Journal of KIIT. Vol. 20, No. 9, pp. 13-18, Sep. 30, 2022. pISSN 1598-8619, eISSN 2093-7571 13
http://dx.doi.org/10.14801/jkiit.2022.20.9.13

한국어 텍스트 질의를 활용한 주식 정보 검색 및 분석 기법

김나영*, 유석종**

Stock Information Retrieval and Analysis using Korean Natural Language Query

Nayoung Kim*, Seok-Jong Yu**

요 약

최근 들어 주식 시장에 대한 관심도는 더욱 높아져 남녀노소 가리지 않고 주식 투자에 뛰어들고 있으나, 주식에 대해 잘 알지 못하는 초보자들은 관련 정보를 검색하고 해석하는 데 어려움을 겪고 있다. 특히, 미국 주식의 경우 모든 정보가 영어와 수치에 의해 이루어져 있으며 필요한 주식 종목 정보를 쉽게 검색할 수 있는 시스템의 필요성이 절실한 실정이다. 이러한 이유로 본 연구에서는 한국어 텍스트를 사용하여 미국 주식 종목 정보를 검색 및 분석할 수 있는 자연 언어 질의 시스템을 제안하고자 한다. 제안 시스템은 BERT 모델과 Random Forest 앙상블 모델을 통해 자연어 질의를 처리하고, 검색 결과를 이해하기 쉬운 한국어로 제공하여 초보자의 투자 판단에 도움을 줄 수 있다. 다양한 유형의 사용자 질의어를 분석하여 적절한 응답이 가능하도록 시스템을 구현하고 성능을 평가하였다.

Abstract

With the explosively increasing interest in US stock, entering the market is on vogue. However, beginners who have little knowledge are having trouble searching related information. Therefore, a easy system for searching US stock market information is indeed necessary, but there are few researches dealing with these trouble. This paper proposes a new text interface that can conveniently look up US stock information. In addition, the proposed interface includes a BERT model and a Random Forest ensemble model to process natural language queries and shows stock information in an easy way written in Korean. We believe that this interface can help beginners with no knowledge in US stock market and struggling to search for related information.

Keywords
stock market, search system, text interface, BERT model, random forest model

* 숙명여자대학교 소프트웨어학부
- ORCID: https://orcid.org/0000-0003-3750-1290
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Jul. 11, 2022, Revised: Aug. 16, 2022, Accepted: Aug. 19, 2022
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Women's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

최근 미국 주식 시장에 대한 관심도가 폭발적으로 증가하여 관련 연구[1]-[3]가 늘어났고, 많은 초보 투자자들이 시장에 뛰어들게 되었다. 그러나 대부분의 투자자들은 주식 시장에 대한 충분한 정보 없이 투자를 시작하고 있다. 특히, 주식 관련 데이터는 그림 1과 같이 그 정보의 종류가 많고 영어로 되어 있어 종목의 상황을 이해하기 어려운 특징이 있다. 최근 토스 증권은 필수 정보만을 제공하는 MTS 앱[1] 서비스를 시작하여 이러한 문제를 개선하고자 하였다. 그러나 너무 많은 정보를 배제하고 있어 종목 분석을 원하는 투자자에게 적합하지 않으며, 주식 투자 시장에 대한 학습을 저해할 수 있다. 본 연구에서는 이와 같은 문제점을 개선하고자 한국어 자연어 질의를 활용한 주식 종목 정보 검색 시스템을 제안하고자 한다.

BERT[4]는 대표적인 자연어 처리 딥러닝 모델로 Google에서 2018년에 소개되었다. BERT는 사전 학습된 레이블이 되어있지 않은 대용량의 데이터로 언어 모델을 학습한다. 그 성능이 탁월하여 많은 자연어 처리 시스템에서 사용하고 있다. 그러나 BERT 모델은 한국어 형태소 분석에는 미숙한 결과를 보이기 때문에 본 연구에서는 한국어 형태소 분석기 패키지인 Konlpy의 Okt 모델을 사용하여 한국어 문장을 처리하고, 자연어 질의 문장에서 키워드를 뽑아내기 위하여 KeyBERT를 사용하였다[5][6].

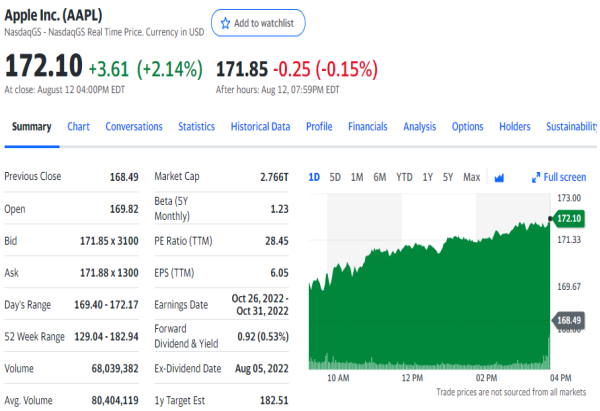

그림 1. Yahoo finance
Fig. 1. Yahoo finance

본 시스템은 자연어로 기술된 사용자의 주식 관련 질의 문장을 종목, ETF, 배당, 섹터 유형으로 분석 처리할 수 있으며, 분석 결과를 한국어로 번역하여 제공할 수 있다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구로 자연어 처리 모델인 BERT와 한국어 형태소 분석에 사용된 Konlpy와 랜덤 포레스트(Random Forest)를 소개한다. 3장에서는 제안하는 자연어 질의 기반 주식 정보 검색 시스템을 기술하고, 4장, 5장에서 각각 실행결과와 결론에 대하여 기술한다.

Ⅱ. 관련 연구

2.1 BERT

자연어 처리를 위한 사전 훈련 언어 모델에는 ELMo[7], OpenAI GPT[8] 등이 있다. 두 방식은 동일한 목적 함수로 학습을 수행할 뿐만 아니라 일방향이나 얕은 양방향성을 가지는 한계점을 보인다. 이 문제점을 해결하기 위해 제안된 BERT[4]는 입력값 전체와 마스킹된 토큰을 한 번에 트랜스포머 인코더에 넣고 원래 토큰값을 예측하므로 깊은 양방향성을 보인다. BERT는 다른 자연어 처리 모델보다 탁월한 성능을 보이며, 사전 학습되어 적은 자원으로도 자연어 처리를 수행하지만, 형태소 단위로 이루어진 한국어 처리에는 미흡한 면을 보인다.

2.2 Konlpy

자연어 처리를 위해서는 말뭉치, 토큰 생성, 형태소 분석, 품사 태깅이 제공되어야 한다. 이러한 작업을 수행하는 대표적인 파이썬 패키지로는 NLTK가 있다[9]. NLTK를 기반으로 개발된 한국어 형태소 분석기가 바로 Konlpy이다. 본 연구에서는 Konlpy를 사용하여 한국어 질문을 분석하였다.

2.3 랜덤 포레스트

본 논문에서는 결정 트리보다 그 성능이 뛰어나고 앙상블 모델로써 널리 쓰이는 랜덤 포레스트 기법[10]을 사용하였다. 지도 학습 머신러닝 기법 중

하나인 결정 트리는 오버피팅(Overfitting)될 가능성이 높다. 이 문제점을 해결하기 위해 제안된 앙상블 모델이 랜덤 포레스트이다. 랜덤 포레스트 기법은 그림 2와 같이 동일 데이터에 대해 여러 개의 결정 트리를 형성한 후 가장 많이 득표한 결과를 최종으로 선택한다.

이 과정에서 훈련 집합의 부분을 랜덤하게 활용하는 배깅(Bagging)을 사용한다. 각 트리의 가지를 치지 않고 데이터가 전부 분류될 때까지 크기를 늘리기 때문에 트리 하나하나는 오버피팅이 날 수 있지만 전체적으로는 분산이 낮아져 오버피팅이 나지 않는다[10]. 본 논문에서는 랜덤 포레스트를 사용하여 뽑아낸 키워드를 기반으로 어떤 유형의 질문인지 파악하였다.

III. 자연어 질의에 의한 종목 정보 검색

기존의 주식 관련 연구들은 주로 자동 매매 시스템이나 주가 예측에 많이 사용되었다. 챗봇 기반의 주식 정보 시스템 연구[11]는 단편적인 정보 제공 수준으로 종목간 비교 분석은 고려하고 있지 않다. 특히, 한국어로 된 자연어 질의 문장을 처리하여 주식 종목 정보를 검색하거나 분석하는 시스템에 관한 연구는 찾아보기 어렵다. 본 논문에서는 미국 S&P500 지수에 포함된 종목 정보 검색과 분석이 가능한 한국어 자연어 처리 시스템을 개발하고자 한다.

3.1 실험 환경

본 시스템은 Pytorch를 활용하여 구현하였으며, Pandas, sklearn, Konlpy 등의 라이브러리를 사용하였다. 미국 주식 데이터 정보는 yfinance API[12]를 통해 수집하였다.

3.2 작동 원리

그림 3은 제안 시스템의 자연어 처리 과정이다. 한국어 질의 문장을 띄어쓰기 단위로 자른 다음 Konlpy의 Okt 모델로 형태소 단위로 토큰화하였다. 토큰화된 문장은 단어 카운트(CountVectorizer)를 거친 후 KeyBERT 모델을 사용하여 키워드를 추출한다. 추출한 키워드들은 코사인 유사도에 따라 중요도를 매긴 후 문장과의 유사도 및 단어의 중요도에 따라 순서화된 키워드 집합으로 출력한다. 질의어 중 영어로 된 주식 종목의 이름이나 섹터 명칭의 경우에는 NLTK와 KeyBERT를 사용하여 인식한다.

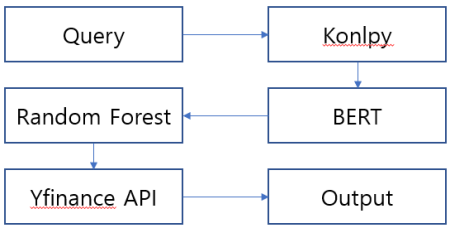

그림 3. 자연어 처리 텍스트 인터페이스의 작동 원리
Fig. 3. Principle of NLP text interface

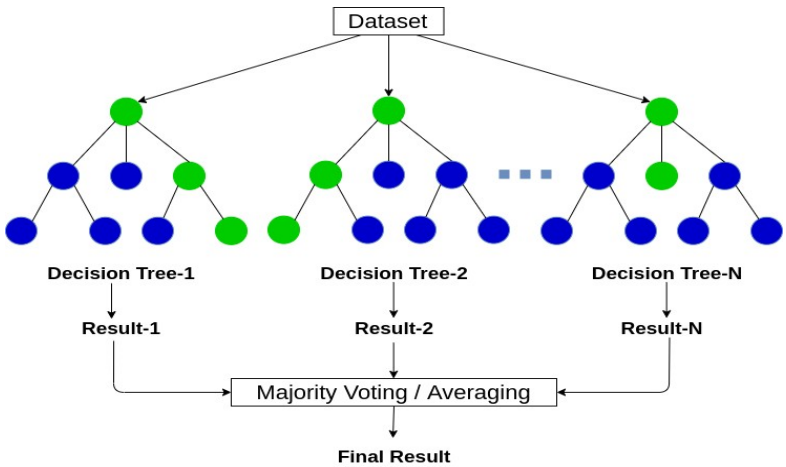

그림 2. 랜덤 포레스트의 작동 원리
Fig. 2. Principle of random forest

랜덤 포레스트 모델을 사용하여 질의어의 유형을 파악한다. 한국어로 된 섹터 명칭은 구글 번역 패키지로 영어로 변환하여 사용하였다.

IV. 실행 결과

표 1은 본 시스템이 처리 가능한 질의 유형이다.

표 1. 질의 유형
Table 1. Applicable series of question

Query category	Detailed query category
Stock	Stock profile
	Stock investigation
	Stock comparison
ETF	ETF stock investigation
Dividend	Dividend investigation
Market	Market investigation

4.1 종목 프로필 질의

종목 프로필 질의는 표 2와 같이 세 가지 유형으로 구분되며 실행결과는 그림 4와 같다.

표 2. 종목 프로필의 세부 질의 유형
Table 2. Applicable series of question in presentation of stock profiling

Type	Korean query	Indicators
Price	애플의 오늘 가격 알려줘.	Closing price, highest price of the year
Volume	테슬라의 거래량 알려줘.	Closing volume
Company	3M은 어떤 회사인지 알려줘.	Company information

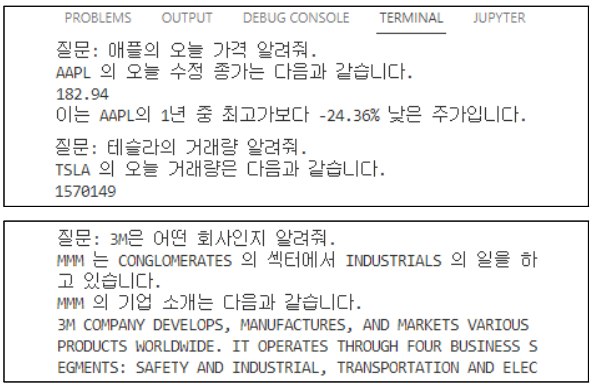

그림 4. 종목 프로필 질의 처리 결과
Fig. 4. Stock profile query result

주가 유형의 경우 수정 종가와 1년 중 최고가와 비교하여 몇 퍼센트가 더 낮은지 분석한다. 거래량 제시의 경우 하루의 거래량을 출력하고, 회사 소개의 경우 그 종목의 사업 종류, 섹터와 기업 소개를 출력한다.

4.2 종목 분석

표 3은 종목 분석 질의 유형이다. 종목 분석을 위해 주가수익비율(PER), 1년간 주가 변화율, 기업 가치, 영업 순이익 비율, 거래량, 그리고 전문가들의 평가를 활용한다. 기업 가치의 경우에는 100만과 10억 단위로 압축하였고, 전문가 평가는 최신 5개만 사용하였다. 그림 5는 질의 실행 결과이다.

표 3. 한 종목 분석 질의 유형
Table 3. Applicable question in profiling stock

Type	Query	Indicators
Investigation	애플 종목을 분석해줘	PER, price change, company value, operating NET value, volume, recommendation

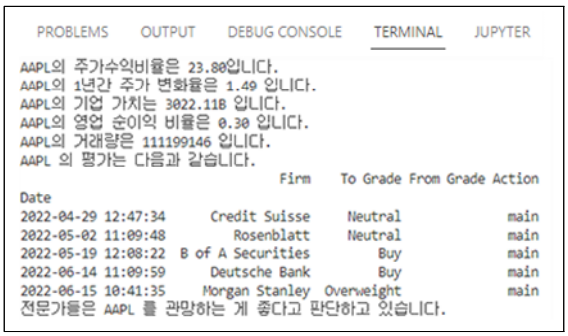

그림 5. 종목 분석 질의 처리결과
Fig. 5. Stock profiling query result

4.3 종목 비교 분석

종목 비교는 표 4와 같은 질의 유형이 가능하며 그림 6과 같이 최대 3개 종목까지 각 지표별 평가 점수를 합산하여 최종 순위를 평가한다.

표 4. 종목 비교 질의 유형
Table 4. Applicable question in comparing stocks

Type	Query	Indicators
Compare	테슬라랑 구글 중에 어떤 종목을 살까?	PER, price change, Company value, operating NET value, volume, recommendation

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    JUPYTER

기업들을 영업 순이익 비율이 작은 순으로 정렬해보았습니다.
TSLA의 영업 순이익 비율은 19.55 입니다. 따라서 1 점을 얻습니다.
기업들을 1년간 주가 변화율이 큰 순으로 정렬해보았습니다.
TSLA의 1년간 주가 변화율은 1.67 입니다. 따라서 1 점을 얻습니다.
기업들을 기업 가치가 큰 순으로 정렬해보았습니다.
TSLA 의 기업 가치는 17.34B 입니다. 따라서 1 점을 얻습니다.
기업들을 영업 순이익 비율이 큰 순으로 정렬해보았습니다.
TSLA 의 영업 순이익 비율은 0.20115 입니다. 따라서 1 점을 얻습니다.
기업들을 거래량이 큰 순으로 정렬해보았습니다.
TSLA 의 거래량은 386539 입니다. 따라서 1 점을 얻습니다.
기업들을 전문가들의 평가가 좋은 순으로 정렬해보았습니다.
TSLA 의 전문가 평가는
                          Firm        To Grade      From Grade Action
Date
2022-06-01 12:08:20  Goldman Sachs          Buy                 main
2022-06-03 11:22:15    Cowen & Co.  Market Perform              main
2022-06-09 08:58:31            UBS          Buy         Neutral   up
2022-06-10 10:30:27       Barclays    Underweight               main
2022-06-13 09:12:55    RBC Capital     Outperform  Sector Perform  up
따라서 1 점을 얻습니다.
해당하는 주식들을 추천 순대로 나열해보았습니다.
TSLA 는 6 점으로 1 위입니다.
위 순으로 구매하시면 되겠습니다.
```

그림 6. 종목 비교 질의 결과
Fig. 6. Stock comparison query result

4.4 ETF 종목 분석

ETF 분석 질의 유형은 표 5와 같으며, 그림 7과 같이 ETF에서 질의 종목의 수정 종가, 거래량, 시장 가치가 높은 순으로 검색한다.

표 5. ETF 분석 질의 유형
Table 5. Applicable question in profiling ETF stocks

Type	Query	Indicators
ETF Investiagation	구글을 보유한 ETF 중에서 가격이 높은 순으로 알려줘.	ETF stocks, closing price
	테슬라가 포함된 거래량이 많은 ETF를 보여줘.	ETF stocks, volume

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    JUPYTER
질문: 구글을 보유한 ETF 중에서 가격이 높은 순으로 알려줘.
해당하는 ETF 중 오늘 수정 종가가 높은 순서대로 5개를 뽑아 정
렬했습니다.
1위: IVV, 403.41
2위: SPY, 401.72
3위: VOO, 369.16
4위: XLG, 304.41
5위: QQQ, 301.94

질문: 테슬라가 포함된 거래량이 많은 ETF를 보여줘.
해당하는 ETF 중 거래량이 큰 순서대로 5개를 뽑아 정렬했습니다
.
1위: SPY, 96829670
2위: QQQ, 79012795
3위: SPLG, 9365960
4위: QYLD, 7401964
5위: VOO, 5292760
```

그림 7. ETF 비교 질의 결과
Fig. 7. ETF profile query result

4.5 섹터 시장 분석

섹터 시장 분석에서는 그림 8과 같이 가격 및 거래량을 포함하는 질의 처리가 가능하다.

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    JUPYTER
질문: 반도체 섹터의 가격이 높은 종목 알려줘.
SEMICONDUCTOR 섹터의 기업 리스트입니다.
주가 수익 비율이 작은 순서대로 정리했습니다.
MU 의 주가 수익 비율은 6.2484794 입니다.
QRVO 의 주가 수익 비율은 7.970342 입니다.
SWKS 의 주가 수익 비율은 8.040404 입니다.
QCOM 의 주가 수익 비율은 11.591924 입니다.
INTC 의 주가 수익 비율은 11.783783 입니다.
따라서 위의 종목들을 구매하면 되겠습니다.

질문: IT 섹터의 거래량이 많은 주식 보여줘.
INFORMATION 섹터의 기업 리스트입니다.
주가 수익 비율이 작은 순서대로 정리했습니다.
DXC 의 주가 수익 비율은 6.8999996 입니다.
IBM 의 주가 수익 비율은 12.970874 입니다.
FIS 의 주가 수익 비율은 13.058187 입니다.
FB 의 주가 수익 비율은 13.957835 입니다.
LDOS 의 주가 수익 비율은 14.582497 입니다.
따라서 위의 종목들을 구매하면 되겠습니다.
```

그림 8. 시장 분석 질의 결과
Fig. 8. Market investigation query result

V. 결론 및 향후 과제

본 연구에서는 자연어 질의로 미국 S&P 500 주식 정보를 손쉽게 검색 분석할 수 있는 시스템을 제안하였다. 이를 위해 KeyBERT와 Konlpy를 사용하여 자연어 처리 모델을 구현하고, 랜덤 포레스트와 yfinance API를 사용하였다. 본 시스템은 영어와 복잡한 수치로 된 미국 주식 정보 데이터를 자연어 문장으로 분석 가능한 장점이 있으며, 이는 한국어 주식 종목에도 확장 적용 가능하다. 그러나 yfinance API와 BERT의 온라인 처리 지연 시간이 존재하여 실시간 응답성이 떨어지는 한계점을 갖는다.

References

[1] Toss Invest, https://tossinvest.com [accessed: May 29, 2022]

[2] J. H. Moon, J. J. Sohn, and S. J. Yu, "Stock Portfolio Analysis and Visualization Services using Gamification", Journal of KIIT, Vol. 20, No. 1, pp. 191-198, Jan. 2022. https://doi.org/10.14801/jkiit.2022.20.1.191.

[3] H. J Jo, J. T. Choi, and J. H. Seo, "OAR Algorithm Technology Based on Opinion Mining Utilizing Stock News Contents", Journal of KIIT, Vol. 13, No. 3, pp. 111-119, Mar. 2015. http://dx.doi.org/ 10.14801/jkiit.2015.13.3.111.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, pp. 3-5, May 2019. https://doi.org/10.48550/arXiv.1810.04805.

[5] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT", Zenodo, v0.3.0, 2020. https://doi.org/10.5281/zenodo.4461265.

[6] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python", Proc. of the 26th Annual Conference on Human & Cognitive Language Technology, pp. 133-136, Oct. 2014.

[7] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations", arXiv: 1802.05365, pp. 2-4, Mar. 2018. https://doi.org/ 10.48550/arXiv.1802.05365.

[8] A. Radford, K. Narasimhan, T. Salimans, and Il. Sutskever, "Improving Language Understanding by Generative Pre-Training", OpenAI, pp. 3-4, 2018.

[9] Bird, Steven, Edward Loper, and Ewan Klein, "Natural Language Processing with Python", O'Reilly Media Inc., 2009.

[10] T. K. Ho, "Random Decision Forests", 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, pp. 278–282, Aug. 1995. https://doi.org/10.1109/ICDAR.1995.598994.

[11] J. Y. Kwon, D. W. Choi, E. S. Kim, and J. H. Moon, "Chatbot-based financial application Using AI Technology", Proc. of the Korea Information Proc. Society Conference, pp. 876–878, Oct. 2019. https://doi.org/10.3745/PKIPS.y2019m10a.876.

[12] Yahoo Finance API, https://pypi.org/project/yfinance/ [accessed: May 29, 2022]

저자소개

김 나 영 (Nayoung Kim)

2022년 9월 현재 : 숙명여자대학교 소프트웨어학부 학사과정
관심분야 : 데이터마이닝, 자연언어처리

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 컴퓨터과학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수
관심분야 : 데이터마이닝, 추천시스템, 정보시각화