한국어 텍스트 분석과 적용: 머신러닝을 통한 증권발행신고서의 비정형화된 텍스트 분석*

김 용 석 (서울대학교 석사과정)

조 성 옥 (서울대학교 교수)**

< 요 약 >

이 논문은 2009년 6월부터 2017년 12월까지 한국증권 시장에 신규 상장된 기업이 제출한 증권 발행신고서의 텍스트를 머신러닝 기법을 통해 긍정적 또는 부정적으로 분류하는 작업을 수행하였다. 분류된 어조에 대한 분석 결과 증권발행신고서 중 투자위험요소에 있어 부정적인 어조가 많은 회사는 최초 공모희망가액에 대비 최종 공모가격 비율이 높게 설정하는 경향이 나타났다. 이런 결과는 최초 공모희망가액 밴드가 낮게 설정되었음을 시사한다. 한편, 증권발행신고서의 어조와 상장 직후 수익률 간에는 통계적으로 유의한 관계가 나타나지 않는다. 이 논문은 한국 금융시장에서 기업이 제출한 비정형화된 텍스트에 나타난 어조를 최초로 분석했다는 점에서 의미가 있다. 또한, 이 논문은 향후 한국어 텍스트 분석을 시도하는 학자들에게 도움이 되고자 비정형화된 텍스트 분석에 사용한 프로그램을 제시한 점에서 의미가 있다.

핵심 단어 : 텍스트 분석, 한국어 어조, 머신러닝, 증권발행신고서, IPO 기업

JEL 분류기호: G14, G23, G24

* 저자들은 논문에 대해 상세하고 유익한 논평을 해 주신 두 명의 심사자들에게 깊이 감사드립니다. 서울대학교 경영연구소의 지원에 감사드립니다.

** 연락담당 저자. 주소: 서울특별시 관악구 관악로 1 서울대학교 경영대학, 08826; E-mail: swjoh@snu.ac.kr, Sungwook.joh@gmail.com; Tel: 02-880-9384; Fax: 02-873-4821.

투고일 2018-10-31; 수정일 2019-01-25; 게재확정일 2019-03-12

1. 서론

최근의 연구자들은 소위 빅데이터의 시대라고 할 만큼 과거에는 상상도 할 수 없을 정도의 많은 양의 데이터에 직면하게 되었다. 정보통신 기술 및 컴퓨터 성능의 발전으로 과거에는 사용되지 못했던 다양한 정보에 대한 분석이 가능해지고 있다. 특히 고차원 모형 분석의 이론적, 실증적 발전에 힘입은 텍스트 분석은 신문이나 정치인의 정치적 편향성(Gentzkow and Shapiro, 2010; Gentzkow et al., 2018), 각 지역의 인종차별주의의 강도(Stephens-Davidowitz, 2014), 경제 정책의 불확실성의 정도(Baker et al., 2016), 증권신고서의 어조(Loughran and McDonald, 2013)처럼 기존에 측정하기 어려운 변수들을 측정하는 데에 활용되는 등 경제학 및 경영학 연구자들의 높은 관심을 받고 있다.¹⁾

그러나, 외국에서의 활발한 연구와는 대조적으로 한국 금융시장의 텍스트 분석은 아직까지 초기단계라고 할 수 있다. 경제 및 경영 분야에서 사용되는 한국어 단어에 대한 평가 및 이에 기초한 단어목록 등의 부족으로 경제, 금융, 회계와 기타 경영분야에서 비정형화된 한국어로 된 자료에 대한 분석을 시도한 논문은 많지 않다. 특히, 말과 글에 담겨있는 감정을 측정하는 어조(Tone) 분석은 금융시장 텍스트 분석에 있어 아직까지 수행된 바 없다.

이 논문은 한국 금융시장에서 기업공개(IPO)를 하는 기업이 제출한 증권발행신고서에 나타난 비정형화된 텍스트에 나타난 어조를 측정하고자 한다. 또한 부정적 또는 긍정적으로 측정된 어조를 이용하여 증권발행신고서 텍스트에 나타난 어조가 IPO 성과변수와 어떤 관계를 갖는지를 분석하고자 한다. 이를 통해 IPO 기업의 어조를 분석한 외국 선행연구의 결과가 국내 금융시장에서 한국어로 작성된 텍스트에서도 관찰되는지를 확인하고자 한다. 또한, 어조를 분석할 때 사용한 프로그램을 학계와 공유함으로써 이 분야에 대한 향후 연구에 공헌하고자 한다.

이 논문은 Li(2010)의 방법론을 따라, 머신러닝 방법론 중 하나인 나이브 베이즈 분류법을 이용해 각 증권발행신고서의 어조를 측정하였다. 보다 구체적으로 설명하자면, 본 연구는 2009년 6월부터 2017년 12월까지 신규상장된 기업이 제출한 증권발행신고서의 투자위험 요소절에서 사용된 문장을 대상으로 어조를 분석하였다. 데이터에서 식별한 약 9만 개의 문장 중 1만 개의 문장을 임의로 추출한 뒤, 저자 및 코더들이 직접 읽고 개별 문장에 대해 해당 문장의 어조를 긍정, 중립, 부정 또는 불확실 등 세 가지 중 하나로 분류해 학습 데이터를 구성하였다. 이 데이터로 학습시킨 모형을 이용해 모든 문장의 어조를 분류했다. 각 증권발행신고서의 어조는 구성하는 문장들의 평균적인 어조로 측정하였다. 측정 결과, 증권발행신고서의 어조는 회사간 이질성이 충분히 큰 것으로 나타났다.

증권발행신고서의 어조와 신규상장된 기업 성과변수 간의 관계를 분석한 결과는 다음과 같다. 첫째, 최초 희망공모가격 밴드 대비 최종 공모가격 비율 즉 공모가격 변화율은 증권

1) 보다 자세한 리뷰는 Gentzkow et al.(2017)를 참고하라.

발행신고서의 어조가 부정적일수록 이 변화율이 높게 나타났다. 증권발행신고서의 어조가 부정적일 때 최종 공모가격이 작은 경향이 있으므로 이는 회사가 최초 희망공모가격 밴드 수준을 낮게 설정하는 경향이 있음을 시사한다. 또한, 이러한 결과는 외국의 선행 연구와도 일치한다(Loughran and McDonald, 2013). 둘째, 외국의 선행연구 결과와는 달리, 상장 직후 수익률과 증권발행신고서의 어조는 통계적으로 유의한 관계가 없었다. 이러한 결과는 수요 조사(bookbuilding) 이후 한국 IPO 시장이 외국, 특히 미국 IPO 시장과 제도적으로 다르게 설정되었기 때문에 나타나는 결과일 수 있다.

이 논문은 다음과 같이 구성된다. 제 2장에서는 선행연구를 검토하고, 본 연구의 가설을 제시한다. 제 3장에서는 연구 방법론을 묘사한다. 제 4장에서는 연구 결과를 제시한다. 제 5장에서는 본 논문을 요약하고, 시사점을 제시한다.

2. 선행연구 검토 및 가설

컴퓨터 성능의 개선과 분석 기술의 향상에 따라, 경제학 및 경영학에서 대규모 텍스트를 분석하기 시작한 연구는 최근 들어 급격히 증가하고 있다. Gentzkow and Shapiro(2010)는 각 정당별 의원들의 의사록을 이용, 미국 내 신문들이 얼마나 정치적으로 편향되어 있는지 측정하였다. Stephens-Davidowitz(2014)은 구글 검색 기록을 통해 미국 내 지역별 인종 차별의식 정도를 측정한 뒤, 이것이 최초의 흑인 대통령인 버락 오바마의 득표율에 어떤 영향을 끼쳤는지 탐색하였다. Baker et al.(2016)은 신문 기사를 통해 측정한 경제 정책의 불확실성 지표를 제시했다. Gentzkow et al.(2018)은 민주당과 공화당 의원들의 말을 서로 비교하면서, 시간이 흐름에 따라 민주당 소속 의원과 공화당 소속 의원이 사용하는 단어나 어구가 이전에는 크게 다르지 않았는데, 점점 달라지고 있다고 보고한다. 그들은 이것이 점차 심화되는 미국 정치에서의 양극화를 나타낸다고 주장한다.

이에 따라 회계학 및 재무학에서 대규모의 기업 및 회계 공시 자료에 대한 분석이 활발히 진행되고 있다. Li(2008)는 대규모 공시 자료를 텍스트 분석한 최초의 연구 중 하나다. 그는 10-K 보고서의 가독성(readability)을 측정하면서, 가독성이 낮은 실적 보고서를 제출한 기업은 그 해 실적이 낮은 경향이 있으며, 이익의 지속성 또한 떨어진다는 것을 보고하였다. Li et al.(2013)은 10-K 보고서의 Management Discussion and Analysis(MD&A) 섹션을 통해 기업이 직면하고 있는 경쟁의 정도를 측정한 뒤, 산업 내 경쟁이 치열할 수록 회사의 한계 투자 수익률이 줄어드는 경향이 있음을 보였다. Buehlmaier and Whited(2018)은 텍스트 분석을 통해 각 기업의 자금조달 제약(financial constraint)을 측정한 뒤, 자금조달 제약이 심한 기업에 대한 주가 수익률이 높아지는 경향이 있으며 이는 위험에 대한 보상으로 판단된다고 보고하였다.

반면 외국의 연구와는 달리 한국 금융시장의 텍스트를 분석한 연구는 아직 초기단계에 머물러 있다. 텍스트 자체에 대한 분석은 아니지만, Park et al.(2015)은 자동화된 알고리즘을

통해 대규모의 공시 자료의 텍스트 분석 가능성을 타진한 최초의 연구다. 이들은 한국의 전자공시시스템이 대규모의 텍스트 분석을 수행하기에 부적합하다고 지적하면서, 그 원인 중 하나로 공시 문서가 PDF 형식으로 제공됨을 지목한 바 있다. PDF 형식의 문서는 기계 판독에 적합하지 않아서, 자동화된 알고리즘으로 분석하기 위해서는 문서를 HTML 또는 TXT 형식으로 추가 변환해야 한다는 것이다. 나아가 이러한 문제를 피하기 위해 웹 크롤링(Web Crawling) 방법을 이용할 수 있지만, 해당 방법은 컴퓨터 처리의 속도가 느리고 연구자에게 추가적인 프로그래밍 지식을 요구한다는 단점이 있다고 지적한다.

한편, Lee et al.(2018)은 음절 숫자와 문장 길이를 고려하는 Fog Index 등으로 연도별 사업보고서의 가독성을 측정하고, 가독성이 낮은 기업은 해당 연도의 이익수준이 낮고, 이익의 지속성이 떨어진다는 것을 확인하였다. Yook(2018)은 네이버 파파고를 통해 한국어를 영어로 번역한 지속가능보고서에 나타난 CEO 메시지를 분석하면서 지속가능성 성과가 낮을 수록 가독성은 떨어지고 긍정적 표현을 많이 사용하며 미래 성과를 더욱 강조하는 경향이 나타난다고 보고한 바 있다.²⁾

위에서 제시한 선구적인 연구들의 존재에도 불구하고, 현재까지 한국 금융시장 텍스트의 어조(Tone)를 체계적으로 분석한 논문은 거의 없다. 가독성과 더불어 텍스트의 어조는 기업의 이익 지속성이나 IPO 성과 등을 예측하는 등, 금융시장을 설명하는 데에 있어 정보를 가지고 있는 것으로 알려져 있다(Li, 2010; Loughran and McDonald, 2013). 어조를 분석한 연구는 크게 두 가지로 나뉘는데, 첫 번째는 부정어, 긍정어 등의 단어목록을 미리 구성한 뒤, 해당 단어가 전체 문서에 등장하는 빈도수를 이용해 측정하는 방식이다(Tetlock, 2007; Tetlock et al., 2008). 특히 Loughran and McDonald(2011)는 금융시장의 특수성을 고려, 텍스트의 어조를 측정할 수 있는 영문 단어목록(word list)을 제시한 바 있으며, 이 단어목록은 이후 재무금융 분야의 텍스트 분석과 관련된 다양한 연구에서 활용되고 있다(e.g., Chen et al., 2014; Garca, 2013).

한국 금융시장 텍스트의 어조 분석이 잘 진행되지 않았던 것은 영어권 국가와는 달리 한국어에 기초하여 설계되고 공인된 단어목록이 부재한 이유가 크다. 일반적으로 활용되는 단어목록을 활용할 수도 있겠지만, Loughran and McDonald(2011)에서 지적하듯이, 단어가 금융시장에서 가지는 의미가 일상생활에서의 의미와 크게 다르기 때문에, 금융시장의 특수성을 고려하지 않은 단어목록을 활용할 경우 상당한 편의가 발생할 수 있다. 예를 들어, 부채를 의미하는 liability의 경우 일상생활에서는 대체로 부정적인 의미를 가지지만, 재무제표에서는 계정과목 중의 하나로 중립적인 의미를 지닌다.

이처럼 단어목록이 부재할 때에 활용할 수 있는 방법이 Antweiler and Frank(2004), Li (2010), Frank and Sanati(2018) 등이 제시한 머신러닝 방법론이다. 제 3.2절에서 자세히

2) Yook(2018)은 지속가능보고서를 구글 번역기 또는 네이버 파파고를 이용해 영문으로 번역한 뒤 감성분석을 수행하였다. 그러나 이 경우의 감성분석은 금융시장의 특수성이 고려되지 않은 것으로, 아래에 후술한 바와 같이 편의를 가질 수 있다.

설명하겠지만, 이 방법은 단어목록이 없을 때에도 활용 가능하며, 단어목록을 직접 설계할 때에 비해 연구자의 자의성이 상대적으로 적게 개입된다. 또 교차 검증 시험 등을 통해, 측정된 어조의 강건성을 추가적으로 시험해 볼 수 있다는 장점도 있다.

이 연구에서는 Li(2010)의 머신러닝 방법론을 이용, 한국 IPO 시장에서 기업들이 제출한 증권신고서의 어조를 측정하고, 텍스트의 어조가 IPO 성과 변수들과 갖는 관계를 분석한다. 상장 이후 시장에서 거래가 이루어지며 형성된 주가가 IPO 공모가격을 초과하는 현상을 뜻하는 IPO 저가발행(underpricing)은 오랜 기간 재무 학자들의 관심을 끌어난 주제이다 (Ibbotson, 1975; Ritter, 1984). 선행 연구들은 IPO 회사와 투자자 사이의 정보비대칭을 IPO 저가발행의 주된 원인으로 지목해 왔다(Rock, 1986). 증권신고서의 텍스트 정보는 수요예측(Benveniste and Spindt, 1989; Benveniste and Wilhelm, 1990), 주관사의 명성 (Carter, 1990)과 더불어 정보비대칭을 완화하는 좋은 수단이 될 수 있다. 또한, 투자에 있어 중요한 정보를 누락할 경우 법적 책임을 지는 발행회사 및 주관사는 증권신고서의 텍스트를 통해 신규상장기업의 발생 가능한 위험 등을 미리 공시할 유인이 있다. 따라서, IPO 저가발행 현상을 연구하는 데에 있어 증권신고서 텍스트 정보를 활용하는 것은 중요하다.

Beatty and Ritter(1986)은 IPO 기업 현금흐름의 불확실성(uncertainty)을 IPO 저가발행의 원인으로 지목하면서, 불확실성이 클 경우 저가발행 폭이 커져 상장 수익률이 높아진다고 주장한 바 있다. Loughran and McDonald(2013)는 증권신고서의 어조가 이러한 회사의 불확실성을 측정한다고 주장하면서, 미국 IPO 시장에서 증권신고서의 어조가 부정적인 회사는 최초 공모 희망가격을 낮게 설정한 결과 공모가격 변화율(offer price revision) 및 상장 직후 수익률(initial return)이 높게 나타나는 경향이 있음을 보인 바 있다. 우리는 이러한 관계가 한국 IPO 시장에서도 관찰되리라고 예상한다.

가설 1: 증권신고서의 어조가 부정적인 회사일 수록 공모가격 변화율이 높게 나타날 것이다.

가설 2: 증권신고서의 어조가 부정적인 회사일수록 상장 직후 수익률이 높게 나타날 것이다.

3. 연구 방법론

3.1 표본

우리는 2009년 6월부터 2017년 12월 사이에 KOSPI 또는 KOSDAQ에 상장된 630개 회사의 증권신고서를 분석 대상으로 삼았다. 2009년 6월 이후로 분석 기간을 한정된 이유는 그 때를 기준으로 증권신고서의 보고 형식이 그 이전과 크게 달라졌기 때문이다. 630개의 신규상장 중, 우리는 98개의 합병목적회사(SPAC), 29개의 외국회사 및 공시가 불충분한 61개의 회사를 제외하였다. 따라서 이 논문의 최종 분석 대상은 442개의 신규상장 회사의 증권신고서이다.

우리는 전체 증권신고서 중 “투자위험요소”란에 한정해서 어조 분석을 수행했다. 투자 위험요소란은 기업이 향후 예상되는 위험요소에 대한 정보를 투자자들에게 제공하는 부분으로, 보고서의 어조에서 드러나는 기업의 정보를 측정하고자 하는 우리의 분석목적에 가장 부합한다고 판단했다.³⁾

각 증권신고서는 DART에서 제공하는 API(Application Programing Interface)를 통해 접근했다. Park et al.(2015)는 DART에서 공시 문서가 PDF 형식으로만 내려 받을 수 있음을 지적하면서, 컴퓨터를 통한 분석이 용이한 TXT 또는 HTML 형식에서의 추가적인 변환이 필요함을 지적한 바 있다.⁴⁾ 이 논문에서는 추가적인 변환 작업 및 이에 따른 텍스트 자료의 변형과 왜곡을 피하기 위해 R을 활용한 웹 크롤링(Web crawling) 방법을 이용해 각 증권신고서 내 투자위험요소란을 HTML 파일로 내려 받았다. 우리는 HTML 파일에서 태그 형식을 모두 제거한 뒤, 하나의 투자위험요소란을 개별문장으로 분절하기 위해 “다.”를 기준으로 문장을 나누었다. 식별된 모든 문장은 문장 기호, 특수 문자 및 숫자를 제거하였다. 위 분석을 위한 R 코드 예제는 부록 <그림 A1>과 같다. 데이터는 <https://github.com/ysugk/DARTcrawl>에서 내려 받을 수 있다.

3.2 어조의 측정

3.2.1 텍스트 분류 모형

기업이 제출한 증권신고서의 어조를 측정하기 위해 이 논문에서는 머신러닝 방법론 중 하나인 나이브 베이즈 분류(the naïve Bayes classification)를 적용하였다. 그 이유는 아래와 같다. 첫째, 나이브 베이즈 분류법은 회계 및 재무 분야의 텍스트 분석 문헌에서 활발히 사용되고 있으며(Buehlmaier and Whited, 2018), 보다 복잡한 방법론만큼이나 잘 작동하는 것으로 알려져 있다(Hastie et al., 2001). 둘째, Loughran and McDonald(2013)과 같이 미리 설계된 단어 목록을 활용한 분석 역시 가능하지만, 회계 및 재무 보고서의 문장 및 단어의 의미는 일반적인 맥락에서의 의미와 많이 다를 수 있기 때문에 이를 고려해 설계된 단어 목록 또는 모형이 아니면 적용하기 어렵다(Loughran and McDonald, 2011). 전술한 바와 같이 한국 금융시장의 텍스트 분석은 아직 초기단계이기에 적합한 단어목록이 없다. 텍스트 분류 모형은 ω 를 텍스트, $v \in V$ 를 미리 정해진 분류 기준이라 할 때,

$$\arg \max_{v \in V} P(v|\omega) \quad (1)$$

를 통해 얻어진 v^* 로 텍스트를 분류하는 방법이다. 이 때 각 텍스트 ω 는 이를 구성하는 J 개

3) 전체 문서 중 특정 절에 한정해 텍스트 분석을 수행한 예로는 Hoberg and Maksimovic(2014), Li et al.(2013), Loughran and McDonald(2011) 등이 있다.

4) PDF 형식의 공시 문서를 내려 받아 분석을 진행한 예로는 Lee et al.(2018)이 있다.

단어들의 집합 $\{\omega_1, \dots, \omega_j\}$ 으로 이해할 수 있다. 예를 들면 “동사의 부채 수준은 경쟁사에 비해 높습니다”라는 문장을 분류할 때 이 문장을 구성하는 단어의 집합 $\{\text{동사의, 부채, 수준은, 경쟁사에, 비해, 높습니다}\}$ 으로 변환할 수 있다는 것이다. 나이브 베이즈 분류 모형은 각 단어가 독립적으로 분포해 있다고 가정하면서, 식 (1)을 베이즈 룰(the Bayes rule)을 이용해

$$\arg \max_{v \in V'} P(v|\omega) = P(\omega|v) \cdot \frac{P(v)}{P(\omega)} = \prod_{j=1}^J P(\omega_j, v) \cdot P(v) \quad (2)$$

로 단순화한다. 따라서, 데이터를 나이브 베이즈 분류 모형을 학습시킨다는 것은 $P(\omega_j|v)$ 와 $P(v)$ 를 데이터에서의 빈도 수를 통해 추정한다는 것을 의미하며, 이 모형을 통해 문장을 분류한다는 것은 이렇게 추정한 $\hat{P}(\omega_j|v)$ 와 $\hat{P}(v)$ 를 이용해 식 (2)로부터 $v^*(\omega)$ 를 추정한다는 것을 의미한다. 따라서 나이브 베이즈 분류의 핵심은 학습 데이터를 잘 구성하는 데에 있다고 하겠다.

3.2.2 학습 데이터의 구성

학습 데이터를 구성하기 위해 우리는 먼저 증권신고서의 투자위험요소란에서 약 90,000개의 문장 중 10,000개의 문장을 임의로 추출하였다. 추출한 문장은 연구자가 직접 읽어가며 각 문장의 어조가 긍정적인지, 중립적인지, 또는 부정적이거나 불확실성을 나타내는지 분류하였다.

<표 1>은 분류한 10,000개 문장을 긍정적 어조와 부정적 어조로 나눈 후 세 낱말로 구성된 구절(Trigram)을 등장 빈도 수를 기준으로 나열한 것이다. 이 표로부터 부정적이라고 분류된 문장은 대개 “부정적인 영향” 등 예상되는 위험을 직접적으로 언급함을, 긍정적이라고 분류된 문장은 “미치는 사항은 없습니다” 등 “이중 부정”을 통해 예상되는 위험의 영향을 한정하는 표현을 포함함을 유추할 수 있다. Trigram을 통해 제시될 경우 부정적인 문장과 긍정적인 문장이 유사하다고 판단할 수 있으나, 이것은 분석대상인 투자위험요소란이 기본적으로 회사의 부정적 측면인 위험요소를 다루고 있기 때문인 것으로 보인다.

Li(2010)를 따라 우리는 나이브 베이즈 분류 모형이 각 문장을 긍정, 중립, 부정 또는 불확실 등 총 세 가지의 어조로 분류하도록 설계한 뒤($V = \{\text{긍정, 중립, 부정 또는 불확실}\}$), 구성된 문장-어조 데이터를 통해 학습시켰다. 학습은 등장하는 전체 단어 중 빈도 수 기준으로 상위 20,000개의 단어만을 사용하도록 했다. 이렇게 학습된 모형은 나머지 80,000개 문장의 어조를 추정하는 데에 사용되었다. 신고서의 어조는 각 신고서 내 서술된 개별 문장에 대해 어조를 평가하고 이를 기초로 신고서의 평균적인 어조로 측정했다. 평균을 계산하기 위해 우리는 긍정적인 문장에는 1, 중립적인 문장에는 0, 부정적이거나 불확실성을 나타내는 문장에는 -1을 부여했다. 그 결과 각 신고서의 평균 어조는 최소 -1에서 최대 1사이의 값을 갖게 되었다.

〈표 1〉 어조(Tone)에 따른 10개 최상위 빈도의 구절(Trigram)

Panel A: 부정적 또는 불확실 어조

Trigram	N
영향을 미칠 수	460
미칠 수 있습니다	380
부정적인 영향을 미칠	349
받을 수 있습니다	156
영향을 받을 수	153
발생할 수 있습니다	107
수익성에 부정적인 영향을	89
변경될 수 있습니다	76
영향을 줄 수	61
투자금액의 일부 또는	61

Panel B: 긍정적 어조

Trigram	N
가져오거나 중요한 영향을	24
미치는 사항은 없습니다	24
변동을 가져오거나 중요한	24
영향을 미치는 사항은	24
이외에 자산 부채	24
작성기준일 이후 본	24
재무제표의 작성기준일 이후	24
중대한 변동을 가져오거나	24
중요한 영향을 미치는	24

〈그림 1〉 증권신고서에 나타난 평균 어조 히스토그램

〈그림 1〉은 신고서 어조의 히스토그램을 보여준다. X축 Tone은 신고서에 사용된 어조의 평균, Y축 Count는 전체 표본 수 대비 해당 구간의 어조가 차지하는 비중(%)을 나타낸다.

첫째, 어조는 평균적으로 음수의 값을 갖고 있으며 이는 평균적으로 부정적인 어조의 문장이 긍정적인 어조의 문장보다 많이 사용되었음을 의미한다. 이는 본 논문의 분석이 증권신고서 가운데 투자위험요소란에 한정된다는 점을 고려할 때 자연스러운 결과로 보인다. 둘째, 어조는 표본 내에서 고루 분포해 있다. 이는 증권신고서의 어조가 회사간 충분히 다르다는 것을 나타내며, 이를 설명변수로 하는 추가적인 분석 또한 가능하다는 것을 시사한다.

3.2.3 어조의 강건성 검토

나이브 베이즈 분류 모형의 성능을 검토하기 위해, 우리는 두 가지 추가적인 분석을 수행했다. 먼저, 우리는 5분할 교차 검증 시험(5-fold cross-validation test)을 통해 모형의 표본 외 성능을 시험했다. 우리는 전체 10,000개의 학습용 데이터를 5분할 한 뒤, 그 중 4개의 부분 데이터로 모형을 학습하고 나머지 1개의 부분 데이터를 통해 이 모형으로 예측된 문장의 어조와 실제 코더가 분류한 어조가 얼마나 일치하는지 검토했다. 검토 결과는 <표 2>와 같다. 나이브 베이즈 분류 모형은 5분할 교차 검증 시험에서 평균적으로 약 75.9%의 정확도를 보였으며, 이는 나이브 베이즈 분류 모형을 활용한 Frank and Sanati(2018)의 Appendix <Table A1>에서 보고되는 교차 검증 시험 정확도인 78%+ -2.5%와 유사한 수치이다. 이를 근거로 우리는 우리의 나이브 베이즈 모형이 표본 외 분류에서도 잘 작동할 것이라고 판단했다.

<표 2> 교차 검증 시험

검증 표본	분할 1	분할 2	분할 3	분할 4	분할 5
정확도	72.95%	75.40%	78.50%	76.45%	75.95%

두 번째로 이 논문에서는 연구자 외에 두 명의 추가적인 코더(코더 A, 코더 B)를 고용해 우리의 학습용 데이터가 지나치게 주관적으로 구성된 것이 아닌지 검토했다. 코더 A는 서울대학교 경영대학 석사과정, 코더 B는 서울대학교 경영대학 학사과정 학생으로, 이들은 각자 독립적으로 어조를 분류하였다. 이들로 하여금 앞에서 논의한 임의로 추출한 10,000개의 문장의 어조를 분류하게 하고 이들의 분류를 이용해 나이브 베이즈 모형을 학습시키고 나머지 문장을 분류했다. <표 3>은 저자 및 각 코더가 분류한 학습 데이터를 가지고 측정한 증권신고서 어조의 상관계수를 나타낸다. 우리는 상관계수가 최소 0.887에 이를 정도로 높다는 것을 확인했다.

<표 3> 상관계수표

	코더 A	코더 B	저자
코더 A	1	0.930	0.887
코더 B		1	0.911
저자			1

4. 증권신고서의 어조와 신규상장 변수

4.1 모형 및 요약통계량

이 절에서 우리는 제 3장에서 측정한 증권신고서의 어조(*Tone*)와 신규상장 변수 간의 관계를 분석한다. 구체적으로, 회귀분석 모형은

$$Y_{ijt} = \alpha + \beta \text{Tone}_{ijt} + X_{ijt}\gamma + \lambda_j + \lambda_t + \epsilon_{ijt} \quad (3)$$

이며, i 는 신규상장기업, j 는 산업, t 는 연도를 나타낸다. λ_j 와 λ_t 는 각각 산업과 연도의 고정효과를 나타내는 더미변수이다. 산업은 한국표준산업분류코드의 대분류 기준을 통해 분류했다. 종속변수로는 공모가격 변화율($100 \times \frac{\text{최종 공모가격}}{\text{공모 희망가액 밴드의 중간값}}$)과 상장 직후 수익률($100 \times \frac{\text{상장 후 주가}}{\text{최종 공모가격}}$)을 분석했다. Loughran and McDonald(2013)은 증권신고서의 어조가 회사의 가치평가 불확실성과 연관된다고 주장하면서, 미국 IPO 시장에서 어조가

〈표 4〉 요약통계량

표본은 2009년 6월부터 2017년 사이에 한국에서 KOSPI나 KOSDAQ에 신규상장된 기업 442개를 포함한다. *Tone*은 제 3절에서 묘사한 방법론으로 측정한 증권신고서의 어조이다. KOSPI는 KOSPI 시장 여부를 나타내는 더미 변수, Market return은 상장 직전 90일간 평균 KOSPI 수익률, Reputation은 직전 연도 IPO 시장에서 주관사 회사의 매출액 비, Age는 기업의 영업 연도 수, Size는 상장 전 기업의 자산 규모의 자연로그값, Leverage는 총 부채를 총 자산으로 나눈 부채비율, Profitability는 영업이익을 총 매출로 나눈 기업의 수익성을 나타낸다. ***, **, *는 각각 1%, 5%, 10% 수준의 통계적 유의성을 나타낸다.

Panel A: 분포

	N	Mean	St. Dev.	Min	Q1	Q3	Max
Tone	442	-0.17	0.12	-0.57	-0.26	-0.08	0.18
KOSPI	442	0.19	0.39	0	0	0	1
Age	442	14.70	9.99	0.88	9.01	16.77	67.25
Size	442	17.97	1.44	15.46	17.04	18.57	24.98
Leverage	442	0.50	0.26	0.03	0.34	0.63	2.90
Profitability	442	0.14	0.15	-0.66	0.08	0.21	0.79
Initial Return	442	0.30	0.43	-0.33	-0.03	0.54	1.60

Panel B: 상관계수표

	Tone of Description (1)	KOSPI (2)	Age (3)	Size (4)	Leverage (5)	Profitability (6)
(1)	1	-0.367***	-0.210***	-0.383***	-0.062	0.050
(2)		1	0.288***	0.709***	0.132***	-0.134***
(3)			1	0.331***	-0.015	-0.158***
(4)				1	0.232***	-0.115**
(5)					1	-0.432***
(6)						1

공모가격 변화율 및 상장 직후 수익률과 음의 관계를 갖는다는 것을 보인 바 있다. 우리는 이러한 관계가 한국 IPO 시장에서도 관찰되리라고 예상했다($\beta < 0$).

X_{jt} 는 통제변수를 나타내는데, 상장 전 기업의 자산 규모의 자연로그값(Size), 총 부채 ÷ 총 자산으로 계산한 부채비율(Leverage), 기업의 연령(Age), 총 매출 대비 영업이익으로 계산한 기업의 수익성(Profitability), KOSPI 시장 여부를 나타내는 더미 변수(KOSPI) 등 기업 특성을 포함한다. 또한, 시황이 좋을 수록 상장 및 신주발행이 활발하다는 Baker and Wurgler(2002)를 따라 상장 직전 90일간 평균 KOSPI 수익률(Market return)을 통제했고, 주관사회사의 명성이 높을 수록 발행가액이 높아진다는 Carter(1990)을 따라 직전 연도 IPO 시장에서 인수주관회사(lead underwriter)의 매출액 비중(Reputation)을 포함한다. 표준오차는 White(1980)의 추정치를 사용했다. 기업 단위 변수는 FnGuide에서 제공하는 DataGuide에서, 그 이외의 변수는 KIND에서 추출하였다.

<표 4>는 이 논문에서 사용된 요약통계량을 보여준다. Panel B의 상관계수표는 KOSDAQ에 상장되는 회사일수록, 영업 연도수가 짧은 회사일수록, 기업의 규모가 작은 회사일수록 증권 신고서의 어조가 긍정적임을 시사한다.

4.2 증권신고서 어조와 공모가격 변화율 및 상장 수익률

<표 5>는 증권신고서의 평균 어조에 대한 회귀분석 결과를 나타낸다. 첫 번째와 두 번째 열은 공모가격 변화율을 종속변수로 하는 회귀분석 결과이다. 두 모형 모두에서 Tone의 회귀계수는 음의 값을 보였으며, 통계적으로도 5% 수준에서 유의했다. 이는 증권신고서의 어조가 부정적인 회사의 공모가격 변화율이 높음을 보여준다. 이러한 결과는 최초 공모 희망가액 밴드를 상대적으로 낮게 형성하는 경향이 있음을 시사한다. 이 결과에 대해서는 두 가지 서로 다른 해석이 가능하다. 하나는 증권신고서의 어조가 그 동안 정량화하기 어려웠던 회사의 정보를 반영하는 것이고 다른 하나는 이것이 경영자의 심리적 감성을 반영한다는 것이다. 우리는 선행연구인 Loughran and McDonald(2013)의 해석을 따라 이 결과를 부정적인 정보가 많은 회사는 증권신고서의 어조가 부정적이기 마련이며, 이런 회사는 공모가격을 설정하는 데에 있어 협상력이 낮기 때문에, 상대적으로 주관사회사에 유리한 방향으로 최초 공모가격이 낮게 형성된 결과라고 해석한다.

<표 5>의 열 (3)과 (4)는 최종 공모가격 대비 상장 당일 증가(%)를 종속변수로 하는 회귀분석 결과이다. Tone의 회귀계수는 음의 값을 보였지만, 통계적으로 유의하지 않았다. 한국 증권시장에는 + -30%의 일일 주가 변동 제한 폭이 있기 때문에 상장 당일의 주가는 정보를 충분히 반영하지 못한다는 비판이 있을 수 있다. 그러나 통계적으로 유의하지 않은 관계는 열 (5)와 (6)에서 상장 3일 후의 증가를 분석하더라도 마찬가지였다. 정리하자면, 우리는 상장 직후 수익률과 증권신고서의 어조 사이의 관계에 있어 선행연구와 부합하는 결과를 찾을 수 없었다.

〈표 5〉 증권신고서 어조와 공모가격 변화율 및 상장 수익률

모든 회귀분석은 2009년 6월부터 2017년 사이에 한국에서 KOSPI나 KOSDAQ에 신규상장된 기업 442개를 대상으로 한다. Tone은 제3절에서 묘사한 방법론으로 측정한 증권신고서의 어조이다. KOSPI는 KOSPI 시장 여부를 나타내는 더미 변수, Market return은 상장 직전 90일간 평균 KOSPI 수익률, Reputation은 직전 연도 IPO 시장에서 주관사 회사의 매출액 비, Age는 기업의 영업 연도 수, Size는 상장 전 기업의 자산 규모의 자연로그값, Leverage는 총 부채를 총 자산으로 나눈 부채비율, Profitability는 영업이익을 총 매출로 나눈 기업의 수익성을 나타낸다. 표준오차는 White(1980)의 추정치를 사용했다. ***, **, *는 각각 1%, 5%, 10% 수준의 통계적 유의성을 나타낸다.

	최종 공모가격/ 공모희망가액 밴드의 중간값(%)		상장 당일 증가/ 최종 공모가격 (%)		상장 3일 후 증가/ 최종 공모가격 (%)	
	(1)	(2)	(3)	(4)	(5)	(6)
Tone	-1.616** (0.723)	-1.691** (0.752)	-0.099 (2.215)	-1.423 (2.250)	-1.499 (2.395)	-2.986 (2.470)
KOSPI	-4.794** (1.950)	-4.527* (2.519)	-12.069** (5.820)	3.164 (7.175)	-8.591 (7.136)	8.872 (8.339)
Market return	13.098 (8.437)	12.352 (8.462)	68.440** (28.276)	74.786*** (27.483)	65.048** (31.942)	72.686** (30.552)
Reputation	0.630** (0.299)	0.564* (0.301)	0.097 (1.601)	0.781 (1.567)	0.243 (1.606)	1.069 (1.567)
Age		-0.072 (0.076)		0.167 (0.195)		0.235 (0.220)
Size		0.192 (0.771)		-7.629*** (2.088)		-8.922*** (2.502)
Leverage		-1.795 (3.763)		2.355 (9.726)		3.486 (12.663)
Profitability		2.805 (7.868)		-31.891** (15.364)		-38.207* (21.494)
Constant	54.494*** (3.793)	52.113*** (13.780)	38.007*** (11.044)	175.227*** (35.636)	11.207 (11.720)	171.146*** (41.060)
Year dummies	Y	Y	Y	Y	Y	Y
Industry-fixed	Y	Y	Y	Y	Y	Y
N	442	442	442	442	442	442
Adjusted R ²	0.063	0.058	0.031	0.062	0.017	0.050

4.3 증권신고서 어조와 장기 주식 성과 및 주가 변동성

〈표 6〉에서는 장기적인 주식 성과 및 현금흐름 추정에서의 불확실 요소와 증권신고서 사이의 관계를 보여준다. 장기적인 주식 성과와 현금흐름 추정에서의 불확실 요소는 각각 상장 후 [+5, +65] 거래일 간의 누적 주가 수익률과 [+5, +65] 거래일 사이의 주가변동성으로 측정했다. 증권신고서의 어조가 회사의 가치평가 불확실성 정도를 반영한다면, 어조가 부정적일 수록 위험 프리미엄과 불확실성을 반영해 상장 후 [+5, +65] 거래일 간의 누적 주가 수익률 및 주가변동성은 높을 것으로 예상된다. 그러나 회귀분석 결과는 장기적인 주식 성과 및 현금흐름 추정에서의 불확실성 요소와 증권신고서 어조 사이의 통계적으로 유의한 관계를 찾을 수 없음을 보여준다.

〈표 6〉 증권신고서 어조와 장기 주식 성과 및 주가 변동성

모든 회귀분석은 2009년 6월부터 2017년 사이에 한국에서 KOSPI나 KOSDAQ에 신규상장된 기업 442개를 대상으로 한다. Tone은 제 3절에서 묘사한 방법론으로 측정된 증권신고서의 어조이다. KOSPI는 KOSPI 시장 여부를 나타내는 더미 변수, Market return은 상장 직전 90일간 평균 KOSPI 수익률, Reputation은 직전 연도 IPO 시장에서 주관사 회사의 매출액 비, Age는 기업의 영업 연도 수, Size는 상장 전 기업의 자산 규모의 자연로그값, Leverage는 총 부채를 총 자산으로 나눈 부채비율, Profitability는 영업이익을 총 매출로 나눈 기업의 수익성을 나타낸다. 표준오차는 White(1980)의 추정치를 사용했다. ***, **, *는 각각 1%, 5%, 10% 수준의 통계적 유의성을 나타낸다.

	상장 후 [5, 65] 거래일 간 주가 수익률		상장 후 [+ 5, + 65] 거래일 간 주가 변동성	
	(1)	(2)	(3)	(4)
Tone of Description	0.715 (1.811)	1.201 (1.865)	0.007 (0.069)	-0.075 (0.070)
KOSPI	3.570 (4.377)	-0.971 (5.449)	-0.973*** (0.169)	-0.092 (0.194)
Market Timing	-71.996*** (21.353)	-72.083*** (21.923)	1.628** (0.783)	2.005*** (0.775)
Reputation	0.353 (0.719)	0.432 (0.752)	-0.015 (0.030)	0.011 (0.029)
Age		0.117 (0.121)		0.0002 (0.007)
Size		1.306 (1.703)		-0.398*** (0.067)
Leverage		1.863 (6.275)		0.224 (0.334)
Profitability		-13.829 (13.159)		-0.784** (0.382)
Constant	50.292*** (9.691)	29.531 (30.533)	5.414*** (0.356)	12.349*** (1.071)
Year dummies	Y	Y	Y	Y
Industry-fixed	Y	Y	Y	Y
N	442	442	442	442
Adjusted R ²	0.035	0.033	0.160	0.233

4.4 강건성 검토

Kim et al.(2018)는 일상생활에서의 긍정어/부정어를 다룬 Harvard IV 단어 리스트를 Google Translate로 번역한 후 수작업으로 단어들의 한글 활용형을 적용시켜서 일상생활의 긍정어/부정어 리스트를 만든 바 있다. 우리는 Kim et al.(2018)의 단어 리스트를 통해 증권 신고서의 전체 단어 수 대비 긍정어/부정어 단어들이 차지하는 비중을 각각 측정하고, 이렇게 구성된 변수를 활용해 위에서 진행된 분석을 다시 수행했다.

〈표 7〉은 나이브 베이즈 분류 모형을 통해 구성한 *Tone* 변수와 Kim et al.(2018)의 단어 리스트를 이용한 일상생활 긍정어 비율(*Positive*) 및 일상생활 부정어 비율(*Negative*)

사이의 상관관계를 보여준다. Tone 변수는 -1과 1 사이의 값을 가지며, 1에 가까울 수록 긍정적, -1에 가까울 수록 부정적이다. 따라서, 만약 Kim et al.(2018)의 단어 리스트를 이용해 측정한 어조가 나이브 베이즈 분류 모형을 통해 구성한 어조와 유사하다면, 긍정어 비율 및 부정어 비율 변수와 Tone은 각각 +와 -의 상관관계를 가질 것으로 예상된다. 분석 결과 긍정어 비율은 Tone과 통계적으로 유의하지 않은 상관관계를, 반면 부정어 비율은 Tone과 통계적으로 유의한 음의 상관관계를 보였다. 한편, 긍정어 비율과 부정어 비율 사이에도 음의 상관관계를 발견할 수 없었다. 이러한 분석 결과는 Kim et al.(2018) 단어 리스트를 통해 구성한 일상생활 부정어 비율은 본 논문에서 측정한 Tone과 비교적 유사하지만, 긍정어 비율의 경우 그렇지 않다는 것을 시사한다.

<표 7> Kim et al.(2018) 단어 리스트 어조의 상관관계

***, **, *표시는 각각 1%, 5%, 10% 수준의 통계적 유의성을 나타낸다.

	Tone	Positive (Kim et al., 2018)	Negative (Kim et al., 2018)
Tone	1	-0.03	-0.44***
Positive(Kim et al., 2018)		1	0.05
Negative(Kim et al., 2018)			1

<표 8>은 공모가격 변화율, 상장 수익률, 장기 주식 성과 및 주식 변동성을 종속변수로 하고, 일상생활 부정어 및 긍정어 비율을 설명변수로 하는 회귀분석 결과를 보여준다. 공모

<표 8> Kim et al.(2018) 단어 리스트 어조와 IPO 변수의 관계

모든 회귀분석은 2009년 6월부터 2017년 사이에 한국에서 KOSPI나 KOSDAQ에 신규상장된 기업 442개를 대상으로 한다. Positive는 Kim et al.(2018)의 긍정어 비율, Negative는 부정어 비율 측정한 증권신고서의 어조이다. KOSPI는 KOSPI 시장 여부를 나타내는 더미 변수, Market return은 상장 직전 90일간 평균 KOSPI 수익률, Reputation은 직전 연도 IPO 시장에서 주관사 회사의 매출액 비, Age는 기업의 영업 연도 수, Size는 상장 전 기업의 자산 규모의 자연로그값, Leverage는 총 부채를 총 자산으로 나눈 부채비율, Profitability는 영업이익을 총 매출로 나눈 기업의 수익성을 나타낸다. 표준오차는 White(1980)의 추정치를 사용했다. ***, **, *는 각각 1%, 5%, 10% 수준의 통계적 유의성을 나타낸다.

	최종 공모가격/ 공모희망가액 밴드의 중간값(%)		상장 당일 종가/ 최종 공모가격 (%)		상장 3일 후 종가/ 최종 공모가격 (%)		상장 후 [5, 65] 거래일 간 추가 수익률		상장 후 [+5, +65] 거래일 간 추가 변동성	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Negative	2.794 (2.310)		-1.204 (6.398)		0.154 (6.943)		-7.142 (5.274)		-0.013 (0.178)	
Positive		1.438 (1.562)		6.911 (5.097)		4.030 (5.592)		-2.303 (3.415)		-0.076 (0.124)
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry-fixed	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N	442	442	442	442	442	442	442	442	442	442
Adjusted R ²	0.051	0.050	0.062	0.066	0.047	0.048	0.037	0.033	0.230	0.231

가격 변화율과 어조 사이의 부정적 관계를 보인 <표 5>와 달리, 일상생활 긍정어/부정어 비율은 고려된 종속변수와 통계적으로 유의한 상관관계를 찾을 수 없었다. 우리는 그 이유로 크게 두 가지를 가능성을 제시한다. 첫째, <표 1>의 구절 분석에서도 나타나듯, 한국어 증권신고서의 긍정적 표현은 부정적인 사건의 가능성을 한정하는 이중 부정의 형태로 나타나는 경우가 많다. 부정어 단어 리스트를 활용한 분석의 경우, 이처럼 이중 부정을 통해 긍정적인 어조를 나타내는 문장을 부정적이라고 분류할 가능성이 높다. 둘째, Loughran and McDonald(2011)에서 지적하듯, 일상생활 단어 리스트가 한국 금융시장 텍스트의 어조를 잘 포착하지 못할 수 있다. 그러나 Tetlock(2007)이 보인 바처럼, 일상생활 긍정어/부정어 리스트 또한 향후 금융시장 텍스트 어조 분석에 활용될 여지가 많은 만큼, 향후 추가적인 분석이 필요한 부분이라 하겠다.

4.5 고찰

이 논문은 한국 증권시장에서 IPO 상장 수익률 및 주가변동성과 증권신고서의 어조 사이의 관계에 있어 미국 증권시장을 대상으로 한 선행연구와 부합하는 결과를 찾을 수 없었다. 그 이유 중 하나로 IPO 제도의 차이를 생각해볼 수 있다. 한국 IPO 시장은 미국과 달리 기관투자자의 수요예측(bookbuilding) 결과를 상장 전에 공시하도록 되어 있고, 이렇게 공시된 기관투자자의 수요예측 절차 참여 결과는 IPO 주식에 대한 시장 평가에 영향을 끼칠 수 있다는 것이다. 따라서 본 연구는 기관투자자의 수요예측 절차 참여 정보를 함께 고려해 증권신고서의 어조와 주가 반응 사이의 관계를 분석할 것을 향후 연구과제로 제시하는 바이다.

5. 결론

이 논문은 2009년 6월부터 2017년 12월에 상장된 한국 기업들의 증권신고서에 사용된 텍스트가 긍정적 또는 부정적 어조인가를 최초로 측정하고, 어조가 이후 IPO 성과변수들과 갖는 관계를 분석하였다. 증권신고서의 어조는 표본 내에서 고루 분포해 있어 기업간 어조의 차이가 있음을 보여준다. 또한, 머신러닝을 통해 분류한 어조에 대한 분석 결과 외국의 선행연구와 일부 일관되는 증거를 찾을 수 있었다. 증권신고서의 어조가 부정적인 회사에서 공모가격 변화율(offer price revision)이 크게 나타나는 경향이 있었다. 하지만, 어조는 상장 직후 수익률과는 통계적으로 유의한 관계가 없었다.

이 논문은 머신러닝 기법을 활용해 한국 금융시장에서 기업들이 제출한 자료를 바탕으로 텍스트의 어조를 최초로 체계적으로 측정했다는 점에서 의의가 있다. 특히, 텍스트 분석 방법론을 구체적으로 예시함으로써 이후 연구자들의 한국어 텍스트 분석에 도움을 주고자 했다.

참 고 문 헌

- 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* 59 (3), pp. 1259–1294.
- Baker, M. and J. Wurgler, 2002, Market Timing and Capital Structure, *The Journal of Finance* 57 (1), pp. 1–32.
- Baker, S. R., N. Bloom, and S. J. Davis, 2016, Measuring Economic Policy Uncertainty, *The Quarterly Journal of Economics* 131 (4), pp. 1593–1636.
- Beatty, R. P. and J. R. Ritter, 1986, Investment banking, reputation, and the underpricing of initial public offerings, *Journal of Financial Economics* 15 (1–2), pp. 213–232.
- Benveniste, L. M. and P. A. Spindt, 1989, How investment bankers determine the offer price and allocation of new issues, *Journal of Financial Economics* 24 (2), pp. 343–361.
- Benveniste, L. M. and W. J. Wilhelm, 1990, A comparative analysis of IPO proceeds under alternative regulatory environments, *Journal of Financial Economics* 28 (1–2), pp. 173–207.
- Buehlmaier, M. M. M. and T. M. Whited, 2018, Are Financial Constraints Priced? Evidence from Textual Analysis, *Review of Financial Studies* 55, pp. 2219–2236.
- Carter, R. B., 1990, Initial Public Offerings and Underwriter Reputation, *The Journal of Finance* 45 (4), pp. 1045–1067.
- Chen, H., P. De, Y. J. Hu, and B. H. Hwang, 2014, Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media, *Review of Financial Studies* 27 (5), pp. 1367–1403.
- Frank, M. Z. and A. Sanati, 2018, How does the stock market absorb shocks?, *Journal of Financial Economics* 129 (1), pp. 136–153.
- Garca, D., 2013, Sentiment during Recessions, *The Journal of Finance* 68 (3), pp. 1267–1300.
- Gentzkow, M. and J. M. Shapiro, 2010, What drives media slant? Evidence from U.S. daily newspapers, *Econometrica* 78 (1), pp. 35–71.
- Gentzkow, M., B. Kelly, and M. Taddy, 2017, Text as Data, *NBER Working Paper*, pp. 1–63.
- Gentzkow, M., J. Shapiro, and M. Taddy, 2018, *Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech*, Cambridge, MA: National Bureau of Economic Research.
- Hastie, T., J. Friedman, and R. Tibshirani, 2001, *The Elements of Statistical Learning*, Springer, pp. 1–533.
- Ibbotson, R. G., 1975, Price performance of common stock new issues, *Journal of Financial Economics* 2 (3), pp. 235–272.

- Kim, Y. H. A., H. G. Kang, and J. K. Lee, 2018, Can Big Data Forecast North Korean Military Aggression?, *Defence and Peace Economics* 29 (6), pp. 666–683.
- Lee, W. J., T. Jung, M. Cho, and S. Y. Lim, 2018, An Explorative Study to Assess Feasibility of Annual Report Readability Research in Korea, *Korean Accounting Review* 43 (4), pp. 37–100.
- Li, F., 2008, Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics* 45 (2–3), pp. 221–247.
- Li, F., 2010, The Information Content of Forward-Looking Statements in Corporate Filings—A Naive Bayesian Machine Learning Approach, *Journal of Accounting Research* 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* 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* 66 (1), pp. 35–65.
- Loughran, T. and B. McDonald, 2013, IPO first-day returns, offer price revisions, volatility, and form S-1 language, *Journal of Financial Economics* 109 (2), pp. 307–326.
- Park, J., H. Kim, and J. Lee, 2015, A Study on the Textual Analysis Research Environment using the DART System in Korea, *Korean Accounting Journal* 24 (4), pp. 199–221.
- Ritter, J. R., 1984, The “Hot Issue” Market of 1980, *The Journal of Business* 57 (2), pp. 215–240.
- Rock, K., 1986, Why new issues are underpriced, *Journal of Financial Economics* 15 (1–2), pp. 187–212.
- Stephens-Davidowitz, S., 2014, The cost of racial animus on a black candidate: Evidence using Google search data, *Journal of Public Economics* 118 (C), pp. 26–40.
- Tetlock, P. C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance* 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* 63 (3), pp. 1437–1467.
- White, H., 1980, A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica* 48 (4), p. 817.
- Yook, K., 2018, CEOs Talk about Social Responsibilities and Sustainable Performance: Applications of Text Mining Approach. *Korean Accounting Journal* 27 (1), pp. 253–279.

〈부록〉

〈그림 A1〉 텍스트 분석을 위한 R코드

```
library(tidyverse)
library(rvest)

prospectus_rcp_no <- read_csv("input/rcp_no.csv", col_types = "cn")

firm <- prospectus_rcp_no$id[1]
rcp_no <- prospectus_rcp_no$prospectus[1]
# Let's start with one example

url <- paste0("http://dart.fss.or.kr/dsaf001/main.do?rcpNo=", rcp_no)
# Now you have the URL address for the targeted document.
# However, the document in DART is organized by multilayer HTML files. Each HTML
file contains the information on each section of the document.
# It means that, you need to collect multiple URL addresses for sections in the document.

html <- read_html(url)
html %>%
  html_nodes("script") %>%
  html_text() %>%
  str_replace_all("WWs+", " ") %>%
  str_split("WW//") %>%
  unlist() %>%
  print()
# If you carefully read this html code, you can find that this code is organized with a
function `viewDoc`. This function contains 6 parameters.
# Using the parameters, you can generate URL addresses for each section and access to
them.

# Lines of codes below are designed to find these parameters for each section using the
regular expression.
text <- html %>%
  html_nodes("script") %>%
  html_text() %>%
  str_replace_all("WWs+", " ") %>%
  str_split("WW//") %>%
  unlist()

tree_mask <- str_detect(text, "treeNodeWWd{1} WW= new TreeWW.TreeNode")
tree <- text[tree_mask]

section_regex <- regex("(?<=text: W")[:print:]]+?(?=W"WW,)")
section <- tree %>%
  str_extract_all(section_regex) %>%
  unlist()

params_regex <- regex("(?<=WW{viewDocWW()[:print:]]+?(?=WW))")
params <- tree %>%
  str_extract_all(params_regex) %>%
  unlist() %>%
```

```

    str_replace_all("", "") %>%
    str_trim("both")

print(params)
# Now we find parameters for every sections in the document.

# Let's make the function.
url_parameter_collect <- function(firm, rcp_no){
  Sys.sleep(2)
  # Not to be blocked from DART, 2-seconds break per one access is necessary.
  print(paste0("This function is now dealing with firm ", firm))
  # To check whether our function is running during the loop.

  url <- paste0("http://dart.fss.or.kr/dsaf001/main.do?rcpNo=", rcp_no)
  html <- read_html(url)

  text <- html %>%
    html_nodes("script") %>%
    html_text() %>%
    str_replace_all("WWs+", " ") %>%
    str_split("WW//") %>%
    unlist()

  tree_mask <- str_detect(text, "treeNodeWWd{1} WW= new TreeWW.TreeNode")
  tree <- text[tree_mask]

  section_regex <- regex("(?<=text: W")[[[:print:]]+?(?=W"WW,)")
  section <- tree %>%
    str_extract_all(section_regex) %>%
    unlist()

  params_regex <- regex("(?<=WW{viewDocWW()[[[:print:]]+?(?=WW))")
  params <- tree %>%
    str_extract_all(params_regex) %>%
    unlist() %>%
    str_replace_all("", "") %>%
    str_trim("both")

  data_frame("id" = firm,
             "section" = section,
             "params" = params) %>%
    separate(params, into = c("rcpNo", "dcmNo", "eleId", "offset", "length", "dtd"), sep=",
  ")
}

prospectus_url_parameter <- map2_dfr(prospectus_rcp_no$id, prospec-
tus_rcp_no$prospectus, url_parameter_collect)
print(prospectus_url_parameter)
# Now you collect parameters for every section in every document.

write_csv(prospectus_url_parameter, "temp/prospectus_url_parameter.csv")

prospectus_url_parameter <- read_csv("temp/prospectus_url_parameter.csv")
# Assume that you are interested in the section `` I . 모집 또는 매출에 관한 일반사항."
```

```
# Then, you can construct the url addresses for this section as below.
targeted_url <- prospectus_url_parameter %>%
  filter(section == " I . 모집 또는 매출에 관한 일반사항") %>%
  mutate(url = paste0("http://dart.fss.or.kr/report/viewer.do?",
    "rcpNo=", rcpNo, "&",
    "dcmNo=", dcmNo, "&",
    "eleId=", eleId, "&",
    "offset=", offset, "&",
    "length=", length, "&",
    "dtd=", dtd))

# First, download HTML files.
html_download <- function(id, url){
  Sys.sleep(2)
  print(paste0("This function is now working on firm ", id))

  read_html(url) %>%
    write_html(file = paste0("temp/HTMLfiles/", id, ".html"))
}

walk2(targeted_url$id, targeted_url$url, html_download)
# Check the directory `HTMLfiles`.
# Using these HTML files, you can do the analysis what you want.
# For example, to extract sentences
HTMLfiles <- dir("temp/HTMLfiles/")

document_tokenize <- function(html){
  path <- paste0("temp/HTMLfiles/", html)

  sentences <- read_html(path) %>%
    html_nodes("p") %>%
    html_text(trim = TRUE) %>%
    str_replace_all("WWs+ ", " ") %>%
    str_subset("다.$") %>%
    str_split("(?<=다WW.)") %>%
    unlist() %>%
    str_replace_all("[:punct:]", "") %>%
    str_trim("both") %>%
    setdiff("")

  data_frame("id" = str_sub(html, 1, 6),
    "sentences" = sentences)
}

sentences <- map_dfr(HTMLfiles, document_tokenize)
print(sentences)
# Now you extarct sentences.

write_excel_csv(sentences, "output/sentences.csv")
```

Text Analysis for IPO firms in Korea: Analysis of Korean Texts in Registration Statements via Machine Learning^{*}

Yongseok Kim

Master Student, Seoul National University

Sung Wook Joh^{}**

Professor, Seoul National University

Abstract

This paper analyzes the texts in registration statements that IPO firms filed when they issued securities in the Korean stock market during June 2009 to December 2017. We classify tones of texts as positive or negative via machine learning. An IPO firm with more negative tones in the risk factor section of registration statement tends to set its final offer price ratio higher than the initial offer price. However, the tone of the texts is not significantly related to the IPO's initial returns.

This is the first paper to analyze tones of texts in the natural language submitted by companies in the Korean financial market. In addition, this paper provides the program used for the analysis of non-standardized text to help scholars analyze Korean texts in the future.

Keywords: Text Analysis; Korean Texts; Machine Learning; IPO; Registration Statement

JEL Classification: G14, G23, G24

* We are grateful to two anonymous referees for their constructive and detailed comments and suggestions. We appreciate support from Institute of Management Research.

** Corresponding Author. Address: Department of Finance, College of Business Administration, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, 08826; E-mail: swjoh@snu.ac.kr, Sungwook.joh@gmail.com; Tel: +82-2-880-9384; Fax: +82-2-873-4821.