한국경영과학회지
제41권 제2호
2016년 5월
http://dx.doi.org/10.7737/JKORMS.2016.41.2.081

온라인상의 기업 및 소비자 텍스트 분석과 이를 활용한 온라인 매출 증진 전략*

김지연 · 조우용 · 최정혜† · 정예림
연세대학교 경영대학

Linking Findings from Text Analyses to Online Sales Strategies

Jeeyeon Kim · Wooyong Jo · Jeonghye Choi · Yerim Chung
Yonsei School of Business

■ Abstract ■

Much effort has been exerted to analyze online texts and understand how empirical results can help improve sales performance. In this research, we aim to extend this stream of research by decomposing online texts based on text sources, namely, companies and consumers. To be specific, we investigate how online texts driven by companies differ from those generated by consumers, and the extent to which both types of online texts have different effects on online sales. We obtained sales data from one of the biggest game publishers and merged them with online texts provided by companies using news articles and those created by consumers in user communities. The empirical analyses yield the following findings. Word visualization and topic analyses show that firms and consumers generate different contexts. Specifically, companies spread word to promote their own events whereas consumers produce online words to share winning strategies. Moreover, online sales are influenced by consumer-generated community topics whereas firm-driven topics in news articles have little to no effect. These findings suggest that companies should focus more on online texts generated by consumers rather than spreading their own words. Moreover, online sales strategies should take advantage of specific topics that have been proven to increase online sales. In particular, these findings give startup companies and small business owners in variety of industries the advantage when they use the online channel for distribution and as a marketing platform.

Keywords : Text Analysis, Online Sales Strategy, News Article, Online Communities

논문접수일 : 2016년 02월 15일 논문게재확정일 : 2016년 05월 09일
논문수정일(1차 : 2016년 04월 26일)
* 본 연구는 연세대학교 경영대학 BK21 Plus(창의성과 기업가정신을 기반으로 지속성장 가능한 스타트업 전문인력 양성팀)의 지원으로 수행되었음.
† 교신저자, jeonghye@yonsei.ac.kr

1. 서 론

1.1 연구배경

기업 및 소비자의 활동이 기업 성과에 미치는 영향은 전통적인 경영학의 주된 관심 분야로, 기업 성과를 좌우할 수 있는 변인들에 대한 다양한 연구가 이루어져 왔다. 그러나 기업 및 소비자가 만들어내는 텍스트 정보에 관한 학문적 탐구는 데이터 구축의 어려움으로 인하여 매우 제한적으로 이루어져 왔다. 하지만, 정보 통신 기술의 발전과 함께 텍스트 데이터의 획득은 보다 용이하게 되었으며, 이와 함께 '온라인 환경'에서의 기업 및 소비자의 구체적인 활동과 콘텐츠로 학문적 관심의 폭이 확장되었다. 예를 들어, 기업 활동 측면에서는 정보 탐색 및 교환, 소비자와의 커뮤니케이션과 같은 활동을 온라인 광고와 뉴스기사 등을 대상으로 탐구하고 있다[2, 6]. 한편, 소비자 활동 측면에서는 소셜 네트워크 상의 상호 작용 및 구전 활동(Word-of-Mouth)에 대하여, 제품 및 서비스에 대한 리뷰와 평가점수 등을 활용하여 적극 연구되고 있다. 기존 오프라인 환경과는 다르게, 온라인 환경에서 보여지는 기업과 소비자 활동의 특징은 모두 '텍스트(text)'로서 포착 가능한 의사 소통이라는 점인데, 본 연구는 온라인 상에서 생성되는 텍스트를 보다 심도 있게 분석하고, 텍스트들이 내포하고 있는 구체적인 콘텐츠가 기업 성과에 미치는 영향에 주목하고자 한다.

본 연구는 온라인 채널에서 기업과 소비자의 활동이 기업 성과와 어떠한 연관성을 갖는지 살펴보았던 기존 연구들을 바탕으로, 관련 이론을 증진시키고 확장하고자 하며, 다음 세 가지를 핵심 연구 목표로 제안한다[15]. 첫째, 본 연구는 온라인 상에서 기업과 소비자들의 활동으로 인하여 발생되는 텍스트를 수집하고, 이것이 단어 단위로는 어떠한 모습을 보이는지 포착한다. 선행 마케팅 연구들에 따르면, 온라인 상에서 이용자(기업 및 소비자)가 발생시킨 의견이나 감정이 기업 성과와 어떠한 관계를 보이는지 분석하는 연구들은 많이 있었다. 특히 텍스트를 처리할 때 필요한 기계학습과정(machine learning)을 최소화하기 위하여, 소비자에게 직접 설문조사를 하거나[3], 평가 점수나 리뷰를 이용자 의견의 긍정·부정으로 나누어 활용하거나 리뷰의 개수를 활용한 연구들이 많이 있었다[11, 19]. 하지만, 이용자의 구체적인 의견이 표현된 텍스트의 콘텐츠를 직접적으로 파악하는 연구는 부족한 실정이다[22]. 둘째, 본 연구는 텍스트 데이터의 관찰의 범위를 넓혀 기업과 소비자가 발생시키는 텍스트의 주제(topic)는 어떠한지 파악한다. 일반적으로 텍스트에서 추출된 키워드의 빈도수를 활용하여 분석하면, 텍스트의 흐름과 단순하고 표면적인 의미들을 파악할 수 있다. 하지만, 이용자의 전체 메시지에서 표현하고 있는 전반적인 내용과 맥락을 충분히 반영하기 어려운 한계가 있다[7]. 때문에, 텍스트의 주제를 추출하여 분석에 반영하는 것은 매우 큰 의미가 있다. 셋째, 본 연구는 구체적인 텍스트 주제(topic)를 파악하여 기업과 소비자들의 활동이 실질적으로 기업의 성과와 어떠한 관계에 있는지 고찰하고자 한다. 기존의 텍스트 마이닝 분석 연구들이 텍스트의 트렌드를 보는데서 그치는 경우가 많았다면, 본 연구는 텍스트 분석 결과를 실증 분석 모형에 포함시킴으로써 모형의 적합성을 높임과 동시에 기업의 성과 예측에 더욱 의미있는 결과를 도출하고자 하였다.

1.2 이론적 배경

기업 성과에 영향을 주는 온라인 텍스트와 관련하여 기존의 마케팅 연구 흐름은 크게 두 가지 차원으로 구분할 수 있는데, 온라인 상에서 발생되고 있는 텍스트의 양(volume)과 텍스트의 방향성(긍정 또는 부정)을 꼽을 수 있다. 텍스트의 양은 발생된 구전의 양을 의미하고, 텍스트의 방향성은 구전이 포함하고 있는 정보의 선호를 의미하며, 주로 긍정·부정 또는 이용자의 평점으로 측정한다. 기업 성과에 대한 텍스트의 영향력을 설명함에 있어 텍스트의 양과 방향성의 비교는 중요한 의미가 있다. 예를 들어, Chevalier and Mayzlin[12]은 도서 리뷰

데이터를 분석하여, 구전의 양과 방향성은 기업 성과를 증가시키는 역할을 한다고 보았다. Liu[19]는 온라인 상 구전의 양은 기업의 성과와 긍정적인 관계에 있지만, 구전의 방향성과 기업 성과의 관계는 혼재되어 있다는 것을 밝혔다. 반면, Chen and Yoon [11]은 기업의 성과와 구전의 방향성은 서로 영향이 없음을 확인하였다. 또한 Pauwels et al.[22]은 소비자들이 얼마나 많이(volume), 어떤 태도로 이야기하는가(valence)보다 어떠한 대화를 나누는가(content)를 아는 것이 더 중요하며 이를 검증하기 위하여, 구전의 효과를 단기적・장기적으로 살펴봄은 물론, 기업 성과에 대한 전통적인 마케팅 활동(paid marketing)과 소셜 미디어(social media)의 직접적・간접적인 효과를 복합적으로 보기도 하였다.

더 나아가, 일부 선행 연구들은 온라인 채널에서 텍스트를 발생시키는 주체를 구분하여 이를 분석하였는데, 대표적으로 기업과 소비자를 구분한 연구들을 꼽을 수 있다. 우선, 소비자들이 포럼이나 게시판에서 생성하는 정보와 기업이 자사 홈페이지 상에서 생성하는 정보를 비교하여, 어떤 정보가 소비자에게 더 큰 영향을 주는지 확인하는 연구가 있었다[10]. 소비자들의 구전 활동은 전통적인 기업의 마케팅 활동과 비교하여 기업 성과에 긍정적인 피드백 매커니즘이 있다는 것을 밝히는 연구도 있었다[14, 23]. 이러한 매커니즘은 구전이 단순히 소비자의 구매를 일으키는 원동력으로써 뿐만 아니라, 기업의 판매 성과에도 직접적인 영향을 미칠 수 있다는 것을 시사한다. 위와 같은 많은 연구들이 텍스트의 효과를 전통적인 기업 마케팅과 마찬가지로 외생적인 요인으로서 파악하였다면, Duan et al.[13]은 기존 연구를 확장하여 구전 자체에서 예측될 수 있는 내생적 요인을 통제함은 물론, 패널 데이터 상에서 소비자 구전과 기업성과 간의 역동적인 변화와 관계를 확인하였다. 또한, Villanueva et al.[24]은 단기적으로는 기업의 마케팅 활동이 기업 성과에 의미 있는 영향을 주지만, 장기적으로는 기업의 마케팅 활동에 비하여 소비자의 구전이 2배 가까이 더 의미 있는 영향을 준다는 것을 밝히기도 하였다.

앞서 언급된 연구들은 대부분 특정 콘텐츠에 대하여 생성된 텍스트의 양이나 조회한 숫자를 변수로서 사용하거나, 연구자 혹은 연구자가 고용한 사람이 그 방향성(긍정 및 부정)을 판단하였다. 이러한 방법으로 텍스트를 활용한 분석은 대량의 텍스트 데이터가 생성되는 온라인 환경에서 모든 텍스트를 파악하고 분석하는 것이 어렵고 비용도 많이 들기 때문에, 실질적으로 적용하기가 쉽지 않다. 또한 이러한 접근은 텍스트의 트렌드와 패턴을 파악할 수 있지만, 텍스트의 구체적인 콘텐츠를 파악하는데는 한계를 갖는다. 다시 말해, 사람의 평가를 바탕으로 한 분석은 비정형 데이터를 생성하는 데 적합할 수 있지만, 태생적으로 오류의 위험성과 시간 및 비용 소요가 많다는 문제점을 갖고 있다. 그러므로 이러한 문제점을 해결한 방법을 활용하여 분석한 연구가 보다 절실한 상황이다. 따라서 이러한 기존 연구의 한계점을 극복하고자 자동화된 텍스트 마이닝 분석이 필요하다.

텍스트 마이닝 분석을 활용한 연구는 주로 계량정보학 및 문헌 정보학 분야에서 활발히 이루어지고 있다. 이 분야의 연구들은 자동화된 텍스트 분석을 통하여 텍스트 데이터의 트렌드와 패턴을 파악하고 있는데, 이는 주어진 텍스트에서 의미 있는 단어들을 추출하고, 단어 사이의 관계를 밝혀내는 것을 핵심으로 한다. 대표적으로, Archak et al.[8]는 제품 리뷰를 텍스트 마이닝 방법으로 다방면에서 분석하였을 때, 소비자들의 제품 선택을 좀 더 성공적으로 예측할 수 있다는 점을 확인하였으며, Netzer et al.[21]은 기존의 트렌드 파악을 위한 텍스트 마이닝 기법 연구들을 확장하여, 텍스트 마이닝 분석과 의미적 네트워크 분석 툴(semantic network analysis tools)과 결합하였다. 이를 통하여 소비자 생성 콘텐츠(consumer-generated contents)를 기반으로 구축된 시장 구조가 전통적인 대규모 판매 데이터 및 설문조사 데이터를 이용하여 유도된 시장 구조와 비슷하다는 것을 밝혀냈다. 본 연구는 이러한 기존 연구들에서 시도했던 텍스트 데이터의 트렌드 파악과 더불어, 텍스트의 특성이 기업 성과에 미치는 영향

을 구체적으로 진단한다는 점에서 그 의의를 찾을 수 있을 것이다. 이를 위해 크게 세 가지의 연구 목적과 질문을 제안하며, 이를 다음 절에서 보다 구체적으로 서술하고자 한다.

1.3 연구 목적 및 연구 질문

본 연구의 목표는 온라인 상에서 기업과 소비자가 발생시키는 텍스트 데이터의 패턴을 파악하고 구체적인 내용을 분석하여, 최종적으로 기업 성과를 예측하는 모형을 설계하는데 있다. 이와 같은 연구 목표를 달성하기 위하여 도출한 구체적인 연구 질문은 다음과 같다. 첫째, 온라인 상에서 기업 또는 소비자가 무엇에 대하여 이야기 하고 있는지 시각적으로 확인할 수 있는가? 둘째, 기업이 작성한 텍스트와 소비자가 작성한 텍스트에서 주요 단어와 토픽들은 각각 어떠한 양상으로 나타나는가? 셋째, 특히 소비자가 작성한 텍스트의 주제들은 실제 기업 매출에 어떠한 영향을 주는가? 이상의 연구 질문에 대한 답을 찾기 위하여, 본 연구는 기업과 소비자들이 작성한 텍스트를 분석하고, 텍스트 분석을 통해 도출된 토픽들을 변수화 하여 실제 기업 데이터에 적용할 것이다.

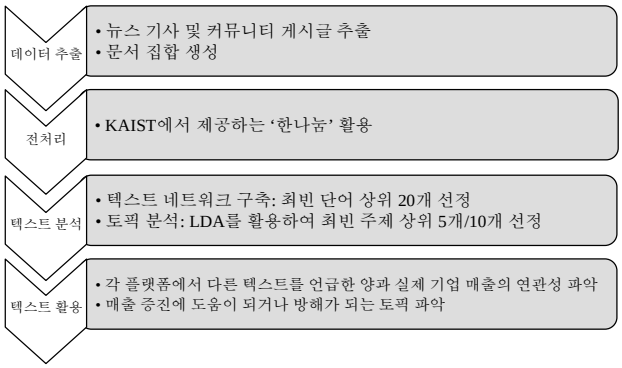
본 연구는 기업 및 소비자 텍스트와 관련된 연구 질문을 해결하기 위해, 온라인 서비스 중 게임 산업 분야의 데이터를 수집하기로 결정하였다. 게임 산업은 시장 내 상호 작용의 정도가 매우 큰 산업으로, 서비스에 대한 소비자들의 적극적인 의사 개진과 소통이 활발한 산업 군 중 하나이다. 또한 해당 게임은 온라인을 통해 의사 소통 하는 것이 보다 친숙한 10대에서 30대 사용자들이 많아, 서비스에 대한 의견뿐만 아니라, 이용자들 간의 소통을 텍스트 형태로 풍부하게 확보할 수 있을 것이라 예상하였다. 이를 위해, 저자들은 국내 온라인 게임 제공업체 중 한 곳과 접촉하였고, 게임 업체의 대표 게임 중 한 게임의 고객 로그 데이터 및 매출 데이터를 확보할 수 있었다. 그리고 게임에 대한 텍스트를 온라인 채널 상에서 수집하여 데이터로 구축하였다. 구체적으로, 해당 게임과 관련된 뉴스 기사와 온라인 커뮤니티 사이트의 게시글을 크롤링(Crawling)하여 수집하였는데, 뉴스 기사는 기업이 온라인 채널에서 소비자에게 정보를 전달하기 위하여 활용하는 수단이고, 온라인 커뮤니티는 소비자들이 정보를 공유하며 기업에 대한 의견을 표출하기 위한 공간으로서 활용하기 때문에 본 연구의 목적에 적합하다고 판단하였다. 이렇게 수집한 데이터는 텍스트 마이닝 분석을 통하여 단어와 주제들을 추출하여 트렌드를 파악함과 동시에, 추출된 정보를 변수화하여 기업 성과 데이터와 결합하였다.

기업 성과 데이터는 일단위(Daily level)로 게임에서 판매된 아이템 매출에 대한 정보와 매출에 영향을 줄 수 있는 접속자들과 관련된 정보를 포함하고 있다. 본 연구에서는 결합된 텍스트 콘텐츠 데이터와 기업 성과 데이터를 회귀 모형으로 실증 분석하였으며, 최종적으로 텍스트 변수가 포함된 모형이 과연 의미 있는 모형인지 검증하기 위하여 적합성(Goodness of fit) 검정을 통하여 확인하고, 10묶음 교차 검증법(10-fold Cross Validation)으로 모형의 타당성을 검증하였다.

이후 이어질 연구 내용은 다음과 같다. 제 2장의 연구 설계에서는 데이터의 수집 및 분석 방법에 대하여 소개할 것이다. 제 3장에서는 데이터의 분석 결과를 제시하고 모형의 적합성을 검증할 것이다. 마지막으로 제 4장에서는 해당 연구의 결론과 시사점을 소개할 것이다.

2. 연구 설계

본 연구의 목적을 달성하기 위한 연구의 흐름은 [그림 1]과 같이 나타내 볼 수 있다. 먼저, 기업의 텍스트 플랫폼인 뉴스 기사와 소비자의 텍스트 플랫폼인 커뮤니티 게시글에서 각각 텍스트를 추출하여 문서 집합을 생성하고, 이를 텍스트 분석에 적합하도록 전처리 한다. 그 이후, 전 처리를 통하여 구축된 텍스트 매트릭스를 활용하여 텍스트 네트워크를 구축하였다. 텍스트 네트워크를 통하여, 기업 또는 소비자들이 온라인 상에서 주로 어떠한 단어를 활



[그림 1] 연구설계

용하여 이야기 하고 있는지 한 눈에 확인할 수 있다. 더 나아가 어떠한 특정 주제들을 이야기하고 있는지 확인하기 위하여 LDA를 활용하여 토픽 분석을 하였다. 최종적으로 추출된 토픽들 가운데 어떠한 주제들이 매출 증진에 도움이 되거나 방해가 되는지 예측해보고자 하였다.

2.1 텍스트 데이터 수집

본 연구에서는 기업과 소비자들이 발생시키는 각기 다른 텍스트들을 비교 분석하기 위하여, 국내 온라인 야구 시뮬레이션 게임인 '프로야구매니저'에 대하여 발생된 텍스트들을 수집하였다. 기업과 소비자들이 발생시키는 텍스트들은 일반적으로 서로 다른 플랫폼에서 생성되는데, 기업은 광고나 뉴스 기사 보도 등을 통하여 해당 기업의 서비스와 상품에 대한 정보를 소비자들에게 전달한다. 반면, 소비자들은 해당 기업의 서비스와 상품을 사용해 본 실사용자들이 커뮤니티를 형성하여 이야기를 나누거나 판매 사이트에 리뷰를 남기고, 개인 소셜 네트워크 상에 정보나 의견을 업로드 하고 공유하기도 한다[9]. 본 연구는 위와 같은 텍스트 생성의 맥락을 최대한 반영하여, 기업 입장이 생성하는 텍스트 데이터는 온라인 뉴스 기사를 대상으로 수집하였고, 소비자들이 생성하는 텍스트 데이터는 해당 기업 서비스와 관련된 의견들이 잘 축적되어 있는 온라인 커뮤니티 사이트의 게시판에서 수집하였다. 텍스트 데이터 수집은 관련 사이트, 즉 온라인 뉴스 기사와 커뮤니티 사이트의 게시글의 HTML 코드를 크롤링한 후, 글의 내용이 담긴 콘텐츠 부분만 추출하여 데이터를 생성하였다. 데이터 수집 기간은 매출 데이터를 갖고 있는 2012년 5월부터 2013년 4월까지 약 1년간이다. 데이터 수집 대상은 뉴스 기사의 경우, 국내 대표 검색 엔진 네이버(www.naver.com) 뉴스 검색을 활용하였고, 대상 게임 이름(프로야구매니저)의 약칭인 '프야매'를 검색하여, 483개의 신문 기사를 확보하였다. 커뮤니티 사이트의 경우에는 프로야구매니저 게임과 관련된 정보를 공유하는 프로야구매니저 인벤(http://bm.inven.co.kr/)의 자유게시판에서 동일한 기간 동안 게시된 5,076개의 글을 수집하였다.

2.2 텍스트 데이터 처리

크롤러(crawler)를 활용하여 수집한 뉴스 기사와 커뮤니티 사이트의 게시글들은 HTML, 스크립트 언어와 순수 텍스트가 혼재되어 있다. 그러므로 불필요한 데이터를 삭제하고 데이터 형식을 통일하여 분석에 필요한 순수 텍스트만을 추출한 후에 텍스트 파일로 저장 후 처리하여 분석에 사용하였다. 이 과정에서 기본적인 전처리 작업은 카이스트(KAIST)에서 제공하는 '한나눔'을 활용하여 자연어를 처리하였으며, 각 단어는 띄어쓰기로 구분되어 있다. 그리고 불필요한 조사 등을 제거하기 위하여 '한나눔'의 명사 어근 추출(Extract Nouns)를 사용하여 명확

한 형태소를 추출하도록 하였다. 또한 일반적인 불용어(stop words) 이외에 뉴스 기사와 커뮤니티에서 사용하는 특정 기호들이나 제목과 콘텐츠 상에 자주 언급되는 단어(e.g., 게임 이름)와 불필요한 단어들도 불용어 사전에 추가하여 빈도 높은 기능어와 주제어로서 가치 없는 기타 고빈어들을 제거하였다. 위와 같은 과정을 거쳐 분석에 활용하기 위하여 최종적으로 추출된 데이터는 [그림 2]와 같다.

[그림 2](a)는 뉴스 기사에서 기사 내용 부분만을 추출한 후, 전처리한 결과이고, [그림 2](b)는 커뮤니티 사이트 게시글의 글 내용 부분만을 추출하여 전처리한 결과이다. 수집한 데이터에서 텍스트를 추출하여 전처리한 결과를 살펴보면, 뉴스 기사의 단어들이 커뮤니티의 단어들에 비하여 정제되어 있으며, 일상적인 단어보다는 좀 더 형식을 갖춘 단어들이 사용되고 있음을 알 수 있다.

2.3 텍스트 데이터 분석 방법

본 연구는 앞서 연구 질문에서도 밝혔듯이, 우선

치어리더 박기량 스마일 게이트 메가포트 야구 시뮬레이션 프로야구매니저 프로야구 이하 실시 인기 치어리더 설문 조사 결과 연속 최고 뽑히 이번 설문 조사 프로야구 대표 치어리더 이르 주제 이달 걸치 진행 롯데

(a) 뉴스 기사에서 텍스트 추출 및 전처리

유창식 호투 광삼형 잘하 넥센 롯데 타격 왠일 막판 전회 점수 안뽑은 이정훈에 대하 예의 오승환에 손승락 탈탈탈 마무리 잔혹사를 보이 두산 어깨 레이드 성공 최고 경기 기아 부모님 미치 소리

(b) 커뮤니티 사이트에서의 텍스트 추출 및 전처리

[그림 2] 텍스트 추출 및 전처리 예시

적으로 텍스트 데이터를 이용하여 기업과 소비자가 온라인 상에서 어떠한 이야기를 하고 있는지 파악하고자 한다. 이를 위해, 단어 등장 빈도수를 기준으로 하여 각 플랫폼(뉴스 기사 사이트와 이용자 커뮤니티 사이트)의 텍스트 중 등장 빈도가 높은 단어들을 추출하고 이를 시각화하고자 한다. 더 나아가, 텍스트의 내용 및 맥락을 충분히 파악하고 텍스트의 구조 및 패턴을 파악하기 위하여, 토픽 분석을 하였다. 토픽 분석은 텍스트 상에서 주로 거론되는 단어들의 군집을 파악하고 각 군집에 대한 토픽들을 레이블링(labling)하며, 선정된 토픽들이 각각 의미하는 바를 분석하는 방법이다. 또한 토픽 분석에서 도출된 주제들을 일자별로 확인하여, 시간 흐름에 따라 변화하는 토픽의 양상을 탐구해 보고자 한다.

2.3.1 텍스트 데이터 시각화

본 연구는 뉴스 기사와 온라인 커뮤니티 간에 다르게 나타나는 중요 단어를 밝혀내는 것을 목표로 하기 때문에, 특정 단어에 가중치를 두지 않고 단순 빈도를 바탕으로 시각화를 진행하였다. 구축한 데이터를 바탕으로 각 플랫폼(뉴스 기사와 커뮤니티 사이트) 간의 특성을 살펴보고자 하였으며, 이를 위하여 텍스트 매트릭스 상에 등장한 단어들을 빈도 수 기준으로 분류하고, 관련 정보들의 단어 간 관계를 시각화하였다. 시각화에는 '단어 동시 출현 네트워크'와 '워드 클라우드(Word Cloud)'를 이용하였다. 핵심 단어의 단순 출현 빈도만을 이용해서는 단어들의 구조를 제대로 파악하기 어렵기 때문에, 핵심 단어와 주변 연관 단어와의 연관성을 종합적으로 고려하여 단어 동시 출현 네트워크를 시각화 하고자 하였다. 워드클라우드는 R프로그램을 이용한 데이터 시각화 기법 중 하나로, 텍스트에 출현하는 단어를 빈도에 비례하는 크기로 표출하는 그림이다. 텍스트 내의 명사들로 구성된 단어 클라우드는 특정 플랫폼에서의 기업 및 소비자들이 발생시키는 텍스트에 대하여, 경제적이고 효과적으로 요약 정보를 제공할 수 있다.

기업과 소비자 간의 텍스트 발생 양상 차이를 살펴보기 위하여, 단어 네트워크를 생성하는 과정은 다음과 같다. 우선, 전처리를 시행한 데이터를 바탕으로 용어-문서 매트릭스(Term Document Matrix)를 구축하였고, 이를 전치(Transpose)하여 동시 출현 단어 매트릭스(Co-occurrence Matrix)를 구축하였다. 이러한 과정을 거쳐 최종적으로 발생된 단어들 가운데 가장 자주 발생되었던 상위 20개의 단위만을 선정한 후 단어 동시 출현 네트워크를 구성하였으며, 단어 등장 최빈값을 기준으로 정렬하여 워드클라우드를 생성하였다.

2.3.2 토픽 분석

토픽 분석은 텍스트 매트릭스 상에서 자주 발생하는 토픽들을 중심으로 군집을 만들고, 각 플랫폼에서 주로 언급되는 토픽을 단어의 분포로 표현함으로써 텍스트의 구조 및 패턴을 파악하고자 하는 방법이다[5, 17]. 본 연구에서는 토픽 분석의 기법 중 기존 연구에서도 많이 사용되고 있는 대표적인 방법인 LDA(Latent Dirichlet Allocation) 알고리즘을 이용하여 분석하였다. LDA 알고리즘은 관찰된 변수인 단어와 콘텐츠(뉴스 기사 또는 게시글)를 통해 보이지 않는 변수를 생성하는 모형이며, 각 콘텐츠의 토픽 비율이 어떻게 구성되는지 결정하는 파라미터를 도출하게 된다[5]. 본 연구에서는 앞서 전처리를 통하여 생성한 데이터를 이용하여 문서-용어 매트릭스(Document Term Matrix)를 구축하였고, 이를 이용하여 LDA 분석을 실시하였다. 이 과정에서 iteration = 5,000, alpha = 0.01, eta = 0.01의 조건을 설정하여 분석하였으며, 그 결과 도출된 자주 언급된 토픽들 가운데 뉴스 기사에서는 상위 5개, 커뮤니티 사이트에서는 상위 10개의 토픽들을 선정하였다.

2.4 기업 성과 데이터 분석 방법

2.4.1 기업 성과 데이터 결합

앞서 밝힌 세 번째 연구 질문인 텍스트 변수가 실제 기업 성과에는 어떠한 영향을 주는지 살펴보기 위하여, 토픽 분석을 통해 구축한 텍스트 데이터와 기업의 성과 데이터를 매칭하였다. 기업 성과 데이

터는 앞서 설명하였던 온라인 야구 시뮬레이션 게임인 '프로야구매니저'의 데이터로서, 2011년 7월부터 2012년 4월까지의 총 296일 분량의 아이템 판매 정보를 담고 있다. 즉, 여기서 기업의 매출 성과는 소비자들의 게임 아이템의 구매를 측정함으로써 파악할 수 있으며, 소비자들이 구매 가능한 게임 아이템은 이용자가 설정한 구단 경영에 이용 가능한 선수 카드, 서브 카드, 부가 상품과 같은 다양한 아이템들을 말한다. 또한 게임 아이템 판매에 영향을 줄 수 있는 게임 접속자수, 접속자 당 게임 시간과 같은 정보 역시 포함하고 있다. 본 연구에서 분석한 기업 성과 데이터는 온라인 상에서 소비자들의 빈번한 구매가 일어나는 제품과 서비스를 대상으로 하였기 때문에, 시시각각 변화하는 소비자들의 반응인 소비자 텍스트가 구매에 어떠한 영향을 주는지 탐구하는데 매우 적합하다.

2.4.2 기업 성과 분석 모형

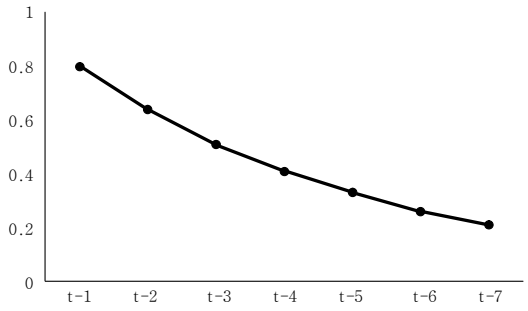
기업의 성과를 예측하기 위한 모형은 식 (1)과 같다. 식 (1)은 오차가 정규분포를 따르는 선형 회귀 모형이다. 이 모형에서 종속변수는 일단위(Daily level)로 t시점에 발생한 아이템 판매 매출을 의미하며, 편향된 분포를 보이고 있기 때문에 로그 변환(log transformation)을 하였다[16, 18]. 판매 성과에 영향을 주는 주요 독립변수로서 앞서 텍스트 분석을 통해 생성하고 추출한 10개의 토픽들을 포함하였다. 토픽 분석을 할 때는 뉴스 기사와 소비자 커뮤니티로부터 추출한 텍스트들의 토픽을 각각 분석하였지만, 본 연구 상에서 추출된 뉴스 기사의 샘플 자체도 크지 않고, 기업 성과에 대하여 소비자 활동이 기업의 전통적 활동에 비하여 더 유의미한 영향을 주므로[14, 23], 기업 성과의 예측력을 더 높이고자 소비자들이 발생시킨 텍스트 토픽들만을 분석 모형에 포함하였다.

$$\log(Sales_t) = \beta_0 + \beta_1 \cdot Number\ of\ Gamers_t + \beta_2 \cdot Game\ Time\ per\ Gamer_i + Topic\ Index1_{\ t} + \cdots + Topic\ Index10_t + \epsilon_t$$

where, $\epsilon_t \sim N(0, \sigma^2)$

$$Topic\ Index\ K_t = \sum_{i=1}^{7} (0.8)^i \times Topic\ K\ Mentioned_{t-i}$$

이 과정에서 토픽의 영향의 강도가 시간의 흐름에 따라 달라질 수 있음을 고려하여, 토픽 지수(Topic Index)로 재구성하여 모형에 포함하였다. 토픽 지수 변수는 식 (2)와 같이 표현할 수 있으며, 커뮤니티 사이트에서 발생하였던 토픽들 중 판매 성과 발생 전 일주일 간 언급된 토픽의 회수를 변수화하여 처리하였다. 관찰할 t시점의 판매 성과를 기준으로 하여, 이전 일주일 간의 토픽들의 영향력은 t시점에서 멀어질수록 약할 것이며, 가까울수록 그 영향력이 강할 것이다. 기업성과 데이터 상에서 아이템 구매의 경향성이 일주일 단위로 반복되며 그 경향성은 선형적으로 감소된다. 일반적으로 광고와 같은 외부적인 활동의 효과가 처음 발현된 이후 관찰 기간의 절반 정도(약 3일)의 시간이 지났을 때, 영향력은 처음 효과의 절반 정도가 된다는 기존 연구를 바탕으로 하여[20], 본 연구에서는 t-1을 0.8의 강도로 영향이 있을 것으로 가정하였다. 이에 따라 일주일 간의 토픽 영향력의 감소는 [그림 3]과 같이 나타난다.



[그림 3] 시점에 따른 토픽 영향력

3. 데이터 분석 결과

3.1 텍스트 데이터 시각화

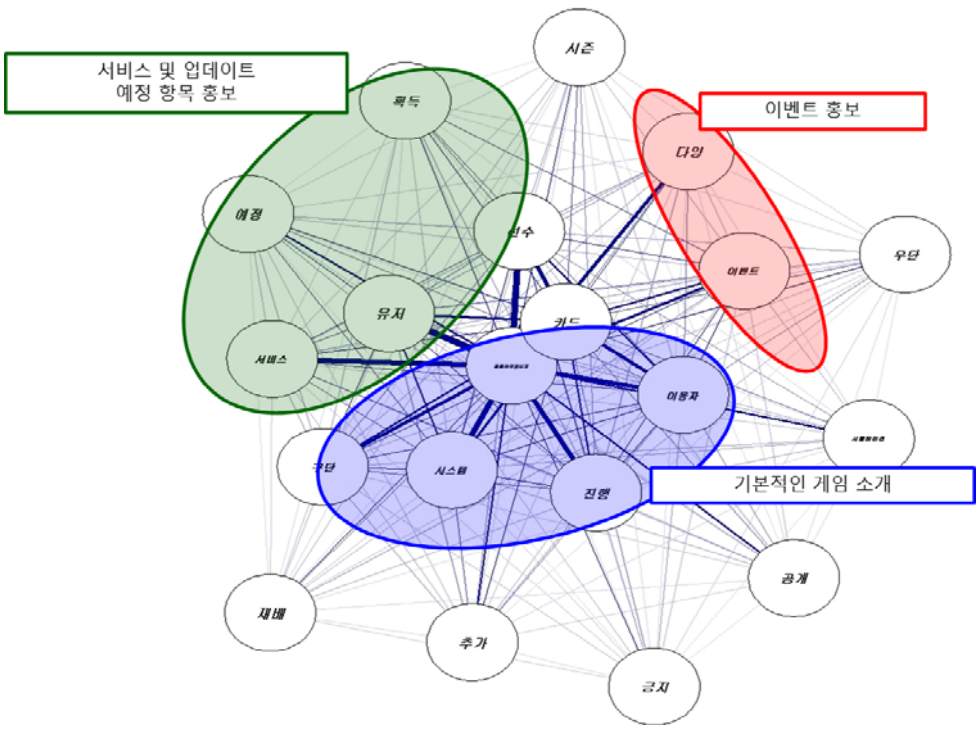
3.1.1 뉴스 텍스트 단어 네트워크

앞서 제 2.2.1절 텍스트 분석 방법에서 언급한 과

정에 따라, 추출한 텍스트를 어근 중심으로 분리하였고, 이를 바탕으로 자주 언급되는 상위 20개 단어를 추출하였다. 뉴스 텍스트의 단어 네트워크를 시각화 한 결과는 [그림 4]에서 확인할 수 있다. 네트워크를 통하여 단어 간의 관계를 분석한 결과, 뉴스 데이터에서 자주 언급되는 핵심어는 1) 기본적인 게임 소개, 2) 서비스 및 업데이트 정보 관련 홍보, 3) 이벤트 홍보 등과 관련되어 있음을 알 수 있다. 이는 뉴스 텍스트 자체가 기업에서 발행한 '보도 자료'를 바탕으로 작성되는 경우가 많기 때문인 것으로 보인다. 즉, 기업은 뉴스 기사를 전통적인 마케팅 활동의 수단으로 활용하고 있으며, 이를 통하여 해당 기업의 서비스와 정보 업데이트 및 이벤트 소식을 소비자들에게 홍보하는 것으로 이해할 수 있다. 또한 핵심어들의 군집에서 파악할 수 있는 사항은 뉴스 기사는 해당 기업의 서비스를 이용하는 소비자들뿐만 아니라, 잠재적인 소비자들에게도 노출될 수 있는 마케팅 활동이므로, 해당 기업 서비스(게임)에 대하여서도 꾸준히 소개하고 있다는 점이다. 다시 말해, 현재 서비스를 이용하고 있는 소비자와 잠재적인 소비자를 뚜렷하게 나누어 홍보하고 있지는 않지만, 현재 서비스를 이용하고 있는 소비자에게는 꾸준히 서비스를 이용하게 하는 동기를, 서비스를 이용하다가 이탈한 고객에 대해서는 다시 서비스를 이용하게 하는 동인을, 아직 서비스를 이용해 보지 않은 고객들에게는 신규로 서비스를 이용하게 할 수 있도록 하는 동기를 제공하고 있는 것이다.

3.1.2 뉴스 텍스트 워드 클라우드

앞서 뉴스 텍스트 키워드 네트워크를 시연하였던 데이터와 동일한 뉴스 텍스트를 바탕으로, 워드클라우드를 추출한 결과는 [그림 5]와 같다. 게임 이름인 '프로, 야구, 매니저'와 같은 단어들은 모든 문



[그림 4] 뉴스 텍스트 키워드 네트워크

[그림 5] 뉴스 텍스트 워드 클라우드

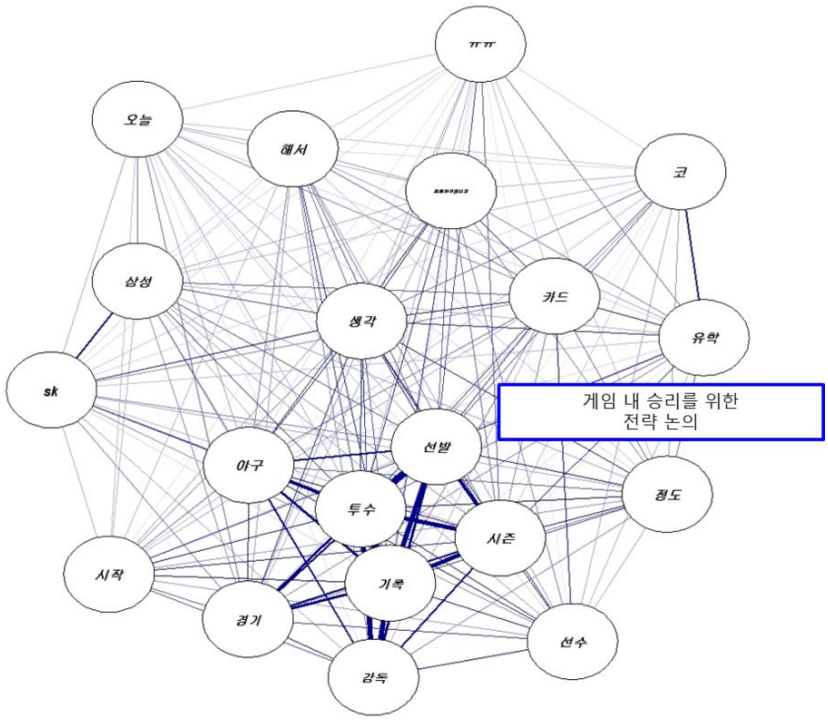
헌에서 당연히 언급 횟수가 높기 때문에, 이를 제외하고 단어를 살펴 보는 것이 합리적인 분석이라 할 수 있다. 위에 표기한 바와 같이 뉴스 텍스트에서 돋보이는 단어는 '이벤트, 서비스, 추가, 모바일, 아이템' 등이며, 앞서 핵심 단어 네트워크에서 확인한 바와 같이 게임 서비스에 이루어지는 업데이트나 이벤트 관련 어구들이 뉴스 텍스트에 활발하게 출현하였다는 것을 확인할 수 있다. 또한 상대적으로 다양한 단어들이 등장하는 것을 알 수 있으며, 특히 야구나 게임 방법에 대한 자세한 콘텐츠를 담은 텍스트 보다는 기업이나 서비스에 대한 표면적인 소개의 단어들이 등장하고 있다는 것을 알 수 있다. 이는 기업이 뉴스 텍스트를 서비스 관련 홍보 수단으로 활용하며 강조하고 싶은 메시지들이 기업과 서비스 및 이벤트에 관한 정보라는 점을 재확인 시켜주는 것이며, 앞선 뉴스 텍스트 키워드 네트워크에서의 해석 신뢰성을 보다 높여준다고 할 수 있다.

3.1.3 커뮤니티 텍스트 단어 네트워크

기업이 작성한 뉴스 텍스트와 달리, 소비자들이 작성한 커뮤니티 텍스트의 경우, 주로 게임 서비스를 효과적으로 이용할 수 있는 방안에 관한 것임을 확인할 수 있었다. 가장 핵심적으로 많이 언급되는 단어는 위에서 확인할 수 있듯, '선발, 투수, 기록, 시즌, 감독, 경기, 선수'이며, 이는 실제 게임 내 승리를 위한 전략과 매우 밀접하게 연관된다고 할 수 있다. 이를 바탕으로, 소비자들은 커뮤니티 텍스트를 통해, 서비스 이용과 관련된 지식, 노하우(Know-how), 전략 등을 활발하게 공유하고 있다고 정리해볼 수 있다. 이미 해당 게임을 사용하고 있는 소비자들이 주로 커뮤니티 내에서 의견을 공유하기 때문에, 뉴스 텍스트에 비해 게임 활용에 대한 보다 깊이 있는 의견 공유가 이루어지며, 선별적으로 정보를 채택한다는 것으로 이해할 수 있다[7].

3.1.4 커뮤니티 텍스트 워드 클라우드

소비자가 작성한 커뮤니티 텍스트로 워드 클라우드를 그려본 결과, [그림 7]와 같은 결과를 얻을 수 있었다. 상대적으로 다양한 단어들이 등장했던 [그림 5]의 뉴스 텍스트 워드 클라우드의 결과와 대조



[그림 6] 커뮤니티 텍스트 키워드 네트워크

[그림 7] 커뮤니티 텍스트 워드 클라우드

되는데, 소비자들은 뉴스 텍스트에 비해 제한된 단어를 활발하게 활용하고 있음을 알 수 있다. 구체적으로, 자주 언급되는 단어는 '선발, 투수, 감독, 수비, 성적, 우승, 기록' 등이 있으며, 이는 앞서 단어 네트워크 분석에서 확인한 바와 같이 소비자들이 서비스를 보다 효율적으로 이용하기 위해 필요한 지식, 노하우(Know-how), 전략 등과 밀접한 관련이 있다는 것을 알 수 있다. 이러한 결과는 앞서 소비자들이 커뮤니티를 바탕으로 서비스 이용과 관련된 지식 및 노하우를 공유한다는 결과의 신뢰성을 한층 더 높여준다고 할 수 있다. 또한 소비자들의 텍스트나 의견의 전파는 기업이 온라인 상에서 전통적인 마케팅 도구로서의 뉴스 기사나 언론을 통하여 활동하는 것에 비하여, 예외적인 흥미 위주의 이슈들을 대상으로 주로 이루어 진다는 것을 확인할 수 있는 것이다[4].

3.2 텍스트 데이터 토픽 분석

앞선, 제 3.1절 텍스트 데이터 시각화의 분석이 텍스트 상에 등장하는 단어들의 빈도수를 통하여 분석하여 상대적으로 단편적인 정보를 보여주는 분석이었다면, 제 3.2절에서 실시하는 텍스트 토픽 분석은 텍스트의 내용과 맥락을 반영하는 분석으로서 상대적으로 더 구체적이며 복합적인 분석 방법이라고 할 수 있다. 본 연구는 앞서 구성한 뉴스 텍스트 데이터와 커뮤니티 텍스트 데이터를 바탕으로, 해당 데이터에서 자주 언급되는 토픽을 추출하여 보다 구체적인 분석을 실시하였다.

3.2.1 뉴스 토픽 분석

뉴스 텍스트 데이터의 경우, 수집한 총 뉴스 기사의 개수가 약 500개 정도이며 중복적인 기사가 많았기 때문에, 토픽 분석을 할 때 풍부한 정보 바탕으로 다양한 토픽을 추출하기에 많은 수는 아니다. 하지만, 앞서 텍스트 키워드 시각화에서 살펴 보았듯이 키워드들은 기업의 이벤트 홍보 등에 집중되어 있으므로, 선정할 토픽의 개수를 조정한다면 기업이 뉴스 기사를 통하여 전달하고자 하는 메시지와 맥락을 충분히 파악할 수 있을 것으로 보았다. 때문에, 본 연구자들은 여러 번의 분석 시도를 통해 뉴스 기사 데이터에서 관련 토픽을 5개로 조정하는 것이 가장 합리적이라는 것을 알 수 있었다. 이를 바탕으로 LDA(Latent Dirichlet Analysis)를 실시하였다. 토픽을 추출하기 위한 분석의 총 반복(iteration)은 5,000번, alpha = 0.01, eta = 0.01의 값을 부여하여 분석을 실시하였다. 먼저, 뉴스 토픽과 관련된 토픽 추출 결과는 <표 1>과 같다.

추출된 토픽은 총 다섯 개이며, 토픽에서 주로 언급된 핵심 어구들을 바탕으로 본 연구자들은 위와 같은 레이블을 달아 두었다. 분석 결과 자체는 앞서 시각화 분석에서 살펴 본 바와 크게 다르지 않았는데, 분석된 총 다섯 개의 토픽 중 네 개가 이벤트 및 서비스 업데이트와 관련된 토픽이었기 때문이

〈표 1〉 뉴스 텍스트의 5대 토픽

토픽	Label	언급된 주요 단어
1	**업데이트 1** (모바일 앱 출시)	서비스, 모바일, 준비, 테스트, 데이터, 관심,
2	**이벤트 1** (올스타 및 레전드카드 출시)	아이템, 이벤트, 올스타, 레전드, 프리미엄, 업데이트, 지급, 서비스
3	이용자 대상 설문조사 결과	이용자, 설문조사, 대상, 프로야구단, 다이노스, 트윈스, 자이언츠, 라이온즈
4	**업데이트 2** (해설위원 관련 업데이트)	해설위원, 업데이트, 발표회, 콘텐츠, 준비, 심재구
5	**이벤트 2** (새로운 홍보모델(서유리))	서유리, 이벤트, 프리미엄, 업데이트, 아이템, 지급

다. 세 번째 토픽의 경우 '이용자 대상 설문 조사'와 관련된 토픽임을 알 수 있었는데, 이 토픽의 원본 텍스트를 역추적 해본 결과, 게임 운영사가 게임 상에서 설문조사 한 결과를 뉴스 기사로 많이 배포했기 때문에, 위와 같은 결과가 나온 것을 확인할 수 있었다. 이를 제외한 네 가지 토픽은 모두 업데이트 및 이벤트와 관련된 토픽들로, 구체적인 사항은 약간씩 다르지만, 뉴스 기사가 게임 운영사의 서비스 홍보 채널로서 활용되고 있다는 것을 일관되게 보여주고 있다.

3.2.2 커뮤니티 토픽 분석

토픽 분석을 대상을 바꾸어, 뉴스 텍스트가 아닌 소비자들이 만들어낸 커뮤니티 텍스트를 분석한 결과는 <표 2>와 같다. 커뮤니티 텍스트의 경우, 뉴스 텍스트보다 많은 약 5,000개의 문헌이 데이터로 포함되었는데, 이는 상대적으로 더 풍부한 정보를 포함하고 있기 때문에, 추출할 토픽의 수를 열 개로 조정하였다. 방법은 동일하게 LDA를 실시하여 분석하였고, 분석을 위한 옵션은 그대로 유지한 채 (iteration = 5,000번, alpha = 0.01, eta = 0.01) 분석을 실시하였다.

추출된 토픽은 총 열 개이며, 앞선 뉴스 토픽 분석과 마찬가지로, 해당 토픽에서 언급된 핵심 단어들을 바탕으로, 레이블을 표기하였다. 분석 결과 커뮤니티에서 언급된 토픽은 크게 세 가지 방향성을 갖고 있었는데, 1) 게임 전략과 관련된 토픽, 2) 이벤트와 관련된 토픽, 3) 실제 야구 이야기로 정리해 볼 수 있다. 먼저, 게임 전략과 관련된 토픽의 경우 압도적으로 많이 언급되고 있는데, 게임 서비스 자체가 다양한 전략을 시행할 수 있는 구조이기 때문에 관련 텍스트가 많이 생성되는 것으로 이해할 수 있다. 구체적으로 소비자들이 공격, 수비, 트레이드, 팀 관리 등과 같은 다양한 차원의 전략을 논의하고 있다는 것을 알 수 있다. 두 번째로, 소비자들은 게임 운영사가 실시하는 이벤트에 대해 주로 논의하고 있는 것도 확인할 수 있는데, 특히 이벤트를 통해 보상 받을 수 있는 '아이템'이라는 단어가 일관되게 언급되고 있음을 확인할 수 있다. 이러한 이벤트와 관련된 주제의 언급은 소비자들이 커뮤니티를

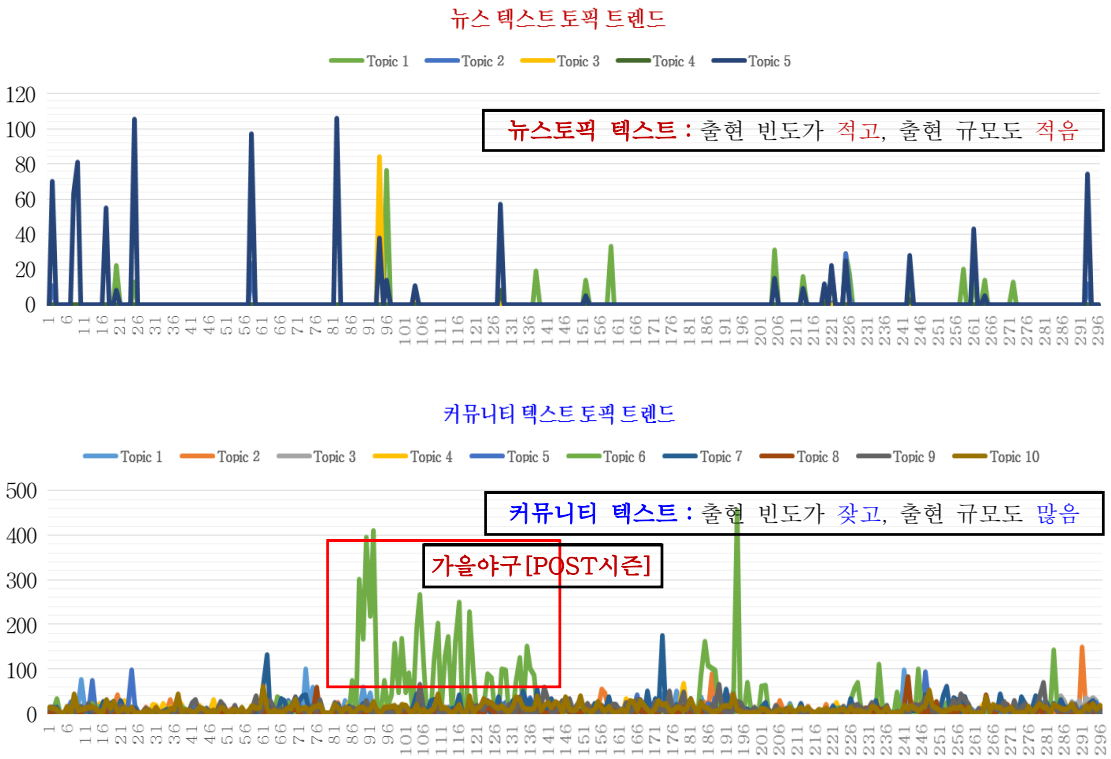
〈표 2〉 커뮤니티 텍스트의 10대 토픽

토픽	Label	언급된 주요 단어
1	**게임전략 1** (방어 위주의 전략)	선발진, 방어율, 마무리, 세이브, 라인업
2	**이벤트 관련 논의 1**	게시판, 이야기, 이벤트, 관련, 사이트, 아이템
3	**게임전략 2** 자신의 팀 구성 자랑	Php, 및 url link들, baseball, 이미지, 최저평점
4	**이벤트 관련 논의 2**	이벤트, 지급, 포인트, 이사회, 아이템, 업데이트
5	**게임전략 3** (넥센 히어로즈 팀 전략)	히어로즈, 넥센, 이택근, 김시진, 선발투수, 올시즌, 트레이드
6	**실제 야구이야기** (한국시리즈)	한국시리즈, 정규시즌, 트윈스, 와이번스, 자이언츠, 라이온즈, 타이거즈, 이글스
7	**게임전략 4** (공격 위주의 전략 논의)	타자, 타석, 타선, 지명타자, 외야수, 유격수, 타율
8	**게임전략 5** (매크로 활용)	매크로, 시뮬레이션, 시스템, 컨디션, 관리
9	**게임전략 6** (리그구성 및 전략 논의)	클래식, 챔피언, 위너스, 메이저, 올스타, 강등
10	**게임전략 7** (선수 채용 관련 논의)	재계약, 라이업, 마지막, 안나온다, 고코(설명 : 高 Cost = 실력이 좋은 선수를 뜻함)

통해 게임 서비스와 관련된 지식 및 노하우를 활발하게 공유하고 있다는 앞선 분석의 결과를 실증적으로 뒷받침해준다고 할 수 있다. 마지막으로, 하나의 토픽에서 현실 세계의 실제 야구에 관한 이야기들을 확인해 볼 수 있다. 이는 해당 게임 자체가 스포츠 게임인 만큼, 사용자들이 현실 프로 야구에도 많은 관심을 갖고 있기 때문인 것으로 생각된다.

3.2.3 시계열 토픽 분석(Time-series Topic Analysis)

분석한 토픽 관련 단어들의 출현 빈도를 점수화하여, 시계열적으로 표현하면 다음과 같은 [그림 8]를 그려볼 수 있다. 추출된 토픽들의 출현을 구체적으로 살펴 보면, 시간의 흐름에 따른 토픽 트렌드를 파악할 수 있는데, 뉴스 토픽들은 커뮤니티 토픽에 비해 간헐적으로 출현하며, 출현하더라도 언급되는 단어의 수가 상대적으로 적다는 것을 확인할 수 있다. 반면, 커뮤니티 토픽의 경우 거의 대부분의 토픽이 지속적으로 출현하고 있어, 출현 빈도가 매우 잦으며, 출현 규모도 특정 토픽의 경우 매우 활발하게 언급되는 것을 알 수 있다. 소비자들의 커뮤니티에서의 텍스트 활동이 기사를 통한 기업의 마케팅 활동에 비하여 비용 소모도 적고, 상황이나 환경 변화에 즉각적으로 반응하기 때문으로 파악된다. 일례로, 커뮤니티 토픽 중 Topic 6(현실 야구에 관한 이야기)의 경우, 특정 기간에 매우 두드러진 출현 빈도를 보여주고 있는데, 해당 기간을 역추적 한 결과, 프로 야구 팀 간의 경쟁이 심화되고, 프로 리그가 후반부로 접어드는 '가을 야구' 시즌임을 확인할 수 있었다.



주) 1. y축은 날짜(day)를 나타내며, x축은 토픽의 출현 빈도를 의미한다.
2. 뉴스 텍스트 토픽과 커뮤니티 텍스트 토픽은 각각 <표 1>과 <표 2>에서 나타난 토픽과 같다.

[그림 8] 텍스트 토픽 트렌드

3.3 기업 성과 예측 분석

3.3.1 기업 성과 예측 모형 분석 결과

본 연구는 앞서 분석한 토픽들을 변수화 하여, 기업 및 소비자가 만들어내는 텍스트가 실제 기업 성과에 어떠한 영향을 미치는지 구체적으로 탐구해 보고자 한다. 앞서 언급한 바와 같이, 뉴스 데이터에서 추출된 토픽들은 출현 빈도가 매우 간헐적이며, 생성되는 양도 많지 않다. 또한 기존 연구에서도 밝히고 있듯이, 소비자들의 구전은 기업의 전통적 활동에 비하여 기업 성과에 더 유의미하고 긍정적인 영향을 준다는 것을 알 수 있다[14, 23]. 그러므로 기업 성과에 영향을 주는 텍스트의 예측력을 높이고자 소비자들이 발생시킨 텍스트 토픽들만을 분석 모형에 포함하였다. 따라서, 본 연구는 소비자들의 텍스트 토픽을 중심으로 기업 성과에 대한 예측을 실시하고자 한다. 이를 위해, 앞서 식 (1)에서 소개한 성과 예측 모형이 설계되었으며, 이에 따른 모형 추정 결과는 다음과 같다.

추정 방법은 MLE(Maximum Likelihood Estimation) 방식을 따랐으며, 본 연구의 핵심은 소비자들이 언급한 텍스트 토픽 중 기업 성과에 유의한 영향을 미치는 변수를 확인하고 이를 구체적으로 해석하는 것이다. 위 표에 굵은 글씨체로 표기해 두었듯, 앞서 확인한 총 10개의 텍스트 토픽 중 4개가 기업 성과에 유의한 영향을 미치는 것으로 확인되었다. 해당 변수들은 텍스트 토픽 중 모두 '게임 전략'과 관련된 것이라는 점에서 공통점이 있으나, 그 세부적인 영향의 방향은 다르다.

먼저, 게임 전략 1의 경우 '선발진, 마무리, 투수, 방어, 방어율' 등의 단어들이 주로 언급되는 토픽이라는 점에서 "방어"와 관련된 게임 전략이라고 정리해 볼 수 있다. 반면, 게임 전략 4의 경우 '타자, 타석, 타순, 지명타자, 타율' 등 실제 게임 플레이 중 "공격"과 관련된 게임 전략이라 정리해 볼 수 있다. 여기서 흥미로운 점은 해당 토픽들의 성과에 관한 영향이 상반된다는 점인데, 방어와 관련된 텍스트 토픽의 경우 실제로 기업 성과 (사용자들의 게임 아이템 구매)에 부정적인 영향(-0.0021, $p < 0.10$)을 미치는 것이 확인된 반면, 공격과 관련된 텍스트 토픽의 경우 기업 성과에 긍정적인 영향(0.0027, $p < 0.01$)을 미치는 것이 확인되었다. 이는 동일한 게임 전략이라 할지라도, 상대적으로 현 상태를 유지하고, 방어하는 등의 보수적인 전략은 게임 아이템 구매를 억제하는 반면, 현 상태에 변화를 꾀하는 공격적인 전략의 경우 게임 아이템 구매를 촉진한다고 해석

〈표 3〉 기업 성과 예측 모형 분석 결과

변수	추정계수	표준오차	p-value
절편	17.2902	0.2464	< .0001
통제 변수 1 : 게임 접속자 수	0.0001	0.0000	< .0001
통제 변수 2 : 1인당 평균 게임 시간	0.0070	0.0038	0.0677
텍스트 토픽 변수 1 : 게임전략 1	**-0.0021**	**0.0012**	**0.0756**
텍스트 토픽 변수 2 : 이벤트 관련 논의 1	0.0009	0.0014	0.5341
텍스트 토픽 변수 3 : 게임전략 2	-0.0083	0.0019	< .0001
텍스트 토픽 변수 4 : 이벤트 관련 논의 2	0.0005	0.0022	0.8362
텍스트 토픽 변수 5 : 게임전략 3	**-0.0044**	**0.0015**	**0.0034**
텍스트 토픽 변수 6 : 실제 야구이야기	0.0003	0.0002	0.1251
텍스트 토픽 변수 7 : 게임전략 4	**0.0027**	**0.0010**	**0.0065**
텍스트 토픽 변수 8 : 게임전략 5	-0.0004	0.0021	0.8642
텍스트 토픽 변수 9 : 게임전략 6	-0.0020	0.0015	0.1854
텍스트 토픽 변수 10 : 게임전략 7	0.0008	0.0020	0.6951

해 볼 수 있을 것이다. 즉, 공격에 관련한 적극적인 내용을 담고 있는 텍스트는 소비자들이 상대적으로 긍정적인 콘텐츠로서 받아들이며, 이 경우 기존 연구에서 밝히고 있는 바와 같이 기업 성과를 증진시킬 수 있다[12]. 하지만, 그 반대의 유지 및 방어와 관련된 상대적으로 소극적이며 부정적인 내용을 담고 있는 콘텐츠와 관련해서는 기업 성과에 부정적인 영향을 미친다고 볼 수 있다.

게임 전략 2의 경우 주로 자신의 팀 구성을 자랑하는 토픽으로, 텍스트와 함께 첨부된 이미지 html tag, url 주소 등이 언급되는 것을 알 수 있었다. 이는 원본 텍스트를 역추적해본 결과 사용자들이 커뮤니티 게시판에 자신의 팀 구성을 캡쳐한 이미지를 공유하고, 실제 자신의 팀 구성을 확인해 볼 수 있는 url을 올려 놓기 때문인 것으로 확인되었다. 자신의 팀을 자랑하는 이 토픽의 경우 사용자들의 게임 아이템 구매에는 부정적인 영향을 미치는 것이 확인되었다(-0.0083, $p < 0.01$). 이는 해당 토픽이 담고 있는 텍스트 정보의 속성과 연관 지어 생각해 볼 수 있는데, 자신의 팀 구성 자체를 자랑하는 사람들의 경우, 일반적으로 매우 잘 관리된 팀 또는 매우 성과가 좋은 팀의 모습을 보여준다는 점에 주목할 필요가 있다. 이렇게 잘 관리된 팀 구성을 다른 사용자들에게 보여주는 경우, 이는 게임 내 더 좋은 성과를 확보할 수 있는 일종의 지침, 전략, 지식이라 할 수 있으며 이러한 지식의 전파는 게임 내 성과 및 실력을 높여줄 수 있는 게임 아이템의 구매를 대체하는 역할을 할 것이라 예상해 볼 수 있다. 이는 게임 내 특정 팀과 연관된 토픽(게임 전략 3)이 기업 성과에 미치는 영향(-0.0044, $p < 0.01$)도 설명할 수 있는데, 특정 팀에 관한 운영 전략이 폭넓게 공유되는 경우, 이 자체가 게임 아이템 구매 없이도 보다 좋은 성과를 낼 수 있도록 도움을 준다고 볼 수 있다. 즉, 팀 운영에 관한 지식 역시, 게임 아이템 수요를 대체하여, 사용자들의 게임 아이템 구매를 억제하는 효과를 낼 수 있다는 것이다.

다만 이벤트와 관련된 논의는 기업 매출을 설명하는데 있어, 유의한 영향을 주지 않는 것으로 나타났다. 일반적으로 이벤트가 할인과 결합된다는 점을 생각해보면, 이는 의아한 결과일 수 있지만 게임 이벤트가 매우 잦기 때문인 것으로 생각된다. 실제로, 해당 게임에서 실시하는 이벤트를 역추적한 결과 한 달 평균 5~6회 정도의 이벤트가 상시적으로 진행되고 있었고, 이는 곧 사용자들이 이벤트에 대해 반응하는 정도가 크게 민감하지 않은 것으로 예상된다. 또한 실제 야구 이야기의 경우, 게임 내 매출과 크게 관련이 없는 것으로 나타났는데, 이는 동일한 야구라 할지라도, 가상의 게임과 현실의 게임은 서로 영향을 미치지 못한다는 것을 의미한다. 다시 말해, 야구 게임을 즐기는 사람들이 현실의 야구에 대한 언급을 활발히 하는 것은 개인들의 단순한 선호와 관심 때문이지, 이 자체가 게임 내 성과를 좌우하는 것은 아니라는 것이다. 마지막으로, 일부 게임 전략(매크로, 리그 구성, 선수 관리)도 게임 매출과 유의하지 않은 결과를 확인할 수 있었는데, 이 내용들은 다른 게임 전략 주제(공격/수비/관리 등)에 비해 다소 부가적인 측면에 치중한 전략이기 때문에, 게임 아이템 구매와는 크게 영향 관계가 없는 것으로 추정된다.

3.3.2 기업 성과 예측 모형 유의성 검정

앞서 예측한 기업 성과 모형의 결과 해석과는 별개로, 텍스트 분석을 포함한 모형 자체가 통계적으로 의미가 있는지를 검증하는 것 역시 필요하다. 다시 말해, 기존에 게임 운영사가 활용할 수 있는 변수들만 가지고 기업 성과를 예측하는 모형과 본 연구에서 활용한 텍스트 토픽 변수들을 추가한 모형을 비교하여, 설명력 자체가 유의하게 증가되는지 검증해볼 필요가 있다. 이러한 검증을 통하여, 텍스트 마이닝 분석이 기업 성과 예측에 의미있는 마케팅 도구로서 활용될 수 있음을 확인할 수 있다. 이를 위해, 본 연구는 카이제곱 적합성 검정(Chi-square goodness of fit)을 바탕으로 기본 모형(모형 (1))과 텍스트 모형(모형 (2))를 비교하고, 10-묶음 교차검증법(10-fold cross validation)을 통하여, 모형의 유의성을 검증하도록 한다.

먼저 카이제곱 적합성 검정 결과는 다음과 같은데, 앞서 언급했듯, 모형의 추정 자체는 MLE 방식으로 구성되어 있기 때문에, 추정 후 각 모형의 로그 가능도(Log likelihood) 값을 얻을 수 있다. 이를 바탕으로, 본 연구는 1) 하루에 접속한 이용자의 수, 2) 이용자 1명 당 게임 시간 변수만 활용한 기본 모형(모형 (1))의 로그 가능도(Log likelihood)와 텍스트 토픽 변수까지 반영하여 복합적으로 구성한 텍스트 토픽 모형(모형 (2))의 로그 가능도(Log likelihood)를 비교하여, 텍스트 변수를 포함한 모형의 설명력이 유의하게 향상되었는지를 확인한다. 실제 검증은 두 모형의 로그 가능도(Log likelihood) 차이를 구하고 -2를 곱하여, 카이제곱 검정을 실시하는 것이다. 검정 결과는 <표 4>에 나타나 있으며, 두 모형의 예측력이 같다는 귀무가설이 매우 유의한 수준에서 기각된다는 것을 알 수 있었다. 이를 통해, 본 연구가 제안하는 텍스트 토픽 변수들은 기업 성과를 예측하는 데 있어 설명력을 유의하게 높여준다고 말할 수 있다.

다음으로, 본 모형이 일부 데이터에 의해 왜곡된 것이 아님을 10-묶음 교차 검증법(10-fold cross validation)을 통해 확인하였다. 구체적으로, 총 300일 분량의 데이터를 무작위로 10개 그룹으로 나누고, 9개의 그룹으로 모형을 만들어 나머지 1개 그룹의 성과를 예측하는 분석을 실시하였다. 이와 관련하여, 사용되는 식은 식 (3)과 같다[1]. 본 모형의 관심 종속 변수는 매출(성과)이며, 연속형 변수이기 때문에, 연속형 변수의 적합성 검증을 하는데 사용되는 RMSE와 MAE 지표를 가지고 본 모형의 설명력이 개선되는지를 확인하였다. 검증 결과는 <표 5>에 나타나 있다. 10-묶음 교차 검증 분석 결과, RMSE와 MAE 지표 모두에서 오차의 크기가 줄어들었음을 확인할 수 있었다. 이는 기업 성과를 예측하는데 있어, 텍스트 토픽 분석을 변수로서 포함하여 분석하면, 예측력이 텍스트 토픽 모형을 포함하지 않았을 때 보다 개선된다는 것을 의미한다.

$$\sum_i (Y_i - a - b_1 X_{1i} - \cdots - b_p X_{\pi})^2 / n$$

〈표 5〉 10-묶음 교차 검증 결과

	모형 (1)	모형 (2)
RMSE (Root Mean Square Error)	0.3620	0.3462
MAE (Mean Absolute Error)	0.2684	0.2551

4. 결론 및 시사점

본 연구는 기본적으로 텍스트 데이터를 분석하되, 이 텍스트와 관련된 단어와 토픽이 생성 주체

〈표 4〉 적합성 검정 결과

	모형 (1)	모형 (2)
변수 구성	Number of Gamers, Game Time per Gamer	Number of Gamers, Game Time per Gamer +**Text Topic Index**
로그 가능도 (Log likelihood)	127.6964	103.5586
자유도	286	276
카이제곱통계량 (Chi square statistic)	48.2756	
자유도 차이	10	
검정 결과	귀무가설 기각[a]	

주) [a] p-value < 0.001.

(기업 및 소비자)에 따라 어떻게 다른지 탐구하였다. 이를 위해, 자주 언급되는 핵심어를 시각적으로 표현하기 위하여, '단어 동시 출현 네트워크'와 '워드 클라우드'를 그려 전반적인 텍스트의 다양한 이슈를 파악하였다. 또한 보다 구체적으로 내용의 맥락을 파악하기 위하여, LDA 방식의 토픽 분석을 실시하였으며, 기업이 만들어낸 텍스트와 소비자가 만들어낸 텍스트의 의미 있는 차이를 진단해 보았다. 마지막으로, 본 연구는 텍스트 토픽 분석에서 확인한 토픽 출현 빈도를 시계열 매출 데이터와 결합하여, 앞서 선행한 특정 텍스트 토픽의 언급이 기업 성과(매출)에는 어떠한 영향을 미치는지 탐구하였다. 본 연구의 핵심적인 발견점은 다음과 같이 정리해볼 수 있다.

첫째, 본 연구는 기업이 생성하는 텍스트와 소비자가 생성하는 텍스트의 유의미한 차이를 확인할 수 있었다. 구체적으로 기업이 생성하는 텍스트는 주로 기업 서비스와 관련하여 이벤트 및 업데이트를 홍보하는 수단으로 활용되고 있었다. 이는 대부분의 기업들이 보도자료의 형식을 통해, 뉴스 기사를 송출하고 언론사들은 이를 크게 편집하지 않고 뉴스 기사로 활용하기 때문인 것으로 보인다. 반면, 소비자 텍스트는 실제 서비스 이용과 관련된 전략, 지식, 노하우(Know-how)를 공유하는 주제가 빈번하게 언급되었다. 이는 소비자들의 적극적인 속성을 매우 잘 나타내는 사례라 할 수 있으며, 이들이 단순히 기업이 제공하는 제품 및 서비스를 이용하기 보다 이와 관련된 자신의 경험, 지식, 의견을 온라인을 통해 활발히 공유한다는 점을 시사한다. 둘째, 앞서 밝혔듯 본 연구가 핵심적으로 진단하고자 하는 것은 기업이 텍스트 분석을 통해 매출을 효과적으로 예측할 수 있는지 여부이다. 이를 위해, 앞서 분석한 뉴스 토픽과 커뮤니티 토픽을 매출 예측에 활용하였는데, 이 중 커뮤니티 토픽(소비자 생성 토픽)만이 매출 예측에 유의미한 값을 갖는 것으로 확인되었다. 이는 기업 텍스트 자체가 간헐적으로 발생하는 점도 있으나, 기업에서 생성하는 텍스트 자체가 매출의 근원이라 할 수 있는 소비자들에게 유의미한 영향을 미치지 못하기 때문인 것으로도 생각된다.

셋째, 소비자들이 생성하는 텍스트 토픽은 매출 예측에 유의미한 설명력을 갖는 것으로 확인되었다. 구체적으로, 소비자들이 생성하는 텍스트 토픽은 매우 다양하나, 그 중 실제 게임 서비스 이용과 관련된 지식(게임 전략)에 관한 토픽이 매출 예측에 유의한 영향을 미치고 있었다. 그러나 영향의 방향성은 구체적인 토픽의 속성에 따라 달라지는데, 같은 게임 전략이더라도, 팀을 방어/유지/관리 하는 토픽들은 게임 아이템을 대체하는 효과를 보여 매출에 부정적인 영향을 미치는 반면, 공격 및 발전에 관한 텍스트 토픽들은 해당 공격을 보다 적극적으로 할 수 있는 게임 아이템에 대한 추가 구매를 유도하는 것으로 확인되었다. 이는 적극적이며 긍정적인 텍스트 콘텐츠는 기업 성과에 긍정적인 영향을 주지만, 소극적이며 부정적인 텍스트 콘텐츠는 기업 성과에 부정적인 영향을 준다는 기존 연구의 일관된 결과임을 보여준다[12].

본 연구의 분석 결과는 전통적인 기업 활동 및 소비자 구전 활동과 관련된 기존 연구에 대한 이해를 높일 뿐만 아니라, 매출에 직접적인 영향을 받을 수 있는 기업에게 실무적인 도움을 전달해줄 것으로 기대된다. 이에 따라 본 연구가 갖는 시사점은 다음과 같이 정리해볼 수 있다. 첫째, 본 연구는 온라인 상에서 소비자들이 무엇에 대해 이야기하고 있는지 분석하는 것은 기업의 입장에서 충분히 의미있는 관찰이라는 것을 보여주었다. 기초적인 코딩만으로도 소비자들의 커뮤니티에 접속하여 제품 및 서비스와 관련된 소비자들의 의견을 취합할 수 있었는데, 이러한 분석은 적은 비용으로도 소비자들의 의견을 손쉽게 파악할 수 있다는 점에서 매우 유용할 것이다. 실제로 많은 기업들이 자사의 제품 및 서비스와 관련된 소비자들의 의견을 파악하기 위해 FGI, 설문조사, 패널 조사 등 매우 큰 비용을 지출하고 있다는 점을 감안한다면 소비자 의견을 추적하는 '텍스트 마이닝' 기법은 매우 획기적으로 비용을 줄일 수 있는 시장 조사 방법 중의 하나일

것이다. 다시 말해, 상대적으로 활용할 수 있는 자산이 제약되어 있는 소상공인이나 스타트업을 하고자 하는 청년들도 텍스트 마이닝 기법을 통하여 보다 쉽게 소비자들의 니즈를 파악할 수 있다는 것을 의미한다. 특히, 본 연구가 사용한 KAIST 한나눔 분석기는 무료로 배포되고 있다는 점을 감안하면, 창업자들이 이와 관련된 지식을 조금만 학습한다면, 고객들의 반응 및 평가를 손쉽게 분석할 수 있어 실무적으로 매우 큰 도움을 얻을 수 있을 것이다.

둘째, 앞서 분석하였듯이 토픽 모델링을 활용한 매출 예측 모형은 성과 예측 오차를 줄여주고 보다 정교한 예측을 가능케 해준다는 장점이 있다. 뿐만 아니라, 특정 텍스트 토픽이 매출 예측에 미치는 영향을 진단하여, 소비자들이 어떠한 텍스트에 대해 생성시킬 때 실제 기업 성과에 영향을 미치는지 구체적으로 확인해 볼 수 있다. 나아가, 기업 실무자는 이러한 결과를 바탕으로 매출 예측에 부정적인 영향을 미칠 수 있는 요인들을 미리 파악하고, 해당 요인이 심각한 문제를 일으키기 전에 미리 선제적으로 대응할 수 있다. 구체적으로, 텍스트 마이닝 분석을 활용한 기업 성과 예측은 관련 정보를 대시보드(dashboard)에 반영하여, 예상 매출을 예측해 보고 발생할 수 있는 위험 요소들을 사전에 차단할 수 있도록 하는 전략으로서 활용될 수 있다. 셋째, 본 연구에서 사용한 방법과 도출한 결과는 온라인 채널을 활용하여 제품과 서비스를 제공하는 다양한 기업에게 시사하는 바가 크다. 본 연구의 분석 대상인 온라인 게임서비스는 다른 카테고리에 비하여 상대적으로 이용자(기업 및 소비자)가 활발하게 텍스트를 생성하며 즉각적인 영향을 주고받는 카테고리이다. 이러한 카테고리를 대상으로 분석한 결과는 온라인 채널을 활용하며 커뮤니케이션 하는 다른 많은 카테고리의 기업들에게도 텍스트 마이닝이 효율적이고 의미 있는 마케팅 활동의 도구로서 활용 가능하다는 것을 보여준다. 특히 온라인 채널을 유통 및 마케팅 플랫폼으로 활용하는 다양한 업종의 소상공인과 청년 창업에도 큰 도움을 줄 것으로 기대한다.

참 고 문 헌

[1] 김민철, 심규승, 한남기, 김예은, 송민, “트위터 상의 악의적 이용 자동분류”, 「한국문헌정보학회지」, 제47권, 제1호(2013), pp.269-286.

[2] 김승경, 이재관, “웹 사이트 디자인의 시각적 요소와 유용성이 성과에 미치는 영향에 관한 연구”, 「한국경영과학회지」, 제32권, 제2호(2007), pp.17-40.

[3] 김승운, 강희택, “온라인 피드백 매커니즘으로서 상품평 게시판의 지각된 효과성과 신뢰, 만족, 이용의도간의 관계구조분석”, 「한국경영과학회지」, 제32권, 제1호(2007), pp.53-69.

[4] 김은미, 이주현, “뉴스미디어로서의 트위터”, 「한국언론학보」, 제55권, 제6호(2011), pp.152-180.

[5] 박자현, 송 민, “토픽 모델링을 활용한 국내 문헌정보학 연구동향 분석”, 「정보관리학회지」, 제30권, 제1호(2013), pp.7-32.

[6] 이동일, 김현교, “모바일, PC 온라인 매체 방문행동이 쇼핑 사이트 방문에 미치는 영향에 대한 동태적 연구”, 「한국경영과학회지」, 제39권, 제4호(2014), pp.85-95.

[7] 진설아, 허고은, 정유경, 송 민, “트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구”, 「정보관리학회지」, 제30권, 제1호(2013), pp.285-302.

[8] Archak, N., A. Ghose, and P.G. Ipeirotis, “Deriving the pricing power of product features by mining consumer review,” *Management Science*, Vol.57, No.8(2011), pp.1485-1509.

[9] Berger, J. and K.L. Milkman, “What makes online content viral?,” *Journal of Marketing Research*, Vol.49, No.2(2012), pp.192-205.

[10] Bickart, B. and R.M. Schindler, “Internet forums as influential sources of consumer information,” *Journal of Interactive Marketing*, Vol.15, No.3(2001), pp.31-40.

[11] Chen, P., S. Wu, and J. Yoon, “The impact of

online recommendations and consumer feedback on sales," *ICIS Proceedings*, Vol.58 (2004), pp.711-724.

[12] Chevalier, J. and D. Mayzlin, "The effect of word of mouth on sales : online book review," *Journal of Marketing Research*, Vol.43, No.3 (2006), pp.345-354.

[13] Duan, W., B. Gu, and A.B. Whinston, "The dynamics of online word-of-mouth and product sales-an empirical investigation of the movie industry," *Journal of Retailing*, Vol.84, No.2(2008), pp.233-242.

[14] Godes, D. and D. Mayzlin, "Using online conversations to study word of mouth communication," *Marketing Science*, Vol.23, No.4 (2004), pp.545-560.

[15] Godes, D. and D. Mayzlin, "Firm-created word-of-mouth communication : evidence from a field test," *Marketing Science*, Vol.28, No.4 (2009), pp.721-739.

[16] Greene, W.H., *Econometric Analysis*, New York : McMillian, 5th ed., 2003.

[17] Griffiths, T. and M. Steyvers, "Finding science topics," *Proceedings of the National Academy of Science of the United States of America*, Vol.101, No.1(2004), pp.5228-5235.

[18] Huang, P., N.H. Lurie, and S. Mitra, "Searching for Experience on the Web : An Empirical Examination of Consumer Behavior for Search and Experience Goods," *Journal of Marketing*, Vol.73, No.1(2009), pp.55-69.

[19] Liu, Y., "Word of mouth for movies : its dynamics and impact on box office revenue," *Journal of Marketing*, Vol.70, No.3(2006), pp.74-89.

[20] Naik, P.A., "Estimating the Half-life of Advertisments," *Marketing Letters*, Vol.10, No.4 (1999), pp.345-356.

[21] Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business : market-structure surveillance through text mining," *Marketing Science*, Vol.31, No.3(2012), pp.5211-543.

[22] Pauwels, K., E.C. Stacey, and A. Lackman, "Beyond Likes and Tweets : Marketing, Social Media Content, and Store Performance," *Marketing Science Institute*, Vol.13(2013), pp. 13-125.

[23] Srinivasan, S.S., R. Anderson, and K. Ponnavolu, "Customer loyalty in e-commerce : an exploration of its antecedents and consequences," *Journal of Retailing*, Vol.78, No.1 (2002), pp.41-50.

[24] Villanueva, J., S. Yoo, and D.M. Hanssens, "The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth," *Journal of Marketing Research*, Vol.45, No.1(2008), pp.48-59.