

경제 분석을 위한 텍스트 마이닝

김수현*, 이영준**, 신진영***, 박기영****

이 연구내용은 집필자 개인의견이며 한국은행의 공식견해와는 무관합니다.
따라서 본 논문의 내용을 보도하거나 인용할 경우에는 집필자명을 반드시
명시하여 주시기 바랍니다.

* 한국은행 경제연구원 국제경제연구실(전화: 02-759-5362, E-mail: soohyonkim@bok.or.kr)

** Precourt Institute for Energy, Stanford University(E-mail: yj.lee@gmail.com)

*** 연세대학교 경영대학(전화: 02-2123-5467, E-mail: jshin@yonsei.ac.kr)

**** 연세대학교 경제학부(전화: 02-2123-5495, E-mail: kypark@yonsei.ac.kr)

본고 작성에 도움을 주신 경제연구원 신운 원장님, 이재량 부원장님과 국제경제연구실 유복근 실장님께 감사드립니다. 아울러 유익한 논평을 주신 경제통계국 손원 차장님, 대외경제정책연구원의 김경훈 부연구위원님, 또한 익명의 심사위원님께서도 감사드립니다. 남은 오류들은 저자들의 책임입니다.

차 례

I . 서론	1
II . 텍스트 마이닝과 텍스트 데이터	3
III . 텍스트 마이닝 방법론	6
IV . 텍스트 마이닝을 활용한 경제분석	32
V . 결론	42
참고문헌	44
부록	53

경제 분석을 위한 텍스트 마이닝

본고에서는 텍스트 분석이 이루어지는 과정, 주요 방법론, 경제 분석에의 적용 사례 등을 상세히 기술하였다. 텍스트 마이닝이란 빅데이터 분석 또는 데이터 과학의 대표적인 분야이며, 머신러닝 등 전산기술을 활용하여 언어에 담긴 정보를 정형적 데이터로 추출해내는 과정이다. 언어는 다의성과 비정형성 등 고유한 특성이 있어 이로부터 원하는 정보만 추출해내기는 쉽지 않다. 텍스트 마이닝을 활용할 경우 기존 경제학적 방법으로 예측이 어려운 실시간 예측 및 불확실성 측정 등이 가능하며, 중앙은행 커뮤니케이션 분석 등 기존 모형의 적용이 어려웠던 주제들에 대한 분석도 가능하다. 전산기술의 발달과 동시에 텍스트 마이닝 기법은 빠른 속도로 발전해가고 있으며, 최근 학계와 중앙은행의 관심도 지속적으로 증가함에 따라 경제학 분야에서도 텍스트 마이닝을 활용한 연구가 주요 학술지에 게재되고 있다. 경제학적 분석에 주로 사용되는 구조화된 수치자료의 한계를 보완함과 동시에 새로운 정보의 원천으로써 텍스트 데이터의 수요는 더욱 증대될 것으로 기대된다.

핵심 주제어: 텍스트 마이닝, 머신러닝, 경제 분석

JEL Classification Numbers: A12, B41, C80

I. 서론

Nate Silver¹⁾는 트위터(Twitter) 게시물을 분석하여 2008년 미국 대통령 선거 결과, 2010 총선, 2012 대선 결과 등을 그 어느 예측가보다 정확히 예측할 수 있었다. 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 인공지능만으로 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 문헌정보학에

1) 미국의 정치평론가, 언론인으로 선거와 정치에 관련된 웹사이트 FiveThirtyEight을 운영한다. 저서로는 “The Signal and the Noise” 등이 있다.

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

서 텍스트 데이터는 오래전부터 연구의 대상이었다. 또한 정치학, 사회학에서도 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 분류모형³⁾(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 연구에도 텍스트 마이닝 기법에 대한 수요가 있을 전망이다. 따라서 본고에서는 텍스트 마이닝이 경제학에 활용될 수 있는 방법론과 사례에 대해 상세하게 논하여 관련 연구

3) 나이브 베이즈 분류모형은 머신러닝의 한 방법으로 베이저안 정리를 응용한 분류기법이다.

에 도움이 되고자 하였다.

본고는 다음과 같이 구성되어있다. 2절에서는 텍스트 마이닝의 배경과 정의에 대해 알아보고 3절에서는 텍스트 마이닝의 방법론에 대해 논하고자 한다. 4절에는 텍스트 마이닝을 활용한 경제학적 연구들을 소개하며 5절에서는 텍스트 마이닝을 활용한 데이터 분석의 한계와 경제학 분야에서 향후 텍스트 데이터 활용 방안 등에 대해 논의한다.

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 사전 정의된 단어로 구성된다고 하였으며, 이렇게 구성된 언어는 만물과 일대일 대응구조를 갖는다고 주장한다. 가까운 사례로 우리가 흔히 접하는 어학사전에는 단어(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 예

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 일맥상

로 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 과정보다 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 인공지능망 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 여러 학문분야가 복합적으로 구성되어 텍스트 마이닝 방법론으로 발전하였다.

오늘날 텍스트 마이닝은 빅데이터 분석 또는 데이터 과학의 대표적인 한 분야⁶⁾로 자리매김하였으며, 머신러닝 등 전산기술을 활용하여 언어에 담긴 정보를 정형적 데이터로 추출해냄으로써 여러 학문분야에 기여하고 있다. 또한 텍스트 마이닝

통한다. 아우구스투스가 당위적(sollen) 언어관을 주장하였다면 비트겐슈타인은 언어의 실존(sein) 측면을 중시하였다고 볼 수 있다.

- 5) 언어에 관한 비트겐슈타인의 관찰은 논리실증주의에 기반하며, 향후 일상철학파라는 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).
- 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 기계학습 분류 방법이 함께 활용된다. 기계학습을 활용한 문서분류 방법은 본고에 기술되어있다.

은 언어학, 전산학, 통계학 등이 모두 상당한 수준으로 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.

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)⁸⁾⁹⁾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 분

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 연구 분야를 지칭하는 넓은 의미로 사용된다.(Hearst, 1999)

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

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

석결과를 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 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 가정하더라도 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)는 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 분석을 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.¹⁰⁾ 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 불용어¹¹⁾(stop words)라고 하며 보통 프로그램 라이브러리나

10) 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.

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

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

여기에서 N 은 총단어의 수, k 는 특정 단어의 순위, s 는 분포의 특징을 나타내는 지수값이다. 따라서 빈번하게 등장하는 단어일수록 내재된 정보의 가치는 희박하다고 볼 수 있다.

패키지에 포함되어 전처리 과정에서 제거할 수 있도록 되어 있다.

1.2 어간추출(stemming)과 표제어추출(lemmatization)

어간추출은 단어의 접사 등을 제거하고 어간을 분리¹²⁾해 내는 작업으로 관련 단어들이 일정하게 어간으로 매핑(mapping)하는 과정이다. 어간추출은 형태론과 정보 검색 등 분야에서 시작되었으며 1960년대부터 컴퓨터를 활용한 어간추출 알고리즘¹³⁾으로 연구되었다. 영문의 경우 동사와 형용사에 접두사(prefix)나 접미사(suffix)가 붙어 수많은 단어가 파생될 수 있다.¹⁴⁾ 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.¹⁵⁾ 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.¹⁶⁾ 따라서 어간추출에 비해 과정이 좀 더 복잡하며, 표제어추출을 위해서는 다음에 서술하는 품사 태깅 과정이 선행되어야 한다.

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

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

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

15) 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/lemmatization.html을 참조

16) 다양한 언어에 대해 어간추출 기법과 표제어추출 기법의 성능을 비교 분석하는 연구가 많이 진행되어있다(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 것이며¹⁷⁾ 품사태깅은 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.¹⁸⁾ 아래는 형태소 분석기로 “한국은행이 12일 금융통화위원회(금통위) 회의를 열고 기준금리를 현행 연 1.50%로 동결했다.” 라는 문장을 형태고 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/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 세

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

18) 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.

종품사태그에 나온 예시와 같이 정의되어 있는데, 형태소 분석기마다 조금 다르게 정의하는 경우도 있으나 세종품사태그와 유사하게 정의하고 있으므로 큰 차이는 없다.

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.¹⁹⁾

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.²⁰⁾

2. 단어표상(Word Representation)

전체리를 통해 필요한 텍스트 데이터만 수집한 후 얻게 되는 단어집합(corpus)

19) 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, 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.

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

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 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 행렬²¹⁾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.²²⁾

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

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

〈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 벡터이다. 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

	문 서 1	문 서 2	문 서 3	문 서 4	문 서 5	문 서 6	문 서 7	문 서 8
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 단어²³⁾ 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.

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

이를 IDF(Inverse Term Frequency)라 한다. 예를 들어 경제와 관련된 문헌들이라면 ‘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 N 건의 문서 중 어떠한 문서(D)에서 특정 단어(t)가 f 번 출현하였고 N 건의 문서 중 t 가 출현한 문서의 수가 n 이라 하면 TF-IDF는 아래와 같이 산출한다.

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

여기에서 T 는 표준화 인자(normalizing factor)로 한 단어가 습관적으로 자주 등장하는 경우 이를 표준화하는 역할을 한다. $\log(N/n)$ 이 가중치이며 t 가 출현한 문서의 수가 적을수록 가중치가 증대됨을 알 수 있다. 위 빈도 행렬의 예에서 T ‘가 문서의 총 단어수라고 하면, 직원’은 D_1 과 D_2 빈도가 각각 2, 1로 높은 편이라 할 수 있다. 그러나 D_1 에서 ‘직원’의 TF-IDF는 아래와 같다.

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

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.

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

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 m 의 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 m 크기의 윈도우에서 동시에 출현하는 빈도를 나타낸다. 동반출현행렬은 그 자체로는 유용한 정보를 얻거나 분석하기에 용이하지도 않을뿐더러 행렬의 차원이 매우 높는데다 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

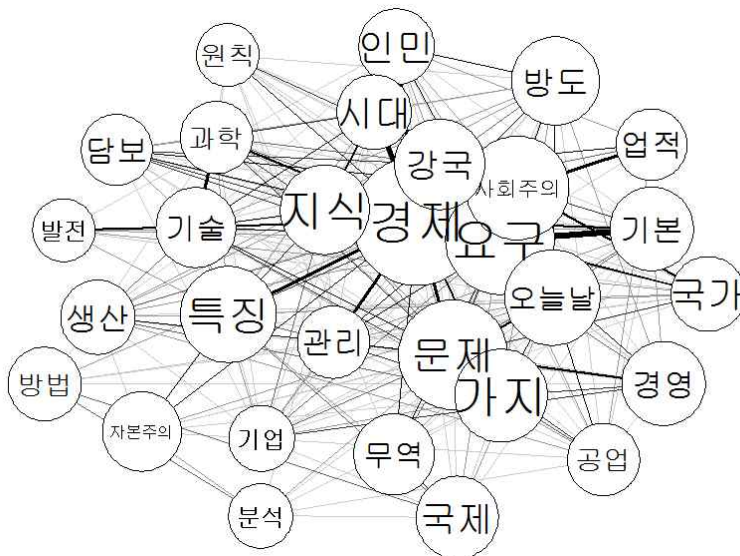
문서의 전처리 과정을 거쳐 단어의 품사별로 빈도를 계산하게 되면, 이를 활용하여 다양한 시각화를 할 수 있다. 이러한 시각화 과정은 문서의 주제, 특징 등을 보다 직관적으로 표현할 수 있어 보편적으로 사용되는 방법이다. 빈도를 기준으로 그래프를 그릴 수도 있겠으나, 워드 클라우드(word clouds)는 빈도에 비례하여 단어의 크기를 나타내어 말뭉치를 시각화하는 방법으로 문서의 특징을 보다 직관적으로 나타내준다. <그림 2>는 2005~2017년 중 한국은행 금융통화위원회 의사록에 나타난 단어중 명사만을 추출하여 그린 워드 클라우드이다.

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

<그림 2> 금통위 의사록 워드 클라우드(Lee, Kim, and Park, 2019a)



<그림 3> 단어들의 동시 출현 관계 다이어그램 예시



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. ____껌이 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 공간²⁴⁾에 위치한 벡터로 변환한다. 초기 신경망 언어모형²⁵⁾은 연산량이 매우 많은 모형이었다. NNLM(neural net language model)²⁶⁾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.

단어의 의미를 힐베르트 공간의 정규화된 벡터로 표현할 경우 내적 또는 cosine 값으로 단어간 유사도²⁷⁾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 , W 라 하자. 확률적 단어표상에서는 <그림 4>와 같은 $K - M + W = Q$ 관계가 성립한다. 아래에서는 확률적 단어표상의 대표적인 word2vec의 Skip-gram과 CBOW에 대해 알아본다.

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

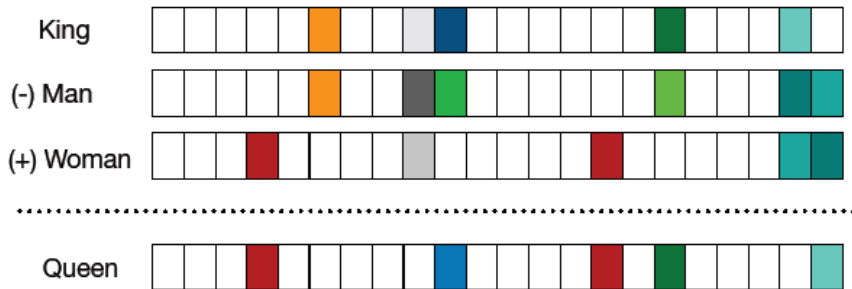
25) 신경망 언어모형은 Rumelhart et al. (1986)에서 시작되었다 볼 수 있다. Rumelhart et al. (1986)에서는 input layer와 output layer만 존재하는 간단한 신경망이다. 오늘날 사용되는 신경망은 input과 output 사이에 다층의 hidden layer가 존재하며 이들 신경망을 추정하는 과정을 학습(learning)이라 한다.

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

27) cosine 유사도는 벡터간 방향의 일치성을 측정함으로써 단어의 의미의 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 정규화 되어있을 경우 cosine 유사도는 두 벡터의 내적으로 표현된다.

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

〈그림 4〉 확률적 단어표상의 예시(Young et al, 2017)



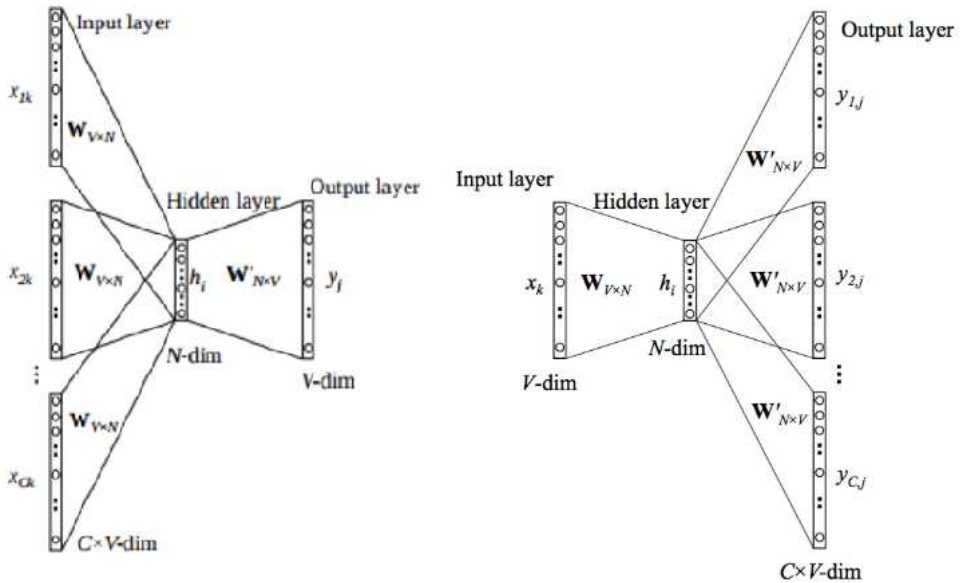
2.2.1 Skip-gram과 CBOW

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 크기만큼 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)²⁸⁾ 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 구성되어있다. 학습된 결과 얻게 되는 단어벡터는 코퍼스 내에 존재하는 V 개의 모든 단어들과의 관계 속에서 해당 단어의 특징 및 의미를

28) 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.

〈그림 5〉 CBOW와 Skip-gram (Mikolov et al, 2013)



내포하는 N 차원의 벡터이다. (〈그림 5〉 참조) 학습은 각각의 입력단어에 대해 아래의 softmax 함수²⁹⁾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 o 는 출력단어, c 는 입력단어(context vector), u_o 와 v_c 는 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.

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

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

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 농담³⁰⁾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 유사하게 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.

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

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

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 학습(추정)한 뒤 검정자료를 통해 분류(예측)가 올바르게 이루어졌는지 여부를 판단할 수 있다. 아래에서는 문서 분류에 활용될 수 있는 지도학습과 비지도학습 기법 대해 알아본다.

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 분류³¹⁾하는데 사용될 수 있다. 로지스틱 회귀는 일반 선형회귀 분석과 유사하지만, 종속변수가 $[0, 1]$ 의 범위 내에 있다는 점이 다르다. 따라서 값의 범위가 $[-\infty, \infty]$ 인 독립변수의 변화에 대응시키기 위해 로지스틱 회귀분석은 아래와 같이 로짓 변화(l)된 종속변수와 독립변수 벡터(X), 회귀계수 벡터(β)의 모형으로 표현한다.

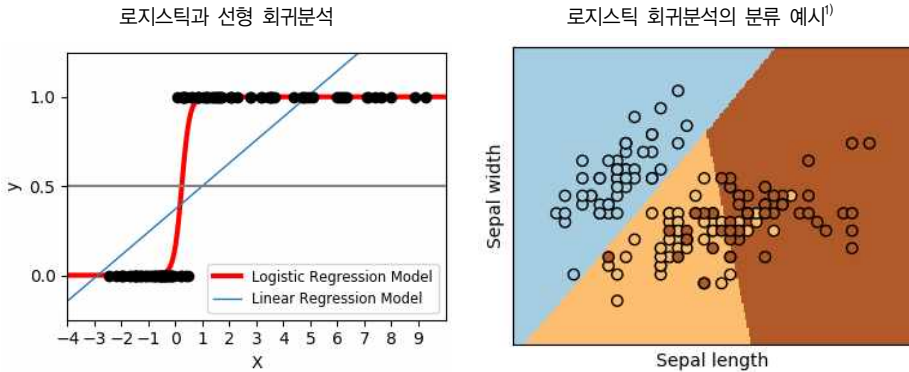
$$l(p) = \log \frac{p(y=1|x)}{1-p(y=1|x)} = \beta^T X$$

여기에서 $p(y=1|x) = 1/(1+e^{-\beta^T X})$ 로 표현할 수 있으며 이를 로지스틱 함수³²⁾(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.

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

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

〈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.

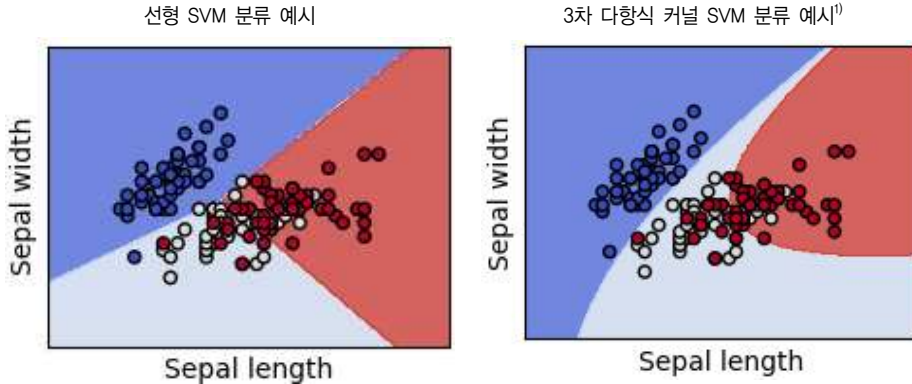
자료: scikit-learn.org

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 같음하고 그것을 통해 초평면을 찾는 커널 트릭(kernel trick)을 사용한다.³³⁾(Grus, 2015)

33) 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.

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

만약 b 값을 적정히 구하게 된다면 초평면의 a 와 $-a$ 와의 거리가 동일하게 되고 y_1 는 $\{-1, 1\}$ 의 값을 갖게 되므로, 위 초평면은 데이터를 두 범주로 구분함과 동시에 초평면이 지나는 벡터³⁴⁾간 거리가 극대화된다. 만약 소프트 마진을 활용하여 초평면을 찾는다면 아래와 같은 초평면이 구해질 수 있다.

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

34) 이를 서포트 벡터라 한다

3.3 나이브 베이즈 분류모형(Naive Bayes Classifier)

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

$$p(y_i | x_{i1} \dots x_{in}) = \frac{p(y_i) p(x_{i1} \dots x_{in} | y)}{p(x_{i1} \dots x_{in})}$$

만약 모든 특성이 나타날 조건부 확률이 독립³⁵⁾이라면 위 식은 아래와 같이 단순화하여 표현할 수 있다.

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

여기에서 $p(x_{i1} \dots x_{in})$ 는 상수이므로

$$p(y_i | x_{i1} \dots x_{in}) \propto p(y_i) \prod_{j=1}^n p(x_{ij} | y)$$

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, 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 기반

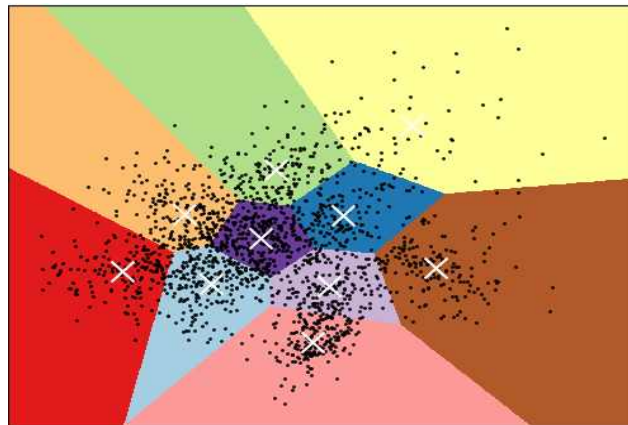
35) 범주에 속할 사전확률(prior)이 동일하거나, 각 단어들이 등장할 조건부 확률이 독립적인 전제 하에 분류하므로 “순진함(Naive)”이라는 수식어가 붙는다고 한다.

하여 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 d 건의 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.

$$c = \operatorname{argmin} \sum_{j=1}^k \sum_{x_i \in c_j} \|x_i - \mu_i\|$$

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 k 를 사전 지정해야하며, 극단치(outlier)에 민감하다는 단점이 있다.

〈그림 8〉 k-평균 군집화 분류



주: 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 분류해내고 있다.

자료: scikit-learn.org

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). 나이트 베이스 분류모형은 베이스 법칙과 단순 가정으로 문서를 분류할 수 있는 지도학습이라면 비모수 베이지안 군집화는 계층적 베이지안 군집화 모형을 사용하는 비지도 학습 기법이다. 앞서 설명한 분류기법이 문서 분포에 대한 가정 없이 사전 지정된 범주의 수에 따라 거리 등의 측도(metric)에 기반하여 분류하는 기법이라면, 비모수 베이지안 군집화에서는 문서가 임의 분포에 따른다는 가정 하에 통계적으로 추론하여 최적 범주의 수를 유동적으로 추정할 수 있다는 점이 다르다.

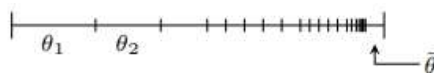
비모수 베이지안 군집화에서는 계산의 단순화 등의 이유로 k 차 심플렉스 위에 정의되는 이산분포인 디리클레 확률과정(Dirichlet process, DP)이 가장 널리 사용된다. i 번째 문서(y_i)가 따르는 임의의 연속 분포 $f(y_i|\theta_i)$ 의 모수 θ_i 는 분포 $G(\theta_i)$ 에서 추출될 수 있으며 모수의 분포 $G(\theta_i)$ 는 디리클레 확률과정을 따르는 계층적 모형을 가정한다.³⁶⁾

$$y_i|\theta_i \sim F(\theta_i), \theta_i \sim G(\theta), G(\theta) \sim DP(\alpha, G_0)$$

이를 디리클레 확률과정 혼합모형(Dirichlet process mixture model; DPM)이라고도 부르며 문서 y_i 가 따르는 분포 $f(y_i|\theta_i)$ 모수 θ_i 의 확률적 동일성 여부로 군집화가 이루어진다. 따라서 서로 다른 문서 y_i 와 y_j 가 동일한 모수에 기반한 분포 $f(y|\theta)$ 에서 추출될 수 있다고 추론될 경우 하나의 범주로 묶이게 된다. 한편 범주의 수가 정해져 있지 않으므로 특성이 매우 다른 문서 y_k 가 존재할 경우 새로운 범주를 대표하는 분포 $f(y_k|\theta_k)$ 가 생성되어 군집화 과정에 유연성을 높일 수 있다.³⁷⁾

36) 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

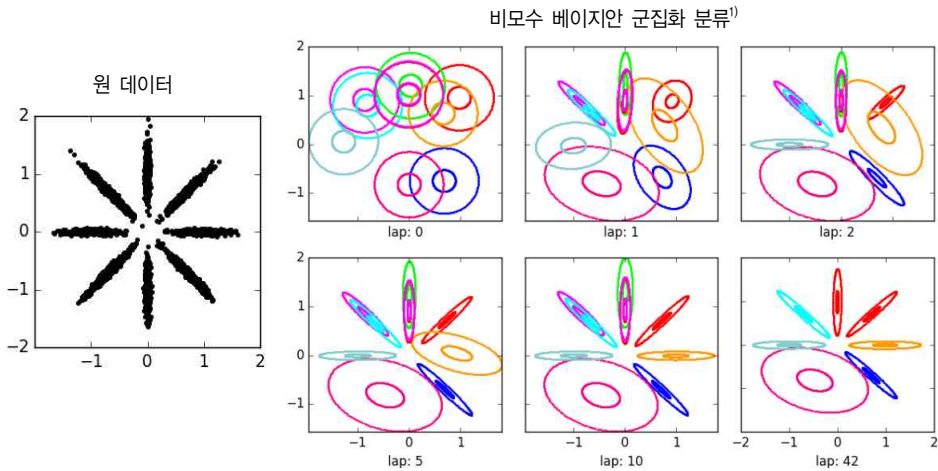
37) 이러한 추정과정을 막대 쪼개기 과정(stick-breaking process), 중국음식점 과정(CRP, Chinese restaurant process) 등으로 부른다. 막대 쪼개기 과정은 범주 구분시 데이터를 아래와 같이 하나의 긴 막대기를 쪼개듯 모수 θ_i 의 범주에 할당하는 방식이다.



중국음식점 과정은 데이터를 범주에 할당하는 방식을 중국음식점에서 손님들이 테이블에 앉는 과정으로 표현한 것이다. 두 방식 모두 새로운 범주가 생성되는데 패넬티가 부여되므로 과적합(overfitting)문

비모수 베이지안 군집화에서는 이산형 분포인 디리클레 분포를 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.

〈그림 9〉 비모수 베이지안 군집화 분류



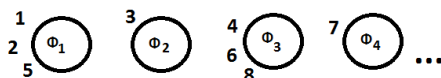
주: 1) Astericks(*) 형태로 분포하도록 임의의 생성된 자료를 비모수 베이지안 군집화 기법으로 사전분포(prior distribution)에서 시작하여 점차 사후분포(posterior distribution)로 추론해가는 과정으로 서로 다른 분포에 속한 점들은 다른 범주로 분류됨

자료: bnpy documentation

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_2 가 주어졌을 때 w_1 , w_2 가 한 문서내 동시 출현할 확률을 $p(w_1, w_2)$, w_1 , w_2 가 출현하는 확률을 각각 $p(w_1)$, $p(w_2)$ 라고 하면 PMI는 아래와 같이 측정한다.

제를 최소화하기 위한 군집화 과정이다.



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

만약 두 단어 w_1 , w_2 가 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_2 로 하여 임의의 w_1 과의 PMI를 측정하고 이를 아래와 같이 비교하여 w_1 의 극성을 분류한다.

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

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 주제를 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 잠재의미분석³⁸⁾(LSA)을 제안하였다. 이는 빈도행렬 등을 구성하는 선형공간에서 본 행렬의 특징을 가장 잘 나타내는 부분공간(linear subspace)을 식별하는 과정이다(Blei et al., 2003). 잠재의

38) 원문에는 잠재의미지수(latent semantic indexing)로 표기되어있으나, 오늘날에는 잠재의미분석으로 더 널리 알려져 있다.

미분석은 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 r 인 $n \times d$ 빈도행렬인 A 의 특잇값($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$)은 AA^T 또는 A^TA 의 고윳값의 제곱근으로 구할 수 있다. 특잇값을 대각원소로 하는 정방행렬을 $D_{r \times r} = \text{diag}(\sigma_1, \dots, \sigma_r)$ 과 AA^T 와 A^TA 를 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 이용³⁹⁾하면 A 는 아래와 같이 특잇값 분해⁴⁰⁾가 가능하다.

$$A = UDV^T$$

잠재의미분석은 아래와 같이 빈도행렬의 특징을 가장 잘 나타내는 k 개의 특잇값만 선택하여 A 를 근사함을 의미한다.

$$A_k = U_k D_k V_k^T$$

잠재의미분석의 수학적 의미는 A 의 열벡터가 U_k 의 열벡터를 기저로한 벡터공간으로 사영(projection)된 것이다.⁴¹⁾ 따라서 위 잠재의미분석에서 $V_k D_k$ 의 행이 각각 문서의 특징을 나타낸다고 볼 수 있다. 잠재의미분석의 수학적 의미는 Eckart and Young(1936)의 정리에 따르면 빈도행렬 A 를 가장 잘 근사하는 행렬 A_k 를 구하는 것이다. 잠재의미분석은 비확률적 과정으로 단어와 문맥 간의 잠재적인 의미

39) 여기에서 U 와 V 의 열벡터들을 각각 left singular vector, right singular vector라 한다.

40) 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.

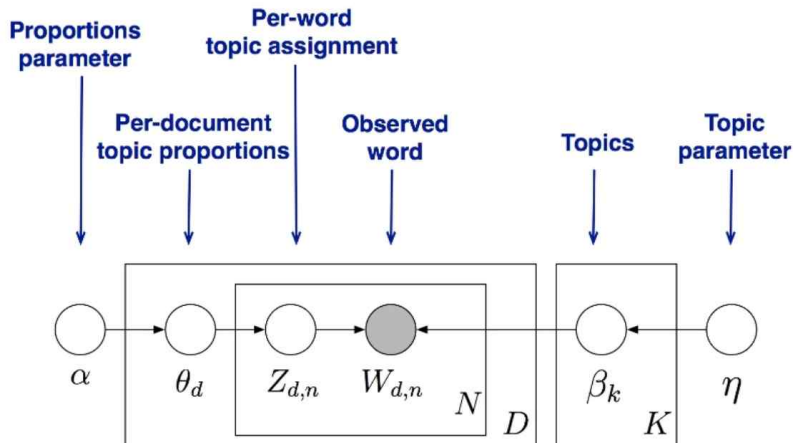
41) 특잇값분해는 원행렬의 성질을 가장 잘 반영하는 저차원의 행렬을 구하는 과정이다. 일상적으로 컴퓨터의 이미지 압축(compression) 기술로도 잘 활용되고 있는데, 특잇값 분해를 통해 이미지를 나타내는 행렬 A 에서 A_k 를 구함으로써 그림의 해상도를 낮출 수 있다.

(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 Allocation)

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 주제를 분류하는 확률모형이다.

〈그림 10〉 LDA 추론(Blei et al., 2003)



LDA는 D 를 말문치 전체 문서의 수, K 는 전체 토픽의 수, N 은 문서 d 의 단어 수라 할 때 우리가 문서 d 에 등장한 n 번째 단어 $w_{d,n}$ 를 관찰하고 〈그림 5〉의 과정으로 토픽의 분포를 추정하는 과정이다. 여기에서 α, β 는 연구자가 임의로 지정한

거나 중층적 비모수 베이지안(hierarchical non-parametric Bayesian) 모형으로 추정할 hyper parameter이다. <그림 5>의 추정과정은 디리클레 분포⁴²⁾에서 임의추출된 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.

〈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

4. 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 특

42) 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

정 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.

4.1 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 쓰이고 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. 김진영, 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. 감성사전을 구축하는 목적은 이를 활용하여 문서의 논조를 분석하고자 함이다. 목적에 따라 측정하고자 하는 논조와 그에 맞는 감성사전을 필요로 함은 당연할 과정일 것이다. 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. Lee, Kim, and Park(2019a)는 이러한 점을 보완하고자 경제·금융 분야에서 활용될 수 있는 감성사전을 구축하였다. 약 20만개 이상의 문서에서 읽은 n-gram과 나이트 베이스 분류모형⁴³⁾을 바탕으로 각각 시장적 접근방법과 사전적 접근방법을 활용하여 감성사전을 구축하였다.⁴⁴⁾

4.2 논조분석(tone analysis)

수작업 또는 기계를 이용하여 감성사전을 구축한 후 이를 활용하여 새로운 문서 샘플의 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 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 CFNAI⁴⁵⁾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.

IV. 텍스트 마이닝을 활용한 경제 분석

여기에서는 경제학적 분석에 텍스트 마이닝을 활용한 사례를 소개하고자 한다. 기존 경제학적 모형이 충격이 발생한 경우 이에 대한 분석이 목적이라면 텍스트

43) 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($p(\text{hawkish})$ 또는 $p(\text{dovish})$)이 동일하다는 가정 하에 단어가 나타날 조건부 확률의 비율로 단어 또는 n-gram의 극성을 측정하는 것이다(Lee, Kim, and Park, 2019a).

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

44) 시장적 접근방법이란 시장에서 관찰되는 특정 이벤트의 시간대에 나타난 기사 혹은 보고서의 감성에 이벤트의 성격을 부여하고 나이브 베이즈 분류모형으로 n-gram의 감성을 분석한다. 사전적 접근방법은 보편적으로 인식될 수 있는 감성을 지닌 seed 단어와 SO-PMI를 통해 감성을 분류하는 방법이다(Lee, Kim, and Park, 2019a).

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

마이닝을 활용한 일부 분석은 충격 자체를 예측하고자 하는데 목적이 있다. 거시경제 관련 분석에서 가장 큰 난점은 충분한 데이터 확보가 어렵다는 점이다. 일례로 한국의 대규모 자본유출 예측 모형을 구축하고자 한다면 과거 대규모 자본유출이 발생한 기간이 1997년 또는 2008년으로 한정되어 예측모형 구축을 위한 충분한 표본을 구하기 어렵다. 따라서 대부분의 경제학적 연구들은 충격 발생에 따른 영향과 파급효과 분석에 방점을 두고 있다. 그러나 예측모형은 실시간으로 매우 높은 정확도로 예측이 가능해야 그 의미가 있다. 미국 연방정부가 구글의 ‘백신’ 또는 ‘독감’ 검색어 추이로 해당 지역의 독감발생 여부 및 백신수요 등을 예측하고 백신을 해당지역으로 수송하는데 활용하고 있는 것과 같이, 경제학에서도 텍스트를 활용하여 거시경제변수의 예측력을 제고할 수 있다.

한편 기존 경제학적 모형으로 분석이 어려웠던 주제들을 텍스트 마이닝으로 다룬 연구들도 다수 이루어지고 있다. 특히 중앙은행 등 거시경제에 큰 영향을 미칠 수 있는 기관들의 커뮤니케이션은 전통적으로 시장참여자들의 큰 관심사가 되었으며, 해당 기관 스스로도 매우 세심히 검토하는 주제라 할 수 있다. 그러나 경제학 연구에 텍스트 분석이 도입되기 전에는 중앙은행 커뮤니케이션에 대한 분석은 대체변수(proxy)를 활용할 수밖에 없는 한계가 있었다(Romer and Romer, 2004; Gurkaynak et al., 2005). 관련 연구에 텍스트 마이닝이 도입된 이후에는 중앙은행 투명성 제고의 결과, 의사록간 유사도 분석, 의사록 논조 변화가 내포하는 통화정책 관련 정보 등 다양한 주제로 연구가 진행되었다. 최근 발표되거나 진행중인 중앙은행 커뮤니케이션 관련 다수의 연구들이 텍스트 마이닝을 도입하고 있다는 사실은 관련 연구에서 텍스트 분석이 매우 유용함을 반증해주는 증거라 할 수 있다.

그럼에도 불구하고 정치학, 언론학, 심리학 등 주요 인문사회 분야에 비해 경제학에의 도입은 비교적 늦게 이루어졌으며, 최근에 들어서야 텍스트 데이터를 활용한 연구가 주요 저널에 게재되고 있다. Gentzkow and Shapiro (2010)는 독자의 정치적 보수 및 진보 성향에 따라 언론이 사용하는 언어가 공화당 또는 민주당 의원들이 사용하는 어휘 및 논조와 유사한 기사를 생산하여 이윤 최적화를 추구한다고 밝히고 있다. Stock and Trebbi (2003)는 도구변수(instrument variable)를 처음 소개한 저자를 구별해내는데 텍스트 마이닝을 활용하였다. 도구변수가 처음 소개된 Philip Wright의 저서 *The Tariff on Animal and Vegetable Oils*의 부록의 저자 Philip Wright인지 혹은 그 아들인 Sewall Wright인지를 논의가 분분하였으나, Stock and Trebbi (2003)는 부록이 본문의 저자 Philip Wright에 의해 집필되었음을 밝히고 있다.

글로벌 금융위기 이후 초저금리 제약(zero lower bound) 하에서 중앙은행의 커뮤니케이션은 양적완화 등 비정상적 통화정책과 함께 중요한 통화정책 수단으로 대두되었다. 사전적 정책방향 제시(forward guidance)는 중앙은행의 중단기 통화정책 목표와 운용방안을 시장에 선제적으로 제시하고 약속(commitment)함으로써 통화정책에 대한 기대를 형성하는 정책적 수단의 일환이다. 사전적 정책방향 제시가 시장의 기대에 효율적으로 영향을 미침으로써 장기금리뿐만 아니라 실물경기 변동에도 영향을 미칠 수 있는 것으로 알려져 있다. 그러나 이러한 커뮤니케이션은 금리나 채권매입량과 달리 수치화된 자료가 아니므로 정량적 분석이 불가능하다. 따라서 텍스트 자료에서 수치화된 지표로 추출하는 분석방법이 필요하다. 이는 앞서 기술한 감성분석 또는 논조분석을 통해 문서의 논조를 정량화하면 여타 거시변수와의 관계 및 설명력 등을 분석해볼 수 있다.

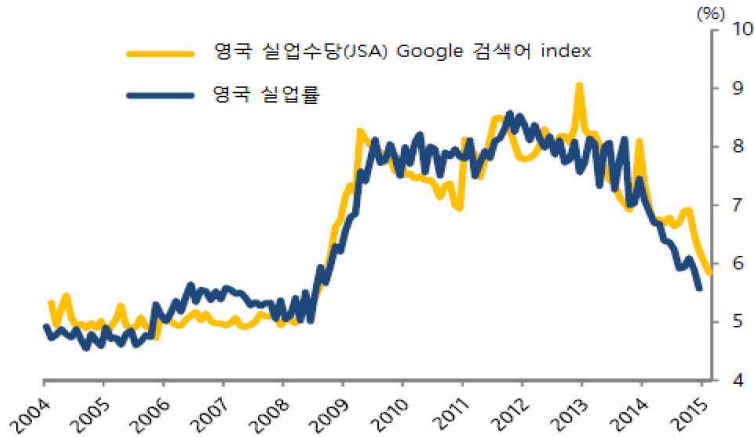
의사록 등 중앙은행의 커뮤니케이션 논조를 분석하는 것은 많은 의미를 내포한다. 우선 중앙은행은 금리결정에 앞서 가능한 공식, 비공식 자료를 모두 검토할 것이므로 가장 우월한 정보집합을 보유하고 있다고 볼 수 있다. 따라서 중앙은행 커뮤니케이션에서 감지된 논조의 변화는 이러한 정보가 종합된 결과로 해석 가능하다. 따라서 중앙은행 커뮤니케이션의 분석은 관측불가능한 종합거시경제지표를 논조라는 잠재변수로 추정하는 과정이다. 이러한 이유로 유럽중앙은행, 일본은행, 캐나다중앙은행 등 주요 중앙은행에서도 커뮤니케이션이 시장에 미치는 영향, 통화정책에 관한 정보 제공의 유효성 등을 텍스트 마이닝 기법으로 분석하고 있다.

아래에서는 텍스트 자료가 지닌 정보를 추출하고 활용한 경제학 분야 연구 사례를 사용된 방법론과 함께 소개하고자 한다. 텍스트 마이닝 방법론이 매우 다양하고 가용한 텍스트 데이터 종류도 다양하므로 본고의 방법론에서 기술하지 못한 분석기법이 사용된 문헌의 경우 방법론 관련 설명을 주석에서 상세히 기술하였다.

1. 검색어 및 소셜 미디어를 활용

구글(Google) 등 검색엔진이 보유한 검색어 데이터베이스를 활용하면 경기상황 및 불확실성 관련 정보를 추출하고 경제현황을 실시간 평가(nowcasting)할 수 있다. 노동시장과 주택시장을 설명하는 모형의 설명변수로 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 11〉 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 검색⁴⁶⁾(Boolean methods)으로 측정하여 정책불확실성지수⁴⁷⁾(EPU: Economic Policy Uncertainty)를 편제하였다. 경제정책 불확실성을 나타내는 정책불확실성지수(EPU)가 상승할 경우 주가변동성이 확대되고 투자와 고용이 위축되는 것으로 나타났다.

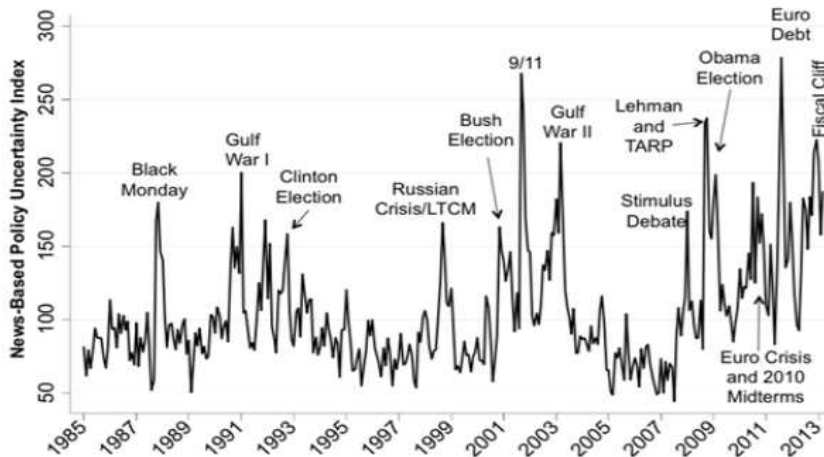
트위터(Twitter) 등 소셜 미디어나 로이터(Reuters) 뉴스 기사 데이터베이스에서 특정 이벤트 또는 심리를 나타내는 단어로 검색하였을 때, 그 결과로 추출되는 게시물 또는 기사의 수를 계산하여 시장의 심리를 측정하기도 한다. 영란은행은 과

46) 불린 검색이란 특정 검색어 포함여부로써 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. 구글 검색에서 검색어 해당 건수를 반환하는 연산도 이러한 불린 검색방법에 속한다.

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

거 스코틀랜드의 분리독립 투표 당시 스코틀랜드 소재 은행의뱅크런 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).

〈그림 12〉 정책불확실성지수(Baker Bloom and Davis, 2013)



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.⁴⁸⁾ 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 동반출현확률⁴⁹⁾(PMI: Point-wise Mutual Information)로 단어의 극성⁵⁰⁾을 측정하였다. 기사의 단어의 극성을 가중평균하여 시기별 심리지수를 편제한 결과 동지수는 주가, 금리, 환율 등에도 유의한 예측력을 지닌다는 결과를 확인하였다.

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 국채 금리에 영향을 미치며, 이로써 FOMC 의결문이 기존 테일러준칙에 사용되는 거시경제변수에 담긴 정보 외에 추가적인 정보를 내포하는 증거가 된다. Lucca and Trebbi(2011)는 핵심어 동반출현확률⁵¹⁾(SO-PMI) 기법으로 FOMC 의결문에 사

48) Bholat et al. (2015)은 다음과 같은 이유로 텍스트 분석을 통한 불확실성 지수가 VIX에 비해 우월하다고 주장한다: ① 금융시장에 국한된 불확실성 보다 넓고 다양한 분야의 불확실성 측정이 가능 ② 자산 가격의 움직임으로 간접적으로 측정된 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)

49) 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.

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

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

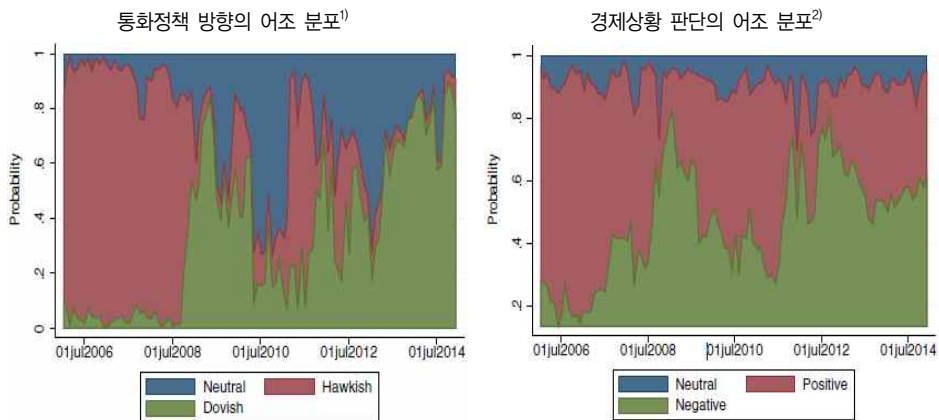
51) 미리 정의된 긍정적 핵심어(seed lexicons)와 부정핵심어를 각각 pw , nw 라 할 때 이들과의 동반출현확률로 단어 w 를 아래와 같이 계산할 수 있는데, 이 경우를 핵심어 동반출현확률(SO-PMI: Semantic Orientation from Point-wise Mutual Information)이라 한다. 핵심어 동반출현확률은 아래와 같이 계산할 수 있다.

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

용된 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 테일러준칙⁵²⁾ (augmented Taylor rule)의 설명변수로 하여 정책금리와 통화정책 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 시장접근법과 사전접근법 감성사전을 활용하였으며, 두 종류의 사전으로 측정한 의사록의 논조의 설명력과 예측력이 유사한 수준으로 밝혀졌다.

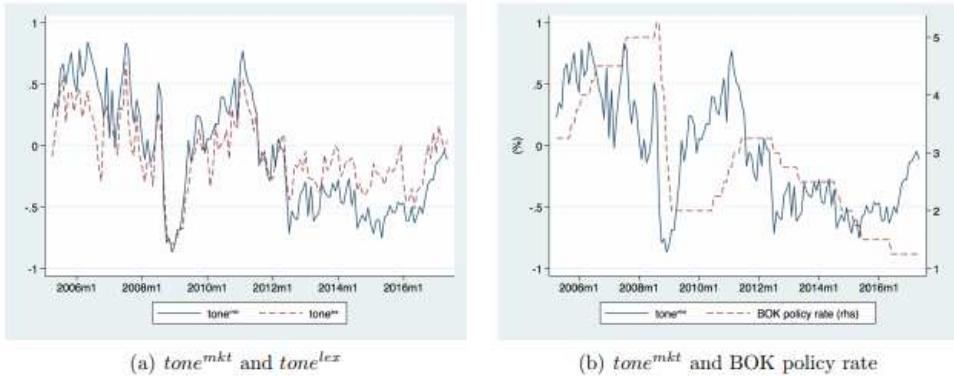
〈그림 13〉 ECB 통화정책 기자회견문 논조 분포(Picault and Renault, 2017)



- 주: 1) 매파적(hawkish, 적색), 중립적(neutral, 청색), 완화적(dovish, 녹색)으로 구분
2) 긍정적(positive, 적색), 중립적(neutral, 청색), 부정적(negative, 녹색)으로 구분

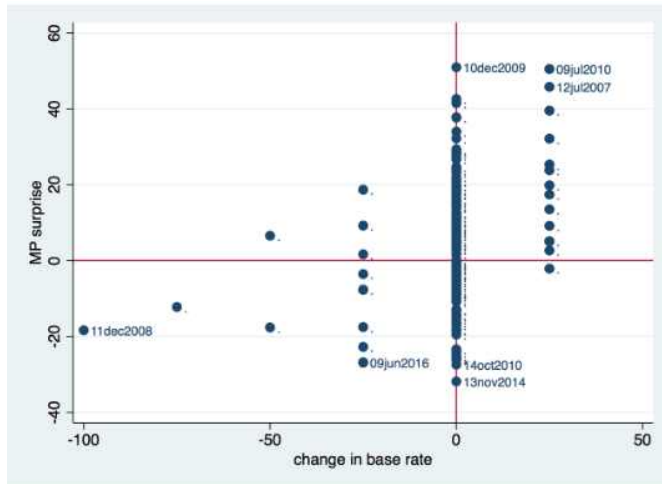
52) 인플레이션갭과 GDP갭이 포함된 기존 테일러 준칙에 ECB 기자회견문에서 추출한 긍정, 부정 논조 지수를 설명변수로 추가하였다.

〈그림 14〉 의사록 논조와 기준금리(Lee et al., 2019a)



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: Latent Dirichlet Allocation)을 활용하여 FOMC의 의결문을 경제현황(economic outlook)과 사전적 정책방향 제시로 구분하고 감성사전을 활용하여 어

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



조를 측정하였다. Hendry and Madeley (2010)는 통화정책방향 의결문에 포함된 지정학적 위험, 주요 대내외 충격요인, 사전적 정책방향 제시 관련 발언 등의 내용이 단기금리 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 회귀 분석을 실시하였다. 그 결과 통화정책회의 의사록에 내포된 낙관적 어조는 장기선 물금리에도 큰 영향을 미치고 불확실성을 감소시켜 금리 변동성을 줄이는 효과가 있음이 확인되었다.

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 단기금리에 영향을 미치는 기준금리 변경과 달리, 통화정책 서프라이즈는 1년 이상 중장기금리에 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.

3.2 거시건전성 정책 관련 커뮤니케이션 분석

중앙은행의 거시건전성 정책관련 커뮤니케이션도 금융시장 움직임에 유의한 영향을 미치며, 시중은행에서 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.

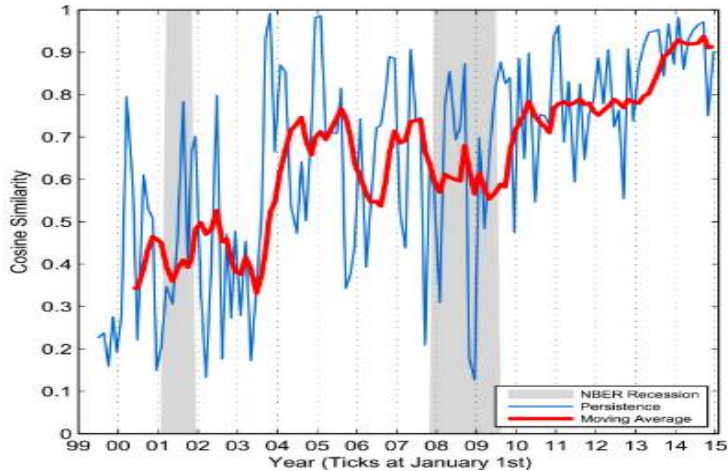
한편 영란은행 건전성 감독기구(PRA: Prudential Regulation Authority)가 은행 및 관련 회사들에 발송한 비밀 서한에는 전신인 금융감독청(FSA: Financial Services Authority)이었을 당시 서한과는 달리 미래지향적, 지도적 어휘의 사용이

대폭 증가하였다는 연구결과도 있다. Bholat et al. (2017)은 기계학습을 활용하여 영란은행 건전성 감독기구의 비밀 서한을 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

중앙은행의 의사록을 분석한 결과 통화정책의 투명성 제고는 통화정책회의의 참여도를 높이는 등 긍정적 효과가 있는 반면 적극적 소수의견의 감소, 의결문의 유사도 상승 등 부작용도 초래 하는 것으로 연구되었다. 의사록이 공개됨에 따라 통화정책 의사결정자들에게는 적극적인 토론으로 회의에 참여할 유인이 증대되고 이는 의사결정구조에 긍정적 효과를 가져왔으나, 본인의 발언에 대한 리스크도 동시에 부담하게 되므로 소수의견을 피력하는 경우가 점차 감소하는 부작용을 초래한 것이다. 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 발언 시간 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 2002년은 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.⁵³⁾

〈그림 18〉 FOMC 의결문의 전기대비 코사인 유사도⁵³⁾ 추이(Acosta and Meade, 2015)

주: 1) TF-IDF로 단어빈도를 정규화하여 코사인 유사성(cosine similarity) 측정

V. 결론

텍스트 데이터는 4차 산업과 빅데이터 분석 기술 발전 등에 힘입어 최근 새로운 정보의 원천으로 인식되고 있다. 앞서 논의한 바와 같이 텍스트 데이터를 정도 있게 분석하면 기존 통계자료에서 얻을 수 없었던 다양한 정보를 취득할 수 있다는 점에서 텍스트 데이터의 장점은 분명히 존재한다. 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 제공할 것이기 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.

그 외에도 텍스트 마이닝을 활용한 분석에는 많은 과제가 남아있다. 텍스트 마이닝과 같은 빅데이터 분석은 기계학습, 인공지능망 또는 딥러닝(deep learning)을

53) 126개 FOMC 의결문의 1,302개 단어의 빈도수를 원소로 하는 1302×126 행렬을 구성하고 벡터공간모형에서 두 의결문(a_i, b_j)의 코사인 유사성(cosine similarity)을 측정

더욱 비중있게 사용하는 추세이다.⁵⁴⁾ 인공지능망도 하나의 모형이라는 점에서 검증이 필요할 것이나, 검증할만한 통계적 방법이 부족하다는 비판이 늘 있어왔다. 따라서 텍스트 마이닝만의 문제는 아닐 것이나, 텍스트 데이터의 확률적 표상의 경우 학습결과와 강건성 등의 통계적 유의성 등을 검증할 대안 마련이 필요하다. 또한 기계학습으로 구축한 감성사전의 경우에도 학습된 사전의 정확성⁵⁵⁾, 일치성 등을 검증할 방법을 강구할 필요가 있다.

54) 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.

55) 일반적으로 머신러닝을 활용한 분석기법에서는 정확도(accuracy)는 혼동행렬(confusion matrix)을 활용한 척도인 정확도(accuracy), 정밀도(precision), 재현율(recall), 특이성(specificity), F1점수(F1 score) 등을 계산함으로써 측정한다. 아래와 같이 혼동행렬을 정의할 때 각 척도의 계산은 다음과 같다.

		Predicted class	
		P	N
Actual class	P	True positives (TP)	False negatives (FN)
	N	False positives (FP)	True negatives (TN)

$$\text{정확도} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{정밀도} = \frac{TP}{TP + FP}$$

$$\text{재현율} = \frac{TP}{TP + FN}$$

$$\text{특이성} = \frac{TN}{TN + FN}$$

$$\text{F1점수} = 2 \times \frac{\text{재현율} \times \text{정밀도}}{\text{재현율} + \text{정밀도}}$$

참고문헌

- 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 감성 분석,” 한국정보과학회 학술발표논문집, 2016, pp. 747-749.
- 김남원, 박진수, “Naive Bayes 방법론을 이용한 개인정보 분류,” 지능정보연구, 제18권 제1호, 2012, pp.91-107.
- 김문형, 장하연, 조유미, 신호필, “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 투자 의사 결정모형,” 지능정보연구, 제18권 제2호, 2012, pp. 143-156.
- 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지,” 제 26회 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 활용한 한의

- 학 원전 연구의 가능성 모색 - 「황제내경(黃帝內經)」에 대한 적용례를 중심으로,” 대한한의학회지 제31권 제4호, 2018, pp. 27-46.
- 손욱, 성병목, 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.
- Acosta, M. (2014), “FOMC Responses to Calls for Transparency,” Finance and Economics Discussion Series 2015-60, Board of Governors of the Federal Reserve System (U.S.).
- Acosta, M., and Meade, E. E. (2015), “Hanging on Every Word: Semantic Analysis of the FOMC's Postmeeting Statement,” FEDS Notes, Washington: Board of Governors of the Federal Reserve System.
- Almatarneh, S., and Gamallo, P. (2018), “Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification,” Conference Paper in Advances in Intelligent Systems and Computing.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014), *Bayesian Data Analysis*, Chapman and Hall/CRC, 3rd edition.
- Baker, S. R., Bloom, N., and Davis S. J. (2016), “Measuring Economic Policy Uncertainty,” *The Quarterly Journal of Economics*, Vol. 131(4), pp. 1593-1636
- Bang, H. C., and Ha, J. (2013), “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).

- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003), "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, Vol. 3, pp. 1137-1155.
- Bholat, D., Hansen, S., Santos, P., and Schonhardt-Bailey, C. (2015), Text Mining for Central Banks: Handbook, Centre for Central Banking Studies (33). pp.1-19, ISSN1756-7270.
- Bholat, D., Brookes, J., Cai, C., Grundy, K., and Lund, J. (2017), "Sending Firm Messages: Text Mining Letters from PRA Supervisors to Banks and Building Societies They Regulate," Staff Working Paper No. 688, Bank of England.
- Born, B., Ehrmann, M., and Fritzsche, M. (2014), "Central Bank Communication on Financial Stability," *Economic Journal*, Vol. 124(577), pp. 701-734.
- Bouma, G. (2009), "Normalized (Pointwise) Mutual Information in Collocation Extraction," Conference Proceedings.
- Cambria, E., and White, B. (2014), "Jumping NLP Curves: A Review of Natural Language Processing Research," *IEEE Computational Intelligence Magazine*, Vol. 9, No. 2, pp. 48-57.
- Chague, F., De-Losso R., Giovanetti, C. B., and Manoel, B. (2013), "Central Bank Communication Affects Long-Term Interest Rates," Working Papers, No. 2013-07, Department of Economics, University of Sao Paulo.
- Choi, H., and Varian, H. (2009), "Predicting Initial Claims for Unemployment Benefits," SSRN: <https://ssrn.com/abstract=1659307>.
- Choi, H., and Varian, H. (2012), "Predicting the Present with Google Trends," Special Issue: Selected Papers from the 40th Australian Conference of Economists, Vol. 88(1), pp. 2-9.
- Cox, D. R. (1958), "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 20(2), pp. 215-242.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990),

- “Indexing by Latent Semantic Analysis,” *Journal of the Association for Information Science and Technology*, Vol. 41(6), pp. 391-407.
- Eckart, C., and Young, G. (1936), “The Approximation of a Matrix by Another of Lower Rank,” *Psychometrika*, Vol. 1, pp. 211-218.
- Fausch, J., and Sigonius, M. (2018), “The Impact of ECB Monetary Policy Surprises on the German Stock Market,” *Journal of Macroeconomics*, Vol. 55, pp. 46-63.
- Fawley, B. W., and Neely, C. J. (2014), “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, pp. 73-109.
- Friedman, M., and Schwartz, A. J. (1963), *A Monetary History of the United States, 1867-1960*, Princeton University Press.
- Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1983), “Human Factors and Behavioral Science: Statistical Semantics: Analysis of the Potential Performance of Key-word Information System,” *The Bell System Technical Journal*, Vol. 62(2).
- Gentzkow, M., and Shapiro, J. M. (2010), “What Drives Media Slant? Evidence From U.S. Daily Newspapers,” *Econometrica*, Vol. 78(1), pp. 35-71.
- Gentzkow, M., Kelly, B., and Taddy, M. (2017), “Text as Data,” NBER Working Paper, No. 23276.
- Gertler, M., and Karadi, P. (2015), “Monetary Policy Surprises, Credit Costs, and Economic Activity,” *American Economic Journal: Macroeconomics*, Vol. 7, pp. 44-76.
- Giannone, D., Reichlin, L., and Sala, L. (2002), “Tracking Greenspan: Systematic and Unsystematic Monetary Policy Revisited,” CEPR Discussion Papers 3550.
- Grus, J. (2015), *Data Science from Scratch*, O’Reilly Media.
- Gurkaynak, R., Sack, B., and Swanson, E. (2005), “Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements,” *International Journal of Central Banking*, Vol. 1.

- Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky D. (2016), “Inducing DomainSpecific Sentiment Lexicons from Unlabeled Corpora,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Conference on Empirical Methods in Natural Language Processing*, pp. 595-605.
- Hansen, S., and McMahon, M. (2016), “Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communications,” *Journal of International Economics*, Vol. 99(1), pp. 114-133.
- Hansen, S., McMahon, M., and Prat, A. (2018), “Transparency and Deliberation within the FOMC: A Computational Linguistics Approach,” *The Quarterly Journal of Economics*, Vol. 133(2), pp. 801-870.
- Hansen, S., and McMahon, M. (2016), “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.
- Hendry, S., and Madeley, A. (2010), “Text Mining and the Information Content of Bank of Canada Communications,” *Staff Working Papers*, Bank of Canada.
- IBM (2015), “Big Data and Analytics,” <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html/>.
- Kawamura, K., Kobashi, Y., Shizume, M., and Ueda, K. (2019), “Strategic Central Bank Communication: Discourse Analysis of the Bank of Japan’s Monthly Report,” *Journal of Economic Dynamics and Control*, Vol. 10, pp. 230-250.
- Kim, Y. (2014), “Convolutional Neural Networks for Sentence Classification,” *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751.
- Kettunen, K., Kunttu, T., and Järvelin, K. (2005), “To Stem or Lemmatize a Highly Inflectional Language in a Probabilistic IR Environment?,” *Journal of Documentation*, Vol. 61(4), pp. 476-496.
- Kroeger, P. (2005), *Analyzing Grammar: An introduction*, Cambridge, Cambridge University Press.

- Kuttner, K. N. (2001), “Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market,” *Journal of Monetary Economics*, Vol. 47, pp. 523-544.
- Landauer, T. K., and Dumais, S. T. (1997), “A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge,” *Psychological Review*, Vol. 104(2), pp. 211-240.
- Lee, Y. (2018), “Introduction to eKoNLPy: A Korean NLP Python Library for Economic Analysis,” <https://github.com/entelecheia/eKoNLPy>.
- Lee, Y., Kim, S., and Park, K. Y. (2019a), “Deciphering Monetary Policy Committee Minutes with Text Mining Approach: A Case of Korea,” *Korean Economic Review*, forthcoming.
- _____ (2019b), “Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea,” BOK Working Paper, No. 2019-11.
- Liu, B. (2009), “Sentiment Analysis and Subjectivity,” in *Handbook of Natural Language Processing*, New York, NY, USA: Marcel Dekker, Inc.
- _____ (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.
- Loughran, T., and McDonald, B. (2011), “When is a Liability not a Liability? Textual analysis, dictionaries, and 10-Ks,” *The Journal of Finance*.
- Lucas, R. E. (1976), “Econometric Policy Evaluation: A Critique,” *Carnegie-Rochester Conference Series on Public Policy*, Vol. 1, pp. 19-46.
- Lucca, D. O., and Trebbi, F. (2011), “Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements,” NBER Working Paper Series.
- MacQueen, J. B. (1967), “Some Methods for Classification and Analysis of Multivariate Observations,” *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 281-297.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press.

- McLaren, N., and Shanbhogue, R. (2011), "Using Internet Search Data as Economic Indicators," Bank of England Quarterly Bulletin, No. 2011Q2.
- Meinusch, A., and Tillmann, P. (2017), "Quantitative Easing and Tapering Uncertainty: Evidence from Twitter," *International Journal of Central Banking*, Vol. 13(4), pp. 227-258
- Mikolov, T., K. Chen, G. Corrado, and J. Dean. (2013), "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781.
- Noh, H., Park, J., Sim, G., Yu, J., and Chung, Y. (2014), "Nonparametric Bayesian Models in Biomedical Research," *The Korean Journal of Applied Statistics*, Vol. 27(6), pp. 867-889.
- Nopp, C., and Hanbury, A. (2015), "Detecting Risks in the Banking System by Sentimental Analysis," Conference on Empirical Methods in Natural Language Processing.
- Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., and Smith, R. (2018), "News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment," Bank of England Staff Working Paper, No. 704.
- Oshima, Y., and Matsubayashi, Y. (2018), "Monetary Policy Communication of the Bank of Japan: Computational Text Analysis," Discussion Papers, No. 1816, Graduate School of Economics, Kobe University.
- Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. (2000), "Latent SEmantic INDEXing: A Probabilistic Analysis," *Journal of Computer and System Science*, Vol. 61(2), pp. 217-315.
- Pescatori, A. (2018), "Central Bank Communication and Monetary Policy Surprises in Chile," IMF Working Paper, No. 18156.
- Picault, M., and Renault, T. (2017), "Words Are Not All Created Equal: A New Measure of ECB Communication," *Journal of International Money and Finance*, No. 79, pp. 136-156.
- Porter, M. F. (1980), "An Algorithm for Suffix Stripping," *Program*, Vol. 14(3), pp.

130-137.

- Pyo, D., and Kim, J. (2017), “News Media Sentiment and Asset Prices: Text-mining Approach,” KIF Working Paper.
- Rapp, R. (2003), “Word Sense Discovery Based on Sense Descriptor Dissimilarity,” In Proc. of the Machine Translation Summit IX, New Orleans.
- Romer, C. D., and Romer, D. H. (2004), “A New Measure of Monetary Shocks: Derivation and Implications,” *American Economic Review*, Vol. 94(4), pp. 1055-1084.
- Rumelhart, D. E., Hintont, G. E., and Williams, R. J. (1986), “Learning Representations by Backpropagating Errors,” *Nature*, Vol. 323(6088), pp. 533-536.
- Salton, G., and McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.
- Shibamoto, M. (2016), “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.
- Sohn, W., Sung, B., and Kwon, H. (2005), “The Financial Markets’ Responses to Monetary Policy Announcements,” *Bank of Korea Economic Analysis* (in Korean).
- Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966), *The general inquirer: A computer approach to content analysis*, MIT Press.
- Tala, F. Z. (2003), “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia,” mimeo.
- Tetlock, P. C. (2007), “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *The Journal of finance*, Vol. 62, pp. 1139-1168.
- Turney, P. D., and Litterman, M. L. (2002), “Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus,” CoRR, vol.cs.LG/0212012.
- Turney, P. D., and Pantel, P. (2010), “From Frequency to Meaning: Vector Space Models of Semantics,” *Journal of Artificial Intelligence Research*, Vol.37, pp.141-188.

- Vapnik, V. N., and Chervonenkis, A. Y. (1964), “A Class of Algorithms for Pattern Recognition Learning”, *Avtomat. i Telemekh.*, Vol. 25(6), pp. 937-945.
- Vozalis, M., and Margaritis, K. G. (2003), “Analysis of Recommender Systems’ Algorithms,” In *Proc. of the 6th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA)*, Athens, Greece.
- Young, T., Hazarika, D., Poria, S., and Cambria, E. (2017), “Recent Trends in Deep Learning Based Natural Language Processing”, *arXiv preprint arXiv:1708.02709*.

부록

〈부표 1〉 세종품사태그

대분류	태그	설명	대분류	태그	설명
체언	NNG	일반 명사	어말 어미	EP	선어말 어미
	NNP	고유 명사		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		XSN	명사 파생 접미사
	VX	보조 용언	접미사	XSV	동사 파생 접미사
	VCP	긍정 지정사		XSA	형용사 파생 접미사
	VCN	부정 지정사	어근	XR	어근
관형사	MM	관형사	부호	SF	마침표물음표,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	불임표(물결,숨김,빠짐)
	JKC	보격 조사		SW	기타기호 (논리수학기호,화폐기호)
	JKG	관형격 조사		NF	명사추정범주
	JKO	목적격 조사	분석 불능	NV	용언추정범주
	JKB	부사격 조사		NA	분석불능범주
	JKV	호격 조사		SL	외국어
	JKQ	인용격 조사	한글 이외	SH	한자
	JX	보조사		SN	숫자
	JC	접속 조사			

Text Mining for Economic Analysis

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

We provide detailed description of how text data analysis is done and review series of studies done through text mining. Natural language can be characterized with ambiguity and obscurity compared to structured data. It is hard to retrieve useful information from text data as it carries natural language itself. Text mining or natural language processing is a multi-discipline area of modern technic in which we can distill and obtain just what we need from text. With the development of AI and machine learning, text mining is becoming one of the high-end technology in various fields of research even in economics. We expect there will be more demand for text data analysis as it will be complementary to traditional structured data and also as a new source of information.

Keywords: Text Mining, Machine Learning, Economic Analysis

JEL classification: A12, B41, C80

* Economist, Economic Research Institute, Bank of Korea, Tel: +82-2-759-5362, E-mail: soohyonkim@bok.or.kr

** Precourt Institute for Energy, Stanford University, E-mail: yj.lee@gmail.com

*** Professor, Yonsei Business School, Tel: +82-2-2123-5467, E-mail: jshin@yonsei.ac.kr

**** Economist, Yonsei School of Economics, Tel: +82-2-2123-5495, E-mail: kypark@yonsei.ac.kr

The views expressed in this paper are those of the author and do not necessarily reflect the official views of the Bank of Korea. When reporting or citing this paper, the authors' names should always be stated explicitly.

BOK 경제연구 발간목록

한국은행 경제연구원에서는 Working Paper인 『BOK 경제연구』를 수시로 발간하고 있습니다. 『BOK 경제연구』는 주요 경제 현상 및 정책 효과에 대한 직관적 설명 뿐 아니라 깊이 있는 이론 또는 실증 분석을 제공함으로써 엄밀한 논증에 초점을 두는 학술논문 형태의 연구이며 한국은행 직원 및 한국은행 연구용역사업의 연구 결과물이 수록되고 있습니다. 『BOK 경제연구』는 한국은행 경제연구원 홈페이지(<http://imer.bok.or.kr>)에서 다운로드하여 보실 수 있습니다.

제2016-1	The Spillover Effects of U.S. Monetary Policy on Emerging Market Economies: Breaks, Asymmetries and Fundamentals	Geun-Young Kim · Hail Park · Peter Tillmann
2	Pass-Through of Imported Input Prices to Domestic Producer Prices: Evidence from Sector-Level Data	JaeBin Ahn · Chang-Gui Park · Chanho Park
3	Spillovers from U.S. Unconventional Monetary Policy and Its Normalization to Emerging Markets: A Capital Flow Perspective	Sangwon Suh · Byung-Soo Koo
4	Stock Returns and Mutual Fund Flows in the Korean Financial Market: A System Approach	Jaebeom Kim · Jung-Min Kim
5	정책금리 변동이 성별·세대별 고용률에 미치는 영향	정성엽
6	From Firm-level Imports to Aggregate Productivity: Evidence from Korean Manufacturing Firms Data	JaeBin Ahn · Moon Jung Choi
7	자유무역협정(FTA)이 한국 기업의 기업내 무역에 미친 효과	전봉걸 · 김은숙 · 이주용
8	The Relation Between Monetary and Macprudential Policy	Jong Ku Kang
9	조세피난처 투자자가 투자 기업 및 주식 시장에 미치는 영향	정호성 · 김순호
10	주택실거래 자료를 이용한 주택부문 거시 건전성 정책 효과 분석	정호성 · 이지은
11	Does Intra-Regional Trade Matter in Regional Stock Markets?: New Evidence from Asia-Pacific Region	Sei-Wan Kim · Moon Jung Choi
12	Liability, Information, and Anti-fraud Investment in a Layered Retail Payment Structure	Kyoung-Soo Yoon · Jooyong Jun
13	Testing the Labor Market Dualism in Korea	Sungyup Chung · Sunyoung Jung

	14	북한 이중경제 사회계정행렬 추정을 통한 비공식부문 분석	최지영
제2016 –15		Divergent EME Responses to Global and Domestic Monetary Policy Shocks	Woon Gyu Choi · Byongju Lee · Taesu Kang · Geun-Young Kim
	16	Loan Rate Differences across Financial Sectors: A Mechanism Design Approach	Byoung-Ki Kim · Jun Gyu Min
	17	근로자의 고용형태가 임금 및 소득 분포에 미치는 영향	최충 · 정성엽
	18	Endogeneity of Inflation Target	Soyoung Kim · Geunhyung Yim
	19	Who Are the First Users of a Newly-Emerging International Currency? A Demand-Side Study of Chinese Renminbi Internationalization	Hyoung-kyu Chey · Geun-Young Kim · Dong Hyun Lee
	20	기업 취약성 지수 개발 및 기업 부실화에 대한 영향 분석	최영준
	21	US Interest Rate Policy Spillover and International Capital Flow: Evidence from Korea	Jieun Lee · Jung-Min Kim · Jong Kook Shin
제2017 –1		가계부채가 소비와 경제성장에 미치는 영향 – 유량효과와 저장효과 분석 –	강종구
	2	Which Monetary Shocks Matter in Small Open Economies? Evidence from SVARs	Jongrim Ha · Inhwan So
	3	FTA의 물가 안정화 효과 분석	곽노선 · 임호성
	4	The Effect of Labor Market Polarization on the College Students' Employment	Sungyup Chung
	5	국내 자영업의 폐업을 결정요인 분석	남윤미
	6	차주별 패널자료를 이용한 주택담보대출의 연체요인에 대한 연구	정호성
	7	국면전환 확산과정모형을 이용한 콜금리 행태 분석	최승문 · 김병국

제2017-8	Behavioral Aspects of Household Portfolio Choice: Effects of Loss Aversion on Life Insurance Uptake and Savings	In Do Hwang
9	신용공급 충격이 재화별 소비에 미치는 영향	김광환 · 최석기
10	유가가 손익분기인플레이션에 미치는 영향	김진용 · 김준철 · 임형준
11	인구구조변화가 인플레이션의 장기 추세에 미치는 영향	강환구
12	종합적 상환여건을 반영한 과다부채 가계의 리스크 요인 분석	이동진 · 한진현
13	Crowding out in a Dual Currency Regime? Digital versus Fiat Currency	KiHoon Hong · Kyoungheon Park · Jongmin Yu
14	Improving Forecast Accuracy of Financial Vulnerability: Partial Least Squares Factor Model Approach	Hyeongwoo Kim · Kyunghwan Ko
15	Which Type of Trust Matters?: Interpersonal vs. Institutional vs. Political Trust	In Do Hwang
16	기업특성에 따른 연령별 고용행태 분석	이상욱 · 권철우 · 남윤미
17	Equity Market Globalization and Portfolio Rebalancing	Kyungkeun Kim · Dongwon Lee
18	The Effect of Market Volatility on Liquidity and Stock Returns in the Korean Stock Market	Jieun Lee · KeeH.Chung
19	Using Cheap Talk to Polarize or Unify a Group of Decision Makers	Daeyoung Jeong
20	패스트트랙 기업회생절차가 법정관리 기업의 이자보상비율에 미친 영향	최영준
21	인구고령화가 경제성장에 미치는 영향	안병권 · 김기호 · 육승환
22	고령화에 대응한 인구대책: OECD사례를 중심으로	김진일 · 박경훈

제2017 -23	인구구조변화와 경상수지	김경근 · 김소영
24	통일과 고령화	최지영
25	인구고령화가 주택시장에 미치는 영향	오강현 · 김솔 · 윤재준 · 안상기 · 권동휘
26	고령화가 대외투자에 미치는 영향	임진수 · 김영래
27	인구고령화가 가계의 자산 및 부채에 미치는 영향	조세형 · 이용민 · 김정훈
28	인구고령화에 따른 우리나라 산업구조 변화	강종구
29	인구구조 변화와 재정	송호신 · 허준영
30	인구고령화가 노동수급에 미치는 영향	이철희 · 이지은
31	인구 고령화가 금융산업에 미치는 영향	윤경수 · 차재훈 · 박소희 · 강선영
32	금리와 은행 수익성 간의 관계 분석	한재준 · 소인환
33	Bank Globalization and Monetary Policy Transmission in Small Open Economies	Inhwan So
34	기존 경영자 관리인(DIP) 제도의 회생기업 경영성과에 대한 영향	최영준
35	Transmission of Monetary Policy in Times of High Household Debt	Youngju Kim · Hyunjoon Lim
제2018 -1	4차 산업혁명과 한국의 혁신역량: 특허자료를 이용한 국가기술별 비교 분석, 1976-2015	이지홍 · 임현경 · 정대영
2	What Drives the Stock Market Comovements between Korea and China, Japan and the US?	Jinsoo Lee · Bok-Keun Yu
3	Who Improves or Worsens Liquidity in the Korean Treasury Bond Market?	Jieun Lee

제2018 -4	Establishment Size and Wage Inequality: The Roles of Performance Pay and Rent Sharing	Sang-yeon Song
5	가계대출 부도요인 및 금융업권별 금융취약성: 자영업 차주를 중심으로	정호성
6	직업훈련이 청년취업을 제고에 미치는 영향	최충 · 김남주 · 최광성
7	재고투자와 경기변동에 대한 동학적 분석	서병선 · 장근호
8	Rare Disasters and Exchange Rates: An Empirical Investigation of Korean Exchange Rates under Tension between the Two Koreas	Cheolbeom Park · Suyeon Park
9	통화정책과 기업 설비투자 - 자산가격경로와 대차대조표경로 분석 -	박상준 · 육승환
10	Upgrading Product Quality: The Impact of Tariffs and Standards	Jihyun Eum
11	북한이탈주민의 신용행태에 관한 연구	정승호 · 민병기 · 김주원
12	Uncertainty Shocks and Asymmetric Dynamics in Korea: A Nonlinear Approach	Kevin Larcher · Jaebeom Kim · Youngju Kim
13	북한경제의 대외개방에 따른 경제적 후생 변화 분석	정혁 · 최창용 · 최지영
14	Central Bank Reputation and Inflation-Unemployment Performance: Empirical Evidence from an Executive Survey of 62 Countries	In Do Hwang
15	Reserve Accumulation and Bank Lending: Evidence from Korea	Youngjin Yun
16	The Banks' Swansong: Banking and the Financial Markets under Asymmetric Information	Jungu Yang

제2018-17	E-money: Legal Restrictions Theory and Monetary Policy	Ohik Kwon · Jaevin Park
18	글로벌 금융위기 전·후 외국인의 채권투자 결정요인 변화 분석: 한국의 사례	유복근
19	설비자본재 기술진보가 근로유형별 임금 및 고용에 미치는 영향	김남주
20	Fixed-Rate Loans and the Effectiveness of Monetary Policy	Sung Ho Park
21	Leverage, Hand-to-Mouth Households, and MPC Heterogeneity: Evidence from Korea	Sang-yoon Song
22	선진국 수입수요가 우리나라 수출에 미치는 영향	최문정 · 김경근
23	Cross-Border Bank Flows through Foreign Branches: Evidence from Korea	Youngjin Yun
24	Accounting for the Sources of the Recent Decline in Korea's Exports to China	Moon Jung Choi · Kei-Mu Yi
25	The Effects of Export Diversification on Macroeconomic Stabilization: Evidence from Korea	Jinsoo Lee · Bok-Keun Yu
26	Identifying Uncertainty Shocks due to Geopolitical Swings in Korea	Seohyun Lee · Inhwan So · Jongrim Ha
27	Monetary Policy and Income Inequality in Korea	Jongwook Park
28	How the Financial Market Can Dampen the Effects of Commodity Price Shocks	Myunghyun Kim
29	Which External Shock Matters in Small Open Economies? US Economic Policy Uncertainty vs. Global Risk Aversion	Youngju Kim · Hyunjoon Lim
30	Do Korean Exports Have Different Patterns over Different Regimes?: New Evidence from STAR-VECM	Sei-Wan Kim · Moon Jung Choi
31	기술진보와 청년고용	심명규 · 양희승 · 이서현

제2018 -32	북한지역 장기주택수요 및 연관 주택건설투자 추정	이주영
33	기업규모간 임금격차 원인 분석	송상윤
34	우리나라 고용구조의 특징과 과제	장근호
35	창업의 장기 고용효과: 시군구 자료 분석	조성철 · 김기호
36	수출입과 기업의 노동수요	음지현 · 박진호 · 최문정
37	청년실업의 이력현상 분석	김남주
38	노동시장 이중구조와 노동생산성: OECD 국가를 중심으로	최충 · 최광성 · 이지은
39	한국과 일본의 청년실업 비교분석 및 시사점	박상준 · 김남주 · 장근호
40	노동시장의 이중구조와 정책대응: 해외사례 및 시사점	전병유 · 황인도 · 박광용
41	최저임금이 고용구조에 미치는 영향	송헌재 · 임현준 · 신우리
42	최저임금과 생산성: 우리나라 제조업의 사례	김규일 · 육승환
43	Transmission of U.S. Monetary Policy to Commodity Exporters and Importers	Myunghyun Kim
44	Determinants of Capital Flows in the Korean Bond Market	Soohyon Kim
45	Central Bank Credibility and Monetary Policy	Kwangyong Park
46	통화정책이 자본유출입에 미치는 영향: 행태방정식 분석	이명수 · 송승주

제2018 -47	Commodities and International Business Cycles	Myunghyun Kim
제2019 -1	Deciphering Monetary Policy Board Minutes through Text Mining Approach: The Case of Korea	Ki Young Park · Youngjoon Lee · Soohyon Kim
2	The Impacts of Macroeconomic News Announcements on Intraday Implied Volatility	Jieun Lee · Doojin Ryu
3	Taking a Bigger Slice of the Global Value Chain Pie: An Industry-level Analysis	Chong-Sup Kim · Seungho Lee · Jihyun Eum
4	Trend Growth Shocks and Asset Prices	Nam Gang Lee
5	Uncertainty, Attention Allocation and Monetary Policy Asymmetry	Kwangyong Park
6	Central Bank Digital Currency and Financial Stability	Ohik Kwon
7	은행의 수익 및 자산구조를 반영한 통화정책 위험선호경로	김익진 · 정호성
8	혁신기업에 대한 산업금융 지원: 이론모형 분석	강경훈 · 양준구
9	가계부채 제약하의 통화정책: 2주체 거시 모형(TANK)에서의 정량적 분석	정용승 · 송승주
10	Alchemy of Financial Innovation: Securitization, Liquidity and Optimal Monetary Policy	Jungu Yang
11	Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea	Youngjoon Lee · Soohyon Kim · Ki Young Park
12	Tracking Uncertainty through the Relative Sentiment Shift Series	Seohyun Lee · Rickard Nyman
13	Intra-firm and Arm's Length Trade during the Global Financial Crisis: Evidence from Korean Manufacturing Firms	Moon Jung Choi · Ji Hyun Eum

제2019-14	특허자료를 이용한 우리나라 지식전파의 지역화 분석	이지홍 · 남윤미
15	Overhead Labour and Skill-Biased Technological Change: The Role of Product Diversification	Choong Hyun Nam
16	Does the Number of Countries in an International Business Cycle Model Matter?	Myunghyun Kim
17	High-Frequency Credit Spread Information and Macroeconomic Forecast Revision	Bruno Deschamps · Christos Ioannidis · Kook Ka
18	경제 분석을 위한 텍스트 마이닝	김수현
