# News and Social Media Text and Investor Expectation*
# 뉴스/소셜 미디어 텍스트와 투자자 기대

Juhwa Lee(First Author)
Sungkyunkwan University
(*ljh3105@skku.edu*)
Doojin Ryu(Corresponding Author)
Sungkyunkwan University
(*sharpjin@skku.edu*)

This paper analyzes whether the investor expectation implied by the text in news articles or social media forums affects stock returns in the Korean market. Our model, trained on 640,457 input articles and forum posts, classifies each post as positive or negative, employing word embedding based on the Word2Vec and bi-directional long short-term memory network to construct the investor expectation indices. We find that the expectation index constructed from news articles and the index from social media forums can explain stock return movements. Interestingly, the investor expectation extracted from social media forums outperforms the expectation from news articles.

Key Words: Interdisciplinarity, Investor Expectation, Machine Learning, News Media Text, Social Media Text, Stock Market

## Ⅰ. 서 론

공학과 응용과학 분야뿐 아니라, 경영학 분야에서도 인공지능(artificial intelligence)과 기계학습(machine learning)에 관한 관심이 커지고 있다. 인공지능은 사람의 행동을 모방하여 판단과 예측을 하고, 기계학습을 통하여 대량의 정보를 활용한다(Russell and Norvig, 2020). 자연어처리(natural-language processing; NLP)는 인공지능/기계학습의 한 분야로, 말 또는 글 형태의 정보를 기계가 인간과 비슷한 방식으로 활용하는 과정을 의미한다. 소셜 네트워크(social network)에 작성된 글을 분석함으로써 소비자가 작성한 평가와 의견으로부터 사업 기회를 찾거나 정책적 관점에서 유권자의 수요를 분석하는 등 활용할 수 있는 범위가 넓다(Hirschberg and Manning, 2015). 자연어처리는 크게 두 과정으로 나뉘는데, 감정분석, 기계 번역, 그리고 유사도 측정 등을 활용하는 자연어 이해(natural-language understanding)와 인공지능 글쓰기, 자동 완성 기능 등을 활용하는 자연어 생성(natural-language generation) 과정으로 구분된다(Sumathy and

Submission Date: 12. 10. 2020 Revised Date: (1st: 02. 03. 2021) Accepted Date: 02. 08. 2021

* This paper is an extended version of Lee's dissertation. The authors are grateful for the helpful comments and suggestions from Keun-Yeong Lee, Young-Han Kim, Shu-Chin Lin, Jinyoung Yu, Karam Kim, and Sohee Shin.

Chidambaram, 2013). 인공지능 개념을 통한 기계학습은 경제 및 금융 분야에도 활용되고 있다(Kim, Cho, and Ryu, 2020; Kim, Ryu, and Cho, 2019). 본 논문은 자연어처리 과정 중 자연어 이해에 집중하며, 기업에 관한 뉴스기사와 투자자가 작성한 글을 바탕으로 긍정·부정 기대 지수를 도출한다. 구체적으로, 인터넷에서 실시간으로 작성되는 텍스트 자료를 활용하여 투자자 기대 지수를 구축하고, 이러한 지수가 업종별 국내 대표기업의 주식수익률을 설명할 수 있는지를 살펴본다.

빅데이터(big data)의 가공 및 활용기술이 발전하면서 글, 그림, 혹은 영상 형태의 비정형 자료를 활용한 연구가 증가하는 추세이다(Kim and Lee, 2013; O'Leary, 2013). 위성사진 자료에 기반한 기계학습 방법을 통하여 농작물 수확의 예측을 시도한 Kim and Lee (2016)의 연구와 경제 분야에서 비정형 자료에 기반한 Word2Vec 임베딩(embedding)[1] 기법을 활용하여, 경제위기에 대한 조기 경보 지표를 도출한 Huang, Simpson, Ulybina, and Roitman (2019)의 연구가 그 예이다. 이들은 뉴스기사를 활용하여 도출한 심리지수가 기존의 경기예측 모형을 개선하며, 재무변수의 변동 및 자금흐름에 대한 분석과 예측에 활용될 수 있는지에 주목한다. Shapiro, Sudhof, and Wilson (2021)은 뉴스기사에 기반한 심리지수가 소비자 심리지수보다 선행하고 있으며 심리 정보를 통해 이자율, 생산 및 소비의 변화를 설명한다고 주장한다.

주식시장에서 투자자는 보유한 주식 지분에 비례하여 기업의 경영에 관여하거나 배당금 혹은 자본이득을 기대한다(Park, Lee, and Lee, 2003). 한편, 투자자는 회사가 경영난을 겪거나 파산하면 투자금액에 따른 책임을 지기 때문에, 기업의 향후 가치변동에 대해 다양한 예측을 시도하게 된다. 이때, 투자자는 투자기간과 위험요인 등의 요소를 고려하여 미래의 현금흐름을 할인하게 되며, 이렇게 산출한 현재가치를 해당 주식의 시장가격과 비교하여 의사결정을 한다(Bodie, Kane, and Marcus, 2021). 합리성 이론에 기반한 고전 재무이론에 따르면, 수익과 위험은 투자자의 의사결정과정에 영향을 미친다. 전통적 금융경제이론과 달리, 행태재무학(behavioral finance)에서는 자산수익률을 결정할 때 투자자의 심리적 편향이 중요한 요소가 된다고 주장한다. Stein (1987)의 연구는 일반적으로 정보가 부족한 개인투자자가 비합리적인 투자결정을 하며, 이러한 투자행위가 주식가격에도 영향을 미친다고 주장한다. De Long, Shleifer, Summers, and Waldmann (1989)은 비합리적인 투자자가 자신의 능력을 과신하거나 불확실한 정보를 바탕으로 투자를 하는 경향이 있음을 제시한다. Baker, Wang, and Wurgler (2008)와 Stambaugh, Yu, and Yuan (2012)은 전문성과 투자경험이 부족하고 심리적 편향에 영향받는 비합리적인 투자자의 거래가 주식가격에 영향을 미친다고 주장한다. Baker and Wurgler (2006, 2007)의 연구에서 투자자 심리지수를 구축하고, 투자자 심리와 주식수익률과의 관계를 규명한 이후, 다수의 실증연구가 주식가격의 움직임과 투자자 심리 및 기대와 관계가 있음을 보이고(Kim, Ryu, and Seo, 2014; Kim, Ryu, and Yang, 2018; Ryu, Kim, and Yang, 2017), 심리지수가 금융시장 투자자의 거래행태에 미치는 영향을 연구한다(Kim and Ryu, 2021; Kim, Ryu, and Yu, 2021; Seok, Cho, and Ryu, 2019a, 2019b).

1) 임베딩이란 사람이 사용하는 말을 숫자의 나열인 벡터(vector) 값으로 변환하여 저장하는 과정을 의미한다(Bengio, Ducharme, Vincent, and Jauvin, 2003).

최근에는 뉴스 정보나 인터넷 커뮤니티, 또는 소셜 네트워크 서비스(social network service; SNS) 등에 게시된 텍스트 자료를 바탕으로 심리지수를 구축하는 연구가 활발하다(Behrendt and Schmidt, 2018; Kim, Lee, Shin, and Park, 2019; Renault, 2017). 특히, 소셜 미디어 및 소셜 네트워크에 작성되는 글에 대해 감정분석 기법을 활용하여 주식가격의 움직임에 대한 예측을 시도한 연구가 있다(Bollen, Mao, and Zeng, 2011; Kim, Jeong, and Lee, 2014). Lee, Kang, Kim, and Lee (2013)는 주식시장과 관련된 텍스트 정보를 수집하고 긍정 및 부정 감정으로 분류하여 투자자 심리지수를 구축하였고, 해당 지수가 주식가격의 움직임을 예측한다고 주장한다. 행동경제학 분야에서도 수치형 자료의 한계에서 벗어나고자 하는 다양한 연구방법이 등장한다. Kalyani, Bharathi, and Jyothi (2016)는 텍스트 자료를 바탕으로 심리지수를 도출하여 주가를 예측하는 연구를 진행하였으며, Random forest, Naive Bayes, Support vector machine과 같은 기계학습 기법을 적용하여 실제 종가 정보와 가장 비슷한 예측 결과를 도출하는 모형을 찾고자 하였다.

본 논문은 네이버 뉴스기사와 네이버 금융 종목토론실에 게시된 텍스트 자료를 활용하여 투자자 기대지수를 구축하고, 해당 지수가 주식수익률과 관계가 있는지 살펴본다. 텍스트 자료를 활용하는 것은 실시간으로 투자자의 심리 정보를 확인할 수 있으며, 투자 판단의 근거를 상세하게 확인해 볼 수 있다는 장점이 있다. 이러한 텍스트 자료를 활용한 연구는 방대한 자료를 수집하고 변환하는 과정을 요구한다. 네이버 종목토론실 및 네이버 뉴스기사의 주식시장과 관련된 문장뿐만 아니라, 영화 평론 긍정 및 부정 분류 문장은 인터넷에서 사용하는 말이 다수 포함되어 있어 기계학습의 정확도를 높일 수 있다고 판단되어 학습에 사용되었다.[2] 또한, 공공데이터포털(https://www.data.go.kr)에서는 사회, 정치 및 경제 등 다양한 주제로 작성된 뉴스를 긍정 및 부정 내용으로 분류한 자료를 제공한다. 이를 활용하여 다양한 주제를 다룬 뉴스가 학습자료에 포함되었다. 대량으로 수집된 텍스트 자료는 불필요한 내용을 포함하는 경우가 있고, 불필요한 내용은 정상적인 기계학습 과정을 방해한다. 이러한 문제를 방지하기 위해, 불용어(stop words) 처리 과정을 적용하였다. 문장의 형태소를 분석하는 기법으로 한글 고유의 특성을 반영한 분석 성능과 작업 시간을 고려하여 '은전한닢(MeCab)'이 사용되었다(Lee, 2018; Park and Cho, 2014). 단어 어휘를 이해하는 방법인 단어 임베딩은 Word2Vec 기법을 적용하였다(Levy and Goldberg, 2014). 문장의 의미를 이해하고 감정분류를 진행하는 신경망 모형은 긍정 및 부정 문장 분류 성과를 높일 수 있는 양방향 LSTM(bi-directional long short-term memory; BLSTM) 모형을 채택하였다. 실증분석 결과, 네이버 금융 종목토론실 작성 글을 활용한 기대 지수가 네이버 뉴스기사를 기반으로 한 기대 지수보다 주식수익률에 대한 설명력이 더 높은 것으로 나타났다.

본 연구의 구성은 다음과 같다. 제2장에서는 투자자 기대 지수를 구축하기 위해 사용한 자료와 모형의 구성 방법을 소개한다. 제3장에서는 분석자료에 관한 설명 및 통계량을 제시하며, 제4장에서는 회귀분석 결과를 제시한다. 제5장에서는 결론을 맺는다.

2) 본 자료는 자연어처리 분석가인 Lucy Park이 https://github.com/e9t/nsmc를 통해 제공하고 있는 공개 데이터를 활용하였다.

## Ⅱ. 표본선정 및 연구방법

### 2.1 연구 자료

논문의 분석대상으로 업종별 대표기업인 삼성전자(Samsung), 셀트리온(Celltrion), SK하이닉스(SK hynix), 현대자동차(HYUNDAI), 엔씨소프트(NCSOFT), 그리고 KB금융(KB)이 선정되었으며, 선정 이유는 다음과 같다. 첫째, 이러한 기업의 주식은 유동성이 매우 풍부하고 개인투자자의 거래 및 관심이 높은 종목이다. 둘째, 유가증권시장에 상장한 대기업을 대상으로 상호 계열 관계가 존재하지 않으면서, 한국표준산업분류 대분류를 기준으로 동일 산업에 소속되지 않은 기업을 선정하였다. 이는 다른 산업에 소속된 회사일지라도 같은 계열의 회사의 경우 투자자들이 동일 기업으로 간주하거나 특정 계열집단 소속회사에 공통적으로 해당하는 정보를 분석대상에 따라 중복하여 활용할 가능성이 있기 때문이다. 셋째, 일 평균 10개 이상의 뉴스와 20개 이상의 토론글이 작성되어 분석자료가 충분한 종목이다. 이는 분석에 적합하지 않거나 불필요한 자료를 제외하더라도 일별 기대 지수 값을 도출하기 위한 최소한의 관측치가 존재해야 하기 때문이다.

텍스트 자료를 얻기 위해 네이버 뉴스기사와 네이버 금융의 종목토론실을 활용하였는데, 그 이유는 다음과 같다. 먼저, 뉴스기사는 새로운 정보나 소식을 전달하는 전통적 대중 매체로 투자자를 포함한 다양한 독자들이 참고하는 자료이다. 또한, 전문 용어의 사용이 많고 비문이나 은어의 사용이 상대적으로 적기 때문에 양질의 학습자료로 활용될 가능성이 크다. 반면, 네이버 종목토론실은 투자자들이 집중하고 있는 이슈에 대한 실시간 정보와 인터넷 이용자의 게시글을 학습자료로 제공할 수 있다. 또한, 투자자가 작성하는 글에는 주관적 가치판단이 담겨 있어 긍정 및 부정으로 분류된 자료로 학습하기에 적합하다. 네이버에 게재된 뉴스기사는 포털 검색란에 기업명을 입력하고 분석하고자 하는 기간 및 페이지 번호를 설정한 후 인터넷 사이트의 구조를 분석하여 수집한다. 네이버 금융 종목토론실에 작성된 글은 인터넷 개발자도구 기능을 통해 필요한 사이트 정보를 분석하여 수집한다.[3] 영화 평론의 긍정 및 부정 문장, 정부 공공데이터포털에서 제공하는 뉴스의 긍정 및 부정 분류 글이 추가적인 학습자료로 사용하며, 특정 단어를 기준에 따라 긍정과 부정으로 분류한 자료를 사용한다.

본 연구의 분석기간은 2018년 1월부터 2020년 2월까지이며, 각 거래일을 구분하는 기준은 장 마감 시간을 고려하여 매 거래일의 15시 30분으로 정하였다. 즉, 15시 30분 이전에 작성된 글을 기반으로 구축한 투자자 기대 지수는 당일 주식수익률을 분석하는 데 사용하고, 장 마감 이후에 작성된 자료를 기반으로 구축한 기대 지수는 다음 날의 주식수익률을 분석하는 데 활용한다.

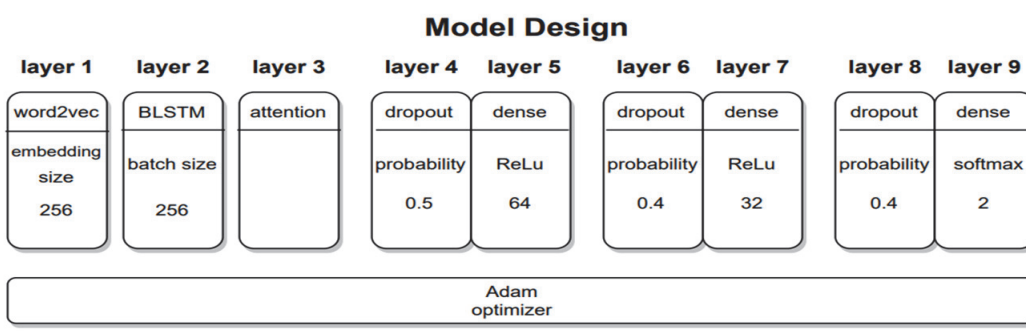
### 2.2 기계학습 모형 설계 과정 및 방법

본 연구를 진행하기 위한 심층학습 알고리즘(algorithm)은 Goodfellow, Bengio, and Courville (2016)과 Luong, Pham, and Manning (2015)의 자료를 참고하여 〈Figure 1〉과 같이 설계되었다. 텍스트에 작성된 단어에 대해 벡터화(vectorization)

3) 네이버 뉴스와 종목토론실 게시글의 경우 무분별한 크롤링을 막기 위한 정기적인 웹사이트 구조 변경 및 중간 경로 폐쇄로 인해 자료의 추가적인 수집은 현재 불가능하다.

과정을 진행하는 임베딩은 Word2Vec 방법 중 Skip-Gram 모형을 채택하여 레이어(layer) 1에 구성하였고, 문장에 대해 분류 학습이 이루어지는 작업은 레이어 2에 배치하였다. 감정분류 학습에는 양방향 LSTM이 사용되었으며, 배치 크기(batch size)는 256개, 학습 횟수는 10회로 적용하였다. 단어 입력 토큰 길이는 학습자료 중 가장 길이가 긴 글을 기준으로 설정하였다. 레이어 3에는 학습 내용의 정보 손실을 줄이고 문장에서 중요한 의미가 있는 단어 위주로 학습할 수 있도록 어텐션(attention) 기법을 적용하였으며(Bahdanau, Cho, and Bengio, 2016; Luong, Pham, and Manning, 2015), 레이어 4, 6, 8에는 드롭아웃(dropout) 규제를 적용하여 학습이 과도하게 진행되어 과잉적합(overfitting) 문제가 발생하지 않도록 하였다(Hinton, Srivastava, Krizhevsky, Sutskever, and Salakhutdinov, 2012). 첫 번째 드롭아웃에서는 0.5의 확률값을 설정하고 두 번째와 세 번째 드롭아웃에서는 0.4로 확률값을 설정한다. 레이어 5, 7, 9에는 입력과 출력을 이어주는 덴스(dense)를 배치하였다. 특히, 레이어 5, 7에는 은닉층을 갖는 다층 퍼셉트론(multilayer perceptron) 신경망 모형을 활용하였고, 첫 번째 덴스에서는 은닉층 계산을 위해 사용된 유닛값(뉴런의 개수)을 64개로, 두 번째 덴스에서는 유닛값을 32개로 설정하였다. 은닉층 활성화 함수로는 경사도 사라짐 문제(vanishing gradient problem)를 방지하기 위해 정규화 선형 유닛(rectified linear unit; ReLu)을 사용하였다(Nair and Hinton, 2010). 레이어 9에는 결과에 대한 확률을 제시해야 하므로 소프트맥스(softmax) 활성화 함수를 적용하여 0과 1 사이의 연속형 변수를 도출하고자 하였다. Word2Vec 임베딩을 위한 임베딩 크기(embedding size)는 256으로 설정한다. 윈도우(window) 값의 입력은 짧은 문장의 경우 2~5 값으로 설정하고, 긴 문장의 경우 5~20 값으로 설정하였을 때 높은 성능을 보고하는 기존 연구결과를 참고하여 5의 값으로 설정한다(Mikolov, Sutskever, Chen, Corrado, and Dean, 2013). 또한, 최적의 하이퍼파라미터(Hyperparameter) 값을 적용하기 위한 옵티마이저(optimizer)로 아담(adaptive moment estimation; Adam)을 사용한다(Kingma and Ba, 2017).

〈Figure 1〉이 제시하는 과정을 요약하면, 단어의 의미는 Word2Vec 과정을 통해 벡터공간(vector space)에 분포된 정보를 바탕으로 이해된다. 그 후, 글을 정방향과 역방향으로 읽고 정답(label) 값을 확인하며 기계학습을 진행하는데, 어텐션 기법을 통




〈Figure 1〉 텍스트 마이닝 학습을 위한 모형의 설계

해 문장 내에서 중요한 의미가 있는 단어 혹은 어절에 높은 가중치를 부여하여 효율적인 학습이 진행되도록 한다(Luong, Pham, and Manning, 2015). 마지막으로, 드롭아웃 규제 적용은 기계학습 과정에서 인간이 의도하지 않은 패턴을 발견하여 학습하는 경우나 과잉적합이 발생하는 것을 방지한다. 구성된 모형을 통해 각 $t$일의 게시글에 대해 각각 0과 1 사이의 연속형 변수를 제시하고 이를 산술평균하여 $t$일의 기대 지수를 도출한다.

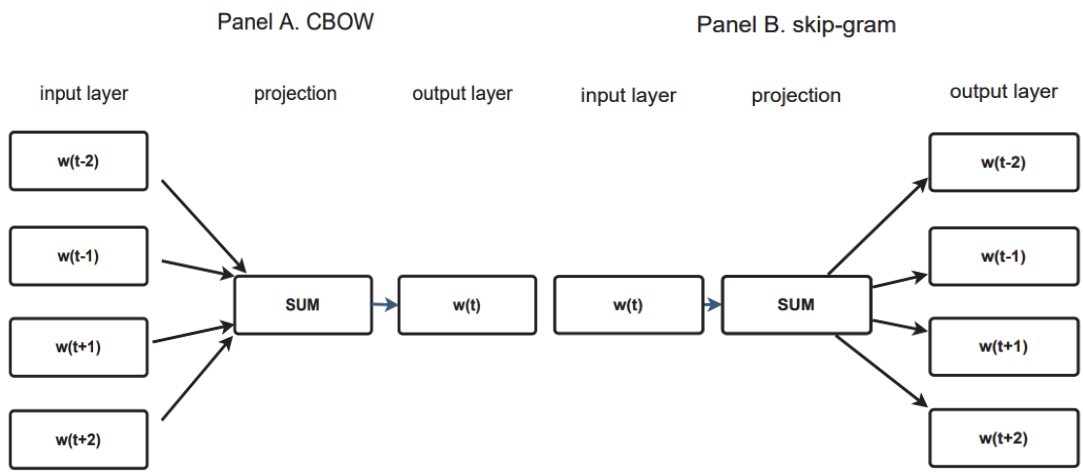
#### 2.2.1 단어 임베딩

컴퓨터는 사람이 사용하는 언어 자체를 읽고 해석하는 것이 아니라 숫자로 변형한 후 이를 읽는 방식으로 이해한다. 임베딩은 벡터화 과정을 거쳐 변환된 숫자를 벡터공간에 입력하고 저장하는 과정을 의미한다. 인간의 언어표현 방법이 다양한 것처럼, 사람이 사용하는 언어표현을 컴퓨터가 이해하는 임베딩 역시 많은 방법이 존재한다. 본 연구에서 채택한, 단어 위주의 임베딩을 활용하는 Word2Vec을 이해하기 위해서는 '분포 가정(distributional hypothesis)'이라는 개념에 대한 이해가 필수적이다. 분포 가정이란 문맥을 통해서 사람들이 사용하는 단어 의미를 이해할 수 있음을 뜻한다. 예를 들면, "*삼성전자 스마트폰 Z플립은 인폴딩 기술을 적용한 폴더블 폰이다.*"라는 문장이 있다면, 'Z플립' 단어를 학습하기 위해 주변에 사용된 단어인 '삼성전자,' '스마트폰,' '인폴딩,' '기술'이라는 단어를 활용하여 이해하게 된다.

〈Figure 2〉는 Mikolov, Sutskever, Chen, Corrado, and Dean (2013)이 제시한 모형을 재구성한 것으로, Word2Vec 기법의 두 모형인 CBOW (continuous bag of words) 모형과 Skip-Gram 모형 예시를 나타낸다. Word2Vec은 〈Figure 2〉에서 제시된 타겟 단어(target word)인 $w(t)$를 기준으로, 주변의 문맥 단어(context word)인 $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$설정 범위까지 학습하기 때문에, 높은 정확도를 나타내면서 빠른 연산속도를 나타낸다는 장점이 있다. Word2Vec의 학습 방법은 CBOW 방법과 Skip-Gram의 두 가지 방법으로 나뉘는데, CBOW는 $t$시점에 등장하는 타겟 단어인 $w(t)$를 기준으로, 전후에 등장하는 문맥 단어에 해당하는 $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$를 통해 타겟 단어를 유추하며 학습하는 방법이다. Skip-Gram은 타겟 단어를 통해 주변 문맥 단어의 의미를 예측하여 학습하는 방법이다. 제시된 두 방법 중에서 본 연구는 기계학습을 위해 Skip-Gram을 채택하였다(Bojanowski, Grave, Joulin, and Mikolov, 2017; Kang and Yang, 2019; Mikolov, Sutskever, Chen, Corrado, and Dean, 2013). 〈Figure 2〉에서 projection 과정은 입력 자료를 벡터값으로 나타낸 후, 출력 자료와 비교하여 학습하고 비슷한 의미를 갖는 단어끼리 가까운 벡터공간에 배치하는 것을 의미한다.

#### 2.2.2 순환 신경망 학습 모형

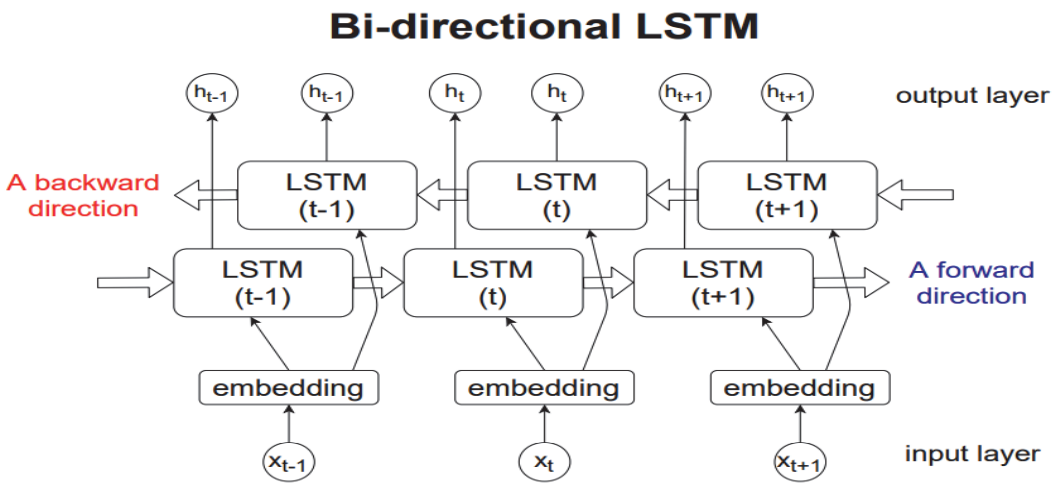
순환 신경망 기법(recurrent neural network; RNN)은 자료를 순서와 흐름에 따라 학습하는 기법을 의미한다. 인간이 사용하는 글에는 앞에서 사용된 단어의 의미가 뒤에 사용된 단어 해석에 영향을 미치는 경우가 많은데, 이 원리를 활용한 알고리즘이 순환 신경망 기법이다(Elman, 1990). 구체적으로, 순환 신경망은 타겟 단어($t$시점) 이전에 사용된 단어($t-1$시점)의 의미가 타겟 단어를 해석하는 데 영향을 미치도록 구성된 기법이다. 일반적인 순환 신경망 기법은 문장이 길어지는 경우, 앞부분에 사




〈Figure 2〉 CBOW 모형과 Skip-Gram 모형 예시

용된 정보에 대한 가중치가 약해지는 단점을 가지는데, 이를 장기 의존성(long-term dependency) 문제라고 한다(Hochreiter, Bengio, Frasconi, and Schmidhuber, 2001). 예를 들어, "*세계 최대 전자박람회 CES 2018에서 삼성전자의 폴더블 스마트폰이 공개될 전망이다. 업계 관계자에 따르면 아웃 폴딩 방식이 아닌 인폴딩 방식의 폴더블 폰이 공개될 것이며 OLED 디스플레이가 적용되어 뛰어난 화질을 자랑한다고 밝혔다. 이어서 LG전자 역시 롤러블 디스플레이가 적용된 프리미엄 TV를 곧 공개할 것이라고 관계자는 설명했다.*"의 예시문장을 통해 설명해보면, 글 앞부분에 사용되었던 단어인 'CES'와 '삼성전자'의 의미는 뒷부분으로 갈수록 약해지게 된다.

이러한 문제를 해결하기 위해, 일반적인 순환 신경망 기법의 한계를 개선한 LSTM 모형이 개발되었다. LSTM 모형은 긴 문장을 분석하고 이해할 때 뒷부분의 내용을 이해하기 위해 앞부분의 내용을 저장하여 활용한다는 특징을 갖는다(Hochreiter and Schmidhuber, 1997). LSTM 모형의 경우 문장의 후반부에 사용된 단어는 이전 시점의 의미를 지속해서 포함하게 되어 많은 정보를 담게 된다. 그러나 이는 앞 시점에 나온 단어일수록 적은 정보를 담고 있으며, 문장 내에서 사용된 단어에 대한 정보 반영이 한쪽으로 치우친 것으로 볼 수 있다(Ko, Yun, and Shin, 2018; Yildirim, 2018). 예를 들어, '삼성전자'라는 단어는 '전자박람회,' 'CES'의 정보를 반영하는 반면, 'LG전자'의 경우, 문장의 뒷부분에 작성되었기 때문에 '삼성전자'에 비해 더 많은 정보를 반영하게 된다. 이러한 이유로 LSTM 모형만 적용한 컴퓨터는 글 또는 문장 내에서 작성된 순서에 따라 반영되는 정보의 양의 차이가 발생하며, 최근에는 개선된 학습 모형인 양방향 LSTM이 주로 사용된다. 본 연구에서 활용한 양방향 LSTM 모형의 구조는 〈Figure 3〉과 같다. 〈Figure 3〉은 Graves and Schmidhuber (2005)가 제시한 모형을 표현한 것으로, 양방향 LSTM은 텍스트 글에서 사용된 단어의 의미를 이해하기 위해 문장을 정방향(forward




〈Figure 3〉 양방향 LSTM 순환 신경망 모형

direction)과 역방향(backward direction)으로 해석함으로써, 전반부와 후반부 시점의 정보를 종합하여 단어의 의미를 이해한다(Graves and Schmidhuber, 2005). 기존 LSTM 모형의 내용에 거꾸로 작동하는 과정을 추가함으로써 다음과 같은 설계가 구성되었다. 정방향과 역방향의 정보를 종합하는 과정은 문장 앞부분 혹은 뒷부분에 사용된 단어가 일방적으로 많은 정보를 갖게 되는 현상을 방지할 수 있다. 은닉상태(hidden state)에 해당하는 $h_t$값은 LSTM($t$) 과정을 통해 얻은 결과를 나타내고, embedding은 투입된 텍스트 정보인 $X_t$의 값들을 벡터화시킨 값이다.

## 2.3 통제변수 선정 및 회귀분석 모형 설계

본 절에서는 Smales (2016, 2020)가 사용한 금융시장 거시경제 변수와 Fama-French (1992, 1993) 3요인 변수를 각각 통제한 모형을 설명한다. 네이버 금융 종목토론실과 네이버 뉴스기사를 통해 만들어진 투자자 기대 지수가 거시경제 변수 혹은 위험요인 설명변수를 통제한 후에도 개별종목 주식수익률을 설명할 수 있는지를 조사한다.

### 2.3.1 금융시장 거시경제 변수 활용 모형

금융시장 거시경제 변수를 통제변수로 설정하고, 기대 지수를 설명변수로 설정한 모형은 개별 주식수익률에 대한 기대 지수의 설명력을 검증한다. Schrimpf (2010)는 금융시장 변수를 활용하여 5개 국가를 대상으로 주식수익률에 대한 예측을 시도하였는데, 단기 이자율과 기간 스프레드는 예측력이 없었으며, 아웃풋 갭(output gap)과 정부채권의 장기 수익률이 주식수익률을 예측함을 보인다. Hjalmarsson (2010)은 역시 금융시장 변수를 활용하여 주식수익률 예측 여부를 분석하였는데, 단기 이자율의 예측력은 발견하지 못하였으나, 기간 스프레드의 경우, 일부 선진국에서 주식수익률에 대한 예측력을 확인하였다. Chun (2020)은 단기 이자율, 기간 스프레드, 신용 스프레드 등의 금융변수를 활용하여 주식

수익률을 예측하였는데, 채무불이행 스프레드, 주가 순자산비율, 배당가격비율, 그리고 배당수익률이 예측력을 갖는 것으로 나타났다. 경기변동을 나타내는 거시경제 변수는 주식수익률을 설명할 수 있는데, Kam and Shin (2017)은 금리, 소비자물가지수, 달러 환율, 산업생산지수 등이 코스피 수익률에 유의한 영향을 미치고 있음을 제시하였다. 이를 고려하여, 본 연구는 국내의 금융시장 거시경제 변수의 영향을 통제한 식 (1)과 같은 분석 모형을 사용한다. 종속변수로는 개별기업 $i$의 $t$시점 일별 주식수익률($Stock\_Return_{i,t}$)을 사용한다. $Sentiment_{i,t}$는 개별기업에 대한 $t$일의 기대 지수를 나타내며, 네이버 금융 종목토론실($OP\_Sent_{i,t}$)과 네이버 뉴스기사($News\_Sent_{i,t}$)를 기반으로 구축한 기대 지수를 활용한다. $\triangle CD(91days)_t$는 CD 91수익률로 무위험 단기 이자율의 일별 변화량을 나타낸다. $\triangle Term\ Spread_t$는 국고채 10년 금리에서 국고채 3년 금리를 차감한 기간 스프레드의 일별 변화량을 나타낸다. $\triangle Credit\ Spread_t$는 회사채 3년물(BBB-) 금리와 회사채 3년물(AA-) 금리를 차감한 신용 스프레드의 일별 변화량을 나타낸다. 투자자의 기대와 공포심리를 나타내는 내재변동성 지수 정보를 활용하되, 본 연구에서는 종속변수가 주식수익률임을 고려하여 *VKOSPI*의 일별 수익률을 $VKOSPI_t$ 변수로 사용했다. 또한, 주식시장의 종합주가지수의 일별 수익률을 $KOSPI_t$변수로 사용했다.

$$\begin{aligned} Stock\_Return_{i,t} &= \alpha + \beta \cdot Sentiment_{i,t} \\ &+ \gamma_1 \cdot \triangle CD(91days)_t + \gamma_2 \cdot \triangle Credit\ Spread_t \\ &+ \gamma_3 \cdot \triangle Term\ Spread_t + \gamma_4 \cdot VKOSPI_t \\ &+ \gamma_5 \cdot KOSPI_t + \epsilon_{i,t} \end{aligned} \quad (1)$$

where $Sentiment_{i,t} \in \{OP\_Sent_{i,t}, News\_Sent_{i,t}\}$

### 2.3.2 Fama-French 3요인 변수 활용 모형

주가의 기대수익률을 설명하는 모형인 CAPM(capital asset pricing model)은 체계적 위험을 바탕으로 주식수익률을 설명한다. CAPM 모형의 경우, 시장 위험을 유일한 위험으로 설정했다는 한계가 있다. 따라서, 본 연구는 Fama-French 3요인 변수를 통제변수로 활용함으로써 기업의 규모와 가치가 주가의 기대수익률에 끼치는 영향을 고려한다.[4] Fama-French 3요인 변수 중에서 무위험 자산 대비 시장 초과수익률을 의미하는 변수 $MKT_t$는 $t$시점 일별 시장포트폴리오(market portfolio)의 대용치인 종합주가지수 수익률에서 무위험 이자율을 차감하여 구하고, 대형주 대비 소형주 초과수익률을 나타내는 $SMB_t$는 시가총액을 기준으로 소형주와 대형주로 나누고 소형주의 평균 수익률에서 대형주의 평균 수익률을 차감하여 계산한다. 성장주 대비 가치주 초과수익률을 의미하는 $HML_t$는 장부가에서 시가를 나누어 가치주와 성장주를 구분한 후, 가치주의 수익률 평균에서 성장주의 수익률 평균을 차감하여 구한

4) 통제변수는 한다파트너스(handapartners.com)에서 제공하는 자료를 사용한다. Fama-French 모형과 기업 특성 정보를 활용하여 자산수익률을 분석한 연구는 선진시장뿐 아니라 신흥시장에서도 수행되었다. Huang (2019)은 중국 주식시장에서 개별 주식수익률에 대한 Fama-French 모형의 설명력을 확인하는 한편, 중국 내 거래소에 따라 해당 모형의 설명력에 차이가 있을 수 있음을 밝혔다. 해당 연구는 중국의 주식시장으로 상해 주식시장과 선전 주식시장을 활용하는데, 두 주식시장의 규모의 차이, 등록된 기업의 수 그리고 기업 특성의 차이가 자산수익률을 분석하는 데 영향을 미친다고 설명하였다. Liu and Gao (2019)는 중국 주식시장에 상장된 기업을 분석한 연구로, 기업이 성장함에 따라 Fama-French 3요인 변수가 기대 주식수익률에 대한 영향이 어떻게 변하는지 설명한다. Hu, Chen, Shao, and Wang (2019)은 Fama-French 모형의 변수 중, 어떤 변수가 중국 주식시장에서 주식수익률에 유의한 설명력을 갖는지를 분석하였다.

다. 본 논문은 종목토론실과 뉴스기사로부터 도출한 기대 지수가 주식수익률에 대해 설명력을 갖는지를 검증하는 연구이므로, 기존의 연구들과 마찬가지로 Fama-French 3요인 변수를 통제변수로 사용하고 기대 지수를 설명변수로 설정하여 식 (2)와 같이 회귀분석을 실시하였다.

$$Stock\_Return_{i,t} = \alpha + \beta \cdot Sentiment_{i,t} + \gamma_1 \cdot MKT_t + \gamma_2 \cdot SMB_t + \gamma_3 \cdot HML_t + \epsilon_{i,t} \quad (2)$$

## Ⅲ. 분석대상 자료의 특성

### 3.1 시계열 및 상관계수 분석

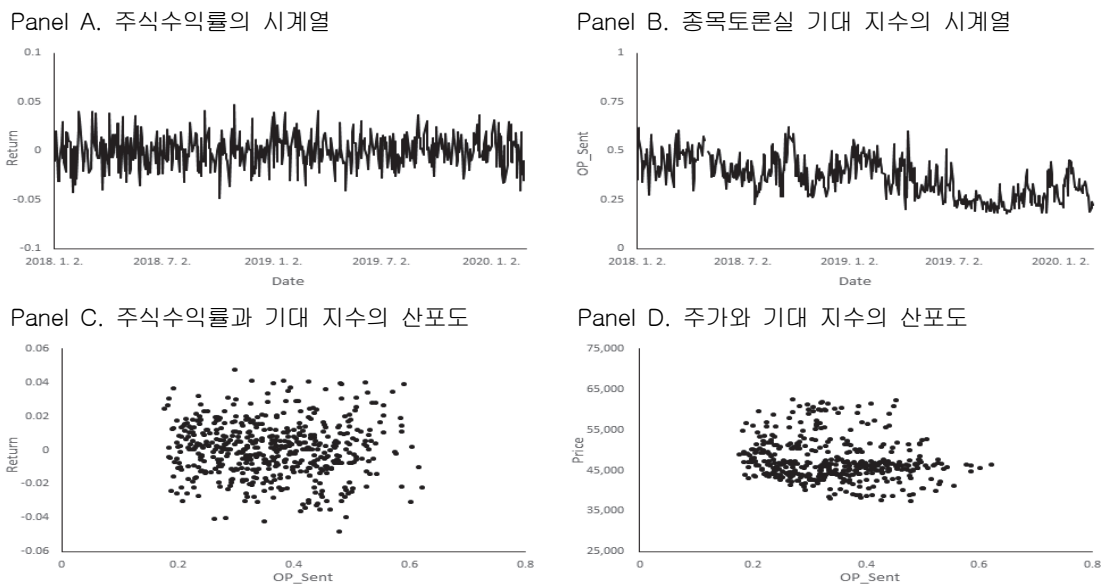
〈Figure 4〉는 2018년 1월부터 2020년 2월 간 삼성전자의 주식수익률과 종목토론실 글을 활용해 도출한 긍정·부정 기대 지수(OP_Sent)의 시계열과 주식수익률(Return)과 기대 지수 및 주가(Price)와 기대 지수 간의 산포도를 제시한다. Panel A는 주식수익률의 시계열을 나타내고, Panel B는 네이버 금융 종목토론실 기대 지수의 시계열을 나타낸다. Panel C는 주식수익률과 종목토론실 기대 지수간의 관계를 산포도로 나타내고, Panel D는 종가[5]와 종목토론실 기대 지수 간의 관계를 산포도로 나타낸다. 주식수익률의 경우 전일 대비 가격 변동을 나타내며 백분율(%) 단위이다. 또한, 기대 지수의 경우 1에 가까울수록 긍정을, 0에 가까울수록 부정을 나타내는 연속형 변수이고 주가는 원화 화폐 단위이다.

〈Table 1〉은 회귀분석에 사용된 기대 지수 변수들의 통계량 정보를 나타낸 것이다. Panel A는 종목토론실의 게시글을 통해 도출된 기업별 긍정·부정 기대 지수($OP\_Sent_{i,t}$)의 통계량을 나타낸다. Panel B는 뉴스기사를 통해 도출된 기업별 기대 지수($News\_Sent_{i,t}$)를 나타낸다. Mean은 평균, Std는 표준오차, Median은 중앙값을 나타낸다. 단위근 검정을 위해 두 가지 방법을 활용하였으며, ADF는 augmented Dickey-Fuller 검정 결과를, PP는 Phillips-Perron 검정 결과를 나타낸다. Obs는 관측값의 수이다. 5%와 1% 수준의 유의성을 **와 ***로 각각 표시한다.

### 3.2 게시글 분석

〈Figure 5〉는 분석대상에 해당하는 6개의 기업 중 삼성전자와 셀트리온 두 기업을 선택하여 분석의 기간인 2018년 1월부터 2020년 2월까지의 뉴스기사와 종목토론실 이용자의 게시물 수를 나타낸다. 가로축은 작성 날짜 정보를, 세로축은 작성된 게시물 수를 나타내며 개수 단위이다. Panel A는 네이버 금융 종목토론실에서 삼성전자와 셀트리온에 관해 작성된 게시글 수를 나타내고, Panel B는 삼성전자와 셀트리온에 대해 작성된 뉴스기사의 수를 보여준다. 삼성전자(Samsung)의 경우 점선을 통해 게시글 수를 표현하였고, 셀트리온(Celltrion)은 실선으로 표시하였다. 〈Figure 5〉는 네이버 금융 종목토론실과 네이버 뉴스기사에 대해 전반적인 게시글 수의 차이가 있음을 보여준다. 또한, Panel A에 나타난 종목토론실 게시글의 경우, 투자자가 관심을 가

5) 삼성전자의 경우 2018년 5월 4일 액면분할을 진행하여 불연속적인 변화가 발생하였는데, 이 때문에 종속변수로 주가를 사용하는 것은 어려움이 있어 활용하지 않는다.




〈Figure 4〉 삼성전자 주식수익률과 종목토론실 긍정·부정 기대 지수 시계열 및 산포도

〈Table 1〉 긍정·부정 기대 지수 기술통계량

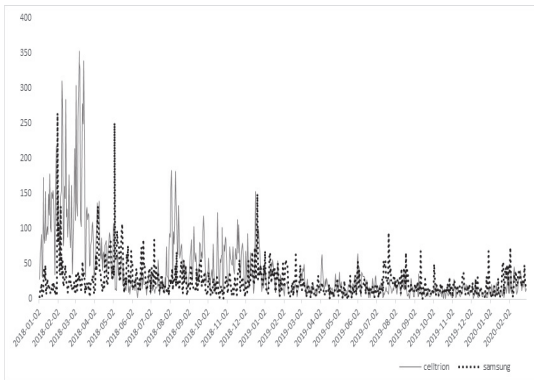
Panel A. 종목토론실 기대 지수 기술통계량 정보

| | Mean | Std | Median | ADF | PP | Obs |
|---|---|---|---|---|---|---|
| $OP\_Sent_{Samsung,t}$ | 0.362 | 0.101 | 0.358 | -4.621*** | -7.911*** | 527 |
| $OP\_Sent_{Celltrion,t}$ | 0.577 | 0.057 | 0.579 | -9.653*** | -13.962*** | 527 |
| $OP\_Sent_{SK\ hynix,t}$ | 0.505 | 0.127 | 0.519 | -6.968*** | -15.362*** | 527 |
| $OP\_Sent_{HYUNDAI,t}$ | 0.358 | 0.110 | 0.351 | -19.342*** | -20.191*** | 527 |
| $OP\_Sent_{NCSOFT,t}$ | 0.624 | 0.167 | 0.627 | -18.879*** | -19.102*** | 527 |
| $OP\_Sent_{KB,t}$ | 0.592 | 0.219 | 0.612 | -17.106*** | -17.113*** | 527 |

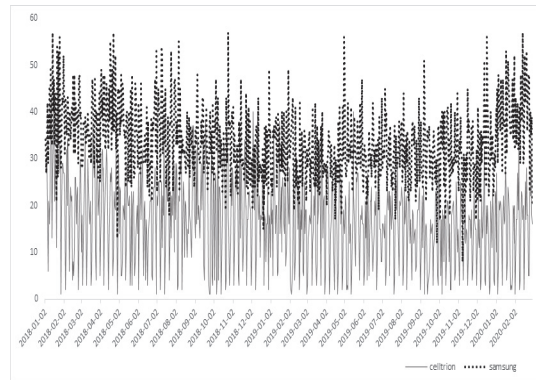
Panel B. 뉴스기사 기대 지수 기술통계량 정보

| | Mean | Std | Median | ADF | PP | Obs |
|---|---|---|---|---|---|---|
| $News\_Sent_{Samsung,t}$ | 0.456 | 0.084 | 0.448 | -3.229** | -10.956*** | 527 |
| $News\_Sent_{Celltrion,t}$ | 0.499 | 0.172 | 0.516 | -12.748*** | -21.624*** | 527 |
| $News\_Sent_{SK\ hynix,t}$ | 0.507 | 0.121 | 0.513 | -8.224*** | -14.904*** | 527 |
| $News\_Sent_{HYUNDAI,t}$ | 0.843 | 0.044 | 0.856 | -13.289*** | -14.666*** | 527 |
| $News\_Sent_{NCSOFT,t}$ | 0.847 | 0.029 | 0.856 | -17.026*** | -17.527*** | 527 |
| $News\_Sent_{KB,t}$ | 0.664 | 0.108 | 0.681 | -11.331*** | -11.562*** | 527 |

Panel A. 종목토론실 게시글 수




Panel B. 뉴스기사 작성 수




〈Figure 5〉 삼성전자/셀트리온 주제로 작성된 일별 게시글 수

질 만한 이슈가 있는 특정 날에는 많은 수의 게시글이 작성되는 반면, 그 외의 상황에서는 작성 게시글의 수가 큰 폭으로 하락하는 것을 알 수 있다. 예를 들어, 삼성전자의 경우, 액면분할이 진행된 2018년 5월 4일 주변에서 게시글 수가 증가한 것을 볼 수 있다. 반면, Panel B의 경우, 뉴스기사는 종목토론실과 비교하면 상대적으로 적은 변동폭을 보이며, 일정한 수준으로 작성되고 있음을 확인할 수 있다.[6)]

네이버 금융 종목토론실에서 삼성전자 기업에 대해 작성 빈도가 높은 일자의 주요 주제는 다음과 같다. 많은 작성 수를 나타낸 상위 3개 날짜의 주요 주제는 액면분할 결정 발표(2018.01.31), 액면분할 상장시작(2018.05.04), 실적부진 전망에 따른 목표주가 하락(2018.12.14)인 것으로 나타났다. 반면 뉴스기사의 작성 빈도에 따른 상위 3개의 주요 주제는 CES 참가에 따른 혁신상 수상(2018.01.14), 유럽 가전 박람회 IFA 참가(2018.04.22), 중소·중견기업에 스마트공장 보급 지원(2018.10.25)인 것으로 나타났다. 셀트리온 기업을 대상으로 작성 빈도가 높은 일자의 상위 3개 주요 주제는 다음과 같다. 종목토론실의 경우, 코스피 200편입 발표(2018.01.14), 코스피 이전 상장(2018.02.10), 코스피 200 편입(2018.03.09)이 작성 빈도가 높은 일자의 주요 주제인 것으로 나타났다. 뉴스기사의 경우, 코스닥 급등에 따른 사이드카 발동(2018.01.12), 제약·바이오주 회계 감리(2018.12.11), 바이오헬스 혁신을 위한 민관 공동 간담회(2019.05.15)가 많이 작성된 상위 3개 일자의 주요 주제인 것으로 나타났다. 위의 내용은 네이버 금융 종목토론실에서 투자자가 작성한 글과 뉴스기사에서 많이 작성한 주요 주제가 다르다는 것을 보여준다. 뉴스기사는 투자자에게 직접 도움이 되는 내용인지 불확실한 내용을 작성 빈도에 따른 상위 주요 주제에 포함하고 있다. 이는 뉴스기사가 일반적으로 주식시장 투자자뿐만 아니라 일반 구독자, 정부 관료, 시민단체 등 다양한 독자들을 대상으로 작성하므로 나타난 특징일 수 있다. 게시글 분석과정을 통해 나타난 다른 성향을 지닌 두 매체 중, 실제 주식수익률에 큰 영향

6) 네이버 금융 종목토론실 게시글의 경우 특정인에 대한 언급 등은 불용어 처리를 함으로써 분석에 반영하지 않는다.

을 미치고 잘 반영할 수 있는 데이터는 무엇인지 나타내고, 그리고 텍스트 자료를 바탕으로 감정분석을 활용하는 방법을 제시하고자 한다.[7)]

### 3.3 학습자료 통계

학습자료는 문장에 포함된 단어를 긍정과 부정으로 구분하여 구성되었다. 〈Table 2〉는 긍정 내용과 부정 내용을 구분하는 데 사용하였던 단어 예시를 나타낸다. 문장에 포함된 단어 정보를 바탕으로 긍정 및 부정 게시글을 분류하며, 하나의 문장에 긍정 분류에 해당하는 단어와 부정 분류에 해당하는 단어가 동시에 포함되어 있으면, 해당 문장은 학습 데이터로 사용하지 않는다.

〈Table 3〉은 기계학습을 위해 사용된 자료의 통계 정보를 나타낸다. 긍정 게시글과 부정 게시글 구분에 따른 전체 글의 개수, 전체 게시글 대비 긍정 및 부정 게시글 비율, 네이버 금융 종목토론실과 뉴스기사의 긍정 및 부정 구분에 따른 게시글의 평균 길이, 그리고 긍정 및 부정 구분에 따른 전체 게시글의 평균 길이 정보를 나타낸다. 긍정 게시물과 부정 게시물의 구성 비율이 큰 차이를 나타내고 있지 않기 때문에 학습 표본데이터의 비율을 조정하는 작업은 진행하지 않았다. 네이버 금융 종목토론실의 게시글 평균 길이는 긍정 게시글의 경우 58글자, 부정 게시글은 72글자였으며, 뉴스기사에서 작성된 글에 비해 훨씬 짧은 것을 확인할 수 있다. 글 길이는 띄어쓰기, 특수문자와 기호를 포함한 글자 수를 나타낸다.

### 3.4 학습 성과 검증

본 논문은 Word2Vec 기법을 통해 타겟 단어와 문맥 단어 정보를 활용하여 단어의 의미를 이해하고 양방향 LSTM 기법을 통해 정방향과 역방향 각각의

〈Table 2〉 긍정/부정 게시글 분류 기준 예시

| 긍정분류 기준 단어 | 부정분류 기준 단어 |
|---|---|
| 주가 상승, 기대, 호황, 흑자, 상장, 출시, 합병, 인수, 급등, 풀매수, 순매수, 홀딩, 돌파, 강세, 재매입, 상한가, 역대, 강력매수, 오름세, 계약, 체결, 이익, 폭등, 쓸어담네, 폭등, 매출증가, 순풍, 수주, 랠리, 정점, 인기, 선방, 안정, 성공, 유상증자, 개발, 도약 | 주가 하락, 조정, 공매도, 불황, 적자, 손절, 급락, 악재, 고전, 탈출, 약세, 하한가, 반토막, 급감, 여파, 퇴출, 중단, 해지, 하락, 과징금, 매각, 실패, 하향, 폭락, 거품, 붕괴, 실망, 위기, 음봉, 악화, 의혹, 파업, 위반, 셧다운, 폐쇄, 공포, 패닉, 손실 |

〈Table 3〉 학습에 사용된 게시글 정보

| | 표본 개수 | 비 율 | 평균 글 길이 (토론) | 평균 글 길이 (뉴스) | 종합 글 평균 길이 |
|---|---|---|---|---|---|
| 긍 정 | 309,931 | 48.39% | 58 | 607 | 362 |
| 부 정 | 330,526 | 51.61% | 72 | 750 | 414 |
| Total | 640,457 | 100% | 65 | 678 | 389 |

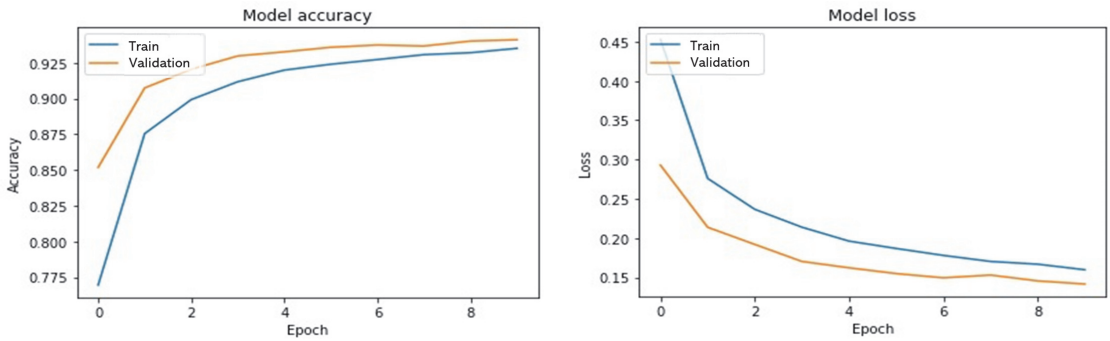
7) 날짜별로 정리된 이슈에 관한 자세한 정보는 개인정보 및 민감한 정보가 포함되어 있을 수 있어 본문에서는 공개하지 않는다. 다만, 합리적인 이유로 저자에게 요청 시 제공할 수 있다.

시점 정보를 반영하였다. 이렇게 학습된 모형은 표본자료에 대해 긍정 및 부정 분류를 시행하게 된다. 모형의 학습성과 및 구성에 대한 평가는 〈Table 4〉와 〈Figure 6〉을 통해 확인할 수 있다. 〈Table 4〉는 특정 단어 입력 시, 도출되는 유사 단어의 정보를 보여준다. 임의로 6개의 단어를 입력하였으며, '세이프가드,' '팹리스'와 같은 전문단어에 대한 학습의 정도를 가늠할 수 있다. 유사 단어는 각 입력어와 유사한 의미로 사용된 단어를 나타낸다. similarity는 0과 1사이의 값을 갖고 도출된 각각의 단어가 입력어에 대해서 갖는 벡터공간에서의 거리를 나타내며 높을수록 입력어와 유사한 단어임을 의미한다. 감정분류의 정확도를 높이기 위해서는 간단한 단어뿐만 아니라 전문성을 요구하거나 어려운 내용이 담긴 단어를 분석하고 이해하는 능력이 필요하다. 높은 학습성과를 위해서는 기법(모형)의 종류도 중요하지만, 학습자료의 양과 품질이 중요시된다. 연구에서 진행된 학습자료가 간단한 내용을 담은 문장만을 이용한 것이 아닌 어려운 내용이 담긴 뉴스기사나 투자자들의 스크랩 인용문 등을 모두 활용하였기 때문에 단어 임베딩이 잘 이루어진 것으로 판단된다.

신경망 학습을 위해 훈련(training) 자료와 검증(validation) 자료는 80:20으로 무작위로 나뉘었고, 추가 학습에도 학습성과 개선이 이루어지지 않는 경우, 조기 종료된다. 〈Figure 6〉은 학습을 통해 얻어지는 지표인 Train과 모형이 얼마나 잘 구성되었는지를 나타내는 지표인 Validation을 나타내고 모형 학습성과를 보여준다. 학습과 테스트를 시행한 후 채점을 한 횟수를 의미하는 훈련 횟수(epoch)가 늘어날수록 학습과 검증의 정확도(accuracy) 값이 개선되는 것을 보여준다. 동시에, 훈련 횟수 증가에 따라 정답에 대한 오차를 나타내는 모형의 손실(loss) 값이 감소한다. 또한, 〈Figure 6〉은 인공신경망 모형에서 자주 발생하는 과잉적합 문제 발생 여부를 보여준다. 과잉적합은 학습자료를 대상으로는 높은 정확도를 나타내지만, 검증 자료에 대해서는

〈Table 4〉 Word2Vec 학습 결과 확인

| 세이프가드 | | 매출 | | 연구 | |
|---|---|---|---|---|---|
| 유사 단어 | similarity | 유사 단어 | similarity | 유사 단어 | similarity |
| 관세부과 | 0.84578 | 매출액 | 0.81356 | 신약개발 | 0.71289 |
| 반덤핑 | 0.82022 | 출하량 | 0.75597 | 개발 | 0.70466 |
| 긴급수입 제한조치 | 0.81713 | 판매량 | 0.73711 | 기술개발 | 0.69440 |
| 수입제한 | 0.79857 | 순이익 | 0.73654 | 학술 | 0.68746 |
| 화이트리스트 | 0.76917 | 영업이익 | 0.72918 | 연구개발 | 0.68699 |
| **금리** | | **세탁기** | | **팹리스** | |
| 유사 단어 | similarity | 유사 단어 | similarity | 유사 단어 | similarity |
| 기준금리 | 0.92846 | 냉장고 | 0.82664 | 파운드리 | 0.72425 |
| 정책금리 | 0.81300 | 건조기 | 0.80609 | 외주 | 0.68308 |
| 콜금리 | 0.80728 | 전자레인지 | 0.75841 | 첨단기술 | 0.66461 |
| 대출금리 | 0.73506 | 드럼세탁기 | 0.73656 | 의료기기 | 0.66315 |
| 예금금리 | 0.73111 | 가전제품 | 0.73191 | CDMO | 0.65895 |




〈Figure 6〉 학습 횟수에 따른 모형 학습성과 및 정확도

낮은 성과를 보여 학습과 검증값 사이에 디커플링(decoupling)의 모습을 보인다. 〈Figure 6〉의 내용은 본 연구에서 진행한 모형에서 과잉적합 문제가 발생하지 않았으며, 기계학습 결과에 대해 신뢰할 수 있음을 나타낸다.

## Ⅳ. 회귀분석 결과

텍스트 기반의 긍정·부정 기대 지수가 주식수익률에 대한 설명력을 갖는지 확인한 결과는 다음과 같다. 〈Table 5〉는 네이버 종목토론실 자료를 바탕으로 만들어진 기대 지수($OP\_Sent_{i,t}$)가 설명변수이고 주식수익률($Stock\_Return_{i,t}$)이 종속변수인 회귀분석의 추정결과를 나타낸다. 〈Table 6〉은 네이버 뉴스기사를 바탕으로 만들어진 기대 지수($News\_Sent_{i,t}$)가 설명변수인 회귀분석 결과를 나타낸다. 모형 (1)은 통제변수가 없는 경우의 회귀분석을 나타내고, 모형 (2)는 금융시장 거시경제 변수를 통제한 후 기대 지수가 기업 주식수익률을 설명하는지를 살펴본다. 모형 (3)은 Fama-French 3요인 변수를 통제변수로 활용하여, 기대 지수의 주식수익률에 대한 설명력을 회귀분석을 통하여 추정한다. $\triangle CD(91days)_t$는 CD금리로 무위험 단기 이자율의 일별 변화량을 나타내고, $\triangle Term\ Spread_t$는 국채(10년) 금리에서 국채(3년) 금리를 차감하여 나타낸 기간 스프레드의 일별 변화량, $\triangle Credit\ Spread_t$는 회사채(BBB-) 금리에서 회사채(AA-)금리를 차감하여 나타낸 신용 스프레드의 일별 변화량, $VKOSPI_t$는 옵션의 내재변동성으로 추정한 코스피 200 변동성 지수 수익률, $KOSPI_t$는 코스피 종합주가지수 수익률을 나타낸다. $MKT_t$는 무위험 자산 대비 시장 초과수익률을 나타내고, $SMB_t$는 대형주 대비 소형주 초과수익률을 나타낸다. $HML_t$는 성장주 대비 가치주 초과수익률을 나타낸다. $Adj.R^2$는 회귀모형의 설명력을 나타낸다. 10%, 5%, 1% 수준의 유의성을 *, **, ***로 표시한다. 네이버 금융 종목토론실 글을 통해 도출한 기대 지수는 기업별 주식수익률에 대한 회귀분석에서, 모두 유의한 양(+)의 계수 값을 갖는다. 이는 종목토론실에서 추출한 기대 지수가 개별종목의 주식수익률을 잘 설명한다는 것을 의미한다. 금융시장 거시경제 변수를 통제변수로 활용한 모형 (2)의 결과보다 Fama-French 3요인 변수를 통제하여 분석한 모형 (3)이 주식수익률에 대해 더 높은 설명력을 갖고 있음을 보여준다.

〈Table 5〉 기업별 주식수익률과 종목토론실 긍정·부정 기대 지수의 회귀분석 결과

| | Samsung | | | Celltrion | | | SK hynix | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| $Intercept$ | -0.012*** | -0.004** | -0.004** | -0.093*** | -0.081*** | -0.081*** | -0.021*** | -0.010*** | -0.009*** |
| | (-4.56) | (-2.31) | (-2.55) | (-7.08) | (-6.42) | (-6.62) | (-5.46) | (-2.95) | (-2.92) |
| $OP\_Sent_{i,t}$ | 0.033*** | 0.013*** | 0.013*** | 0.161*** | 0.141*** | 0.141*** | 0.043*** | 0.022*** | 0.021*** |
| | (4.83) | (2.80) | (3.08) | (7.11) | (6.50) | (6.66) | (5.79) | (3.38) | (3.33) |
| $\triangle CD(91days)_t$ | | -0.030 | | | 0.032 | | | -0.041 | |
| | | (-0.94) | | | (0.38) | | | (-0.77) | |
| $\triangle Term\ Spread_t$ | | -0.057* | | | 0.013 | | | -0.047 | |
| | | (-1.66) | | | (0.15) | | | (-0.80) | |
| $\triangle Credit\ Spread_t$ | | -0.058 | | | 0.401 | | | -0.000 | |
| | | (-0.32) | | | (0.85) | | | (-0.00) | |
| $VKOSPI_t$ | | 0.005 | | | -0.004 | | | 0.001 | |
| | | (0.79) | | | (-0.23) | | | (0.11) | |
| $KOSPI_t$ | | 1.355*** | | | 1.034*** | | | 1.391*** | |
| | | (22.55) | | | (6.49) | | | (13.43) | |
| $MKT_t$ | | | 1.215*** | | | 0.902*** | | | 1.251*** |
| | | | (24.90) | | | (6.55) | | | (13.96) |
| $SMB_t$ | | | -0.423*** | | | 0.303** | | | -0.287*** |
| | | | (-8.91) | | | (2.25) | | | (-3.34) |
| $HML_t$ | | | -0.387*** | | | -0.764*** | | | -0.523*** |
| | | | (-6.43) | | | (-4.47) | | | (-4.78) |
| $Adj.R^2$ | 0.0409 | 0.5679 | 0.6419 | 0.0863 | 0.1732 | 0.2159 | 0.0583 | 0.3464 | 0.3842 |
| | HYUNDAI | | | NCSOFT | | | KB | | |
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| $Intercept$ | -0.010*** | -0.008*** | -0.008*** | -0.007** | -0.007** | -0.008** | -0.007*** | -0.005*** | -0.005*** |
| | (-3.67) | (-3.35) | (-3.19) | (-2.25) | (-2.25) | (-2.55) | (-3.31) | (-2.88) | (-2.84) |
| $OP\_Sent_{i,t}$ | 0.026*** | 0.023*** | 0.022*** | 0.013*** | 0.013*** | 0.014*** | 0.010*** | 0.008*** | 0.008*** |
| | (3.70) | (3.52) | (3.47) | (2.62) | (2.71) | (2.96) | (3.10) | (2.77) | (2.85) |
| $\triangle CD(91days)_t$ | | 0.078 | | | 0.035 | | | -0.024 | |
| | | (1.61) | | | (0.63) | | | (-0.57) | |
| $\triangle Term\ Spread_t$ | | -0.058 | | | -0.012 | | | 0.015 | |
| | | (-1.10) | | | (-0.19) | | | (0.32) | |
| $\triangle Credit\ Spread_t$ | | 0.243 | | | -0.051 | | | 0.088 | |
| | | (0.89) | | | (-0.16) | | | (0.38) | |
| $VKOSPI_t$ | | 0.001 | | | -0.012 | | | 0.002 | |
| | | (0.07) | | | (-1.16) | | | (0.21) | |
| $KOSPI_t$ | | 0.827*** | | | 0.497*** | | | 0.798*** | |
| | | (8.99) | | | (4.74) | | | (10.08) | |
| $MKT_t$ | | | 0.899*** | | | 0.442*** | | | 0.854*** |
| | | | (11.37) | | | (4.84) | | | (12.37) |
| $SMB_t$ | | | -0.185** | | | -0.201** | | | -0.163** |
| | | | (-2.38) | | | (-2.23) | | | (-2.41) |
| $HML_t$ | | | 0.544*** | | | -0.465*** | | | 0.307*** |
| | | | (5.51) | | | (-4.07) | | | (3.56) |
| $Adj.R^2$ | 0.0236 | 0.181 | 0.2344 | 0.0111 | 0.071 | 0.1054 | 0.0161 | 0.2153 | 0.247 |

〈Table 6〉 기업별 주식수익률과 뉴스기사 긍정·부정 기대 지수의 회귀분석 결과

| | Samsung | | | Celltrion | | | SK hynix | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| $Intercept$ | -0.012*** | -0.003 | -0.003 | 0.001 | 0.003 | 0.003 | -0.023*** | -0.006* | -0.007** |
| | (-3.21) | (-1.32) | (-1.10) | (0.14) | (0.75) | (0.75) | (-5.62) | (-1.77) | (-2.04) |
| $News\_Sent_{i,t}$ | 0.028*** | 0.009 | 0.007 | -0.001 | -0.005 | -0.006 | 0.046*** | 0.015** | 0.016** |
| | (3.32) | (1.61) | (1.41) | (-0.14) | (-0.64) | (-0.76) | (5.93) | (2.14) | (2.39) |
| $\triangle CD(91days)_t$ | | -0.022 | | | 0.032 | | | -0.033 | |
| | | (-0.68) | | | (0.37) | | | (-0.60) | |
| $\triangle Term\ Spread_t$ | | -0.056 | | | 0.019 | | | -0.048 | |
| | | (-1.60) | | | (0.20) | | | (-0.80) | |
| $\triangle Credit\ Spread_t$ | | -0.120 | | | 0.334 | | | -0.049 | |
| | | (-0.67) | | | (0.69) | | | (-0.16) | |
| $VKOSPI_t$ | | 0.003 | | | 0.002 | | | 0.002 | |
| | | (0.56) | | | (0.10) | | | (0.17) | |
| $KOSPI_t$ | | 1.362*** | | | 1.169*** | | | 1.400*** | |
| | | (22.48) | | | (7.09) | | | (13.21) | |
| $MKT_t$ | | | 1.232*** | | | 1.019*** | | | 1.246*** |
| | | | (25.26) | | | (7.14) | | | (13.49) |
| $SMB_t$ | | | -0.420*** | | | 0.313** | | | -0.299*** |
| | | | (-8.80) | | | (2.23) | | | (-3.45) |
| $HML_t$ | | | -0.385*** | | | -0.758*** | | | -0.535*** |
| | | | (-6.31) | | | (-4.26) | | | (-4.86) |
| $Adj.R^2$ | 0.0188 | 0.5636 | 0.6367 | -0.0019 | 0.1067 | 0.1501 | 0.0611 | 0.3378 | 0.3779 |
| | HYUNDAI | | | NCSOFT | | | KB | | |
| | (1) | (2) | (3) | (1) | (2) | (3) | (1) | (2) | (3) |
| $Intercept$ | -0.044*** | -0.030** | -0.025* | 0.005 | 0.011 | 0.008 | -0.012*** | -0.010** | -0.008** |
| | (-2.94) | (-2.13) | (-1.87) | (0.20) | (0.43) | (0.35) | (-2.80) | (-2.53) | (-2.02) |
| $News\_Sent_{i,t}$ | 0.052*** | 0.035** | 0.030* | -0.005 | -0.011 | -0.009 | 0.017*** | 0.014** | 0.011** |
| | (2.92) | (2.13) | (1.90) | (-0.16) | (-0.39) | (-0.31) | (2.65) | (2.44) | (1.97) |
| $\triangle CD(91days)_t$ | | 0.077 | | | 0.033 | | | -0.026 | |
| | | (1.57) | | | (0.59) | | | (-0.63) | |
| $\triangle Term\ Spread_t$ | | -0.063 | | | -0.017 | | | 0.029 | |
| | | (-1.17) | | | (-0.29) | | | (0.62) | |
| $\triangle Credit\ Spread_t$ | | 0.287 | | | -0.129 | | | 0.142 | |
| | | (1.04) | | | (-0.41) | | | (0.60) | |
| $VKOSPI_t$ | | 0.002 | | | -0.011 | | | 0.003 | |
| | | (0.25) | | | (-1.11) | | | (0.38) | |
| $KOSPI_t$ | | 0.832*** | | | 0.501*** | | | 0.801*** | |
| | | (8.96) | | | (4.74) | | | (10.10) | |
| $MKT_t$ | | | 0.899*** | | | 0.445*** | | | 0.857*** |
| | | | (11.23) | | | (4.82) | | | (12.36) |
| $SMB_t$ | | | -0.169** | | | -0.184** | | | -0.148** |
| | | | (-2.16) | | | (-2.02) | | | (-2.18) |
| $HML_t$ | | | 0.557*** | | | -0.458*** | | | 0.304*** |
| | | | (5.60) | | | (-3.97) | | | (3.50) |
| $Adj.R^2$ | 0.0141 | 0.1688 | 0.2221 | -0.0019 | 0.0581 | 0.0906 | 0.0113 | 0.2127 | 0.2409 |

한편, 〈Table 6〉의 결과는 뉴스기사를 기반으로 기대지수를 도출하면, 셀트리온, 엔씨소프트의 기대지수는 주식수익률을 유의하게 설명하지 못하는 것으로 나타난다. 삼정전자 기대 지수의 경우, 통제변수를 포함하여 분석하면 주식수익률에 대한 설명력이 사라진다.

주식수익률에 대한 기대 지수의 추정된 계수가 유의한 양(+)의 값을 갖는 것은 전일 종가 발표 이후 당일 종가 발표까지 기대 지수가 증가할 때, 주식수익률 역시 증가한다는 것을 의미한다. 〈Table 5〉와 〈Table 6〉을 통해 살펴보면, 네이버 금융 종목토론실 게시글을 활용하여 구축한 기대 지수가 네이버 뉴스기사를 기반으로 도출된 기대 지수보다 모든 기업에서 더 높은 설명력과 유의성을 나타낸다. 전반적으로 종목토론실 자료를 기반으로 구축한 기대 지수가 뉴스기사에 기반한 기대 지수보다 주식수익률에 대한 설명력이 높다. 이는 〈Figure 5〉에 나타난 두 매체의 특성 차이 때문일 수 있다. 네이버 금융 종목토론실 이용자는 해당 기업에 관해 이목을 끌만한 투자 이슈가 생겼을 때 많은 글을 작성하는 경향이 있다. 투자자가 직접 작성한 글은 투자자들의 관심 혹은 집중하고 있는 내용을 나타내어 정확한 긍정 및 부정 표현을 포함하고 있을 가능성이 크다. 반면, 네이버 뉴스기사는 다양한 독자들을 상대로 작성하게 되어 투자자에게 필요하지 않은 정보를 포함할 수 있다. 따라서 투자와 직접 연관되는 내용의 비중 감소는 감정분류 과정을 어렵게 만들 수 있다. 또한, 〈Table 3〉에서 확인한 바와 같이 두 매체의 게시글 길이 차이가 결과에 영향을 미쳤을 수 있다. 네이버 금융 종목토론실 글의 평균 길이는 65글자(긍정은 58글자, 부정은 72글자)인 반면, 뉴스기사의 경우 평균 글의 길이가 678글자(긍정은 607글자, 부정은 750글자)인 것으로 나타났다. 글의 길이가 길어지면 긍정 및 부정 표현이 불분명해져 기대 지수를 도출에 방해가 될 가능성이 크다. 마지막으로, 언론 매체의 특성상 객관적이고 감정이 배제된 글쓰기가 주로 이루어지기 때문에(Kim and Lee, 2018), 뉴스기사 자료를 바탕으로 긍정 및 부정 글을 분류하는 데 어려움이 있을 수 있다.

본 연구를 통해 투자자의 심리기대가 금융 자산의 수익률에 미치는 영향을 간접적으로 확인할 수 있었으며, 국내 기업을 대상으로 투자자 긍정·부정 기대 지수를 도출하기에 더 적합한 자료는 투자자들이 직접 작성하여 남긴 종목토론실 글임을 알 수 있었다. 그러나 뉴스기사도 기대 지수를 도출하는 과정에서 필요한 양질의 학습자료를 제공할 수는 있다. 자연어처리에 해당하는 감정분류 성능은 양질의 글을 학습 데이터로 활용할 수 있는지에 따라 영향을 받는데, 투자자가 작성하는 내용만으로 전문 용어나 시사 단어 등에 대한 학습 제공량이 부족할 수 있으므로, 뉴스기사 역시 기계학습에 기여할 가능성을 배제할 수는 없다.

## Ⅴ. 결 론

투자에 영향을 주는 정보들은 경제 주체들의 투자심리에 영향을 끼친다. 일반적으로 수치형 정보를 바탕으로 자료 분석이 진행되지만, 주식시장에 참가하는 투자자의 게시글과 뉴스기사 역시 투자심리에 영향을 미치는 중요한 정보를 포함하고 있을 가능성이 크다. 이러한 이유로 텍스트 자료에 대한 자연어처리를 통해 글에 나타나 있는 의미를 분석하고 패턴을 연구하려는 시도가 늘고 있으며, 본 연구는 주요 기업 6곳을 선정하여 해당 기업을 대상으로 작성

되는 텍스트 자료를 활용한 감정분석을 시행하였다.

단어의 의미를 학습하는 단어 임베딩 방법은 문맥 정보를 활용하여 학습하는 기법인 Word2Vec을 적용하였고, 유사한 의미를 나타내는 단어는 밀집 벡터공간에서 큰 값을 갖도록 구성하였다. 양방향 LSTM 기법을 적용하여 글 속에 존재하는 단어의 의미를 별개로 해석하는 것이 아닌 정방향과 역방향으로 각각 학습시켜 기계학습의 성능을 높이고, 어텐션 기법을 통해 효율적인 학습이 이루어지도록 하였다. 학습된 모형을 통해, 네이버 금융 종목토론실과 네이버 뉴스기사 자료를 감정분류하고 각각의 기대 지수를 도출하였다. 본 연구는 금융시장 거시경제 변수와 Fama-French 3요인 변수를 통제한 후, 주식수익률에 대한 도출된 개별기업의 기대 지수의 설명력을 분석하였다. 회귀분석 결과, 주식 투자자들이 작성한 네이버 금융 종목토론실 텍스트 정보가 네이버 뉴스기사보다 높은 설명력과 유의성을 갖고 있음을 확인할 수 있었다. 그러나 뉴스기사 역시 기계학습에 기여할 가능성을 배제할 수는 없으므로, 추후 연구에서는 금융 종목토론실 자료와 뉴스기사를 혼합하여 정교하게 활용한다면 신경망 기법의 성능을 높일 수 있을 것이다.

본 연구는 뉴스와 소셜 미디어 텍스트 자료 분석에 관한 알고리즘 설계 방법을 개괄적으로 소개하고, 기계학습에 대한 방법론을 국내 주식시장에 적용한 융합연구라는데 의의가 있다. 특히, 투자자가 직접 작성한 텍스트 정보를 활용하여 기대 지수를 구축하며, 이를 통해 주식수익률을 설명한 기여가 있다. 추후 연구에서는 정규시간 외의 시간대에 작성된 텍스트 정보를 기반으로 주가 등락을 예측하거나, 감정분류 성과 향상을 위해 여러 기법을 활용하여 분석한다면 확장이 가능할 것이다.

## 참고문헌


Bahdanau, D., K. Cho, and Y. Bengio (2016), "Neural machine translation by jointly learning to align and translate," *arxiv. https://arxiv.org/abs/1409.0473v7*

Baker, M., J. Wang, and J. Wurgler (2008), "How does investor sentiment affect the crosssection of stock returns?," *Journal of Investment Management*, 6(2), pp.57-72.

Baker, M. and J. Wurgler (2006), "Investor sentiment and the cross section of stock returns," *Journal of Finance*, 61(4), pp.1645-1680.

Baker, M. and J. Wurgler (2007), "Investor sentiment in the stock market," *Journal of Economic Perspectives*, 21(2), pp.129-152.

Behrendt, S. and A. Schmidt (2018), "The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility," *Journal of Banking and Finance*, 96, pp. 355-367.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin (2003), "A neural probabilistic language model," *Journal of Machine Learning Research*, 3, pp.1137-1155.

Bodie, Z., A. Kane, and A. Marcus (2021), *Investments*, 12th Edition, New York: McGraw Hill.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017), "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, 5, pp.135-146.

Bollen, J., H. Mao, and X. Zeng (2011), "Twitter mood predicts the stock market," *Journal of*

*Computational Science*, 2(1), pp.1-8.

Chun, S. (2020), "Predicting Korean stock market return with financial and macro variables: Focusing on in-sample and out-of-sample tests," *Journal of Insurance and Finance*, 31(1), pp.87-113.

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1989), "The size and incidence of the losses from noise trading," *Journal of Finance*, 44(3), pp.681-696.

Elman, J. L. (1990), "Finding structure in time," *Cognitive Science*, 14(2), pp.179-211.

Fama, E. F. and K. R. French (1992), "The cross-section of expected stock returns," *Journal of Finance*, 47(2), pp.427-465.

Fama, E. F. and K. R. French (1993), "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33 (1), pp.3-56.

Goodfellow, I., Y. Bengio, and A. Courville (2016), *Deep learning*, Cambridge: MIT Press.

Graves, A., and J. Schmidhuber (2005), "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, 18(5-6), pp.602-610.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov (2012), "Improving neural networks by preventing co-adaptation of feature detectors," *arxiv. https://arxiv.org/abs/1207.0580*

Hirschberg, J. and C. D. Manning (2015), "Advances in natural language processing," *Science*, 349(6245), pp.261-266.

Hjalmarsson, E. (2010), "Predicting global stock returns," *Journal of Financial and Quantitative Analysis*, 45(1), pp.49-80.

Hochreiter, S., Y. Bengio, P. Frasconi, and J. Schmidhuber (2001), "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in J. Kolen and S. Kremer (Eds.), *A Field Guide to Dynamical Recurrent Networks,* IEEE Press. pp.237-244.

Hochreiter, S. and J. Schmidhuber (1997), "Long short-term memory," *Neural Computation*, 9(8), pp.1735-1780.

Hu, G. X., C. Chen, Y. Shao, and J. Wang (2019), "Fama - French in China: Size and value factors in Chinese stock returns," *International Review of Finance*, 19(1), pp.3-44.

Huang, C., S. Simpson, D. Ulybina, and A. Roitman (2019), "News-based sentiment indicators," *IMF Working Papers No. 19/273*.

Huang, T. L. (2019), "Is the Fama and French five-factor model robust in the Chinese stock market?" *Asia Pacific Management Review*, 24(3), pp.278-289.

Kalyani, J., H. N. Bharathi, and R. Jyothi (2016), "Stock trend prediction using news sentiment analysis," *arxiv. https://arxiv.org/abs/1607.01958*

Kam, H. and Y. Shin (2017), "The impact of macro-economic variables on stock returns in Korea," *Korean Journal of Business Administration*, 30(1), pp.33-52.

Kang, H. and J. Yang (2019), "Optimization of Word2Vec models for Korean word embeddings," *Journal of Digital Contents Society*, 20(4), pp.825-833.

Kim, E. J. and H. S. Lee (2013), "A study on alternative design research model using unstructured online data: Through design ethnography methodology," *Design Convergence Study*, 12(5), pp.205-223.

Kim, H., H. Cho, and D. Ryu (2020), "Corporate default predictions using machine learning: Literature review," *Sustainability*, 12(16), 6325.

Kim, H., D. Ryu, and H. Cho (2019), "Corporate default predictions and machine learning," *Korean Journal of Financial Engineering*, 18(3), pp.131-152.

Kim, J. S., D. Ryu, and S. W. Seo (2014), "Investor sentiment and return predictability of disagreement," *Journal of Banking and Finance*, 42, pp.166-178.

Kim, K. and S. Lee (2018), "News audiences perceptual biases and assessment of news fairness: An analysis of the influences of trust for media, message bias, self-categorization, and self-enhancement," *Communication Theories*, 14 (3), pp. 145-198.

Kim, K. and D. Ryu (2021), "Does sentiment determine investor trading behaviour?" *Applied Economics Letters*, Forthcoming.

Kim, K., D. Ryu, and H. Yang (2018), "Investor sentiment indices and the cross-section of stock returns of individual firms," *Korean Management Review*, 47(5), pp.1231-1260.

Kim, K., D. Ryu, and J. Yu (2021), "Do sentiment trades explain overconfidence around analyst recommendation revisions?" *Research in International Business and Finance*, 56, 101376.

Kim, N. and Y.-W. Lee (2016), "Machine learning approaches to corn yield estimation using satellite images and climate data: A case of Iowa state," *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, 34(4), pp.383-390.

Kim, S., Y. Lee, J. Shin, and K. Y. Park (2019), "Text mining for economic analysis," *BOK Working Paper 2019-18*.

Kim, Y., S. Jeong, and S. Lee (2014), "A study on the stock market prediction based on sentiment analysis of social media," *Entrue Journal of Information Technology*, 13(3), pp.59-70.

Kingma, D. P. and J. Ba (2017), "Adam: A method for stochastic optimization," *arxiv. https://arxiv.org/abs/1412.6980v9*

Ko, S.-J., H.-Y. Yun, and D.-M. Shin (2018), "Electronic demand data prediction using bidirectional long short term memory networks," *Journal of Software Assessment and Valuation*, 14(1), pp.33-40.

Lee, D. H., H. G. Kang, S. H. Kim, and C. M. Lee (2013), "Autocorrelation analysis of the sentiment with stock information appearing on big-data," *Korean Journal of Financial Engineering*, 12(2), pp.79-96.

Lee, Y. (2018), "Introduction to eKoNLPy: Korean NLP python packagage for economic analysis," Available at https://github.com/entelecheia/eKoNLPy.

Levy, O. and Y. Goldberg (2014), "Linguistic regularities in sparse and explicit word representations," *Proceedings of the 18th Conference on Computational Natural Language Learning*, pp.171-180.

Liu, H. and Y.-C. Gao (2019), "The impact of corporate lifecycle on Fama-French three-factor model," *Physica A*, 513, pp. 390-398.

Luong, M.-T., H. Pham, and C. D. Manning, (2015), "Effective approaches to attention-based neural machine translation," *arxiv. https://arxiv.org/abs/1508.04025v5*

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013), "Distributed representations

of words and phrases and their compositionality," *Advances in Neural Information Processing Systems 26*.

Nair, V. and G. E. Hinton (2010), "Rectified linear units improve restricted Boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning*, pp.807-814.

O'Leary, D. E. (2013), "Artificial intelligence and big data," *IEEE Intelligent Systems*, 28(2), pp.96-99.

Park, E. L. 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, K. S., E. J. Lee, and I. M. Lee (2003), "Determinants of dividend policy of Korean firms," *Asian Review of Financial Research*, 16(2), pp.195-229.

Renault, T. (2017), "Intraday online investor sentiment and return patterns in the U.S. stock market," *Journal of Banking and Finance*, 84, pp. 25-40.

Russell, S. J. and P. Norvig (2020), *Artificial intelligence: A modern approach*, 4th Edition, Pearson.

Ryu, D., H. Kim, and H. Yang (2017), "Investor sentiment, trading behavior and stock returns," *Applied Economics Letters*, 24(12), pp.826-830.

Schrimpf, A. (2010), "International stock return predictability under model uncertainty," *Journal of International Money and Finance*, 29(7), pp.1256-1282.

Seok, S. I., H. Cho, and D. Ryu (2019a), "Firm-specific investor sentiment and the stock market response to earnings news," *North American Journal of Economics and Finance*, 48, pp.221-240.

Seok, S. I., H. Cho, and D. Ryu (2019b), "Firm-specific investor sentiment and daily stock returns," *North American Journal of Economics and Finance*, 50, 100857.

Shapiro, A. H., M. Sudhof, and D. J. Wilson (2021), "Measuring news sentiment," *Journal of Econometrics*, Forthcoming.

Smales, L. A. (2016), "Trading behavior in S&P 500 index futures," *Review of Financial Economics*, 28, pp.46-55.

Smales, L. A. (2020), "News sentiment as an explanation for changes in the VIX futures basis," *Journal of Investing*, 29(4), pp.92-102.

Stambaugh, R. F., J. Yu, and Y. Yuan (2012), "The short of it: Investor sentiment and anomalies," *Journal of Financial Economics*, 104(2), pp. 288-302.

Stein, J. C. (1987), "Informational externalities and welfare reducing speculation," *Journal of Political Economy*, 95(6), pp.1123-1145.

Sumathy, K. L. and M. Chidambaram (2013), "Text mining: Concepts, applications, tools and issues - An overview," *International Journal of Computer Applications*, 80(4), pp.29-32.

Yildirim, O. (2018), "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification," *Computers in Biology and Medicine*, 96, pp.189-202.

• The author Juhwa Lee graduated from the School of Business Administration, College of Business & Economics, Chung-Ang University. He has got a Master's degree in Economics at Sungkyunkwan University. His current research interests are machine learning, big data analysis, financial management, and behavioral finance.

• The author Doojin Ryu is a full/tenured professor of economics at Sungkyunkwan University. He graduated from Seoul National University (School of Electrical Engineering), and has got a Ph.D. degree at KAIST. He was a research fellow at the National Pension Service, an assistant professor at Hankuk University of Foreign Studies, and a full/tenured professor at Chung-Ang University. Prof. Ryu is currently an editor of Investment Analysts Journal (SSCI) and a subject editor of Emerging Markets Review (SSCI), Journal of Multinational Financial Management (SSCI), and Emerging Markets Finance & Trade (SSCI). He is an editorial board member of Journal of Futures Markets (SSCI) and Asian Business & Management (SSCI).