한국증권학회지 제50권 2호 (2021) 135-170 https://doi.org/10.26845/KJFS.2021.04.50.2.135

Building the Korean Sentiment Lexicon for Finance (KOSELF)*

Su-Ji Cho, *Ph.D. Candidate, Dankook University*
Heung-Kyu Kim, *Professor, Dankook University*
Cheol-Won Yang**, *Professor, Dankook University*

〈**Abstract**〉

This study aims to verify and establish a Korean sentiment lexicon suitable for corporate financial analysis. When analyzing existing sentiment lexicons, the KOSAC and KNU (Kunsan University) dictionaries developed based on Korean are weak because they are used for general purposes. The Harvard IV and Loughran and McDonald (2011) have the disadvantage of being translated from English. In this study, the Korean Sentiment Lexicon for Finance (KOSELF) is constructed and presented. To verify its usefulness, text data from about 20,000 analyst reports published in Korea from 2016 to 2018 are collected from the Hankyung Consensus web page. After calculating the sentiment variables of negative and positive word frequency using five sentiment lexicons for each report, the recommendation and target price changes are regressed on these sentiment variables. The sentiment variables from the newly-constructed KOSELF in this study have a significant relationship with the analyst's recommendation and target price change. Even when the sentiment variables calculated through the other four sentiment lexicons are added, it shows better performance. Our work has practical significance in that it proposes a Korean sentiment dictionary that can be used for finance.

Keywords: Text Analysis; Sentiment Lexicon; Analyst Report; Recommendation; Target Price

JEL Classification: G10, G11, G14, G23

* This paper was supported by the research fund of the National Research Foundation of Korea (NRF-2019S1A5A2A03038389).

** Corresponding Author. Address: School of Business Administration, Dankook University, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, Korea, 16890; E-mail: yang@dankook.ac.kr; Tel: +82-31-8005-3437; Fax: +82-31-8021-7208.

Received: October 31, 2020; Revised: January 21, 2021; Accepted: February 8, 2021

기업 재무분석을 위한 한국어 감성사전 구축*

조 수 지 (단국대학교 박사과정)
김 홍 규 (단국대학교 교수)
양 철 원 (단국대학교 교수)**

〈 요 약 〉

본 연구는 기업 재무분석에 적합한 한국어 감성사전을 검증하고 구축하는 것을 목적으로 하고 있다. 기존 감성사전을 분석하였을 때, 한국어 기반으로 개발한 KOSAC과 KNU(군산대) 감성사전은 일반용이라는 약점을 가지며, Harvard IV(HV)와 Loughran and McDonald(LM)(2011)는 영어를 단순 번역하였다는 단점을 지니고 있다. 본 연구는 이를 보완한 한국어 금융 감성사전(KOSELF, Korean Sentiment Lexicon for Finance)을 새롭게 구축하여 제시하였다. 감성사전을 검증하기 위해 한경 컨센서스에서 제공하는 2016년부터 2018년까지 한국에서 발행된 약 2만 개의 애널리스트 보고서 자료를 사용하였다. 보고서 별로 5개 감성사전을 통해 계산한 부정어, 긍정어 비율 등의 감성변수와 목표주가 및 추천의견 변경과의 관계를 검증하였다. 본 연구에서 새롭게 구축한 KOSELF 감성변수는 애널리스트 목표주가 및 추천의견 변경과 유의미한 관계를 가졌으며, 나머지 4개의 감성사전을 통해 계산한 변수들을 추가하였을 때도 우수한 성과를 보여주었다. 본 연구는 재무, 금융 분야에서 활용할 수 있는 한국어 감성사전을 제안하였다는 점에서 의의를 지닌다.

핵심 단어 : 텍스트 분석, 감성사전, 애널리스트 보고서, 추천의견, 목표주가

JEL 분류기호: G10, G11, G14, G23

* 본 논문에 대해 유익한 조언을 해 주신 익명의 심사자들께 감사드립니다. 애널리스트 보고서 자료 구축을 위해 조교로서 수고해 준 민경수, 최보미, 정희조에게 고마움을 표합니다. 이 연구는 한국연구재단 연구지원사업의 지원을 받아 수행되었습니다(NRF-2019S1A5A2A03038389).

** 연락담당 저자. 주소: 경기도 용인시 수지구 죽전로 152 단국대학교 경영경제대학 경영학부, 16890; E-mail: yang@dankook.ac.kr; Tel: 031-8005-3437; Fax: 031-8021-7208.

1. 서론

빅데이터 시대에 가장 규모가 크면서도 새로운 정보를 제공해 줄 수 있는 것이 텍스트 자료이다. 기업분석에 있어서도 지금까지는 주로 기업에 대한 정형화된 수치 자료를 사용하였다. 예를 들어, 기업이 스스로 공시하는 사업보고서 자료, 애널리스트 보고서 자료 중에서도 기업에 대한 분석과 예측에 사용되는 것은 주로 수치 자료였다. 하지만 사업보고서와 애널리스트 보고서에서 가장 많은 부분을 차지하는 것은 텍스트이다. 지금까지 텍스트는 분석에서 제외되어 왔지만, 이제는 이런 텍스트들을 분석하기 시작하였으며 거기서 수치 자료에서 제공하는 것 이상의 정보를 얻어내기 시작하였다(Tetlock et al., 2008).

그렇지만 한국은 기업 재무분석과 연관하여 텍스트에 대한 연구가 거의 이루어지지 않고 있다. 이는 두 가지 장벽 때문이라 사료된다. 첫째, 기업분석 및 예측을 위해서는 텍스트의 계량화가 필요하다. 텍스트라는 비정형화된 자료를 그대로 사용할 수는 없으니 이를 정량화해야 한다. 다른 수치자료에 비해 한 단계 작업이 더 필요한 것이다. 둘째, 한국어 텍스트의 경우, 언어의 장벽으로 인해 미국의 연구를 그대로 적용할 수 없다. 특히 한국어는 영어와 전혀 다른 언어체계를 가지고 있다.[1] 그에 맞는 새로운 방법론을 개발해야 한다. 수치 분석은 미국의 연구를 그대로 한국 자료로 바꾸어 검증하는 것이 가능하지만, 텍스트의 경우는 적용되지 않는다.

위 두 가지를 생각할 때 현재 시점에서 가장 필요한 것이 텍스트의 계량화를 위한 한국어 감성사전을 개발하는 것이다. 특히 Loughran and McDonald(2011)는 기업분석에 사용될 감성사전은 일반 사전과 달라야 함을 주장하고 있다. 이런 점을 고려하여 본 연구는 기업 재무분석에 적합한 한국어 감성사전을 검증하고 구축하는 것을 목적으로 하고 있다. 현재 한국어 감성사전이 존재하지 않는 것은 아니지만, 일반용 사전에 한정되어 있으며 재무에 특화된 사전은 존재하지 않는다. 따라서 본 연구는 기존 감성사전을 검토한 후에 단점을 보완하여 기업 재무분석에 적합한 금융 감성사전을 구축하고자 한다.

본 연구는 다음과 같은 측면에서 기존 연구와 차별점을 가지며, 문헌에 기여할 수 있으리라 사료된다. 첫째, 한국어에 적합한 금융 감성사전을 제안하였으며, 이를 KOSELF(Korean Sentiment Lexicon for Finance)라 명명하였다.[2] 이는 실용적인 면에서 큰 의미가 있다. 기본적으로 기업 재무분석 및 증권분석에 사용될 수 있을 것으로 기대되며 그 외에도 다양한 금융분야에도 적용될 수 있을 것이다.

둘째, 본 연구는 애널리스트 보고서 자료를 사용하여 감성사전을 검증하였다. 감성사전의 효율성을 검토하기 위해서는 벤치마크가 필요한데, 본 연구는 애널리스트 보고서의 추천의견과 목표주가를 사용하였다. Loughran and McDonald(2011)는 영어 금융 감성사전을 검증하기 위한 벤치마크로 시장수익률이나 SUE(standardized unexpected earnings)을 사용하였다.

1) 언어학적으로 분류를 해 볼 때 일반적으로 영어는 인도유럽어족(Indo-European Language)이고 한국어는 우랄알타이 어족(Ural-Altaic language)에 속하는 것으로 인식된다.

2) 본 연구를 통해 구축된 KOSELF 단어목록은 다음의 웹 사이트에 공개되어 있다. (https://sites.google.com/view/cheolwon-yang/koself?authuser=0)

하지만 시장수익률이나 SUE는 텍스트 감성에 의해만 결정되기 보다는 다른 요인들에 의해서도 영향을 받기 때문에 텍스트 감성의 벤치마크로 사용하기에는 부족함이 있다. 수익률은 시장참여자들의 컨센서스가 반영된 것이므로 특정 텍스트의 감성에 대한 완벽한 벤치마크는 아니다. 즉, 텍스트 작성자 외의 다른 시장참여자의 의견도 반영되어 있는 것이다. 하지만 애널리스트 보고서는 텍스트를 작성한 애널리스트 자신이 추천의견이나 목표주가를 제시하기 때문에 이런 편의(bias)에서 자유롭다. 즉, 작성자 자신이 부정적인 의견을 가지고 있다면 자신이 작성한 텍스트 속에 서술되어 있을 것이며 동시에 추천의견이나 목표주가에 반영될 것이다. 이런 연구 설계를 통해 감성사전을 더 정확하게 검증할 수 있다.

본 연구에서는 문헌들을 고려하여 기존 4개의 감성사전을 검증하였다. 한국어 기반으로 서울대에서 개발한 KOSAC(Korean Sentiment Analysis Corpus)과 군산대에서 개발한 KNU(Kunsan University) 감성사전 2개를 고려하였다. 영어 감성사전인 Harvard IV(HV)와 Loughran and McDonald(LM)(2011)를 구글번역기를 사용하여 한국어로 번역하여 추가하였다. KOSAC과 KNU는 한국어 사전이지만 일반용이라는 약점을 가지고 있다. HV와 LM은 영어사전을 단순 번역하였다는 단점을 지니고 있다. 본 연구는 이를 보완한 기업 재무분석을 위한 한국어 감성사전(KOSELF)을 새롭게 구축하여 제시하였다.

감성사전을 검증하기 위해 한경 컨센서스에서 제공하는 2016년부터 2018년까지 한국에서 발행된 약 2만 개의 애널리스트 보고서 자료를 사용하였다. 이들 애널리스트 보고서를 통해 추천의견, 목표주가, 텍스트를 추출하였다. 각 감성사전을 사용하여 부정어 비율과 긍정어 비율 등의 감성변수를 보고서 별로 계산하였다. 이후 감성변수들과 애널리스트 보고서의 추천의견 및 목표주가와의 관계를 회귀분석 방법을 통해 검증하였다.

본 연구의 주요 결과는 다음과 같다. 첫째. 기존의 일반 한국어 사전은 기업분석에 적합하지는 않았다. 애널리스트 보고서 텍스트를 대상으로 빈도분석을 실시한 결과, 부정어 상위 30위에 해당하는 예상, 수준, 기준, 반영, 현재, 이유, 발생, 계획 등의 단어는 중립적인 성격이 강했다. 경상, 화재, 환자, 스포 등의 단어는 기업 재무분석에서 부정적인 의미와는 거리가 있었다. Loughran and McDonald(2011)의 주장과 동일하게, 한국어에서도 기업 재무분석을 위해서는 새로운 금융 감성사전이 필요함을 보여주고 있다.

둘째, 본 연구에서 새롭게 구축한 KOSELF 감성사전 변수들이 애널리스트의 추천의견 변경 및 목표주가 변경과 유의미한 관계를 가졌다. KOSELF 감성변수는 기존의 KOSAC, HV, LM 감성사전을 통해 계산한 변수들을 추가하였을 때도 여전히 통계적 유의성을 유지하며 더 우수한 성과를 보여주었다. 특히 목표주가 변경보다 추천의견을 변경한 경우 본문에 드러난 감성을 더욱 잘 추출하는 것을 확인하였는데, 일반적으로 투자자들이 추천의견 변경에 더욱 민감하게 반응함을 고려하면 KOSELF의 성능이 더욱 효과적으로 작용하는 것으로 판단된다. 이러한 결과는 감성변수를 다양하게 변화시켰을 때에도 동일한 양상을 보였다.

본 연구는 다음과 같이 구성된다. 제 2장은 관련 문헌에 대해 소개한다. 제 3장은 연구에 사용된 자료와 변수의 구성, 감성사전 검증을 위한 벤치마크에 대해 설명하며, 제 4장은 실증분석 결과를 제시한다. 제 5장에서는 본 연구를 마무리한다.

2. 문헌연구

2.1 텍스트 분석

빅데이터 기술의 향상에 따라 재무학에서도 텍스트를 분석하기 시작한 연구들이 나오고 있다. 텍스트 어조를 다루고 있는 연구는 크게 두 가지로 구분된다. 첫째는 긍정어, 부정어 등의 단어사전을 만들 후, 해당 단어가 문서에서 발생하는 빈도수를 이용하여 측정하는 방식이다. Tetlock(2007)은 Wall Street Journal(WSJ) 컬럼의 텍스트 어조를 정량적으로 측정하였으며, 비관적 어조가 이 후의 주식시장의 하락을 예측하며, 비관적 어조가 특이하게 낮거나 높은 후에 시장 거래량이 증가함을 발견하였다. Tetlock et al.(2008)은 WSJ과 Dow Jones News Service(DJNS)에서 개별 기업에 관한 부정적 단어들을 추적하여 변수를 산출하였다. 이 뉴스들의 부정적 어조가 회사이익과 주가의 하락을 예측하였으며, 특히 뉴스가 회사의 본질가치와 연관되어 있을 때 가장 강한 예측력을 보였다. Loughran and McDonald(2011)는 금융시장의 특수성으로 인해 일반 사전의 목록을 사용하면 정확성에 문제가 있을 수 있음 제시하였다. 그들은 금융 분야에 특화된 영문 단어목록을 제시하고, 이를 사용하면 정확성을 높일 수 있음을 보여주었다. 이후 Chen et al.(2014), Garca(2013)는 이 단어목록을 재무 분야의 텍스트 분석에 활용하였다.

둘째는 머신러닝 방법론이다. Antweiler and Frank(2004)는 Yahoo! Finance와 Raging Bull의 메시지 보드 텍스트를 나이브 베이지 알고리즘(Naive Bayes algorithm)을 사용하여 호황신호(bullishness signal) 변수를 계산하였으며, 이 변수가 시장 변동성을 예측할 수 있음을 보여주었다. Li(2010)과 Li et al.(2013)는 나이브 베이지안 머신러닝 알고리즘을 사용하여 10-K 보고서의 Management Discussion and Analysis(MD&A) 섹션을 분석하여 기업의 미래예견 지수와 기업을 둘러싸고 있는 경쟁의 정도를 측정하였다. Buehlmaier and Whited(2018)은 기업 연차보고서의 텍스트를 분석하여 기업의 자금조달 제약(financial constraint)을 측정하는 변수를 산출한 뒤, 이 변수를 사용하여 자금조달 제약이 심한 기업일수록 주가 수익률이 높아짐을 보여주었다.

한국에서 텍스트 분석을 재무에 적용한 경우는 많지 않다. Kim and Joh(2019)은 머신러닝을 통해 한국기업의 증권발행신고서의 비정형화된 텍스트 분석하였으며 이를 통해 IPO 수익률을 설명하였다. 한국 재무 분야에서 텍스트 분석이 활발하지 않은 이유 중 하나는 영어권 국가와 달리 한국어에 기반한 감성 사전이 부족한 것도 있다. 본 연구는 이런 점에서 재무분야의 연구에 도움이 될 수 있는 기업 재무분석을 위한 감성사전을 구축하여 향후 연구의 기반을 제공하고자 한다. 이를 위해 기존의 감성사전에 대한 문헌연구도 필요할 것이다. 이에 대해서는 다음에 제시하였다.

2.2 감성사전에 대한 연구

감성사전(Sentiment Lexicon)은 긍정, 부정 또는 중립으로 분류된 어휘 목록을 의미한다. 이를 통해 특정 문장 또는 글 수준의 감성 분석을 수행할 수 있다. 감성사전을 영어권에서 가장 발달하였기 때문에 이에 대해 먼저 설명한 후 한국어 감성사전에 대해 살펴보고자 한다.

가장 널리 활용되는 영어 감성사전은 SentiWordNet이다. 이는 워드넷(WordNet)의 synset이라 불리우는 유의어 집합에 대해, 해당 단어의 유의어, 반의어로 집합을 확장한 후, 단어들을 분류기로 학습하여 긍정, 부정, 객관성에 대한 값을 부여하였다(Baccianella et al., 2010). 하지만 재무 분야에서 Harvard IV에서 제공하는 긍정어와 부정어 리스트를 사용하여 감성을 측정한 연구들이 더 많다(Tetlock, 2007; Tetlock et al., 2008). Harvard IV은 하버드 대학에서 개발한 일반적 용도의 감성사전으로 긍정(positive), 부정(negative) 외에도 즐거움(pleasure), 고통(pain), 선(virture), 악(vice) 등 여러 감성 목록을 제공하고 있다. 재무분석에서는 보통 긍정과 부정 두 범주를 주로 사용하며, 여기여 더해서 강한(strong)과 약한(week), 능동적(active)과 수동적(passive) 범주에 해당하는 단어들도 함께 사용하기도 한다.

한국어에서 감성은 어떻게 측정할 것인지 여러 대안들을 생각할 수 있다. 첫째, 영어 감성사전을 번역하여 사용하는 것을 생각할 수 있다. 널리 활용되는 영어 감성사전을 구글로 번역하여 한글사전으로 사용할 수 있다. Kim et al.(2018)는 일반적인 긍정어/부정어를 다룬 Harvard IV 단어 목록을 Google Translate로 번역한 후 수작업으로 단어들의 한글 활용형을 적용시켜서 한국어 긍정어/부정어 목록을 만들었다. 이들은 이를 한국의 뉴스 기사에 적용하여 여기서 도출된 감성이 실제 북한의 도발을 예측할 수 있는지 분석하였다.

하지만 영어를 번역하여 한국어에 그대로 적용할 때 문제점이 발생할 수 있다. 첫째, 언어 단위가 다를 수 있다. 예를 들어, 한국어 단어가 영어에서는 구로 존재할 수 있다. 한국어 '신물나다'는 영어로 'sick of'라는 구로 존재한다. 둘째, 감성 정도에서 차이가 있을 수 있다. 화를 의미하는 한국어 '노발대발하다'는 감성 정도가 7.7점으로 아주 높지만 이에 대응하는 영어 단어 'infuriate'는 2.5점으로 감성 정도가 그리 크지 않다. 셋째, 의미의 범위 자체가 서로 다를 수 있다. 예를 들어, 영어 'angry'는 한국어로 '화나다', '노하다', '분하다', '약오르다', '성질나다' 등 다양한 한국어 어휘로 존재할 수 있다. 반대로 한국어 단어가 여러 개의 영어와 대응되는 경우도 존재할 것이다(Park et al., 2018).

대안으로 한국어의 특성을 살린 새로운 감성 사전을 개발하는 것을 생각할 수 있다. 이미 이러한 노력들은 이미 진행되어 왔다. Shin et al.(2016a)은 SentiWordNet을 기반으로 감성 어휘를 추출한 후 한국어 Deco 사전을 확장하는 방식으로 DecoSelex라는 감성사전을 구축하였다. 하지만 구축된 DecoSelex 감성 사전은 현재 공개되고 있지 않다. 오픈 한글(http://openhangul.com/restrict)은 집단지성을 사용하여 참여자가 단어에 대해 긍정, 부정, 중립을 투표하고 누적함으로 단어에 대한 감성 정보를 제공해주는 오픈 서비스로 다양한 분양의 감성사전을 제공하였지만, 이 역시 지금은 오픈 서비스의 한계적인 문제로 운영 중단된 상태이다(An and Kim, 2015). KOSAC(Korean Sentiment Analysis Corpus)은 서울대에서 개발한 말뭉치를 사용한 한국어 감성사전으로, 형태소 단위의 감정 특성을 제공하는 것이 특징이다(Shin et al., 2012; Shin et al., 2016b). KNU-한국어 감성사전은 군산대에서 개발한 감성사전으로, 표준국어대사전의 뜻풀이의 감성을 Bi-LSTM을 활용하여 긍정과 부정으로 분류하였다. 이 사전은 감성 어휘를 1-gram, 2-gram, 어구 그리고 문형 등 다양한 형태로 제공하고 있으며, SentiWordNet, SenticNet 등 외부 소스를 활용하여 감성 어휘를 확장하고 온라인 상 사용되는

신조어, 이모티콘도 포함하고 있는 것이 특징이다(Park et al., 2018). 본 연구는 현재 일반인에게 공개되어 사용 가능한 KOSAC과 KNU-한국어 감성사전 2개를 모두 사용하여 기업재무분석에서의 활용도를 검증하였다.

여기서 추가로 고려해야 할 요소 중 하나는 범용 감성사전을 사용할 것인지, 특정 분야에 적합하게 개발된 감성사전을 사용할 것이지 여부이다. 재무와 같은 특정 분야에서 텍스트의 감성을 측정하고 싶다면 범용 감성사전으로는 불완전할 수 있다. Loughran and McDonald (2011)는 1994~2008년 미국 기업의 사업보고서 텍스트를 범용 감성사전인 Harvard IV를 사용하여 분석하였을 때, 부정어로 특정된 단어들의 4분의 3 정도가 재무적 관점에서 보았을 때 부정적인 의미를 가진 단어가 아님을 발견하였다. 그들을 재무적 특성을 고려한 새로운 부정어 목록 사전을 구축하였으며, 이 단어 목록을 사용한 부정어 변수가 주가 수익률이나 기업 이익발표치 등을 더 잘 예측함을 보여주었다. 이는 범용 한국어 감성사전이 있지만 재무분석을 위해서는 더 특화된 사전이 필요할 수 있음을 의미한다.

이런 문헌들을 고려하여 본 연구에서는 기존 4개의 감성사전을 검증하였다. 한국어 기반으로 개발된 KOSAC과 KNU-한국어 감성사전 2개에, 영어 감성사전인 Harvard IV(HV)와 Loughran and McDonald(LM)(2011)를 구글번역기를 사용하여 한국어로 번역한 사전 2개를 추가하였다. KOSAC과 KNU는 한국어 사전이지만 범용이라는, HV와 LM은 영어사전을 단순 번역하였다는 단점을 각각 지니고 있다.[3] 본 연구에서는 이러한 문제점을 보완한 기업 재무분석을 위한 한국어 감성사전(KOSELF)을 새롭게 구축하여 제시하였다.

3. 자료

3.1 자료 및 변수

본 연구에서 분석 대상이 되는 텍스트는 애널리스트 보고서이다. 애널리스트 보고서 자료는 한경 컨센서스 웹 사이트를 통하여 추출하였다(http://consensus.hankyung.com). 2016년부터 2018년까지 3년 표본기간 중에서 코스피 200에 포함된 종목을 대상으로 애널리스트 보고서 pdf 파일을 파이썬 프로그래밍을 통해 자동으로 크롤링하였다. 여기에는 보고서 대상 기업, 애널리스트의 소속 증권사명, 작성 애널리스트 이름, 최종 추정일, 투자의견, 목표주가 등도 포함되어있다. 각 보고서에 위의 모든 항목이 다 존재하는 것은 아니다. 예를 들어, 투자의견이 있더라도 목표주가나 EPS 예측치가 존재하지 않는 관측치도 있으며, 반대로 투자의견이 없더라도 목표주가나 보고서 제목이 있는 경우도 있다. 이후 각 보고서 본문 내용에 대한 텍스트 변환 작업을 수동으로 진행하였다. 텍스트 변환 작업에서 추출한 내용은 보고서 제목 및 부제목을 제외한 본문 전문이다.

3) 기존 사전의 단점들은 <표 3>의 애널리스트 보고서 전체 표본을 대상으로 각 감성사전의 부정어와 긍정어 30순위 목록을 통해 확인할 수 있다. 이에 대한 자세한 내용은 '4.2 기업 재무분석을 위한 한국어 감성사전의 필요성 및 구축' 부분에 설명하였다.

통제변수로 사용하기 위해 기업의 특성변수도 필요하다. 기업의 특성변수는 Lee and Park (2019)의 선행연구를 참고하여 기업 시가총액, 거래량, 베타값, 자기자본의 시장가 대비 장부가치, YoY 이익증가율, 레버리지 관련 변수들을 FnGuide에서 제공하는 DataGuide에서 구하였다. 기업의 규모는 52주 평균 시가총액에 로그를 취하여 사용하였다. 거래량은 52주 평균 거래량에 마찬가지로 로그를 취한 값을 사용하였으며, 자기자본의 시장가 대비 장부가치(BM)는 자기자본의 연말 장부가치에 시가총액을 나누어서 구하였다. 레버리지(leverage)는 기업의 총부채를 총자산으로 나누어 계산하였다. 또한 해당 기업에 대해 당해 보고서를 발행한 애널리스트 수와 해당 연도에 발행된 총 보고서 수를 추가로 계산하여 사용하였다.

애널리스트 보고서의 텍스트는 실제 분석에 사용하기 이전 다음과 같은 전처리 과정을 거쳤다. 먼저 텍스트에 존재하는 특수문자 및 영문자, 숫자를 모두 제거하였다. 다음으로 향후 감성사전 단어와 애널리스트 보고서 내 본문과의 비교를 위해 토큰화(tokenizing)하였으며, 결과적으로 문장을 단어 및 형태소 기준으로 분리하였다. 토큰화는 한국어 자연어 처리를 위한 형태소 분석기로 널리 사용되어지는 KoNLPy(Korean NLP in Python)에 포함된 Komoran(KOrean MORphological ANalyzer) 형태소 분석 모듈을 사용하였다. KoNLPy(Park and Cho, 2014)는 서울대학교에서 한국어 정보처리를 위해 개발된 파이썬 패키지로, 형태소 분석을 통해 문장을 형태소 단위로 토큰화하고 각 토큰에 해당하는 품사(명사, 동사, 수사 등)를 함께 태그하여 반환한다. 기초적인 분석 단계에서 형태소 분석 및 품사 태깅(tagging) 과정을 실시하지 않고 본문 즉, 애널리스트 보고서 원문 자체와 감성사전 내 등록 단어를 직접적으로 비교할 수 있을 것이다. 그러나 이 경우 단어 본래의 의미를 손실하여 단순히 음절 자체의 존재여부를 비교하는 문제가 발생한다. 예를 들어 LM사전의 '비하(degradation)'의 단어는 사전에서 부정적 감성을 가지는 키워드이나, 본문 내 단순 출현여부만을 집계할 시 '준비하고'라는 의미적으로 일치하지 않는 어절에서 단어의 부정적 감성이 추출되는 문제가 발생한다. 따라서 형태소 분석을 통해 '준비/NNG(일반명사) + 하/XSV(동사 파생 접미사) + 고/EC(연결 어미)'와 같이 어절을 형태소 단위로 분할함으로써 이러한 오집계를 방지하고자 하였다. 이와 유사하게 다수의 선행연구(An and Kim, 2015; Lee, 2011)에서 이러한 감성어 사전을 활용한 감성 추출을 위해 형태소 분석이 선행되어야 함을 언급하고 있다.

애널리스트 보고서의 감성 변수는 긍정어 비율에서 부정어 비율을 차감한 의견변수(OPN) 변수를 도입하였다. 애널리스트 보고서의 의견변수(OPN)는 다음과 같이 정의된다.

$$OPN_i = POS\%_i - NEG\%_i \tag{1}$$

여기서, 부정어과 긍정어 변수는 다음과 같이 구하였다.

$$NEG\% = \frac{\text{부정어수}}{\text{부정어수} + \text{긍정어수}} \tag{2}$$

$$POS\% = \frac{\text{긍정어수}}{\text{부정어수} + \text{긍정어수}} \tag{3}$$

애널리스트 보고서별로 감성사전을 적용하여 대응하는 부정어와 긍정어의 수를 찾아낸 후 식 (2), (3) 그리고 (1)에서 정의된 변수들을 계산하였다. 또한 이 과정에서 단어의 길이가 1인 '힘', '꿈' 등의 단어는 제외하였다. NEG% 또는 POS% 변수가 클수록 해당 애널리스트 보고서가 부정 또는 긍정의 어조가 강함을 의미하게 된다. 따라서 OPN > 0의 경우 해당 애널리스트 보고서가 긍정의 어조이며, 수치가 클수록 강한 긍정의 어조임을 의미한다. 반대로 OPN < 0이면서 그 수치가 클수록 강한 부정의 어조임을 나타낸다. 분모로 전체 단어 수를 넣기도 하지만 사전을 구성하는 단어 크기의 영향을 제거하기 위해 부정어와 긍정어의 수의 합을 분모로 사용하였다. 전체 단어 수를 분모로 사용한 감성사전 변수에 대해서도 나중에 강건성 분석을 시행하였다.

3.2 감성사전 검증을 위한 벤치마크

애널리스트 보고서는 애널리스트가 직접 추천의견이나 목표주가를 갱신하며 이를 뒷받침하는 자신의 논리를 텍스트로 작성한다는 측면에서 감성사전을 검증하기 좋은 자료가 된다. 애널리스트 자신이 제시한 추천의견이나 목표주가가 자신의 텍스트의 감성에 대한 정확한 벤치마크가 될 수 있다.

본 연구에서는 주요 벤치마크로 추천의견과 목표주가 중에서 목표주가를 우선적으로 고려하였다. 추천의견은 보통 5개의 범위로 구분되어 추천의견이 변경되는 경우가 희박한데 비해 목표주가는 더 자유롭게 변경될 수 있는 여지가 많다. <표 1>의 추천의견 변경과 목표주가 변경의 요약통계량을 통해서도 이를 확인할 수 있다. 따라서 목표주가 변경을 벤치마크로 사용하는 것이 텍스트의 감성을 검증하기에 더 좋다고 판단하였다.

목표주가 변경은 이전 목표주가대비 애널리스트가 새롭게 제시한 목표주가의 변화를 의미한다. 이에 대한 측정치인 목표주가 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 이전 보고서의 목표주가를 차감한 값을 이전 목표주가로 나눈 비율이다.

나중에 강건성 검증을 위해 추천의견 변경도 벤치마크로 사용하였다. 추천의견을 사용하기 위해 다음과 같이 방법으로 계량화하였다. 추천의견을 매도, 비중축소, 중립, 매수, 적극매수의 5개 범위로 구분하였으며, 각 범주에 대해 매도 = 1, 비중축소 = 2, 중립 = 3, 매수 = 4, 적극매수 = 5 값을 임의로 부여하였다. 추천 투자등급은 증권사마다 차이가 존재하지만, Lee and Choi (2003), Kim and Eum(2006), Kim(2010)의 분류방식을 그대로 차용하였다.

추천의견 자체가 의미가 있기도 하지만, 재무 연구에서는 추천의견이 변경되는 경우가 더 많은 정보력을 가지고 있음을 지지하고 있다. 따라서 검증을 위한 벤치마크로 애널리스트 보고서의 추천의견 변경을 사용하였다. 추천변경(ΔRECOMM)은 특정 기업의 해당 애널리스트 보고서 추천의견 값에서 이전 보고서의 추천의견 값을 차감하여 구하였다. 따라서 추천변경 -2는 추천의견이 2단계 하락한 경우를, -1은 1단계 하락한 경우를, 0은 그래도 유지되는 경우를, 1은 1단계 상승한 경우를, 2는 2단계 상승한 경우를 나타낸다.

〈표 1〉 애널리스트 보고서의 요약통계량

본 표는 2016년부터 2018년까지의 애널리스트 보고서의 추천의견 및 목표주가의 요약통계량을 보여주고 있다. 애널리스트 보고서는 추천의견에 따라 매도, 비중축소, 중립, 매수, 적극매수의 5개 범위로 구분하였으며, 각 범주에 다음의 값을 임의로 부여하였다. 매도 = 1, 비중축소 = 2, 중립 = 3, 매수 = 4, 적극매수 = 5. 추천변경(ΔRECOMM)은 특정 기업의 애널리스트 보고서 추천의견 값에서 이전 보고서의 추천의견 값을 차감한 수치이다. 따라서 추천변경 -2는 추천의견이 2단계 하락한 경우를, -1은 1단계 하락한 경우를, 0은 그래도 유지되는 경우를, 1은 1단계 상승한 경우를, 2는 2단계 상승한 경우를 나타낸다. 목표주가의 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 이전 보고서의 목표주가를 차감한 비율이다. 분석대상 보고서의 목표주가 변화율 분포에 따라 목표주가가 동일한 경우 3의 값을 기준으로 다음과 같이 나누었다. 목표주가 하향인 경우 중위수를 기준으로 1, 2그룹으로, 목표주가 상향인 경우 중위수를 기준으로 4, 5그룹으로 구분하였다. 괄호 안의 값은 각 그룹의 평균을 보여준다.

패널 A: 애널리스트 보고서 수집 및 분석 제외대상

	제외 샘플 수	샘플 수	
최초 수집 애널리스트 보고서		46,750	
(-) 분기별 1회 미만 발행 보고서	25,585	21,165	
(-) 투자의견 결측 보고서	272	20,893	
(-) 투자의견 Not Rated 보고서	55	20,838	
(-) 본문 텍스트 수집 불가 보고서	3,207	17,631	
(-) 분석기간 내 최초발행 보고서	1,174	16,457	(최종 분석대상)

패널 B: 분석대상 보고서의 요약통계량

	매도	비중축소	중립	매수	적극매수	의견 없음	결측치	총계
리포트수	1	16	1,286	16,229	78	348	3,207	21,165
(%)	0.006	0.091	7.303	92.158	0.443			100
ΔRECOMM								
-2	1	0	0	0	0			1
-1	0	3	248	16	0			267
0	0	10	930	14,948	59			15,947
1	0	0	1	228	10			239
2	0	0	0	2	1			3
전체	1	13	1,179	15,194	70	-		16,457
ΔTPRC								
그룹 1(-22%)	1	1	192	1,165	2			1,361
2(-7%)	0	3	113	1,184	0			1,300
3(0%)	0	8	645	9,932	57			10,642
4(7%)	0	0	97	1,453	5			1,555
그룹 5(22%)	0	1	101	1,437	6			1,545
관측치	1	13	1,179	15,194	70	-		16,457
(평균)	(-11.5%)	(-1.1%)	(-4.5%)	(0.4%)	(2.8%)			(-2.8%)

4. 실증분석 결과

4.1 요약통계량

<표 1>은 2016년부터 2018년까지의 애널리스트 보고서를 통해 추출한 변수들에 대한 요약

통계량을 보여주고 있다. 추천의견의 분포를 기준으로 하여 벤치마크 변수들의 분포를 제시하였다. 2016년부터 2018년까지 코스피200 종목을 대상으로 발행된 총 46,750개 보고서 중에서 분석기간 내 12번 이상 즉, 평균적으로 분기별 1회 이상 해당 종목에 대해 추천의견을 제시한 21,165개의 보고서를 분석대상으로 하였다. 이 중에서 분석기간 내 최초로 발행되어 전기(t-1) 정보를 추출할 수 없는 1,174개 경우를 제외하였으며, 투자의견이 아예 존재하지 않는 272개 보고서, Not Rated로 제시된 55개 보고서를 제외하였다. 또한 일반적 정보는 존재하나 pdf를 통해 텍스트 원문을 추출할 수 없거나, 인코딩 문제로 분석이 불가한 3,207개 경우를 결측치로 제외하였다. 따라서 최종적으로 분석에 사용한 애널리스트 보고서 수는 총 16,457개의 보고서이다. 추천의견이 존재하는 애널리스트 보고서는 매도, 비중축소, 중립, 매수, 적극매수의 5개 범위로 구분하였다. 이중 매도는 1건(0.006%), 비중축소는 16건(0.091%)으로 중립보다 낮은 추천의견을 제시하는 경우가 매우 적게 나타났다. 전체 추천의견 중 매수가 16,229건(92.158%)로 대부분을 차지하였다.

추천변경(ΔRECOMM) 변수는 이전 애널리스트 보고서와의 차이를 측정하기 때문에 관측치가 축소된다. 분석기간 내 최초로 발행되어 전기(t-1) 정보를 추출할 수 없는 1,174개 경우를 제외하고 총 16,457개가 남는다. 매도 의견을 보면 중립 의견에서 2단계 내려온 보고서 1개이다. 가장 관측수가 많은 것은 매수 의견이 그대로 유지되고 있는 경우로 보고서 수가 14,948개이다. 그 다음은 중립 의견이 그대로 유지되고 있는 경우가 930개이다. 매수 의견에서 중립 의견으로 1단계 하향한 경우가 248개, 반대로 중립 의견에서 매수 의견으로 1단계 상향한 경우가 228개로 1단계 의견 변경은 주로 매수와 중립 의견을 한 번씩 변경하여 제시하는 경우가 다수임을 알 수 있다. 적극매수 의견은 중립 의견에서 2단계 상향한 경우가 1개, 매수 의견에서 1단계 상향한 경우가 10개, 적극매수 의견이 그대로 유지되고 있는 경우가 59개로 대부분이었다.

다음은 목표주가 변화율(ΔTPRC)의 분포를 5개의 그룹으로 나누어서 보여주고 있다. 그룹 3인 목표주가 변화가 없는 표본을 의미하며 10,642개로 가장 많다. 하지만 목표주가가 하향한 표본은 2,661개, 상향한 표본은 3,100개로 그 수가 추천의견 변경에 비해 월등하게 많다. 애널리스트들이 추천의견에 비해 목표주가는 상대적으로 더 자유롭게 변경하고 있음을 확인할 수 있다.

목표주가 변화율(ΔTPRC)을 추천의견 별로도 보여주고 있다. 각 열의 값은 추천의견 범주에 해당하는 보고서들의 목표주가 변화율(ΔTPRC)의 평균값을 나타낸다. 목표주가의 변화율의 평균을 보면, 매도가 -0.115%, 비중축소가 -0.011%, 중립이 -0.045%, 매수가 0.004%, 적극매수가 0.028%로 추천의견 등급이 높을수록 마찬가지로 목표주가 변화율도 높아지는 경향이 있다.

<표 2>는 애널리스트 보고서 텍스트를 사용하여 계산한 한국어 감성변수의 요약통계량을 보여주고 있다. 패널 A는 OPN 변수를 보여주고 있다. LM의 OPN의 평균은 0.002로 가장 작은데 비해 KNU의 OPN의 평균이 0.692로 가장 크다. KOSELF의 OPN의 평균은 0.319으로 5개 사전의 중간에 해당한다. 표준편차에 있어서는 KOSELF가 0.645로 다른 사전들에 비해 월등히 크다. 이는 KOSELF가 텍스트의 부정과 긍정 어조를 넓은 범위의 수치 안에서 균형있게 잘 반영하고 있음을 의미한다.

〈표 2〉 한국어 감성사전 변수의 요약통계량

본 표는 2016년부터 2018년까지 발행된 애널리스트 보고서 본문 텍스트를 사용하여 추출한 한국어 감성사전별 긍·부정어 감성변수의 요약통계량을 보여주고 있다. 패널 A는 각 감성사전별 POS%에서 NEG%를 차감한 OPN 변수의 평균값을 보고하고 있다. 따라서 OPN < 0인 경우 보고서의 부정어 비율이 긍정어 비율보다 높으며, OPN > 0인 경우 반대로 긍정어 비율이 부정어 비율보다 높은 경우에 해당한다. OPN = 0인 경우 부정어와 긍정어 비율이 동일하여 중립적 보고서임을 의미한다. 패널 B는 감성사전 변수의 평균 및 산포에 대한 통계량을 보고하고 있다. 감성변수는 보고서에 대해 각 사전별로 부정어와 긍정어 출현빈도를 총 부정어 및 긍정어 출현빈도로 나누어 비율(%)로 나타내었다. 패널 C는 분석 대상이 되는 기업특성 변수로서 종목 거래량(52주 평균에 로그를 취한 값), 기업의 규모(52주 평균 시가총액에 로그를 취한 값), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수에 대한 요약통계량을 보고하고 있다.

	평균	표준편차	1사분위	중간값	3사분위
패널 A: OPN 변수					
KOSAC OPN	0.157	0.198	0.027	0.160	0.292
KNU OPN	0.692	0.341	0.520	0.778	1.000
HV OPN	0.364	0.258	0.200	0.368	0.538
LM OPN	0.002	0.367	-0.250	0.000	0.250
KOSELF OPN	0.319	0.645	0.000	0.400	1.000
패널 B: 감성사전 변수					
KOSAC NEG%	0.422	0.099	0.354	0.420	0.486
KNU NEG%	0.154	0.171	0.000	0.111	0.240
HV NEG%	0.318	0.129	0.231	0.316	0.400
LM NEG%	0.499	0.183	0.375	0.500	0.625
KOSELF NEG%	0.340	0.323	0.000	0.300	0.500
KOSAC POS%	0.578	0.099	0.514	0.580	0.646
KNU POS%	0.846	0.171	0.760	0.889	1.000
HV POS%	0.682	0.129	0.600	0.684	0.769
LM POS%	0.501	0.183	0.375	0.500	0.625
KOSELF POS%	0.660	0.323	0.500	0.700	1.000
패널 C: 기업특성 변수					
LN_거래량	12.491	1.211	11.707	12.408	13.301
Beta	0.894	0.489	0.556	0.847	1.232
LN_시가총액	15.584	1.399	14.539	15.583	16.528
BM	0.939	0.581	0.500	0.847	1.250
Leverage	0.513	0.206	0.348	0.515	0.636
YoY_Growth	-1.317	286.676	-0.185	0.064	0.323
애널리스트수	179	136	80	153	234
리포트수	9.35	5.42	5.00	9.00	12.00

패널 B는 각 사전별로 부정어와 긍정어 수를 구하고 이를 사용하여 부정과 긍정의 감성변수를 계산하였다. 5개의 사전 모두에서 평균적으로 긍정어 비율이 부정어보다 커 애널리스트 보고서에서는 긍정적 단어를 더 많이 사용하는 것으로 나타났다. 먼저 부정어 감성변수(NEG%)를

살펴보면 재무 특화사전인 LM에서 0.499로 가장 부정적 감성을 추출하는 것이 나타났다. 특히 KNU의 경우 평균이 0.154, 1사분위 값이 0으로 부정적 감성을 적게 반영하고 있다. 긍정어 감성변수(POS%)는 부정어 감성변수와 반대로 나타나고 있다.

4.2 기업 재무분석을 위한 한국어 감성사전의 필요성 및 구축

기존 감성사전의 현황을 살펴보기 위해 애널리스트 보고서 전체 표본을 대상으로 감성사전의 부정어와 긍정어의 빈도에 대해 살펴보았다. <표 3>은 각 한국어 감성사전별로 30순위 빈도에 해당하는 단어 목록을 보고하고 있다. 패널 A는 부정어 순위 목록을 보여준다. KOSAC은 한국어 사전이지만 일반 사전으로써 기업 재무분석을 하기에는 부족함을 드러낸다. 1위부터 8위인 예상, 증가, 유지, 수준, 지속, 기준, 반영, 수익 등의 단어는 중립적인 성격이 강할뿐만 아니라, 오히려 긍정적인 성향에 더 가까워보인다. 중국, 현재, 안정, 전략, 이유, 기업, 환경 등도 중립적인 단어이다. KNU 감성사전도 경상, 벗어나, 화재, 지지, 환자, 스포, 잉여 등의 단어는 기업 재무분석에서 부정적인 의미와는 거리가 있다. HV는 일반어 사전이며 한국어로 번역된 영어 사전이라는 한계를 가지고 있다. 발생, 평균, 가장, 계획, 견인, 상태, 거래 등의 단어는 기업 재무분석에 있어서는 중립적이거나 오히려 긍정적일 것으로 보인다. 상대, 제외 등의 단어는 각각 영단어 opponent, exclusion을 구글 및 파파고(네이버) 번역기로 번역한 것인데 한글에서는 적합하지 않아 보인다. LM은 영어로 기업 재무분석에 적합하게 구축되었다. 이에 걸맞게 감소, 하락, 사의, 조정, 하향 등의 단어들이 높은 순위를 보이고 있다. 하지만 확대, 매수, 전환, 강화, 해소, 개발 등의 단어는 각각 영단어 escalate, bribe, diversion, tightening, dissolution, exploitation을 번역한 것인데, 한국어로는 부자연스러워 보인다. 이는 원 단어를 번역함에 있어 확대(악화), 매수(뇌물), 전환(우회하다, 바꾸다), 강화(조이다, 긴축), 해소(파경, 해산, 해체), 개발(착취) 등 한 영단어 당 다양한 한글로 나타남에 따라 발생하는 문제라고 할 수 있다.

패널 B는 긍정어의 빈도에 따른 순위 목록을 보여주고 있는데, 부정어와 비슷한 성향을 보여주고 있다. KOSAC, KNU, HV는 일반어 사전으로써의 단점을 가지고 있다. KOSAC의 의견, 기록, 기존, 가격, 판매, 국내, 수요, 요인, 규모, 평가, 필요, 가치, 공급, 생산, KNU의 이익, 수익, 가치, 함께, 대상, HV의 이익, 추정, 효과, 관련, 동기, 다만, 가치, 고려, 지배 등은 기업 재무분석 시에는 중립적으로 사용될 단어들이다. 부정어에서와 마찬가지로 HV와 LM은 영어 사전으로써의 단점을 가지고 있다. 이상, 제외, 매우, 그럼에도 불구하고 등의 단어들은 한국어에서는 부정어와도 결합될 수 있어 긍정어라 하기 어려운 측면이 있다.

본 연구는 이러한 각 사전의 단점을 극복하고 장점을 결합하여 기업 재무분석에 적합한 한국어 전용 감성사전을 구축하고자 하였다. 구체적으로는 다음과 같은 과정을 따라서 새로운 감성사전을 구축하였으며, 이를 KOSELF(Korean Sentiment Lexicon for Finance)라고 명명하였다. 먼저 애널리스트 보고서에서 사용하는 단어의 출현빈도 분석을 수행하고, 분석 결과에 기반하여 재무 분야에 적절하다고 판단되는 단어를 추출함으로써 초기 seed 사전을 구성하였다. 이후 국문 감성사전인 KOSAC과 KNU와 영문 재무 특화사전인 LM의 구성단어를 비교하여 이

〈표 3〉 한국어 감성사전별 부정어/긍정어 상위 30개 리스트

본 표는 한국어 감성사전별 부정어 및 긍정어의 출현빈도 상위 30개 단어 목록을 보고하고 있다. 분석 표본은 2016년부터 2018년까지 코스피 200 종목에 대해 발행된 애널리스트 보고서의 본문이다. 전처리 후 토큰화한 보고서 본문 텍스트에서 각 단어가 출현한 문서의 빈도와 비중(%)을 함께 보고하고 있다. 패널 A와 B는 각각 부정어와 긍정어 리스트를 나타낸다. KOSAC과 KNU는 국문 일반사전, HV는 영문 일반사전, LM은 영문 재무특화사전, KOSELF는 본 연구진의 국문 재무특화사전을 나타낸다. KOSELF 단어목록은 다음 웹 사이트를 참고하시오(https://sites.google.com/view/cheolwon-yang/koself?authuser=0).

패널 A: 부정어

순위	KOSAC				KNU				HV				LM				KOSELF			
	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)
1	예상	12,325	0.05	0.05	부진	5,849	0.32	0.32	감소	7,758	0.08	0.08	감소	7,758	0.08	0.08	감소	7,758	0.10	0.10
2	증가	11,939	0.05	0.10	부담	3,277	0.18	0.50	하락	7,015	0.07	0.15	확대	7,180	0.07	0.15	하락	7,015	0.09	0.19
3	유지	11,751	0.05	0.15	부정	1,390	0.08	0.58	비용	6,213	0.06	0.21	하락	7,015	0.07	0.22	부진	5,849	0.07	0.26
4	수준	8,623	0.04	0.19	부족	791	0.04	0.62	우려	4,113	0.04	0.25	매수	5,321	0.05	0.27	하향	3,412	0.04	0.30
5	지속	8,157	0.04	0.23	경상	620	0.03	0.65	발생	3,850	0.04	0.29	조정	4,768	0.05	0.32	전환	3,325	0.04	0.34
6	기준	7,313	0.03	0.26	손해	529	0.03	0.68	평균	3,762	0.04	0.33	하향	3,412	0.03	0.35	부담	3,277	0.04	0.38
7	반영	7,187	0.03	0.29	벗어나	504	0.03	0.71	이상	3,729	0.04	0.37	전환	3,325	0.03	0.38	축소	3,128	0.04	0.42
8	수익	6,815	0.03	0.32	훼손	389	0.02	0.73	전환	3,325	0.03	0.40	부담	3,277	0.03	0.41	적자	2,996	0.04	0.46
9	비용	6,213	0.03	0.35	위축	383	0.02	0.75	부담	3,277	0.03	0.43	축소	3,128	0.03	0.44	둔화	2,451	0.03	0.49
10	부진	5,849	0.03	0.38	어려움	286	0.02	0.77	적자	2,996	0.03	0.46	적자	2,996	0.03	0.47	제외	2,442	0.03	0.52
11	이후	5,564	0.02	0.40	힘들	252	0.01	0.78	가장	2,681	0.03	0.49	지연	2,588	0.03	0.50	손실	2,439	0.03	0.55
12	중국	5,242	0.02	0.42	한계	246	0.01	0.79	역시	2,553	0.03	0.52	강화	2,568	0.03	0.53	악화	1,637	0.02	0.57
13	현재	4,469	0.02	0.44	최악	222	0.01	0.80	제한	2,468	0.02	0.54	둔화	2,451	0.02	0.55	매각	1,622	0.02	0.59
14	우려	4,113	0.02	0.46	불안	218	0.01	0.81	제외	2,442	0.02	0.56	제외	2,442	0.02	0.57	부정	1,390	0.02	0.61
15	안정	4,094	0.02	0.48	실망	167	0.01	0.82	손실	2,439	0.02	0.58	손실	2,439	0.02	0.59	지연	1,294	0.02	0.63
16	상황	4,062	0.02	0.50	화재	148	0.01	0.83	계획	2,138	0.02	0.60	악화	1,637	0.02	0.61	경쟁사	1,288	0.02	0.65
17	발생	3,850	0.02	0.52	지지	146	0.01	0.84	인상	1,963	0.02	0.62	매각	1,622	0.02	0.63	불가피	1,264	0.02	0.67
18	이상	3,729	0.02	0.54	위기	135	0.01	0.85	경쟁	1,947	0.02	0.64	해소	1,533	0.02	0.65	절감	1,166	0.01	0.68
19	부담	3,277	0.01	0.55	충격	129	0.01	0.86	지배	1,729	0.02	0.66	부정	1,390	0.01	0.66	문제	849	0.01	0.69
20	제한	2,468	0.01	0.56	피해	123	0.01	0.87	악화	1,637	0.02	0.68	완화	1,353	0.01	0.67	비하	797	0.01	0.70
21	전략	2,433	0.01	0.57	주춤	123	0.01	0.88	상대	1,462	0.01	0.69	개발	1,318	0.01	0.68	부족	791	0.01	0.71
22	이유	2,099	0.01	0.58	아쉬움	109	0.01	0.89	부정	1,390	0.01	0.70	경쟁사	1,288	0.01	0.69	중단	693	0.01	0.72
23	장기	1,956	0.01	0.59	약점	99	0.01	0.90	견인	1,367	0.01	0.71	불가피	1,264	0.01	0.70	수치	676	0.01	0.73
24	경쟁	1,947	0.01	0.60	환자	91	0.00	0.90	지연	1,294	0.01	0.72	절감	1,166	0.01	0.71	위험	659	0.01	0.74
25	공장	1,938	0.01	0.61	소외	91	0.00	0.90	포인트	1,193	0.01	0.73	문제	849	0.01	0.72	공격	640	0.01	0.75
26	변화	1,907	0.01	0.62	걱정	87	0.00	0.90	서비스	1,157	0.01	0.74	비하	797	0.01	0.73	정체	610	0.01	0.76
27	일부	1,875	0.01	0.63	스포	86	0.00	0.90	상태	1,124	0.01	0.75	부족	791	0.01	0.74	결국	602	0.01	0.77
28	기업	1,819	0.01	0.64	마이너스	82	0.00	0.90	자본	928	0.01	0.76	종료	784	0.01	0.75	처리	550	0.01	0.78
29	확인	1,789	0.01	0.65	불황	78	0.00	0.90	거래	924	0.01	0.77	중단	693	0.01	0.76	재평가	540	0.01	0.79
30	지배	1,729	0.01	0.66	잉여	77	0.00	0.90	문제	849	0.01	0.78	수치	676	0.01	0.77	손해	529	0.01	0.80

〈표 3〉 한국어 감성사전별 부정어/긍정어 상위 30개 리스트(계속)

패널 B: 긍정어

순위	KOSAC				KNU				HV				LM				KOSELF			
	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)	단어	빈도	비중(%)	누적(%)
1	전망	12,238	0.04	0.04	이익	14,201	0.20	0.20	이익	14,201	0.08	0.08	이익	14,201	0.21	0.21	이익	14,201	0.16	0.16
2	목표	11,331	0.04	0.08	개선	10,384	0.15	0.35	증가	11,939	0.07	0.15	개선	10,384	0.15	0.36	개선	10,384	0.11	0.27
3	대비	11,173	0.04	0.12	기대	8,120	0.11	0.46	상승	9,403	0.05	0.20	상승	9,403	0.14	0.50	상승	9,403	0.10	0.37
4	개선	10,384	0.03	0.15	수익	6,815	0.10	0.56	추정	9,114	0.05	0.25	안정	4,094	0.06	0.56	확대	7,180	0.08	0.45
5	의견	9,907	0.03	0.18	긍정	4,378	0.06	0.62	기대	8,120	0.05	0.30	회복	3,944	0.06	0.62	안정	4,094	0.05	0.50
6	시장	9,124	0.03	0.21	안정	4,094	0.06	0.68	효과	6,895	0.04	0.34	이상	3,729	0.06	0.68	회복	3,944	0.04	0.54
7	성장	8,583	0.03	0.24	가치	2,808	0.04	0.72	조정	4,768	0.03	0.37	진행	2,954	0.04	0.72	이상	3,729	0.04	0.58
8	기대	8,120	0.03	0.27	강세	1,758	0.02	0.74	추가	4,692	0.03	0.40	강화	2,568	0.04	0.76	진행	2,954	0.03	0.61
9	기록	7,557	0.02	0.29	할인	1,388	0.02	0.76	긍정	4,378	0.02	0.42	제외	2,442	0.04	0.80	가치	2,808	0.03	0.64
10	확대	7,180	0.02	0.31	함께	1,303	0.02	0.78	관련	4,137	0.02	0.44	달성	2,428	0.04	0.84	강화	2,568	0.03	0.67
11	효과	6,895	0.02	0.33	정상	1,253	0.02	0.80	안정	4,094	0.02	0.46	매력	2,426	0.04	0.88	달성	2,428	0.03	0.70
12	기존	5,915	0.02	0.35	성공	1,188	0.02	0.82	회복	3,944	0.02	0.48	매우	1,193	0.02	0.90	매력	2,426	0.03	0.73
13	영향	5,290	0.02	0.37	적극	1,100	0.02	0.84	발생	3,850	0.02	0.50	성공	1,188	0.02	0.92	확보	2,243	0.02	0.75
14	가격	5,082	0.02	0.39	수혜	1,051	0.01	0.85	이상	3,729	0.02	0.52	기회	818	0.01	0.93	강세	1,758	0.02	0.77
15	판매	4,671	0.02	0.41	충분히	1,026	0.01	0.86	동기	3,724	0.02	0.54	향상	723	0.01	0.94	완화	1,353	0.01	0.78
16	최근	4,467	0.01	0.42	능력	799	0.01	0.87	다만	3,168	0.02	0.56	발전	629	0.01	0.95	매우	1,193	0.01	0.79
17	비중	4,393	0.01	0.43	향상	723	0.01	0.88	진행	2,954	0.02	0.58	강점	341	0.01	0.96	성공	1,188	0.01	0.80
18	긍정	4,378	0.01	0.44	최고	691	0.01	0.89	가치	2,808	0.02	0.60	우수	323	0.00	0.96	적극	1,100	0.01	0.81
19	국내	4,294	0.01	0.45	대상	689	0.01	0.90	고려	2,577	0.01	0.61	호황	307	0.00	0.96	추천	1,095	0.01	0.82
20	수요	4,020	0.01	0.46	상승세	679	0.01	0.91	강화	2,568	0.01	0.62	해결	235	0.00	0.96	수혜	1,051	0.01	0.83
21	요인	3,908	0.01	0.47	꾸준히	667	0.01	0.92	구조	2,553	0.01	0.63	그럼에도 불구하고	219	0.00	0.96	충분히	1,026	0.01	0.84
22	규모	3,854	0.01	0.48	발전	629	0.01	0.93	달성	2,428	0.01	0.64	선호	197	0.00	0.96	추진	924	0.01	0.85
23	평가	3,201	0.01	0.49	완성	536	0.01	0.94	매력	2,426	0.01	0.65	협력	192	0.00	0.96	기회	818	0.01	0.86
24	축소	3,128	0.01	0.50	이벤트	479	0.01	0.95	시현	2,052	0.01	0.66	독점	180	0.00	0.96	향상	723	0.01	0.87
25	필요	3,059	0.01	0.51	흥행	388	0.01	0.96	보수	1,991	0.01	0.67	풍부	156	0.00	0.96	최고	691	0.01	0.88
26	예정	3,041	0.01	0.52	우수	323	0.00	0.96	확인	1,789	0.01	0.68	인기	155	0.00	0.96	증대	686	0.01	0.89
27	본격	2,888	0.01	0.53	상위	290	0.00	0.96	지배	1,729	0.01	0.69	혁신	143	0.00	0.96	상승세	679	0.01	0.90
28	가치	2,808	0.01	0.54	신뢰	267	0.00	0.96	자산	1,694	0.01	0.70	장점	139	0.00	0.96	발전	629	0.01	0.91
29	공급	2,627	0.01	0.55	안전	219	0.00	0.96	회사	1,480	0.01	0.71	확신	128	0.00	0.96	바닥	537	0.01	0.92
30	생산	2,596	0.01	0.56	희망	214	0.00	0.96	유사	1,440	0.01	0.72	획득	124	0.00	0.96	충족	507	0.01	0.93

중 저자들 간의 합의에 의해서 중립적이라고 판단되는 단어들은 제거하고 기업 재무분석에 적합한 한국어 단어들을 추가하였다. 최종 KOSELF 감성사전은 부정어 47개와 긍정어 48개로 각각 구성되었으며, 구체적인 단어목록은 웹 싸이트(https://sites.google.com/view/cheolwon-yang/koself?authuser=0)를 통해 공개하였다.

<표 3> KOSELF 감성 사전을 사용하여 애널리스트 전체 표본을 대상으로 추출한 부정어와 긍정어의 30순위 단어 목록을 보여주고 있다. 부정어 리스트를 살펴보면 감소, 하락, 부진, 하향, 적자, 부담, 축소, 둔화, 손실, 악화, 매각, 지연, 불가피, 절감 등 기업 재무분석에서 부정적 의미를 포함하는 단어들이 반영되어 나타났다. 실제 보고서 원문에서 '실적 부진', '적자 전환', '성장세 둔화' 등 일반적으로 bi-gram 차원에서 널리 사용되어지는 부정적 단어들 중 실적, 성장세 등 중립적 단어는 제외하고 집계된 것을 알 수 있다. 다음으로 KOSELF 긍정어 리스트를 살펴보면 이익, 개선, 상승, 확대, 안정, 회복, 달성, 매력, 확보, 적극, 추천, 수혜, 최고, 상승세 등 마찬가지로 기업 재무분석에서 활용되어지는 긍정적 감성의 단어들이 두드러지게 나타난다.

4.3 목표주가 변경을 사용한 한국어 감성사전의 성과 검증

애널리스트 보고서에서 목표주가의 변경은 일종의 투자지표로서 해석될 수 있다. 특히 국내 금융시장의 경우 매수 추천의견을 가지는 보고서가 대부분임을 고려하면, 동일한 매수 추천의견이라고 하더라도 애널리스트는 자신의 목표주가를 소폭 상향하거나, 하향 제시함으로써 해당 종목에 대한 수정된 의견을 간접적으로 반영할 수 있다. 따라서 직전 목표대비 현재의 목표주가 변경 비율은 애널리스트의 감성의 중요한 벤치마크가 될 수 있다. 아래와 같이 목표주가 변화율을 종속변수로 사용한 회귀분석 모형을 검증하였다. 설명변수는 감성사전 5개를 사용하여 계산한 OPN 변수들이다. 이를 통해 어떤 감성사전을 통해 추출한 애널리스트의 감성이 목표주가 변경과 더 밀접한 관계가 있는지 검증할 수 있을 것이다.

$$\Delta TPRC_i = \alpha_0 + \alpha_1 OPN_i + \alpha_2 Control\ Variables + \epsilon_i \quad (4)$$

여기서 ΔTPRC는 목표주가 변화율이며, OPN은 POS%에서 NEG%를 차감한 의견변수이다. NEG%(POS%)는 보고서 텍스트의 부정어(긍정어) 수를 부정어와 긍정어의 합으로 나눈 비율이다. 통제변수(Control Variables)는 Lee and Park(2019)을 참조하여 다음의 변수를 사용하였다. Nreport는 해당 연도에 해당 기업에 대한 보고서의 수, Nanalyst는 해당 연도에 해당 기업에 대해 보고서를 발표한 애널리스트의 수, 거래량은 해당 연도의 거래량에 로그값을 취한 값, 기업규모(Size)는 연말의 주가에 상장주식수를 곱한 값에 로그를 취한 값, B/M은 자기자본의 장부가치를 연말의 시가총액으로 나눈 값이다. 이외에도 기업의 이익 성장률(YoY Growth), 베타(beta), 그리고 레버리지 비율(leverage)을 추가하였다. 산업효과와 연도효과를 제거하기 위하여 한국산업표준분류 중분류코드와 각 연도를 더미변수로 변환하여 통제변수로 사용하였다.

회귀분석에 앞서 <표 4>에 목표주가 변경에 따른 감성사전의 부정어 및 긍정어 빈도에 대한 평균값을 나타내었다. 목표주가 변경의 정도를 5개 그룹으로 범주화하여 제시하였으며, 목표주가 변경치가 음의 방향으로 클수록 그룹 1, 반대로 양의 방향으로 클수록 그룹 5의 값을

가지도록 하였다. 대부분의 보고서는 목표주가 변경되지 않아 0인 경우가 많은데 이에 대해서는 또한 그룹 3에 할당하였다.

패널 A의 사전별 OPN의 평균값을 살펴보면 목표주가가 양의 방향으로 커질수록, 즉 그룹 1에서 그룹 5로 이동할수록 모든 사전에서 증가하는 경향을 보인다. 특히 KOSELF에서 0.02에서 0.57로 상승하여 변화의 크기가 가장 뚜렷하다. LM은 목표주가가 하락하는 그룹 1과 2에서

〈표 4〉 목표주가 변경에 따른 감성사전 변수

본 표는 애널리스트 보고서의 목표주가 변경에 따른 각 사전별 감성변수의 분포를 보여주고 있다. 애널리스트 보고서의 목표주가 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 직전 보고서의 목표주가를 차감한 값을 이전 목표주가로 나눈 비율이다. 표는 목표주가 변화율을 5개 범주 그룹으로 나누어 서술하였으며 1, 2그룹은 ΔTPRC < 0인 경우를, 3그룹은 ΔTPRC = 0인 경우를, 4, 5그룹은 ΔTPRC > 0인 경우를 나타낸다. 패널 A는 각 감성사전별 POS%에서 NEG%를 차감한 OPN 변수의 평균값을 보고하고 있다. 따라서 OPN < 0인 경우 보고서의 부정어 비율이 긍정어 비율보다 높으며, OPN > 0인 경우 반대로 긍정어 비율이 부정어 비율보다 높은 경우에 해당한다. OPN = 0인 경우 부정어와 긍정어 비율이 동일하여 중립적 보고서임을 의미한다. 패널 B는 각 사전별 부정어에 대한 평균값을, 패널 C는 각 사전별 긍정어에 대한 평균값을 나타내고 있다. 감성변수는 보고서에 대해 각 사전별로 부정어와 긍정어 출현빈도를 총 부정어 및 긍정어 출현빈도로 나누어 비율(%)로 나타내었다.

	KOSAC	KNU	HV	LM	KOSELF
패널 A: OPN(평균)					
전체	0.16	0.69	0.36	0.00	0.32
ΔTPRC					
1	0.13	0.57	0.30	-0.19	0.02
2	0.13	0.59	0.31	-0.16	0.06
3	0.15	0.69	0.36	0.02	0.34
4	0.18	0.77	0.43	0.09	0.45
5	0.20	0.79	0.44	0.10	0.57
패널 B: 부정어 빈도%(평균)					
전체	0.42	0.15	0.32	0.50	0.34
ΔTPRC					
1	0.43	0.21	0.35	0.60	0.49
2	0.43	0.20	0.34	0.58	0.47
3	0.42	0.15	0.32	0.49	0.33
4	0.41	0.12	0.29	0.46	0.27
5	0.40	0.10	0.28	0.45	0.22
패널 C: 긍정어 빈도%(평균)					
전체	0.58	0.85	0.68	0.50	0.66
ΔTPRC					
1	0.57	0.79	0.65	0.40	0.51
2	0.57	0.80	0.66	0.42	0.53
3	0.58	0.85	0.68	0.51	0.67
4	0.59	0.88	0.71	0.54	0.73
5	0.60	0.90	0.72	0.55	0.78

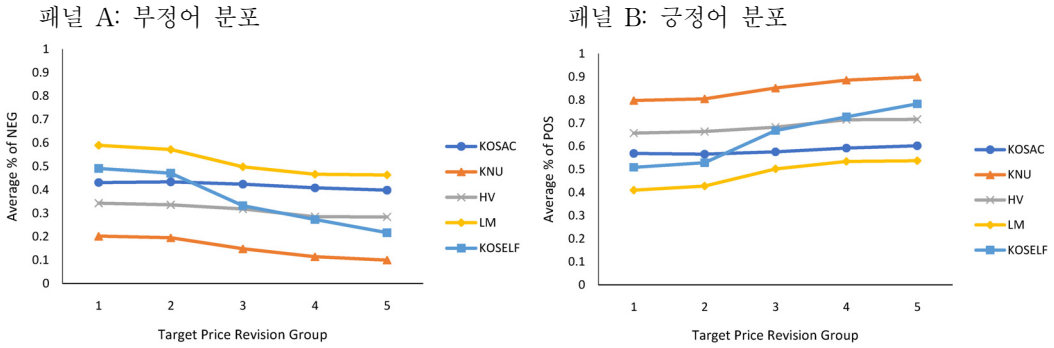
OPN의 평균값이 음(-)의 값을 갖는 것이 다른 사전에 비해 우수한 점이라 생각된다. KOSAC, KNU, HV는 KOSELF에 비해 선명한 변화 패턴을 보이지 못한다.

패널 B의 사전별 부정어 비율의 평균값을 살펴보면 목표주가가 그룹 1에서 그룹 5로 커질수록 부정어 빈도가 모든 사전에서 감소하는 경향을 보인다. 특히 KOSELF에서 이러한 경향이 뚜렷하게 나타나며, LM, KNU 사전 또한 유사한 패턴을 보인다. 다만 KOSAC과 HV 사전에서는 상대적으로 약한 감소추세를 가지는 것으로 나타났다. 패널 B의 긍정어 비율의 평균값에서도 동일한 결론을 얻을 수 있었다.

이는 <그림 1>에서 제시한 그래프에서도 확인할 수 있다. 부정어와 긍정어 모두에서 KOSELF가 다른 사전에 비해 목표주가 변경에 따라 가장 선명한 변화를 보여주고 있다. 부정어에 있어서는 LM의 수치가 KOSELF보다 높으며, 긍정어에 있어서는 KNU의 수치가 KOSELF보다 높다. 그 외에 있어서는 KOSELF가 수치 자체에 크기에 있어서도 우수함을 보여준다.

<그림 1> 목표주가 변경에 따른 감성변수의 분포

본 그림은 목표주가 변경에 따른 사전별 긍·부정어 감성변수의 분포를 나타낸다. 각각 KOSAC(서울대 일반사전), KNU(군산대 일반사전), HV(하버드 영어사전), LM(재무 특화 영어사전)의 감성변수 평균값의 분포를 나타내고 있으며, 본 연구에서 제안한 한국어 재무 특화 사전인 KOSELF 또한 포함하고 있다. 감성변수는 전처리를 거친 애널리스트 보고서 본문에 각 사전을 구성하는 긍·부정어의 출현빈도수 비율이다. 목표주가의 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 이전 보고서의 목표주가를 차감한 비율이다. 목표주가 변화율을 5개 범주 그룹으로 나누어 서술하였으며 1, 2그룹은 목표주가 변화율이 음(-)인 경우를, 3그룹은 목표주가 변화가 없는 경우를, 4, 5그룹은 목표주가 변화율이 양(+)인 경우를 나타낸다.



<표 5>는 목표주가 변화율을 종속변수로 사용한 회귀분석을 결과를 보여주고 있다. 분석 결과를 살펴보면 모형 (1)부터 (5)까지 각 감성사전에 기반한 OPN이 양(+)의 유의미한 값을 가지고 있다. 회귀계수의 크기를 비교하면, KOSELF(β = 0.198, t = 24.633), 그리고 LM(β = 0.177, t = 22.848), KNU(β = 0.140, t = 17.831), HV(β = 0.123, t = 15.487), KOSAC(β = 0.088, t = 10.649) 순으로 모두 목표주가 변화율에 유의한 양(+)의 관계를 가지고 있다. 특히 모형 (6)의 회귀분석 변수선택 결과를 살펴보면 KOSELF 사전에 기반한 의견변수의 회귀계수(β = 0.134, t = 15.319)가 가장 커서 다른 변수에 비해 상대적으로 높은 영향력을 보여 주고 있다.

〈표 5〉 회귀분석 결과: 목표주가 변경과 오피니언(OPN) 변수의 관계

본 표는 애널리스트 보고서의 목표주가 변화율(ΔTPRC)을 종속변수로 한 회귀분석 결과를 나타낸다. 애널리스트 보고서의 목표주가 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 직전 보고서의 목표주가를 차감한 값을 이전 목표주가로 나눈 비율이다. 독립변수로서 각 감성사전별 POS%에서 NEG%를 차감한 OPN 변수를 사용하였다. POS%(NEG%)는 애널리스트 보고서 텍스트의 각 사전별 긍정어(부정어) 수를 긍정어와 부정어 수 총합으로 나눈 비율이다. 따라서 OPN < 0인 경우 보고서의 부정어 비율이 긍정어 비율보다 높으며, OPN > 0인 경우 반대로 긍정어 비율이 부정어 비율보다 높은 경우에 해당한다. OPN = 0인 경우 부정어와 긍정어 비율이 동일하여 중립적 보고서임을 의미한다. 통제변수로서 로그를 취한 종목 거래량(52주 평균), 로그를 취한 기업의 규모(52주 평균 시가총액), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수를 사용하였다. 또한 표준산업분류 중분류코드와 연도 더미변수를 통해 산업효과와 연도효과를 통제하였다. 모형 (1)부터 모형 (5)는 각 감성사전 변수를 설명변수로 하는 회귀분석 결과를 나타낸다. 모형 (6)은 감성사전 변수에 대한 독립변수의 전진선택 결과를 나타낸다. 표는 각 변수에 대한 회귀계수와 함께 t-값을 괄호 안에 서술하고 있다. 표에서 ***, **, *은 각각 1%, 5%, 10% 유의수준 하에서 통계적으로 유의함을 의미한다.

변수	(1)	(2)	(3)	(4)	(5)	(6)
OPN_KOSAC	0.086***					0.028***
	(10.649)					(3.366)
OPN_KNU		0.140***				0.062***
		(17.831)				(7.140)
OPN_HV			0.123***			0.006
			(15.487)			(0.649)
OPN_LM				0.177***		0.106***
				(22.848)		(11.886)
OPN_KOSELF					0.196***	0.134***
					(24.633)	(15.319)
Nreport	-0.135***	-0.144***	-0.118***	-0.122***	-0.136***	-0.149***
	(-5.982)	(-6.398)	(-5.248)	(-5.459)	(-5.976)	(-6.613)
Nanalyst	0.219***	10.228***	0.227***	0.222***	0.222***	0.223***
	(9.691)	(10.130)	(10.075)	(9.933)	(9.787)	(9.925)
BM	-0.089***	-0.085***	-0.071***	-0.083***	-0.083***	-0.079***
	(-10.444)	(-10.007)	(-8.301)	(-9.860)	(-9.678)	(-9.202)
LN_거래량	0.046***	0.039***	0.053***	0.049***	0.053***	0.051***
	(4.977)	(4.200)	(5.763)	(5.376)	(5.661)	(5.543)
LN_시가총액	-0.063***	-0.057***	-0.081***	-0.075***	-0.058***	-0.044***
	(-4.713)	(-4.325)	(-6.157)	(-5.746)	(-4.287)	(-3.299)
YoY Growth	0.006	0.005	0.006	0.004	0.005	0.004
	(0.728)	(0.647)	(0.725)	(0.548)	(0.578)	(0.468)
beta	0.058***	0.049***	0.052***	0.051***	0.052***	0.044***
	(7.133)	(5.998)	(6.428)	(6.381)	(6.339)	(5.401)
leverage	0.033***	0.031***	0.019**	0.017*	0.014	0.025***
	(3.540)	(3.360)	(2.071)	(1.892)	(1.528)	(2.657)
Adj. R^2	0.030	0.043	0.038	0.054	0.062	0.080
N	15,921	15,844	15,912	15,908	15,060	15,025
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

4.4 추천의견 변경을 사용한 성과 검증

본문의 분석에서 감성사전을 검증하기 위한 벤치마크로 목표주가 변경을 사용하였다. 여기서는 대안으로 추천의견 변경을 사용하여 감성사전의 성과를 비교하였다. 먼저 추천의견 변경에 따른 부정어와 긍정어의 빈도의 패턴을 간략하게 살펴보고자 한다. <표 6>은 추천의견 변경에 따른 사전별 감성변수의 분포를 보여주고 있다. 패널 A의 사전별 OPN의 평균값을 살펴보면 추천의견이 상향될수록 OPN도 증가하는 사전은 HV와 KOSELF 2개 뿐이다. 특히 KOSELF

〈표 6〉 추천의견 변경에 따른 감성변수의 분포

본 표는 애널리스트 보고서의 추천의견 변경 단계에 따른 각 사전별 감성변수의 분포를 보여주고 있다. 애널리스트 보고서는 추천의견에 따라 매도, 비중축소, 중립, 매수, 적극매수의 5개 범위로 구분하였으며, 각 범주에 다음의 값을 임의로 부여하였다. 매도 = 1, 비중축소 = 2, 중립 = 3, 매수 = 4, 적극매수 = 5. 추천의견 변경(ΔRECOMM)은 특정 기업의 애널리스트 보고서 추천의견 값에서 이전 보고서의 추천의견 값을 차감한 수치이다. 패널 A에서 사용한 감성변수는 보고서에 대해 각 사전별로 부정어와 긍정어 출현빈도를 총 부정어 및 긍정어 출현빈도로 나누어 POS%에서 NEG%를 계산한 후, 차이를 계산한 OPN 변수이다. 패널B는 부정어 출현비율인 NEG%를, 패널C는 긍정어 출현비율인 POS%의 추천의견에 따른 사전별 평균값을 보고하였다.

	KOSAC	KNU	HV	LM	KOSELF
패널 A: OPN(평균)					
전체	0.16	0.69	0.36	0.00	0.32
ΔRECOMM					
-2	0.33	0.88	0.26	0.17	-1.00
-1	0.12	0.55	0.30	-0.19	-0.04
0	0.16	0.69	0.36	0.00	0.32
1	0.16	0.69	0.36	-0.01	0.45
2	0.19	0.47	0.37	0.11	0.92
패널 B: 부정어 빈도%(평균)					
전체	0.42	0.15	0.32	0.50	0.34
ΔRECOMM					
-2	0.33	0.06	0.37	0.41	1.00
-1	0.44	0.23	0.35	0.59	0.52
0	0.42	0.15	0.32	0.50	0.34
1	0.42	0.15	0.32	0.50	0.27
2	0.41	0.26	0.31	0.44	0.04
패널 C: 긍정어 빈도%(평균)					
전체	0.58	0.85	0.68	0.50	0.66
ΔRECOMM					
-2	0.67	0.94	0.63	0.59	0.00
-1	0.56	0.77	0.65	0.41	0.48
0	0.58	0.85	0.68	0.50	0.66
1	0.58	0.85	0.68	0.50	0.73
2	0.59	0.74	0.69	0.56	0.96

에서 -2로 두 단계 하락할 때 -1.00, -1로 한 단계 하락할 때 -0.04로 둘 다 음(-)의 값을 가지고 있는 유일한 사전이다. 이런 점들은 KOSELF가 다른 사전에 비해 우수한 점이라 생각된다.

패널 B의 부정어 빈도를 살펴보면 KOSELF를 제외한 기존 4개 사전에서 추천의견 변경이 없는 경우와 1단계 상향의 경우는 부정어 빈도가 동일하여 문맥의 차이가 거의 존재하지 않는다고 볼 수 있다. 또한 앞선 기초통계 분석에서 2단계 하향을 한 보고서 샘플 수가 1개임을 고려하면, KOSAC, HV, LM, KOSELF 사전은 추천의견 단계를 상향할수록 부정어 빈도가 선형적으로 감소하는 경향을 보였다. 동일한 맥락에서 패널 B에 서술한 긍정어 빈도의 경우 추천의견 단계를 상향할수록 긍정어 빈도가 선형 증가하는 경향을 보였다. 다만 부정어와 긍정어 빈도의 감소 또는 증가 추세가 본 연구진이 제시한 KOSELF 사전에서 가장 가파른 추이변화를 보였다.

이러한 추세는 <그림 2> 추천의견 변경에 따른 상자도표(box plot)을 통해서도 확인할 수 있다. 변수의 정의 상 NEG%와 POS%는 상하반전시킨 그래프이기 때문에 NEG% 하나의 상자도표만을 보고하였다. 2단계 하향 및 상향은 각각의 케이스 수가 상대적으로 적음을 감안하여 1단계 하향 및 상향 그룹에 포함하여 나타내었다. 따라서 -1은 추천의견이 하향한 모든 경우를, +1은 추천의견이 상향한 모든 경우를 포함한다. 상자도표를 보면, KOSAC, KNU, HV의 경우 +1에서 -1로 가면서 부정어비율(NEG%)이 증가하는 경향이 명확하지 않다. 반면 LM과 KOSELF는 -1로 가면서 NEG%이 증가하는 현상이 뚜렷하게 나타난다.

지금부터는 여러 통제변수들을 사용한 로지스틱 회귀분석을 통해 각 감성사전 변수와 추천의견 변경과의 관계에 대해 살펴보고자 한다. 종속변수는 애널리스트 보고서의 추천의견 변경 변수이다. 설명변수는 감성사전 5개를 사용하여 계산한 감성 변수들이다. 다음은 구체적인 회귀분석식을 보여주고 있다.

$$DOWN_i = \alpha_0 + \alpha_1 OPN_i + \alpha_2 Control\ Variables_i + \epsilon_i \quad (5)$$

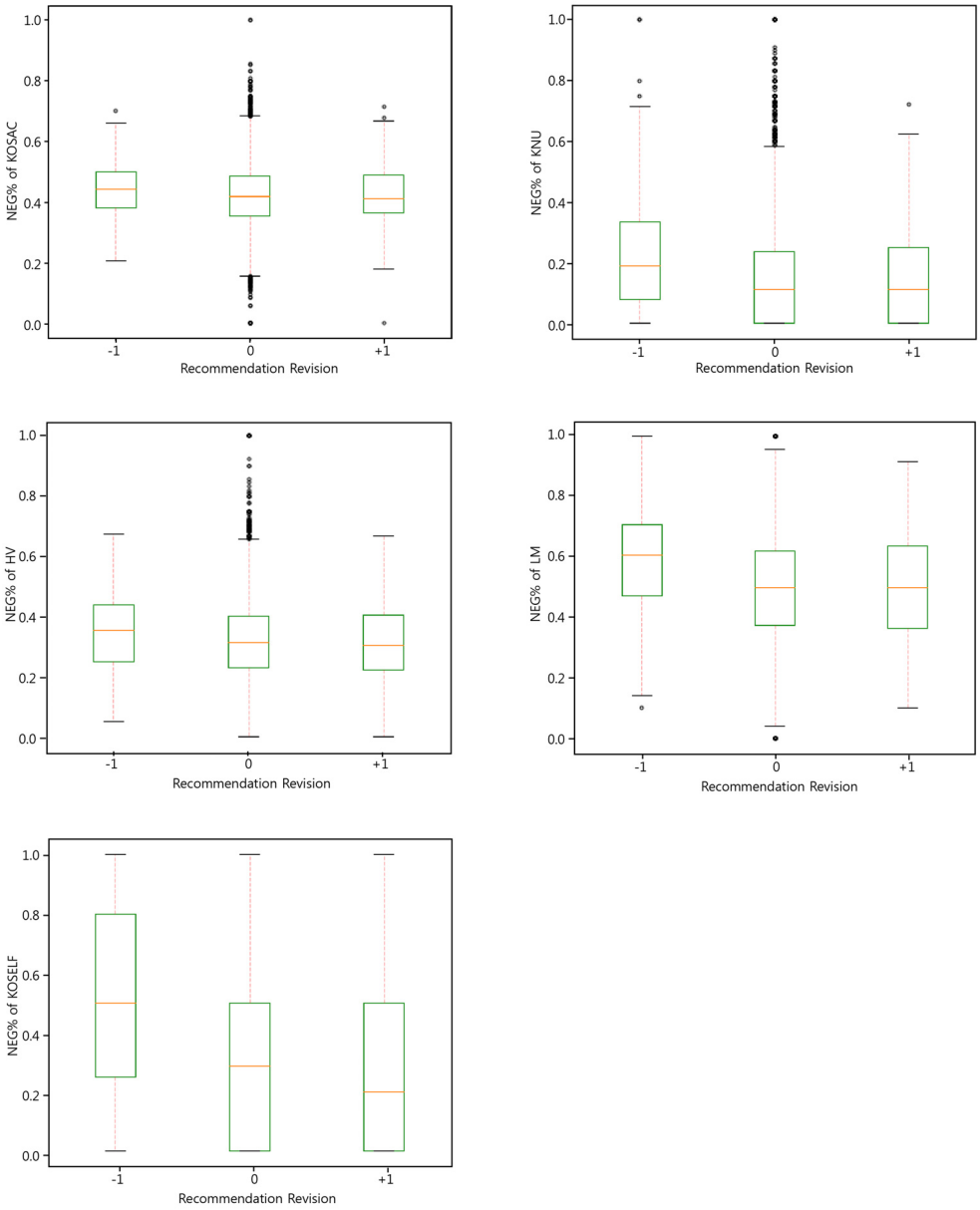
$$UP_i = \alpha_0 + \alpha_1 OPN_i + \alpha_2 Contraol\ Variables_i + \epsilon_i \quad (6)$$

여기서 DOWN은 추천의견이 하향(downgrade)이면 1, 아니면 0이 부여되는 더미변수이다. UP은 추천의견이 상향(upgrade)이면 1, 아니면 0이 부여되는 더미변수이다. OPN은 POS%에서 NEG%를 차감한 의견변수이다. NEG%(POS%)는 보고서 텍스트의 부정어(긍정어) 수를 부정어와 긍정어의 합으로 나눈 비율이다. 통제변수는 앞의 본문과 동일하다.

<표 7>은 로지스틱 회귀분석 결과를 보여주고 있다. 패널 A는 부정어, 패널 B는 긍정어 검증 결과를 나타내고 있으며, 모형 (1)부터 (5)까지는 5개 사전기반 OPN 변수에 대한 로지스틱 회귀분석 결과를 나타내고 있다. 마지막 모형 (6)은 5개 사전기반 OPN 변수 모두를 동시에 회귀분석한 결과이다. 각 모형의 분석결과를 살펴보면 다음과 같다. 패널 A에서 5개 사전기반 OPN 변수는 추천의견 하향보고서에서 모두 통계적으로 유의한 영향을 미치는 것으로 나타났다. 특히 LM은 β = -1.466(p = 0.000)으로 하향인 보고서에서 가장 많이 출현하는 사전임이 나타났다. 모형 (6)에서는 한국어와 영문 일반사전인 KOSAC, HV사전은 탈락하였으며, 역시 재무 특화 영문사전인 LM과 본 연구진의 KOSELF가 상대적으로 높은 영향력을 보였다.

〈그림 2〉 추천의견 변경에 따른 감성변수의 분포

본 그림은 애널리스트 보고서의 추천의견 변경 단계에 따른 사전별 감성변수의 분포를 나타낸다. 각각 KOSAC(서울대 일반사전), KNU(군산대 일반사전), HV(하버드 영어사전), LM(재무 특화 영어사전)의 감성변수 평균값의 분포를 나타내고 있으며, 본 연구에서 제안한 한국어 재무 특화 사전인 KOSELF 또한 포함하고 있다. 감성변수는 전처리를 거친 애널리스트 보고서 본문에 각 사전을 구성하는 긍·부정어의 출현빈도수 비율이다. 추천의견은 매도(1단계)부터 적극매수(5단계)까지 점수를 부여하고 특정 기업의 애널리스트 보고서 추천의견 값에서 이전 보고서의 추천의견 값을 차감한 수치이다.



〈표 7〉 로지스틱 회귀분석 결과: 추천의견 변경과 OPN 변수의 관계

본 표는 애널리스트 보고서의 추천의견 변경 여부를 종속변수로 한 OPN 변수의 성과를 검증한 로지스틱 회귀분석 결과를 나타낸다. 패널 A의 부정어 검증의 경우 종속변수는 추천의견이 하향이면 1, 아니면 0의 값을 가지는 더미변수이다. 패널 B의 긍정어 검증의 경우 종속변수는 추천의견이 상향이면 1, 아니면 0의 값을 가지는 더미변수이다. 독립변수로서 각 감성사전별 POS%에서 NEG%를 차감한 OPN 변수를 사용하였다. POS%(NEG%)는 애널리스트 보고서 텍스트의 각 사전별 긍정어(부정어) 수를 긍정어와 부정어 수 총합으로 나눈 비율이다. 따라서 OPN < 0인 경우 보고서의 부정어 비율이 긍정어 비율보다 높으며, OPN > 0인 경우 반대로 긍정어 비율이 부정어 비율보다 높은 경우에 해당한다. OPN = 0인 경우 부정어와 긍정어 비율이 동일하여 중립적 보고서임을 의미한다. 통제변수로서 로그를 취한 종목 거래량(52주 평균), 로그를 취한 기업의 규모(52주 평균 시가총액), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수를 사용하였다. 또한 표준산업분류 중분류코드와 연도 더미변수를 통해 산업효과와 연도효과를 통제하였다. 모형 (1)부터 모형 (5)는 각 OPN 변수를 설명변수로 하는 로지스틱 회귀분석 결과를 나타낸다. 모형 (6)은 감성사전 변수에 대한 독립변수의 전진선택 결과를 나타낸다. 표는 각 변수에 대한 회귀계수와 함께 p-값을 괄호 안에 서술하고 있다. 표에서 ***, **, *은 각각 1%, 5%, 10% 유의수준 하에서 통계적으로 유의함을 의미한다.

패널 A: 추천의견 하향인 경우 종속변수 = 1

변수	(1)	(2)	(3)	(4)	(5)	(6)
OPN_KOSAC	-1.034***					0.307
	(0.002)					(0.579)
OPN_KNU		-1.071***				-0.452**
		(0.000)				(0.012)
OPN_HV			-0.985***			2.444
			(0.000)			(0.118)
OPN_LM				-1.466***		-0.970***
				(0.000)		(0.000)
OPN_KOSELF					-0.814***	-0.572***
					(0.000)	(0.000)
Nreport	-0.008***	-0.008***	-0.008***	-0.008***	-0.008***	-0.008***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.002)
Nanalyst	0.070	0.069	0.073	0.065	0.061	0.055
	(0.131)	(0.136)	(0.114)	(0.158)	(0.192)	(0.237)
BM	0.205	0.239*	0.146	0.164	0.187	0.160
	(0.136)	(0.083)	(0.298)	(0.237)	(0.184)	(0.258)
LN_거래량	0.050	0.029	0.035	0.040	0.024	0.028
	(0.537)	(0.722)	(0.662)	(0.622)	(0.768)	(0.740)
LN_시가총액	0.165	0.146	0.177*	0.146	0.158	0.124
	(0.100)	(0.146)	(0.077)	(0.150)	(0.125)	(0.234)
YoY Growth	-0.001	-0.001	-0.001	-0.001	0.000	0.000
	(0.813)	(0.821)	(0.822)	(0.836)	(0.855)	(0.861)
beta	-0.125	-0.120	-0.122	-0.096	-0.108	-0.089
	(0.406)	(0.432)	(0.419)	(0.525)	(0.489)	(0.572)
leverage	-0.380	-0.429	-0.322	-0.355	-0.071	-0.158
	(0.487)	(0.435)	(0.552)	(0.507)	(0.900)	(0.778)
Adj. R^2	0.068	0.081	0.070	0.092	0.094	0.109
N	16,066	15,988	16,457	16,053	15,194	15,158
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

〈표 7〉 로지스틱 회귀분석 결과: 추천의견 변경과 OPN 변수의 관계(계속)

패널 B: 추천의견 상향인 경우 종속변수 = 1

변수	(1)	(2)	(3)	(4)	(5)	(6)
OPN_KOSAC	0.085					0.205
	(0.812)					(0.650)
OPN_KNU		−0.098				0.243
		(0.631)				(0.622)
OPN_HV			−0.221			1.352
			(0.407)			(0.245)
OPN_LM				−0.194		−0.500**
				(0.291)		(0.015)
OPN_KOSELF					0.337***	0.443***
					(0.004)	(0.000)
Nreport	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***	−0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Nanalyst	0.098**	0.099**	0.100**	0.099**	0.101**	0.099**
	(0.043)	(0.041)	(0.039)	(0.041)	(0.046)	(0.049)
BM	−0.067	−0.061	−0.075	−0.067	−0.006	−0.018
	(0.656)	(0.686)	(0.620)	(0.656)	(0.969)	(0.907)
LN_거래량	0.040	0.036	0.033	0.034	0.002	0.005
	(0.641)	(0.679)	(0.704)	(0.695)	(0.979)	(0.958)
LN_시가총액	0.062	0.059	0.060	0.056	0.091	0.079
	(0.570)	(0.592)	(0.585)	(0.608)	(0.423)	(0.491)
YoY Growth	0.000	0.000	0.000	0.000	0.000	0.000
	(0.871)	(0.874)	(0.876)	(0.875)	(0.838)	(0.830)
beta	0.462***	0.462***	0.463***	0.466***	0.589***	0.595***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.000)	(0.000)
leverage	0.061	0.037	0.032	0.025	0.057	−0.015
	(0.919)	(0.951)	(0.957)	(0.966)	(0.928)	(0.981)
Adj. R^2	0.064	0.063	0.064	0.064	0.072	0.075
N	16,066	15,988	16,057	16,053	15,194	15,158
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

패널 B의 추천의견 상향인 보고서를 1로 설정한 로지스틱 회귀분석 결과, KOSELF를 제외한 4개 사전 기반 감성변수가 모두 통계적으로 유의하지 않았다. KOSELF의 경우만 $\beta = 0.337$($p = 0.004$)로 추천의견 상향과 유미한 관계를 보여주었다. 이런 결과는 KOSELF의 긍정어 목록이 다른 사전에 비해 우수한 성과를 내고 있음을 시사한다.

위의 결과들을 통해 본 연구에서 구축한 KOSELF가 한국 기업분석을 위한 재무사전으로써 유용함을 확인할 수 있었다. 기존의 KOSAC, KNU와 같은 한국어 감성사전이 있지만 일반사전으로써 기업 재무분석에 활용하기에는 한계를 지니고 있다. 이런 점들을 개선하였기 때문에 좋은 성과를 얻은 것이라 사료된다.

4.5 강건성 검증

지금까지의 분석에 대해 몇 가지의 강건성 검증(robustness check)을 실시하였다. 먼저

〈표 8〉 강건성 검증: 목표주가 변경과 부정어 감성변수(NEG%)의 관계

본 표는 애널리스트 보고서의 목표주가 변화율(ΔTPRC)을 종속변수로 한 회귀분석 결과를 나타낸다. 애널리스트 보고서의 목표주가 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 직전 보고서의 목표주가를 차감한 값을 이전 목표주가로 나눈 비율이다. 독립변수로서 NEG%는 애널리스트 보고서 텍스트의 부정어 수를 부정어 수와 긍정어 수 총합으로 나눈 비율이다. 통제변수로서 로그를 취한 종목 거래량(52주 평균), 로그를 취한 기업의 규모(52주 평균 시가총액), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수를 사용하였다. 또한 표준산업분류 중분류코드와 연도 더미변수를 통해 산업효과와 연도효과를 통제하였다. 모형 (1)부터 모형 (5)는 NEG% 변수를 설명변수로 하는 회귀분석 결과를 나타낸다. 모형 (6)은 부정어 감성사전 변수(NEG%)에 대한 독립변수의 전진선택 결과를 나타낸다. 표는 각 변수에 대한 회귀계수와 함께 t-값을 괄호 안에 서술하고 있다. 표에서 ***, **, *은 각각 1%, 5%, 10% 유의수준 하에서 통계적으로 유의함을 의미한다.

변수	(1)	(2)	(3)	(4)	(5)	(6)
NEG%_KOSAC	-0.086***					-0.028***
	(-10.649)					(-3.366)
NEG%_KNU		-0.140***				-0.062***
		(-17.831)				(-7.140)
NEG%_HV			-0.123***			-0.006
			(-15.487)			(-0.649)
NEG%_LM				-0.177***		-0.106***
				(-22.848)		(-11.886)
NEG%_KOSELF					-0.196***	-0.134***
					(-24.633)	(-15.319)
Nreport	-0.135***	-0.144***	-0.118***	-0.122***	-0.136***	-0.149***
	(-5.982)	(-6.398)	(-5.248)	(-5.459)	(-5.976)	(-6.613)
Nanalyst	0.219***	0.228***	0.227***	0.222***	0.222***	0.223***
	(9.691)	(10.130)	(10.075)	(9.933)	(9.787)	(9.925)
BM	-0.089***	-0.085***	-0.071***	-0.083***	-0.083***	-0.079***
	(-10.444)	(-10.007)	(-8.301)	(-9.860)	(-9.678)	(-9.202)
LN_거래량	0.046***	0.039***	0.053***	0.049***	0.053***	0.051***
	(4.977)	(4.200)	(5.763)	(5.376)	(5.661)	(5.543)
LN_시가총액	-0.063***	-0.057	-0.081***	-0.075***	-0.058***	-0.044***
	(-4.713)	(-4.325)	(-6.157)	(-5.746)	(-4.287)	(-3.299)
YoY Growth	0.006	0.005	0.006	0.004	0.005	0.004
	(0.728)	(0.647)	(0.725)	(0.548)	(0.578)	(0.468)
beta	0.058***	0.049***	0.052***	0.051***	0.052***	0.044***
	(7.133)	(5.998)	(6.428)	(6.381)	(6.339)	(5.401)
leverage	0.033***	0.031***	0.019**	0.017*	0.014	0.025***
	(3.540)	(3.360)	(2.071)	(1.892)	(1.528)	(2.657)
Adj. R^2	0.030	0.043	0.038	0.054	0.062	0.080
N	15,921	15,844	15,912	15,908	15,060	15,025
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

본문에서 사용한 OPN 변수에 대한 대안 측정치를 고려하여 보았다. 첫 번째는 OPN을 구성하는 POS%와 NEG%를 분리하여 각각의 설명력에 대해 검증하였다.

<표 8>은 목표주가 변화율에 대해 부정어 비율(NEG%) 변수를 회귀분석하였다. 긍정어 비율(POS%)에 대한 결과는 생략하였다. 변수의 정의 상 POS% = 1 – NEG%이기 때문에 동일한 결과를 갖기 때문이다. 5개 사전 모두의 부정어 비율(NEG%)이 목표주가 변화과 유의미한 음(–)의 관계를 가지는 것으로 나타났다. 모형 (6)에서 5개의 사전을 모두 사용하였을 때, KOSELF가 β = –0.134(t = –15.319)로 가장 큰 회귀계수를 가지고 있었다. 한국어 일반 감성 사전인 KOSAC과 KNU 또한 목표주가 변화에 유의하게 나타났으나 상대적으로 영향력이 작았다. 별도의 표로 서술하지는 않았으나 긍정어 검증 결과 또한 동일하게 나타났다.

<표 9>는 추천의견 변경에 대해 부정어 비율(NEG%)과 긍정어 비율(%POS) 변수를 회귀분석하였다. 다음과 같은 두 개의 식을 검증하였다.

$$DOWN_i = \alpha_0 + \alpha_1 NEG\%_i + \alpha_2 Control\ Variables_i + \epsilon_i \quad (7)$$

$$UP_i = \alpha_0 + \alpha_1 POS\%_i + \alpha_2 Controal\ Variables_i + \epsilon_i \quad (8)$$

여기서 DOWN은 추천의견이 하향(downgrade)이면 1, 아니면 0이 부여되는 더미변수이다. UP은 추천의견이 상향(upgrade)이면 1, 아니면 0이 부여되는 더미변수이다. NEG%(POS%)는 보고서 텍스트의 부정어(긍정어) 수를 부정어와 긍정어의 합으로 나눈 비율이다. 통제변수는 앞의 본문과 동일하다.

패널 A에서 5개 사전기반 NEG% 변수는 추천의견 하향보고서에서 모두 통계적으로 유의한 영향을 미치는 것으로 나타났다. 특히 LM은 β = 2.932(p = 0.000)으로 하향이 아닌 보고서에 비해 하향인 보고서에서 가장 많이 출현하는 사전임이 나타났다. 모형 (6)의 변수선택 과정을 거친 결과 한국어와 영문 일반사전인 KOSAC, HV사전은 탈락하였으며, 역시 재무 특화 영문사전인 LM과 본 연구진의 KOSELF가 상대적으로 높은 영향력을 보였다.

반면 긍정어에 대한 로지스틱 회귀분석 결과, KOSELF만 유의미한 양(+)의 회귀계수를 보여주었다. 모형 (6)에서 모든 사전 변수를 동시에 사용한 경우에도 KOSELF만이 β = 0.886(p = 0.000)으로 긍정어 비율(POS%)이 추천의견 상향과 유의한 관계를 보여주었다. KOSELF의 회귀계수 크기에 있어서는 긍정어 비율(POS%)보다는 부정어 비율(NEG%)에서 더욱 큰 관계를 가지는 것을 볼 수 있다.

설명변수에 대한 두 번째 강건성 검증은 보고서별 부정어 또는 긍정어 출현빈도를 해당 보고서에 나타난 전체 단어수로 나누어 총 단어수 대비 부정어 또는 긍정어의 비율을 감성변수로 사용하였다. 구체적인 정의는 아래와 같다.

$$NEG\%_{all} = \frac{\text{부정어 수}}{\text{전체 단어 수}} \quad (9)$$

$$POS\%_{all} = \frac{\text{긍정어 수}}{\text{전체 단어 수}} \quad (10)$$

<표 9> 강건성 검증: 추천의견 변경과 감성변수(NEG%, POS%)의 관계

본 표는 애널리스트 보고서의 추천의견 변경 여부를 종속변수로 한 감성사전 변수들의 성과를 검증한 로지스틱 회귀분석 결과를 나타낸다. 패널 A의 부정어 검증의 경우 종속변수는 추천의견이 하향이면 1, 아니면 0의 값을 가지는 더미변수이다. 패널 B의 긍정어 검증의 경우 종속변수는 추천의견이 상향이면 1, 아니면 0의 값을 가지는 더미변수이다. 독립변수로서 NEG%(POS%)는 애널리스트 보고서 텍스트의 부정어(긍정어) 수를 부정어와 긍정어 수의 총합으로 나눈 비율이다. 통제변수로서 로그를 취한 종목 거래량(52주 평균), 로그를 취한 기업의 규모(52주 평균 시가총액), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수를 사용하였다. 또한 표준산업분류 중분류코드와 연도 더미변수를 통해 산업효과와 연도효과를 통제하였다. 모형 (1)부터 모형 (5)는 각 감성사전 변수를 설명변수로 하는 로지스틱 회귀분석 결과를 나타낸다. 모형 (6)은 감성사전 변수에 대한 독립변수의 전진선택 결과를 나타낸다. 표는 각 변수에 대한 회귀계수와 함께 p-값을 괄호 안에 서술하고 있다. 표에서 ***, **, *은 각각 1%, 5%, 10% 유의수준 하에서 통계적으로 유의함을 의미한다.

패널 A: 부정어(NEG%)를 사용한 로지스틱 회귀분석 결과

변수	(1)	(2)	(3)	(4)	(5)	(6)
NEG%_KOSAC	2.067***					0.307
	(0.002)					(0.579)
NEG%_KNU		2.143***				0.904**
		(0.000)				(0.012)
NEG%_HV			1.971***			2.444
			(0.000)			(0.118)
NEG%_LM				2.932***		1.941***
				(0.000)		(0.000)
NEG%_KOSELF					1.635***	1.144***
					(0.000)	(0.000)
Nreport	-0.008***	-0.008***	-0.009***	-0.008***	-0.008***	-0.008***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.002)
Nanalyst	0.070	0.069	0.073	0.065	0.061	0.055
	(0.131)	(0.136)	(0.114)	(0.158)	(0.192)	(0.237)
BM	0.205	0.239*	0.146	0.164	0.187	0.160
	(0.136)	(0.083)	(0.298)	(0.237)	(0.184)	(0.258)
LN_거래량	0.050	0.029	0.035	0.040	0.024	0.028
	(0.537)	(0.722)	(0.662)	(0.622)	(0.768)	(0.740)
LN_시가총액	0.165*	0.146	0.177*	0.146	0.158	0.124
	(0.100)	(0.146)	(0.077)	(0.150)	(0.125)	(0.234)
YoY Growth	-0.001	-0.001	-0.001	-0.001	0.000	0.000
	(0.813)	(0.821)	(0.822)	(0.836)	(0.855)	(0.861)
beta	-0.125	-0.120	-0.122	-0.096	-0.108	-0.089
	(0.406)	(0.432)	(0.419)	(0.525)	(0.489)	(0.572)
leverage	-0.380	-0.429	-0.322	-0.355	-0.071	-0.158
	(0.487)	(0.435)	(0.552)	(0.507)	(0.900)	(0.778)
Adj. R^2	0.068	0.081	0.070	0.092	0.094	0.109
N	16,066	15,988	16,057	16,053	15,194	15,158
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

<표 9> 강건성 검증: 추천의견 변경과 감성변수(NEG%, POS%)의 관계(계속)
패널 B: 긍정어(POS%)를 사용한 로지스틱 회귀분석 결과

변수	(1)	(2)	(3)	(4)	(5)	(6)
POS%_KOSAC	0.170					0.205
	(0.812)					(0.650)
POS%_KNU		-0.196				0.243
		(0.631)				(0.622)
POS%_HV			-0.442			1.352
			(0.407)			(0.245)
POS%_LM				-0.389		-1.001**
				(0.291)		(0.015)
POS%_KOSELF					0.673***	0.886***
					(0.004)	(0.000)
Nreport	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Nanalyst	0.098**	0.099**	0.100**	0.099**	0.101**	0.099**
	(0.043)	(0.041)	(0.039)	(0.041)	(0.046)	(0.049)
BM	-0.067	-0.061	-0.075	-0.067	-0.006	-0.018
	(0.656)	(0.686)	(0.620)	(0.656)	(0.969)	(0.907)
LN_거래량	0.040	0.036	0.033	0.034	0.002	0.005
	(0.641)	(0.679)	(0.704)	(0.695)	(0.979)	(0.958)
LN_시가총액	0.062	0.059	0.060	0.056	0.091	0.079
	(0.570)	(0.592)	(0.585)	(0.608)	(0.423)	(0.491)
YoY Growth	0.000	0.000	0.000	0.000	0.000	0.000
	(0.871)	(0.874)	(0.876)	(0.875)	(0.838)	(0.830)
beta	0.462***	0.462***	0.463***	0.466***	0.589***	0.595***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.000)	(0.000)
leverage	0.061	0.037	0.032	0.025	0.057	-0.015
	(0.919)	(0.951)	(0.957)	(0.966)	(0.928)	(0.981)
Adj. R^2	0.064	0.063	0.064	0.064	0.072	0.075
N	16,066	15,988	16,057	16,053	15,194	15,158
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

<표 10>에서는 위의 새로운 감성변수를 활용하여 로지스틱 회귀분석을 시행하였다. 패널 A의 부정어 검증 결과 5개 사전에 기반한 전체 단어수 대비 부정어 출현비율이 모두 추천의견 하향변경에 통계적으로 유의한 영향을 미쳤다. 그러나 모형 (6)의 변수선택 과정을 거친 결과 KOSELF에 대해서만 유의수준 1% 내에서 통계적 유의성이 유지되는 것으로 나타났다. 반대로 패널 B의 긍정어 검증 결과 5개 사전 추천의견 상향조정 여부에 모두 유의하지 않은 것을 확인할 수 있는데, 이러한 결과는 <표 7>의 결과와 유사하게 해석할 수 있다. 전체 단어수 대비 긍정어 비율이 추천의견을 하향조정하는 경우 의미있게 낮지만, 추천의견 변경이 없는 경우와 상향 조정하는 경우에서 서로 유사하기 때문에 종속변수를 상향 조정여부로 설정한 모형에서는 통계적으로 유의하지 않은 것이라 사료된다.

〈표 10〉 강건성 검증: 추천의견 변경과 전체 단어수 대비 감성변수의 관계

본 표는 애널리스트 보고서의 추천의견 변경 여부를 종속변수로 한 감성사전 변수들의 성과를 검증한 로지스틱 회귀분석 결과를 나타낸다. 패널 A의 부정어 검증의 경우 종속변수는 추천의견이 하향이면 1, 아니면 0의 값을 가지는 더미변수이다. 패널 B의 긍정어 검증의 경우 종속변수는 추천의견이 상향이면 1, 아니면 0의 값을 가지는 더미변수이다. 독립변수로서 NEG%all(POS%all)는 애널리스트 보고서 텍스트의 부정어(긍정어) 수를 전체 단어 수로 나눈 비율이다. 통제변수로서 로그를 취한 종목 거래량(52주 평균), 로그를 취한 기업의 규모(52주 평균 시가총액), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수를 사용하였다. 또한 표준산업분류 중분류코드와 연도 더미변수를 통해 산업효과와 연도효과를 통제하였다. 모형 (1)부터 모형 (5)는 각 감성사전 변수를 설명변수로 하는 로지스틱 회귀분석 결과를 나타낸다. 모형 (6)은 감성사전 변수에 대한 독립변수의 전진선택 결과를 나타낸다. 표는 각 변수에 대한 회귀계수와 함께 p-값을 괄호 안에 서술하고 있다. 표에서 ***, **, *은 각각 1%, 5%, 10% 유의수준 하에서 통계적으로 유의함을 의미한다.

패널 A: 부정어(NEG%all)를 사용한 로지스틱 회귀분석 결과

변수	(1)	(2)	(3)	(4)	(5)	(6)
NEG%all_KOSAC	3.143					0.332
	(0.108)					(0.564)
NEG%all_KNU		34.959***				0.125
		(0.000)				(0.723)
NEG%all_HV			14.624***			3.284*
			(0.000)			(0.070)
NEG%all_LM				22.644***		0.156
				(0.000)		(0.692)
NEG%all_KOSELF					28.374***	28.374***
					(0.000)	(0.000)
Nreport	-0.008***	-0.008***	-0.009***	-0.008***	-0.008***	-0.008***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Nanalyst	0.067	0.071	0.075	0.064	0.067	0.067
	(0.148)	(0.129)	(0.103)	(0.168)	(0.154)	(0.154)
BM	0.223	0.210	0.133	0.120	0.049	0.049
	(0.107)	(0.130)	(0.344)	(0.392)	(0.729)	(0.729)
LN_거래량	0.051	0.049	0.041	0.035	0.026	0.026
	(0.529)	(0.546)	(0.614)	(0.675)	(0.754)	(0.754)
LN_시가총액	0.177*	0.160	0.170*	0.142	0.133	0.133
	(0.077)	(0.112)	(0.091)	(0.166)	(0.199)	(0.199)
YoY Growth	-0.001	-0.001	0.000	-0.001	0.000	0.000
	(0.821)	(0.827)	(0.852)	(0.837)	(0.905)	(0.905)
beta	-0.120	-0.112	-0.121	-0.085	-0.072	-0.072
	(0.427)	(0.461)	(0.424)	(0.580)	(0.643)	(0.643)
leverage	-0.326	-0.394	-0.387	-0.603	-0.643	-0.643
	(0.553)	(0.474)	(0.477)	(0.268)	(0.243)	(0.243)
Adj. R^2	0.065	0.076	0.075	0.100	0.116	0.116
N	16,066	16,066	16,066	16,066	16,066	16,066
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

〈표 10〉 강건성 검증: 추천의견 변경과 전체 단어수 대비 감성변수의 관계(계속)
패널 B: 긍정어(POS%all)을 사용한 로지스틱 회귀분석 결과

변수	(1)	(2)	(3)	(4)	(5)	(6)
POS%all_KOSAC	-0.169					0.009
	(0.926)					(0.926)
POS%all_KNU		-3.052				0.971
		(0.324)				(0.324)
POS%all_HV			-1.407			0.367
			(0.544)			(0.545)
POS%all_LM				-2.636		0.634
				(0.426)		(0.426)
POS%all_KOSELF					-1.692	0.324
					(0.569)	(0.569)
Nreport	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Nanalyst	0.098**	0.098**	0.098**	0.098**	0.098**	0.098**
	(0.043)	(0.042)	(0.042)	(0.042)	(0.042)	(0.043)
BM	-0.069	-0.068	-0.073	-0.066	-0.066	-0.068
	(0.647)	(0.652)	(0.628)	(0.659)	(0.660)	(0.651)
LN_거래량	0.039	0.036	0.038	0.040	0.039	0.039
	(0.647)	(0.671)	(0.657)	(0.643)	(0.653)	(0.646)
LN_시가총액	0.060	0.061	0.061	0.059	0.060	0.061
	(0.579)	(0.572)	(0.577)	(0.586)	(0.579)	(0.575)
YoY Growth	0.000	0.000	0.000	0.000	0.000	0.000
	(0.870)	(0.881)	(0.867)	(0.881)	(0.879)	(0.871)
beta	0.462***	0.466***	0.464***	0.466***	0.465***	0.462***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
leverage	0.057	0.091	0.073	0.072	0.064	0.059
	(0.924)	(0.879)	(0.902)	(0.904)	(0.914)	(0.921)
Adj. R^2	0.064	0.064	0.064	0.064	0.064	0.064
N	16,066	16,066	16,066	16,066	16,066	16,066
Industry Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes

4.6 제안

본 연구에서는 기존 감성사전이 가지는 한계점을 보완하고자 재무분야에 특화된 한국어 감성사전 KOSELF를 구축하였으며, 실제 애널리스트 보고서의 텍스트를 통해 사전의 성능을 검증함으로써 애널리스트 보고서의 정량지표가 아닌 문맥에서 사용되는 단어의 정성적 의미 즉, 보고서가 내포하고 있는 긍정 또는 부정적 감성을 측정하고자 하였다. 본 논문의 결과를 종합하여 보면, KOSELF 감성사전이 KOSAC, KNU, HV와 같은 일반용 감성사전보다는 확실히 우수하다고 할 수 있지만, 동일한 재무특화 사전인 LM을 뛰어 넘는 성과를 보여주고 있다고 하기는 어렵다. 긍정어와 부정어로 구분하여 보면, KOSELF는 긍정어 부분에서, LM은 부정어 부분에서 상대방보다 더 우수한 성과를 보여주고 있다.

〈표 11〉 추가분석: KOSELF-LM 사전의 성과 검증

본 표는 본 연구진이 제시하는 재무 특화 한글사전인 KOSELF와 재무 특화 영문사전인 LM사전을 결합하여 나타낸 KOSELF-LM 감성변수의 성과를 검증한 결과를 나타낸다. 패널 A는 애널리스트 보고서의 추천의견 변경 여부를 종속변수로 하는 로지스틱 회귀분석 결과를 나타낸다. 부정어 검증에서는 추천의견이 하향이면 1, 아니면 0의 값을 가지고 긍정어 검증에서는 추천의견이 상향이면 1, 아니면 0의 값을 가지는 더미변수이다. 패널 A는 독립변수로서 NEG%(POS%)를 가지며, 이는 애널리스트 보고서 텍스트의 KOSELF-LM 부정어(긍정어) 수를 부정어와 긍정어 수 총합으로 나눈 비율이다. 통제변수로서 로그를 취한 종목 거래량(52주 평균), 로그를 취한 기업의 규모(52주 평균 시가총액), BM(시가총액 대비 자기자본의 연말 장부가치), 레버리지(총부채 대비 총자산), 연평균 이익증가율, 해당 종목에 대해 해당 연도에 보고서를 발행한 애널리스트 수 및 해당 연도의 총 보고서 수를 사용하였다. 또한 표준산업분류 중분류코드와 연도 더미변수를 통해 산업효과와 연도효과를 통제하였다. 모형 (1), (3)은 KOSELF-LM 사전의 부정어(긍정어) 변수를 설명변수로 하는 로지스틱 회귀분석 결과를 나타낸다. 모형 (2), (4)는 전체 감성사전 변수에 대한 독립변수의 전진선택 결과를 나타낸다. 패널 A는 각 변수에 대해 회귀계수와 함께 괄호 안의 p-값을 서술하고 있다. 패널 B는 애널리스트 보고서의 목표주가 변화율(ΔTPRC)을 종속변수로 한 회귀분석 결과를 나타낸다. 목표주가 변화율(ΔTPRC)은 특정 기업의 애널리스트 보고서 목표주가에서 직전 보고서의 목표주가를 차감한 값을 이전 목표주가로 나눈 비율이다. 패널 B의 독립변수, 통제변수는 패널 A와 동일하다. 패널 B는 각 변수에 대한 회귀계수와 함께 t-값을 괄호 안에 서술하고 있다. 표에서 ***, **, *은 각각 1%, 5%, 10% 유의수준 하에서 통계적으로 유의함을 의미한다.

패널 A: 추천의견 변경에 대한 로지스틱 회귀분석 결과

변수	추천의견 하향		추천의견 상향	
	(1)	(2)	(3)	(4)
NEG%(POS%)_KOSAC		0.054		0.002
		(0.815)		(0.968)
NEG%(POS%)_KNU		0.490		0.227
		(0.484)		(0.633)
NEG%(POS%)_HV		-1.864***		0.789
		(0.002)		(0.374)
NEG%(POS%)_LM		-1.482**		1.315
		(0.032)		(0.252)
NEG%(POS%)_KOSELF-LM	4.040***	6.000***	-0.149	0.201
	(0.000)	(0.000)	(0.678)	(0.654)
Nreport	-0.008***	-0.007***	-0.009***	-0.009***
	(0.001)	(0.002)	(0.001)	(0.001)
Nanalyst	0.062	0.061	0.099**	0.098**
	(0.183)	(0.188)	(0.040)	(0.042)
BM	0.098	0.159	-0.065	-0.060
	(0.480)	(0.252)	(0.664)	(0.690)
LN_거래량	0.003	-0.009	0.033	0.037
	(0.967)	(0.913)	(0.700)	(0.669)
LN_시가총액	0.136	0.131	0.059	0.061
	(0.181)	(0.199)	(0.590)	(0.576)
YoY Growth	0.000	0.000	0.000	0.000
	(0.917)	(0.937)	(0.877)	(0.874)
beta	-0.098	-0.113	0.462***	0.458***
	(0.526)	(0.470)	(0.004)	(0.004)
leverage	-0.444	-0.546	0.031	0.037
	(0.410)	(0.317)	(0.958)	(0.951)
Adj. R^2	0.121	0.129	0.063	0.063
N	16,048	15,977	16,048	15,977
Industry Dummy	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes

〈표 11〉 추가분석: KOSELF-LM 사전의 성과 검증(계속)

패널 B: 목표주가 변경에 대한 회귀분석 검증 결과

변수	(1)	(2)
NEG%_KOSAC		-0.020(-2.415)**
NEG%_KNU		-0.030(-3.319)***
NEG%_HV		0.015(1.562)
NEG%_LM		-0.009(-0.744)
NEG%_KOSELF-LM	-0.227(-29.370)***	-0.209(-22.758)***
Nreport	-0.140(-6.321)***	-0.146(-6.583)***
Nanalyst	0.227(10.303)***	0.227(10.258)***
BM	-0.067(-8.041)***	-0.068(-8.076)***
LN_거래량	0.054(6.027)***	0.053(5.852)***
LN_시가총액	-0.066(-5.126)***	-0.059(-4.505)***
YoY Growth	0.003(0.368)	0.003(0.375)
beta	0.051(6.344)***	0.048(6.038)***
leverage	-0.034(-4.308)***	0.036(3.924)
Adj. R^2	0.074	0.076
N	15,903	15,833
Industry Dummy	Yes	Yes
Year Dummy	Yes	Yes

이런 점을 고려하여 실무적 사용을 위해 KOSELF의 확장안으로서 재무 특화 영문사전인 LM과의 결합을 통해 KOSELF-LM 사전을 구축하였으며, 이에 따른 추가적인 성과 검증을 진행하였다. LM 사전에서 제공하고 있는 1,074개 부정어와 476개 긍정어 리스트 중 한글로 번역하였을 때 그 의미가 적절하지 않다고 판단되는 일부 단어를 제외하고 총 909개 부정어와 395개 긍정어를 기존 KOSELF 사전의 구성 단어와 결합하였다. 예를 들면, LM 사전의 부정어 중 'misprice(가격을 잘못 매기다)', 'confess(고백)', 'challenge(도전)', 'concede(시인)' 등을 제외하였다. 긍정어에서는 'tremendous(가공할)', 'upturn(전복)', 'pleasure(쾌락)' 등의 단어를 제외하였다.

<표 11>은 목표주가 변화율에 따른 회귀식과 추천의견 변경에 따른 회귀식에서 감성변수를 KOSELF-LM을 통해 측정한 NEG%(POS%) 비율로 변경하여 분석한 결과에 대해 서술하고 있다. 먼저 패널 A의 로지스틱 회귀분석 결과를 살펴보면 부정어 검증결과에서 KOSELF-LM에 기반한 감성변수가 $\beta = 4.040(p = 0.000)$로 추천의견 하향여부에 영향을 미치고 있는 것을 확인할 수 있다. 그러나 긍정어 검증결과에서는 의미 있는 결과를 보이지 않았다.

패널 B의 목표주가 변화율에 대한 회귀분석 결과도 마찬가지로 KOSELF-LM의 부정어 감성변수가 $\beta = -0.227(t = -29.370)$로 유의하게 나타났으며, 긍정어 또한 같은 결과를 얻을 수 있었다. 따라서 본 연구에서 제안하는 한국어 재무 특화 사전인 KOSELF가 기업 재무분석에서 추천의견의 변경이나, 목표주가의 변경과 같은 정량지표와 마찬가지로 애널리스트가 본문에서 사용하는 여러 단어들의 감성을 효과적으로 측정할 수 있는 것으로 나타났다.

5. 결론

금융시장에서 애널리스트는 특정 종목에 대한 공시정보를 해석함과 동시에 비공시정보를 발굴하는 역할을 수행하며, 이러한 정보를 애널리스트 보고서에 나타냄으로써 시장가격이 해당 정보를 반영할 수 있도록 한다. 이 과정에서 추천의견이나 목표주가와 같은 정량지표는 가장 직접적으로 애널리스트가 가진 정보를 반영할 수 있는 지표로서 작용한다. 그럼에도 불구하고 애널리스트는 금융시장에서 자신의 명성을 유지하기 위해, 또는 소속 증권사 및 다수의 이해관계자와의 밀접한 관계에 따라 추천의견 및 목표주가의 조정을 매우 보수적으로 제시하는 경향이 있으며, 이러한 사실은 특히 하향 조정의 경우 두드러지는 것이 다수의 선행연구에서 확인된 바 있다. 이 경우 애널리스트는 본문의 구성과, 사용하는 다양한 단어를 통해 자신이 가지는 의견을 보다 간접적으로 반영할 수 있으며, 이는 일종의 정성적 지표로서 해석되어질 수 있다. 그러나 기존의 일반 감성사전은 '수익', '증가', '기존' 등의 재무 분야의 중립 단어를 긍정어로 포함하는 등 기업 재무분석에 사용하기에 적합하지 않은 한계점이 존재한다. 또는 '경상(가벼운 부상)', '환자' 등의 단어가 부정어를 구성하는 등 분야의 특성을 반영하지 않는다는 단점 또한 존재한다. 기존 재무 분야에 특화된 Loughran and McDonald(2011)의 사전이 존재하나, 이는 영문 사전이기 때문에 한글로 번역하는 과정에서 마찬가지로 원문의 의미를 제대로 반영하지 못하는 경우가 다수 나타난다.

이러한 문제점을 보완하기 위하여 본 연구에서는 한글 재무 특화사전인 KOSELF를 구성하였다. KOSELF의 긍정어 및 부정어를 통해 애널리스트 보고서에 나타난 감성을 변수화하여 측정하였으며, 이를 추천의견 변경단계와 목표주가 변화율을 통해 검증하였다. 결과적으로 KOSELF는 기존의 사전들로 측정한 감성변수보다 효과적으로 애널리스트 보고서 본문에 나타난 긍·부정적 감성을 추출하는 것으로 나타났다. 특히 목표주가 변경보다 추천의견을 변경한 경우 본문에 드러난 감성을 더욱 잘 추출하는 것을 확인하였는데, 일반적으로 투자자들이 추천의견 변경에 더욱 민감하게 반응함을 고려하면 KOSELF의 성능이 더욱 효과적으로 작용하는 것으로 판단된다. 이러한 결과는 감성변수를 다양하게 변화시켰을 때에도 동일한 양상을 보였다.

그럼에도 불구하고 본 연구는 다음과 같은 한계점을 가진다. 첫째로, 애널리스트 보고서와 감성사전의 긍·부정어를 비교하는 과정에서 원문을 그대로 비교하는 것이 아닌 형태소 분석을 통한 토큰화(tokenizing) 단계를 거쳤으나, 이 경우 형태소 분석기의 성능에 따라 기존의 단어 의미와 별개의 감성이 추출되기도 한다. 예를 들어, 기업 '두산 모트롤'의 경우 '모/MM(관형사) + 트롤/NNP(고유명사)'로 분해되거나, 방한용 의류소재인 '패트롤'의 경우 '패/NNG(일반명사) + 트롤/NNP(고유명사)'로 분해된다. 이처럼 경우에 따라, 특히 고유명사나 외래어의 경우 원문장을 분해하는 과정에서 본래의 의미와는 다른 별개의 형태소 집합으로 분해될 수 있으며, 이 과정에서 실제 단어와는 관련 없는 긍정 또는 부정의 감성을 잘못 추출할 우려가 있다. 따라서 향후 형태소 분석 시 특정 단어들을 사용자 사전(user dictionary)의 형태로 추가 구성함으로써 토큰화 과정의 오류를 줄이고 보다 고도화된 형태소 분석 및 단어의 감성 추출을 진행할 수 있을 것이다.

둘째로, 본 연구에서는 애널리스트 보고서와 감성사전을 구성하는 텍스트에 대해 한 개의 토큰 즉, 1-gram에 기반한 분석을 전제로 하였다. 그러나 '단축(curtailment)', '감소(decline)', '부담(burden)', '문제' 등의 1-gram 차원에서 부정적 감성을 지니는 단어들이 '시간 단축', '위험/리스크 감소', '부담 없이 즐길 수 있는', '문제 해결' 등 n-gram 차원으로 확장할 시 긍정적 감성을 지니는 것으로 해석될 수 있다. 따라서 향후 재무분야에 특화된 n-gram 차원의 감성사전을 구축함으로써 이 같은 한계점을 보완할 수 있다. 마지막으로, 본 연구에서 구축한 KOSELF 감성사전은 현재 긍·부정어에 대해 동일한 감성점수를 부여하고 있다. 그러나 이에 대해 보다 세분화된 감성 점수를 부여함으로써 텍스트가 가지는 감성을 보다 효과적으로 측정할 수 있을 것으로 기대된다.

한국어를 위한 금융 감성사전(KOSELF)의 현재 버전이 완벽하다고 할 수는 없다. 앞으로도 애널리스트 보고서 외의 다양한 표본들을 추가하여 사전을 더 보완해 나갈 계획이다. 또한 구축된 사전을 공개하여 여러 사람들이 한국어 감성 분석에 활용할 수 있도록 도움을 주는 것을 목표로 하고 있다.

References

An, J., and H. W. Kim, 2015, Building a Korean Sentiment Lexicon Using Collective Intelligence, *Journal of Intelligence and Information Systems*, Vol. 21 (2), pp. 49-67.

Antweiler, W., and M. Z. Frank, 2004, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *The Journal of Finance*, Vol. 59 (3), pp. 1259-1294.

Baccianella, S., A. Esuli, and F. Sebastiani, 2010, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *Proceedings of the International Conference on Language Resources and Evaluation*, Vol. 10 (2010), pp. 2200-2204.

Buehlmaier, M. M. M., and T. M. Whited, 2018, Are Financial Constraints Priced? Evidence from Textual Analysis, *The Review of Financial Studies*, Vol. 31 (7), pp. 2693-2728.

Chen, H., P. De, Y. J. Hu, and B. H. Hwang, 2014, Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media, *The Review of Financial Studies*, Vol. 27 (5), pp. 1367-1403.

Garca, D., 2013, Sentiment during Recessions, *The Journal of Finance*, Vol. 68 (3), pp. 1267-1300.

Kim, D. S., and S. S. Eum, 2006, The Impact of Analysts' Revisions in Their Stock Recommendation and Target Prices on Stock Prices, *Asia-Pacific Journal of Financial Studies*, Vol. 35 (2), pp. 75-108.

Kim, S. S., 2010, Analyst Recommendation Change and Fund Performance in Korea Fund Stock Market, *Korean Journal of Business Administration*, Vol. 23 (3), pp. 1351-1370.

Kim, Y. H., H. G. Kang, and J. K. Lee, 2018, Can Big Data Forecast North Korean Military Aggression?, *Defence and Peace Economics*, Vol. 29 (6), pp. 666-683.

Kim, Y., and S. W. Joh, 2019, Text Analysis for IPO firms in Korea: Analysis of Korean Texts in Registration Statements via Machine Learning, *Korean Journal of Financial Studies*, Vol. 48 (2), pp. 215-235.

Lee, E., and C. G. Park, 2019, Does Adoption of K-IFRS Increase Upward Bias in Analysts' Earnings Forecasts?, *The Korean Journal of Financial Management*, Vol. 36 (1), pp. 179-205.

Lee, J. S., 2011, Three-Step Probabilistic Model for Korean Morphological Analysis, *Journal of KIISE: Software and Applications*, Vol. 38 (5), pp. 257-268.

Lee, W. H., and S. M. Choi, 2003, The Effect of Changes in Analysts' Investment Recommendation Ranking on Stock Returns and Trading Volumes, *Journal of Korean Securities Association*, Vol. 32, pp.1-44.

Li, F., 2010, The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research*, Vol. 48 (5), pp. 1049-1102.

Li, F., R. Lundholm, and M. Minnis, 2013, A Measure of Competition Based on 10-K Filings, *Journal of Accounting Research*, Vol. 51 (2), pp. 399-436.

Loughran, T., and B. McDonald, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance*, Vol. 66 (1), pp. 35-65.

Park, E. J., and S. Cho, 2014, KoNLPy: Korean Natural Language Processing in Python, *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology*, pp. 133-136.

Park, S. M., C. W. Na, M. S. Choi, D. H. Lee, and B. W. On, 2018, KNU Korean Sentiment Lexicon: Bi-LSTM-based Method for Building a Korean Sentiment Lexicon, *Journal of Intelligence and Information Systems*, Vol. 24 (4), pp. 219-240.

Shin, D. H., D. H. Cho, and J. S. Nam, 2016a, Building the Korean Sentiment Lexicon DecoSelex for Sentiment Analysis, *Journal of Korealex*, Vol. 28, pp. 75-111.

Shin, H. P., M. H. Kim, and S. Z. Park, 2016b, Modality-based Sentiment Analysis through the Utilization of the Korean Sentiment Analysis Corpus, *Eoneohag*, Vol. 74, pp. 93-114.

Shin, H., M. Kim, Y. M. Jo, H. Jang and C. Andrew 2012, Annotation Sceheme for Constructing Sentiment Corpus in Korean, In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pp. 181-190.

Tetlock, P. C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance*, Vol. 62 (3), pp. 1139-1168.

Tetlock, P. C., M. S. Tsechansky, and S. Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals, *The Journal of Finance*, Vol. 63 (3), pp. 1437-1467.