ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 June: 28(2): 237~262 http://dx.doi.org/10.13088/jiis.2022.28.2.237

증권신고서의 TF-IDF 텍스트 분석과 기계학습을 이용한 공모주의 상장 이후 주가 등락 예측*

양수연
KAIST 경영대학원 경영공학부
(olivia.yang@kaist.ac.kr)

이채록
(부산대학교 경영학과)
li016010@naver.com

원종관
부산대학교 경영학과
(jongkwan1@pusan.ac.kr)

홍태호
부산대학교 경영학과
(hongth@pusan.ac.kr)

본 연구는 개인투자자들의 투자의사결정에 도움을 주고자, 증권신고서의 TF-IDF 텍스트 분석과 기계학습을 이용해 공모주의 상장 5거래일 이후 주식 가격 등락을 예측하는 모델을 제시한다. 연구 표본은 2009년 6월부터 2020년 12월 사이에 신규 상장된 691개의 국내 IPO 종목이다. 기업, 공모, 시장과 관련된 다양한 재무적 및 비재무적 IPO 관련 변수와 증권신고서의 어조를 분석하여 예측했고, 증권신고서의 어조 분석을 위해서 TF-IDF (Term Frequency - Inverse Document Frequency)에 기반한 텍스트 분석을 이용해 신고서의 투자위험요소란의 텍스트를 긍정적 어조, 중립적 어조, 부정적 어조로 분류하였다. 가격 등락 예측에는 로지스틱 회귀분석(Logistic Regression), 랜덤 포레스트(Random Forest), 서포트벡터머신(Support Vector Machine), 인공신경망(Artificial Neural Network) 기법을 사용하였고, 예측 결과 IPO 관련 변수와 증권신고서 어조 변수를 함께 사용한 모델이 IPO 관련 변수만을 사용한 모델보다 높은 예측 정확도를 보였다. 랜덤 포레스트 모형은 1.45%p 높아진 예측 정확도를 보였으며, 인공신공망 모형과 서포트벡터머신 모형은 각각 4.34%p, 5.07%p 향상을 보였다. 추가적으로 모형간 차이를 맥니마 검정을 통해 통계적으로 검증한 결과, 어조 변수의 유무에 따른 예측 모형의 성과 차이가 유의확률 1% 수준에서 유의했다. 이를 통해, 증권신고서에 표현된 어조가 공모주의 가격 등락 예측에 영향을 미치는 요인이라는 것을 확인할 수 있었다.

주제어 : 주가 예측, IPO, TF-IDF, 기계학습, 어조 텍스트 분석

논문접수일 : 2022년 6월 16일 논문수정일 : 2022년 6월 21일 게재확정일 : 2022년 6월 21일
원고유형 : 학술대회용 Fast-Track 교신저자 : 홍태호

1. 서론

기업공개(Initial Public Offering: IPO)란 주식회사가 최초로 외부투자자에게 주식을 발행하거나 기존 발행주식을 공개적으로 매각함으로써 불특정 다수에게 주식이 분산되도록 하는 것을 의미한다(Kim, 2008). IPO를 통해 기업은 대규모 자금조달, 합리적 경영, 경영 투명성 제고, 기업 홍보 등의 효과를 기대할 수 있다. 최근 이러한 IPO 시장을 향해 투자자들의 이목이 집중되고

* 본 논문은 제11회 DB금융경제 공모전 우수상 수상작 “Random Forest와 TF-IDF 기반 텍스트 분석을 이용한 IPO 주식의 상장일 가격 등락 예측”을 발전시켰음.

있으며, COVID-19의 영향으로 공모 기업 수가 줄었음에도 불구하고 2020년 한 해 동안의 공모 금액은 5.9조원으로 3개년 내 최고치를, 평균 기관수요예측 경쟁률과 평균 일반청약 경쟁률은 사상 최고치를 경신한 바 있다(Park and Han, 2021). 금융위원회가 일반청약자 참여기회 확대 방안을 발표하면서 IPO 시장을 향한 개인투자자들의 관심은 더욱 커졌다. 2021년 1월부터는 개인투자자들에게 배정되는 공모주 물량이 기존 20%에서 최대 30%까지 대폭 확대되었으며, 최소 청약증거금을 납입한 모든 청약자들에게 동등하게 배정 기회를 제공하는 균등방식이 공모주 배정 물량 중 절반 이상에 도입되어 공모주에 대한 개인투자자들의 접근성이 더욱 높아질 전망이다.

한편 공모주는 기존의 주식에 비해 가격변동성이 크다는 위험성을 수반하며 물량 또한 한정적이다. 따라서 공모주 투자 시에는 시장 및 회사에 대한 가치 분석이 매우 중요하다. 기관투자자의 경우 기업설명활동(Investor Relations: IR)과 수요예측(book building) 과정에 참여하기 때문에 기업의 적정가치를 파악할 수 있는 기회가 많은 편이지만, 일반 개인투자자는 청약을 통해 참여 여부만을 결정할 수 있으며, 참고할 수 있는 정보도 기업의 투자설명서와 기업설명활동 자료에 국한되어 있다(Cho et al., 2020). 따라서 공모주 투자 수익률 및 주가 변동 방향을 예측하여 개인투자자들의 투자 의사결정에 도움을 줄 수 있는 연구가 중요해지고 있다. 이에 따라, 다양한 재무 및 비재무 변수를 이용하여 IPO 주식 가격을 예측하는 연구가 활발하게 진행되고 있다. 전통적인 통계적 분석 방법론보다는 기계학습, 딥러닝과 같은 인공지능 방법론을 이용한 연구가 증가하는 추세이며, 수치형 변수 뿐만 아니라 텍스트 분석을 통해 얻어진 투자설명서의 어조를 변수로 사용한 연구도 증가하고 있다(Loughran and McDonald, 2013).

본 연구는 IPO 주식 가격을 예측해온 최근 선행연구의 흐름을 국내 IPO 주식에도 적용시켜 예측 성과를 향상시키고자 한다. 선행 연구에서 IPO 주식 수익률에 영향을 미치는 것으로 주장하고 있는 개별 요인들을 종합하고, 증권신고서의 어조를 추가 변수로 사용하여 국내 IPO 주식 상장 이후 거래일 기준 5일 후의 가격 등락을 예측해보고자 한다. 가격의 등락 여부를 예측하는 것이 실제 투자에서 활용도가 높기 때문에 본 연구에서는 수익률이 아닌 가격 등락 여부를 예측하며(Shin et al., 2018), 연구 표본은 2009년 6월부터 2020년 12월 사이에 신규상장된 국내 IPO 종목이다. 증권신고서의 어조 분석을 위해서는 TF-IDF (Term Frequency - Inverse Document Frequency)에 기반한 텍스트 분석을 수행하고, 가격 등락 예측에는 선행연구에서 이용되고 있는 로지스틱 회귀분석, 랜덤 포레스트, 서포트벡터머신, 인공신경망 기법을 사용한다. 본 연구는 텍스트 분석과 기계학습을 함께 이용하여 IPO 주식의 가격 등락을 예측한다는 점에서 선행연구들과 차별화된다.

본 논문의 구성은 다음과 같다. 2장에서는 IPO 관련 선행연구들을 살펴보고, 3장에서는 본 연구에서 사용한 방법론에 대해 서술한다. 4장에서는 IPO 데이터를 바탕으로 수행한 실험 및 실험 결과를 설명하고, 마지막 5장에서는 연구 결과에 대한 요약과 함께 연구의 의의 및 한계를 제시한다.

2. 선행연구

2.1. 기업공개(Initial Public Offering: IPO)

신규상장을 승인받은 기업은 증권신고서를 공시하고, 기관투자자들을 대상으로 기업설명회를 실시한 이후 수요예측 과정을 거친다. 수요예측 과정에서는 국내외 기관투자자들의 희망 공모가격과 배정물량 등을 파악하고 협의를 거쳐 공모가격이 결정된다(Kim, 2016). 이후 일반투자자들을 대상으로 청약 및 납입 절차가 이루어지고 나면 기업은 신규상장신청서를 거래소에 제출하고, 신규상장이 이루어지면 기준가격을 결정해 매매거래를 체결할 수 있게 된다. 공모주의 시초가격은 상장일 아침 공모가격의 -10% ~ +100% 범위에서 결정되며, 상장일 중 가격 변동은 시초가격의 ±30%의 범위에서 이루어진다. 일반적으로 초기에는 기업의 정보가 충분히 반영되지 않아 주식이 저평가되지만, 이후 이 부분이 점차 해소되면서 주가가 상승하는 경우가 많다(Cho et al., 2020).

IPO 저가발행은 전세계 IPO 시장에서 일반적으로 나타나는 이상현상으로, 우리나라에서는 평균 30%대, 미국에서는 10~20%의 저가발행이 발생한다고 알려져 있다(Park and Jeon, 2015). Ibbotson(1975)의 연구를 시작으로, IPO 저가발행의 원인을 밝혀 내기 위한 다수의 연구가 진행되어왔다. 주로 제시된 고전적 가설은 정보 비대칭으로 인한 저평가 가설이다(Benveniste and Spindt, 1989; Rock, 1986; Ruud, 1993). IPO 과정에는 공모기업과 기관투자자, 일반투자자, 주관증권사 등 여러 이해관계자들 사이에 상당한 정보 비대칭이 발생하는데, 이로 인한 위험부담을 최소화하기 위해 공모가격이 의도적으로 낮게 발행된다는 것이 저평가 가설의 설명이다(Park et al., 2016). 그러나 이 가설은 IPO 주식의 장기 저성과 현상을 규명하지 못한다는 한계가 있다(Park et al., 2016). 이에 최근에는 IPO 저가발행을 투자자의 행태론적 관점에서 설명하려는 연구들이 주목받고 있다(Chan, 2010; Derrien, 2005; Kim and Jung, 2010). Miller(1977)는 투자자의 낙관적 기대가 가격 형성에 영향을 미쳐 상장 초기에 높은 수익률이 발생하게 되고, 시간이 흘러 이 현상이 차츰 사라지면서 장기 저성과 현상을 초래한다고 설명하였다.

저가발행과 유사한 맥락에서, IPO 주식의 수익률을 결정하는 요인들에 대해 진행된 선행연구도 다수 존재한다. 수익률에 영향을 주는 정보는 크게 재무적 정보와 비재무적 정보로 나눌 수 있다. 기존에는 재무제표에서 쉽게 얻을 수 있으면서도 기업의 다양한 특성을 반영할 수 있는 재무적 정보를 많이 사용했다(Kim et al., 2013). 재무적 정보에는 기업의 과거 실적, 현재 재무상태, 미래 전망 등이 해당되고, 대표적으로 자산, 매출액, 당기순이익, 부채비율, 자기자본이익률(ROE), 총자산이익률(ROA) 등이 있다. 최근에는 재무적 정보와 비재무적 정보를 함께 이용하여 IPO 주식의 수익률 혹은 가격 등락을 예측하는 연구도 진행되고 있는데, 비재무적 정보에는 공모 규모, 상장일 유통가능물량 및 비율, 공모종목밀도, 할인율, 기관수요예측 경쟁률, 일반 청약 경쟁률, 보호예수 확약 비율 등이 있다(Shin et al., 2018). 선행연구들은 주로 가격 예측에 영향을 미치는 개별 요인에만 주목해 분석한 경향이 있다는 점을 지적하며 Shin et al.(2018)은 다양한 개별 요인들을 종합하고, 개인투자자들이 쉽게 수집할 수 있는 비재무적 자료를 이용하여 상장일의 IPO 주가 방향성을 예측한 바 있다.

〈Table 1〉 Studies on Predicting the IPO Price Using Machine Learning

Authors	Method	Main Findings
Muditomo and Broto(2021)	Decision Tree	The decision tree model was able to predict whether the Indonesian IPO performance during the COVID-19 pandemic would be underpriced or not.
Ly and Nguyen(2020)	Random Forest, Decision Tree, Naïve Bayes, & Logistic Regression	The logistic regression model trained on the sentiment of prospectuses could predict U.S. IPO stock price movements at an accuracy of 9.6% higher than chance.
Katsafados et al.(2020)	SVM, Logistic Regression, Random Forest, & Neural Network	Models that use both textual data from S-1 filings and financial information as input variables showed an accuracy of more than 70% in predicting U.S. IPO underpricing.
Baba and Sevil(2020)	Random Forest	The random forest model outperformed robust regression models in predicting Turkish IPO initial returns, and the most important predictors were IPO proceeds and IPO volume.
Cho et al.(2020)	Neural Network, Genetic Algorithm, Decision Tree, Logistic Regression, & Discriminant Analysis	The neural network model showed the best accuracy in predicting whether or not a Korean IPO stock would offer a specific amount of return.
Shin et al.(2018)	SVM, Logistic Regression, Discriminant Analysis, Decision Tree, Neural Network, & Case-Based Reasoning	Six different models were trained on non-financial input variables to predict short-term Korean IPO stock price movements, and the SVM model showed the highest prediction accuracy and sensitivity among other models.
Gandoman et al.(2017)	SVM & Particle Swarm Optimization Algorithm	The combination of SVM and PSO algorithm markedly increased the accuracy of Iranian IPO price prediction, compared to the model where SVM was used alone.
Basti et al.(2015)	Decision Tree & SVM	Market sentiment, sales, and total assets turnover ratios were the most crucial factors that affect the short-term performance of Turkish IPOs.
Esfahanipour et al.(2015)	Neural Network	A meaningful relationship between the probability of withdrawal and underpricing was found, so a radial basis neural network model was applied to predict the probability of withdrawal and underpricing of Iranian IPOs.
Luque et al.(2012)	Genetic Algorithm	The Genetic Algorithm model had two significant advantages over other machine learning models in the prediction of underpricing of U.S. IPOs: predictive performance and robustness to outliers.
Kaohua and Liwen(2012)	Rough Set & SVM	Rough set theory was used to analyze the influencing factors of Chinese IPO underpricing, and SVM was applied to verify the results by predicting IPO underpricing using different combinations of attributes.
Reber et al.(2005)	Neural Network & Genetic Algorithm	Multilayer perceptron models were developed to predict the mispricing of U.K. IPOs, and the genetic algorithm was used to select appropriate input variables.

2.2. 인공지능 기법을 이용한 IPO 주가 예측

최근 IPO 주식의 수익률 및 가격 등락 예측에 인공지능 기법을 활용하는 연구들이 주목받고 있으며, 인공지능 방법론은 기존의 전통적인 통계 방법론보다 높은 예측 정확도를 보이며 우수성을 인정받고 있다. 기계학습과 딥러닝 기법을 이용하여 IPO 주식의 수익률 혹은 가격 등락 여부를 예측한 연구들을 <Table 1>에 요약하였다.

미국, 영국, 인도네시아, 터키, 이란, 중국 등 다양한 국가에서 연구가 진행되고 있으며, 예측에 사용된 주요 알고리즘은 로지스틱 회귀분석과 같은 통계적 기법부터, 인공신경망, 랜덤 포레스트와 같은 기계학습 기법, 다층 퍼셉트론, 방사 신경망과 같은 딥러닝 기법 등이 존재한다.

2.2.1. **로지스틱 회귀분석**(Logistic Regression: LR)

로지스틱 회귀분석은 통계적 분석 방법론의 일종으로, 예측 및 분류에 사용되는 기법이다. 종속변수의 예측 값이 항상 0과 1 사이의 확률 값을 가지기 때문에 이진분류 문제에 특화된 방법론이다. 종속변수가 범주형 데이터라는 점에서 선형 회귀분석과 차이가 있으며, 구현이 쉽고 결과 해석이 용이하다는 장점이 있다. 예측 정확도도 상당히 높은 편이어서 결정 값이 이항 값일 때 예측모델로 사용된다(Lee et al., 2019).

2.2.2. **랜덤 포레스트**(Random Forest: RF)

랜덤 포레스트는 Breiman(2001)이 개발한 분류기법으로서, 전통적인 의사결정나무(Decision Tree) 기법을 다수의 나무로 확장시킨 앙상블(ensemble) 방법이다. 랜덤 포레스트의 핵심은 배깅과 입력변수들에 대한 무작위 추출이다(Kim and Ahn, 2016). 배깅은 여러 개의 학습용 데이터를 생성하여 다수의 모델을 만들었다가 최종적으로는 모델들을 하나의 분류기로 종합하는 방식을 의미한다. 랜덤 포레스트는 서로 상관성이 없는 수십 혹은 수백 개의 의사결정나무 모델을 결합함으로써 분류 정확도를 높인다(Kim and Won, 2019). 랜덤 포레스트의 성과에 영향을 미치는 변수에는 트리의 개수(T)와 최대 허용 깊이(D) 등이 있다. 랜덤 포레스트의 가장 큰 장점은 입력변수가 많을 때에도 높은 정확도로 예측이 가능하다는 것이다. 또한 다른 기법들에 비해 과적합문제가 적으며, 잡음이나 결측치 및 이상치의 영향도 적게 받는다. 데이터의 빈도가 불균형한 이항분류의 예측에 우수한 성능을 보인다고 보고되기도 하였다(Brown and Mues, 2012).

2.2.3. **서포트벡터머신**(Support Vector Machine: SVM)

서포트벡터머신은 서로 다른 두 집합이 있을 때 동질적인 데이터들이 놓이는 하나 혹은 복수의 초평면을 구하는 분류 또는 예측 알고리즘이다(Cortes and Vapnik, 1995). 저차원의 입력변수를 가진 비선형 문제도 고차원의 특징공간으로 매핑시켜 선형 문제로 재정의할 수 있도록 하는 특징을 가진다(Hong and Kim, 2010). 서포트벡터머신은 인공신경망에 필적할 만한 예측력을 가지면서도 예측모형의 구조를 이해하기 쉽다는 장점이 있다(Kim and Ahn, 2011). 또한 적은 수의 데이터를 이용하는 경우에도 과적합의 가능성이 적으며 우수한 예측성과를 보인다(Ahn and Kim, 2014). 서포트벡터머신의 또다른 장점은 다른 기계학습 기법에 비해 조정해야 할 파라미터의 수가 적다는 점이다. 최근에는 전통적인 방식의 서포트벡터머신 대신 유전자 알고리즘을 파라미터 최적화에 사용한 연구도 다수 존재한다(Ahn, 2014; Manurung et al., 2017). 본 연구에서도 유전자 알고리즘을 이용하여 서포트벡터머신 모형의 변수 및 파라미터 조합을 설정한다.

2.2.4. **인공신경망**

인공신경망은 McCulloch and Pitts가 1943년에

인간 뇌의 신경조직에 영감을 얻어 개발한 인공지능 방법론이다(Shin et al., 2018). 인공신경망의 구조는 입력층, 은닉층, 출력층으로 이루어져 있으며, 각 층에는 하나 혹은 복수의 노드(node)가 존재한다. 노드의 입력값에 가중치를 곱하고 편차를 더한 후에 활성화 함수를 적용한 값이 출력되어 다음 층으로 전달되며(Kim et al., 2021), 학습은 최적의 가중치를 찾고 출력값과 목표값의 오차가 최소화될 때까지 반복된다. 인공신경망은 복잡하고 방대한 비선형 데이터를 분석하는 데에 적합한 기법이다. 변수들 간의 선형적 관계나 확률분포를 가정하지 않는 특징이 있어 전통적인 통계 기법에 비해 적용 범위가 광범위하기도 하다(Cho et al., 2020; Lee et al., 2019). 이처럼 인공신경망은 강력한 예측력과 범용성을 가지고 있어 다양한 분야의 연구에서 이용되고 있다(Shin et al., 2018). 인공신경망의 성능을 결정짓는 요인에는 은닉층의 개수, 은닉층의 노드 수, 학습 반복 횟수, 학습률, 모멘텀 등이 있다(Park et al., 2006). 최근에는 인공신경망에 유전자 알고리즘을 결합하여 파라미터 최적화를 실시한 연구들이 다수 존재한다(Seo, 2015). 본 연구에서도 유전자 알고리즘을 통해 인공신경망 모형의 변수 선정 및 파라미터 조합 최적화를 진행한다.

2.3. 텍스트 분석을 이용한 IPO 주가 예측

텍스트 분석은 기존에 측정하기 어려웠던 요인들을 정량적인 값으로 변환하여 모형에서 분석할 수 있게 해준다는 강점이 있어 경제학 및 경영학 분야에서 다방면으로 이용되고 있다(Kim and Joh, 2019). 일례로 Jung and Park(2019)는 증권형 크라우드펀딩 투자설명서의 텍스트 분석을 진행한 결과, 투자설명서 어조가 투자자들의 투자 의사결정에 영향을 미친다고 보고하였다. 전자공시제도의 보급으로 기업 공시자료에 일반투자자들의 접근이 용이해짐에 따라, 공시자료의 텍스트를 분석하는 연구도 주목받기 시작하였다(Kim et al., 2015). 특히 증권신고서의 텍스트 정보는 정보 비대칭 문제를 완화할 수 있기 때문에 IPO 저가발행 현상 및 수익률에 관한 연구에서 매우 중요한 수단이 될 수 있다(Kim and Joh, 2019). 이러한 텍스트 정보를 이용한 선행연구들은 대부분 상장기업의 투자설명서에 나타난 텍스트의 어조를 활용하였으며, 텍스트 어조 변수를 재무 변수와 함께 사용할 경우 예측력이 향상된다고 주장한 연구들도 있다(Katsafados et al., 2020; Tao et al., 2018).

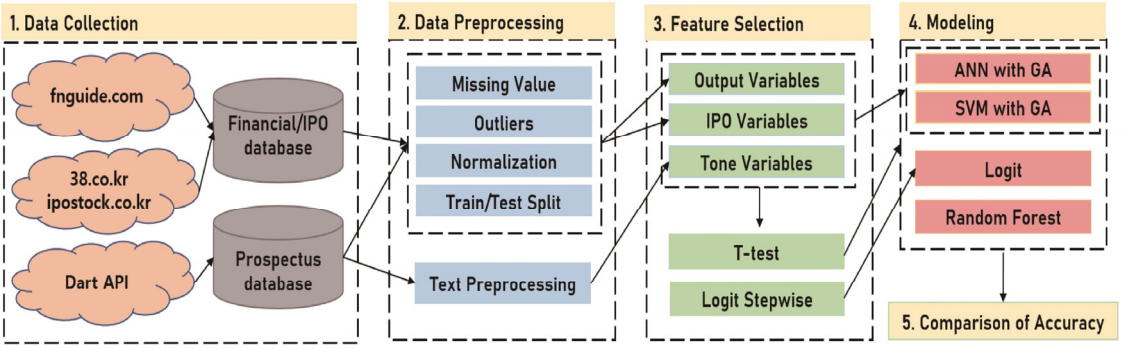
어조 분석 방법에는 크게 사전기반 방식(lexicon-based approach)과 기계학습을 이용하는 방식(machine learning-based approach)이 존재한다. 사전기반 방식은 긍정어, 부정어의 종류와 긍정, 부정의 정도를 파악할 수 있는 감성사전(sentiment dictionary)을 미리 구축해두고 각 단어가 텍스트에 등장하는 빈도수를 측정해 어조를 파악한다(Tetlock et al., 2008). 이러한 방식은 분석에 적용하기 용이하다는 특징을 가지지만, 감성사전에 대한 의존도가 높아 사전 구축에 상당한 노력이 필요하며 특정 도메인에 의존적인 사전을 사용할 경우 신뢰도가 떨어질 수 있다(Seo and Kim, 2016). Loughran and McDonald(2011)는 금융 텍스트 분석에 특성화된 영어 감성 단어 목록을 구축한 바 있다. 이후 많은 연구자들이 이들의 단어 목록을 이용하여 텍스트 분석을 진행하였으며(Garcia, 2013; Katsafados et al., 2020; Ly and Nguyen, 2020), 단어 목록을 사용하여 분석한 증권신고서의 어조와 IPO 수익률 간에는

유의미한 관계가 존재한다고 많은 연구에서 실증 연구를 통해 주장하고 있다(Fuksa, 2013; Hanley and Hoberg, 2010; Jegadeesh and Wu, 2013; Ly and Nguyen, 2020). Loughran and McDonald (2013)는 미국 증권신고서에 나타난 어조가 부정적일수록 상장 직후 수익률과 공모가격 변화율이 높아질 수 있다고 주장하였다. 또한 중국 증권보고서 및 IPO와 관련해서도 어조가 부정적이고 불확실성을 나타낼수록 IPO의 초기수익률이 높다고 주장한 연구가 존재한다(Yan et al., 2019). Tao et al. (2018)은 투자설명서의 경영진단의견서 부분이 기업공개활동에 가장 큰 영향력을 미치고, 부정적인 어조가 높은 수익률을 야기할 수 있다고 주장하였다.

사전기반 방식의 어조분석이 사전에 정교하게 구축된 일종의 규칙 기반이라면, 기계학습을 이용하는 방식은 단어 목록이 없는 상황에서도 사용할 수 있는 방법이다. 학습용 데이터로 텍스트 분류 모형을 구축해 놓으면, 새로 수집한 텍스트 데이터의 분류는 모형이 자동적으로 실행해준다. 텍스트 분류 모형으로 가장 널리 사용되는 기계학습 기법에는 서포트벡터머신, 나이브 베이즈, 의사결정나무 등이 있다(Kolchyna et al., 2015). 기계학습을 이용하는 방식은 별도의 감성사전을 구축할 필요가 없고, 사전기반 방식보다 연구자의 자의성이 덜 개입된다는 장점이 있다(Kim and Joh, 2019). 또한, 감성사전을 이용할 때보다 기계학습을 이용했을 때 감성분석의 성능이 뛰어날 수 있다는 것이 여러 연구를 통해 주장되고 있다(Kolchyna et al., 2015; Nguyen et al., 2018). Li(2010)는 기존 연구들이 사용해온 감성사전의 오류를 지적하고, 기업보고서의 어조와 내용이 기업 성과와 어떤 관련이 있는지 나이즈 베이즈(Naïve Bayes) 기법을 이용하여 분석한 바 있다. 분석 대상은 미국의 분기보고서(10-Q)와 사업보고서(10-K)의 MD&A에 포함된 미래 관련 공시 텍스트였다. Buehlmaier and Whited(2018) 또한 같은 기법으로 기업의 연차보고서 속 어조와 주가 수익률 간의 관계를 분석하였다.

하지만 아직까지 한국어 텍스트 분석을 이용한 IPO 관련 연구는 많지 않다. 큰 이유 중 하나는 국내 금융시장과 한국어의 특수성을 고려하여 설계된 단어 목록이 존재하지 않기 때문이다(Kim and Joh, 2019). Joh and Kim(2018)은 이런 점을 지적하며, 기계학습을 사용해 국내 증권신고서에 나타난 텍스트의 어조를 분석한 바 있다. 이들은 투자위험요소란의 어조와 길이가 IPO 주식의 장기 성과를 예측하는 데에 중요한 수단이 된다고 주장하였다. 후속 연구에서는 투자위험요소란의 어조가 부정적인 회사일수록 공모가격 변화율이 커지는 경향을 발견하기도 하였다(Kim and Joh, 2019). 이러한 연구결과는 앞서 언급한 Loughran and McDonald(2013)의 연구결과와 동일하다.

하지만, Joh and Kim(2018)의 연구는 Loughran and McDonald(2013)를 비롯한 선행연구들과 달리 증권신고서 어조가 상장 직후의 수익률에 미치는 영향을 찾아볼 수 없다고 보고하였다. 또한 Joh and Kim(2018)은 단순히 기업의 재무적인 요소와 주식시장 관련 요인을 변수로 사용하였으며, 예측에 회귀분석만을 이용하였다. 본 연구는 선행연구들을 발전시켜 IPO와 관련된 다양한 변수들을 활용하고, 증권신고서 텍스트 분석을 통해 생성한 어조 변수들을 추가적으로 사용하여 IPO 주식 상장 이후 가격의 등락 여부를 로지스틱 회귀분석, 랜덤 포레스트, 서포트벡터머신, 인공신경망 기법으로 예측해보고자 한다.



〈Figure 1〉 Research Framework

〈Table 2〉 Distribution of Sample IPOs by Year

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Observations	34	70	70	23	37	44	72	68	60	75	70	68	691

3. 연구 프레임워크

3.1. 데이터 수집 및 전처리

본 연구에서 분석에 사용된 데이터는 2009년 6월부터 2020년 12월 사이의 기간 동안 코스피(KOSPI) 또는 코스닥(KOSDAQ)에 신규 상장된 IPO 종목이다. 기간을 2009년 6월 이후로 설정한 이유는 이 시점부터 증권보고서의 보고 형식이 변경되었기 때문이다. 최초 수집한 표본은 총 1,017개였으나, 이 중 기업인수목적회사, 증권신고서를 확인할 수 없는 기업, 변수의 누락이 있는 경우를 표본에서 제외하였다. 분석에 적합하지 않은 표본을 제외하고 최종적으로 분석에 사용된 표본은 691개였다. 표본의 연도별 분포는 <Table 2>와 같다.

신규상장 기업 명단과 이들의 재무 및 주가 자료는 FnGuide에서 수집하였다. 각 기업의 증권신고서는 '전자공시시스템(DART)'에서 웹 크롤링을 통해 수집하였으며, 증권신고서에서는 투자위험요소란의 텍스트와 IPO 관련 정보를 수집하였다. 누락된 정보는 '아이피오스탁(http://www.ipostock.co.kr/)'과 '38커뮤니케이션(http://www.38.co.kr/)' 웹사이트에서 추가로 수집하였으며, 변수 간에 차이나는 스케일(scale)을 조정해주기 위해 최소-최대 정규화(min-max normalization)를 실행하였다.

3.2. 변수 선정

3.2.1. 출력변수

본 연구는 상장 이후 IPO 주가의 등락 여부를 예측하고자 한다. 따라서 출력변수를 '상장 당일 시초가 대비 상장 이후 5거래일 종가의 상승 여부'로 설정하였다. 결과값의 부호에 따라 주가가

상승하거나 동일한 259개(약 37%)의 주식에는 1을, 하락한 432개(약 63%)의 주식에는 0을 부여하였다. 수요예측이나 청약에서 배정받지 못한 투자자들은 상장일 시초가에 매수하게 되는 경우가 많다(Shin et al., 2018). 특히 최근에는 일반 청약 경쟁률이 높아지면서 상장일 이전에 충분한 물량을 배정받지 못하는 개인투자자들이 많아지고 있는 추세이다. 따라서 본 연구는 개인투자자들의 공모주 투자에 실질적인 도움이 될 수 있도록 공모가가 아닌 상장일 시초가를 기준으로 하여 가격 등락 여부를 예측하였다. 또한 국내 주식시장의 가격제한폭 제도를 고려하여, 상장 당일이 아닌 상장일 기준 5거래일 이후 종가의 상승 여부를 알아보는 것으로 설정하였다.

$$\frac{\text{상장}+5\text{거래일 종가}-\text{시초가}}{\text{시초가}} \geq 0\text{이면}$$

$$Y = 1,\ \frac{\text{상장}+5\text{거래일 종가}-\text{시초가}}{\text{시초가}} < 0\text{이면 } Y = 0$$

3.2.2. 어조변수

증권신고서 내 투자위험요소란에 한정하여 긍정적 어조, 중립적 어조, 부정적 어조의 비율을 측정하였다. 투자위험요소란은 “투자자가 특정 유가증권을 매입함에 따라 노출되는 가격하락 또는 원리금 회수위험 등을 초래하는 제반 요인”에 대해 기재한 부분이다(Kim, 2008). 증권신고서 내 다른 부분들은 현황 중심으로 기술되는 반면, 투자위험요소란은 투자자의 관점에서 서술된 부분이다(Kim, 2008). 관련 선행 연구에서도 증권신고서의 텍스트 분석에 투자위험요소란이 사용된 바(Kim and Joh, 2019), 투자위험요소란은 본 연구의 목적인 증권신고서 어조 분석에 가장 적합하다고 판단하였다. 어조 변수를 생성하는 과정을 살펴보면, 먼저 DART API를 통해 수집한 증권신고서에 존재하는 HTML태그를 제거하고, 텍스트 데이터에 대해 전처리 작업을 실시하였다. Python의 한글 자연어 처리 패키지인 KoNLPy에 존재하는 형태소 분석기(Kkma: 꼬꼬마 한글 형태소 분석기)를 이용하여 텍스트를 문

〈Table 3〉 Examples of Positive/Neutral/Negative Sentences from Prospectuses

긍정적 어조의 예시
- 2015 년 말 제조사업부분의 분할에 따라 선수금 등이 감소하여 부채비율이 개선되었습니다. - 그 결과 2014 년 상반기 해외 매출 비중이 8,765 에 달하고 있으며 향후에도 해외 매출 비중은 지속적으로 증가할 것으로 예상됩니다.
중립적 어조의 예시
- 당사는 언어 데이터를 생성 및 판매하는 사업을 주요 사업으로 영위하고 있으며 이러한 사업모델을 가지고 영업을 하는 경쟁사는 호주의 상장 사인 Appen 등이 있는 것으로 판단됩니다. - 자발적 보호 예수 자 주식 1,155,172 주 또한 최대주주 등과 동일하게 1 년 간 보호 예수 되며 상장 일로부터 6 개월이 경과한 경우 매 1 월마다 최초 보유주식 등의 100 분의 5 에 상당하는 부분까지 매각할 수 있습니다.
부정적 어조의 예시
- 이러한 조치는 무거운 벌금 처벌 금지 또는 사업활동의 제한을 포함하며 당사의 사업 운영결과 또는 재무상태에 상당한 악영향이 될 수 있습니다. - 저임금을 바탕으로 높은 성장세를 구가하고 있는 중국 등 신흥국 기업의 시장 진입으로 인한 경쟁구도 변화도 주의해 보아야 할 사항인 점 투자자들께서는 유의하시기 바랍니다.

<Table 4> Sentence Classification Results

Tone	Positive	Neutral	Negative	Total
Number of Sentences	3,990 (29.45%)	9,697 (49.98%)	5,713 (20.57%)	19,400

장별로 분절하였다. 형태소 분석기의 불완전함으로 인해 잘못 분리된 문장들을 제거하기 위하여 글자 수가 25자 이하인 문장들은 삭제하였다. 그 결과, 총 194,479개의 문장을 얻을 수 있었다. 모든 문장에서 특수 문자, 문장 부호, 괄호 속 단어 등은 제거하였다.

이어 어조 분석을 실시하였는데, 국내에는 금융 텍스트 분석에 적합한 감성 사전이 존재하지 않으므로 Joh and Kim(2018), Katsafados et al. (2020), Kim and Joh(2019)의 논문을 참조해 기계학습 기반의 어조 분석을 실시하였다. 따라서 먼저 학습용 데이터를 제대로 구축하는 것이 무엇보다 중요한 작업이었다. 앞서 전처리를 완료한 19만여 개의 문장 중 기업마다 10%씩 임의 추출한 문장 19,400개를 임의 추출하였고, 이어 추출된 문장들을 하나씩 읽으면서 긍정적 어조(1), 중립적 어조(0), 부정적이거나 불확실성을 띄는 어조(-1)로 직접 분류하였다. 저자들이 함께 합의를 거쳐 어조 분류를 실시하였으며, 실제 증권신고서 투자위험요소란에서 찾아볼 수 있는 문장들의 예시는 다음과 같다.

분류 결과는 <Table 4>와 같다. 중립적 어조가 약 49.98%로 가장 높은 비율을 차지하였는데, 이는 정보 전달을 위한 문장들이 많기 때문이다. 긍정적 어조에 비해서는 부정적 혹은 불확실성을 띄는 어조의 비율이 더 높았다.

이어 단어 벡터 도출 과정을 진행했는데, 문장들로 이루어진 텍스트 데이터에 unigram과 bigram을 적용시켜 1어절과 2어절 단위의 데이터로 변환하였다(Katsafados et al., 2020). 이 과정에서 한국어 형태소 분석기(Open Korean Text: OKT)를 결합하여 문장을 형태소 단위로 분리하였다. 동시에 품사 태깅을 통해 동사, 명사, 부사, 형용사만 학습용 데이터로 이용될 수 있게 하였다. 또한, 형태소들의 표현 방식을 일치시키기 위해 정규화와 어간 추출을 동시에 진행하였다. 분석에 불필요한 단어, 즉 불용어(예: ‘년도’, ‘분기’, ‘당사’, ‘것이’, ‘있는’, ‘하다’ 등의 82개 단어)를 추가로 제거하였다.

불용어 처리 후에는 기계학습 모형이 학습할 수 있는 형태로 텍스트 데이터를 변환시켜주었다. 수치화된 단어 벡터를 생성하기 위해서는 Bag-of-Words (BOW) 모델을 이용하였다(Lee and Kim, 2019). Bag-of-Words는 단어들의 순서를 고려하지 않고 빈도수에만 기반하여 텍스트 데이터를 수치화하는 방법이다. 본 연구에서는 Bag-of-Words를 만들기 위해 TF-IDF (Term Frequency - Inverse Document Frequency)를 사용하였다. TF-IDF방법은 단어의 빈도(Term Frequency: TF)와 역 문서 빈도(Inverse Document Frequency: IDF)를 곱한 값을 이용하여 각 단어의 중요도를 계산해준다. 특정 문서에서만 빈도수가 높은 단어들은 중요도가 높지만, 모든 문서에서 빈도수가 높은 단어들은 중요도가 낮다고 간주하는 원리이다(Katsafados et al., 2020). 예를 들어, 증권신고서 상에서 반복적으로 사용되지만 크게 중요하지 않은 단어 혹은 조사 등은 낮은 가중치를 부여받게 된다(Lee and Kim, 2019). TF-IDF는 감

〈Figure 2〉 Word Cloud of Positive Sentences

〈Figure 3〉 Word Cloud of Neutral Sentences

〈Figure 4〉 Word Cloud of Negative Sentences

성 분석(sentiment analysis)을 진행한 많은 연구에서 높은 정확도를 보이는 기법이다(Imamah and Rachman, 2020; Katsafados et al., 2020; Prastyo et al., 2020). 이에 본 연구는 증권신고서

〈Table 5〉 5-Fold Cross-Validation Results

	Set 1	Set 2	Set 3	Set 4	Set 5	Mean
Accuracy (%)	73.14	75.70	73.84	70.36	71.29	72.87

의 어조를 분석하는 데에 있어서 TF-IDF를 이용하는 것이 효과적일 것이라고 판단하였다.

한편, 텍스트 데이터가 어조별로 제대로 분류되었는지 확인해보기 위하여 워드 클라우드(word cloud)를 통해 빈도분석의 결과를 확인하였다(<Figure 2>, <Figure 3>, <Figure 4>).

마지막으로 서포트벡터머신모형을 통해 학습 및 분류 작업을 수행하였다. 기계학습 모형이 각 기업의 투자위험요소란 텍스트에 대해 긍정적 어조, 중립적 어조, 부정적 어조의 비율을 각각 계산해 도출하도록 설계하였는데, 과적합을 방지하기 위하여 빈도수가 상위 10,000위 내에 있는 단어들만 학습용 데이터로 삼았다(Joh and Kim, 2018; Mai et al., 2019). 실험에서 사용된 입력변수의 크기는 최대 feature 수 10,000개, unigram을 통한 1어절 단위 데이터 3,419개, bigram을 통한 2어절 단위 데이터 6,581개였다. 텍스트 분류 모형으로는 선형(linear) 서포트벡터머신을 이용하였다. 서포트벡터머신은 적은 수의 데이터로 신속하게 분류를 실행한다는 장점이 있으며(Ahn and Kim, 2014), 다수의 텍스트 분석 연구에서 TF-IDF 방법과 함께 사용되었다(Dadgar et al., 2016; Islam et al., 2017). 또한, Martens and Provost(2014)의 텍스트 분류 연구에서 선형 커널함수를 이용한 서포트벡터머신 모형이 RBF (Radial Basis Function) 커널함수를 이용한 서포트벡터머신 모형보다 더 높은 예측 정확도를 보인 바 있다. 그밖에도 다수의 연구에서 서포트벡터머신이 텍스트 분류 예측에 사용되었다(Fang et al., 2014; Moraes et al., 2013; Sun et al., 2009).

모형의 성능과 강건성을 검증하기 위하여 5분할 교차 검증(5-fold cross-validation)을 실시하였다. 5분할 교차 검증은 전체 학습용 데이터를 5개로 분할하고 각각의 부분 데이터에 대해 예측을 실시함으로써 예측의 정확성과 객관성을 높이는 방법이다. 교차 검증을 실시한 결과, <Table 5>에서 확인할 수 있듯이 평균 예측 정확도는 약 72.87%였다. 이는 나이브 베이즈 분류 기법을 이용해 증권신고서 어조를 측정한 Kim and Joh(2019)의 연구에서 보고된 정확도인 75.9%와 유사하다. SVM 모델의 성능을 확인한 뒤에는 학습용 데이터를 제외한 나머지 약 175,000개의 문장들을 검증용 데이터로 사용하여 텍스트 분류를 진행하였다.

3.2.3. IPO 관련 변수

IPO 관련 변수는 기업 관련, 공모 관련, 시장 관련 변수로 나누어 살펴볼 수 있다(Baba and Sevil, 2020; Muditomo and Broto, 2021; Perera and Kulendran, 2016). <Table 6>는 변수 선정을 거치기 이전의 전체 입력변수 목록이며, <Table 7>는 변수들의 기술 통계량을 요약하고 있다. 우선 기업 관련 변수는 주로 재무제표 상에서 얻을 수 있는 정보로, 기업의 역량과 성과를 판단할 수 있는 가장 기본적인 지표이다. 특히, 기업연령은 여러 연구에서 IPO 성과에 영향을 미치는 주요 요인 중 하나라고 주장되고 있다(Baba and

〈Table 6〉 Variable Description

Characteristics		Name	Description
IPO-Related Variables (Quantitative Data)	Firm-specific	offer_date	Offering date (YYYYMMDD)
		firm_age	Number of years elapsed since IPO
		asset	Natural logarithm of total assets
		liability	Natural logarithm of total liabilities
		equity	Natural logarithm of total equity
		sales	Natural logarithm of sales
		operating_income	Natural logarithm of operating income
		net_income	Natural logarithm of net income
		CFO	Natural logarithm of cash flow from operating activities
		debt_ratio	Debt ratio = total liabilities / total equity
		ROA	Return on assets = (net income / assets) * 100
		ROE	Return on equity = (net income / equity) * 100
		EBITDA	Natural logarithm of earnings before interest, taxes, depreciation, and amortization
		EAR	Equity-to-asset ratio = total equity / total assets
	Issue-specific	offer_price	Natural logarithm of public offering price
		tot_offer_amt	(Natural logarithm of) Total offering amount = number of shares * offering price
		inst_alloc	IPO allotment status of institutional investors
		inst_compete	IPO subscription rate of institutional investors
		private_compete	IPO subscription rate of private investors
		lock_up	Lock-up ratio = number of shares locked up / total number of shares
	Market-specific	is_kospi	KOSPI = 1, KOSDAQ = 0
		kosdaq_15_return	Returns of the KOSDAQ index for 15 days before the IPO date
Tone Variables (Qualitative Data)		%Positive	Percentage of positive sentences in the “Risk Factors” section of each prospectus
		%Negative	Percentage of negative sentences in the “Risk Factors” section of each prospectus
		%Neutral	Percentage of neutral sentences in the “Risk Factors” section of each prospectus

Sevil, 2020; Basti et al., 2015; Chun et al., 2013; Park et al., 2016).

한편, 공모 관련 변수 중 하나인 공모금액은 많은 IPO 연구에서 확정공모가와 함께 중요하게 인식되고 있다(Chen and Cheng, 2012; Han, 2015; Quintana et al., 2018). 본 연구에서는 공모금액이 작을수록 물량 확보가 어려워 경쟁률이 높아지고, 이로 인해서 높은 수익률로 이어지게 될 것이라고 예상하였다(Baek and Jeong, 2018). 또한, 기관투자자 배정비율, 기관경쟁률, 개인경쟁률

〈Table 7〉 Descriptive Statistics of Variables

Var. No.	Name	Mean	SD	Min	Max
V1	firm_age	17.13	9.88	1.00	64.75
V2	asset	24.80	1.36	21.59	32.52
V3	liability	23.93	1.62	19.80	32.43
V4	equity	25.57	1.14	0.00	30.85
V5	sales	24.35	3.00	0.00	32.12
V6	operating_income	26.08	1.01	0.00	29.18
V7	net_income	25.23	1.01	0.00	28.32
V8	CFO	27.96	1.07	0.00	29.31
V9	debt_ratio	1.41	2.57	-11.68	31.86
V10	ROA	0.05	0.37	-3.84	0.74
V11	ROE	0.06	2.46	-59.04	5.49
V12	EBITDA	25.89	1.02	0.00	29.77
V13	EAR	52.29	21.81	0.00	98.04
V14	offer_price	9.25	1.05	0.00	12.36
V15	tot_offer_amt	23.98	1.12	21.10	29.22
V16	inst_compete	5.02	1.51	0.00	7.30
V17	inst_alloc	60.33	25.94	0.00	80.00
V18	private_compete	5.26	1.99	0.00	11.29
V19	lock_up	8.86	15.63	0.00	86.16
V20	is_kospi	0.16	0.37	0.00	1.00
V21	kosdaq_15_return	0.00	0.46	-2.35	1.24
V22	%Positive	0.20	0.06	0.03	0.41
V23	%Negative	0.28	0.07	0.04	0.52
V24	%Neutral	0.52	0.08	0.33	0.93
Target Variable		0.37	0.48	0.00	1.00

은 모두 투자자들의 공모주에 대한 관심도를 나타내는 변수들이다(Baek and Jeong, 2018). 이러한 변수들의 값이 높을수록 상장 초기 수익률도 높아질 것이라고 예상하였다(Min, 2017). 마지막으로 의무보유확약 비율 또한 기업에 대한 매력도를 의미하는 지표이기에 공모 후 가격에 영향을 미치는 요소라고 판단하였다(Baek and Jeong, 2018). 의무보유확약을 통해 기관투자자들은 매력적인 기업에 대해 더 많은 물량을 확보할 수 있기 때문이다.

〈Table 8〉 IPO Variable Selection Results

Variable	t-test	LR	RF	GA+SVM	GA+ANN
offer_date	O	X	O	X	X
firm_age	O	O	O	O	X
asset	O	X	O	O	O
liability	O	X	O	X	O
equity	O	O	O	X	O
sales	O	O	O	X	X
operating_income	O	O	O	X	O
net_income	O	X	O	O	O
CFO	O	X	O	O	X
debt_ratio	O	O	O	O	O
ROA	O	O	O	O	O
ROE	O	X	O	X	X
EBITDA	O	X	O	O	O
EAR	O	O	O	O	O
offer_price	O	X	O	O	X
tot_offer_amt	O	O	O	O	O
inst_alloc	O	X	O	X	O
inst_compete	O	O	O	O	X
private_compete	O	X	O	O	X
lock_up	O	X	O	X	X
is_kospi	O	X	O	O	O
kosdaq_15_return	O	O	O	O	X

마지막으로, 시장 관련 변수 중에서 코스피시장 여부를 하나의 중요한 변수로 고려하였다(Kim and Joh, 2019). 코스닥 상장기업은 코스피에 비해 소규모기업이 많아서 시가총액과 공모금액이 작은 경우가 많다(Baek and Jeong, 2018). 이는 높은 수익률로 이어질 가능성이 충분하기 때문에 상장 당시의 주식시장 상황 또한 IPO 주식 가격에 영향을 미칠 수 있다고 판단하였다(Chun et al., 2013; Kim and Joh, 2019; Tao et al., 2018). 연구의 표본 중 약 84%가 코스닥 시장 상장기업이라는 점을 고려해 상장일 전 코스닥 수익률을 하나의 변수로 고려하였다.

IPO 관련 변수는 독립표본 t-test와 단계적 로지스틱 회귀분석, 유전자 알고리즘을 이용하여 추출된 변수들을 예측 모델에 활용하였다. 먼저 각 입력변수에 대해 출력변수(IPO 주식 가격의 등락 여부)와 독립표본 t-test를 실시하여 통계적 유의성을 검정하였고, 검정 결과 모든 변수들이

유의수준 5%에서 유의하다는 것을 확인하였다.

다음으로 IPO 주가 등락 여부 예측에 사용할 로지스틱 회귀분석 모형에 적합한 변수를 선정하기 위해 단계적 로지스틱 회귀분석을 수행하였다. 단계적 로지스틱 회귀분석은 독립변수를 선정하기 위해 주로 사용되는 방법으로, 최적의 로지스틱 회귀모형이 구성되었을 때의 변수를 추출한다. 후방 소거법(backward elimination)을 실시한 결과, 유의미하다고 판단된 변수는 '기업연령', '자본총계', '매출', '영업이익', '부채비율', 'ROA', 'EAR', '총공모금액', '기관수요예측경쟁률', '상장 전 15일간 코스닥 수익률'로 총 10개였다.

또한 유전자 알고리즘을 통해 서포트벡터머신 모형과 인공신경망 모형에 사용될 변수들을 선정하였다. 유전자 알고리즘은 확률적 탐색이나 학습 및 최적화를 위한 기법으로, 다윈의 적자생존과 멘델의 유전 법칙을 바탕으로 하고 있다(Hong and Shin, 2003). 기존 선형 모델로 해결 불가능한 복잡한 문제들에 적합한 인공지능 방법론으로써 다수의 연구에서 최적의 입력변수를 선정하는 데에 사용되었다(Ahn, 2014; Cho et al., 2020; Hong and Shin, 2003). 본 연구에서도 모형의 분류 정확도를 목적함수로 하여, 유전자 알고리즘의 적응도(fitness value)가 가장 우수한 최적의 변수군을 선정하였다. 본 연구에서는 population size 30, generation count 30, mutation rate 0.2, crossover rate 0.5와 0.8로 설정하였다. 서포트벡터머신 모형에서 사용된 변수들은 '기업연령', '자산총계', '당기순이익', 'CFO', '부채비율', 'ROA', 'EBITDA', 'EAR', '공모금액', '총공모금액', '기관수요예측경쟁률', '개인경쟁률', '코스피여부', '상장 전 15일간 코스닥 수익률'로 총 14개였다. 한편, 인공신경망 모형에서 사용된 변수들은 '자산총계', '부채총계', '자본총계', '영업이익', '당기순이익', '부채비율', 'ROA', 'EBITDA', 'EAR', '총공모금액', '기관투자자배정비율', '코스피여부'로 총 12개였다. 기계학습 모형에서 각 변수의 사용 여부를 <Table 8>에 정리하였다.

3.3. 기계학습을 이용한 IPO 주가 등락 예측

최종적으로 선정된 IPO 관련 변수들과 증권신고서 어조 변수들을 이용하여 상장 이후 IPO 주식 가격 등락 여부를 예측하는 기계학습 모형을 구축하였다. 예측에 사용된 기계학습 기법은 로지스틱 회귀분석, 랜덤 포레스트, 서포트벡터머신, 인공신경망이었다. 랜덤 포레스트 모형의 최적 파라미터 선정에는 격자검색기법을 사용하였으며, 최대 나무 깊이는 10, 최소 분할 샘플 데이터 수는 8, 결정 트리의 개수는 100으로 설정하였다. 서포트벡터머신 모형과 인공신경망 모형의 파라미터 조합 최적화에는 유전자 알고리즘을 사용하였다. 각 모형의 예측 성능은 분류 정확도를 통해 비교해보았다. 추가적으로, 증권신고서 텍스트에 나타난 어조가 IPO 주가 등락 여부 예측력을 향상시킬 수 있는지 알아보기 위해 아래와 같이 두가지 Group으로 구별하여 기계학습 모형들의 성능을 확인하였다.

- Group 1: IPO 관련 변수
- Group 2: IPO 관련 변수 & 증권신고서 어조 변수

본 연구에서 학습용 데이터는 상장일자가 2019년 이전에 해당하는 표본으로 삼았고, 2019년 이후에 상장된 기업 표본은 검증용 데이터로 사용하였다. 총 691개의 표본 중 학습용 데이터의 개수는 553개로 전체의 약 80%였으며, 검증용 데이터의 개수는 138개로 전체의 약 20%에

〈Table 9〉 The Distribution of Stock Price Movement After IPO in Training and Test Set

	Training Set	Test Set	Total
Up (1)	200	59	259
Down (0)	353	79	432

〈Table 10〉 Comparison of Performance Results

Group	LR (%)	RF (%)	SVM (%)	ANN (%)
Group 1: Only IPO-Related Variables	Test: 57.25 (Train: 65.45)	Test: 59.42 (Train: 64.73)	Test: 52.90 (Train: 79.86)	Test: 57.25 (Train: 64.48)
Group 2: IPO-Related Variables + Tone Variables	Test: 57.25 (Train: 65.09)	Test: 60.87 (Train: 64.74)	Test: 57.97 (Train: 81.22)	Test: 61.59 (Train: 66.97)

해당하였다. <Table 9>는 학습용 데이터와 검증용 데이터에서 출력변수(상장 이후 5거래일의 종가)가 각각 0과 1인 표본의 개수를 보여준다.

4. 분석결과

Group 1에서는 IPO 관련 변수만 사용한 반면, Group 2에서는 Group 1에서 사용한 IPO 관련 변수와 더불어 3개의 증권신고서 어조 변수도 사용하였다. 다양한 기계학습 기법의 예측 정확도를 비교한 결과는 <Table 10>과 같다. Group 1의 예측 정확도는 랜덤 포레스트 모형에서 59.42%로 가장 높았다. 이어서 인공신경망 모형과 로지스틱 회귀분석 모형이 57.25%, 서포트벡터머신 모형은 52.90%의 예측 정확도를 보였다. Group 2에서는 로지스틱 회귀분석 모형을 제외한 모든 모형에서 예측 정확도가 Group 1에서보다 향상되었다. 대표적으로 서포트벡터머신 모형은 Group 1에 비해 예측 정확도가 5.07% 포인트 향상되어 Group 2에서 57.97%라는 정확도를 보였다. Group 2에서 랜덤 포레스트 모형과 인공신경망 모형의 결과는 각각 60.87%와 61.59%로, Group 1보다 각각 1.45% 포인트, 4.34% 포인트가 향상된 값이었다. 반면, Group 2에서 로지스틱 회귀분석 모형은 예측 정확도가 57.25%로, Group 1에서의 결과와 동일했다.

한편, 증권신고서 어조 변수 사용에 따른 모형 성능 차이가 유의한지 검정하기 위해 맥니마 검정(McNemar Test)을 실시하였다. <Table 11>은 맥니마 검정에 의한 유의확률 결과를 정리한 표이다. 랜덤 포레스트, 서포트벡터머신, 인공신경망 기법 모두에서 유의확률이 0.01보다 작게 나왔다는 결과를 통해, 증권신고서 어조를 추가적으로 사용하면 분류 성능이 향상된다는 것을 확인하였다.

〈Table 11〉 McNemar Test Results

RF	ANN	SVM
21.441*** (0.000)	8.1*** (0.004)	7.111*** (0.008)

Values in table mean test statistics, and values in parentheses mean p-values.
*Significance level 10%: *, 5%: **, 1%: ****

5. 결론

본 연구는 다양한 IPO 관련 변수와 증권신고서 어조 변수를 이용하여 IPO 주식의 상장 이후 가격 등락 여부를 예측하였다. 예측 모형을 개발하기 위해서 로지스틱 회귀분석, 랜덤 포레스트, 서포트벡터머신, 인공신경망 기법과 TF-IDF 기반의 텍스트 분석을 이용하였다. 분석 결과, IPO 변수만을 사용한 모형보다 증권신고서 어조 변수와 IPO 변수를 함께 사용한 모형에서 예측 정확도가 더 높았는데, 이를 통해 증권신고서 어조가 IPO 주식의 상장 가격 등락 여부 예측에 유의미한 영향을 끼친다는 것을 확인할 수 있었다. 본 연구가 제안한 모형 중 가장 좋은 성능을 보인 인공신경망 모형은 IPO 변수와 증권신고서 어조 변수를 함께 사용했으며, 61.59%의 예측 정확도를 보였다.

본 연구의 학술적 의의는 다음과 같다. 첫째, 본 연구는 기계학습 기법을 이용해 IPO 주식의 상장 이후 가격 등락 여부를 예측했고, 로지스틱 회귀분석과 같은 통계적 기법을 사용할 때보다 예측 성과를 향상시켰다. 다양한 기계학습 기법의 예측 정확도를 비교해본 결과, 유전자 알고리즘을 활용한 인공신경망 모형이 IPO 주식의 상장 이후 가격 등락 여부 예측에 가장 효과적임을 확인하였다. 둘째, 텍스트 분석을 통해 증권신고서의 비정형화된 텍스트에 나타난 어조를 분석한 변수를 이용해 예측 성과를 향상시켰다. 본 연구는 한국어 금융 텍스트 분석에 적합한 감성사전이 잘 구축되어 있지 않은 상황에서 사전 기반이 아닌 기계학습을 이용하여 어조 분석을 실시하였다. 이를 통해 증권신고서 어조 변수를 추가적으로 이용하면 IPO 주식 가격 등락 여부 예측력이 좋아진다는 것을 확인할 수 있었다. 국내 증권신고서 텍스트의 어조를 분석하여 IPO 주식 가격의 등락을 예측한 연구는 매우 드물다. 따라서 본 연구는 텍스트 분석과 기계학습을 함께 이용한 IPO 연구라는 점에서 의의를 지닌다.

본 연구의 실무적 의의는 다음과 같다. 본 연구가 제시하는 IPO 주식 가격 등락 예측 모델은 IPO 투자자들의 투자의사결정에 기여할 수 있다. 국내 IPO 시장은 일반 주식 시장에 비해 정보가 한정적이며 불확실성이 높은 편이다(Cho et al., 2020). 본 연구의 예측 모델은 최대 61.59%의 뛰어난 예측 정확도를 보이기 때문에 투자 여부 등 관련 의사결정에 도움을 줄 수 있을 것이라 기대한다. 또한 본 연구는 개인투자자들이 쉽게 접근할 수 있는 증권신고서에서 얻을 수 있는 정보를 변수로 이용해 IPO 가격의 등락 여부를 예측하였다. 개인투자자의 경우, IPO 투자 시 기업 투자자에 비해 기업에 대해 얻을 수 있는 정보가 다소 한정적이며 투자 수익에 대한 불확실성이 더 크다. 그럼에도 불구하고 IPO 투자에 참여하는 개인투자자들이 증가하는 추세인데, 본 연구는 이러한 점에서 개인투자자들에게 특히 유용하다는 의의를 지닌다.

본 연구의 한계 및 추후 연구 방향은 다음과 같다. 본 연구는 IPO 가격의 등락 여부만을 고려하였기 때문에 실질적인 수익률을 예측하지는 못하였다는 한계를 지닌다. 추후 수익률 예측 등의 연구를 진행해 한계를 보완할 수 있다. IPO 주식 상장 이후 +5거래일 가격의 상승·하락을 예측하였지만 그 이후의 장기적인 성과까지 살펴보지는 못하였다는 점도 한계이다. 따라서 향후에는 본 연구를 토대로, IPO 관련 변수와 증권신고서 어조 변수가 IPO 주식의 단기 및 장기 수익률까지도 예측할 수 있는지 검증해볼 필요가 있다. 또한, 본 연구에서는 표본의 약 84%가 코

스닥 시장 상장기업이라는 점을 고려해 입력변수 중 하나로 상장일 전 코스닥 수익률을 사용하였으나, 코스피 IPO 종목의 경우 코스닥 수익률은 관련이 없는 변수일 가능성이 높다. 따라서 추후에는 코스닥 시장 상장기업과 코스피 시장 상장기업을 따로 살펴봄으로써 각 시장의 특성을 제대로 반영할 필요성이 있어 보인다. 마지막으로, 증권 신고서 텍스트 분석 시에 데이터 특성 및 기간에 따라 효과적인 방법에 차이가 있을 수 있기 때문에 향후 금융 분야에 특화된 감성사전을 이용한 어조 분류 등과 비교 분석이 필요할 수 있다. 또한, 증권신고서 뿐만 아니라 시장조사 보고서와 같은 다양한 텍스트의 주제, 시제, 가독성 등을 이용한 분석이 이루어질 수도 있을 것이다.

참고문헌(References)

Ahn, C. K. and D. Kim, "Efficient variable selection method using conditional mutual information," *Journal of the Korean Data and Information Science Society*, Vol.25, No.5(2014), 1079~1094.

Ahn, H. C., "Optimization of Multiclass Support Vector Machine using Genetic Algorithm: Application to the Prediction of Corporate Credit Rating," *Information Systems Review*, Vol.16, No.3(2014), 161~177.

Baba, B. and G. Sevil, "Predicting IPO initial returns using random forest," *Borsa Istanbul Review*, Vol.20, No.1(2020), 13~23.

Baek, J. S. and M. S. Jeong, "A study on the effect of competition rate of subscription and guarantee rate on the underpricing in the initial public offerings," *The Korean Finance Association*, (2018), 1943~1980.

Basti, E., C. Kuzey, and D. Delen, "Analyzing initial public offerings' short-term performance using decision trees and SVMs," *Decision Support Systems*, Vol.73, (2015), 15~27.

Benveniste, L. M., and P. A. Spindt, "How investment bankers determine the offer price and allocation of new issues," *Journal of Financial Economics*, Vol.24, No.2(1989), 343~361.

Breiman, L, "Random forests," *Machine Learning*, Vol.45, No. 1(2001), 5~32.

Brown, I. and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, Vol.39, No.3(2012), 3446~3453.

Buehlmaier, M. M. M. and T. M. Whited, (2018) "Are Financial Constraints Priced? Evidence from Textual Analysis," *The Review of Financial Studies*, Vol.31, No.7(2018), 2693~2728.

Chan, Y, "Retail Trading and IPO Returns in the Aftermarket," *Financial Management,* Vol.39, No.4(2010), 1475~1495.

Chen, Y. S. and C. H. Cheng, "A soft-computing based rough sets classifier for classifying IPO returns in the financial markets," *Applied Soft Computing*, Vol.12, No.1(2012), 462~475.

Cho, D. H., H. S. Ryou, S. H. Jung, and K. J. Oh, "Using AI to develop forecasting model in IPO market," *Journal of the Korean Data and Information Science Society*, Vol.31, No.3 (2020), 579~590.

Chun, K. M., I. H. Gee, and H. U. Lee, "The Effect of IPO Subscription Rates for Institutional Investors and Private Investors

on IPO Firm Performance: The Moderating Role of Competition and On-line Reviews," *Korean Journal of Business Administration*, Vol.26, No.5(2013), 1149~1176.

Cortes, C. and V. Vapnik, "Support-vector networks," *Machine learning*, Vol.20, No.3(1995), 273~297.

Dadgar, S.M., M.S. Araghi., and M.M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification" *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, (2016), 112~116.

Derrien, F, "IPO Pricing in "Hot" Market Conditions: Who Leaves Money on the Table?," *The Journal of Finance*, Vol.60, No.1(2005), 487~521.

Esfahanipour, A., M. Goodarzi., and R. Jahanbin, "Analysis and forecasting of IPO underpricing," *Neural Computing and Applications*, Vol.27, No.3(2015), 651~658.

Fang, F., K. Dutta, and A. Datta, "Domain Adaptation for Sentiment Classification in Light of Multiple Sources," *INFORMS Journal on Computing*, Vol.26, No.3(2014), 586-598.

Fuksa, M, "Sentiment and the Performance of Initial Public Offerings," *Available at SSRN 2243379*, (2013).

Gandoman, S. H., N. Kiamehr., and M. Hemetfar, "Forecasting Initial Public Offering Pricing Using Particle Swarm Optimization (PSO) Algorithm and Support Vector Machine (SVM) In Iran," *Business and Economic Research*, Vol.7, No.1(2017), 336.

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

Han, G. S. "A Study on the Underpricing of IPOs in Korea Capital Market," *Korean International Accounting Review*, Vol.59, (2015), 125~146.

Hanley, K. W. and G. Hoberg, "The Information Content of IPO Prospectuses," *Review of Financial Studies*, Vol.23, No.7(2010), 2821~2864.

Hong, S. H. and K. S. Shin, "Using GA based Input Selection Method for Artificial Neural Network Modeling: Application to Bankruptcy Prediction," *Journal of Intelligence and Information Systems*, Vol.9, No.1(2003), 227~249.

Hong. T. H. and E. M. Kim, "The Prediction of Purchase Amount of Customers Using Support Vector Regression with Separated Learning Method," *Journal of Intelligence and Information Systems*, Vol.16, No.4(2010), 213~225.

Ibbotson, R. G, "Price performance of common stock new issues," *Journal of Financial Economics*, Vol.2, No.3(1975), 235~272.

Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regression," *2020 6th Information Technology International Seminar (ITIS), Surabaya, Indonesia*, (2020), 238~242.

Islam, M., F.E. Jubayer, and S.I. Ahmed, "A support vector machine mixed with TF-IDF algorithm to categorize Bengali document," *2017 International Conference on Electrical, Computer and Communication Engineering*, (2017), 191~196.

Jegadeesh, N. and D. Wu, "Word power: A new approach for content analysis," *Journal of Financial Economics*, Vol.110, No.3(2013), 712~729.

Joh, S. W. and Y. Kim, "Is Textual Information Informative to Informed Investors? Evidence from Bidding Information of Institutional Investors in IPOs," *The Thirteenth Conference on Asia-Pacific Financial Markets, Seoul, Korea,* (2018).

Jung, J. Y. and K. W. Park, "A Study on Investor Protection through Morphological Analysis of Equity Crowdfunding Investment Manual," *Journal of Information Technology Services*, Vol.18, No.5(2019), 165~182.

Kaohua, Y. and Liwen, Z, "Analysis of influencing factors of IPO underpricing based on rough set and support vector machine," *2012 International Conference on Information Management, Innovation Management and Industrial Engineering*, Vol. 3, (2012), 244~248.

Katsafados, A. G., I. Androutsopoulos., I. Chalkidis., E. Fergadiotis., G. N. Leledakis., and E. G. Pyrgiotakis, "Textual information and IPO underpricing: A machine learning approach," *MPRA Paper 103813, University Library of Munich, Germany*, (2020).

Kim, D. Y. and D. E. Won, "An AI Model for Short-term KOSPI Prediction: A Machine Learning-Based Model Using Random Forest Technique," *Quantitative Issue*, Samsung Securities Research Center, (2019).

Kim, H. A. and S. C. Jung, (2010) "The Effect of Optimistic Investors' Sentiment on Anomalous Behaviors in the Hot Market IPOs," *The Korean Journal of Financial Management*, Vol.27, No.2(2010), 1~33.

Kim, H. H. "Suggestions for Disclosure of Risk Factors in Investment," *Practical Explanations of Securities*, Korean Listed Companies Association, (2008), 105~108.

Kim, H. J., J. W. Park, and J. W. Lee, "A Study on the Textual Analysis Research Environment using the DART System in Korea," *Korean Accounting Journal*, Vol.24, No.4(2015), 199~221.

Kim, I. H., "Ways to Go Public: Choice of IPO, Sellout, and Reverse Takeover," *The Korean Finance Association*, (2008), 958~1008.

Kim, J., S.M. Jun., S. Hwang., H.K. Kim., J. Heo., and M.S. Kang, "Impact of Activation Functions on Flood Forecasting Model Based on Artificial Neural Networks," *Journal of The Korean Society of Agricultural Engineers*, Vol.63, No.1(2021), 11~25.

Kim, K. J. and H. C. Ahn, "Optimization of Support Vector Machines for Financial Forecasting," *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 241~254.

Kim, K. Y., G. R. Lee, and S.W. Lee, "A Comparative Analysis of Artificial Intelligence System and Ohlson model for IPO firm's Stock Price Evaluation," *Journal of Digital Convergence*, Vol.11, No.5(2013), 145~158.

Kim, S. J. and H. C. Ahn, "Application of Random Forests to Corporate Credit Rating Prediction," *The Journal of Business and Economics*, Vol.32, No.1(2016), 187~211

Kim, T. H., "A Study on the present situations and some proposal for improvements of IPO Regulations and Systems," *Journal of Business Administration & Law*, Vol.26, No.4(2016), 201~238.

Kim, Y. S. and S. W. Joh, "Text Analysis for IPO firms in Korea: Analysis of Korean Texts in Registration Statements via Machine Learning," *Korean Journal of Financial Studies*, Vol.48,

No.2(2019), 215~235.

Kolchyna, O., T. T. Souza., P. Treleaven., and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," *arXiv preprint arXiv:1507.00955*, (2015).

Lee, H. S., S. H. Jeong, and K. J. Oh, "A study on the prediction of Korean NPL market return," *Journal of Intelligence and Information Systems*, Vol.25, No.2(2019), 123~139.

Lee, S. W. and J. H. Kim, "A Study on the Extraction of Psychological Distance Embedded in Company's SNS Messages Using Machine Learning," *Information Systems Review*, Vol.21, No.1(2019), 23~38.

Li, F, "The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach," *Journal of Accounting Research*, Vol.48, No. 5(2010), 1049~1102.

Loughran, T. and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, Vol.66, No.1(2011), 35~65.

Loughran, T. and B. McDonald, "IPO first-day returns, offer price revisions, volatility, and form S-1 language," *Journal of Financial Economics*, Vol.109, No.2(2013), 307~326.

Luque, C., D. Quintana., and P. Isasi, "Predicting IPO Underpricing with Genetic Algorithms," *International Journal of Artificial Intelligence*, Vol.8, No.12(2012), 133~146.

Ly, T. H. and K. Nguyen, "Do Words Matter: Predicting IPO Performance from Prospectus Sentiment," *2020 IEEE 14th International Conference on Semantic Computing*, (2020), 307~310.

Mai, F., S. Tian., C. Lee., and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *European Journal of Operational Research,* Vol.274, No.2(2019), 743~758.

Manurung, J., H. Mawengkang., and E. Zamzami, "Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment," *Journal of Physics: Conference Series*, Vol.930, No. 1(2017).

Martens, D. and F. Provost, "Explaining Data-Driven Document Classifications," *MIS Quarterly*, Vol.38, No.1(2014), 73-100.

Miller, E. M, "Risk, uncertainty, and divergence of opinion," *Journal of Finance*, Vol.32, No.4 (1977), 1151~1168.

Min, J. H, "IPO Stock Performance of Institutional and Individual Investors," *Journal of CEO and Management Studies*, Vol.20, No.3(2017), 75~98.

Moraes, R., J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications*, Vol.40, No.2(2013), 621-633.

Muditomo, A. and A. S. Broto, "IPO Performance Prediction During Covid-19 Pandemic in Indonesia Using Decision Tree Algorithm," *Journal of Finance and Banking*, Vol.25, No.1(2021), 132~143.

Nguyen, H., A. Veluchamy., M. Diop., and R. Iqbal, "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Science Review*, Vol.1, No.4(2018), 7.

Park J. S. and B. H. Han, "Analyzing the IPO

Market of 2020 and Forecasting the Market of 2021," *KOSDAQ Venture,* Eugene Research Center, (2021), 16-150.

Park, J. W., G. C. Jung, and J. E. Cho, "Institutional Investor Trading and IPOs Performance," *Korean Journal of Financial Studies*, Vol.45, No.1(2016), 171~192.

Park K. J. and J. Q. Jeon, "The Effect of IPO Syndicates on Underwriting Services: Focusing on Multiple Lead Underwriters and Co-Managers," *Korean Journal of Financial Studies,* Vol.44, No.1(2015), 189~219.

Park, S. G., H. J. Lee, H. J. Sim, J. Y. Lee, and J. E. Oh, "Construction of Sound Quality Index for the Vehicle HVAC System Using Regression Model and Neural Network Model," *Korean Society for Noise and Vibration Engineering,* (2006), 1308~1313.

Perera, W. and N. Kulendran, "Short-run underpricing and its determinants: Evidence from Australian IPOS," *Corporate Ownership and Control*, Vol.13, No.3(2016), 502~517.

Prastyo, P.H., I. Ardiyanto., and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, (2020), 1~6.

Quintana, D., F. Chávez., R. M. Luque Baena., and F. Luna, "Fuzzy techniques for IPO underpricing prediction," *Journal of Intelligent & Fuzzy Systems*, Vol.35, No.1(2018), 367~381.

Reber, B., B. Berry., and S. Toms, "Predicting mispricing of initial public offerings," *Intelligent Systems in Accounting, Finance, and Management*, Vol.13, No.1(2005), 41~59.

Rock, K, "Why new issues are underpriced," *Journal of Financial Economics*, Vol.15, No.1−2 (1986), 187~212.

Ruud, J. S, "Underwriter price support and the IPO underpricing puzzle," *Journal of Financial Economics*, Vol.34, No.2(1993), 135~151.

Seo, K. K., "Sales Prediction of Electronic Appliances using a Convergence Model based on Artificial Neural Network and Genetic Algorithm," *Journal of Digital Convergence*, Vol.13, No.9(2015), 177~182.

Seo, S. H. and J. T. Kim "Research Trends in Deep Learning-Based Sentiment Analysis," Journal of Korea Multimedia Society, Vol.20, No.3(2016), 8~22.

Shin, S. H., H. J. Lee, and J. J. Ahn, "A study on initial price change prediction of IPO shares using non-financial information," *Journal of the Korean Data and Information Science Society*, Vol.29, No.2(2018), 589~616.

Sun, A., E. P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decision Support Systems*, Vol.48, No.1(2009), 191-201.

Tao, J., A. V. Deokar., and A. Deshmukh, "Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach," *Journal of Business Analytics*, Vol.1, No.1(2018), 54~70.

Tetlock, P. C., M. Saar-Tsechansky., and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," *The Journal of Finance*, Vol.63, No.3(2008), 1437~1467.

Yan, Y., X. Xiong., J. G. Meng., & G. Zou, "Uncertainty and IPO initial returns: Evidence from the Tone Analysis of China's IPO Prospectuses," *Pacific-Basin Finance Journal*, Vol.57, (2019), 101075.

Abstract

The prediction of the stock price movement after IPO using machine learning and text analysis based on TF-IDF

Suyeon Yang* · Chaerok Lee** · Jonggwan Won** · Taeho Hong***

There has been a growing interest in IPOs (Initial Public Offerings) due to the profitable returns that IPO stocks can offer to investors. However, IPOs can be speculative investments that may involve substantial risk as well because shares tend to be volatile, and the supply of IPO shares is often highly limited. Therefore, it is crucially important that IPO investors are well informed of the issuing firms and the market before deciding whether to invest or not. Unlike institutional investors, individual investors are at a disadvantage since there are few opportunities for individuals to obtain information on the IPOs. In this regard, the purpose of this study is to provide individual investors with the information they may consider when making an IPO investment decision.

This study presents a model that uses machine learning and text analysis to predict whether an IPO stock price would move up or down after the first 5 trading days. Our sample includes 691 Korean IPOs from June 2009 to December 2020. The input variables for the prediction are three tone variables created from IPO prospectuses and quantitative variables that are either firm-specific, issue-specific, or market-specific.

The three prospectus tone variables indicate the percentage of positive, neutral, and negative sentences in a prospectus, respectively. We considered only the sentences in the Risk Factors section of a prospectus for the tone analysis in this study. All sentences were classified into ‘positive’, ‘neutral’, and ‘negative’ via text analysis using TF-IDF (Term Frequency – Inverse Document Frequency). Measuring the tone of each sentence was conducted by machine learning instead of a lexicon-based approach due to the lack of sentiment dictionaries suitable for Korean text analysis in the context of finance. For this reason, the training set was created by randomly selecting 10% of the sentences from each prospectus, and the

* School of Management Engineering, College of Business, KAIST
** School of Business, Pusan National University
*** Corresponding author: Taeho Hong
School of Business, Pusan National University, Busan, Korea
Tel: *** - **** - **** E-mail: hongth@pusan.ac.kr

sentence classification task on the training set was performed after reading each sentence in person. Then, based on the training set, a Support Vector Machine model was utilized to predict the tone of sentences in the test set. Finally, the machine learning model calculated the percentages of positive, neutral, and negative sentences in each prospectus.

To predict the price movement of an IPO stock, four different machine learning techniques were applied: Logistic Regression, Random Forest, Support Vector Machine, and Artificial Neural Network. According to the results, models that use quantitative variables using technical analysis and prospectus tone variables together show higher accuracy than models that use only quantitative variables. More specifically, the prediction accuracy was improved by 1.45% points in the Random Forest model, 4.34% points in the Artificial Neural Network model, and 5.07% points in the Support Vector Machine model. After testing the performance of these machine learning techniques, the Artificial Neural Network model using both quantitative variables and prospectus tone variables was the model with the highest prediction accuracy rate, which was 61.59%. The results indicate that the tone of a prospectus is a significant factor in predicting the price movement of an IPO stock. In addition, the McNemar test was used to verify the statistically significant difference between the models. The model using only quantitative variables and the model using both the quantitative variables and the prospectus tone variables were compared, and it was confirmed that the predictive performance improved significantly at a 1% significance level.

Key Words : Stock Price Prediction, IPO, TF-IDF, Machine Learning, Tone Analysis

Received : June 16, 2022 Revised : June 21, 2022 Accepted : June 22, 2022
Corresponding Author : Taeho Hong

저 자 소 개

양 수 연
현재 KAIST 경영공학부 석사과정에 재학 중이다. 부산대학교 경영학과에서 학사학위를 취득하였다. 주요 관심분야는 비즈니스 애널리틱스, 딥러닝, AI 등이다.

이 채 록
부산대학교 경영학과에서 학사 학위를 취득하였다. 주요 관심분야는 데이터 엔지니어링, 빅데이터, 데이터베이스, AI 등이다.

원 종 관
부산대학교 경영학과에서 학사학위를 취득하였다. 현재 부산대학교 경영학과 경영정보 전공 석사과정에 재학 중이다. 주요 관심분야는 지능형 테크핀, 암호화폐 예측, 신용평가, 딥러닝, AI 등이다. 딥러닝 연구결과를 정보시스템연구, 지식경영연구 등에 게재하였다.

홍 태 호
현재 부산대학교 경영학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공하여 공학석사와 공학박사를 취득하였다. 주요 관심분야는 비즈니스 애널리틱스, 딥러닝, 오피니언 마이닝, CRM 등이다. 주요 논문을 Expert Systems, Expert Systems with Applications, Information Processing & Management, Asia Pacific Journal of Information Systems, 지능정보연구, 정보시스템연구 등에 게재하였다.