(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0098427
(43) 공개일자 2022년07월12일

(51) 국제특허분류(Int. Cl.)
G06Q 40/04 (2012.01) *G06F 40/284* (2020.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06Q 40/04 (2013.01)
G06F 40/284 (2020.01)
(21) 출원번호 10-2021-0000135
(22) 출원일자 2021년01월04일
심사청구일자 2021년01월04일

(71) 출원인
연세대학교 원주산학협력단
강원도 원주시 흥업면 연세대길 1
(72) 발명자
안재준
서울특별시 동작구 보라매로5가길 7 캐릭터그린빌 1903호
탁근주
경기도 광주시 태봉로 61 2단지성원아파트 202동 1305호
(뒷면에 계속)
(74) 대리인
오영진

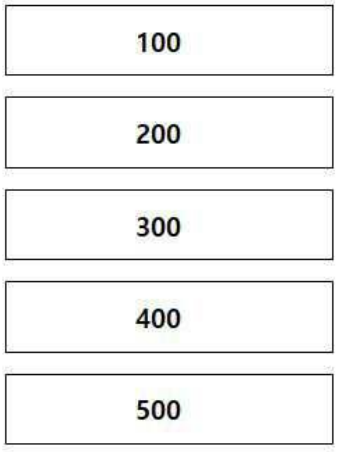
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **특정 도메인의 감성사전을 활용한 주가 등락 예측 장치**

(57) 요 약

본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치는 미리 설정된 기간 동안의 주가지수 및 환율정보를 포함하는 일별 수치데이터를 수집하는 수치데이터 수집모듈; 미리 설정된 기간 동안의 일별 뉴스기사를 수집하는 뉴스기사 수집모듈; 상기 일별 뉴스기사의 제목에 기초하여 일별 감성지수를 산출하는 감성지수 산출모듈; 상기 일별 수치데이터 및 상기 일별 감성지수에 기초하여 데이터셋을 생성하는 데이터셋 생성모듈; 및 상기 데이터셋 생성모듈에 기계학습을 수행하여 t일 시점 대비 t+1일 시점의 종가의 등락을 예측하는 등락예측 알고리즘을 생성하는 등락예측 알고리즘 생성모듈;을 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 20/00 (2021.08)
(72) 발명자
최세환
서울특별시 동대문구 한천로8가길 17-1

박성종
경기도 고양시 일산동구 산두로 54 정발마을3단지 건영빌라 302동 104호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711118846
과제번호	2020-51-0194
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	정형 시계열데이터 분석을 위한 합성곱 기반의 새로운 딥러닝 모형 개발
기 여 율	1/1
과제수행기관명	연세대학교 원주산학협력단
연구기간	2020.06.01 ~ 2021.02.28

명 세 서

청구범위

청구항 1

미리 설정된 기간 동안의 주가지수 및 환율정보를 포함하는 일별 수치데이터를 수집하는 수치데이터 수집모듈;

미리 설정된 기간 동안의 일별 뉴스기사를 수집하는 뉴스기사 수집모듈;

상기 일별 뉴스기사의 제목에 기초하여 일별 감성지수를 산출하는 감성지수 산출모듈;

상기 일별 수치데이터 및 상기 일별 감성지수에 기초하여 데이터셋을 생성하는 데이터셋 생성모듈; 및

상기 데이터셋 생성모듈에 기계학습을 수행하여 t일 시점 대비 t+1일 시점의 종가의 등락을 예측하는 등락예측 알고리즘을 생성하는 등락예측 알고리즘 생성모듈;

을 포함하는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치.

청구항 2

청구항 1에 있어서, 상기 감성지수 산출모듈은,

미리 구축된 주식 관련 도메인에 특화된 감성사전을 저장하는 감성사전 저장유닛;

상기 감성사전에 기초하여 상기 일별 뉴스기사의 제목에 극성을 부여하는 라벨링유닛;

극성이 부여된 상기 일별 뉴스기사를 BERT 모형에 적용 가능하도록 전처리하는 전처리유닛;

전처리된 상기 일별 뉴스기사를 상기 BERT 모형에서 지도학습을 수행하여 일별 감성지수를 산출하는 감성지수 출력유닛;

을 포함하는 것을 특징으로 하는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치.

청구항 3

청구항 2에 있어서,

상기 전처리유닛은 상기 일별 뉴스기사를 토큰 임베딩, 세그먼트 임베딩 및 포지션 임베딩으로 구분되도록 전처리를 수행하는 것을 특징으로 하는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치.

청구항 4

청구항 2에 있어서,

상기 감성지수 출력유닛은, 일별 상기 뉴스기사의 긍정도의 평균값을 해당일의 감성지수로 산출하는 것을 특징으로 하는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치.

청구항 5

청구항 1에 있어서,

상기 데이터셋 생성모듈은 t-4일부터 t일까지의 5일간의 거래일 정보를 활용하기 위하여 슬라이딩 윈도우 기법을 적용하여 상기 데이터셋을 생성하는 것을 특징으로 하는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치.

청구항 6

청구항 1에 있어서,

상기 등락예측 알고리즘 생성모듈은, XGBoost 모델에서 상기 데이터셋을 학습시켜 생성되는 것을 특징으로 하는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 주가 등락 예측 장치에 관한 것으로 구체적으로는 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치에 관한 것이다.

배 경 기 술

[0003] Covid-19의 여파로 경기 침체가 장기화되면서 국내 주식시장에 2030세대의 투자가 급증하는 새로운 현상이 나타나고 있으며, 특히 2020년 기준 20대 신규 투자자의 비중은 26%로, 과거 2년 평균 비중(22.9%)에 비해 늘었으며, 빚을 내면서까지 주식투자에 참여하는 현상을 일컫는 이른바 '빚투' 현상이 급증하고 있다.

[0004] 투자에 대한 잘못된 접근으로 인한 사례도 속출하면서 투자 열풍이 새로운 사회적 문제로 떠오르고 있지만, 한편으론 주식시장에 대한 관심도 이전에 비해 증가한 모습도 나타나고 있다.

[0005] 주식시장은 기업의 내부적 요인뿐만 아니라 금리, 환율과 같은 거시경제적 변수와 국내·외 정치 및 경제적 상황과 같은 외부적인 요인에도 영향을 받기 때문에 주가를 예측하는 연구는 다양한 접근법으로 시도되어 왔다.

[0006] 하지만 주가가 매 시점 이미 새로운 정보를 반영하고, 독립적으로 움직이는 확률보행적 특성을 가지고 있다는 점에서, 여전히 주식시장의 미래를 예측하는데에는 많은 어려움이 따르고 있다.

[0007] 이러한 문제점을 극복하기 위한 방안으로 기존에는 사용하지 않던 다양한 형태의 데이터를 주가 예측 분야에 활용하고 있는데, 특히 데이터 저장 및 처리장치의 발전과 인공지능 기술의 발달로 온라인 뉴스나 소셜 미디어 정보같은 빅데이터를 활용하는 사례가 많아지고 있다.

[0008] 뉴스를 통해 사회에서 발생하는 현상에 대한 설명이나 관련 정보를 획득할 수 있다는 점에서 뉴스와 주가는 밀접한 관계를 가진 것으로 볼 수 있으며, 뉴스기사는 개인의 심리와 사회적 분위기에 대한 정보를 담고 있기 때문에 시장의 상황을 대변하고 있다고도 볼 수 있다.

[0009] 대부분의 기존 연구들은 다양한 기술적 지표들의 조합을 적용하여 분석에 이용하거나 거시 경제지표 등의 정보를 더하는 정도에 머무르는 상태이며, 특히 금융 연구분야에서는 여러 금융 지표들을 분석할 때 뉴스나 SNS 등의 비정형 정보에 대한 영향을 고려한 연구는 드문 상황이다.

[0010] 인공지능 분야의 이론 및 기술적 발전으로 머신러닝 및 딥러닝 모형이 주가 및 방향성 예측에 사용되고 있지만 수많은 시장의 변수가 직간접적으로 서로 영향을 미치며, 불규칙적으로 변동하기 때문에 이를 이용한 주가 및 방향성 예측에는 한계가 존재한다.

[0011] 한편, 하기 선행기술문헌은 사용자 참여형 주가 예측 서비스 제공 방법에 관한 내용을 개시하고 있으며, 본 발명의 기술적 요지는 개시하고 있지 않다.

선행기술문헌

특허문헌

[0013] (특허문헌 0001) 대한민국 공개특허공보 제10-2015-0094923호

발명의 내용

해결하려는 과제

[0014] 본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치는 전술한 문제점을 해결하기 위하여 다음과 같은 해결과제를 목적으로 한다.

[0015] 종래 대비 더욱 정확하게 주가의 등락을 예측할 수 있는 주가 등락 예측 장치를 제공하는 것이다.

[0016] 본 발명의 해결과제는 이상에서 언급된 것들에 한정되지 않으며, 언급되지 아니한 다른 해결과제들은 아래의 기재로부터 당해 기술분야에 있어서의 통상의 지식을 가진 자에게 명확하게 이해되어 질 수 있을 것이다.

과제의 해결 수단

[0018] 본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치는 미리 설정된 기간 동안의 주가지수 및 환율정보를 포함하는 일별 수치데이터를 수집하는 수치데이터 수집모듈; 미리 설정된 기간 동안의 일별 뉴스기사를 수집하는 뉴스기사 수집모듈; 상기 일별 뉴스기사의 제목에 기초하여 일별 감성지수를 산출하는 감성지수 산출모듈; 상기 일별 수치데이터 및 상기 일별 감성지수에 기초하여 데이터셋을 생성하는 데이터셋 생성모듈; 및 상기 데이터셋 생성모듈에 기계학습을 수행하여 t일 시점 대비 t+1일 시점의 종가의 등락을 예측하는 등락예측 알고리즘을 생성하는 등락예측 알고리즘 생성모듈;을 포함한다.

[0019] 상기 감성지수 산출모듈은, 미리 구축된 주식 관련 도메인에 특화된 감성사전을 저장하는 감성사전 저장유닛; 상기 감성사전에 기초하여 상기 일별 뉴스기사의 제목에 극성을 부여하는 라벨링유닛; 극성이 부여된 상기 일별 뉴스기사를 BERT 모형에 적용 가능하도록 전처리하는 전처리유닛; 전처리된 상기 일별 뉴스기사를 상기 BERT 모형에서 지도학습을 수행하여 일별 감성지수를 산출하는 감성지수 출력유닛;을 포함하는 것이 바람직하다.

[0020] 상기 전처리유닛은 상기 일별 뉴스기사를 토큰 임베딩, 세그먼트 임베딩 및 포지션 임베딩으로 구분되도록 전처리를 수행하는 것이 바람직하다.

[0021] 상기 감성지수 출력유닛은, 일별 상기 뉴스기사의 긍정도의 평균값을 해당일의 감성지수로 산출하는 것이 바람직하다.

[0022] 상기 데이터셋 생성모듈은 t-4일부터 t일까지의 5일간의 거래일 정보를 활용하기 위하여 슬라이딩 윈도우 기법을 적용하여 상기 데이터셋을 생성하는 것이 바람직하다.

[0023] 상기 등락예측 알고리즘 생성모듈은, XGBoost 모델에서 상기 데이터셋을 학습시켜 생성되는 것이 바람직하다.

발명의 효과

[0025] 본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치는 주가와 관련된 수치정보 및 뉴스기사정보가 추가된 기계학습모형을 활용하되, 뉴스기사정보의 경우 도메인에 특화된 감성사전을 활용하여 감성지수를 산출함으로써 종래 대비 높은 예측 성능을 발휘할 수 있는 효과가 있다.

[0026] 본 발명의 효과는 이상에서 언급된 것들에 한정되지 않으며, 언급되지 아니한 다른 효과들은 아래의 기재로부터 당해 기술분야에 있어서의 통상의 지식을 가진 자에게 명확하게 이해되어질 수 있을 것이다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치를 간략히 도시한 블록도이다.

도 2는 도 1의 수치데이터 수집모듈에 의하여 수집된 데이터의 일 예를 도시한 도표이다.

도 3은 도 1의 뉴스기사 수집모듈에 의하여 수집된 데이터의 일 예를 도시한 도표이다.

도 4는 도 1의 감성지수 산출모듈의 세부 구성을 간략히 도시한 블록도이다.

도 5는 도 4의 감성사전 저장유닛에 저장된 감성단어의 일 예를 도시한 도표이다.

도 6 내지 도 10은 본 발명의 일 실시예에 따른 특정 도메일의 감성사전을 활용한 주가 등락 예측 장치에서 적용되는 BERT 모형을 설명하기 위한 이미지이다.

도 11은 도 1의 감성지수 산출모듈의 산출물을 도시한 도표이다.

발명을 실시하기 위한 구체적인 내용

[0029] 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하되, 도면 부호에 관계없이 동일하거나 유사한 구성 요소는 동일한 참조 번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.

[0030] 또한, 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다. 또한, 첨부된 도면은 본 발명의 사상을 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 발명의 사상이 제한되는 것으로 해석되어서는 아니 됨을 유의해야 한다.

[0032] 이하 도 1 내지 도 11을 참조하여 본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치에 대하여 설명하도록 한다.

[0033] 본 발명의 일 실시예에 따른 특정 도메일의 감성사전을 활용한 주가 등락 예측 장치는 도 1에 도시된 바와 같이 수치데이터 수집모듈(100), 뉴스기사 수집모듈(200), 감성지수 산출모듈(300), 데이터셋 생성모듈(400) 및 등락 예측 알고리즘 생성모듈(500)을 포함하도록 구성된다.

[0034] 수치데이터 수집모듈(100)은 미리 설정된 기간 동안의 주가지수 및 환율정보를 포함하는 일별 수치데이터를 수집하는 기능을 수행하며, 이러한 수치데이터로는 도 2에 도시된 바와 같이 코스피 지수의 시가, 고가, 저가, 종가, 거래량 및 기술적 지표(MACD, OBV, CCI)와, 거시경제지표 변수로 활용할 수 있는 원-달러 환율, 원-위안 환율, 원-엔 환율 등을 들 수 있다.

[0035] 뉴스기사 수집모듈(200)은 미리 설정된 기간 동안의 일별 뉴스기사를 수집하는 기능을 수행하며, 예를 들면 국내 대표 포털사이트인 NAVER에서 증권 섹션으로 범위를 한정한 후 수집기간을 지정하고 '코스피'를 검색어로 입력했을 때 검색된 뉴스기사를 수집한 결과물은 도 3과 같다.

[0036] 감성지수 산출모듈(300)은 뉴스기사 수집모듈(200)에서 수집한 일별 뉴스기사의 제목에 기초하여 일별 감성지수를 산출하는 기능을 수행하는데, 이러한 감성지수 산출모듈(300)은 일별 뉴스기사의 제목을 BERT 모델을 활용하여 감성분석을 수행한다.

[0037] BERT(Bidirectional Encoder Representations from transformers) 모델은 2018년 Google에서 개발한 자연어처리 사전 훈련 모형이며, 기존에 자연어처리에 활용하였던 통계기법, 딥러닝 모형들보다 좋은 성능을 내고 있기 때문에 많은 자연어 처리분야에서 각광을 받고 있는 모형이다.

[0038] BERT의 구조는 Transformer에 기반을 두며, Transformer의 Encoder 부분만을 사용하는 구조로서, 도 6은 Transformer의 구조이며, 인코더-디코더 모델을 나타내고 있다.

[0039] Transformer는 기존 인코더-디코더 모델과는 다르게 합성곱신경망이나 순환신경망 대신에 Self-Attention 개념을 도입했다는 특징이 있다.

[0040] BERT는 도 7에 도시된 Transformer의 인코더 블록 N개를 가지고 있으며, Base모델은 12개, Large모델은 24개로 구성되어 있다.

[0041] 인코더 블록은 입력 시퀀스를 N번 만큼 재귀적으로 반복처리하며 동일한 연산을 수행한 후에 입력값을 더해주어 기존의 학습된 정보를 보존하고, 추가적으로 학습을 수행하는 형태인 Residual Connections로 처리하는 구조로 구성되어 있다.

[0042] 이때 BERT는 Transformer와는 달리 입력시퀀스에 대하여 Positional Encoding을 사용하지 않고, Token Embedding, Segment Embedding, 그리고 Position Embedding 3개의 Embedding을 합산한 결과를 취하여 인코더 블록의 입력값으로 한다.

[0043] 도 8은 Multi-Head Attention 부분의 내부 구조를 도식화한 형태이며, 서로 다른 가중치 행렬을 이용하여 Attention을 h번 계산하고 이를 연결한 결과를 가지며, 하기 수식 (1)과 같이 표현되어진다.

[0045] <수식 (1)>

[0046] $$MultiHead(Q, K, V) = [head_1; \ldots; head_h] W^O$$
$$where,\ \ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

[0048] Transformer에서 Q는 디코더의 Hidden state, K는 인코더의 Hidden state, V는 K에 Attention을 부여받은 Normalized Weights를 의미하며, 초기화은 V와 K가 동일하게 적용된다.

[0049] 하지만 BERT의 경우에는 같은 구성으로 이루어진 입력값 Q, K, V가 각각 다른 초기화 과정을 거치게 되고, 이러한 구성을 통해 BERT에서는 동일한 토큰이 문장 내에 있는 다른 토큰에 대한 Self-Attention 효과를 가지게 된다.

[0050] Sefl-Attention은 Q=K=V인 경우의 attention을 의미하며, scaled Dot-Product를 수행하고 softmax를 취해주면 feed-forward 신경망에 입력시킬 수 있는 형태의 벡터가 출력되는데, 하기 수식 (2)는 출력되는 Attention score에 대한 수식을 나타내며, 이때 dk는 k의 차원을 나타낸다.

[0052] <수식 (2)>

[0053] $$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})^V$$

[0055] 예를 들어, BERT base모델의 경우 h=12이므로 전체 토큰을 12등분으로 분리시키고, scaled Dot-Product Attention을 적용하여 합치는 과정을 거치게 되며, 최종적으로 어텐션 결과를 Feed-forward Network로 통과시켜 가장 높은 확률을 지닌 값을 최종 결과값으로 하여 산출하게 된다

[0056] 상술한 BERT 모형을 적용하여 감성분석을 수행하는 감성지수 산출모듈(300)은 구체적으로 도 4에 도시된 바와 같이 감성사전 저장유닛(310), 라벨링유닛(320), 전처리유닛(330) 및 감성지수 출력유닛(340)을 포함하도록 구성될 수 있다.

[0057] 감성사전 저장유닛(310)은 미리 구축된 주식 또는 금융 관련 도메인에 특화된 감성사전을 저장하는 기능을 수행하며 도메인 정보가 고려된 감성사전 내의 감성단어는 도 5에 도시된 바와 같다.

[0058] 라벨링유닛(320)은 감성사전 저장유닛(310)에 저장된 감성사전에 기초하여 일별 뉴스기사의 제목에 극성을 부여하는 기능을 수행하며, 구체적으로 뉴스기사를 어절단위로 분리시킨 단어들에 대하여 극성을 부여하게 된다.

[0059] 전처리유닛(330)은 극성이 부여된 일별 뉴스기사를 BERT 모형에 적용 가능하도록 전처리를 수행하는 구성으로, 일별 뉴스기사의 문장을 도 9에 도시된 바와 같이 Token embedding, Segment embedding, Position embedding 세 가지의 embedding으로 표현시키도록 전처리를 수행한다.

[0060] Token embedding은 Word piece embedding 방식을 사용하여 각 토큰의 의미가 표현되며, Segment embedding은 [SEP]라는 구분자로 구분하여 서로 다른 문장 간 구분되어져 있는 의미를 내포하고 있는 임베딩 표현이며, Position embedding은 단어들의 절대적인 위치 정보를 나타내는 임베딩 표현이다.

[0061] 감성지수 출력유닛(340)은 전처리된 일별 뉴스기사를 BERT 모형에서 지도학습을 수행하여 일별 감성지수를 산출하기 위한 감성지수예측 알고리즘을 생성하는 기능을 수행한다.

[0062] 즉, 감성지수 출력유닛(340)에서는 3개의 임베딩 표현으로 변환된 뉴스기사제목들은 라벨링된 극성과 함께 지도학습 과정을 거치며 입력값으로 들어가는 해당 기사제목이 긍정적인 기사일 경우의 확률값을 출력하게 된다.

[0063] 도 10은 BERT를 활용한 여러가지 전이학습 방법 중 본 발명에서 수행하는 과정과 같은 Single sentence classification task를 수행하는 과정이 도식화된 그림이며, 최종 출력된 기사별 확률값들은 도 11에 도시된 바와 같이 출력되며, 일자별로 확률값들의 평균을 구하여 각 일자별 감성지수를 출력하게 된다.

[0064] 데이터셋 생성모듈(400)은 수치데이터 수집모듈(100)에서 수집한 일별 수치데이터 및 감성지수 산출모듈(300)에서 산출한 일자별 뉴스기사의 감성지수에 기초하여 주가의 등락을 예측하기 위한 등락예측 알고리즘을 생성하는 기능을 수행한다.

[0065] 특히 본 발명에서는 5일간의 코스피 지수의 시가, 고가, 저가, 종가, 거래량 및 기술적지표, 해당 거래일의 환율 변수와 일자별 감성지수의 조합을 통해 t시점 대비 t+1시점 종가의 등락을 예측하기 위하여 Tree 기반 Boosting 기법인 XGBoost를 사용한다.

[0066] XGBoost(eXtreme Gradient Boosting)는 2016년 chen과 Guestrin이 제시한 Tree기반의 부스팅 기법을 개선한 알고리즘으로, 부스팅은 약한 분류기를 순차적으로 개선해 나감으로써 강한 분류기를 만들어 나가는 배깅 기법과 더불어 대표적인 앙상블 기법 중 하나이다.

[0067] 기존 부스팅 기법의 경우에는 순차적인 학습 방법으로 인해 계산시간이 오래걸린다는 단점이 존재했지만, XGBoost는 계산을 병렬로 처리시키는 구조로 구성함으로써 기존의 단점을 극복하였으며, 또한 CART(Classification And Regression Tree)모형을 기반으로 분류기를 생성하기 때문에 범주형, 연속형 자료에 대한 학습이 모두 가능한 모형이다.

[0068] 하기 수식 (3)은 tree기반 모델의 앙상블 모형을 표현하는 수식으로 $\hat{y}_i$ 는 입력 x_i 에 대한 예측값을 의미하며, f_k 는 알고리즘에서 생성한 결정트리를 의미하며, K는 전체 트리의 개수이며 F는 생성 가능한 결정트리의 집합을 나타낸다.

[0070] <수식 (3)>

[0071] $$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F$$

[0073] 수식 (3)을 이용하여 XGBoost 모형의 목적함수를 표현하면 하기 수식 (4)와 같이 나타낼 수 있다.

[0075] <수식 (4)>

[0076] $$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

[0078] 첫 번째 항인 $l(y_i, \hat{y}_i)$ 은 손실함수로 실제값인 y_i 와 예측값인 $\hat{y}_i$ 의 차이를 나타내며, 두 번째항인 $\Omega(f_k)$ 는 regularization term으로 모형의 복잡성을 정의하는 매개변수 항이 되며, 모형의 복잡성을 결정하는 요인은 트리의 깊이와 노드의 개수, 그리고 노드의 점수가 된다.

[0079] XGBoost 모형은 파라미터 조정을 통해 데이터의 특성과 목적에 맞게 설정할 수 있다. 먼저 부스터의 종류 및 병렬처리에 활용할 스레드 개수 등을 설정할 수 있으며, 부스팅 파라미터로써 학습률, tree의 깊이, 노드의 개수 등 tree의 구조와 관련된 값을 지정할 수 있다.

[0080] 나아가 마지막으로 출력되는 값의 형태에 따라 목적함수의 종류도 다양하게 설정할 수 있다.

[0081] 본 발명의 경우 2017년 1월 2일부터 2019년 12월 30일까지의 데이터는 학습을 위한 훈련 데이터로 사용하였고, 2020년 1월 2일부터 2020년 6월 30일까지의 데이터는 검증을 위한 평가 데이터로 사용하였다.

[0082] 또한 t+1시점의 종가의 등락을 예측하기 위하여 데이터셋 생성모듈(400)은 t-4시점부터 t시점까지 5일간의 거래일의 정보를 활용하기위해 Sliding Window 기법을 적용하여 데이터셋을 구축하였다.

[0083] XGBoost의 매개변수 최적화는 Gridsearch Optimization을 통해 최적화 하였으며, 최적화된 파라미터는 하기 표 1과 같다.

표 1

[0085]

Parameter	Value
Booster	gbtree
eta(learning rate)	0.02
max_depth	6
Subsample	0.5
colsample_bytree	0.7
Objective	binary:logistic

[0087] 상기 변수 조합으로 이루어진 데이터셋을 이용하여 실험한 결과, 훈련 데이터셋의 정확도는 92%, F1 score는 0.914를 나타냈으며 평가 데이터셋의 경우 정확도는 82%, F1 score 0.84를 나타내었다.

[0088] 반면 범용 감성사전을 이용하여 라벨링 된 기사제목을 이용하여 일자별 감성지수를 만들어 변수로 활용하였을 때, 훈련 데이터셋의 경우에는 76%의 정확도와 0.79의 F1 score, 평가 데이터셋의 결과로는 57%의 정확도와 0.64의 F1 score를 보였다.

[0089] 마지막으로 뉴스기사 정보를 활용하여 추가된 변수가 모형의 성능에 영향을 미치는지 검증하기 위해 일자별 감성지수 변수를 제거하고 나머지 변수만으로 실험한 결과, 훈련 데이터셋은 60%의 정확도와 0.66의 F1 score를 보였으며, 평가 데이터셋의 경우에는 53%의 정확도와 0.60의 F1 score를 나타냈다. 각 변수조합에 따른 평가 데이터셋에 대한 정확도와 F1 score 및 민감도와 특이도 결과는 하기 표 2와 같다.

표 2

[0091]

	Accuracy	F1 score	sensitivity	specificity
Case1 도메인 특화 감성지수 추가	0.823	0.853	0.800	0.863
Case2 범용 감성지수 추가	0.579	0.642	0.608	0.533
Case3 수치형 변수 조합만 이용	0.521	0.476	0.619	0.467

[0093] 본 발명의 일 실시예에 따른 특정 도메인의 감성사전을 활용한 주가 등락 예측 장치는 기술적 지표 및 거시경제 변수와 도메인에 특화된 감성사전을 이용한 뉴스기사제목 감성분석을 수행하여 산출된 결과를 일별 감성지수로 변수에 추가하여 코스피 지수의 방향성을 예측할 수 있는 장치를 제시한다.

[0094] 실험 결과, 단순히 수치적 정보만을 이용한 기계학습 모형보다는 뉴스기사정보가 추가된 기계학습모형에서 더욱 우수한 성능을 나타내었으며, 범용 감성사전을 이용하여 감성분석을 수행한 결과를 활용한 경우보다는 도메인에 특화된 감성사전을 활용하여 산출된 일별 감성지수 변수가 추가되었을 때 더욱 높은 예측 성능을 나타내었다.

[0095] 본 발명에서 뉴스기사를 분석하여 주가 예측을 시도하는데 있어 관련 업계 종사자 및 전공자를 대상으로 주식시장에 대한 특성이 고려된 감성사전을 구축하고, 이를 감성 분석에 활용하였을 때, 모형의 예측 성능이 향상됨을 확인하였다.

[0096] 이러한 결과는 감성을 분류하는 연구에서는 분석의 대상 또는 분석의 목적이 고려된 감성사전을 구축하고, 활용해야 한다는 점을 시사하고 있다.

[0097] 기존에 사용된 통계적 기법 및 기계 학습을 이용한 사전 구축 방법은 문맥과 언어의 특성을 제대로 반영하지 못하고 있기 때문에 감성분석 시, 부정확한 극성정보가 담긴 텍스트 정보를 이용하게 된다.

[0098] 이러한 이유로 특정한 주제 내에서 텍스트 분석이 수행될 경우에는 감성분석에 활용되는 감성사전의 더욱 정밀하게 작성되어야 한다.

[0099] 그리고 본 연구결과에서 시사하는 바와 같이 인간의 감정과 사회문화적 뉘앙스를 담고 있는 텍스트의 극성이 정밀하게 라벨링이 될 경우 텍스트 분석을 수행할 때 텍스트가 담고 있는 정보를 효과적으로 사용할 수 있다는 점을 확인할 수 있다.

[0101] 본 명세서에서 설명되는 실시예와 첨부된 도면은 본 발명에 포함되는 기술적 사상의 일부를 예시적으로 설명하는 것에 불과하다. 따라서 본 명세서에 개시된 실시예들은 본 발명의 기술적 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이므로, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것이 아님은 자명하다. 본 발명의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당해 기술분야에 있어서의 통상의 지식을 가진 자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시예는 모두 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0103] 100: 수치데이터 수집모듈

200: 뉴스기사 수집모듈

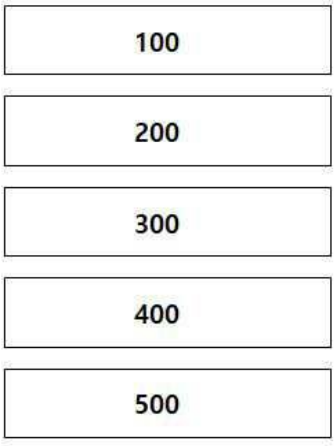
300: 감성지수 산출모듈

400: 데이터셋 생성모듈

500: 등락예측 알고리즘 생성모듈

도면

도면1



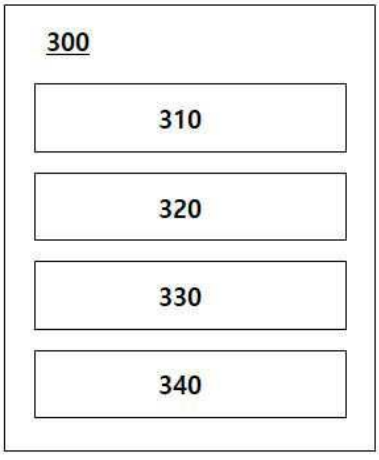
도면2

일자	시가	고가	저가	종가	거래량	MACD	CCI	OBV	원-달러 환율	원-위안 환율	원-엔 환율
2017-01-02	2022.23	2031.79	2015.68	2026.16	229874000	8.63	-4.69	355988 24000	1208	173.1	1033.49
2017-01-03	2034.31	2044.07	2028.47	2043.97	268127000	9.13	59.09	358669 51000	1203.5	172.76	1025.7
2017-01-04	2046.29	2046.29	2040.61	2045.64	371488000	9.54	93.35	362384 39000	1206.4	173.83	1022.07
					⋮						
2020-06-26	2137.47	2142.04	2115.25	2134.65	761043000	32.37	-13.34	667935 42000	1200.6	169.54	1121.06
2020-06-29	2105.54	2120.5	2087.84	2093.48	643155000	27.05	-94.9	661503 87000	1198.6	169.61	1119.3
2020-06-30	2124.38	2134.38	2108.26	2108.33	708604000	23.77	-61.87	668589 91000	1203	170.13	1116.53

도면3

date	url	news_title
2017.01.01	https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=015&aid=0003708240	[대도약2017] 코스피지수 최고치 2350선 기대…IT·수출주가 상승 견인할 듯
2017.01.01	https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=018&aid=0003715367	[펀드와치]'1월 효과' 기대에 중소형株 펀드 '방긋'
2017.01.01	https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=009&aid=0003865651	[위클리 펀드] 국제유가 상승에 브라질펀드 5%↑
		⋮
2020.06.30	https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=277&aid=0004708391	작년 기업 감사보고서 정정 1319건...전년比 14%↓
2020.06.30	https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=018&aid=0004676554	[마켓인]'코로나 국면에 승부수'…부동산 자산운용사 3色 경쟁
2020.06.30	https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=009&aid=0004605886	펀드 양도세공제 '全無'…간접투자 고사위기

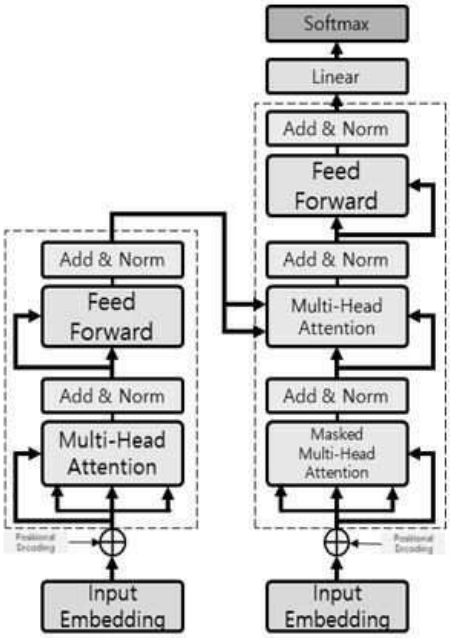
도면4



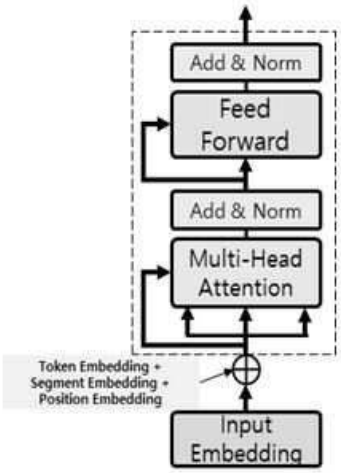
도면5

긍정 단어	부정 단어
장밋빛	주가하락
호황에	갑질
회복세	경기민감
긍정적	경제위기
기대감	고민
기대되는	곤두박질
⋮	⋮
효과에	폭풍
후끈	폭풍전야
훨훨	피난처
흥행	하락에도
힘받나	하락장서도
힘받는	하락폭

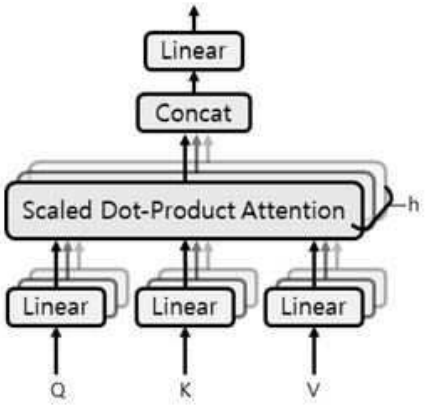
도면6



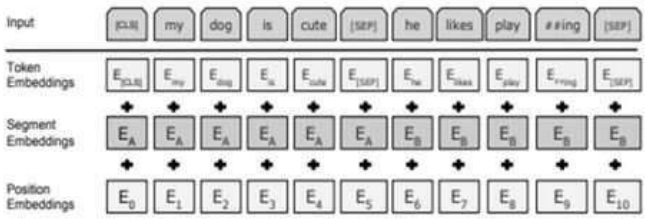
도면7


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Token Embedding +
Segment Embedding +
Position Embedding
Input Embedding

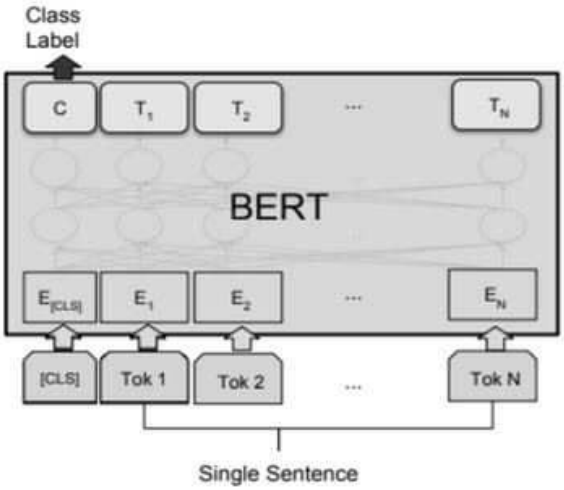
도면8


Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
Q
K
V

도면9


Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
Segment Embeddings
Position Embeddings

도면10



도면11

date	text	prob
2017.01.01	[대도약2017] 코스피지수 최고치 2350선 기대…IT·수출주가 상승 견인할 듯	0.977136
2017.01.01	[펀드와치]'1월 효과' 기대에 중소형株 펀드 '방긋'	0.890693
2017.01.01	[위클리 펀드] 국제유가 상승에 브라질펀드 5%↑	0.84537
	⋮	
2020.06.30	[김인경의 쬬!금융]마이너스 금리 4년..투자처 없는 日은행	0.073215
2020.06.30	투자처 노리는 예탁금 50兆 육박… 단기성 자금 공모주시장 달군다	0.580111
2020.06.30	코로나19 재확산공포… 코스피 2100선 무너져	0.025823