https://doi.org/10.5392/JKCA.2020.20.08.001

뉴스 데이터를 활용한 텍스트 감성분석에 따른 지역 산업생태계 위기 예측 - 광주 지역 자동차 산업을 중심으로 -

Crisis Prediction of Regional Industry Ecosystem based on Text Sentiment Analysis Using News Data - Focused on the Automobile Industry in Gwangju -

김현지[*], 김성진[**], 김한국[*]
과학기술연합대학원대학교 데이터 및 HPC 과학/한국과학기술정보연구원 기술사업화센터[*],
한국과학기술정보연구원 기술사업화센터[**]

**Hyun-Ji Kim(hyunjikim@kisti.re.kr)[*], Sung-Jin Kim(sungjin.kim@kisti.re.kr)[**],
Han-Gook Kim(hgkim712@kisti.re.kr)[*]**

요약

지역 산업생태계의 노후화 문제가 점차 심각해지면서, 지역 산업생태계의 쇠퇴를 측정하고 재생하기 위한 연구가 활발히 이루어지고 있다. 하지만 지역 산업생태계 위기 예측에 관한 연구는 거의 이루어지지 않고 있다. 위기는 단기간에 걸쳐 급진적으로 나타나는데, 사후대응으로는 역부족인 경우가 대다수이므로 위기가 발생하기 전에 대응해야 한다. 즉, 지역 산업생태계의 위기를 조기에 파악하여 선제적인 대응을 하는 것이 장기적인 관점으로 바라봤을 때 더욱 필요하고 요구된다는 것이다. 이에 본 연구는 대용량의 뉴스 데이터를 활용하여 뉴스의 감성 점수에 따른 지역 산업생태계의 위기 예측 가능성을 점검하였다. Google 감성분석 API를 사용하여 뉴스 감성 분석을 실행하였고 이를 월별로 정리하여 감성 분석 결과 실제 이벤트 간의 연관관계를 확인하였다.

■ 중심어 : | 텍스트마이닝 | 감성분석 | 자연어처리(NLP) | 지역산업생태계 |

Abstract

As the aging problem of the regional industry ecosystem has gradually become serious, research to measure and regenerate the regional industry ecosystem decline has been actively conducted. However, little research has been done on regional industry ecosystem crises. Crisis emerges radically over a short period of time, and it is often impossible to respond by post-response, so you must respond before the crisis occurs. In other words, it is more necessary and required when looking at the crisis early and taking a proactive response from a long-term perspective. Therefore, it is necessary to develop a predictive model that can proactively recognize and respond to the crisis in the regional industry ecosystem. Therefore, this study checked the possibility of predicting the risk of regional industry and market according to the emotional score of the news by using large-scale news data. News sentiment analysis was performed using the Google sentiment analysis API, and this was organized by month to check the correlation between actual events.

■ keyword : | Text mining | Sentiment Analysis | Natural Language Processing | Regional Industry Ecosystem |

* 본 연구는 한국과학기술정보연구원 '데이터 기반 기술사업화 지원체제 구축(K-20-L03-C05-S01)'과제의 지원으로 수행되었음
접수일자 : 2020년 07월 17일 심사완료일 : 2020년 08월 12일
수정일자 : 2020년 08월 12일 교신저자 : 김한국, e-mail : hgkim712@kisti.re.kr

I. 서 론

지역경제 발전은 국가 경제성장의 원동력으로서 꼭 필요하다. 지역경제의 위기로 인해 지역 간 발전 격차가 확대되면 이는 곧 국가 경제의 성장 침체로 이어지게 된다.

이 중 본 연구에서 주목한 광주광역시는 1966년 아시아 자동차공장 기공이 이루어지면서 국내 최초의 자동차생산도시로 출발하게 되었다[1]. 광주광역시는 기아자동차, 금호타이어 등 285개 자동차 및 부품 관련 업체가 자리 잡고 있는 한국 제2의 자동차 생산도시로 지역 경제의 심장 역할을 함으로써, 2016년 기준 완성차 62만대 생산, 매출 9조 5000억 원, 고용인력 7800명 등으로 지역 경제의 약 40%를 차지하고 있다[2]. 자동차 산업은 기계, 전기, 전자, 철강 등 연관 산업의 발달에 큰 영향을 주는 주력산업인만큼 지역 경제뿐만 아니라 국내 산업에서도 아주 중요한 역할을 한다[3].

하지만 우리나라 자동차 산업은 대한민국의 주력산업으로 국가발전을 주도해왔으나, 2018년에 생산량이 연간 400만 대에 밑돌면서 멕시코에 밀려 세계 7위로 전락하는 등 위기 상황이 지속되고 있어 광주처럼 자동차 산업을 기반으로 하는 도시는 지역 경제에 미치는 영향이 매우 크다. 또한, 자동차 대기업과 부품업체와의 극심한 임금 격차가 존재하는 한 부품업체의 독자적 발전은 힘든 상황으로[4], 임금을 기존 자동차 업계의 절반 수준으로 낮춰 기업의 경쟁력을 높이고 일자리를 늘려 지역경제를 살리기 위한 정책을 시도하고 있으나 여전히 위태로운 상황이다.

이런 경제적인 리스크를 더욱 최소화하기 위해서는 사후대응보다는 사전대응이 필요하며 조기예측을 통해 더욱 큰 위험에 대해 대비하여야 한다.

뉴스는 국내외에서 발생하는 사건들을 살피는 데에 가장 적합하다[5]. 산업 및 지역에 관한 많은 사건들의 흐름을 보고, 해당 산업이 위기에 처할 것을 미리 예측하여 큰 손실을 방지하는 데에 있어 뉴스 데이터는 좋은 정보원이 된다. 따라서 본 연구에서는 뉴스 데이터를 활용한 텍스트 마이닝을 통해 지역 산업생태계 위기를 예측하는 연구를 진행해보았다.

본 논문은 다음과 같이 구성된다. 2장에서는 텍스트 마이닝을 통한 감성분석과 산업 및 지역 생태계, 위기 예측 관련 연구를 중심으로 살펴본다. 3장에서는 대용량 뉴스 데이터의 수집, 전처리 과정, 감성분석, 경제지표 및 산업단지활동지표의 수집 및 분석방법에 관해 설명한다. 4장에서는 본 연구에서 대용량 뉴스 데이터의 감성정보 변화와 경제지표 및 산업단지활동지표를 비교하며 연관관계를 확인한 결과를 다룬다. 마지막으로 5장에서는 연구 내용을 요약하고, 본 연구의 결론과 한계점, 향후 연구 과제를 제시하며 마무리한다.

II. 이론적 고찰

1. 텍스트 마이닝을 통한 감성분석

텍스트 마이닝이란 비정형 텍스트 데이터에서 새롭고 유용한 정보를 찾아내는 기술이며, 비정형 데이터 예시로는 자연어처리(Natural Language Processing) 기술에 기반을 두고 데이터를 가공한다. 즉, 전처리를 통해 비정형 데이터에서 정형화된 데이터로 바꾸어 특징을 추출하는 과정을 뜻한다. 본 연구에서는 텍스트 마이닝을 통한 감성분석을 진행하였다.

감성분석(Sentiment analysis)이란 텍스트에 표현된 개체와 그 속성에 대한 의견, 감성, 평가, 태도 등을 분석하여 텍스트에 드러난 감성을 분류하는 것이다[6]. 감성분석은 크게 2가지로 나눌 수 있는데, 분석 데이터에 레이블(Label)이 있는 경우와 없는 경우에 따라 지도학습(Supervised learning)과 비지도학습(Unsupervised learning)으로 나눌 수 있다. 본 연구에서 사용된 텍스트 데이터는 뉴스 데이터는 레이블이 없는 경우이기 때문에 비지도 학습에 해당한다. 비지도학습을 통한 텍스트의 감성분석에는 자연어처리 방식이나 문장의 패턴을 이용하는 방법, 단어 간 상관관계 분석을 통해 문장의 극성을 구분하는 방법, 극성이 부여된 단어들도 구성된 감성사전을 이용해 분류하는 방법 등이 있다[7].

본 연구에서는 비지도학습을 통한 텍스트 감성분석을 활용하였고, 수많은 텍스트 데이터 중 뉴스기사를 사용하였다. 뉴스기사는 대중이 관심을 보이는 부분이나 대중에게 알려야 하는 혹은 대중이 알아야만 하는 정보를 정제하여 표현한 글이다[8]. 즉, 뉴스기사는 현

실에서 일어나는 각종 상황에 대한 설명과 앞으로 정치, 경제, 사회, 기업 등과 관련하여 미래에 어떤 변화가 발생하고 진행이 될 것인지에 대한 정제된 정보이기에 가장 신뢰할 수 있는 중요한 정보라 말할 수 있다[9].

2. 산업 및 지역 생태계 관련 연구

최근 몇 년 사이에 산업위기지역 문제는 대한민국의 중요한 사회적 이슈로 대두되어, 산업위기대응 특별지역, 고용위기지역 등의 형태로 지원정책이 수립 및 추진되고 있다. 이러한 맥락에서 이종호(2019)는 탈공업화, 주력산업의 쇠퇴, 산업구조조정 등으로 산업 및 고용의 위기에 직면한 지역들을 대상으로 지역산업정책 추진 경험이 풍부한 유럽 선진국들의 사례를 고찰하고, 정책적 시사점을 제시하였다[10].

전지혜(2019)는 구미지역을 둘러싼 산업 환경의 변화와 그런 환경 변화 속에서의 구미지역의 산업위기 실태를 분석하였다. 이때, 구미지역 산업위기 극복에 있어서 환경 변화에 대응·적응할 수 있는 회복력의 강화가 요구되는데, 이를 위해서는 기업과 지역차원에서 각각의 방안이 필요하다. 먼저 기업차원에서는 융·복합 기술에 기반을 둔 혁신역량을 향상하고 사업다각화의 실현이 이루어져야 하며, 지역차원에서는 기업과 산업의 자생적인 공진화를 통한 혁신 생태계 조성을 목표로 하는 산업위기대응특별지역 제도의 선정이 이루어져야 한다고 하였다[11].

이 밖에도 산업위기지역을 정량적으로 진단하는 척도로 정성훈(2019)의 연구에서는 산업현황(산업성장률, 생산액, 사업체 수, 수출입 추이 등), 기업현황(기업경기실사지수 등), 고용현황(취업률, 실업률 등), 투자현황(자본재 수입액, 기계류 수입액 등), 부동산 현황(주택매매가격지수, 전세가격지수 등), 소비현황(대형소매점 판매액 지수, 소매판매액 지수 등), 노동시장 현황(비정규직 비중, 임금 수준, 근로시간, 퇴직 등)을 보편적으로 활용하고 있다[12].

더 나아가 조성철(2019)은 정성적인 차원에서 정책환경 변화, 지역 자체 산업역량 부족, 공간분업으로 인한 지역 불균형 발전[13]과 표준화된 생산기능의 자동화나 해외이전, 대·중소기업 간의 수직적인 관계, 쇠퇴한 도시에서 나타나는 각종 위험성과 숙련 노동의 이탈로 진단했다[14].

위 선행연구에서 살펴보았듯이, 현재 산업 및 지역 생태계 위기 관련 연구에서 뉴스 데이터를 활용하여 산업위기지역을 진단한 사례를 찾아보기 어렵다는 것을 알 수 있고, 새로운 관점으로 해당 연구를 진행했다는 점에서 시사하는 바가 크며, 앞으로 대용량 뉴스 데이터를 활용한 산업 및 지역 생태계 위기에 관한 연구에 기초로 활용될 것이다.

3. 뉴스 데이터를 활용한 예측 관련 연구

인터넷 기술의 발전과 인터넷상 데이터의 급속한 증가로 인해 데이터의 활용 목적에 적합한 분석방안 연구들이 활발히 진행되고 있다. 최근에는 텍스트 마이닝을 활용한 연구들이 이루어지고 있는데, 특히 문서 내 텍스트에 해당하는 문장 혹은 어휘의 극성 분포(긍정, 부정)에 따라 의견을 스코어링(scoring)하는 감성분석과 관련된 연구들도 다수 이루어지고 있다[15]. 하지만 대부분 투자 및 재테크에 관련된 주제인 '주가지수'와 '부동산'에 관련된 내용이 대부분이며 주된 연구는 다음과 같다.

동일한 어휘의 극성이 해석하는 사람의 입장에 따라 또는 분석 목적에 따라 서로 상이하게 해석되는 현상은 지금까지 다루어지지 않은 어려운 이슈로 알려져 있는데, 구체적으로는 주가지수의 상승이라는 한정된 주제에 대해 각 관련 어휘가 갖는 극성을 판별하여 주가지수 상승 예측을 위한 감성사전을 구축하고, 이를 기반으로 한 뉴스 분석을 통해 주가지수의 상승을 예측한 연구가 있다[16]. 또한, 뉴스 콘텐츠를 분석하기 위해 빅데이터 감성분석 기법을 적용하여, 주가지수의 등락을 예측하는 지능형 투자의사결정 모형을 제시하였으며, 이렇게 도출된 모형은 여러 유형의 뉴스 중에서 시황·전망·해외 뉴스가 주가지수 변동을 가장 잘 예측하는 것으로 나타났다[17]. 추가로, 인터넷 뉴스 매체별 주가지수 예측 모델을 수립하고 적중률 분석을 수행한 연구가 있다[18].

또 다른 연구로는 온라인 언급이 기업 성과에 미치는 영향을 분석한 것이 있는데, 특정 주제에 대한 사전구축이 아닌 온라인 뉴스 정보를 활용한 기업의 어휘사전 구축을 통해 개별 기업의 주가 등락 예측에 대한 분석

을 수행하였으며, 향후 감성사전 구축 시 불필요한 어휘가 추가되는 문제점을 보완한 연구 수행을 통하여 주가 예측정확도를 높이는 방안을 모색하는 것에 기여한다[19]. 여기까지 언급한 연구들의 주제는 주가 예측이었다면, 다음은 부동산이다. 대용량의 부동산 관련 기사를 수집하여 빅데이터 기술을 통해 부동산시장을 분석할 수 있을 것이란 가정하에 웹 크롤링(Web Crawling)에 의해 수집한 부동산 관련 뉴스기사의 감성분석을 통해 산출한 데이터와 실거래가 데이터가 비교적 높은 상관관계가 있음을 발견할 수 있었다[20].

하지만 본 연구에서는 뉴스 데이터를 활용하여 주가지수나 부동산 관련 주제가 아닌 산업 및 지역 생태계 위기 예측을 위해 감성분석을 활용하였다는 점에서 기존의 연구와 큰 차별성이 있다.

Ⅲ. 연구 방안

1. 실험 설계

본 연구는 1차 분석과 2차 분석으로 2가지 실험 설계로 진행되었다. 대용량 뉴스 데이터를 수집하여, Google Sentiment Analysis API를 활용한다. 이를 통해 뉴스 감성분석을 실행하고, 이를 월별로 정리하여 감성 분석 결과 실제 이벤트 간의 연관 관계를 확인하여, 뉴스 데이터를 통한 지역 산업 위기 예측 가능성을 점검해보았다.

1차 분석에서는 “자동차”, “산업”과 관련된 뉴스를 수집하여, 해당 데이터의 뉴스 감성 정보를 수집하고, 주요 이벤트와 감성지수 사이의 연관관계를 정성적으로 평가한다. 2차 분석에서는 “자동차”, “광주”와 주요 일보, 지역 신문, 경제지의 뉴스를 수집하여 감성정보와 경제지표 및 산업단지활동지표 사이의 연관성을 정량적으로 평가하였다. 1차 분석에서는 보다 큰 규모의 산업 생태계를 살펴보았다면, 2차 분석에서는 “광주”라는 지역 생태계를 알아보고자 분석 대상을 좁혀서 진행한 것이라고 보면 된다.

2. 데이터 수집 및 분석방법

2.1 데이터 수집방법

표 1. 1차 분석

기간	2018년 01월 01일 ~ 2019년 08월 15일
키워드	“자동차”, “산업”
출처	네이버 게시 뉴스
데이터 수	31,479건
전처리여부	X

1차 분석은 “자동차”, “산업”을 키워드로 하여 2018년 1월 1일부터 2019년 8월 15일까지 게시된 뉴스를 수집하였다. 수집데이터는 총 31,479건이며, 이는 네이버 뉴스 중 네이버 게시 뉴스에 해당한다[표 1].

표 2. 2차 분석

기간	2016년 01월 01일 ~ 2018년 12월 31일
키워드	“자동차”, “광주”
출처	네이버 뉴스 〈주요 주간지〉 : 조선일보, 중앙일보, 동아일보, 경향신문, 한겨레 〈경제/IT〉 : 서울경제, 매일경제, 한국경제 〈지역지〉 : 광주드림, 광주매일신문, 전남일보, 전라일보, 전북도민일보, 전북일보
데이터 수	5,803건
전처리여부	인사, 동정 등과 관련된 뉴스 제거 정부정책 전문 등이 포함된 뉴스 제거

2차 분석은 “자동차”, “광주”를 키워드로 기간은 2016년 1월 1일부터 2018년 12월 31일까지다. 수집데이터는 총 5,803건이며, 네이버 뉴스 데이터를 수집하였다. 해당 데이터는 주요 주간지 (조선일보, 중앙일보, 동아일보, 경향신문, 한겨레), 경제/IT(서울경제, 매일경제, 한국경제)와 지역지(광주드림, 광주매일신문, 전남일보, 전라일보, 전북도민일보, 전북일보)로 구성된다. 해당 데이터는 전처리 과정을 통해 인사, 동정 등과 관련된 뉴스와 정부 정책 전문 등이 포함된 뉴스는 제거하였다[표 2].

월별 경제지표(경제활동인구, 고용률, 실업률, 실업자, 취업자)와 산업단지활동지표(위기지수, 고용현황, 수출실적, 가동률, 생산실적, 입주기업 수)는 국가통계포털(KOSIS, Korean Statistical Information Service)에서 제공하는 데이터를 기본으로 광주광역시의 2016년 1월부터 2018년 12월까지의 데이터를 수집하였다.

2.2 데이터 분석방법

키워드에 해당하는 네이버 뉴스 데이터를 수집하여 저장된 뉴스 데이터 중 마지막 글자 150자는 제외하였다. 이유는 '저작권자', '무단전재 및 재배포 금지', '언론사명', '기자명', '메일주소' 등이 들어가 있기 때문이다. 해당 연구에 있어서 하단에 공통으로 들어가 있는 그와 같은 내용들은 불필요하기에 모두 제거를 해주었다. 그 외 데이터 전처리 과정은 Google Natural Language API가 지원해주는 내에서 처리를 진행한 후, 감성 정보를 분석하였다.

감성 점수 중 스코어는 -1과 +1 사이의 값을 가진다. -1은 부정(negative), +1은 긍정(positive)의 감성 분석 결과가 생성되었다. 이때, magnitude는 전반적인 감성 강도를 나타내며, 감성 강도가 높은 데이터를 따로 뽑게 되면 그것에 맞게 감성 강도가 높은 정보만 추출할 수 있다.

본 연구에서는 위와 같은 분석방법을 통해 감성 정수와 각 경제지표 및 지역 산업단지활동지표 간의 상관도를 고려하여 감성 점수와 지역 경제활동 간의 상관관계를 파악하고자 한다.

Ⅳ. 연구 결과

1. 1차 분석

1.1 일별 데이터 뉴스 데이터 변화량

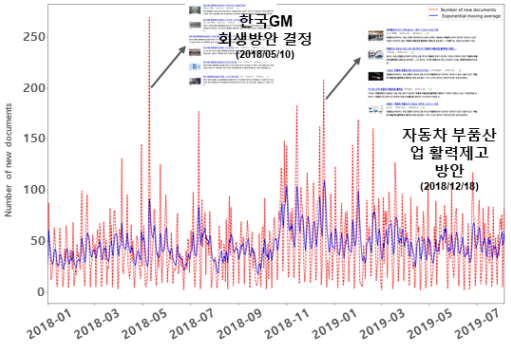

그림 1. 1차 분석 - 일별 데이터 뉴스 데이터 변화량

2018년 한국 GM 이슈 등 관련된 이슈가 생길 때 많은 뉴스들이 생성된 것으로 판단된다. 따라서, 특정 이슈가 있을 때 데이터의 양으로 거시적인 유추를 해볼 수 있다고 볼 수 있다[그림 1].

1.2 일별 데이터 뉴스 감성정보 변화

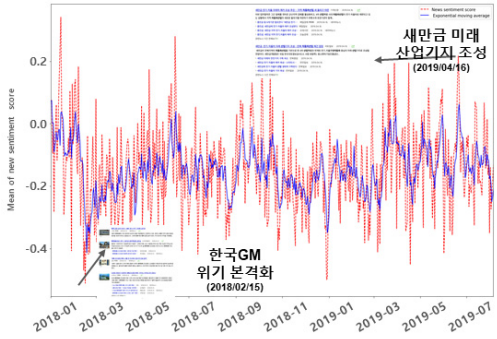

그림 2. 1차 분석 - 일별 데이터 뉴스 감성정보 변화

[그림 2]는 1차 분석방법을 토대로 일자별 데이터를 뉴스의 Negative와 Positive를 일자별로 분석해본 결과이다. 2018년부터 한국 GM으로 인해 2018년 전반기는 비교적 부정적인 경향이 많았다. 하지만 2018년 5월 이후 GM 문제가 일부 해결되며 긍정적인 뉴스가 나오는 날짜가 증가하는 결과를 볼 수 있었다. 대규모 뉴스 데이터의 감성분석은 특정 사건들과 맞물려 의미 있는 결과를 도출할 때 도움이 될 것으로 판단된다.

2. 2차 분석

2.1 일별 데이터 뉴스 데이터 변화량

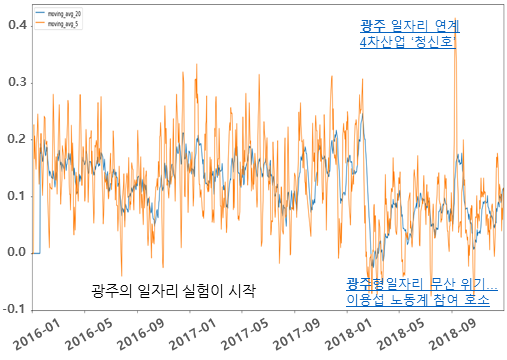

그림 3. 2차 분석 - 일별 데이터 뉴스 감성정보 변화

광주의 경우 지난 2년간 연속적으로 "광주형 일자리" 사업과 관련하여 뉴스가 나오고 있으며, 비교적 최근

광주 일자리에 대한 긍정적 평가가 나오며 실현의 기대가 높아지고 있다는 것을 알 수 있었다[그림 3].

2.2 월별 경제 및 산업지표와 감성정보 비교

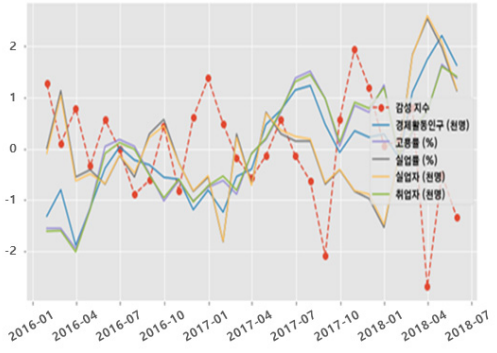

그림 4. 월별 경제지표 및 감성정보 변화 추이

[그림 4]를 통해 경제활동 지표의 경우 2016년 초기보다 2018년 이후로 갈수록 지표가 나아지는 것을 확인할 수 있었다. 0을 기준으로 하였을 때 위는 긍정, 아래는 부정이라고 판단하면 된다. 그리고 감성분석 데이터와 함께 월별 경제지표를 비교해보았다. 감성지수의 경우 이동 평균 기법을 활용하여 1개월을 하나의 Window로 가정하고, 2~6개월 간의 평균을 이동 평균 변수로 추가하였다. 동일한 방법으로 time lag 정보를 활용하여 1개월 전부터 6개월 전까지의 기록을 변수로 추가하여 분석을 수행하였고, 월별 감성지수와 각 경제지표 간의 상관관계를 분석하여 감성지수의 경제지표 영향도를 검토하고자 하였다.

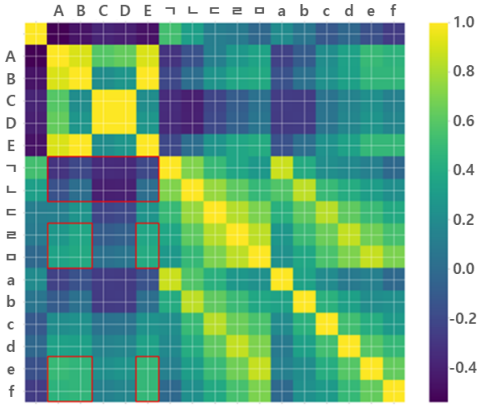

그림 5. 감성분석 지표와 경제지표 간의 상관관계 도표

표 3. 감성분석 지표 기호 및 설명

기호	설명	기호	설명
A	경제활동인구(천명)	ㄱ	senti_rolling_2
B	고용률(%)	ㄴ	senti_rolling_3
C	실업률(%)	ㄷ	senti_rolling_4
D	실업자(천명)	ㄹ	senti_rolling_5
E	취업자(천명)	ㅁ	senti_rolling_6
F	입주기업 수	a	senti_shift_1
G	생산실적	b	senti_shift_2
H	가동률	c	senti_shift_3
I	수출실적	d	senti_shift_4
J	고용현황	e	senti_shift_5
K	위기지수	f	senti_shift_6

[표 3]은 감성분석 결과인 [그림 5]와 [그림 7]의 기호와 그에 대한 설명을 나타낸 것이다. [그림 5]와 [그림 7]의 가로축과 세로축은 월별 경제 및 산업단지활동지표와 이동 평균, 이동 지수를 뜻한다. A~E까지는 경제지표, F~K까지는 산업단지활동지표이며, 앞에서 언급한 2~6개월 간의 이동 평균 변수를 senti_rolling_2와 같은 형태로 각각 표기하였고, 기호는 ㄱ~ㅁ으로 나타내었다. 같은 방식으로 1개월 전부터 6개월 전까지의 기록을 변수로 추가한 time lag 데이터는 senti_shift_1과 같은 형태로 표기하였으며 기호는 a~f로 나타내었다.

상관도는 [그림 5]와 [그림 7]의 세로축을 보면 되는데, 색깔이 밝을수록 상관도가 높고 어두울수록 상관도가 낮은 것이며, 0을 기준으로 위는 양의 상관관계이고 아래는 음의 상관관계이다. 이처럼 시간에 따른 감성지수의 변화를 통해 경제 및 산업지표 중 어떤 것과 가장 상관관계가 높은지 알 수 있다.

결과적으로 상관도 데이터의 경우 실업률과 실업자수를 중심으로 보았을 때, 2~3개월 차이의 초기 감성지수의 경우에는 음의 상관관계가 강하게 나타나는 것을 확인할 수 있었고, 2~3개월 이후부터 6개월 이전의 감성지수는 대체적으로 양의 상관관계가 나타났다. 이 밖에 다른 지표들을 보았을 때도 2개월 정도 차이에 근접할수록 음의 상관관계가 나타났으며, 시간이 지날수록 대체로 양의 상관관계를 띄는 것으로 나타났다. 즉, 감성지수의 경우 단기간의 영향이 나타나는 것이 아니라 몇 개월 후부터 영향이 나타나는 것으로 판단할 수 있었다[그림 5].

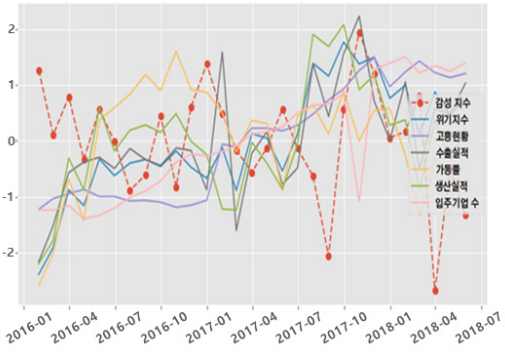

그림 6. 월별 광주 산업단지활동지표 및 감성정보 변화 추이

[그림 6]은 감성분석 데이터와 광주 산업단지 간의 활동지표를 비교한 것으로, 추가된 변수는 경제지표와 동일하며 감성지수의 변화와 각 산업단지활동지표 간의 상관관계를 비교 분석해보았을 때, 전반적으로 Window 사이즈가 5~6 사이에 가장 높은 상관도를 보이는 것을 알 수 있었다.

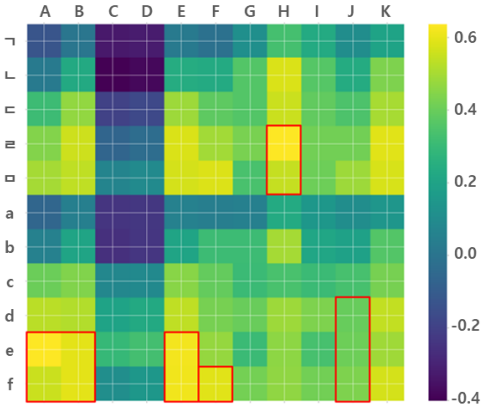

그림 7. 감성분석 지표와 경제 및 산업단지활동지표의 상관관계 도표

즉, 전체적인 지표들과 감성지수의 이동 평균 및 이동 지수를 비교했을 때 5개월 이후에서 본격적인 감성지수의 효과가 나타나는 것으로 판단된다. 특히, 산업단지 활동지표면에서는 “입주기업 수”, “생산실적”, “가동률” 등이 가장 먼저 영향을 받는 모습을 보였고, 다음으로는 “수출실적” 및 “고용현황”이 순차적으로 영향을 받아 전체적인 지표가 향상되는 것으로 판단된다[그림 7].

V. 결 론

지역 산업생태계가 쇠약해지는 문제가 커지고 있으나, 미리 조기에 측정해보는 지역 산업생태계 위기에 관한 연구는 거의 이루어지지 않고 있다. 위기는 사후에 대응하기에는 역부족이므로 선제적으로 위기를 인지하고 대처할 수 있는 예측 연구가 필요한데, 본 연구는 대용량 뉴스 텍스트 데이터를 활용하여 지역 산업생태계의 위기를 조기에 알아본다는 점에서 큰 의의가 있다. 수집한 데이터를 불필요한 부분을 제거하는 전처리 과정을 거친 후, Google API를 활용한 감성분석 기법을 통해 월별로 정리하여 감성 분석 결과와 실제 이벤트 간의 연관관계를 확인해보았다.

그 결과, 감성지표는 시차를 두고 경제지표에 영향을 준다는 결론을 도출할 수 있었고, 특히 5개월 이상의 데이터를 활용한 경우, 감성지표와 경제지표 간의 양의 상관관계를 찾을 수 있을 것으로 기대된다.

지역 산업생태계 분야에서 뉴스 데이터를 활용한 국내연구는 초기 단계지만, 앞으로 그 활용도는 더욱 크게 확대될 것으로 사료되며, 특히 이런 뉴스 데이터의 경우 실시간으로 정보가 전달이 될 수 있고 이슈를 빠르게 반영한다는 점에서 그 유용성이 크다고 볼 수 있다.

하지만 이런 유용성에도 불구하고 데이터 전처리 과정을 수작업으로 진행하였기 때문에 작업자 의존성이 높다는 한계점을 가짐과 동시에 광주 지역만을 대상으로 하였기 때문에 연구 결과의 보편적 적용 측면에서는 한계를 가지고 있다.

따라서 향후에는 검색 결과로 확보된 데이터에 대해 공통적인 기준으로 노이즈 제거를 할 수 있는 모델 연구가 필요하며, 실험의 유효성을 높이기 위해 광주 외 다른 지역 데이터 대상으로 동일 프로세스로 작업하여 결과치를 비교하는 추가 연구가 필요하다. 또한, 데이터 적합성 제고를 위해 추가적인 경제지표의 발굴 및 비교분석이 필요하며, 보다 수월한 데이터 확보를 위해 산업생태계 위기 관련 뉴스에 대한 자동 분류기 개발 등이 필요하다.

참 고 문 헌

[1] http://www.jiat.re.kr/jaiicmgr/upload/news/6e1037d1-c7e1-4c73-8cec-39b733f7fc28.pdf

[2] http://www.donga.com/news/article/all/20170705/85219452/4

[3] 한국은행, *전북지역 자동차산업 현황과 대응전략*, 2018.

[4] https://www.gwangjuin.com/news/articleView.html?idxno=202221

[5] 이민철, 김혜진, "텍스트 마이닝 기법을 적용한 뉴스 데이터에서의 사건 네트워크 구축," 지능정보연구, 제24권, 제1호, pp.183-203, 2018.

[6] Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vol.10 pp.79-86, 2002.

[7] 송민채, 신경식, "뉴스기사를 이용한 소비자의 경기심리지수 생성," 지능정보연구, 제23권, 제3호, pp.1-27, 2017.

[8] 송치영, "뉴스가 금융시장에 미치는 영향에 관한 연구," 국제경제연구, 제8권, 제3호, pp.1-34, 2005.

[9] J. Dimmick, Y. Chen, and Z. Li, "Competition between the Internet and traditional news media: The gratification opportunities niche dimension," The Journal of Media Economic, Vol.17, Issue 1, pp.19-3, 2004.

[10] 이종호, 장후은, "유럽의 산업위기지역 지원정책 추진 동향 및 시사점," 한국경제지리학회지, 제22권, 제3호, pp.246-257, 2019.

[11] 전지혜, 이철우, "한국 산업위기지역의 현 주소: 구미지역 산업 환경과 위기실태," 한국경제지리학회지, 제22권, 제3호, pp.291-303, 2019.

[12] 정성훈, "한국 산업위기지역에 대한 정책적 진단과 처방," 한국경제지리학회지, 제22권, 제3호, pp.237-245, 2019.

[13] 이두희, "지역산업위기 원인과 극복방안," 한국경제지리학회 춘계 학술대회 발표자료, 2019.

[14] 조성철, "지역별 제조업 고용변화에 대한 자동화와 세계의 영향," 한국경제지리학회지, 제22권, 제3호, pp.274-290, 2019.

[15] 정지선, 김동성, 김종우, "온라인 언급이 기업 성과에 미치는 영향 분석 : 뉴스 감성분석을 통한 기업별 주가 예측," 지능정보연구, 제21권, 제4호, pp.37-51, 2015.

[16] 유은지, 김유신, 김남규, 정승렬, "주가지수 방향성 예측을 위한 주제지향 감성사전 구축 방안," 지능정보연구, 제19권, 제1호, pp.95-110, 2013.

[17] 김유신, 김남규, 정승렬, "뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형," 지능정보연구, 제18권, 제2호, pp.143-156, 2012.

[18] 천세원, 김유신, 정승렬, "뉴스 콘텐츠의 오피니언 마이닝을 통한 매체별 주가상승 예측정확도 비교 연구," 한국지능정보시스템학회 학술대회논문집, pp.133-137, 2013.

[19] 정지선, 김동성, 김종우, "온라인 언급이 기업 성과에 미치는 영향 분석 : 뉴스 감성분석을 통한 기업별 주가 예측," 지능정보연구, 제21권, 제4호, pp.37-51, 2015.

[20] 경정익, 이국철, "Textmining에 의한 부동산 빅데이터 감성분석 모형 개발," 주택연구, 제24권, 제4호, pp.115-136, 2016.

저 자 소 개

김 현 지(Hyun-Ji Kim) 정회원

- 2017년 2월 : 건국대학교 의학공학부 의용메카트로닉스(공학사)
- 2017년 9월 ~ 현재 : 과학기술연합대학원대학교 한국과학기술정보연구원 데이터 및 HPC 과학 석사과정

〈관심분야〉 : 기술사업화, 산업시장분석, 데이터 및 텍스트 마이닝

김 성 진(Sung-Jin Kim) 정회원

■ 2004년 8월 : 포항공과대학교 산업공학과(공학사)
■ 2011년 8월 : 포항공과대학교 산업경영공학과(공학박사)
■ 2011년 8월 ~ 2013년 5월 : GS칼텍스
■ 2013년 7월 ~ 현재 : 한국과학기술정보연구원 선임연구원

〈관심분야〉 : 기술사업화, 산업시장분석, 지식경영

김 한 국(Han-Gook Kim) 정회원

■ 2007년 3월 : 동경공업대학교 경영공학(공학박사)
■ 2009년 9월 ~ 현재 : 한국과학기술정보연구원 선임·책임연구원
■ 2017년 3월 ~ 현재 : 과학기술연합대학원대학교 데이터 및 HPC 과학 겸임교수

〈관심분야〉 : 기술사업화, 데이터 활용 산업시장분석, 기술가치평가, 텍스트마이닝