한국빅데이터학회지
제3권 제1호, 2018, pp. 13-20

Word2Vec을 활용한 뉴스 기반 주가지수 방향성 예측용 감성 사전 구축

News based Stock Market Sentiment Lexicon Acquisition Using Word2Vec

김다예 · 이영인†

연세대학교 정보대학원

요 약

주식 시장에 대한 예측은 오랜 기간 많은 이들의 꿈이었다. 하지만 수많은 노력에도 불구하고 주식 시장을 정확하게 예측하기란 쉬운 일이 아니었다. 본 연구는 주식 시장의 방향성에 주목하여 이 방향성을 예측할 수 있는 감성사전을 구축하는 새로운 방법을 제시한다. 이를 위해 2015년 1월 1일부터 2017년 12월 31일까지 3년간의 증시 뉴스 25,000여 건의 데이터를 수집하여, 문맥을 고려하기 위한 Word2Vec을 적용하였다. 이를 바탕으로 뉴스에 감성분석을 실시하여 KOSPI 종가 지수를 예측해 보았다.

■ 중심어 : 주가예측, Word2Vec, 자연어처리, 텍스트마이닝, 뉴스, 감성사전

Abstract

Stock market prediction has been long dream for researchers as well as the public. Forecasting ever-changing stock market, though, proved a Herculean task. This study proposes a novel stock market sentiment lexicon acquisition system that can predict the growth (or decline) of stock market index, based on economic news. For this purpose, we have collected 3-year's economic news from January 2015 to December 2017 and adopted Word2Vec model to consider the context of words. To evaluate the result, we performed sentiment analysis to collected news data with the automated constructed lexicon and compared with closings of the KOSPI (Korea Composite Stock Price Index), the South Korean stock market index based on economic news.

■ Keyword : Stock Prediction, Word2Vec, Natural Language Processing, Text Mining, News, Sentiment Lexicon

Ⅰ. 서 론

주식 시장에 대한 예측은 오랜 기간 많은 이들의 꿈이었다. 이에 주식 시장을 예측하기 위한 다양한 시도가 있었지만 워낙 변화무쌍한 주식 시장을 제대로 예측하기란 어려운 일이었다. 정보 기술의 발달로 4차 산업혁명 시대가 개막하고 각종 빅데이터 기술이 발달하고 있는 오늘날에도 발전된 기술을 이용해 주식 시장을 예측하려는 활발한 시도가 있어왔다. 하지만 수많은 노력에도 불구하고 여전히 주식 시장을 정확하게 예측하기란 쉬운 일이 아니었다.

초창기의 주식 시장 예측 연구는 주가가 뉴스 같은 새로운 정보에 크게 영향을 받는다는 EMH (Efficient Market Hypothesis)를 기반으로 했으며 [9], 뉴스의 무작위성 때문에 정확도가 50% 이상이

2018년 08월 11일 접수; 2018년 08월 11일 수정본 접수; 2018년 08월 31일 게재 확정
† 교신저자 yilee2000@gmail.com

되어서는 안 된다고 여겨졌다[18]. 하지만 다양한 연구들이 주식 시장의 가격은 예측이 가능함을 보이며 EMH를 반증했다[8, 9]. 최근 들어 온라인 소셜 미디어가 많은 이용자들을 끌어 모으자, 이를 활용한 주가예측 연구들이 진행되었다. 트위터의 감정상태와 주가지수의 영향을 연구하기도 하고[8, 24] 소셜 미디어 데이터를 활용해 주식 시장 감성 사전을 구축하기도 하였다[16]. 물론 전통의 뉴스 데이터를 활용한 연구들도 여전히 진행되었다[23, 27, 29].

주식 시장을 예측하는 방법 또한 다양했는데, 감성분석(Sentiment Analysis)도 그 중 하나이다[7, 20, 21, 23]. 감성사전은 감성분석을 위해 필요한 단어들의 사전으로, 각 분야마다 활용되는 단어가 다르기에 새롭게 만들어야 한다. 예를 들어 감자라는 단어가 일반 문서에서 등장한다면 식품인 감자를 나타내 중성인 단어가 되겠지만, 감자가 금융 관련 문서에 등장한다면 이는 주식회사나 유한 회사가 결손을 보전하거나 과대 자본을 시정하기 위하여 법원에 등록되어 있는 자본의 총액을 줄이는 일[29]을 나타내 전혀 다른 감성을 나타내게 된다. 그렇기 때문에 각 도메인마다 그 특성에 맞는 감성사전 구축이 필요하다.

국내 주식 시장에 특화된 감성사전 구축을 위한 연구도 활발하게 진행되었다[1, 3, 4, 6]. 하지만 기존 연구에서는 단어의 문맥이 충분히 고려되지 않았으며, 사용된 데이터의 크기 또한 크지 않았다. 이에 본 연구는 3년치 증시 주요 뉴스 데이터를 활용하여 주가지수의 방향성을 예측하기 위한 감성사전을 문맥을 고려하여 자동으로 구축해보고자 한다.

Ⅱ. 관련 연구

2.1 주식 시장 예측 연구

수많은 연구자들이 주식 시장을 예측하고자 연구를 진행했다. 초기의 주식 시장 예측 연구의 대세는 Fama(1965)가 제시한 EMH(Efficient Market Hypothesis)였다[9]. 주가가 뉴스 같은 새로운 정보에 크게 영향을 받는다는 내용인데, 뉴스는 무작위성을 보이기 때문에, 결국 주식 시장 예측 정확도는 50% 이상이 될 수 없다는 것이었다[18]. 하지만 연구자들의 노력은 계속되어, EMH가 옳지 않다는 것이 밝혀졌다[8, 18].

뉴스를 이용한 주식 시장 예측 연구들은 계속해서 진행되었다. Li et al.[7]는 뉴스가 주가에 미치는 영향을 감성분석을 통해 연구했고[6], Li et al.[6]은 뉴스와 대중의 감정이 주가에 미치는 영향에 대해 연구했다[7]. Wu et al.[22]은 경제 뉴스로 대만 주식 시장을 예측하는 연구를 진행하였다[8]. 뉴스뿐만 아니라 온라인 소셜 미디어를 이용한 주가 예측 연구들도 등장했다. Bollen et al.[8]은 트위터의 감정상태가 다우존스 지수에 미치는 영향을 연구하여 86.7%의 정확도로 일간 상승과 하락을 예측하였다[3]. Zhang et al.[24]은 트위터 메시지의 희망과 공포가 다우존스, 나스닥, Standard & Poor's 500 등의 종합 주가지수에 미치는 영향에 대해 연구하여 감정상태와 다음날 종합 주가지수와 관계가 있음을 밝혀냈다[24]. Oliveira et al.[16]은 주식시장 중심 마이크로 블로그 서비스인 StockTwits 데이터를 활용해 감성사전을 구축하기도 하였다.

최근 들어서는 딥러닝 알고리즘을 활용해 주식 시장을 예측하려는 시도들이 많이 포착되고 있다. Hiransha et al.[10]은 MLP(Multilayer Perceptron), RNN(Recurrent Neural Networks), LSTM(Long Short-Term Memory), CNN(Convolutional Neural Network)을 활용해 인도의 NSE(National Stock Exchange)와 NYSE(New York Stock Exchange) 지수를 예측했다[10]. Zhou et al.[25]은 LSTM과 CNN을 변형한 GAN-FD 모델을 제시해 중국의 주가지수를 예측했고, Matsubara et al.[14]은 뉴스 기사를 토대로 Support Vector Machine과

MLP를 이용해 닛케이 225 지수와 Standard & Poor's 500 지수를 예측했다[14].

2.2 감성분석 연구

감성분석은 오피니언 마이닝이라고도 불리며, 비정형 텍스트 데이터로부터 특정 상품이나 개념에 대한 사람들의 생각, 감정, 태도와 같은 주관적인 반응을 분석해내는 과정을 의미한다[1, 17]. 이러한 감성분석을 통해 주식 시장을 예측하려는 시도 또한 많이 있었다. Antweiler and Frank[7]는 야후 금융과 Raging Bull의 게시판 메시지 150만 개에 대한 감성분석을 진행했고, Schumaker et al.[21]는 금융 뉴스에 대한 감성분석을 시도해 투자자들이 좋은 뉴스를 보면 팔고, 나쁜 뉴스를 보면 사는 것이 좋을 수 있다고 제안했다. Yu et al.[23]은 블로그, 포럼, 트위터 등의 소셜 미디어와 New York Times, Wall Street Journal, The Economist 등을 포함한 10대 전통적 미디어에 대한 감성분석을 진행해 기업의 주가를 예측하였다.

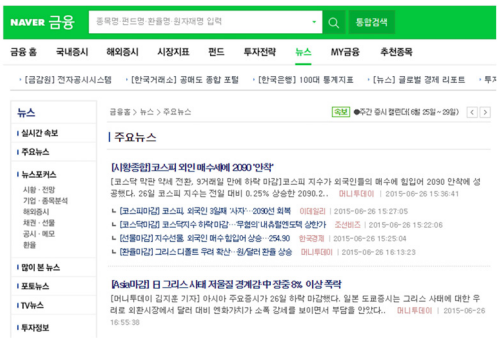
감성분석을 통해 국내 주식 시장을 예측하려는 연구도 활발하게 진행되었다. 안성원, 조성배[3]는 Bag of Words 모델과 Naïve Bayesian 분류 기법을 기반으로 뉴스를 긍정과 부정으로 분류해 신규 발행된 뉴스가 주가 상승 또는 하락에 영향을 미치는지 예측하는 알고리즘을 제안했고, 유은지 외[4]는 익일 주가지수와 빈도수를 기반으로 감성사전을 구축하여 기사 1,000여건을 대상으로 감성분석을 진행했다. 홍태호 외[6]은 뉴스 감성분석과 SVM을 이용해 다우존스 지수와 Standard & Poor's 500지수를 예측했고, 김재봉, 김형중[1]은 증권정보공유에 특화된 웹사이트인 Paxnet의 게시글을 활용해 감성사전을 구축하고 감성분석을 시행하였다.

하지만 기존 연구에서는 단어의 문맥이 충분히 고려되지 않았으며, 사용된 데이터의 크기 또한 크지 않았다. 이에 본 연구는 3년치 증시 주요 뉴스 데이터를 활용하여 주가지수의 방향성을 예측하기 위한 감성사전을 문맥을 고려하여 자동으로 구축해보고자 한다.

Ⅲ. 방법론

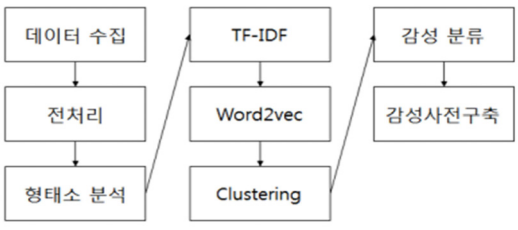
3.1 데이터 수집

네이버 금융[27]의 주요 뉴스에는 <그림 1>에서 볼 수 있듯이 하루에 올라오는 증시 뉴스 중 네이버 측에서 엄선한 약 40개의 뉴스 기사가 노출된다. 이를 기준으로, 증시관련 일간 주요 뉴스를 2015년 1월 1일부터 2017년 12월 31일까지 수집했다. 하나의 주제로 묶여 올라온 기사의 경우 대표 기사만 수집하였으며, 수집도구로는 Python의 기본 라이브러리와 beautifulsoup을 사용했다. 그 결과, 24,995개의 기사가 수집되었고, 중복 및 결측치를 제외한 50개 언론사의 24,985개의 기사가 분석에 활용되었다.



〈그림 1〉 네이버 금융 주요뉴스 화면[27]

3.2 연구 방법



〈그림 2〉 전체 분석 과정

전체 분석 과정은 <그림 2>와 같다. 데이터 수집이 완료된 후, 수집된 데이터를 기반으로 전처리를 진행하였다. 기사 하단의 불필요한 광고 등 문구를 삭제하고 이를 기반으로 형태소 분석을 실시하였다. 형태소 분석기는 Python의 KoNLPy를 이용해 꼬꼬마 형태소 분석기를 사용하였다[2]. 이후 2글자 이상인 단어만 추출하고 불용어 사전을 구축하여 불용어와 숫자, 특수문자 등을 제거하였다.

3.2.1 TF-IDF

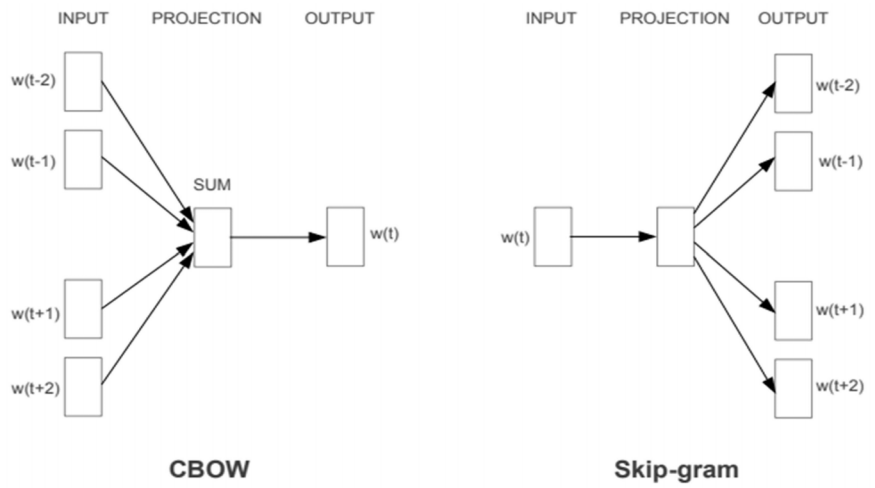
TF-IDF의 TF는 Term Frequency의 약자로, 단어의 빈도를 나타내며, IDF는 Inverse Document Frequency의 약자로, 역문서빈도를 나타낸다. TF-IDF는 TF와 IDF를 곱한 값으로, 한 단어의 특정 문서에서 중요도를 나타내는 값이다[19]. 본 연구에서는 Python의 오픈소스 라이브러리 gensim[30]을 이용해 TF-IDF를 산출하고, 0.1 이상의 값을 가지는 단어들만을 골라내 3,688,238개의 단어로 추려냈다.

3.2.2 Word2Vec

Word2vec은 구글의 Mikolov 등이 2013년 제안한 개념으로, 신경망 분석에 기반을 둔 비지도 학습 기법이다[15]. 이 모델은 각 단어들이 학습 문헌 내에서 가지는 의미를 다차원의 벡터 값을 통해 수치적으로 표현하는 것을 목표로 하였다[5]. 여기에는 두 가지 아키텍쳐가 존재하는데, 하나는 CBOW이고 다른 하나는 Skip-gram이다.

<그림 3> 좌측에 있는 CBOW 모델은 그림과 같이 현재 단어를 문맥에 기반해 예측하는 모델이고, 우측의 Skip-gram 모델은 현재 단어가 주어졌을 때 그 주변 단어들을 예측하는 모델이다.

본 연구에서는 두 가지 모델을 모두 채택해 진행하였다. Skip-gram의 경우 CBOW보다 많은 컴퓨팅 파워를 필요로 하였기에, 일부 연산은 메모리 부족 에러로 아쉽게도 진행하지 못하였다. 3,688,238개의 단어 중 고유한 단어의 수는 41,116개였는데, 모두 사용할 경우 역시 메모리 부족 에러가 발생하였다. 이에 최소 등장 횟수로 사용 단어 수를 조절하였다. 벡터 사이즈의 경우 너무 낮은 값을 넣게 되면 단어들이 한 곳으로 몰려 클러스터로 나누기에 무의미한 결과를 보여줬으므로, 최대한의 차이점을 주기 위해 컴퓨팅 파워가 허락하는 한 높여보았다. 이를 위해 Python의 gensim 라이브러리를 활용하였다. 생성된 Word2vec 모델별 정보는 다음의 <표 1>과 같다.



〈그림 3〉 Word2Vec의 두 모델[15]

〈표 1〉 Word2Vec 모델별 세부정보

모델 번호	최소 등장 횟수	사용 단어수	벡터 사이즈	모델
1	2	25,902	8,000	CBOW
2	2	25,902	8,000	Skip-gram
3	4	20,923	10,000	CBOW
4	4	20,923	10,000	Skip-gram
5	6	14,911	20,000	CBOW
6	6	14,911	20,000	Skip-gram

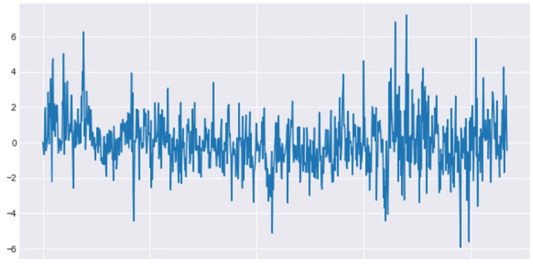
3.2.3 클러스터링

Word2vec 결과를 바탕으로, 클러스터링을 진행하였다. 클러스터링에는 K-means 알고리즘을 사용하였다. K-means 알고리즘은 가장 잘 알려진 분할적 클러스터링 방법으로, 다양한 분야에서 독립적으로 생겨난 역사를 지니고 있다. 비록 50년전에 등장하긴 했지만, 쉽게 사용 가능하고, 간단하고, 효율적이며, 경험적으로 성공적이었기에 여전히 클러스터링에 가장 많이 사용되고 있는 알고리즘이다. 주어진 K값에 따라 선정된 클러스터의 중앙과 데이터의 유클리디언 거리를 구하는 방식으로 K개의 클러스터를 나누게 된다[11].

본 연구에서 K값은 양성, 중성, 음성을 나타내기 위해 3으로 설정하였다. 그 결과 3개의 클러스터가 형성되었다. 이를 위해 Python의 오픈소스 라이브러리 Scikit-learn[28]을 활용하였다.

3.2.4 감성사전 구축

Word2vec을 기반으로 구축된 감성사전을 평가하기 위하여 감성분석을 진행하였다. 2015년 1월 1일부터 2017년 12월 31일까지의 KOSPI 종가 데이터와 비교해 보았다. KOSPI 종가 데이터는 한국거래소 KRX의 Marketdata에서 엑셀 파일 형태로 다운로드 가능하다[28]. 다운로드 받은 엑셀 파일에서 일자, 현재지수 외에도 대비, 등락률, 배당수익률, 추가이익비율, 주가자산 비율 등등 다양한 지수를 열람 가능하지만 해당 감성사전은 주가지수의 방향성에 초점을 맞추고 있기 때문에, 방향성 비교를 위해 대비를 활용하였다.



〈그림 4〉 모델 1 감성분석 결과

Ⅳ. 결 과

〈표 2〉 Word2Vec 모델별 세부정보

모델	정확도	정밀도	재현율	F값
1	56.99%	60.81%	53.83%	57.10%
2	52.65%	55.56%	54.85%	55.20%
3	54.00%	57.55%	51.53%	54.37%
4	52.92%	55.47%	58.16%	56.79%
5	55.09%	59.27%	49.74%	54.09%
6	53.32%	56.19%	55.61%	55.90%

모델 평가 결과, 대체로 50%대의 비슷비슷한 정확도를 보였다. 이 중, 단어수 25,902, 벡터 사이즈 8,000에 CBOW 아키텍쳐를 활용한 모델 1이 가장 높은 정확도를 보여줬으며, 재현율을 제외한 정밀도와 F값에서도 가장 높은 수치를 보여주었다. 반면에 가장 고른 클러스터 분포를 보여주었던 단어수 14,911, 벡터 사이즈 20,000에 Skip-gram 아키텍처를 사용한 모델 6은 무난한 성능을 보여주었지만, 재현율을 제외하고는 모델 1에 비해 낮은 수치를 보여주었다. 수치상으로는 재현율이 가장 큰 차이를 보였는데 모델 5의 재현율이 유일하게 50% 이하의 값을 나타냈으며, 모델 4의 재현율은 58%로 약 9% 정도의 차이를 보였다.

V. 결 론

본 연구는 주가 방향성 예측용 감성사전을 자동으로 구축하기 위해 진행된 것으로, 3년치 증시 주요 뉴스 데이터를 활용하여 TF-IDF로 중요 단어를 추려낸 이후 단어의 문맥을 고려할 수 있는 Word2Vec을 기반으로 하여 주가 방향성 예측용 모델을 생성하고, 이 모델을 적용해서 주가의 방향성을 예측하였다.

본 연구의 의의는 25,000여 건의 기사를 토대로 문맥을 고려하여 단어 감성사전을 자동으로 구축하는 방법을 제시하였다는 것이다. 전통적으로 주식 시장 예측에 활용된 뉴스 데이터를 활용하면서도 Word2vec을 활용해 단어의 문맥을 본격 적용해 자동으로 감성 사전을 구축한 것이다. 그러므로 이를 다른 분야에도 활용을 한다면 문맥을 고려한 각 분야의 감성사전을 자동으로 구축해 해당 분야와 관련된 학계와 산업계 및 정부에서 유용하게 쓰일 수 있을 것으로 보인다.

그러나 이러한 장점들에도 불구하고 본 연구에는 다음과 같은 한계점이 있다. 전체 주가지수의 방향은 예측했지만, 실제 투자자들에게 도움이 될 수 있는 특정 종목에 대한 방향은 예측은 수행하지 못하였다.

후속 연구로는 컴퓨팅 파워를 확보하여 본 연구에서 진행하지 못했던 연산을 수행해 모델을 제작할 수 있을 것이다. 또한 해당 모형을 좀 더 보완하여 전자, 제약 등 특정 분야의 주가에 대한 감성사전을 제작해 주식 시장이라는 광범위한 도메인을 모두 아우르는 것이 아닌, 보다 전문화된 감성사전을 구축하는 연구가 진행될 수 있을 것으로 보인다.

참 고 문 헌

[1] 김재봉, 김형중, “주가지수 방향성 예측을 위한 도메인 맞춤형 감성사전 구축방안”, 한국디지털콘텐츠학회논문지, 제18권, 제3호, pp.585-592, 2017.

[2] 박은정, 조성준, “KoNLPy : 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제26회 한글 및 한국어 정보처리 학술대회논문집, 2014.

[3] 안성원, 조성배, “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측”, 한국정보과학회 학술발표논문집, 제37권, 제1C호, pp.364-369, 2010.

[4] 유은지, 김유신, 김남규, 정승렬, “주가지수 방향성 예측을 위한 주제지향감성사전 구축 방안”, 지능정보연구, 제19권, 제1호, pp.95-110, 2013.

[5] 한남기, “word2vec 학습 자질을 사용한 새로운 한글 개체명 인식 모델 제안”, 석사학위논문, 연세대학교, 서울, 2016.

[6] 홍태호, 김은미, 차은정, “뉴스 감성분석과 SVM을 이용한 다우존스 지수와 S&P500 지수 예측”, 인터넷전자상거래연구, 제17권, 제1호, pp.23-36, 2017.

[7] Antweiler, W. and M. Frank, “Is all that talk just noise? The information content of internet stock message boards”, *The Journal of Finance*, Vol.59, No.3, pp.1259-1294, 2004.

[8] Bollen, J., H. Mao, and X. Zeng, “Twitter mood predicts the stock market”, *Journal of Computational Science*, Vol.2, No.1, pp.1-8, 2011.

[9] Fama, E., “The behavior of stock-market prices”, *The Journal of Business*, Vol.38, No.1, pp.34-105, 1965.

[10] Hiransha, M., E. Gopalakrishnan, K. Vijay, and K. Soman, “NSE Stock Market Prediction Using Deep-Learning Models”, *Procedia Computer Science*, Vol.132, pp.1351-1362, 2018.

[11] Jain, A., “Data clustering: 50 years beyond K-means”, *Pattern Recognition Letters*, Vol.31,

No.8, pp.651-666, 2010.

[12] Li, Q., T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements", *Information Sciences*, Vol.278, pp.826-840, 2014.

[13] Li, X., H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis", *Knowledge-Based Systems*, Vol.69, No.1, pp.14-23, 2014.

[14] Matsubara, T., R. Akita, and K. Uehara, "Stock Price Prediction by Deep Neural Generative Model of News Articles", *IEICE Transactions on Information and Systems*, Vol.E101, No.4, pp.901-908, 2018.

[15] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv preprint, arXiv:1301.3781 [cs.CL], 2013.

[16] Oliveira, N., P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures", *Decision Support Systems*, Vol.85, pp.52-73, 2016

[17] Pang, B. and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2, pp.1-135, 2018.

[18] Qian, B. and K. Rasheed, "Stock market prediction with multiple classifiers", *Applied Intelligence*, Vol.26, No.1, pp.25-33, 2007.

[19] Salton, G. and M. McGill, *Introduction to Modern Information Retrieva*, McGraw-Hill, 1986.

[20] Schumaker, R. and H. Chen, "Textual Analysis of stock market prediction using breaking financial news : The AZFin Text System", *ACM Transactions on Information Systems*, Vol.27, No.2, pp.1-19, 2009.

[21] Schumaker, R., Y. Zhang, C. Huang, and H. Chen, "Evaluating sentiment in financial news articles", *Decision Support Systems*, Vol.53, No.3, pp.458-464, 2012.

[22] Wu, G., T. Hou, and J. Lin, "Can economic news predict Taiwan stock market returns?", *Asia Pacific Management Review*, 2018.

[23] Yu, Y., W. Duan, and Q. Cao, "The impact of social and conventional media on firm equity value : a sentiment analysis approach", *Decision Support Systems*, Vol.55, No.4, pp.919-926, 2013.

[24] Zhang, X., H. Fuehres, and P. Gloor, "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear", *Procedia-Social and Behavioral Sciences*, Vol.26, pp.55-62, 2011.

[25] Zhou, X., Z. Pan, H. Guyu, T. Siqi, and C. Zhao, "Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets", *Mathematical Problems in Engineering*, Vol.2018, 2018.

[26] http://marketdata.krx.co.kr/mdi#document=030402.

[27] http://finance.naver.com/news/mainnews.nhn.

[28] http://scikit-learn.org/stable/index.html.

[29] https://ko.dict.naver.com/detail.nhn?docid=867600.

[29] https://radimrehurek.com/gensim/.

저 자 소 개

김 다 예(Daye Kim)

- 2012년 : 연세대학교 중어중문학과 (문학사)
- 2017년~현재 : 연세대학교 정보대학원 비즈니스 빅데이터 분석 트랙 (석사 과정)
- 관심분야 : Big Data Analytics, Data Mining, Deep Learning

이 영 인(Youngin Lee)

- 2004년 : 서울시립대학교 전산통계학과 (이학사)
- 2017년~현재 : 연세대학교 정보대학원 비즈니스 빅데이터 분석 트랙 (석사 과정)
- 관심분야 : 빅데이터 분석, 딥러닝, 블록체인