「정보시스템연구」 제31권 제3호
한국정보시스템학회
2022년 9월, pp. 67~87
http://dx.doi.org/10.5859/KAIS.2022.31.3.67

BERT를 활용한 미국 기업 공시에 대한 감성 분석 및 시각화

김효곤* · 유동희**

〈목 차〉

Ⅰ. 서 론

연간 보고서, 분기 보고서와 같은 기업 공시는 경영 전반에 대한 기업의 견해를 담은 공식 문서로서 주주, 투자자 등에게 유용한 정보를 제공한다. 미국의 경우 공개 기업(Public Company)은 Form 10-K(연간 보고서), Form 10-Q(분기 보고서)를 작성하여 미국 증권거래위원회(United States Securities and Exchange Commission, SEC)에 정기적으로 제출해야 한다. SEC 보고서에는 기업의 재무제표와 같은 정량적 데이터뿐만 아니라 사업 개요, 업황, 인식한 위험 요인, 향후 계획 등 기업 관점에서 기술한 정성적 데이터를 포함하고 있으므로 기업 분석을 위한 가치 있는 데이터로 간주되고 있다(Stephany et al., 2020).

대·내외적으로 큰 경영 변화를 겪고 있는 기업은 투자자의 올바른 의사결정 지원을 위해 해당 기업과 관련된 최신 정보를 제공해야 하는 윤리적 책임을 가지고 있다(Yuthas et al., 2002). 산업계 전반에 충격을 준 COVID-19 팬데믹에 대응하여 많은 기업이 COVID-19 팬데믹에 대한 견해를 기업 공시에 신속하게 담았다. Larcker et al.(2020)의 조사에 따르면 당시 수집한 3,644개 SEC 보고서에서 COVID-19가 언급된 건이 2020년 1월 4.8%, 2월 35.5%, 3월 76.6%, 4월 100%, 5월 99.9%로 가파르게 증가하였다. 따라서 2020~2021년에 공시된 SEC 보

* 경상국립대학교 기술경영학과, hyogon.kim@gnu.ac.kr(주저자)
** 경상국립대학교 경영정보학과 및 경영경제연구소, dhyoo@gnu.ac.kr(교신저자)

고서를 분석한다면 COVID-19 팬데믹에 대한 미국 기업의 견해를 시간 흐름에 따라 살펴볼 수 있을 것으로 판단된다.

그러나 SEC 보고서는 구조화되지 않은 문서, 즉 비정형 데이터로서 방대한 양의 수치 자료와 문자로 구성되어 있다. SEC 보고서, 뉴스, 인터넷 게시물과 같은 비정형 텍스트 데이터로부터 패턴, 관계 등 유용한 정보를 추출하는 것을 텍스트 마이닝 또는 텍스트 데이터 마이닝이라 한다(Ignatow and Mihalcea, 2017). 텍스트 마이닝 기술 중 하나인 감성 분석(Sentiment Analysis)은 텍스트로부터 정서적 상태, 주관적 정보 등을 체계적으로 추출, 식별, 분류, 정량화하는 것을 목표로 한다.

감성 분석은 소셜 미디어(SNS) 상의 소비자, 고객의 경험, 주장과 같이 원천적으로 감성을 표현하기 위해 생산된 데이터를 긍정, 부정 등으로 분류하여 기업 경영에 활용하는 것이 일반적이며 널리 인기를 얻고 있다(홍태호 등, 2018; Capuano et al., 2021). 하지만 SEC 보고서, 기업의 사회적 책임(CSR) 보고서와 같이 작성자가 의견 및 감성을 덜 명시적으로 표현하는 문서에 대한 감성 분석도 시도되고 있다. 예를 들어 SEC 보고서는 실적 예측(Azimi and Agrawal, 2021), 재무 위험 예측(Wang et al., 2013), 부실 기업 식별(Gandhi et al., 2019) 등 기업 평가를 위한 감성 분석 연구에서 활용되었다. 또한 CSR 보고서의 내용을 감성 분석하여 총자산이익률(ROA)과 같은 지표 값을 예측하기도 하였다(Che et al., 2020; Myšková and Hájek, 2018).

효과적인 데이터 분석을 위해 감성 분석을 포함한 텍스트 마이닝의 분석 기법도 점차 진화하고 있다. 특히 딥러닝과 자연어 처리(Natural Language Processing, NLP)의 발전으로 인해 다양한 분류, 예측 문제에서 괄목할 만한 성과가 나타나고 있다. 최근에는 문장의 전후 관계, 상황과 같은 문맥적 의미를 고려하는 BERT(Bidirectional Encoder Representations from Transformers)와 같은 언어 모델을 활용하여 보다 복잡한 감성 분류 문제에 도전하는 연구들이 수행되고 있다.

한편 오늘날의 빅데이터 시대에는 기초 데이터가 방대함에 따라 감성 분석을 비롯한 데이터 분석에 있어 분석 과정의 효율성과 더불어 분석 결과에 대한 효과적 정보 전달이 강조되고 있다(Ali et al., 2016; Qin et al., 2020). 이때 다양한 시각적 요소의 조합을 통한 데이터 시각화는 분석 결과의 의미를 보다 명확하게 전달해 주어 사람들이 데이터를 쉽게 이해하도록 하며 데이터에 숨겨진 추세, 패턴, 이상값 등을 더욱 쉽게 식별할 수 있도록 돕는 것으로 알려져 있다(Heer et al., 2010).

현재 텍스트 마이닝 및 NLP 기술을 통한 감성 분석의 활성화에도 불구하고 SEC 보고서와 같이 공식적인 문서에 대한 감성 분석 연구는 상대적으로 적다. 더욱이 문서에 함축된 감성이 시간의 흐름에 따라 어떻게 변화하는가를 밝힌 연구 또한 많지 않다. 본 연구는 COVID-19 팬데믹에 대한 미국 기업의 견해와 평가를 감성과 연관 지어 정량적으로 측정하고, 다양한 방법으로 시각화하여 기존 연구 공백을 채움은 물론 투자자, 소비자, 행정가 등에게 시사점과 통찰을 제공하고자 한다. COVID-19 팬데믹과 같이 통제할 수 없으며 전방위로 영향을 미치는 사건에 대한 기업의 반응과 반응의 변화를

탐색하는 것은 개별 기업은 물론 산업 전반을 이해하는데 도움이 될 것이다.

연구의 목적 달성을 위한 본 연구의 세부 과제는 다음과 같다. 첫째, SEC 보고서에 드러난 COVID-19 팬데믹 관련 텍스트를 정량화한다. 둘째, COVID-19 팬데믹 텍스트를 감성 분석하고 시간의 흐름에 따른 변화를 추적한다. 셋째, COVID-19 팬데믹으로 인한 감성의 분포가 업종에 따라 차이를 보이는지 살펴본다. 여기에서 연구의 대상 및 시간적 범위는 COVID-19 팬데믹의 전 세계적 확산이 시작된 2020년 1월 1일부터 2021년 12월 31일까지의 기간 중 미국 SEC에 제출 및 공시된 Form 10-K와 Form 10-Q로 하였다. 기술적인 측면에서는 BERT 언어 모델, 구체적으로 금융 분야 데이터의 분석을 목적으로 Fine-tuning 한 FinBERT(Araci, 2019)를 활용하였다. 본 연구의 데이터 처리, 분석, 시각화는 모두 파이썬(Python) 및 관련 라이브러리를 사용하여 구현하였다.

Ⅱ. 문헌 연구

2.1 미국 기업 공시 개요 및 특징

SEC는 1934년 증권 거래법(Securities Exchange Act of 1934)을 근거로 설립되었다. 투자자를 보호하고 공정하며 질서 있는 증권 시장을 유지하기 위해 규정을 만들고 모든 공개된 기업이 규정에 따라 SEC 보고서를 작성 및 공시하도록 관리, 감독하고 있다(SEC, 2022b). SEC 보고서는 작성 목적에 따라 수십여 종의 서식이 있으며 이 중 Form 10-K와 Form 10-Q는 재무 및 투자 관점에서 자주 언급되었다. Chi and Shanthikumar(2018)는 개인 투자자의 거래는 구글 검색량보다 Form 10-K와 Form 10-Q의 검색량과 관련이 더 높음을 주장하였다. 투자 목적 외에도 SEC 보고서는 연구 자료로서 상당한 가치를 가지는 것으로 평가되고 있다(Stephany et al., 2020; Yuthas et al., 2002).

Form 10-K는 연간 보고서(Annual Report)라고도 불리며 감사를 마친 재무제표는 물론 상당한 양의 정성적, 정량적 기업 정보가 종합적으로 기술되어 있다. 기업에 따라 Form 10-K와는 별도로 사진, 그림 등을 포함하여 심미적으로 디자인한 '연간 보고서'란 제목의 별도 책자를 발간하기도 한다. 연간 보고서와 Form 10-K의 내용은 동일하지 않을 수 있으나 보통 Form 10-K가 더 많은 정보를 담고 있다(SEC, 2022a). 일반적으로 Form 10-K는 Item 1~15까지의 항목으로 구성되어 있으며, 기업의 회계 연도 마감 후 작성 및 공시되어야만 한다. 학계에서는 다른 공시 자료에 비해 많은 내용을 담고 있는 Form 10-K를 대상으로 한 연구가 활발한 편이다.

Form 10-Q는 분기 보고서(Quarterly Report)라고도 불리며 기업은 회계 연도 마감 분기를 기준으로 직전 분기까지 매년 3회 작성 및 제출해야 한다. Form 10-K와 비교 했을 때 재무제표에 대한 감사를 받지 않는다는 차이가 있으며 정보의 양 또한 상대적으로 적은 편이다. 하지만 Form 10-Q는 시간 흐름에 따른 기업의 변화를 연속적으로 볼 수 있도록 해주어 투자 관점에서 중요한 자료라 여겨지고 있다. 일반적인 Form 10-Q는 Part I-Financial Information(Item

1~4)과 Part II-Other Information(Item 1~6)으로 구성되어 있다.

2.2 SEC 보고서에 대한 감성 분석 연구

감성 분석은 문맥으로부터 주관적 정보를 추출하는 텍스트 마이닝 방법이다. 일반적인 감성 분석의 목표는 텍스트에 드러난 견해의 극성 탐지(Polarity Detection)로서, 극성은 긍정, 부정, 중립이 될 수 있다. 이외에도 분석의 목적에 따라 기쁨, 슬픔, 화남과 같은 보다 상세한 감정이나 정서적 상태를 분류하기도 한다. Facebook, Instagram, Twitter와 같은 SNS 플랫폼, 기타 온라인 후기를 통해 소비자, 고객 등이 게시한 내용을 감성 분석하여 전략 기획, 조직 의사 결정, 고객 관계 관리(CRM) 등 기업 경영에 활용하는 것이 인기를 얻고 있으며 관련 연구도 활발하다(사공원 등, 2016; 이선민 등, 2021). 반대로 기업을 평가하는 수단으로써 감성 분석을 활용할 수도 있는데, 아래는 이 같은 관점에서 SEC 보고서를 대상으로 하여 감성 분석한 연구들의 주요 내용 및 성과를 요약한 것이다.

Wang et al.(2013)은 금융 관련 감성 어휘를 정의하였고 SVM(Support Vector Machine) 기반의 회귀 및 순위 분석을 통해 기업의 재무적 위험을 예측하였다. 1996~2006년까지 공시된 3만 개의 Form 10-K를 수집하였으며, 수집된 자료로부터 유일 키워드, 감성 키워드 등 1,546개의 키워드로 구성된 금융 감성 용어집을 제작했다. 감성 분석 결과를 통해 금융 감성 용어의 빈도와 기업의 재무적 위험 간에 상관이 있음을 주장했다.

Chouliaras(2015)는 Form 10-K에 드러난 긍정 및 부정의 감성을 시계열로 구분하여 변화량을 산출하였으며 이를 주식 수익률 예측에 영향을 미치는 변수로 가정하였다. 감성 분류를 위해 선행 연구(Garcia, 2013; Loughran and McDonald, 2011)에서 제시한 감성 단어 사전을 사용하였으며, 시간의 흐름에 따른 감성 변화, 시가총액, 수익률 등을 변수로 하는 회귀 모델을 제시하였다. 감성의 변화에 따라 목표 변수인 수익이 증가하거나 감소할 수 있음이 연구의 결과로 나타났다.

Kang et al.(2018)은 SEC 보고서를 작성하는 사람의 주관이 보고서에 담길 수 있음을 가정하였으며, 편향된 주관으로 인해 독자가 보고서 내용을 부정확하게 해석할 수 있음을 문제로 삼았다. 이를 증명하기 위해 연구는 Form 10-K의 서술 논조에 주목하였으며 논조와 기업 실적 간의 관계를 분석하였다. 1996~2010년까지 공시된 10만 개의 Form 10-K를 수집한 후 Huang et al.(2014)가 제시한 긍정 및 부정의 키워드를 활용하여 Form 10-K에 드러난 감성의 분포를 측정하였다. 상관 분석을 위해 서술 논조, 기업 규모, 주가 수익 등을 변수로 하는 최소 제곱 회귀 모델을 제시하였다. 연구를 통해 일반적으로 긍정의 논조가 우세한 경우 실적 역시 좋은 것이 확인되었다. 하지만 긍정의 논조가 과한 경우 오히려 실적과 부정의 관계가 나타날 수 있음도 보여주었다.

Lee et al.(2018)은 54개 기업이 공시한 Form 10-K를 대상으로 텍스트 마이닝하여 패턴과 판매 실적 간의 관계를 규명하였다. Form 10-K의 항목 중 기업이 인지한 위험 요인을 기술하기 위한 'Item 1A. Risk Factors'의 내용을 TF-IDF

방법으로 분석하여 400개의 키워드를 추출하였다. 연구는 관계 규명을 위해 문단 길이, 문장의 수, 복합 연간 성장률(CAGR) 등을 변수로 두고 상관, 회귀, 계층 군집 등 다양한 분석을 실시하였다. 감성 분석을 위해서는 'RSentiment' 패키지를 사용하였으며 감성 키워드와 CAGR 간에 약한 상관이 있음을 언급하였다.

Gandhi et al.(2019)은 1997~2014년까지의 약 6천 개의 Form 10-K를 분석하여 미국 은행의 부실 예측을 시도하였다. Form 10-K 상에 드러난 부정적 키워드의 분포를 알기 위해 Loughran and McDonald(2011)의 감성 단어 사전 활용하였다. 연구는 부정적 감성이 지배적인 경우 상장 폐지의 가능성이 증가하고 배당금 지급 확률 및 자산 수익률이 감소하는 등 기업 재무에 부정적인 영향을 미침을 밝혔다.

Azimi and Agrawal(2021)은 감성과 기업 실적 간의 관계를 규명하기 위해 1994~2017년 동안 공시된 Form 10-K 6만 개를 분석하였다. 키워드 추출 및 벡터 변환에는 Word2Vec을 이용하였으며 변환된 데이터를 LSTM 구조의 신경망 모델로 학습하여 약 90%의 정확도를 달성하였다. 연구는 긍정과 부정의 감성 모두가 기업의 실적을 예측함에 있어 유용하며, 긍정의 감성은 수익성과 현금 흐름에 긍정적 영향을 미치고 부정의 감성은 반대의 역할을 한다고 주장하였다. 이때 긍정의 감성 보다 부정의 감성이 더욱 큰 효과를 가진다고 하였다. 아울러 SEC 보고서의 텍스트 마이닝 시 흔히 사용하는 항목인 'Item 1A. Risk Factors' 외에도 문서 전반에 대한 빅데이터 처리의 필요성을 논하였다.

Hao and Pham(2022)은 COVID-19 팬데믹 기간인 2020년 한 해 동안 공시된 Form 10-Q 7,895개를 분석하였으며 기업 공시 내용과 시장의 불확실성과의 관련성을 탐색하였다. COVID-19 관련 키워드, 긍정 및 부정의 키워드, 수익률, 시장 분석가(애널리스트)의 예측 등을 변수로 하는 회귀 모델을 만들고 검증하였으며, COVID-19 팬데믹과 같은 위기 상황에서 기업 공시가 시장 참여자에게 일시적으로 도움을 준다고 주장하였다.

이상으로 감성 분석 기법을 활용하여 SEC 보고서를 분석한 선행 연구의 일부를 살펴보았다. 단어 분포, 문맥을 통해 감성을 정량화하여 변수로 만들고 기업의 재무 및 위험 요인과의 관계를 규명하거나 향후 실적 및 성과를 예측하는 등의 시도가 이어지고 있음을 알 수 있었다. 기술적인 측면에서는 사전에 정의한 감성 단어 사전(용어집)을 활용하여 감성 키워드를 추출하는 방법을 여러 연구에서 주로 사용하였다. 딥러닝이 대두하면서부터는 CNN, RNN, LSTM은 물론 BERT와 같은 신경망 기반 기술이 분류나 예측에서 높은 성능을 보여주어 주목을 받고 있다(Birjali et al., 2021; Wankhade, 2022). 특히 BERT는 최근 감성 분석 연구 주제로 인기를 얻고 있는 Aspect Based Sentiment Analysis(ABSA)에서 빈번히 활용되고 있었다(Zhao and Yu, 2021).

Ⅲ. 데이터 분석

3.1 데이터 분석 기준

기초 데이터인 Form 10-K 및 Form 10-Q로부터 본 연구의 관심 텍스트를 추출하고 시간

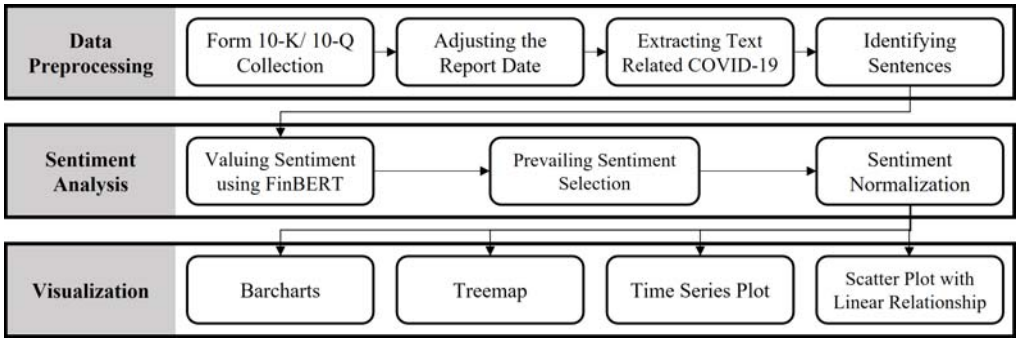
단위로 분석하기 위한 분석 기준을 정의할 필요가 있었다. 첫째, SEC 보고서 상의 수많은 텍스트로부터 COVID-19 팬데믹 관련 텍스트를 추출하는 기준이 필요하였다. 둘째, 감성의 정량화를 위해 감성 분석 최소 단위(수준)와 범위를 정해야만 했다. 셋째, 데이터를 시계열로 표현하기 위한 최소 시간 단위(Time Frame)가 필요하였다. 넷째, 거시적 관점에서 업종별 상대 비교와 효과적인 시각화를 위해 각 기업이 속한 444개 Standard Industrial Classification (SIC)를 축소할 필요가 있었다. 이와 같은 이유로 데이터 및 분석 기준을 정의하였으며 <표 1>로 정리하였다.

3.2 분석 절차

본 연구의 데이터 분석 절차를 나타내면 <그림 1>과 같다. 데이터 분석 절차는 크게 3단계로 구성하였다. 첫째, 데이터 처리 단계에서는 원시 데이터 수집, 데이터 시간(시점) 조정, COVID-19 관련 텍스트 추출, 문장 식별을 수행한다. 둘째, 감성 분석 단계에서는 FinBERT 활용 감성 값 산정, 감성 정규화, 우세 감성 분류를 수행한다. 셋째, 시각화 단계에서는 Barcharts, Treemap, Time Series Plot 등 다양한 표현 기법을 활용하여 시각화한다.

<표 1> 연구의 데이터 및 분석 기준 정의

정의 항목	정의 내용	상세 기준
COVID-19 팬데믹 키워드	covid, corona, pandemic, outbreak	SEC 보고서 및 COVID-19 팬데믹을 다룬 영문 뉴스, 인터넷 게시물 등 참고
감성 분석 수준	문장 단위 감성 분석	많은 수의 단어로 인한 분석 잡음 발생 가능성 및 분류한 감성의 모호함 최소화
감성 분석 범위	COVID-19 팬데믹 관련 키워드를 포함한 문장만 분석 대상으로 한정	COVID-19와 관련이 있는 주변 문장의 정보량을 잃긴 하지만 기초 데이터의 절대량 (1백만 건 이상)이 적지 않음을 고려
시간 단위 정의	보고 날짜(회계 마감일)가 속한 분기 + 1분기 (예: 2020.4Q → 2021.1Q)	데이터 생산 주기(분기), 미국의 주(State) 마다 다른 회계 연도 마감, 회사 규모에 따른 보고서 제출 시점 차이 등을 고려하여 통일
업종 분류	미국 표준 분류 체계에 따라 9개 업종으로 구분	미국 노동부 Standard Industrial Classification (SIC) 매뉴얼의 Division Structure를 참고하여 444개 소분류(SIC)를 9개 표준 대분류로 축소



<그림 1> 분석 절차

3.3 전처리

3.3.1 SEC 보고서 수집

SEC는 온라인 데이터베이스인 Electronic Data Gathering, Analysis, and Retrieval System (EDGAR)을 통해 기업 공시 데이터를 개방하고 있다. EDGAR는 기업 공시 조회를 위한 방법으로 일반적인 검색 외에도 프로그래밍을 위한 API(Application Programming Interface), Submission Bulk와 같은 대규모의 이력 데이터도 추가로 제공하고 있다.

연구의 기초 데이터인 SEC 보고서 수집에 앞서 연구의 시간적 범위인 2020~2021년 동안에 공시 이력을 가진 기업의 목록을 Submission Bulk로부터 추출하였다. SEC는 개별 기업을 고유하게 구분하기 위해 10자리의 숫자로 구성된 Central Index Key(CIK)를 사용한다. Submission Bulk는 CIK로 구분된 파일의 묶음으로서 개별 CIK 파일을 통해 특정 기업의 SEC 보고서 제출 이력을 확인할 수 있다. CIK 파일은 JSON(Javascript Object Notation) 구조를 따르며 SEC 보고서의 제출 이력에 대한 메타데이터 역할을 함은 물론 기업 명칭, 사업장 주소 등의 추가적인 정보도 담고 있다.

연구의 시간적 범위 내 Form 10-K 및 Form 10-Q를 제출한 이력을 가진 기업의 수는 8,409개였다. 해당 기업이 공시한 SEC 보고서의 수는 51,478개, 데이터 크기는 183GB였다. 보고서 서식 별로는 Form 10-K는 13,709개(26.6%), Form 10-Q는 37,769개(73.4%)로 확인되었다.

3.3.2 데이터 기준 시간 조정

COVID-19 팬데믹의 경과에 따른 데이터 분석을 위해서는 데이터가 가진 시간의 일관성이 확보되어야 한다. 하지만, 미국의 주(State)마다 다른 회계 연도 마감, 기업 규모에 따라 상이한 SEC 보고서 마감 기한, 특별한 사유에 따른 마감 기한 연장 가능성 등으로 인해 보고서 제출 시점은 보고서가 담고 있는 시의성과 무관하게 연중 언제라도 될 수 있다. 따라서 앞서 정의한 바와 같이 데이터의 생산 주기인 분기 별로 시간 기준을 통일하는 작업을 수행하였다.

시간 기준의 통일을 위한 기준 시간 데이터는 SEC 보고서의 보고 날짜(Report Date 또는 Period of Report)를 사용하였으며, 시간 기준을 조정하는 다음의 규칙을 모든 SEC 보고서에 일괄적으로 적용하였다.

- 1단계) SEC 보고서의 보고 날짜가 속한 분기를 계산한다.
- 2단계) 1단계)에서 계산한 분기가 1, 2, 3 분기 중 하나일 경우 1분기를 더하고 4분기일 경우 분기를 1로 둔다.
- 3단계) 2단계)에서 계산한 분기가 2, 3, 4 분기 중 하나일 경우 보고 날짜의 연도를 그대로 사용하고, 1분기일 경우 연도에 1을 더한다.

3.3.3 COVID-19 팬데믹 관련 텍스트 추출

EDGAR에서 취득할 수 있는 일반적인 SEC 보고서는 HTML 형식으로 구성되어 있으며 본 연구에서 관심 있는 COVID-19 팬데믹 관련 텍스트 외에도 재무제표와 같은 다양한 정보가 HTML 태그와 함께 섞여 있다. 우선 파이썬의 Beautiful Soup 라이브러리를 활용하여 HTML 태그를 비롯한 불필요한 데이터를 모두 제거하였다.

다음으로 NTLK(Natural Language Toolkit) 라이브러리의 Sentence Tokenizer를 활용하여 각 보고서의 텍스트를 문장 단위로 추출하였으며 추출한 문장 수는 38,319,705개였다. 다시 정규식으로 앞서 정의한 COVID-19 팬데믹 관련 키워드(covid, corona, pandemic, outbreak)를 포함한 텍스트를 파싱하였다. 결과적으로 총 1,126,660개의 텍스트가 추출되었으며 이는 전체 데이터의 2.9%에 해당하였다.

3.3.4 문장 식별

NLTK의 Sentence Tokenizer는 텍스트를 문장 단위로 구분해주지만 문법적으로 올바른 문장인지를 판정하지는 않는다. 그로 인해 소제목, 예를 들어 'COVID-19 Update'와 같이 감성 분석의 대상으로서 가치가 없는 텍스트가 나올 수 있다. 불필요한 텍스트 분석으로 인한 데이터 과잉을 피하고 유용한 정보만을 취하기 위해 문장 여부를 판정하는 추가 작업을 수행하였다.

영문법적으로 완전한 문장을 구분하는 것이 최선이나 완전한 문장을 식별하는 것은 쉽지 않다. 따라서 본 연구에서는 영어 문장이 가지는 최소한의 조건을 고려하여 문장을 식별하는 방법을 택하였다. 일반적인 문장은 최소한 한 개 이상의 주어와 술어로 구성된 절을 가짐에 착안하여 문장 요소 중 명사, 대명사, 고유 명사 중 하나 이상을 포함하면서, 동시에 동사, 조동사 중 하나 이상을 포함한 텍스트를 추출하였다. 텍스트 추출 및 토크나이징(Tokenizing) 처리에는 NLP 라이브러리인 spaCy를 활용하였다. 이 방법을 통해 앞서 추출한 1,120,660개의 텍스트 중 3,313개의 불완전한 텍스트를 제거하였으며 나머지 1,123,347개의 텍스트를 대상으로 감성 분석을 실시하였다.

3.4 감성 분석

3.4.1 FinBERT를 활용한 감성 분석

BERT는 2018년 구글이 공개한 Transformer Attention Mechanism의 언어 모델이다(Devlin et al., 2018). BERT는 등장과 동시에 다양한 NLP 문제에서 기존의 기술과 방법을 능가하는 성능을 보이며 주목을 받았다(Dor et al., 2020). BERT는 사전 학습된 언어 모델(Pre-trained Language Model, PLM)로서 특정 작업 수행에 적합하도록 손쉽게 Fine-tuning 할 수 있으며, 이 같은 장점으로 인해 BERT를 기반으로 한 5,000개 이상의 PLM이 만들어졌다(Arslan et al., 2021). BERT는 감성 분석에서도 뛰어난 성능을 보여주었는데, 몇몇 비교 실험에서 규칙, 어휘 기반의 기존 방법은 물론 같은 신경망 구조의 RNN 및 LSTM보다 성능이 우수한 것으로 나타났다(Alaparthi and Mishra, 2020; Dhola and Saradva, 2021).

본 연구는 SEC 보고서에 등장하는 금융 관련 전문 용어의 매끄러운 처리를 위해 FinBERT (Araci, 2019)를 사용하였다. FinBERT는 46,143개의 금융 관련 문서에서 추출한 2,900만개 이상의 방대한 단어를 추가로 학습한 언어 모델이다. FinBERT는 금융 분야의 감성 분류 문제에서 기존의 BERT보다 높은 분류 정확도를 보여주었다(Araci, 2019).

3.4.2 우세 감성 분류

FinBERT는 감성을 분류하고자 하는 텍스트

를 입력으로 받고 세 가지 감성, 즉 긍정, 부정, 중립의 감성 값을 각각 출력한다. 세 가지 감성 값의 합은 항상 1.0이 된다. 본 연구는 가장 높은 값을 가진 감성을 해당 텍스트의 감성으로 분류하고자 하였으나, 경우에 따라 어느 감성 값도 0.5 보다 크지 않을 수 있다. 따라서 감성의 절대적 우열을 가리기 위해 세 가지 감성 중 하나의 예측 값이 0.5 이상으로 나타난 건에 한하여 유효하게 감성 분석되었다고 판단하였다. 그리고 이 감성을 우세 감성이라 정의하였고, 우세 감성이 나타나지 않은 텍스트는 데이터 시각화 단계에서 제외하였다.

3.4.3 감성 분석 결과의 정규화

본 연구에서 수행한 감성 분석의 결과는 분기의 시간 단위로 나타난다. 분기 별 감성의 절대량은 기업이 해당 시점에 COVID-19 팬데믹에 대해 얼마만큼의 관심을 두었는지를 보여준다고 가정하고 있다. 하지만 감성 분포의 경향을 분기 별로 비교하거나, 9개 대분류 업종 별 기업 수가 다른 상황에서 업종 간 감성 분포를 비교하는 것에 절대량을 사용할 수 없다. 이에 따라 분기 별, 업종 별 감성 분포의 비교가 가능하도록 감성 분석 결과를 정규화하였다. 정규화 긍정, 부정, 중립 감성은 0~1까지의 값을 가지며 세 감성의 합은 항상 1.0이 된다. 결과적으로 당해 분기와 비교 분기 간에 특정 감성이 어느 정도의 비율(%)을 차지하는지 상대적으로 비교할 수 있다.

Ⅳ. 분석 결과

4.1 감성 정량화 및 빈도 분석

<표 2>는 수집한 SEC 보고서 중 COVID-19 팬데믹 관련 키워드(이하 COVID-19 키워드)를 담고 있는 보고서의 수, 그리고 이 수와 전체 보고서 수와의 양적 차를 보여주고 있다. COVID-19 키워드를 언급한 보고서 수가 2020

<표 2> Form 10-K 및 Form 10-Q 내용 상 COVID-19 팬데믹 관련 키워드 유무 비교

연도	분기	Form 10-K			Form 10-Q		
		전체 보고서(A)	COVID-19 키워드 포함 보고서(B)	차이 (A-B)	전체 보고서(C)	COVID-19 키워드 포함 보고서(D)	차이 (C-D)
2020	1분기	5,461	3,125	2,336	1,222	141	1,081
	2분기	457	370	87	5,454	5,075	379
	3분기	386	344	42	5,540	5,320	220
	4분기	374	349	25	5,680	5,459	221
2021	1분기	5,781	4,922	859	1,192	1,093	99
	2분기	491	432	59	5,998	5,696	302
	3분기	391	366	25	6,256	5,964	292
	4분기	368	350	18	6,427	6,114	313
합계		13,709	10,258	3,451	37,769	34,862	2,907

년 1분기에 절대적으로 적은 것이 눈에 띄며, 이것은 당시 COVID-19의 직·간접적 파급이 막 시작한 시기임을 그 이유로 유추할 수 있다.

<표 3>은 COVID-19 키워드를 언급한 보고서의 수, 보고서 당 문장의 수, 평균 문장의 수를 분기에 따라 정리한 결과이다. Form 10-K에 언급된 COVID-19 관련 평균 문장의 수는 2020년 2분기부터 증가하기 시작하였으며 2021년 2분기에 최고치를 보인 후 감소하였다. Form 10-Q에 언급된 COVID-19 관련 문장의 수는 평균은 물론 문장의 절대량 역시 2020년과 비교해 2021년에 전반적으로 감소한 것을 알 수 있었다.

<표 4>는 COVID-19 팬데믹 관련 전체 문장에 대한 감성을 FinBERT로 예측한 결과를 보

<표 3> Form 10-K 및 Form 10-Q의 문장 및 평균 문장 수의 분기별 분포

연도	분기	Form 10-K			Form 10-Q		
		보고서 수	문장 수	평균 문장	보고서 수	문장 수	평균 문장
2020	1분기	3,125	24,104	7.713	141	515	3.652
	2분기	370	9,460	25.568	5,075	140,306	27.647
	3분기	344	11,010	32.006	5,320	170,062	31.967
	4분기	349	10,896	31.221	5,459	168,223	30.816
2021	1분기	4,922	191,441	38.895	1,093	23,239	21.262
	2분기	432	17,621	40.789	5,696	103,387	18.151
	3분기	366	11,627	31.768	5,964	114,544	19.206
	4분기	350	10,676	30.503	6,114	116,236	19.011
합계 / 평균		10,258	286,835	27.962	34,862	836,512	23.995

<표 4> FinBERT로 예측한 감성 분류 결과 및 정규화 감성 분포

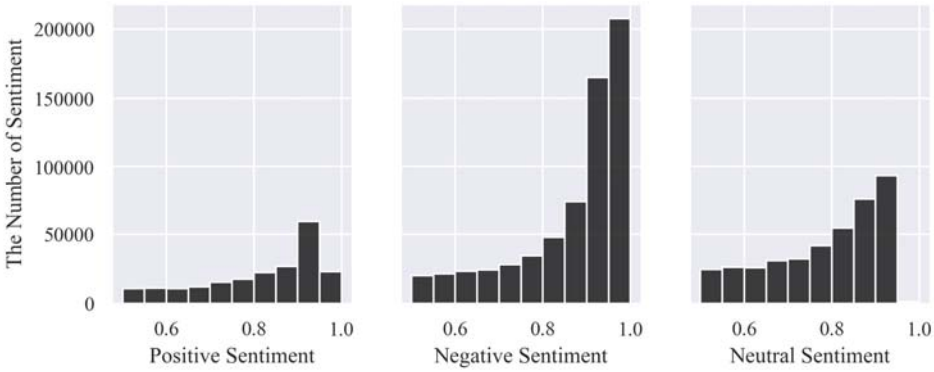
연도	분기	긍정(Positive)		부정(Negative)		중립(Neutral)	
		감성량	정규화 감성	감성량	정규화 감성	감성량	정규화 감성
2020	1분기	1,304	0.054	15,032	0.623	7,773	0.322
	2분기	17,839	0.122	80,865	0.553	47,361	0.324
	3분기	25,591	0.144	97,568	0.552	53,576	0.303
	4분기	26,751	0.153	94,392	0.540	53,485	0.306
2021	1분기	33,481	0.159	109,814	0.524	65,981	0.315
	2분기	20,368	0.172	58,135	0.493	39,384	0.334
	3분기	25,061	0.203	57,823	0.470	40,087	0.325
	4분기	24,172	0.195	58,504	0.473	40,972	0.331
합계 / 평균		174,567	0.159	572,133	0.522	348,619	0.318

여주고 있으며, 이것은 SEC 보고서를 제출한 이력이 있는 기업 전체의 평균적인 감성 분포라 할 수 있다. 1,123,347개의 문장 중 28,028개는 우세 감성 없었으며 이 데이터는 이후 시각화 단계에서 고려하지 않았다. 감성의 양은 부정이 긍정 및 중립보다 절대적으로 많았으며, 이 패턴은 8개 분기 동안 변함이 없었다. 정규화 감성의 시간 흐름에 따른 변화는 감성 별로 차이를 보였다. 부정의 감성 비율은 점차 감소하고 있으며, 반대로 긍정의 감성 비율은 점차 증가하고 있음을 알 수 있었다. 중립 감성에서는 증가 또는 감소 패턴이 보이지 않았다.

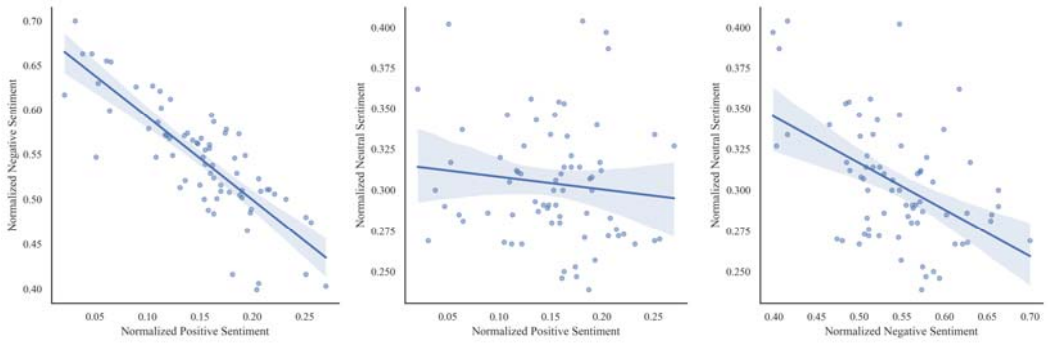
4.2 업종 전체 감성 분석

<그림 2>는 세 가지 감성의 빈도를 나타낸 히스토그램이다. <표 4>에서 확인한 바와 같이 부정 감성의 절대량이 많은 것을 시각적으로 확인할 수 있다. 감성의 예측값은 전반적으로 0.8 이상의 구간에 분포하고 있는데 이것은 FinBERT가 본 연구의 데이터인 COVID-19 관련 문장에 대하여 비교적 명료하게 감성 분류하고 있음을 나타낸다.

<그림 3>은 정규화 감성 간의 선형 관계에 대한 도표이다. 긍정과 부정 감성 간의 선형 상

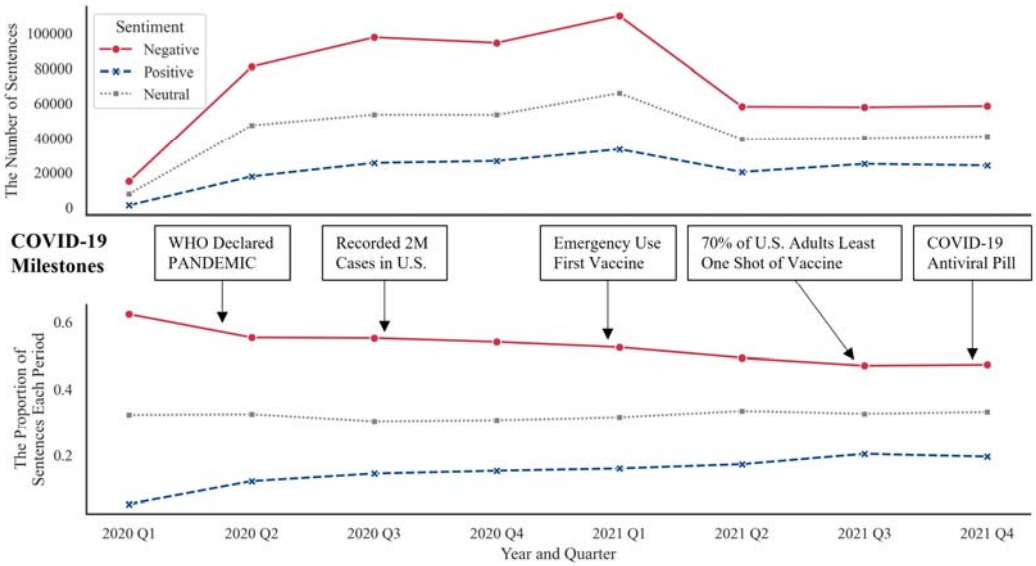


<그림 2> COVID-19 관련 문장에 대한 FinBERT의 감성 예측값 분포



긍정(X축) - 부정(Y축) 긍정(X축) - 중립(Y축) 부정(X축) - 중립(Y축)

<그림 3> 정규화 감성 간 선형 상관



<그림 4> COVID-19 팬데믹 관련 감성의 시계열 및 Key Milestones

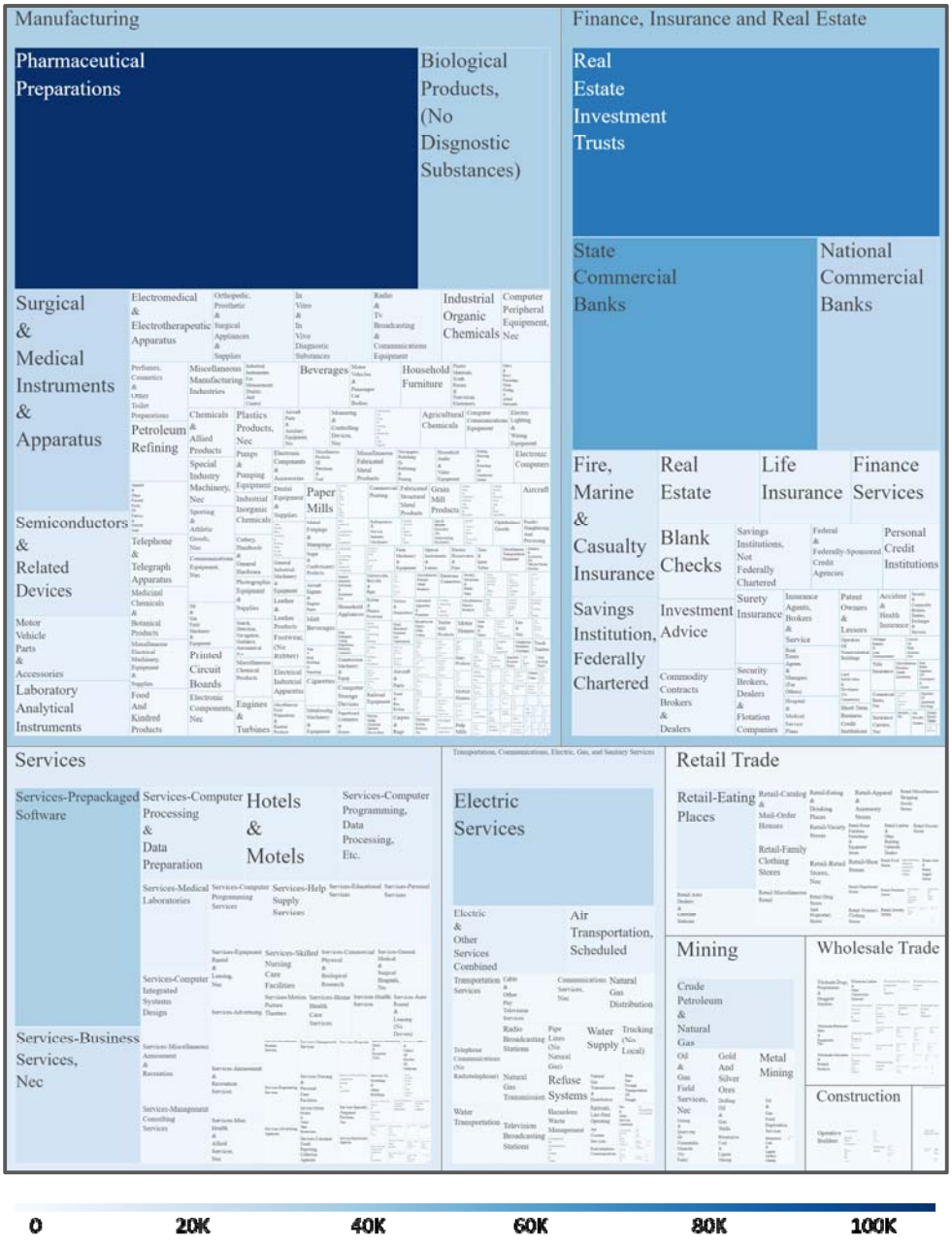
관(좌측 도표)이 가장 강함이 보였다. 세 개의 도표가 나타내는 의미를 종합하면, 최대 1.0의 값을 갖는 정규화 감성에서 특정 감성의 변화로 인한 상쇄가 필연적인데, 이때 부정의 감성이 감소하는 경우 이 감소량은 긍정, 중립 순으로 배분된다고 판단할 수 있다.

<그림 4>는 시간(분기)을 가로축, 감성량을 세로축에 놓은 시계열 도표이다. 도표의 상단은 세 가지 감성의 절대량을 나타내며 하단의 도표는 정규화 감성을 나타내고 있다. 중앙의 COVID-19 Milestones는 COVID-19의 최초 발병 후 WHO의 팬데믹 선언, 확진자 및 사망자 증가, 백신 개발 및 접종 등 주요 이슈를 시간 흐름에 따라 나열한 것이다(Whiting and Wood, 2021).

먼저 상단의 감성의 절대량을 보면, 앞서 언급한 바와 같이 2020년 2분기부터 급격하게 증가하면서 2021년 1분기 정점에 다다랐음을 직관적으로 알 수 있다. 다음으로 하단의 정규화 감성에 대한 시계열을 보면 긍정의 감성은 점차 상승, 부정의 감성은 점차 하락하는 모습을 시각적으로 확인할 수 있다. 중립 감성은 큰 변화를 보이지 않고 있다.

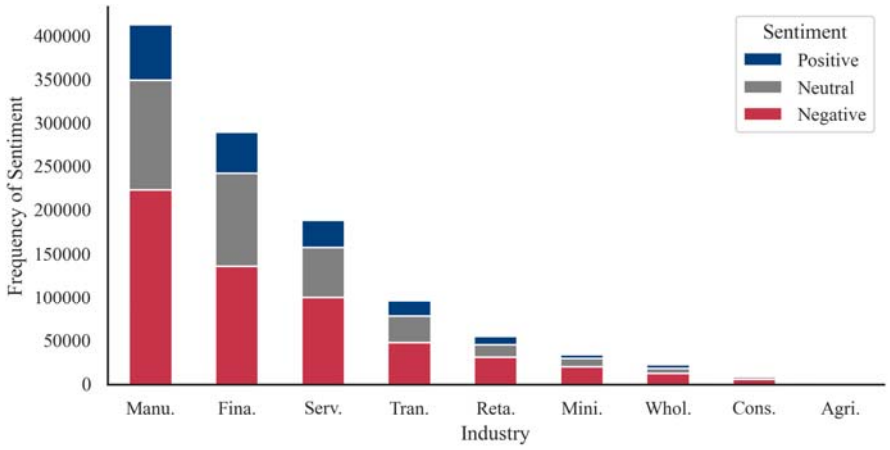
4.3 업종 간 감성 분석 비교

<그림 5>는 9개 대분류 업종 및 하위 소분류 업종의 계층적 구조에 따라 COVID-19 관련 문장의 분포를 시각화한 것이다. 사각형의 크기가 클수록 해당 업종의 SEC 보고서에 COVID-19 관련 문장이 많이 언급되었음을 의미한다. 가장 많은 기업이 속한 Manufacturing 업종의 크기가 가장 크며 다음으로 Finance, Insurance and Real Estate 업종, Service 업종 순임을 확인할 수 있다.



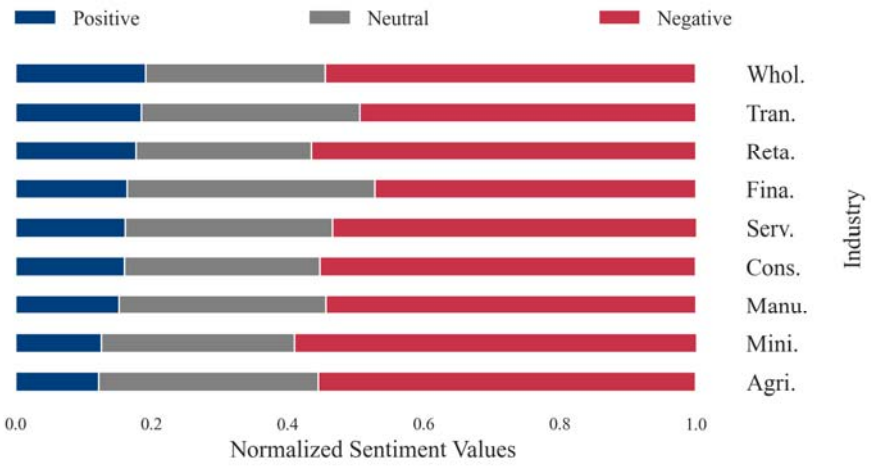
<그림 5> 업종 대분류 및 소분류에 따라 COVID-19 관련 문장의 양적 크기 비교

<그림 6>과 <그림 7>은 업종 별 감성을 비교할 수 있도록 시각화한 도표이다. <그림 6>은 감성의 절대량을 내림차순으로 보여주고 있다. 가장 많은 기업이 속한 Manufacturing 업종에서 COVID-19 관련 문장의 절대량이 많음을 알 수 있다.



Note. Agri. = Agriculture, Forestry and Fishing; Mini. = Mining; Cons. = Construction; Manu. = Manufacturing; Tran. = Transportation, Communications, Electric, Gas, and Sanitary Services; Whol. = Wholesale Trade; Reta. = Retail Trade; Fina. = Finance, Insurance and Real Estate; Serv. = Services

<그림 6> 감성 절대량의 업종 별 비교

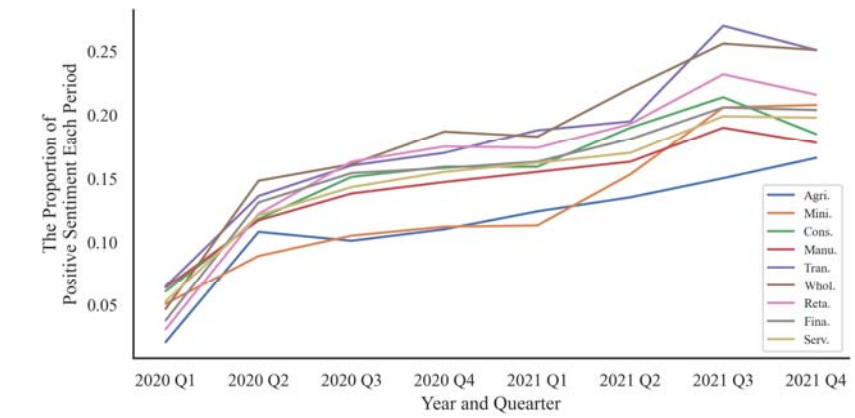
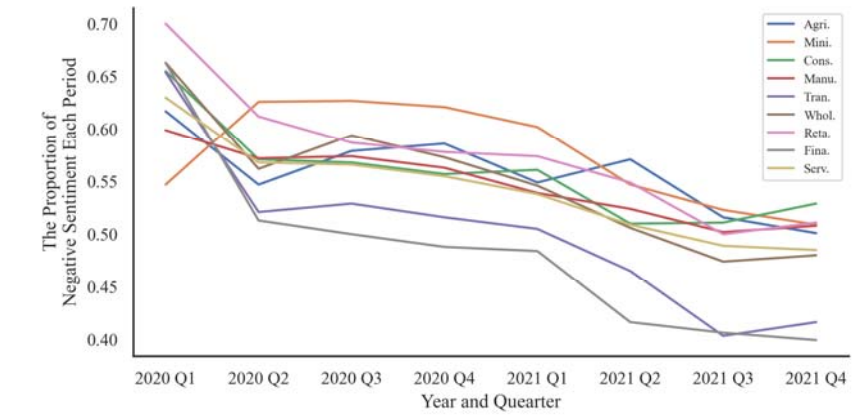
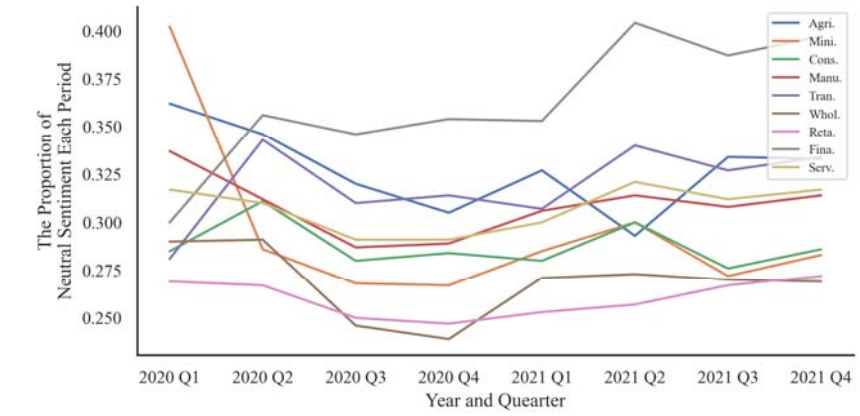


<그림 7> 정규화 감성의 업종 별 비교

<그림 7>은 업종 별 정규화 감성을 나타내며 긍정 감성 비율로 내림차순 한 것이다. 전체 감성량을 100%로 두었을 때 업종 별 감성의 상대적인 크기를 시각적으로 비교할 수 있다. 긍정 감성 비율이 가장 높은 곳은 Wholesale Trade 업종, 가장 낮은 곳은 Agriculture, Forestry and Fishing 업종이었다.

<그림 8>은 세 가지 감성의 시계열을 업종 별로 비교한 도표이다. 긍정 및 부정의 감성의 시계열패턴은 업종에 따라 다소 차이가 있으나,

상승 및 하락의 큰 흐름은 유사한 것으로 판단 라 양상이 달랐다.
할 수 있었다. 중립 감성의 시계열은 업종에 따

긍정 (Positive)	
부정 (Negative)	
중립 (Neutral)	

<그림 8> 업종 별 정규화 감성의 시계열 비교

Ⅴ. 결 론

본 연구에서는 FinBERT 언어 모델을 활용한 SEC 보고서의 감성 분석을 통해 2020~2021년 동안의 COVID-19 팬데믹 상황에 대한 기업 및 업계의 견해를 정량화하였으며, 분석 결과의 효과적인 전달을 위해 다양한 방법으로 시각화하였다. 공식적인 문서인 SEC 보고서를 대상으로 한 감성 분석 및 감성 분류, 감성의 시계열 변화 추적, 적절한 데이터 시각화를 목표로 하여 기존 연구와의 차별성을 꾀하였다.

SEC의 온라인 데이터베이스인 EDGAR로부터 51,478개의 SEC 보고서(Form 10-K 및 Form 10-Q)를 수집하였고, 파이썬 및 NLP 라이브러리를 활용하여 COVID-19 팬데믹 관련 텍스트(문장) 1,123,347개를 추출하였다. 추출한 문장에 대한 기업의 견해는 긍정, 부정, 중립으로 분류할 수 있다. 본 연구는 사전 학습된 언어 모델인 BERT를 금융 감성 분류에 적합하도록 Fine-tuning한 FinBERT(Araci, 2019)를 사용하여 감성 분류 및 정량화를 시도하였다. 다량의 기업 공시를 데이터로 한 분석 및 시각화를 통해 밝힌 핵심 내용을 요약하면 아래와 같다.

첫째, 2020~2021년 동안 SEC 보고서 상에서 언급되었던 COVID-19 팬데믹 관련 텍스트의 양적 변화를 관찰하였다. 2020년 한 해 동안 COVID-19 관련 텍스트 양은 급격히 증가하였으며, 2021년 1분기를 정점으로 감소한 것을 수치와 도표로 확인할 수 있었다.

둘째, COVID-19 팬데믹의 경과에 따라 미국 기업 공시에 드러난 전반적인 감성의 변화가 관찰되었다. 가장 많은 비중을 차지한 부정 감성은 8개 분기에 걸쳐 서서히 감소한 반면, 긍정의 감성은 서서히 증가하는 것을 시각적으로 확인할 수 있었다. 아울러 긍정과 부정 감성 간의 선형 관계 분석을 통해 두 감성이 높은 강도로 상쇄함을 알 수 있었다. 중립 감성에서는 주목할만한 패턴이나 관계를 볼 수 없었다.

셋째, COVID-19 팬데믹에 대한 업종 별 감성의 절대적, 상대적 차이를 살펴볼 수 있었다. 많은 기업이 속한 Manufacturing 업종의 텍스트 절대량이 가장 많았으며, 정규화 긍정 감성의 비율이 가장 높은 곳은 Wholesale Trade 업종, 가장 낮은 곳은 Agriculture, Forestry and Fishing 업종이었다. McKinsey Global Institute and Oxford Economics(2020)는 COVID-19 팬데믹 이전 수준의 경기 회복 기간에 대한 분석에서 Wholesale Trade, Retail Trade 등은 타 업종과 비교해 상대적으로 빠른 회복세를 보일 것이고, Manufacturing, Mining 등은 회복에 5년 이상이 소요될 것이라 예측하였다. 이런 점은 본 연구의 업종 별 긍정과 부정 감성 분포를 경기 회복 기대감에 비유하였을 때 일정 부분 일치한다고 할 수 있다. 업종 별 정규화 감성의 시계열 도표를 보면 긍정은 상승, 부정은 하락하는 패턴이 전 업종에서 나타났으나, 중립 감성에서는 업종 간 유사점을 볼 수 없었다.

본 연구의 이론적 시사점은 다음과 같다. 첫째, COVID-19 팬데믹과 같은 갑작스러운 사건의 발생과 기업 활동과의 관계에 대한 깊이 있는 논의에 앞서, 본 연구에서 수행한 텍스트 마이닝 기반의 데이터 분석 및 시각화를 통해 정보와 통찰은 물론 추가 연구를 위한 단초를 얻을 수 있다(Ignatow and Mihalcea, 2017). 둘째, SEC 보고서와 같이 감성을 나타내기 위함이 아

닌 데이터에 대한 감성 분석을 통해 감성의 양적 분포는 물론 시계열 변화 및 패턴도 관찰할 수 있음을 보였으며, 이 같은 시의성은 기존 연구에서 언급한 SEC 보고서의 정보 매체로서의 가치(Stephany et al., 2020; Yuthas et al., 2002)를 방증한다.

실무적 시사점으로는 첫째, SEC 보고서를 시계열로 분석함에 있어 기초 데이터 수집, 시간 데이터 조정, 업종 분류 등 적절한 전처리 과정과 분석 방법론을 제시하였다. 둘째, COVID-19 팬데믹으로 인한 미국 산업의 변화 모습을 다양한 시각화로 풀어 보여줌으로써 투자자, 소비자 등의 이해를 도움은 물론 향후 변화의 방향성을 유추할 수 있는 가이드 역할을 하였다.

본 연구의 한계는 다음과 같다. 첫째, SEC 보고서는 기업이 작성하며 기업에 따라 COVID-19 팬데믹과 같은 위험 요인을 평가하는 기준에 차가 있을 수 있다. 또한 기업에 따라 보고서 분량, 내용의 충실성 등 편차가 있다. 이런 이유로 감성의 분포가 과대 또는 과소 편향할 수 있다. 둘째, 감성의 양적 분석에 초점을 맞춤으로 인해 텍스트의 질적 평가를 하지 못하였다. 항공, 여행 업계와 같이 COVID-19로 거의 완전한 수요 붕괴에 이른 업종(Dube et al., 2021)의 견해, 즉 감성에 대한 평가는 달리함이 타당할 수 있다. 아울러 감성 분포와 변화 정도가 산업별로 차이를 보인 점에 대한 추가 연구를 실시한다면 본 연구에서 드러난 양적 결과와의 관련성을 모색해 볼 수 있을 것이다. 셋째, FinBERT 언어 모델을 통한 기계적인 감성 분류 결과의 부정확성이다. 본 연구에서 사용한 약 112만개의 데이터의 경우 감성 분류를 위한 정답이 존재하지 않기 때문에 분류 결과를 정확하게 평가할 수 없다는 한계가 있다.

참고문헌

사공원, 하성호, 박경배, “온라인 후기에 내재된 고객의 감성분석과 LQI 차원별 호텔서비스 품질 평가,” 정보시스템연구, 제25권, 제3호, 2016, pp. 217-245.

이선민, 천세진, 박상언, 이태욱, 김우주, “동적 토픽 모델링과 감성 분석을 이용한 COVID-19 구간별 비대면 근무 부정요인 검출에 관한 연구,” 정보시스템연구, 제30권, 제4호, 2021, pp. 277-301.

홍태호, 나우한잉, 임강, 박지영, “LDA를 이용한 온라인 리뷰의 다중 토픽별 감성분석 - TripAdvisor 사례를 중심으로,” 정보시스템연구, 제27권, 제1호, 2018, pp. 89-110.

Alaparthi, S., and Mishra, M., “Bidirectional Encoder Representations from Transformers (BERT): A Sentiment Analysis Odyssey,” 2020, *arXiv preprint arXiv:2007.01127*.

Ali, S. M., Gupta, N., Nayak, G. K., and Lenka, R. K., “Big Data Visualization: Tools and Challenges,” In 2016 2nd *International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, December 2016, pp. 656-660.

Araci, D. “FinBERT: Financial Sentiment

Analysis with Pre-trained Language Models," 2019, *arXiv preprint arXiv:1908.10063.*

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., and Goujon, A., "A Comparison of Pre-Trained Language Models for Multi-class Text Classification in the Financial Domain," In *Companion Proceedings of the Web Conference 2021,* April 2021, pp. 260-268.

Azimi, M., and Agrawal, A., "Is Positive Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning," *The Review of Asset Pricing Studies,* Vol 11, No. 4, 2021, pp. 762-805.

Birjali, M., Kasri, M., and Beni-Hssane, A. "A Comprehensive Survey on Sentiment Analysis: Approaches, Challenges and Trends," *Knowledge-Based Systems,* Vol. 226, 2021, 107134.

Capuano, N., Greco, L., Ritrovato, P., and Vento, M., "Sentiment Analysis for Customer Relationship Management: An Incremental Learning Approach," *Applied Intelligence,* Vol. 51, No. 6, 2021, pp. 3339-3352.

Che, S., Zhu, W., and Li, X., "Anticipating Corporate Financial Performance from CEO Letters Utilizing Sentiment Analysis," *Mathematical Problems in Engineering,* 2020, 2020.

Chi, S., and Shanthikumar, D. M., "Do Retail Investors Use SEC Filings? Evidence from EDGAR Search," *Evidence from EDGAR Search (October 25, 2018),* 2018.

Chouliaras, A., "The Pessimism Factor: SEC EDGAR Form 10-K Textual Analysis and Stock Returns," *Available at SSRN 2627037,* 2015.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805,* 2018.

Dhola, K., and Saradva, M., "A Comparative Evaluation of Traditional Machine Learning and Deep Learning Classification Techniques for Sentiment Analysis," In *2021 11th International Conference on Cloud Computing, Data Science and Engineering,* IEEE, January 2021, pp. 932-936.

Dor, L. E., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., and Slonim, N., "Active Learning for BERT: An Empirical Study," In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* November 2020, pp. 7949-7962.

Dube, K., Nhamo, G., and Chikodzi, D., "COVID-19 Pandemic and Prospects for Recovery of the Global Aviation

Industry," *Journal of Air Transport Management*, Vol. 92, 2021, 102022.

Gandhi, P., Loughran, T., and Mcdonald, B., "Using Annual Report Sentiment as a Proxy for Financial Distress in US Banks," *Journal of Behavioral Finance*, Vol. 20, No 4, 2019, pp. 424-436.

Garcia, D., "Sentiment during Recessions," *The Journal of Finance*, Vol. 68, No. 3, 2013, pp. 1267-1300.

Hao, J., and Pham, V. T., "COVID-19 Disclosures and Market Uncertainty: Evidence from 10-Q Filings," *Australian Accounting Review, Forthcoming*, 2022.

Heer, J., Bostock, M., and Ogievetsky, V., "A Tour Through the Visualization Zoo," *Communications of the ACM*, Vol. 53, No. 6, 2010, pp. 59-67.

Huang, X., Teoh, S. H., and Zhang, Y., "Tone Management," *The Accounting Review*, Vol. 89, No. 3, 2014, pp. 1083-1113.

Ignatow, G., and Mihalcea, R., An Introduction to Text Mining: Research Design, Data Collection, and Analysis, Sage Publications, New York, 2017.

Kang, T., Park, D. H., and Han, I., "Beyond the Numbers: The Effect of 10-K Tone on Firms' Performance Predictions Using Text Analytics," *Telematics and Informatics*, Vol. 35, No. 2, 2018, pp. 370-381.

Larcker, D. F., Lynch, B., Tayan, B., and Taylor, D. J., "The Spread of COVID-19 Disclosure," *Rock Center for Corporate Governance at Stanford University Closer Look Series: Topics, Issues and Controversies in Corporate Governance No. CGRP-84*, 2020.

Lee, B., Park, J. H., Kwon, L., Moon, Y. H., Shin, Y., Kim, G., and Kim, H. J., "About Relationship Between Business Text Patterns and Financial Performance in Corporate Data," *Journal of Open Innovation: Technology, Market, And Complexity*, Vol. 4, No. 1, 2018, 3.

Loughran, T., and Mcdonald, B., "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, Vol. 66, No. 1, 2011, pp. 35-65.

Myšková, R., and Hájek, P., "Sustainability and Corporate Social Responsibility in the Text of Annual Reports-The Case of the IT Services Industry," *Sustainability*, Vol. 10, No. 11, 2018, 4119.

Qin, X., Luo, Y., Tang, N., and Li, G., "Making Data Visualization More Efficient and Effective: a Survey," *The VLDB Journal*, Vol. 29, No. 1, 2020, pp. 93-117.

Stephany, F., Stoehr, N., Darius, P., Neuhäuser, L., Teutloff, O., and Braesemann, F., "The Corisk-Index: A Data-Mining Approach to Identify Industry-Specific Risk Assessments Related to

COVID-19 in Real-Time," *arXiv preprint arXiv:2003.12432*, 2020.

Wang, C. J., Tsai, M. F., Liu, T., and Chang, C. T., "Financial Sentiment Analysis for Risk Prediction," In *Proceedings of The Sixth International Joint Conference on Natural Language Processing*, October 2013, pp. 802-808.

Wankhade, M., Rao, A. C. S., and Kulkarni, C., "A Survey on Sentiment Analysis Methods, Applications, and Challenges," *Artificial Intelligence Review*, 2022, pp. 1-50.

Yuthas, K., Rogers, R., and Dillard, J. F., "Communicative Action and Corporate Annual Reports," *Journal of Business Ethics*, Vol. 41, No. 1, 2002, pp. 141-157.

Zhao, A., and Yu, Y., "Knowledge-Enabled BERT for Aspect-Based Sentiment Analysis," *Knowledge-Based Systems*, Vol. 227, 2021, 107220.

McKinsey Global Institute and Oxford Economics, COVID-19 Recovery in Hardest-hit Sectors Could Take More than 5 Years, McKinsey & Company, July 29, 2020, Retrieved June 10, 2022, Available: https://www.mckinsey.com/featured-insights/coronavirus-leading-through-the-crisis/charting-the-path-to-the-next-normal/covid-19-recovery-in-hardest-hit-sectors-could-take-more-than-5-years.

U.S. Securities and Exchange Commission, How to Read a 10-K/10-Q, Retrieved April 7, 2022a, Available: https://www.sec.gov/oiea/investor-alerts-and-bulletins/how-read-10-k10-q.

U.S. Securities and Exchange Commission, Rules and Regulations for the Securities and Exchange Commission and Major Securities Laws, Retrieved April 7, 2022b, Available: https://www.sec.gov/about/laws/secrulesregs.htm.

Whiting, K. and Wood, J., Two Years of COVID-19: Key Milestones in the Pandemic, World Economic Forum, December 2021, Retrieved April 7, 2022, Available: https://www.weforum.org/agenda/2021/12/covid19-coronavirus-pandemic-two-years-milestones/.

김 효 곤 (Kim, Hyo Gon)

동의대학교에서 컴퓨터 통계학 학사를 취득하였다. 현재 경상국립대학교 기술경영학과 석사과정에 있으며, 한국토지주택공사 차장으로 재직 중이다. 주요 관심 분야는 GIS, 인공지능, 빅데이터 분석 등이다.

유 동 희 (Yoo, Dong Hee)

고려대학교에서 경영학사와 경영학 박사학위를 취득하였다. 현재 경상국립대학교 경영정보학과에서 교수로 재직하고 있으며, 주요 관심 분야는 빅데이터 분석, 인공지능, 지식 그래프, 지능형시스템 등이다.

<Abstract>

Sentiment Analysis and Data Visualization of U.S. Public Companies' Disclosures using BERT

Kim, Hyo Gon · Yoo, Dong Hee

Purpose

This study quantified companies' views on the COVID-19 pandemic with sentiment analysis of U.S. public companies' disclosures. It aims to provide timely insights to shareholders, investors, and consumers by analyzing and visualizing sentiment changes over time as well as similarities and differences by industry.

Design/methodology/approach

From more than fifty thousand Form 10-K and Form 10-Q published between 2020 and 2021, we extracted over one million texts related to the COVID-19 pandemic. Using the FinBERT language model fine-tuned in the finance domain, we conducted sentiment analysis of the texts, and we quantified and classified the data into positive, negative, and neutral. In addition, we illustrated the analysis results using various visualization techniques for easy understanding of information.

Findings

The analysis results indicated that U.S. public companies' overall sentiment changed over time as the COVID-19 pandemic progressed. Positive sentiment gradually increased, and negative sentiment tended to decrease over time, but there was no trend in neutral sentiment. When comparing sentiment by industry, the pattern of changes in the amount of positive and negative sentiment and time-series changes were similar in all industries, but differences among industries were shown in neutral sentiment.

Keyword: Sentiment Analysis, Visualization, BERT, SEC, Disclosure, Form 10-K, Form 10-Q, COVID-19

* 이 논문은 2022년 4월 20일 접수, 2022년 6월 8일 1차 심사, 2022년 8월 2일 게재 확정되었습니다.