

키워드 기반 뉴스 요약과 투자자 감정 분석을 제공하는 서비스 구현

김태환*, 강창협*, 나지운*, 조형욱*, 정석원**

*목포대학교 정보보호학과(학부생), **목포대학교 정보보호학과(교수)

e-mail : (obtusa07, skwldns3, whguddnr9, ajh_2568, jsw)@mokpo.ac.kr

Implementation of service providing keyword-based news summary and stock investor sentiment analysis

Taehwan Kim, Jiun Na, Hyounguk Jo, JaeHyun Kang, Seok Won Jung
Department of Information Security Engineering, Mokpo National University

요 약

최근 코로나19 국면으로 접어들며, 주식시장에 개인투자자가 급증하였다. 그러나 신규 개인 투자자의 수익률은 열악한 상태로 나타난다. 금융 투자에 있어 정보의 중요성은 날이 갈수록 높아져가고 있다. 최근 헤지펀드에서는 전통적 금융 데이터뿐만 아니라 대체 데이터를 이용하여 투자에 나서고 있다. 본 논문에서는 포털 사이트의 종목토론실의 투자자들의 감성분석과 뉴스의 요약 정보를 조합하여 주식시장 투자자에게 투자정보를 제공하는 서비스를 제안한다. 개인 투자자에 따른 적합한 정보와 종목의 감성 정보를 동시에 제공하여 효율 높은 투자를 유도하도록 기대한다.

1. 서 론

자본시장 연구원의 "코로나19 국면의 개인투자자: 투자행태와 투자 성과" 보고서에 따르면, 코로나19가 발생한 후 주식 시장 활동 계좌 수가 843만 개가 증가함과 동시에 87조 원의 주식 순매수 금액으로 유례없는 신규 투자자들의 증가가 있었다[1]. 하지만, 이러한 신규 투자자들의 수익률은 동기간 기준으로 거래 비용 차감 후 -1.2%이며 특히, 시장포트폴리오의 수익률을 차감한 초과 수익률은 차감 후 -17.6%이다. 이러한 열악한 수익률에 대해 보고서에서는 해결책으로 정교한 주식 투자 관리 서비스와 투자자의 과신을 막는 것을 주장하고 있다[1]. 최근의 헤지펀드에서는 전통적인 금융데이터 뿐만 아니라 대체 데이터를 이용하여 투자에 나서고 있다[2]. 이에 주목하여 본 논문에서는 대체 데이터와 전통적 데이터를 융합하여 사용자들에게 투자에 도움이 되는 새로운 지표를 제시하는 것을 목표로 한다.

본 논문의 구성은 다음과 같다. 2장은 기존 서비스의 문제점을 포함한 관련 연구, 3장은 제안하는 서비스 개요를 설명하고, 4장에서는 결론을 맺는다.

2. 관련 연구 및 서비스

2.1 자연어 처리와 감성분석

자연어 처리란 머신러닝을 이용하여 텍스트의 구조와 의미를 파악하고 이를 활용하여 정보를 추출하여 이용하는 기술이라 할 수 있다. 하위분류로는 자연어 인식과 자연어 생성이 존재한다[3]. 본 논문에서는 두 가지 분야를 모두 활용한다.

감성 분석은 텍스트에 들어있는 의견이나, 감성, 평가 태도 등의 주관적인 정보를 컴퓨터를 통해서 분석하는 과정이다[4].

2.2 기존 정보 제공 서비스

종토넷은 각종 종목토론실의 글을 읽어와 인기 있는 글을 메인 화면에 출력해 주는 서비스를 제공한다. 2시간 내의 코스피, 코스닥의 인기 글, 오늘 하루의 코스피, 코스닥의 인기 글, 인기 있는 토론 종목 서비스를 제공한다[5].

빅카인즈는 매일 수집된 뉴스 중 상위 8개의 이슈를 보여주는 오늘의 이슈 기능과 빅카인즈에 기사를 공급하는 언론사별 최신 뉴스를 보여주는 언론사별 뉴스 기능이 있다. 또한 내가 등록한 관심 키워드와 관련된 뉴스를 제공하는 나의 관심 뉴스 기능을 제공하고, 수집된 뉴스에서 주요 인물, 장소명과 기관명을 분석한 후 해당 키워드가 포함된 뉴스 건수가 높은 순으로 보여주는 서비스들을 제공한다[5].

종토넷 서비스의 경우는 원 글을 그대로 제공하기만 하고 분석 기능이 없어 종목에 대한 투자자들의 감성을 알 수 없다. 빅카인즈 서비스의 경우에는 기사에 대한 요약 서비스가 없어 글의 핵심을 알기 위해서는 본문을 읽어야 한다는 문제점이 있다.

3. 제안 서비스 개요

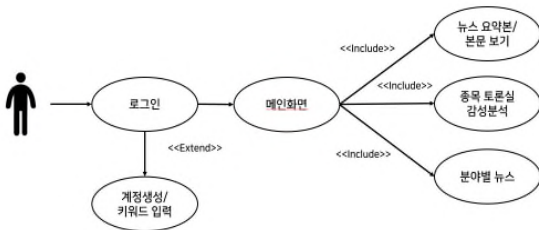
본 논문에서는 종토넷의 분석기능 부재와 빅카인즈의 요약 서비스 부재라는 단점을 보완하고, 이들의 장점을 합쳐서 키워드 기반 뉴스를 제공함과 동시에 투자자의 감성 분석 정보를 함께 제공하는 서비스를 제안하고 이의 구현 결과를 제시한다.

3.1 서비스의 기능

본 논문에서 제안하는 키워드 기반 뉴스 요약과 투자자 감성 분석 제공 서비스의 유스 케이스 다이어그램은 그림 1과 같다. 제공하는 기능은 뉴스 요약본/본문 보기, 종목 토론실 감성 분석, 분야별 뉴스 등이다.

본 서비스는 그림 2와 같은 과정을 통해 사용자에게 제공된다.

크롤러를 통해 뉴스 본문을 크롤링하고, 크롤링 된 뉴스에 대해 뉴스를 요약하고, 키워드를 추출한다.



(그림 1) 유스 케이스 다이어그램

텍스트 마이닝을 DBMS에 저장하고, 크롤링 된 종목 토론실 게시 글에 대한 감성 분석 결과를 DBMS에 저장한다. 위의 과정을 스케줄링 패키지를 통해 매 시간마다 반복하여 사용자에게 서비스를 제공한다.



(그림 2) 서비스 흐름도

3.2 뉴스 요약 서비스

본 논문에서는 텍스트를 요약하기 위해 단순하면서도 효과적인 TextRank 알고리즘을 채용하였다[7].

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

(그림 3) Text Rank 알고리즘 수식

TextRank는 PageRank 알고리즘을 활용한 것으로 문서 내의 문장을 이용하여 문장의 Ranking을 계산하는 알고리즘이다. 이를 통해 Rank가 매겨진 상위의 문장을 설정한 비율로 선정하여 요약된 텍스트를 제공한다. TextRank 알고리즘을 사용하기 위하여 속도가 빠른 GENSIM 라이브러리를 이용하였다.

그림 4와 같이 요약된 기사는 데이터 베이스에 저장되며, 사용자가 등록한 키워드에 맞는 뉴스 요약 기사를 본문과 함께 제공한다.



(그림 4) 뉴스 요약기능 구현 결과

3.3 키워드 추출

개인 투자자들이 원하는 키워드의 뉴스를 제공하려면 뉴스의 본문에서 키워드를 추출해야 한다. 이 기능을 구현하기 위해 KR-WordRank 라이브러리를 사용하였다[8]. 비지도 학습 방법인 KR-WordRank 라이브러리는 한국어 텍스트에서 필요한 단어와 키워드를 자동으로 추출한다. 본 논문에서는 크롤링 된 뉴스 기사에서 KR-WordRank 알고리즘을 이용해 뉴스의 키워드를 추출 후 뉴스 요약 테이블에 저장해 사용자가 지정한 키워드의 뉴스를 제공한다.

3.4 감성 분석 서비스

현재 종목의 긍정 부정을 분석하기 위해 머신러닝을 통한 분석을 진행하였다. 본 논문에서는 LSTM 모델을 이용한 지도학습을 진행하였다. 지도학습을 위한 데이터는 비정형 감성사전을 구축하였다. 감성사전은 KNU 한국어 감성사전을 기반으로 표준국어대사전, 고려한국어대사전을 기준으로 진행하였다. 또한, 계층적 샘플링을 진행하여 학습 및 테스트를 진행하였다. 학습을 위한 데이터 전처리 과정으로는 의미 없는 단어 제거, 토큰화 및 정수 인코딩, 문장 패딩 및 최대 시퀀스 길이 결정의 단계로 나뉘어 진행하였다.

학습의 정확도를 위해 의미 없는 데이터를 제거하였으며, Konlpy의 Mecab 함수를 이용해 부울 용어를 제거 후 컴퓨터가 이해할 수 있는 정수로 인코딩을 진행하였다. 마지막으로 제각각인 문장의 길이에 대한 최대 시퀀스를 설정하였다.

이를 통하여 테스트 데이터의 정확도는 약 96%가 나왔으며 이 모델을 통하여 종목토론실의 감성을 분석해 그림 5와 같이 사용자에게 제공한다. 또한 종목토론실의 인기 키워드를 한눈에 보여주기 위해 Konlpy의 Okt 함수를 사용하였다. 전체 종목에 대해 모든 게시글을 Okt를 사용하여 형태소 분석 후 KR-WordRank를 통해 추출하였다. 이는 그림 5와 같이 종목토론실의 감성 분석 기능과 함께 제공한다.



(그림 5) 감성분석 기능 구현 결과

4. 결론

전통적 데이터만을 제공하던 기존의 서비스들과 달리 본 논문에서 제시한 서비스는 사용자에게 요약된 금융 투자 정보를 제공한다. 또한 대체 데이터인 투자자들의 감성분석 결과를 이용하여 투자에 나설 수 있게 도와준다. 이로 인해, 과잉 정보를 적절하게 판별하여 나눌 수 없는 개인 투자자들에게 다른 투자자들의 관점을 확인시킴으로써 투자정보를 신중하게 판단하고 이용할 수 있도록 일조한다. 향후, 증권사의 보고서와 DART의 공시 자료 등 전통적 금융 투자에 대한 관점 데이터를 확충함과 동시에 다양한 대체 데이터를 이용하여 제안 서비스를 고도화할 계획이다. 그러면 신규 개인 투자자들이 투자에 필요한 정보를 보다 엄격하게 검증하여 투자할 수 있도록 도움을 주어 개인 투자자들의 열악한

수익률을 개선할 수 있을 것으로 기대한다.

참 고 문 헌

- [1] 김민기, 김준석. *코로나19 국면의 개인투자자: 투자행태와 투자 성과*. 자본시장연구원, 2021.
- [2] 퀀트대디, "대체 테이터와 퀀트 투자." 금융 머신러닝의 시대, Oct 17, 2021 접속, <https://brunch.co.kr/@stochastic73/6>.
- [3] 구글 클라우드, "자연어 처리란?", Oct 17, 2021 접속, <https://cloud.google.com/learn/what-is-natural-language-processing?hl=ko>.
- [4] Yngie, "감성 분석 (Sentiment Analysis)." 2021 접속, https://yngie-c.github.io/nlp/2020/07/31/sentiment_analyses/.
- [5] 종토넷, "종토넷." Oct 17, 2021 접속, <https://jongto.net/>.
- [6] 빅카인즈, "빅카인즈." Oct 17, 2021 접속, <https://www.bigkinds.or.kr/>.
- [7] Mihalcea, Rada, and Paul Tarau. "Texttrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. pp.404-411, 2004.
- [8] Ssoyoung, "[NLP]자연어처리_키워드추출." Oct 17, 2021 접속, <https://soyoung-new-challenge.tistory.com/45>.
- [9] 온병원, 박상민, 나철원, "KNU 한국어 감성사전." Oct 17, 2021 접속, <http://dilab.kunsan.ac.kr/knusl.html>.
- [10] konlpy.org, "KoNLPy: 파이썬 한국어 NLP." Oct 17, 2021 접속, <https://konlpy.org/ko/latest/>.