀에 대해서 답변을 드리면요. N-gram부분에서 품사에 제한을 주면 차원을 줄이기 때문에 효율성이 높아진다, 그 부분 맞습니다. 그래서 저희가 이제 분석을 할 때 품사를 5가지로 제한을 합니다. 이는 명사, 동사, 형용사, 부사입니다. 그거 하고 부정어(negation)까지 넣어서 5개로 축약을 하면 차원(dimension)이 많이 줄어듭니다. 그 5가지만 분석하고 있기 때문에 이미 실제(practice)에서 적용을 많이 하고 있는 부분이고요. 방법론 평가 문제가 텍스트 마이닝뿐만 아니라 머신 러닝 전체적인 문제입니다. 사실 이게 검증이나 평가(evaluation)를 어떻게 하느냐, 특히 신경망 부분에서 블랙박스(black box) 아니냐 하는 비판들이 많이 있습니다. 그 부분 역시 텍스트 마이닝에서 해결해 나가야 할 문제라고 생각하고요. 일단 머신러닝에서 쓰는 오차행렬(confusion matrix)이라고 하죠. 그걸 갖고 평가를 하고 있는 것으로 알고 있습니다. 그 이상의 것은 제가 구체적으로 찾아보겠습니다. 세 번째 객관적 증거일 수 있으나 자의적인 과정이 들어갈 수 있다. 맞는 말씀입니다. 제가 지금 개인적으로 하고 있는 것은 북한 말을 가지고 하고 있는데 북한말은 품사분석기가 인식을 못합니다. 그래서 그 북한말을 남한어로 바꿔줘야 되거든요.

그 바꿔주는 과정에서 최대한 객관적으로 하겠지만 번역하기가 힘든 것은 제가 아는 단어를 넣었습니다. 그런데서 자의성이 들어갈 수가 있고요. 그런 부분이 있고, 형태소 부분에도 그렇습니다. 아까 신진영 교수님께서 잘 말씀해 주셨지만 지금 현재 아까 말씀드린 eKoNLPy가 올라가 있거든요. 이영준 박사님이 Github에 올리셨는데 저희가 논문을 다 쓰고 출간되고 난 다음에 저희가 발견을 했지만 사소한 문제가 있었던 것 같습니다. 왜냐면 형태소 분석기라고 하는 것은 애초에 만들어질 때 머신 러닝을 통해서 알고리즘으로 만들어지는데요 eKoNLPy 그것을 업데이트를 한 것이거든요. 그러다보니 충돌되는 부분에서 나오는 문제점이 있을 수 있습니다. 그래서 형태소분석을 하면 예시해드린 그런 문장에서, “한국은행이 금리를 1.5% 올렸다.” 거기 형태소 분석에서는 공공요금이라는 형태소가 나옵니다. 그런 것을 수동적으로 저희가 다 찾아서 해결을 했겠지만 미흡한 부분이 있고 그런 부분을 보시고 찾으셔서 업데이트를 해주시는 것은 연구자들에게 부탁드리는 부분이기도 합니다만 저희도 계속 노력을 하겠습니다. LDA 말씀하셨는데 해석이 어렵다. 물론 그렇습니다. 그게 하이퍼 파라미터(hyper parameter)에 따라서 주제의 개수같은 것을 정하는 주제가 몇 개까지 찾을(detect) 거냐는 것을 미리 정하는데 하이퍼 파라미터(hyper parameter)를 몇 개로 정하냐에 따라서 이게 좀 예민하거든요. 그래서 그 하이퍼 파라미터(hyper parameter)도 추정하는 방법으로는 non-pararmetric Bayseian 방법이 있습니다. 그걸로 하면 파라미터 개수도 추정을 해줍니다. 그런 방식으로 해서 회색지대를 줄일 수 있을 것 같습니다. 그리고 김태경 교수님 코멘트에 대해 말씀드리겠습니다. KoNLP를 만드셨다고 들었습니다.

김태경(수원대) : 콜라보레이션을 같이 했어요.

발표자(김수현) : 아 library를 만드셨다고요. 만나 뵙게되어 영광입니다. 경영학에서 주로 연구를 하고 계시다고, 네 그래서 요즘에 텍스트마이닝이 인공지능 분야에서도 각광을 받고 있는 것이 최근에 그 컴퓨터공학에서의 노벨상이라고 불리는 튜링상 수상자들 3분이 전부 다 텍스트마이닝 하시는 분들이거든요. 굉장히 트렌드를 잘 반영하고 있다고 볼 수 있습니다. 전산언어학(computational

linguistics)이라고 해서 미국에서는 주로 스탠퍼드 대학에서 많이 하는데 거기서 자연어 처리(natural language processing)이라고도 많이 합니다. NLP라고도 많이 쓰거든요. 그래서 NLP라는 말이 들어가는데 한국에서는 이제 학교에서 전산언어학이라는 말을 많이 쓰는 것 같아요. 서울대에서도 전산언어 연구소라고 하거든요. 신효필 소장님이 꽤 오래 연구를 하고 계시고, 비용적인 것을 고려해야 된다. 텍스트데이터를 잘못 검색하게 되면 돈을 많이 줘야 합니다. 로이터 이런데서 API를 쓴다거나 하면 한달에 200만원 300만원씩 줘야하는 문제가 있기 때문에 이걸 잘 피해가는 것이 텍스트 분석을 하는데 있어서 과제입니다.

김동헌 교수님 질문에 대해서 말씀드리겠습니다. 비용 편익 부분을 말씀하셨는데 어떤 부분이 부족한가? 구글 트렌드로 해보셨다고 해서 잘 아시겠지만 정형적 데이터가 잡지 못하는 것은 타이밍인 것 같습니다. 뭔가를 예측하고자 할 때 우리가 어떤 이벤트가 생겼다고 할 때 통계자료는 최소 한달 내지 분기 뒤에 나오기 때문에 시의성에 대한 부분이 있고요. 두번째로 기존 데이터들은 범주가 정해져 있습니다. 예측을 하고자 할 때 데이터를 쓰고자 하면 정해진 범주의 데이터만 쓸 수 있습니다. 그래서 그 데이터는 함의하고 있는 뜻이 그 범주에 대한 것이지 여러 가지 정보를 담고 있지 못합니다. 그런 측면에서 텍스트데이터는 좀 더 폭이 넓고 시의성 있는 자료가 될 수 있다는 것이고요. 그리고 두 번째는 프리퀀시를 따질 때 텍스트 데이터는 연구자 임의로 프리퀀시를 정할 수 있습니다. 기존 데이터는 한달이면 한달, 분기면 분기 이렇게 돼있어서 그 프리퀀시를 바꿀 수 없는데 텍스트데이터는 하루 혹은 하루 반 이런 식으로 끊어서 얼마든지 프리퀀시를 만들 수 있다는게 장점으로 볼 수 있을 것 같습니다. 답변이 됐는지 모르겠습니다.

전현배 교수님 답변 드리겠습니다. 실증분석에서 복제가 가능해야 된다. 아까 신진영 교수님이 말씀해주신 대로 저희 사전으로 올렸고요. 형태소 분석기뿐 아니라 감성 사전도 올라가 있습니다. 의사록은 당연히 한국은행 홈페이지에 올라가 있고요. 만약에 의사록을 다운받으실 수 있으면 이영준 박사님이 올려놓은 라이브러리(library)를 가지고 복제(replication) 할 수 있도록 설명도 다 올려놨습니다. 그래서 해보시면 될 것 같습니다. 실증분석파트는 굉장히 쉬운 회귀분석이기 때문에 그거는 Matlab이나 Stata로 해보시면 될 것 같고요.

그리고 이명헌 교수님께서 질문해 주셨는데, 경제학설사 부분에 인용은 가능합니다. 실제로 저희 논문에 보시면 도구변수를 누가 처음 발명을 했느냐에 대해서 그 책의 부록에 도구변수가 들어있는데요. 사람들은 그 부록은 아들이 썼다. 두 분이 다 경제학자인데, 제가 지금 성함은 기억이 안 납니다만 아들이 썼다라는 이야기가 있고, 본문과 부록까지도 그 아버지 경제학자가 썼다는 부분이 있었는데 그걸 텍스트마이닝으로 답을 냈습니다. 부록도 아버지가 다 쓴게 맞다. 그렇게 해서 최초 개발한 사람이 아버지 경제학자다.

플로어 : 그게 맞는지는 실험적으로 검증할 방법이 있나요?

발표자(김수현) : 네 그것은 교과서를 보시면 많이 나와 있습니다. 가장 처음 보시는 예제들이거든요. 샬럿 브론테의 소설 제인 에어 이런거 라이브러리에 다 올라와 있습니다. 샬럿 브론테의 소설을 모두 다운로드 받아 소설간 어체, 주제 유사성 분석 등을 해보거든요. 임의의 텍스트를 받았을 때 저자가 누구이냐, 이런 것들까지 찾을 수 있고요 그런 종류의 예제가 많이 있습니다. 물론 같은 방법론을 학설사에서도 많이 사용할 수 있지 않을까하는 생각이 듭니다.

응용적인 부분도 추가해주면 좋겠다고 말씀해 주셨는데요, 경제학에서 텍스트 분석의 과제에 대해서 좀 더 고민해보도록 하겠습니다. 그리고 박정수 교수님 말씀해 주신 것도 비슷한 맥락이라고 제가 이해를 했습니다. 예시를 찾을 수 있는지 한번 생각해 보겠습니다. 그리고 마지막으로 통화정책의 키워드와 시장의 감성(sentimental)을 말씀 해주셨는데, 어떤 특정한 키워드에 집중해서 통화정책 효과를 유도할 수도 있을 것 같습니다. 그런 특정 키워드나 주제에 관련된 감정(sentimental)을 측정하는 기법이 있습니다. Aspect-based Sentiment Analysis 라고 하여서 소위 ABSA라고 부릅니다. 통화정책 쪽에서는 그런 기법을 활용한 워킹페이퍼는 아직 없는 것 같습니다.이런 기법을 써서 미국 FOMC 의사록을 분석하는 것이 저희 공저자분들과 추진하고 있는 주제입니다. 답변이 되었으면 좋겠습니다. 감사합니다.

사회자(김동헌) : 네 감사합니다. 저희가 텍스트마이닝에 대한 주제가 우리 분석

의 틀에서 조금 더 일찍 다뤄졌으면 하는 생각을 했는데요. 그럼에도 불구하고 이런 시점에서 중요한 이 주제를 가지고 같이 논문도 하게 되고 토론도 하게 돼서 유익한 시간인 것 같습니다. 마무리하겠습니다. 발표해주신 김수현 박사님 그리고 토론해주신 최동욱 교수님 김태헌 교수님 모두 감사드리고요. 저희가 시간이 조금 경과됐지만 한 5분 정도 쉬신 다음에 다음 세션을 진행하도록 하겠습니다.