BOK 이슈노트

AI 알고리즘을 이용한 산업 모니터링: 증권사 리포트 텍스트 분석

서범석
한국은행 조사국 거시모형팀 과장
Tel. 02-759-4248
bsseo@bok.or.kr

2023년 2월 17일

AI 등 통계 기법을 이용하면 무수히 많은 사람의 언어를 종합해서 빠르게 분석하는 것이 가능하다. 본 연구는 기업분석 전문가인 증권사 애널리스트들의 기업 평가 보고서 12만 8천건을 빅데이터로 구축하고, 이로부터 유의미한 산업별 모니터링 정보를 추출하였다. 이를 위해 자연어처리(natural language processing) 등 다양한 통계 기법을 이용하여 텍스트 정보를 가공하였으며, 동 과정의 알고리즘화를 통해 사람의 개입 없이 산업 동향을 파악하는 방안을 점검하였다.

구체적으로 본 연구는 증권사 보고서에 나타나는 숫자 정보는 모두 제거하고, 오직 텍스트에 나타나는 정성적(qualitative) 정보만을 이용하여 애널리스트들의 생각을 취합하였다. 이를 통해 애널리스트들이 평가하는 기업 업황을 산업별 · 지역별로 추정하고, 업황의 변동요인을 통계 알고리즘을 이용하여 요약 · 정리하였다. 또한 증권사 보고서의 텍스트 분석을 통해, 환율, 금리 등 주요 경제 이슈에 대한 전문가들의 견해를 알고리즘으로 취합하고, 취합 결과를 새로운 정량 지표로 제시하였다.

분석 결과 새롭게 제시한 텍스트 지표는 GDP, BSI 등 거시경제 지표를 예측하는 데 매우 유용한 것으로 나타났고, 산업별 변동요인 파악에도 효과적인 것으로 나타났다. 특히 텍스트 지표와 경기선행지수 순환변동치와의 Granger 인과관계를 분석해 보면, 코스피 컨센서스 전망치에는 나타나지 않는 경기선행지수로의 일방향적 인과관계가 텍스트 지표에는 존재하는 것으로 나타난다. 이러한 결과는 애널리스트들이 제시하는 텍스트 정보에 숫자가 전달하지 못하는 새로운 정보가 반영되고 있을 가능성을 시사한다.

본 연구에서 제시한 텍스트 분석 과정은 통계 알고리즘을 이용하여, 웹 스크래핑(web-scraping)을 통한 보고서의 입수부터, 업황 파악과 변동요인 분석, 그리고 분석 결과의 시각화까지 모두 알고리즘화가 가능하도록 설계되었다.

기술 발전에 따른 자동화와 효율성 제고가 가속화되고 있다. 자연어처리를 이용한 경제분석은 아직 연구 초기 단계이지만, GPT 모형 등 최근의 기술 발전 속도를 생각하면 빠른 시일 내에 스스로 정보를 취합하고 경제 판단을 내릴 수 있는 통계 모형의 실현이 가능할 것으로 판단된다. 따라서 본 연구에서 제시한 알고리즘 등 텍스트 분석 연구를 지속해 나갈 필요가 있으며, 경제 분야의 연구 효율이 개선될 수 있도록 AI 등의 통계 기법을 계속 발전시켜 나가야 할 것이다.

■ 논고 작성에 커다란 도움을 주신 통계연구반 김민수 반장, 조형배 조사역과 유익한 논평을 주신 기업통계팀 김대진 팀장, 지역연구지원팀 김정성 팀장, 국제경제연구실 한바다 과장, 은행분석팀 유재원 과장께 감사를 표합니다. 본문에 남아있는 오류는 저자의 책임임을 밝힙니다.

한국은행

Ⅰ. 검토 배경

AI를 이용한 언어분석 기술이 각광을 받고 있다. 정보를 주고받는 가장 기본적인 수단이 사람의 언어라는 점에서, 언어분석 기술은 경제 분야에서도 활용 가치가 매우 높다. 무수히 많은 사람의 언어를 종합해서 이해할 수 있다면, 향후 경기 향방에 대한 거시적 판단에 큰 도움이 될 것은 자명하다.

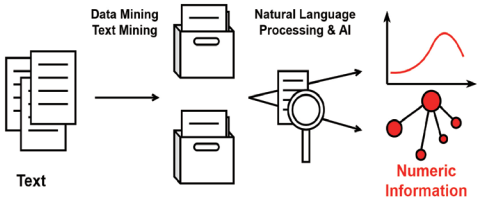
본 연구는 기업분석 전문가인 증권사 애널리스트들의 기업평가 보고서를 빅데이터로 구축하고, 이로부터 유의미한 경제적 정보를 추출하고자 하였다. 본 연구는 수치로 나타나는 지표는 모두 제외하고, 오로지 텍스트에 나타나는 정성적(qualitative) 정보만을 이용하여 애널리스트들의 생각을 취합하였다.

텍스트는 수치로 표현하기 힘든 미묘한 뉘앙스를 전달하고 저자의 주관적 견해와 관심사 등을 반영한다. 본 연구는 사람이 읽고 정리할 수 없을 정도의 거대한 양의 텍스트를 다양한 통계 기법을 이용하여 분석하였다. 이를 통해 숫자가 전달하기 어려운 정보를 취합하고, 취합한 정보를 경제 분석에 활용하는 새로운 통계 기법들을 제시하였다.

구체적으로, 본 연구는 텍스트 분석을 통해 애널리스트들이 평가하는 기업 업황을 산업별 · 지역별로 추정하고, 업황의 변동요인을 알고리즘을 이용하여 요약 · 정리하였다. 특히 변동요인은 산업별 · 지역별 이슈를 한눈에 파악할 수 있도록 테이블 형태로 시각화하여 제시하였다.

또한 증권사 보고서의 텍스트를 분석하면 환율, 금리 등 주요 경제 이슈에 대한 전문가들의 견해를 종합해서 파악하는 것이 가능하다.

〈그림 1〉 증권사 리포트 텍스트 정보의 정량화



이러한 견해를 정량지표로 분석할 수 있도록 주요 이슈에 대한 영향도 및 평가 지표의 작성 방법을 제시하였다.

한편 산업 간 공통으로 나타나는 업황 변화 요인을 살펴보면, 산업 간 경영환경의 유사도 또한 유추해 볼 수 있다. 산업 간 유사도를 추정하는 새로운 정량지표의 작성 방법을 텍스트 분석을 통해 제시하였다. 텍스트에서 추출한 산업 간 유사도는 산업 간 유사한 정도뿐만 아니라 그 이유에 대한 정보도 제공이 가능하다.

본 연구에서 추정한 모든 텍스트 지표들은 통계 알고리즘을 이용하여 사람의 개입 없이 추정할 수 있도록 Python 프로그램을 바탕으로 설계하였다. 이를 위해 본 연구에서는 자연어처리(natural language processing) 기술과 데이터 마이닝 기법 등 다양한 통계 기법을 활용하였다. 구체적으로 텍스트 데이터의 논조 파악을 위해 트랜스포머(transformer) 기반의 감성분석(sentiment analysis) 모형을 활용하였고, 키워드 빈도 분석, 동적인자모형(dynamic factor model) 기반의 시계열 분석, 네트워크 데이터 분석 등을 활용하여 텍스트 데이터로부터 유의미한 경제적 정보를 추출하였다.

경제금융 보고서의 텍스트 분석을 연구한 기존 연구들은 대부분 10-K 보고서 등 사업보고서에 대해 감성분석 모형을 적용하거나(Humpherys et al. 2011, Wu et al.

2012), 토픽 모델링을 적용하여 주가 예측력을 검증하는 데 초점이 맞춰져 있다(Guo et al. 2016, Lewis and Young 2019). 본 연구는 기존 연구와 달리, 정밀한 데이터 마이닝 기법들을 적용하여 산업 분석에 유용한 텍스트 지표들을 제시하고, 이 과정을 사람의 개입 없이 알고리즘화하는 방안을 찾는 데 주안점을 두었다.

방대한 양의 텍스트 정보를 알고리즘으로 취합할 수 있다면, 기업 정보의 1차 생산자인 애널리스트들의 생각을 실시간으로 취합할 수 있고, 이는 정보의 2차 가공자인 경제 분석 연구자들의 업무 효율을 크게 개선시킬 수 있을 것으로 판단한다.

특히 대용량 텍스트의 경우 사람이 모두 읽고 분석하는 것은 불가능에 가깝다. 그러나 AI 등 통계 알고리즘을 이용하면 웹 스크래핑(web-scraping)을 통한 문서의 입수부터, 업종별 업황 파악과 변동요인 분석, 그리고 분석 결과의 시각화까지 십여분이면 충분하기에 업무 효율의 획기적인 개선을 기대할 수 있을 것으로 판단한다.

본 논고는 다음과 같이 구성하였다. Ⅱ장에서 자연어처리를 활용한 경제 분석 사례를 살펴보고, 일반적인 자연어처리 기술 등 선행 연구를 소개하였다. 이어지는 Ⅲ장에서 증권사 보고서의 입수 과정과 텍스트 데이터의 추출 과정을 자세히 설명하였다. 그런 다음 Ⅳ장에서 새로운 텍스트 지표의 작성 방법을 제시하였고, Ⅴ장에서 이들 지표의 타당성을 검증하였다. 마지막으로 Ⅵ장에서 본 논고의 시사점 및 향후 발전방향을 정리하고 본 논고를 마무리하였다.

Ⅱ. 선행 연구

1. 경제 분석을 위한 자연어 처리

최근 Python 기반 프로그래밍 환경과 자연어처리(natural language processing) 기술이 비약적으로 발전하면서, 경제 분석에 텍스트 데이터를 접목하려는 시도가 크게 늘고 있다.

해외에서는 이미 2010년대 중반부터 텍스트를 이용한 경제 분석이 활발하게 이루어졌다. 텍스트 데이터의 원천으로는 언론의 견해를 반영하는 뉴스 텍스트(Baker et al 2016, Shapiro et al 2020), 대중의 시각을 전달하는 SNS나 검색 정보(Sun et al. 2016), 기업 회계 보고서 및 평가 보고서(Lewis and Young 2019) 등 다양한 텍스트가 활용되었다. 대부분의 연구는 사전접근법(lexical approach) 또는 통계 모형에 기반한 감성 분석(sentiment analysis)과 토픽 모델링의 방법을 이용하여 텍스트를 분석하고 있다. 최근에 와서는 보다 복잡한 형태의 인공신경망 모형을 이용한 연구도 활발하게 진행되고 있다(Hajek and Olej 2013).

텍스트는 비정형 데이터로 사실상 전달하는 정보의 범위에 한계가 없다. 따라서 기존 연구들은 텍스트를 분석하여, 물가, 주가 등 가격 지표를 예측하거나(Kalamara et al. 2022, Li et al. 2020), 위기 지표, 불확실성 지표 등 새로운 경제정보를 산출하고(Li et al 2009, Baker et al 2016), 사기 탐지(fraudulent detection)(Li et al. 2020), 기업의 지속가능성 분류(Te Liew et al. 2014), 신용 평가(credit scoring)(Yap et al. 2011)에 이용

하는 등 다양한 목적을 위해 텍스트를 활용하고 있다.

국내에서는 한국은행이 선도적으로 경제 분석을 위한 텍스트 데이터의 활용을 확대하고 있다. Lee et al(2019)은 뉴스 텍스트 분석을 통해 통화정책 서프라이즈 지수를 산출하였고, 서범석(2022 a)은 뉴스 텍스트를 이용한 경기 예측 방법 및 물가, 주가, 주택가격 등 15개 부문의 경제 부문별 텍스트 지표의 작성 방법을 제시하였다. 또한 Seo et al(2022 b)은 트랜스포머(transformer) 기반의 자연어처리 기법을 이용하여 뉴스심리지수를 작성하였고, 한승욱 외(2022)는 비슷한 방법으로 인플레이션 어조지수를 평가하였다.

최근에는 뉴스 데이터 이외에도 트위터 데이터를 이용한 핀테크 트렌드 분석(김도희 · 김민정 2022), ESG 보고서를 이용한 ESG 방향성 평가(박수빈 · 이용규 2022) 등 다양한 텍스트 분석이 국내에서 시도되고 있다.

다만 국내외 대부분의 연구는 텍스트에서 추출한 정보를 일회성으로 평가하는 데 보다 초점이 맞춰져 있다. 텍스트가 정보를 전달하는 가장 기초적인 수단이라는 점에서, 텍스트 분석은 Jeseena and David(2014), Lewis and Young(2019) 등이 지적한 바와 같이, 경제 정보를 취합하고 분석하는 과정을 자동화하는 데도 매우 긴요하다. 특히 최근의 ChatGPT 등 발전된 자연어처리 기술은 텍스트 분석 기술이 경제 분석의 자동화에 커다란 혁신을 가져올 수 있음을 시사한다(Siegele 2022). 블룸버그 저널리스트 Weisenthal(2022)은 그의 기사 "This AI Chatbot is a Shockingly Competent Macro Pundit"에서 텍스트 분석 기술이 경제 분석 자동화에 유용할 수 있음을 구체적으로 보여주고 있다.

본 연구는 기존 연구와 달리 산업별 업황 분석 및 변동요인 파악에 유용한 텍스트 지표를 제시하고, 추정 과정의 알고리즘화 가능성을 논함으로써, 향후 경제 분석의 자동화 연구에 도움이 되고자 하였다.

2. 자연어 처리 방법론

기본적인 자연어처리 방법론을 적용하기 위해서는 텍스트 데이터를 정량 데이터로 전환하는 전처리 과정이 필요하다. 이는 보통 텍스트 데이터를 입수하는 과정, 동일한 단어를 알고리즘이 동일하게 인식할 수 있도록 단어를 형태소로 분해하는 과정(tokenization), 형태소 데이터를 범주형 데이터 형태인 Bag of words 데이터로 전환하는 과정(integer encoding or embedding) 등을 포함한다. 이에 대한 자세한 설명은 서범석 외(2022 b), 김수현 외(2018) 등이 소개하고 있다.

최근의 자연어처리 방법론은 특정한 목적 없이 대용량 텍스트의 문장 패턴을 학습시킨 인공신경망 모형을 먼저 추정하고, 이를 여러 목적에 맞게 활용하는 추세이다. 이런 거대 인공신경망 모형으로는 BERT(Bidirectional Encoder Representations from Transformers), GPT(Generative Pre-trained Transformer) 등의 구조가 주로 활용되고 있고, 감성분석, 문장생성 등의 분야에서 현재 상업화가 가능한 수준까지 빠르게 발전하고 있다. 이에 대한 자세한 설명은 〈부록 2〉에 기술하였다.

본 연구에서는 증권사 리포트에 나타나는

문장의 논조를 평가하기 위해 BERT 모형을 이용하였고, 변동요인 파악을 위해서는 키워드 빈도 분석의 방법론을 이용하였다. 키워드 빈도 분석은 문장구조 패턴은 무시하고 키워드 패턴만을 고려하여 문장에서 중요한 정보를 추출한다. 키워드 빈도 분석은 학습데이터 없이 비지도(unsupervised) 형식으로 활용이 가능하므로 학습데이터 구성을 위한 시간과 비용을 들이지 않고 활용할 수 있는 장점이 있다 [1].

특히 본 연구의 변동요인 분석을 위해서는 연속된 3개의 단어(형태소) 패턴을 이용하여 텍스트를 분석하는 방법인 Trigram 방법론을 이용하였다. 경제경영 텍스트의 경우 합성어가 많으므로, 1개의 단어를 이용하여 키워드 패턴을 분석하는 Unigram에 비해 Trigram을 이용하는 것이 효과적이다. Trigram을 이용하면 경제적 의미가 낮은 일상 표현 등 불용어(stopwords)를 제거하는 과정이 수월해지는 장점이 있다.

Ⅲ. 증권사 애널리스트 리포트 텍스트 데이터

이제 본 연구에서 이용한 증권사 기업평가 보고서의 입수 과정과 이들 보고서의 텍스트 데이터 변환 과정을 자세히 소개한다.

1. 애널리스트 리포트의 입수

웹 스크래핑(web-scraping) 기술을 이용하여 2019년에서 2022년 중 발간된 증권사 기업 평가 보고서 12만 8천건을 수집하였다. 이는 52개 증권사 1,079명의 애널리스트가 작성한 2,283개 기업에 대한 기업 분석 보고서로 월평균 2천 문건에 해당한다.

애널리스트 보고서 수집시 산업 및 거시 분석 보고서는 제외하고, 개별 기업 분석 보고서만 수집하였다. 이는 개별 기업 분석 결과를 산업 및 거시 분석으로 확장하여 활용하기 위함이다. 예를 들어, 환율 상승이 수출기업에

〈표 1〉 증권사 애널리스트 리포트 텍스트 데이터

(업종별 대상 기업 수)

업 종	기업수
전자/영상/통신장비	257
정보통신업	186
전문/과학/기술	170
기타기계/장비	170
의료물질/의약품	155
화학물질/제품	124
도매/소매	112

(업종별 리포트 수)

업 종	리포트 수
전자/영상/통신장비	12,102
정보통신업	10,618
화학물질/제품	9,152
금융	8,227
도매/소매	6,537
의료물질/의약품	4,976
전문/과학/기술	4,456

(업종별 유효 문장 수)

업 종	문장 수
전자/영상/통신장비	172,713
정보통신업	154,058
화학물질/제품	107,047
의료물질/의약품	87,188
금융	74,255
전문/과학/기술	73,208
도매/소매	69,124

(분기별 대상 기업 / 작성 증권사 / 리포트 / 유효 문장 수)

	19q1	19q2	19q3	19q4	20q1	20q2	20q3	20q4	21q1	21q2	21q3	21q4	22q1	22q2	22q3	22q4
기업 수	784	849	777	911	763	824	833	990	981	986	941	1,023	983	1,073	860	1,058
증권사 수	35	35	34	34	34	35	34	34	34	35	36	35	38	42	41	40
리포트 수	5,946	7,052	6,698	6,764	6,146	6,983	6,403	7,863	6,669	7,046	6,597	6,195	5,873	6,479	6,004	6,576
문장 수	85,515	87,689	84,207	84,191	80,561	90,341	90,849	103,091	92,311	87,828	87,673	81,963	80,175	91,709	81,486	101,810

1) 향후 본 연구에서 추출한 키워드를 지도학습 데이터로 이용하여 BERT 등 통계 모형에 활용하면 경제 키워드 추출 모형을 구성하는 데 도움이 될 것으로 사료된다.

유리한 것은 자명한 사실이나, 구체적으로 어떤 기업에 얼마나 유리하며, 이를 국가 전체로 취합할 경우 어떤 효과가 나타나는지 구체화하는 것은 쉽지 않다. 개별 기업에 대한 미시적 평가를 바탕으로 국가 전체에 대한 거시적 효과를 분석할 경우, 보다 구체적인 경제 분석이 가능한 이점이 있다. 따라서 미 · 거시 분석의 연계가 가능하도록, 본 연구에서는 개별 기업 보고서를 텍스트 데이터의 기초자료로 구성하였다.

2. 텍스트 데이터 추출

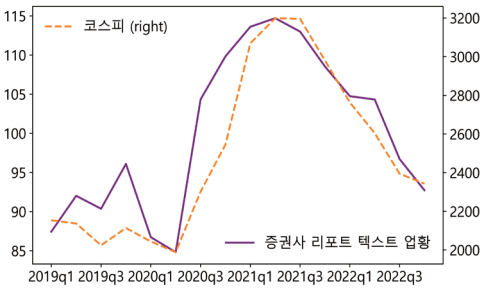
수집한 보고서는 2,283개 기업에 대한 기업 이슈, 평가, 전망, 투자 의견 등을 반영한다. 이들 보고서는 산업별 분석이 용이하도록 한국거래소 기준 156개 업종으로 구분하였고, 이를 다시 한국은행 기업경영분석 통계 기준 40여개 업종으로 분류하였다. 또한 산업 분석을 지역별로도 수행할 수 있도록 2,283개 기업의 지역정보를 메타데이터로 연결하여, 기업을 본사 소재지 기준 지역별로도 분류하였다.

입수한 보고서는 신속한 분석을 위해 보고서 작성 시점을 기준으로 분기별로 취합하여 분석하였다. 증권사 기업 평가 보고서는 월별로 취합하는 것이 가능하지만, 시점별 자료의 양, 시계열의 변동성, 계산 편의 등을 고려하여 분기별로 취합하였다. 또한 보고서 작성 시점에 사용된 정보는 모두 해당 시점의 최신 정보라는 가정하에 보고서에서 언급하는 대상 시점이 아닌, 보고서 발간 시점을 기준으로 자료를 취합하였다.

입수한 보고서는 제 Ⅱ장에서 소개한 전처리 과정을 거쳐 분석이 용이한 Bag of words 데이터로 전환하였다. 이 과정에서 12만 8천건의 보고서에서 약 16백만개의 문장을 추출하였다. 이들 문장 중 중복 문장, 단순 수치 언급, 이해관계 고지 등 대상 기업에 대한 정성 정보를 포함하지 않는 문장을 제거하여 약 145만개의 유효 문장을 선별하였다. 이는 글자수 기준으로 약 3천만자에 해당하며 200자 원고지 기준 16만장 분량에 해당하는 텍스트이다.

이런 대용량 텍스트를 사람이 모두 읽고 분석하는 것은 불가능에 가깝다. 그러나 Python 등 통계 프로그램을 이용하면 웹 스크래핑을 통한 문서의 입수부터, 전처리까지 모두 알고리즘화가 가능하다.

〈그림 2〉 전산업 텍스트 업황($TBCI_{t,t}$)과 코스피 지수



Ⅳ. 텍스트 기반 산업 모니터링 정보 추출

이번 장에서는 본격적으로 Ⅲ장에서 추출한 텍스트 데이터를 바탕으로 경제 모니터링 지표를 작성하는 방법들을 제시하고, 동 방법들의 알고리즘화 가능성을 평가하였다.

1. 텍스트 기반 업황 지수(Text-based Business Confidence Indicator, TBCI)

텍스트는 수치로 파악하기 힘든 저자의 견해를 잘 보여준다. 따라서 기업 업황에 대한 전문가들의 견해를 취합하고자, 앞서 추출한 텍스트데이터에 감성분석(sentiment analysis) 모형을 적용하였다. 이를 위해 트랜스포머(transformer) 기반의 통계 모형을 구축하고, 문장을 긍정, 부정 또는 중립으로 분류하였다. 감성분석 모형은 공개된 문장을 미리 학습한 Pre-trained 모형에 Seo et al.(2022 b)이 이용한 뉴스 데이터를 일부 학습시켜 추정하였다. 감성분석 모형에 대한 자세한 설명은 Seo et al.(2022 b)을 참고하기 바란다.

구체적으로 업종 i의 t기 텍스트 기반 업황 지수(Text-based Business Confidence Indicator), $TBCI_{i,t}$를 다음과 같이 추정하였다.

$$TBCI_{i,t} = \frac{X_{i,t} - \overline{X}_{i,.}}{s_{i,.}} \times 10 + 100,$$

$$X_{i,t} = \frac{\sum_{s \in S_{i,t}} I(s \in P) - I(s \in N)}{\sum_{s \in S_{i,t}} I(s \in P) + I(s \in N)},$$

$$\overline{X}_{i,.} = \sum_{t}^{T} X_{i,t} / T,$$

$$s_{i,.} = \sqrt{\sum_{t}^{T} (X_{i,t} - \overline{X}_{i,.})^2 / T - 1}.$$

여기서 P와 N은 각각 가능한 모든 긍정 및 부정 문장의 집합을 의미하고, $S_{i,t}$는 증권사 리포트에서 추출한 업종 i의 t기 문장 샘플 중복집합(multiset), $I(s \in A)$는 샘플 s가 집합 A에 속할 경우 1, 속하지 않을 경우 0을 반환하는 지시함수를 의미한다.

즉, 업황 $TBCI_{i,t}$는 업종 i와 관련한 증권사 리포트의 문장 논조를 분류한 뒤, 산출한 긍정 문장과 부정 문장 수의 차이를 두 수의 합계로 나누어 추정하고, 장기평균을 이용해 표준화하여 산출하였다.

텍스트 업황은 텍스트를 입력 데이터로 받아서 업황을 출력하는 통계 모형에 의해 추정하므로 동 과정은 자동화가 가능하다.

2. 기업경영환경 변화 요인표 (Business Condition Factors, BC-factors)

텍스트 지표는 기초 자료가 문장으로 이루어지는 만큼, 변동요인 파악이 매우 용이하다. 증권사 리포트에서 추출한 유효 문장에 대해 키워드 빈도 분석을 적용하였다. 이를 통해 기업 업황에 영향을 미치는 기업경영환경 변화 요인(Business Condition Factors, BC-factors)을 다음과 같이 추정하였다.

먼저 효과적인 요인 파악을 위해 경제적으로 의미 없는 불용어(stopwords)를 분석 과정에서 제거해야 한다. 전체 문장을 대상으로 키워드를 분석하면, 일반적인 회계표현이

〈표 2〉 요인 평가 문장 예시

	요인	관계언	평가
1	코로나 바이러스 영향	으로(로)	유통업계 방문객이 감소
2	판관비 감소도 가능하기	때문(에)	추가 이익 증가를 예상
3	중국전지 판매가 호조를 기록	(하)면서	흑자 전환에 성공할 전망
4	NIM 하락폭 축소에	따라(서)	이자이익 증가폭 확대 예상

많이 나타나므로 요인 파악이 쉽지 않다. 따라서 문장 구조를 분해하여 관계언을 중심으로 유효 문장을 〈표 2〉와 같이 요인과 평가로 분해하였다. 그런 다음, 요인 표현에서 일반명사, 고유명사, 동사, 형용사를 제외한 모든 품사를 제거하고, 자주 등장하는 일반적 회계표현 등의 불용어[2]를 제거한 뒤, Trigram 방법론을 이용하여 기업경영환경 변화요인을 추정하였다.

예를 들어, '신모델 출시 효과로 전분기 대비 판매 호조를 보이면서 매출과 영업이익에서 견조한 성장세'와 같은 예시 문장은 다음과 같이 분해가 된다.

(요인)		(평가)
신모델 · 출시 · 효과	→	판매 · 호조 · 보이
판매 · 호조 · 보이	→	매출 · 영업이익 · 견조, 영업이익 · 견조 · 성장, 견조 · 성장 · 세

† 불용어: 전분기대비

추출한 요인 키워드는 업종 및 분기별로 취합한 뒤, 알고리즘을 이용해서 다음과 같이 상위 5개를 선정하여 〈그림 3〉의 테이블 형태로 출력하였다. 이때 비슷한 이슈들은 동일한 색상으로 출력되도록, 요인 정보를 키워드 중심으로 분류하여 시각화하였다.

$$BC-factors^{5}_{i,t} = \{(\omega^{[K]},\cdots,\omega^{[K-4]}) \mid \omega^{[k]} = f(v_{(k)}),\ v = m_{U_{i,t} \setminus W}(\omega),\ v_{(k)}\ is\ the\ k^{th}\ order\ statistic\ of\ v\}.$$

여기서 $U_{i,t}$는 증권사 리포트에서 추출한 업종 i의 t기 Trigram 패턴 샘플의 중복집합, $m_A(\omega)$는 multiplicity function, 즉 A에 나타나는 ω의 개수, W는 Trigram 패턴 불용어 집합, K는 불용어를 제외한 Trigram 패턴의 수, 즉 $K = |Supp(U_{i,t} \setminus W)|$를 의미한다.

〈그림 3〉과 같은 경영환경 변화요인표는 산업별 주요 이슈를 개괄적으로 파악하는 데 매우 유용하다. 또한 추정 과정은 알고리즘을 이용하므로 자동화가 가능하다. 그러나 키워드만으로 산업별 이슈의 구체적 내용을 파악하는 것은 어려우며, 구체적 경제 분석을 위해서는 이슈와 연결된 문장을 직접 확인하는 과정이 필요하다. 그럼에도 불구하고 텍스트 분석을 이용하면 증권사 보고서의 문장을 이슈별로 모아서 살펴볼 수 있으므로, 구체적 경제 분석을 위해서도 동 방법론의 효용성은 매우 높을 것으로 사료된다.

〈그림 3〉 정보통신업의 기업경영환경 변화 요인표($BC-factors^{5}_{i,t}$) 예시[1,2]

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	3845 48 / 32	3992 61 / 33	4753 55 / 31	4533 65 / 29	4749 58 / 29	5205 69 / 30	4923 69 / 30	5468 83 / 32	5261 82 / 33	4728 80 / 30	4923 86 / 33	5542 102 / 33	6405 95 / 32	5195 105 / 36	4506 77 / 34	5248 97 / 31
1	마케팅비용증가	마케팅비용증가	의료정보시스템	마케팅비용증가	온라인쇼핑	온라인쇼핑	온라인쇼핑	의료정보시스템	웹툰웹소설	자사주소각	망연계솔루션	빅데이터분석	온라인게임	온라인게임	공급망관리	의료정보시스템
2	검사막모바일	감가상각증가	온라인게임	5G가입증가	검사막모바일	이동전화매출	5G가입증가	온라인게임	취약점진단	중간지주사	인공지능빅	온라인쇼핑	모바일게임시장	공간정보플랫폼	게임매출감소	데이터센터화재
3	온라인게임	5G가입증가	5G가입증가	온라인교육	5G가입증가	마케팅비용증가	온라인교육	중단사업손익	중간지주사	지배구조개편	지능빅데이터	전자지급결제	의료정보시스템	검사막모바일	미디어콘텐츠사	게임매출감소
4	사막온라인	선택약정할인	마케팅비용증가	이동전화매출	자율주행자동	5G가입증가	인공지능기술	지배구조개편	5G가입증가	자사주매입	기존게임매출	마이데이터사업	검사막모바일	온라인쇼핑	시장성장둔화	웹툰웹소설
5	홈쇼핑송출	검사막모바일	선택약정요금	온라인쇼핑	데이터법통과	빅데이터분석	카메라모듈검사	온라인쇼핑	기업가치상승	기업가치상승	검사막모바일	유무선통신스	튜디오연결편성	정보보호산업	이동전화매출	기존게임매출

주: 1) N: 해당 분기 및 업종에 나타나는 총 문장 수, C: 기업 수, A: 보고서 작성기관 수　　2) 키워드 순서(1~5)는 많이 언급된 상위 5개 키워드의 내림차순 순서를 의미

2) 내년상반기, 올해하반기, 분석보고서, 당사자료작성, 일평균거래, 주요변동사항, 기타포괄손익 등 다수

3. 주요 이벤트의 영향도 (Text-based Impact of an Event Indicator, TIEI)

증권사 리포트에 나타나는 특정 키워드의 언급 빈도를 살펴보면, 특정 이벤트가 업종별로 얼마나 주요하게 언급되고 있는지를 쉽게 파악할 수 있다. 따라서 키워드 언급빈도를 이용하여 주요 이벤트의 업종별 영향도(Text-based Impact of an Event Indicator, TIEI)를 다음과 같이 정의하고 정량지표로 추정하였다.

$$TIEI_{i,t}^{A} = \frac{\sum_{s \in S_{i,t}} I(s \in A_w)}{|S_{i,t}|} \times 100.$$

여기서 A_w는 단어군 w를 포함하는 가능한 모든 문장의 집합을 의미하고, $S_{i,t}$와 $I(\cdot)$는 앞에서 설명한 바와 같다.

즉, 주요 이벤트의 업종별 영향도는 증권사 리포트에 나타나는 총 문장 수 대비 특정 단어를 포함하는 문장의 수로 추정하였다. 추정 결과는 영향도가 높은 업종을 기준으로 정렬한 뒤 시계열 그래프로 출력하여, 각 업종의 관점에서 영향도가 높은 업종을 먼저 보여주도록 설계하였다. 예를 들어 '러우전쟁'의 업종별 영향도 $TIEI_{i,t}^{러우전쟁}$는 증권사 리포트에서 '러시아', '우크라이나', '러우'를 포함하는 문장의 상대 빈도로 계산하였으며 〈그림 4〉와 같이 추정해 볼 수 있다.

업종별 영향도는 알고리즘으로 추정하므로 특정 이슈가 주어지면, 그에 대한 영향도를 자동으로 추정하는 것이 가능하다.

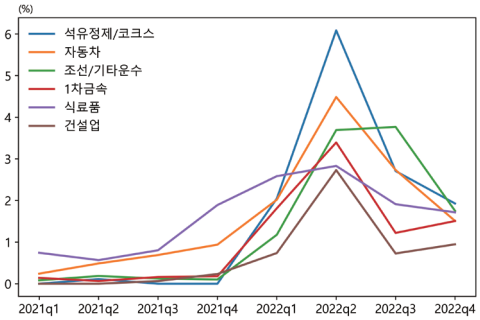
4. 주요 이벤트에 대한 평가 (Text-based Evaluation of an Event Indicator, TEEI)

논조 정보를 이용하면 주요 이벤트에 대한 애널리스트들의 평가를 정량적으로 파악하는 것도 가능하다. 특정 이벤트가 업종별로 호재인지 악재인지 판단하기 위해 업종별 평가(TEEI, Text-based Evaluation of an Event Indicator)를 다음과 같이 정의하고 정량지표로 추정하였다.

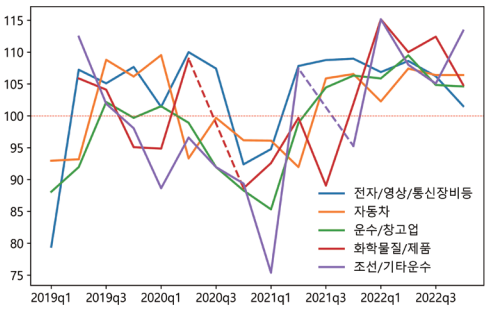
$$TEEI_{i,t}^{A} = \frac{X_{i,t} - \overline{X}_{i,\cdot}}{s_{i,\cdot}} \times 10 + 100,$$

$$X_{i,t} = \frac{\sum_{s \in S_{i,t}} I(s \in P \cap s \in A_w) - I(s \in N \cap s \in A_w)}{\sum_{s \in S_{i,t}} I(s \in P \cap s \in A_w) + I(s \in N \cap s \in A_w)}.$$

〈그림 4〉 업종별 러 · 우전쟁 영향도($TIEI_{i,t}^{러우전쟁}$) 예시



〈그림 5〉 업종별 환율에 대한 평가($TEEI_{i,t}^{환율}$)[1)]



주 : 1) 점선은 결측구간을 선형보간(linear interpolation)했음을 의미하며 선이 없는 구간은 자료가 없음을 의미

여기서 P, N, A_w, $S_{i,t}$와 $I(\cdot)$는 앞에서 설명한 바와 같다.

즉, 업종별 평가는 특정 단어를 포함하는 긍·부정 문장 수의 차이를 그 합계로 나누어 추정하였다. 추정 결과는 긍·부정 문장 수를 기준으로 업종을 정렬한 뒤 시계열 그래프로 출력하여, 시장의 관심이 높은 업종을 먼저 보여주도록 설계하였다.

텍스트 데이터는 기초 자료가 문장으로 이루어지는 만큼 특정 평가에 대한 기초 문장을 확인하면 평가 결과뿐만 아니라 구체적인 평가 내용도 파악할 수 있다. 예를 들어 '환율'의 업종별 평가 $TEEI_{i,t}^{환율}$을 〈그림 5〉와 같이 살펴보면, 각 시점별로 환율 흐름이 호재로 작용하고 있는지 악재로 작용하고 있는지 유추해 볼 수 있고, 관련 기초 문장을 확인해보면 업종별로 다르게 나타나는 환율의 구체적 영향까지 파악할 수 있다. 자세한 분석 결과는 제 Ⅴ장 텍스트 정보의 타당성 검증을 참고하기 바란다.

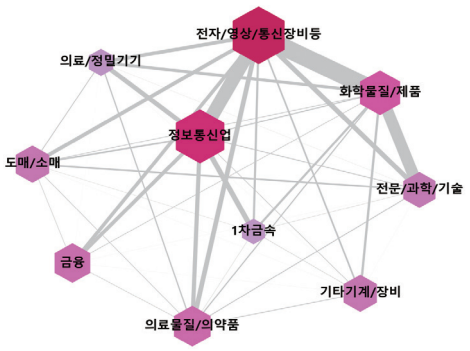
앞서 설명한 업종별 영향 지표와 평가 지표는 일반적으로 특정 이벤트의 산업별 영향을 정량화해서 비교하는 것이 어렵다는 점에서 그 효용성이 매우 높다. 영향 및 평가 지표는 특정 이슈에 대한 애널리스트들의 생각을 서베이로 조사한 것과 비슷하게 결과를 활용할 수 있다. 또한 특정 이슈가 주어지면 알고리즘을 이용하여 즉시 계산이 가능하다는 점에서 자동화 가능성도 매우 높다. 다만 구체적 평가 내용을 파악하기 위해서는 평가와 연결된 기초 문장들을 직접 확인하는 과정이 필요하다.

5. 공통요인 기반 산업별 유사도 (Business Condition Similarity, BC-Similarity)

업종 간 공통으로 나타나는 기업경영환경 변화요인을 살펴보면, 외부 요인에 대한 언급 빈도를 기준으로 두 업종 간의 경영환경 유사도(Business Condition Similarity, BC-similarity)를 〈그림 6〉과 같이 유추해볼 수 있다. 예를 들어, '전자/영상/통신장비'와 '화학물질/제품'에서 '전기차', '배터리', '원자재' 등이 공통으로 많이 언급되고 있다면, 두 업종의 경영환경은 전기차 수요 등 비슷한 요인에 의해 영향받을 것이란 점을 쉽게 알 수 있다. 이러한 정보는 산업 분석 및 전망 등에 유용하게 활용될 수 있다. 또한 텍스트를 바탕으로 추정한 업종별 유사도는 유사한 정도뿐만 아니라 요인 파악이 용이하다는 점에서 활용 가능성이 더욱 높다.

산업간 유사도를 추정하기 위해 Ⅳ-2에서 구한 요인 표현을 기준으로, 각 업종쌍 (i,j)에 대해 공통으로 나타나는 요인들의 분포를 Kullback-Leibler(KL) Divergence로 추정하였다.

〈그림 6〉 산업 간 유사도($BC-similarity_{(i,j),t}$) 예시[1]



주: 1) 노드(node)의 크기는 해당 기간 중 증권사 리포트에 나타나는 업종별 문장 수에 비례하고 연결선(edge)의 굵기는 두 업종간 공통요인 분포의 KL divergence 역수에 비례

$$d_{KL}(P_{i,t}\|P_{j,t}) = \sum_{u \in (U_{i,t}\cup U_{j,t}) \setminus W} P_{i,t}(u)\log\left(\frac{P_{i,t}(u)}{P_{j,t}(u)}\right).$$

$$BC-similarity_{(i,j),t} = \frac{2}{d_{KL}(P_{i,t}\|P_{j,t}) + d_{KL}(P_{j,t}\|P_{i,t})}.$$

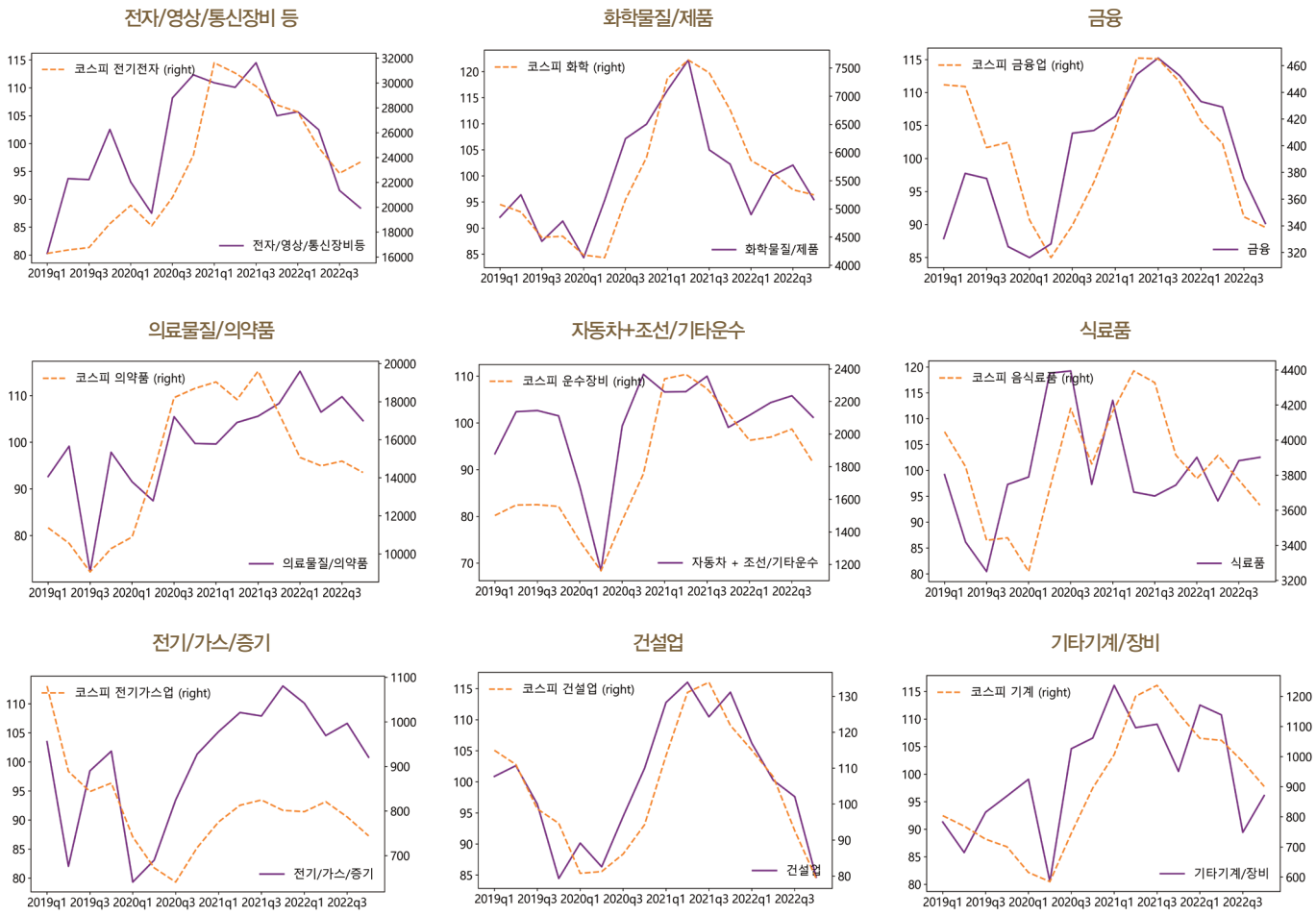
여기서 $U_{i,t}$와 $U_{j,t}$는 각각 업종 i와 j의 Trigram 패턴 샘플의 중복집합, $P_{i,t}(u)$는 Trigram 패턴 u가 업종 i와 관련한 증권사 리포트 텍스트에서 나타나는 상대 비율을 계산한 것이다.

산업 간 유사도는 위에서 추정한 KL Divergnece의 역수를 이용하여 다음과 같이 추정하였다.

이렇게 구한 산업 간 유사도는 네트워크 데이터로 구성할 수 있다. 〈그림 6〉은 문장 수 기준 상위 업종 10개의 산업 간 유사도를 네트워크 데이터로 시각화하여 출력한 것이다.

산업간 유사도 지표는 KL divergence에 의해 동일한 방식으로 추정하므로 추정 과정의 자동화가 가능하다.

〈그림 7〉 산업별 텍스트 업황 지수와 코스피 산업별 지수 비교



주: 1) 실선은 증권사 리포트 텍스트로부터 추정한 업종별 업황 지수($TBCI_{i,t}$)이며 점선은 관련 업종의 코스피 산업별 지수를 나타냄

〈표 3〉 업종별 텍스트 업황 vs 코스피 산업별지수 간의 시차 상관계수[1,2]

(원 계 열)

시차	전자/영상/통신장비	화학물질/제품	금융	의료물질/의약품	자동차+조선/기타운수	식료품	전기/가스/증기	건설업	기타기계/장비
-2	0.779	0.813	0.498	0.243	0.328	0.608	0.391	0.621	0.678
-1	0.843	0.908	0.594	0.349	0.644	0.542	0.506	0.890	0.729
0	0.736	0.810	0.569	0.585	0.714	0.242	0.250	0.883	0.648
1	0.490	0.456	0.401	0.621	0.428	-0.415	-0.319	0.633	0.412

(1 차 차 분 계 열)

시차	전자/영상/통신장비	화학물질/제품	금융	의료물질/의약품	자동차+조선/기타운수	식료품	전기/가스/증기	건설업	기타기계/장비
-2	0.448	0.481	0.458	0.168	0.208	0.120	-0.041	0.56	0.021
-1	0.345	0.483	0.367	-0.241	0.377	0.227	0.514	0.748	0.289
0	0.319	0.585	0.496	0.326	0.536	0.374	0.625	0.525	0.285
1	-0.044	0.099	0.294	0.053	0.022	-0.641	-0.461	0.568	-0.022

주: 1) 음(-)의 시차(분기)는 증권사 리포트 업황이 코스피 산업별 지수에 선행함을 의미
2) 빨간색은 고려한 시차(-2~+1분기) 중 시차상관계수가 가장 높음을 의미

Ⅳ. 텍스트 정보의 타당성 검증

이번 장에서는 제 Ⅳ장에서 제시한 텍스트 지표들의 타당성을 다양한 방식으로 검증한다. 업황 지수 $TBCI_{i,t}$는 주가지수, GDP, BSI 등 금융·경제 지표와의 비교를 통해 선행성과 예측력 등을 평가하였고, 기업경영환경 변화요인과 주요 이벤트에 대한 업종별 영향도, 평가 등은 과거 이슈와의 정성적 비교를 통해 지표의 타당성을 검증하였다.

1. 산업별 업황 및 변화요인 분석

(업종별 업황 추정)

먼저 업종별 텍스트 업황($TBCI_{i,t}$)의 추정 결과를 살펴보자. 〈그림 7〉의 분석 결과는 텍스트 업황이 관련 산업 코스피 지수의 흐름을 잘 추정하고 있음을 보여준다. 〈표 3〉에서 텍스트 업황과 코스피 산업별 지수의 상관관계를 살펴보면 상관계수가 0.5~0.9 수준으로 매우 높게 나타나고, 일부 업종 및 구간에서 텍스트 지표가 다소 선행하는 모습을 보이는 것을 알 수 있다. 다만 이러한 상관성과 선행성은 업종별로 다소 차이가 크게 나타난다.

텍스트 업황이 숫자 정보는 제외하고 오로지 텍스트 정보만을 이용하여 추정된 것을 감안하면 〈그림 7〉과 〈표 3〉의 분석 결과는 놀라운 수준이다. 다만 1차 차분을 이용하여 트렌드를 제거한 뒤, 두 지표의 상관계수를 구해보면 〈표 3〉과 같이 상관계수가 다소 낮아지는 것으로 나타난다. 이는 증권사 리포트에 나타나는 텍스트 정보가 단기 변동보다는 트렌드 파악에 보다 유용할 수 있음을 시사한다.

(전산업 업황 추정)

산업 구분 없이, 전체 유효 문장을 기준으로 전산업 업황을 추정한 뒤, 코스피 및 거시지표와의 상관관계를 살펴보았다. 〈표 4〉의

〈표 4〉 거시 정량지표와의 시차 상관계수[1,2]

(전산업 텍스트 업황)

시차	경기선행지수 순환변동치[3]	GDP 실질SA 전기비	GDP 실질원계열 동기비	BSI 전산업 업황실적[3]	BSI 전산업 매출실적[3]	BSI 전산업 업황전망[3]	BSI 전산업 매출전망[3]	코스피[3]
−2	0.673	0.043	0.709	0.735	0.817	0.741	0.794	0.743
−1	0.910	0.195	0.675	0.844	0.790	0.792	0.740	0.908
0	0.898	0.627	0.444	0.733	0.564	0.557	0.428	0.916
1	0.661	0.290	−0.043	0.268	0.122	0.046	−0.046	0.698

(코스피 지수)

시차	경기선행지수 순환변동치[3]	GDP 실질SA 전기비	GDP 실질원계열 동기비	BSI 전산업 업황실적[3]	BSI 전산업 매출실적[3]	BSI 전산업 업황전망[3]	BSI 전산업 매출전망[3]	코스피[3]
−2	0.531	0.131	0.648	0.760	0.840	0.787	0.842	0.618
−1	0.845	0.164	0.739	0.860	0.860	0.835	0.818	0.882
0	0.976	0.394	0.614	0.803	0.707	0.686	0.605	1.000
1	0.882	0.362	0.252	0.539	0.396	0.345	0.247	0.882

(당분기 코스피 시장 컨센서스[4])

시차	경기선행지수 순환변동치[3]	GDP 실질SA 전기비	GDP 실질원계열 동기비	BSI 전산업 업황실적[3]	BSI 전산업 매출실적[3]	BSI 전산업 업황전망[3]	BSI 전산업 매출전망[3]	코스피[3]
−2	0.473	0.122	0.575	0.687	0.783	0.733	0.803	0.558
−1	0.792	0.067	0.703	0.825	0.846	0.806	0.808	0.818
0	0.950	0.394	0.666	0.815	0.730	0.732	0.651	0.983
1	0.894	0.357	0.311	0.601	0.456	0.403	0.306	0.901

(1분기 앞 코스피 전망 시장 컨센서스[5])

시차	경기선행지수 순환변동치[3]	GDP 실질SA 전기비	GDP 실질원계열 동기비	BSI 전산업 업황실적[3]	BSI 전산업 매출실적[3]	BSI 전산업 업황전망[3]	BSI 전산업 매출전망[3]	코스피[3]
−2	0.205	0.072	0.393	0.559	0.681	0.606	0.705	0.274
−1	0.515	0.130	0.540	0.683	0.772	0.713	0.781	0.584
0	0.786	0.077	0.655	0.798	0.819	0.769	0.772	0.818
1	0.945	0.335	0.663	0.805	0.748	0.738	0.683	0.974

주: 1) 음(−)의 시차(분기)는 텍스트 업황, 코스피 지수, 코스피 컨센서스 등 고려한 지표가 거시 정량지표에 선행함을 의미
2) 빨간색은 고려한 시차(−2~+1분기) 중 시차상관계수가 가장 높음을 의미
3) 분기평균 기준
4) 각 분기 약 2개월이 지난 시점에서 추정한 당분기 코스피 컨센서스
5) 전망 대상시점과 거시지표 발표시점을 기준으로 비교

분석 결과는 텍스트 업황($TBCI_{i,t}$)이 GDP(실질원계열 전년동기비), BSI 전산업 업황·매출 실적 및 전망 등에 뚜렷한 선행성을 보이고 있음을 보여준다.

〈표 4〉와 같이 텍스트에서 추출한 업황 대신 코스피 지수를 이용하여 동일한 비교를 해보면, 텍스트 업황이 코스피보다 거시 지표에 대한 선행성을 잘 보여주고 있음을 확인할 수 있다. 이러한 결과는 당분기 또는 1분기 앞 코스피 컨센서스 전망치와 비교해도 마찬가지이다. 〈표 5〉는 텍스트 업황과 경기선행지수 순환변동치 사이의 인과관계를 Granger 검정을 통해 분석한 것이다. 분석 결과 원계열에서 코스피나 코스피 컨센서스 전망치에는 보이지 않는 경기선행지수로의 일방향적 인과관계가 텍스트 업황에서 나타나고 있음을 확인할

〈표 5〉 경기선행지수 대비 Granger 검정 결과

비교 지표			원계열 p-value	차분계열 p-value
전산업 텍스트 업황	→	선행지수 순환변동치	0.001	0.026
	←		0.653	0.314
코스피	→		0.425	0.015
	←		0.416	0.825
코스피 당분기 시장 컨센서스	→		0.886	0.298
	←		0.004	0.041
코스피 1분기 앞 시장 컨센서스	→		0.050	0.855
	←		0.000	0.002

주: 1) 빨간색은 유의수준 α=0.05 하에서 텍스트 업황 등 고려한 지표로부터 경기선행지수 순환변동치로의 일방향적인 인과관계가 존재함을 의미

수 있다. 차분계열에서도 코스피 컨센서스 전망치에는 나타나지 않는 일방향적 인과관계가 텍스트 업황과 코스피에는 나타난다. 이러한 결과는 애널리스트들이 제공하는 텍스트 정보에 숫자가 전달하는 정보량 이상의 정보가 반영되고 있음을 시사한다.

텍스트 업황의 변동성이 코스피 산업전망에 유의미한 추가적 정보를 제공하는지 살펴보기 위해 통계 모형을 이용하여 〈표 6〉과 같이 텍스트 업황의 예측력을 평가해보았다. 이를 위해 월평균 코스피 산업별지수와 텍스트 업황 지수를 동적인자모형(Dynamic Factor Model, DFM)으로 구축하고 텍스트 업황 지수 반영시 예측 오차가 줄어드는지 점검하였다. 공정한 평가를 위해 19.1월부터 22.6월까지의 월자료를 이용하여 모형을 학습시킨 뒤 22.7월부터 9월에 대해 예측오차를 평가하는 표본외 성능검증(out of sample testing)을 수행하였다.

분석 결과는 〈표 6〉과 같이 텍스트 업황이 각 코스피 산업전망 예측치를 크게 개선하지는 못하는 것으로 나타난다. 그러나 전기전자 등 일부 업종에서 예측 오차가 하락한 점을 고려하면, 애널리스트가 제시하는 텍스트가 시장 변동성이 설명하지 못하는 정보를 일부 제공했을 가능성이 있으며, 이에 대해서는 향후 보다 면밀한 분석이 필요할 것으로 사료된다.

(업종별 기업경영환경 변화 요인 추정)

업종별 경영환경 변화요인($BC-factors_{i,t}^{5}$)의 추정 결과를 과거 이슈와 비교하여 정성적으로 검증해보았다. 일부 업종의 결과를 살펴보면, 〈표 7〉과 같이 알고리즘으로 추출한 텍스트 정보가 경영환경 변화요인 파악에 매우 효과적인 것을 알 수 있다.

- 전자/영상/통신장비를 살펴보면 20.4분기 이전에는 스마트폰이 주요 이슈를 차지했으나 동 분기 이후 전기차 등 자동차가 주요 이슈로 부상한 것을 확인할 수 있다. 또한 반도체 공급부족 이슈가 21.2분기에서 22.1분기 중 지속됐음을 알 수 있다. 마찬가지로 패널 가격이 19년 상반기 하락을 지속하다가 20.1분기 들어 상승 전환한 후 21.4분기부터 다시 하락하여 디스플레이

〈표 6〉 텍스트 업황 반영에 따른 DFM 예측오차 비교[1)]

	코스피	음식료품	화학	의약품	기계	전기전자	의료정밀	운수장비	유통업	통신업	금융업
코스피 산업지수	2.9	0.4	7.9	6.5	0.3	2.0	3.7	2.7	4.2	0.4	9.5
코스피 산업지수+텍스트 업황	2.8	0.3	7.8	6.8	0.2	1.6	2.8	3.6	4.5	1.1	10.3

주: 1) 22.7~9월 코스피 및 코스피 산업별 지수 예측치의 월평균 절대 백분율 오차(mean absolute percentage error, MAPE)(%)

〈표 7〉 예시 업종의 기업경영환경 변화 요인표[1,2]

(전자/영상/통신장비등)

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	4457 77 / 31	5276 88 / 32	5329 102 / 31	6576 103 / 31	5659 99 / 30	6276 96 / 31	5731 107 / 32	7010 116 / 33	6275 117 / 33	5041 111 / 34	5751 115 / 31	5588 127 / 30	6220 121 / 34	6572 134 / 38	5019 109 / 37	5881 120 / 35
1	미중무역	미중무역	미중무역	5G스마트폰	5G스마트폰	스마트폰수요	위성통신안테나	5G스마트폰	패널가격상승	반도체공급부족	폴더블스마트폰	패널가격하락	전기자동차	스마트폰수요	스마트폰수요	적외선영상센서
2	스마트폰수요	중국스마트폰	저가스마트폰	인쇄회로기판	전기차시장	폴더블스마트폰	5G스마트폰	스마트폰출시	스마트폰카메라	패널가격상승	인쇄회로기판	반도체공급부족	전기차시장	완성차업체	고객스마트폰	완성차업체
3	중국스마트폰	스마트폰업체	일본수출규제	미중무역	폴더블스마트폰	스마트폰출시	스마트폰출시	반도체후공정	전기차시장	자동차전장	패널가격상승	폴더블스마트폰	반도체공급부족	자동차전지	완성차업체	폴더블스마트폰
4	패널가격하락	패널가격하락	글로벌스마트폰	휴대카메라모듈	저가스마트폰	글로벌스마트폰	스마트폰수요	스마트폰수요	전장부품사업	완성차업체	소연료전지	위성통신안테나	완성차업체	스마트폰카메라	글로벌경기침체	재고조정영향
5	프리미엄스마트폰	저가스마트폰	리튬이온전지	폴더블스마트폰	패널가격상승	신모델출시	미중무역	신모델출시	자동차부품	부품공급부족	제품믹스개선	부품공급부족	자동차부품	원자재가격상승	원자재가격상승	전방수요부진

(도매/소매)

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	2472 36 / 29	2198 37 / 28	2197 36 / 28	2339 42 / 28	2648 36 / 26	2414 34 / 30	2690 42 / 28	2837 41 / 30	2830 45 / 31	2762 46 / 29	2346 50 / 29	2549 45 / 28	2494 51 / 33	2553 53 / 33	2249 41 / 31	2169 46 / 30
1	줄기세포치료	식자재유통	식자재유통	온라인채널	식자재유통	닭가슴살	긴급재난지원	전기자전거	태양광발전	온라인채널	거리두기단계	위드코로나전환	중고차시장	거리두기해제	거리두기해제	물가상승
2	식자재유통	최저임금인상	중국전자상거래법	탈독관리총국	코로나바이러스	근거리쇼핑	점포구조조정	식자재유통	거리두기강화	점포구조조정	거리두기강화	원자재가격상승	식자재유통	식자재유통	물가상승	온라인사업
3	최저임금인상	온라인채널	최저임금인상	대구경북지역	자재유통시장	긴급재난지원	주택거래량	구조조정효과	식자재유통	온라인쇼핑	온라인채널	시장상황변화	물가상승	물가상승	가격원화	해외여행수요
4	추석시점차이	미세먼지영향	미중무역	국내건설경기	자사주소각	거리두기완화	재난지원사용	전동킥보드	닭가슴살	구조조정효과	거리두기완화	소비심리회복	완성차업체	원자재가격상승	원화약세	여행수요회복
5	중국전자상거래법	온라인유통	자회사실적	건설자재유통	자재유통사업	온라인매출	으뜸효율가전	점포구조조정	경북지역중심	백신접종상승	시장상황변화	백신접종상승	중고차판매	거리두기완화	온라인사업	사회지배구조

(정보통신업)

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	3845 48 / 32	3992 61 / 33	4753 55 / 31	4533 65 / 29	4749 58 / 29	5205 69 / 30	4923 69 / 30	5468 83 / 32	5261 82 / 33	4728 80 / 30	4923 86 / 33	5542 102 / 33	6405 95 / 32	5195 105 / 36	4506 77 / 34	5248 97 / 31
1	마케팅비용증가	마케팅비용증가	의료정보시스템	마케팅비용증가	온라인쇼핑	온라인쇼핑	온라인쇼핑	의료정보시스템	웹툰웹소설	자사주소각	망연계솔루션	빅데이터분석	온라인게임	온라인게임	공급망관리	의료정보시스템
2	검사막모바일	감가상각증가	온라인게임	5G가입증가	검사막모바일	이동전화매출	5G가입증가	온라인게임	취약점진단	중간지주사	인공지능빅	온라인쇼핑	모바일게임시장	공간정보플랫폼	게임매출감소	데이터센터화재
3	온라인게임	5G가입증가	5G가입증가	온라인교육	5G가입증가	마케팅비용증가	온라인교육	중단사업손익	중간지주사	지배구조개편	지능빅데이터	전자지급결제	의료정보시스템	검사막모바일	디어콘텐츠사업	게임매출감소
4	사막온라인	선택약정할인	마케팅비용증가	이동전화매출	자율주행자동	5G가입증가	인공지능기술	지배구조개편	5G가입증가	자사주매입	기존게임매출	마이데이터사업	검사막모바일	온라인쇼핑	시장성장둔화	웹툰웹소설
5	홈쇼핑송출	검사막모바일	선택약정요금	온라인쇼핑	데이터법통과	빅데이터분석	카메라모듈검사	온라인쇼핑	기업가치상승	기업가치상승	검사막모바일	유무선통신스	튜디오연결편집	정보보호산업	이동전화매출	기존게임매출

(금융업)

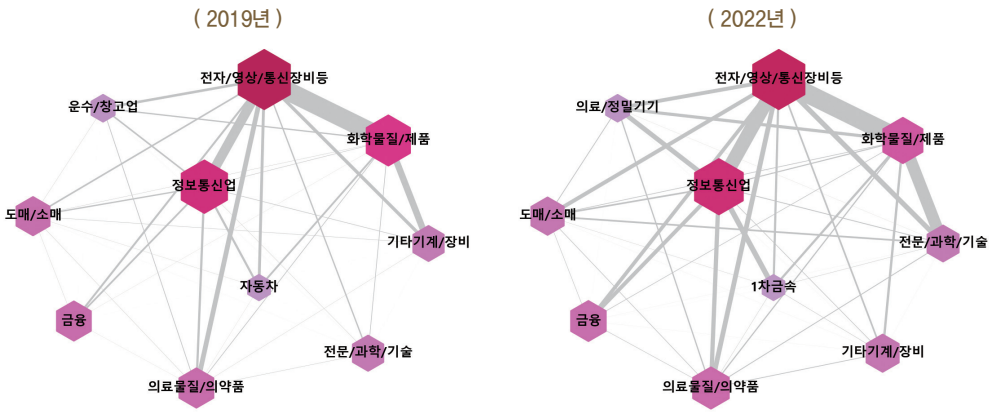
분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	2236 47 / 24	2815 53 / 25	2299 46 / 25	2406 50 / 25	1953 47 / 23	3201 53 / 25	2487 51 / 27	2819 60 / 25	2077 51 / 23	2732 59 / 25	2559 59 / 28	2368 62 / 26	2178 59 / 24	3363 72 / 28	2666 58 / 27	2972 67 / 26
1	주식시장하락	신종자본증권	시장금리하락	시장금리하락	부동산규제	시장변동확대	거래대금증가	책임준비전입	관련추가충당	시장금리상승	인터넷전문은행	기준금리인상	거래대금감소	기준금리인상	공동재보험	이자이익증가
2	거래대금감소	염가매수차익	기준금리인하	안심전환대출	기준금리인하	기준금리인하	장기위험손해	거래대금증가	책임준비전입	내부등급법승인	기준금리인상	이자이익증가	기준금리인상	주택담보대출	추가충당적립	조달비용상승
3	장기위험손해	자본비율하락	파생결합증권	운용자산이익	보험손해상승	금융시장변동	보험영업이익	기준금리인하	장기위험손해	거래대금증가	거래대금감소	거래대금감소	이자이익증가	거래대금감소	기준금리인상	공동재보험
4	차보험손해	인터넷전문은행	지주회사체제	파생결합증권	시장금리하락	퇴직연금사업	손해사업비율	신용공여잔고	시장금리상승	이자이익증가	중금리대출	중금리대출	희망퇴직비용	가계대출규제	거래대금감소	조달금리상승
5	하회위험요소	가맹수수인하	증시변동확대	책임준비전입	금융상품판매	부동산규제	기준금리인하	증시추가상승	선제충당적립	조달금리하락	이자이익증가	주택담보대출	기술사업금융	시장변동확대	이자이익증가	부동산시장

주: 1) N: 해당 분기 및 업종에 나타나는 총 문장 수, C: 기업 수, A: 보고서 작성기관 수　　2) 키워드는 순서(1~5)는 많이 언급된 상위 5개 키워드의 내림차순 순서를 의미

제조업 업황에 큰 영향을 미쳤음을 유추할 수 있는데, 이는 국제 패널가격 추이와 일치한다.

- 도매/소매의 경우에는 2019년도에 최저임금 인상이, 그리고 2020년 이후부터 최근까지는 코로나 사태가 주요 기업경영환경 변화요인으로 언급되고 있음을 확인할

〈그림 8〉 공통요인 기반 산업 간 유사도[1,2)]



주: 1) 노드(node)의 크기는 해당 기간 중 증권사 리포트에 나타나는 업종별 문장 수에 비례하고 연결선(edge)의 굵기는 두 업종간 공통요인 분포의 KL divergence 역수에 비례

수 있다. 또한 코로나 사태 2분기 이후인 20.3분기부터 점포 구조조정 등이 도소매업 주요 이슈로 부상한 것을 경영환경 변화 요인표를 통해 유추해볼 수 있다.

- 정보통신업은 마케팅 비용 증가가 19.1분기에서 20.2분기 사이의 주요 이슈로 언급되었다. 또한 5G 도입에 따른 가입자 증가가 19.1분기 이후 주요하게 언급되었으나 21.2분기부터는 5G에 대한 언급이 줄어든 것을 확인할 수 있다. 정보통신업의 경우 20.4분기와 21.2분기 사이에 지배구조 개편 이슈가 있었음을 동 자료를 통해 유추할 수 있다.

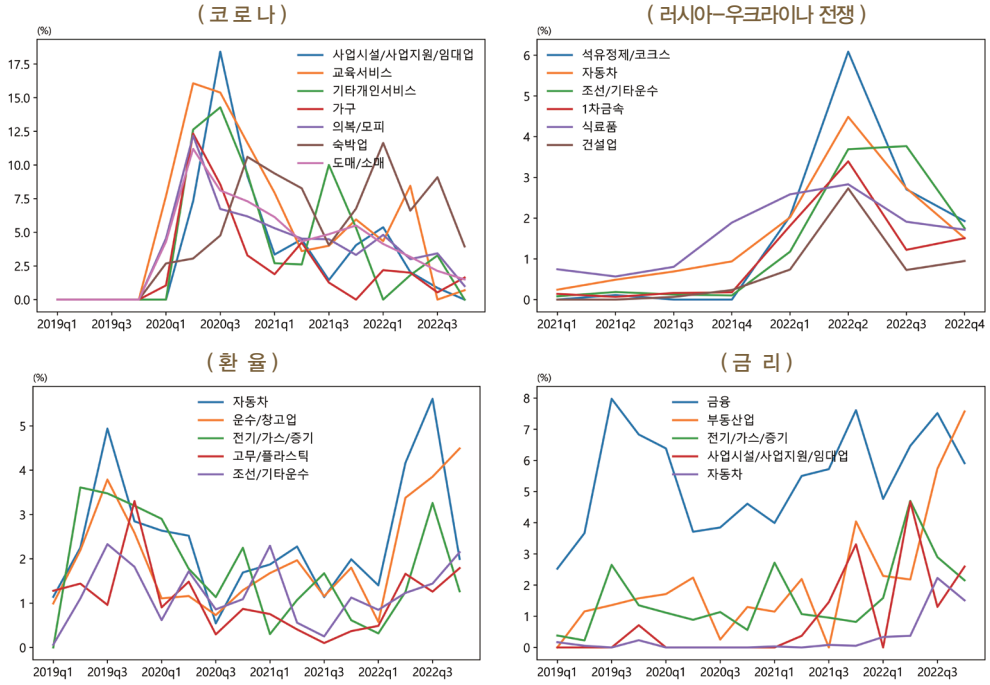
- 금융업은 20.2분기 코로나 발생 이후 거래대금 증가가 지속되다가 21.3분기부터 거래대금 감소가 주요 이슈로 언급되고 있음을 알 수 있고, 21.2분기 이후에는 금융업의 이자이익이 지속적으로 증가하고 있음을 함께 유추할 수 있다.

(공통요인 기반 산업 간 유사도 추정)

기업경영환경 변화요인을 바탕으로 추정한 업종별 유사도 지표를 살펴보았다. 추정 결과를 보면 업종간 유사도는 2019년 대비 2022년에 전자/영상/통신장비와 정보통신업 간의 연계성이 소폭 높아지고, 화학물질/제품과 전문/과학/기술 간 연계성이 크게 증가한 것으로 나타난다.

전자/영상/통신장비와 정보통신업 간 연계는 사회적 책임경영, 매출 성장세, 글로벌 경기 침체 등이 공통으로 언급된 것으로 보아, 두 업종에서 거시 경제적 요인의 영향도가 높아졌기 때문으로 보인다. 반면 화학물질/제품과 전문/과학/기술 간 연계는 연구개발비, 건강기능식품, 온라인쇼핑 등이 주로 언급된 것으로 보아, 두 업종의 사업 영역 간 유사도가 높아진 것 때문으로 사료된다.

〈그림 9〉 주요 경제 이슈의 업종별 영향도



주: 1) 증권사 리포트에 나타나는 총 문장 수 대비 특정 단어를 포함하는 문장의 상대빈도를 이용하여 추정한 지표로, 영향도가 높은 업종을 기준으로 정렬하여 각 업종의 관점에서 영향도가 높은 5개 업종을 시각화

2. 주요 이벤트의 산업별 영향 분석

(주요 이벤트의 업종별 영향도)

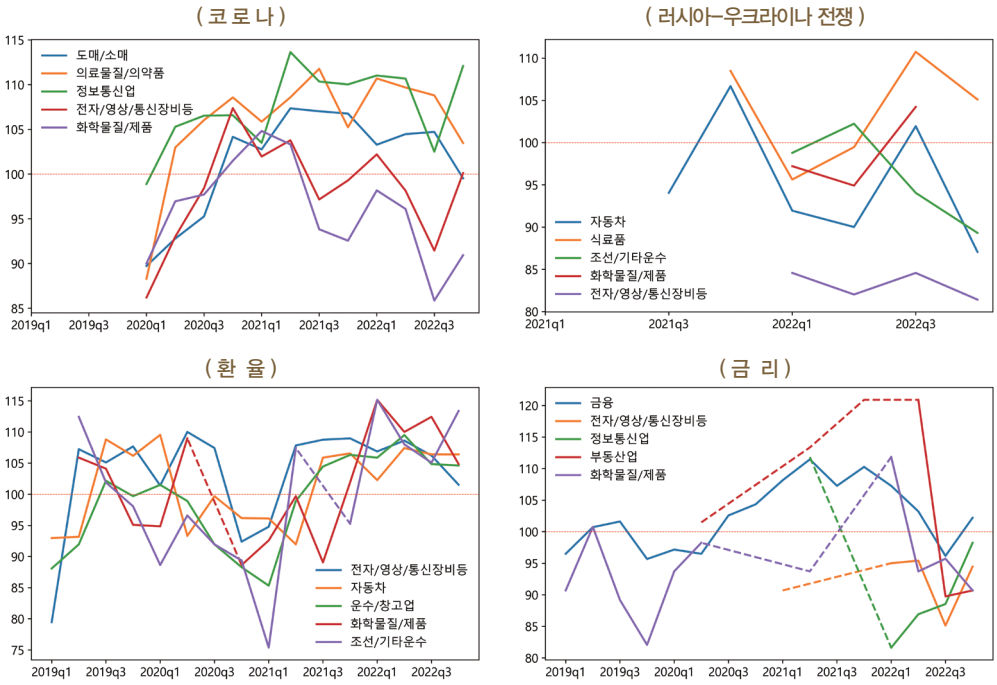
업종별 영향도($TIEI^{A}_{i,t}$)의 검증을 위해 코로나, 러우 전쟁, 환율, 금리를 기준으로 업종별 영향도를 추정한 뒤, 과거 이슈와의 비교를 통해 검증해보았다. 〈그림 9〉의 결과를 살펴보면 동 지표가 특정 이벤트에 대한 업종별 영향을 정량화하여 보여주는 데 매우 유용한 것을 알 수 있다.

- 코로나의 업종별 영향도를 살펴보면 20.2분기에 교육서비스가 가장 큰 영향을 받은 것으로 보이며 동 분기 이후 대부분의 업종에서 코로나 영향이 지속적으로 줄어든 반면 숙박업, 교육서비스 등은 거리두기 해제 이후인 22.1분기와 2분기에도 코로나가 여전히 주요 기업경영환경 변화요인으로 작용하고 있음을 유추할 수 있다.

- 러우 전쟁의 업종별 영향도를 보면 석유정제/코크스, 자동차, 조선/기타운수 순으로 그 영향이 크게 나타난 것으로 보이며, 특히 조선/기타운수는 여타 업종이 22.2분기 이후 영향도가 낮아지는 것과 달리 22.3분기까지 높은 영향이 지속됐음을 유추해 볼 수 있다.

- 환율의 업종별 영향도를 살펴보면 최근의 환율 흐름은 자동차, 운수/창고업, 전기/가스/증기 순으로 높은 영향을 미친 것으로 보이며 특히 운수/창고업의 경우 22.1분기 이후 그 영향이 지속적으로 커지고 있음을 유추할 수 있다.

〈그림 10〉 주요 경제 이슈에 대한 업종별 평가



주: 1) 특정 단어를 포함하는 긍 · 부정 문장 수의 차이를 그 합계로 나누어 산출한 지표로, 긍 · 부정 문장 수를 기준으로 업종을 정렬하여 시장의 관심이 높은 5개 업종을 시각화
2) 점선은 결측구간을 선형보간(linear interpolation)했음을 의미하며 선이 없는 구간은 자료가 없음을 의미
3) 지표가 100 보다 크면 해당 이벤트가 긍정적 100보다 작으면 부정적임을 의미

• 금리의 업종별 영향도를 보면 금융의 경우 기준금리 변동 시기마다 금리의 영향이 높게 나타나는 것을 확인할 수 있고, 최근의 금리 상승 구간을 살펴보면 부동산업과 자동차에서 금리의 영향도가 높아진 것을 유추할 수 있다.

(주요 이벤트에 대한 업종별 평가)

앞에서와 같이 코로나, 러우 전쟁, 환율, 금리를 기준으로 특정 이벤트에 대한 업종별 평가($TEEI^{A}_{i,t}$)를 추정하고 이를 정성적으로 검증해보았다. 〈그림 10〉의 결과를 살펴보면 동 지표가 특정 이벤트에 대한 시장의 평가를 정량화하여 보여주는 데 매우 유용한 것을 알 수 있다. 특히 텍스트 지표는 기초 자료가 문장으로 이루어지는 만큼 각 평가와 관련한 기초 문장을 확인하면 구체적 평가 내용도 알 수 있어 경제분석이 매우 용이해짐을 알 수 있다.

• 코로나의 업종별 평가를 살펴보면 의료물질/의약품의 경우 위생제품 생산의 영향으로, 정보통신업의 경우 비대면 서비스 증가의 영향으로 코로나가 기업경영 호재로 작용한 반면, 도매/소매는 20.4분기 이후에나 코로나로 인한 경영환경 악화에서 회복한 것으로 보이고, 전자/영상/통신장비와 화학물질/제품은 22.3분기 코로나로 인한 중국봉쇄조치의 부정적 영향이 컸던 것으로 나타난다.

• 러우 전쟁의 업종별 평가를 살펴보면 대부분의 업종에서 22.2분기 이후 회복세를

보이다가 최근 들어 다시 부정적 평가가 커진 것을 알 수 있다. 특히 조선/기타운수는 22.2분기 이후 러우 전쟁의 부정적 평가가 지속적으로 확대되고 있는 것으로 나타난다.

- 환율의 업종별 평가를 살펴보면 자동차의 경우 19.1분기 신흥국 통화 약세가 업황에 부정적 평가를 미친 것으로 나타나고, 조선/기타운수는 21.1분기 원달러 환율 하락으로 인한 재고자산 평가손실이 크게 반영된 것을 알 수 있다. 22.2분기 이후 최근의 환율 흐름은 대부분의 업종에서 긍정적 요인으로 평가받고 있는 것으로 나타난다.
- 금리의 업종별 평가를 살펴보면 금융의 경우 20.2분기 이후 낮은 금리 수준이 긍정적 요인으로 작용해 왔으나 최근 금리상승과 함께 금리가 부정적 요인으로 전환된 것을 알 수 있다. 또한 부동산업의 경우 22.2분기 이후 금리에 대한 평가가 급격하게 부정적으로 변하였으며, 테크 기업 위주의 정보통신업은 이에 앞서 21.2분기에서 22.1분기 사이에 금리에 대한 평가가 부정적으로 바뀌었음을 유추할 수 있다. 다만 22.4분기에는 대부분의 업종에서 금리에 대한 부정적 평가가 다소 완화된 것으로 나타난다.

Ⅵ. 시사점

본 연구는 증권사 기업평가 보고서의 텍스트 분석을 통해 경제 분석에 유용한 새로운 모니터링 지표들을 제시하고 과거 데이터를 바탕으로 검증하였다.

증권사 애널리스트 리포트는 특정 산업을 오래 연구한 기업분석 전문가들이 작성하므로, 동 보고서에 나타나는 텍스트 정보를 종합하여 가공 · 활용하는 방안을 지속적으로 연구할 필요가 있다.

증권사 리포트에 나타나는 텍스트 데이터는 가공하는 방식에 따라 다양한 미시적 · 거시적 연구가 가능하며 산업 관련 동향 파악 및 요인 분석에 효과적으로 활용될 수 있다. 특히 텍스트 데이터는 발간일 기준으로 취합이 가능하므로, 여타 공식 통계보다 신속하게 분석할 수 있는 장점이 있다. 또한 텍스트 데이터는 수치화하기 힘든 여러 주제에 대한 전문가들의 견해를 취합하여 보여준다. 이러한 점에서 텍스트를 바탕으로 작성한 지표들은 숫자가 반영하지 못하는 새로운 정보를 정량화하여 제공할 여지가 크다.

따라서 향후 텍스트 업황 지표를 산업 전망 모형에 외생변수로 추가하여 활용하는 방안이나, 산업간 유사도 지표를 바탕으로 경제분석 모형을 개발하는 방안 등 다양한 연구를 지속적으로 검토할 필요가 있다.

구체적으로 본 연구는 다음과 같은 점에서 효용성을 갖는다. 먼저 월평균 2,000개에 달하는 애널리스트 보고서를 사람이 모두 읽고 이해하는 것이 매우 어렵다는 점에서, 본 연구에서 제시한 방법론은 서베이 등 조사방법론을 이용하지 않고 전문가들의 생각을 취합하는 방식으로 그 효용성이 매우 높다. 또한 본 연구에서 제시한 방법론은 미국 IB 리포트 등 여러 보고서에도 동일하게 적용이 가능하며 알고리즘을 통한 자동화가 가능하다는 점에서 기업분석 등 경제 분석 업무의 효율을 제고하는 데

크게 기여할 수 있다.

다만, 본 논고에서 제시한 텍스트 분석 방법론은 다음 사항에 유의하여 사용해야 한다. 먼저 텍스트 지표는 구체적 사안에 대한 정밀한 논조 파악이나 고차원적 경제 분석에는 적합하지 않을 수 있다. 텍스트 분석은 텍스트 데이터의 특성상 노이즈가 많이 포함되고 저자의 선입견 등을 반영하게 된다. 따라서 텍스트 정보는 공식 통계와 일치하지 않을 수 있으며, 근거가 불명확한 정보도 포함될 수 있다. 그러나 이러한 정보가 애널리스트들이 생각하는 경제 인식을 그대로 보여주는 것이라면, 동 정보는 오히려 경기 판단을 위한 유용한 심리 지표로써 공식 통계와 함께 보조적으로 활용될 수 있다.

향후 텍스트를 이용한 보다 깊이 있는 경제 분석을 위해서는 텍스트에 나타나는 정보를 경제이론 등 배경지식과 연결하여 분석할 필요가 있다. 이러한 분석을 위해서는 GPT(Generative Pre-trained Transformer) 등과 같은 거대 통계 모형의 구축이 필요할 것으로 사료된다. 경제 분석을 위한 GPT 등의 개발은 아직 요원하지만, 최근의 기술발전 속도를 생각하면 매우 빠른 시일 내에 사람의 개입 없이도 스스로 정보를 취합하고 비교적 복잡한 경제 판단을 내릴 수 있는 통계 모형의 실현이 가능할 것으로 생각된다.

기술 발전에 따른 자동화와 효율성 제고는 거부할 수 없는 시대적 흐름이다. 경제 분야에서도 연구 효율이 개선될 수 있도록 AI 등 통계 기법 개발을 크게 장려해 나갈 필요가 있다.

〈부록 1〉

지역별 산업 업황 및 변동요인 분석

증권사 리포트에 나타나는 2,283개 기업을 본사 소재지 기준으로 분류하면 업종별 업황($TBCI_{g,i,t}$) 및 기업경영환경 변화요인($BC-factors_{g,i,t}$)을 지역별로도 동일하게 추정해 볼 수 있다.

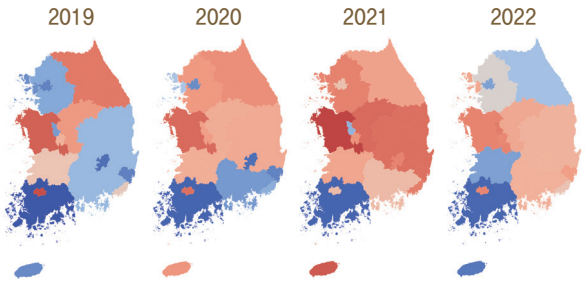
$TBCI_{g,.,t}$, $BC-factors_{g,.,t}$는 지역 g에 소재한 모든 업종의 t기 문장 샘플 중복집합 $S_{g,.,t}$와 Trigram 샘플 중복집합 $U_{g,.,t}$에 대해서 업황 및 변동요인 산출식을 적용하여 추정하였다. 이렇게 추정한 지역별 기업 업황 분포는 기업경영분석 통계의 지역별 총자산증가율 분포와 가장 높은 상관관계를 보인다. 지역별 업황을 시점별로 추정하여 시계열 지표로 살펴보면 전라남도의 경우 한국전력 업황이 지속적으로 낮게 평가됨에 따라 여타 지역에 비해 낮게 나타나며 충청남도의 경우 덕산네오룩스, 이녹스첨단소재 등 전기전자 부품 기업의 업황 호조로 업황이 높게 나타나는 것을 확인할 수 있다. 지역별 업황($TBCI_{g,.,t}$)을 BSI 매출실적과 비교해보면 대체로 상관계수가 0.5~0.9로 높게 나타난다.

〈표 A1〉 공식 통계와의 상관관계[1)]

공식 지역별 통계	상관계수
총자산증가율	0.68
기업순이익률	0.44
총자산순이익률	0.41
부가가치율	0.35
매출액순이익률	0.28

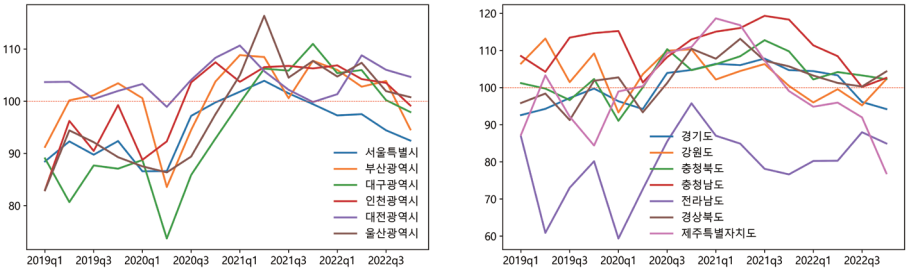
주: 1) 2020년 기준

〈그림 A1〉 증권사 리포트에 나타나는 지역별 업황 분포[1)]



주: 1) 붉은색은 업황이 긍정적, 파란색은 업황이 부정적임을 의미

〈그림 A2〉 지역별 기업 업황[1)]



주 : 1) 지표가 100 보다 크면 긍정적 100보다 작으면 부정적임을 의미

〈표 A2〉 지역별 텍스트 업황(TBCI) vs 기업경기실사지수(BSI) 매출실적 시차 상관계수[1,2),3)]

시차	부산	대구	인천	울산	경기	충북	전북	경북
−2	0.486	0.609	0.775	0.555	0.706	0.717	0.599	0.526
−1	0.551	0.814	0.734	0.725	0.471	0.642	0.305	0.602
0	0.445	0.893	0.588	0.766	0.145	0.409	0.021	0.671
1	0.000	0.764	0.391	0.582	−0.099	0.098	−0.108	0.243

주: 1) 음(−)의 시차는 증권사 리포트 업황이 코스피 산업별 지수에 선행함을 의미
2) 빨간색은 고려한 시차(−2~+1분기) 중 시차상관계수가 가장 높음을 의미 3) 분기평균 기준

기업경영환경 변화요인($BC-factors_{g,\ldots,t}$)을 지역별로 살펴보면 각 지역의 소재 기업에 따라 서로 다르게 나타나는 지역별 이슈를 파악하는 것도 가능하다.

- 대전광역시는 KT&G의 전자담배 사업과 바이오벤처기업의 신약 연구개발 이슈가 전기간에 걸쳐 나타나며 시간에 따른 변동이 적은 모습이다.
- 강원도는 20.2분기 이후 보툴리눔톡신 등 미용 의료 관련 이슈가 주로 언급되고 있으며, 예술/스포츠/여가 업종 비중이 높게 나타남에 따라 코로나 이슈도 많이 언급되고 있는 것을 확인할 수 있다.
- 제주도는 19.3분기와 20.1분기 사이 일본 불매운동으로 인한 일본노선 축소가 주요 경영환경 이슈로 언급된 것을 확인할 수 있고, 20.2분기 이후부터는 제주에 본사를 두고 있는 카카오의 경영환경 이슈가 주요하게 언급되고 있는 것을 알 수 있다.

〈표 A3〉 예시 지역의 기업경영환경 변화 요인표[1,2)]

(대전광역시)

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	662 14 / 27	680 16 / 27	703 13 / 26	829 18 / 25	987 20 / 25	882 17 / 28	964 20 / 26	1304 24 / 28	1049 24 / 26	839 26 / 28	839 19 / 30	1135 28 / 26	656 21 / 25	988 27 / 27	817 19 / 24	1226 26 / 22
1	암조기진단	궐련전자담배	자가면역질환	클린룸제어	코로나바이러스	단일클론항체	인조대리석	자동차센서	후보물질발굴	외관검사장비	체외진단시장	효소바이오소재	프리필드실린지	인조대리석	이족보행로봇	적외영상센서
2	궐련전자담배	전자담배판매	피하주사제형	궐련전자담배	카본블록필터	후보물질발굴	외관검사장비	신약연구개발	임상상진행	면역화학진단	반도체공급부족	디스플레이산업	환경오염제어	자동차센서	정밀지향마운트	암조기진단
3	미중무역	자가면역질환	열관리시스템	산업클린룸	영상보안장비	전자담배수출	환경접착소재	자가면역질환	유전검사서비스	국내임상완료	체외진단기기	화학오염모니터링	박물방출제형	대리석시장지	향마운트시장	보툴리반도체수출
4	인조대리석	바이오마커기반	면역질환치료	자가면역질환	크린텍카본블록	신생항원발굴	머신비전기술	후보물질발굴	유전빅데이터	동물백신제조	전자약전문	자가면역질환	인체적용시험	확대기판증설	자가면역질환	영상센서기술
5	미중무역전쟁	삼중음성유방암	위성영상판매	자율주행차	삼중음성유방암	세대형체의약	산업머신비전	외관검사장비	외관검사장비	열관리시스템	궐련전자담배	신약연구개발	자동차센서	분자진단키트	피부장벽기술	미중무역

(강 원 도)

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	365 7 / 17	158 8 / 22	204 5 / 16	449 10 / 17	192 4 / 16	569 8 / 21	241 6 / 16	476 8 / 15	309 9 / 16	356 12 / 21	219 7 / 20	299 10 / 21	421 11 / 17	245 9 / 20	224 8 / 20	310 12 / 16
1	환자감시장치	세무회계사무	수입고객제품	자동제세동기	신종코로나	건강기능식품	특신제제	환자감시장치	미용성형시장	정부지원사업	거리두기단계	국가출하승인	환자감시장치	체류인원제한	자가면역질환	자가면역질환
2	자동심장충격기	마이크로바이옴	세무사무제품	환자감시장치	세무회계사무	폐광지역개발	메디컬에스테틱	생체신호측정	저온냉동고	이해관계상생	바우처지원사업	해출채권팩토링	감시장치심장	영업시간단축	혈장단백질	면역질환불륙버스
3	오염제어기술	자사주매입	빅데이터	체외진단기기	최근주가하락	프로바이오틱스	에스테틱전문	항체진단키트	연구장비산업	거리두기단계	영업시간연장	전기화물차	장치심장충격기	코로나이전수준	거리두기해제	에이프릴바이오
4	통신단말이용	강원랜드기초자	프로바이오틱스	중국위생허가	제품세무회계	품목허가취소	디지털뉴딜정책	클린룸소모	가열자력교반	재생의학전문	거리두기완화	체권팩토링사업	스마트팩토리	거리두기해제	치료효능단백	면역원발생
5	기술관련산업	핀테크신규	테크핀모델	응급의료기기	세전이익폐광	히알루론산필러	필러특신	신호측정진단	성형시장적용	의학전문회사	화장주요유통	톨리놀톡신제재	생체신호측정	매출성장둔화	단백질결합	식약처행정처분

(제주특별자치도)

분기	2019q1	2019q2	2019q3	2019q4	2020q1	2020q2	2020q3	2020q4	2021q1	2021q2	2021q3	2021q4	2022q1	2022q2	2022q3	2022q4
N C / A	408 3 / 25	468 3 / 24	464 3 / 23	348 3 / 24	506 3 / 23	556 4 / 23	393 2 / 24	427 3 / 25	431 4 / 25	529 5 / 22	382 6 / 22	392 4 / 22	444 3 / 20	592 3 / 23	565 4 / 21	882 5 / 25
1	카풀서비스	웹툰웹소설	한일관계	일본노선감편	오픈뱅킹도입	중소형광고	카카오페이지	자회사상장카	지노복합리조트	자회사상장	웹툰카카오웹툰	채널비즈계정	플레이른블록체인	롯데관광개발	오픈채팅광고	데이터센터화재
2	신규광고모델	광고상품출시	주주적격심사	일본노선수요	중소형광고	금융상품판매	프렌프리미엄지면	페이모빌리티	카카오페이지	웹툰웹소설	다음웹툰카카오	상생비용부담	블록체인	카카오톡플랫폼	쏘카카세어링	카카오기초자산
3	카지노이전확정	신규광고상품	한일관계악화	카카오페이지	뱅킹수수료감소	교환배송선물	하반기게임즈	체크인카카오	롯데관광개발	가상화폐시장	플랫폼사업대하	톡채널비즈	주주환원정책	주요사업성장	카세어링사업	주주가치제고
4	지방발노선확대	카풀서비스	펌뱅킹수수료	원달러환율	펌뱅킹수수료	코로나확산이후	생활밀착서비스	시장금리하락	주요자회사	카카오페이스	토리뮤직미디어	게임매출	트래블버블체결	플랫폼기타매출	전기자전거	지급수수료증가
5	광고대행수수료	주식채권펀드	원달러환율	오픈뱅킹도입	일본불매운동	카카오기초자산	광고카카오페이	결제금융서비스	카카오페이지엠	카뱅금융사업	멜론유료가입	회사카카오게임	메타버스구축	카카오카카오톡	카카오오픈채팅	광모빌리티플랫폼

주: 1) N: 해당 분기 및 업종에 나타나는 총 문장 수, C: 기업 수, A: 보고서 작성기관 수 2) 키워드는 순서(1~5)는 많이 언급된 상위 5개 키워드의 내림차순 순서를 의미

〈부록 2〉 GPT 등 문장 생성 모형의 경제분석 활용 가능성

최근의 자연어처리 방법론은 언어분석에 매우 탁월한 성능을 보이는 인공신경망 모형을 기반으로 발전하고 있다. 인공신경망 모형은 조각별선형(piecewise linear) 함수의 구조를 갖는 비모수(non-parametric) 통계 모형으로, 선형모형이 추정하지 못하는 다양한 경우의 수를 다룰 수 있다(Seo et al., 2022 a). 따라서 자연어처리를 위해 인공신경망 모형을 이용할 경우, 모형의 파라미터 수를 크게 늘리면, 방대한 양의 문장 패턴을 모형으로 추정할 수 있다.

이때 인공신경망은 여러 구조로 모형을 구성할 수 있는데, 텍스트 분석을 위해서는 트랜스포머(transformer)의 구조가 널리 활용되고 있다. 트랜스포머는 번역을 위해 고안된 구조로 텍스트를 입력 받는 인코더(encoder)와, 이를 번역하여 다른 언어로 출력하는 디코더(decoder)로 이루어진다. 여기서 언어의 문맥을 파악하는 데 유용한 인코더의 구조만 분리하여 활용하는 모형이 BERT(Bidirectional Encoder Representations from Transformers), 언어를 생성해서 출력하는 데 유용한 디코더의 구조만 분리하여 활용하는 모형이 GPT(Generative Pre-trained Transformer)이다. 이들 거대 인공신경망 모형은 특정한 목적 없이 문장 패턴을 먼저 학습시키고(pre-train), 이후 다양한 목적에 따라 모형을 변형시킨 후, 목적에 맞는 데이터를 추가로 학습(fine-tuning)시켜 주로 이용한다.

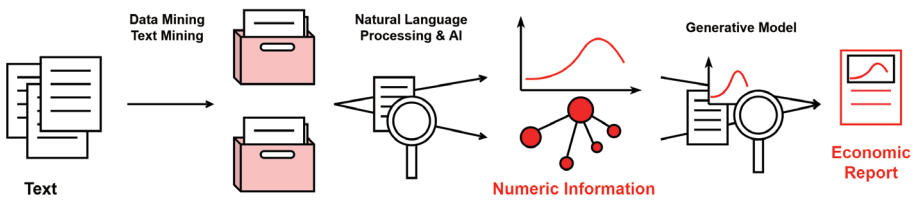
최근 문장 생성 모형(generative model)으로 인기를 끌고 있는 GPT 모형은 다양한 학습 문장의 패턴을 분석하여, 특정 주제가 주어지면 그와 비슷한 학습데이터의 문장 패턴을 찾아내서로 조합함으로써 새로운 문장을 만들어 낸다. 따라서 방대한 문장 패턴을 학습한 GPT 모형은 수려한 문장을 만들어 내는 데 탁월한 성능을 보이며, 이는 일반 사람의 문장력을 능가할 수 있다.

GPT 모형의 경제 분야 활용 가능성을 살펴보면, GPT 모형은 문장의 단어나 어구를 자주 쓰이는 표현으로 교정하거나, 특정한 목적의 수려한 문장을 만들어 내는 데 사람에 버금가는 성능을 보일 것으로 사료된다(Aldrick, 2023). 또한 통계 수치를 읽어주는 문장을 만들거나, 특정 경제 용어를 설명하는 간단한 문장을 만드는 데도 활용이 가능할 것으로 판단한다.

문장 생성 모형인 GPT 모형을 경제 분석에 직접적으로 활용할 수는 없다. 그러나 인터넷에 올라오는 경제 뉴스, 회계 보고서, 분석 보고서 등을 실시간으로 학습하도록 알고리즘을 구성하면, 특정 주제와 유사한 문장 패턴을 조합하는 방식으로, 경제 현안에 대한 설명문을 최신 정보를 바탕으로 만들어 내는 것이 기술적으로 가능하다. 본 연구에서 제시한 분석 알고리즘과 GPT 등의 문장 생성 모형을 연결하면, 웹 스크래핑(web-scraping)을 통한 문서의 입수부터, 업종별 업황 파악과 변동요인 분석, 그리고 분석 결과의 시각화에 더해, 분석 결과의 문서화도 가능할 것으로 판단한다.

상용화가 가능한 수준의 문장 생성 모형을 만드는 것은 어려운 일이나, 방대한 양의 경제 텍스트를 인터넷에서 손쉽게 수집할 수 있는 점을 고려하면, 간단한 경제적 표현을 생성하는 GPT 모형을 추정하는 일은 비교적 어렵지 않을 것으로 판단된다.

〈그림 B1〉 GPT 등 문장 생성 모형의 활용 예시



〈참고문헌〉

김도희 · 김민정 (2022). 텍스트마이닝을 활용한 핀테크 및 디지털 금융 서비스 트렌드 분석. 디지털융복합연구, 20(3), 131-143.

김수현 · 이영준 · 신진영 · 박기영 (2019). 경제분석을 위한 텍스트마이닝. BOK 경제연구, 2019(18).

박수빈 · 이용규 (2022). 텍스트 마이닝을 활용한 금융업 세부 업종간 ESG 보고서 비교 분석. 국가정책연구, 36(1), 31-56.

서범석 (2022 a). 뉴스 텍스트를 이용한 경기 예측: 경제 부문별 텍스트 지표의 작성과 활용. BOK 이슈노트, 2022(18)

서범석 · 이영환 · 조형배 (2022 b). 기계학습을 이용한 뉴스심리지수(NSI)의 작성과 활용. 국민계정리뷰, 2022(1), 68-90.

한승욱 · 김태완 · 이현창 (2022). 인공지능 언어모형을 이용한 인플레이션 어조지수 개발 및 시사점. BOK 이슈노트, 2022(38)

Ahmad, K. (2006). Multi-lingual sentiment analysis of financial news streams. PoS, 001.

Aldrick, P. (2023). ChatGPT will be the calculator for writing, top economist says. Bloomberg Article. 2023-1-19.

Bharti SK, Babu KS (2017) Automatic keyword extraction for text summarization: a survey. CoRR. abs/1704.03242.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4), 1593-1636.

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science, 2(3), 153-170.

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. Financial Innovation, 6(1), 1-25.

Hájek, P., & Olej, V. (2013, September). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In International conference on engineering applications of neural networks (pp. 1-10). Springer, Berlin, Heidelberg.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011).

Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, 50(3), 585-594.

Jaseena, K. U., & David, J. M. (2014). Issues, challenges, and solutions: big data mining. CS & IT-CSCP, 4(13), 131-140.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: economic forecasting using newspaper text. Journal of Applied Econometrics, 37(5), 896-919.

Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. Accounting and Business Research, 49(5), 587-615.

Lee, Y. J., Kim, S., & Park, K. Y. (2019). Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea. Bank of Korea WP, 11.

Li, N., Liang, X., Li, X., Wang, C., & Wu, D. D. (2009). Network environment and financial risk using machine learning and sentiment analysis. Human and Ecological Risk Assessment, 15(2), 227-252.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 57(5), 102212.

Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. International Journal of Forecasting, 32(2), 243-256.

Seo, B., Lin, L., & Li, J. (2022 a). Mixture of Linear Models Co-supervised by Deep Neural Networks. Journal of Computational and Graphical Statistics. 31(4).

Seo, B., Lee, Y. & Cho, H. (2022 b). Machine-learning-based news sentiment index (NSI) of Korea. Bank of Korea WP, 15.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. Journal of econometrics.

Siegele, L. (2022). How the tech behind ChatGPT could change the world – an updated episode from our archive. The Economist Article. 2022-12-27.

Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis, 48, 272-281.

Te Liew, W., Adhitya, A., & Srinivasan, R. (2014). Sustainability trends in the process industries: A text mining-based analysis. Computers in Industry, 65(3), 393-400.

Türegün, N. (2019). Text mining in financial information. Current analysis on economics & finance, 1, 18-26.

Weisenthal, J. (2022). This AI chatbot is a shockingly competent macro putdit – How sustainable is Japanese public dept? OpenGPT has the answer. Bloomberg Article. 2022-12-1.

Wu, J. L., Su, C. C., Yu, L. C., & Chang, P. C. (2012). Stock price predication using combinational features from sentimental analysis of stock news and technical analysis of trading information. International proceedings of economics development and research, 55, 39-43.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, 38(10), 13274-13283.