한양대학교

SW프로젝트 요약서

프로젝트 기간	**2023.03.01 - 2023.10.31 (총8개월)**
프로젝트 팀원	김나현**(컴퓨터소프트웨어학부, 4학년)** 이수빈**(컴퓨터소프트웨어학부, 4학년)**
지도교수	채동규 교수님
프로젝트 명	**Transformer, LSTM, SVM** 앙상블 모델과 뉴스 데이터와 기술 지표를 활용한 주가 예측 프로젝트
프로젝트 내용	Transformer, LSTM, SVM을 결합한 앙상블 모델과 뉴스 데이터와 기술 지표를 활용하여 주가 예측 문제를 해결하였다. 관련 선행 연구를 분석하면서 이전 논문에서 시도되지 않은 ML 모델과 데이터를 사용하지 않기 위해, {Ensemble 모델, 주가, 뉴스 감성 점수, 기술 지표}의 조합으로 ML 모델링 및 실험을 진행하였다.
기대효과 및	선행 연구에서 시도되지 않은 {Ensemble 모델, 주가, 뉴스 감성

개선방향	점수, 기술 지표}의 조합을 사용하여 ML 모델링을 진행하였다. 최종 모델 실험 결과 51%의 정확도를 얻게 되었다. 향후 연구를 통하여 모델의 개선방안을 논의하고 정확도를 향상시킬 수 있는 방안을 통해 모델의 Quality를 향상시킬 계획이다.

한양대학교

SW프로젝트 결과보고서

프로젝트명	**Transformer, LSTM, SVM** 앙상블 모델과 뉴스 데이터와 기술 지표를 활용한 주가 예측 프로젝트
프로젝트 요약	AI 기반의 주가 예측을 목표로 진행함. 이전 논문에서 시도되지 않은 ML 모델과 데이터를 사용하지 않기 위해, {Ensemble 모델, 주가, 뉴스 감성 점수, 기술 지표}의 조합으로 ML 모델링 및 실험을 진행함. 10/11월에 걸쳐 일정 기간 실험 결과 51%의 정확도를 얻어, 주가 예측에는 위의 조합을 사용할 수 없다는 결론으로 귀결됨.
프로젝트 기간	**2023.03.01 – 2023.10.31 (**총 **8**개월**)**
산출물	졸업 작품 (O), 졸업 논문 ( )

학과	학번	학년	이름	연락처

한양대학교

컴퓨터소프트웨어	2019080737	4	김나현	chtlaalways@gmail.com 010-5549-3498
컴퓨터소프트웨어	2019011449	4	이수빈	1092soobin2@gmail.com 010-2110-8266

한양대학교

목 차

1. 프로젝트 개요

1.1 프로젝트 목적 및 동기

- 프로젝트의 정의

우리가 선택한 졸업프로젝트 주제는 '인공지능 기반 주식 가격 예측 프로젝트'로서 관심 종목에 대한 주가 차트 및 외부 요인(뉴스 등)에 대한 다음 시점의 주식 가격이 상승할지 하락할지 등을 자동으로 예측하는 인공지능 모델을 연구하는 것이다.
우리는 Transformer, LSTM, SVM을 결합한 앙상블 모델을 이용하여 뉴스 데이터와 기술 지표를 활용한 주가 예측 프로젝트를 진행하였다.

- 프로젝트의 기술적 배경

- 딥러닝 기반 시계열 예측
    - AI 모델 : CNN, RNN, LSTM 등

- 프로젝트의 필요성

주식 가격은 여러 외부 요인이 개입되기 때문에 예측이 쉽지 않다. 따라서 인공지능을 이용하여 주식 가격을 예측하려는 시도가 2000년대 초반부터 꾸준히 진행되어 왔다. 초기에는 주식 시장에서의 패턴 인식을 위해 전통적인 통계 모델이 주로 사용되었으며 빅데이터의 등장과 함께 많은 양의 데이터를 처리할 수 있는 LSTM, CNN 모델 등이 적용되어 왔다.

선행 연구를 분석해 봤을 때, 주가에 영향을 미치는 요소는 매우 다양하며 주가 예측을 위해 사용할 수 있는 데이터의 종류는 다양하다는 것을 알 수 있었다. 기존의 많은 연구들은 주식 동향에 영향을 미치는 요인 중 하나인 텍스트 데이터 중 뉴스 데이터와 같은 단일한 변수를 선택하여 주가를 예측하였다.

그러나 실제 세계에서는 텍스트 데이터뿐만 아니라 거래량, 기술 지표, 펀더멘털 등

다양한 요소가 함께 작용하여 주식 동향이 결정된다. 각각의 요인들이 주가 분석에 영향을 미치는 정도가 다르므로, 여러 요인을 조합하여 주가 동향 분석에 사용할 경우 예측 정확도를 보완하고 시너지가 날 수 있을 것으로 판단하였다.

따라서 이벤트의 발생 및 투자 심리를 대변할 수 있는 텍스트 데이터와 기술 지표, 거래량 등이 포함된 시계열 데이터, 두 가지 종류의 데이터를 활용하기로 결정하였다.

또한 사용될 딥러닝 모델에 관한 선행 연구 분석을 통하여 자연어 처리에서 두각을 보이는 트랜스포머 모델을 주식 가격 예측에 활용하는 선례는 적으며, 트랜스포머 모델에 감성분석 점수와 기술 지표 데이터를 적용한 연구는 없음을 알 수 있었다.

따라서 우리는 자연어 처리에서 현재 두각을 보이는 모델인 트랜스포머 모델과, 전통적인 데이터 분석 모델인 SVM, LSTM을 결합한 Hard-Voting 앙상블 모델을 제안하고자 한다.

1.2 프로젝트 목표

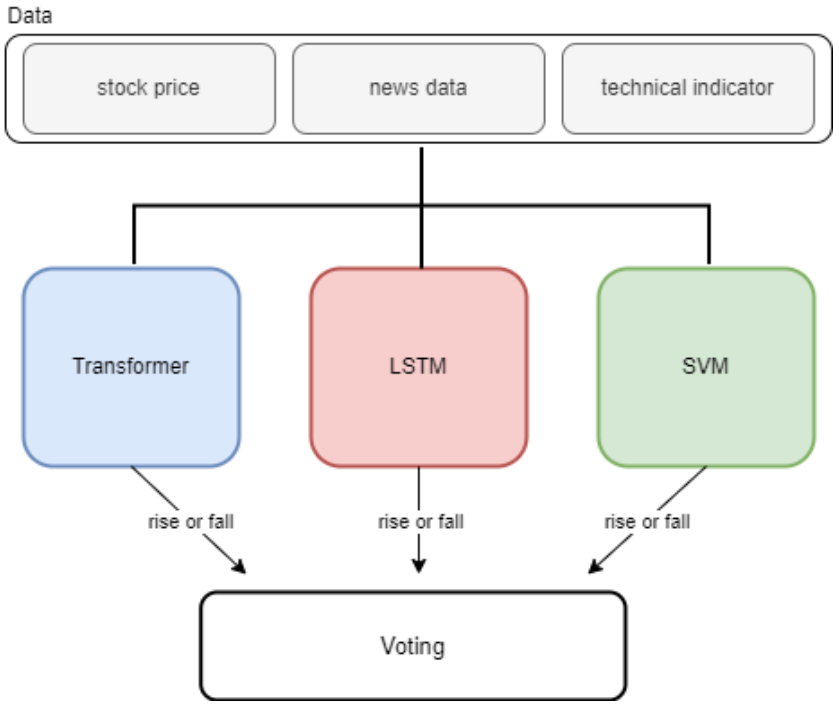
- 프로젝트에서 달성하고자 하는 최종목표는 기존 연구와 차별화된 방법(사용 모델, 적용 데이터 등)을 사용하여 정확도가 높은 모델을 구현하는 것이다.

- 프로젝트의 최종목표를 달성하기 위한 세부목표는 다음과 같다.

1. 기존 연구 분석을 통하여 딥러닝 주식 예측 모델을 이해하고 선행 연구의 한계점을 분석한다.
2. 기존에 없었던 연구 방향을 제시하고 코드로 구현한다.
    a. 기술 지표 데이터, 감성 분석 데이터를 얻어 모델에 적용한다.
    b. 정확도를 향상시키기 위한 다양한 방법을 실험하여 적용한다.

2. 프로젝트 내용

- 모델 디자인은 다음과 같다.



세부적으로는 먼저 도서, 논문, 인터넷 검색 등을 통한 기초 기술에 대한 조사를 진행하였고 다음과 같은 내용을 순차적으로 진행하였다.

1. 예측에 사용할 데이터 수집
    a. 예측 종목 선정
    b. 기술 지표 데이터 수집
    c. 감성 분석 데이터 수집
2. 개별 모델 구현
    a. 트랜스포머, LSTM, SVM 모델 구현
    b. 각 모델에 알맞는 데이터 전처리 방법 사용
    c. 하이퍼파라미터 튜닝
        i. 학습률(learning rate), 배치 크기(batch size), node size, Input-window Size, Output-window Size 등.
    d. 개별 모델 평가

3. 앙상블 모델 구현

    a. 트랜스포머, LSTM, SVM 모델 hard-voting

- 현재 기술의 한계성을 극복하기 위한 연구개발 추세

1. 데이터의 품질 및 다양성:

- 데이터 획득과 전처리: 정확하고 신뢰할 수 있는 데이터를 확보하기 위해 데이터 품질을 높이는 노력이 진행되고 있다. 또한 텍스트 데이터나 이미지 데이터 등의 비정형 데이터를 보다 효과적으로 활용하여 예측 모델의 성능을 향상시키는 연구가 진행되었다.
- 또한 금융 시장의 특성을 고려한 시계열 데이터 처리 방법 및 모델링에 대한 연구가 진행 중이다.

2. 모델의 복잡성 및 해석 가능성:

- 블랙박스로 간주되던 딥러닝 모델의 해석 가능성을 높이기 위한 연구가 활발히 이루어지고 있다. 모델의 의사 결정 과정을 설명하고, 예측에 영향을 미치는 요인들을 파악하는 것이 중요한 과제로 인식되고 있으며 '설명 가능한 인공지능 모델' 연구가 주목받고 있다.

3. 프로젝트의 기술적 내용

- SW 및 **HW** 개발환경 소개

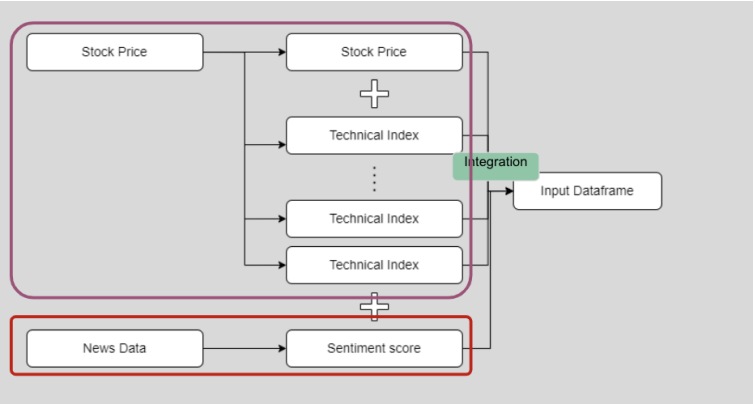
- 환경 및 에디터: Colab, VScode
- 앙상블 모델: Python 3.9.7, Pytorch, Keras
- 뉴스 데이터 수집: Python 3.9.7, NLTK, Beautiful Soup 4, Google Trans
- 웹페이지: AWS EC2, Github Pages, Javascript, Python, Flask

- 적용기술에 대한 소개 및 적용 방법

- Python: Pytorch, bs4 등 구현에 필요한 모듈이 다양하게 지원되는 AI 분야에서 널리 사용되는 언어
- Pytorch: 트랜스포머 모델을 구현하기 위한 Library
- Keras: LSTM, SVM 모델을 구현하기 위한 Library
- NLTK: 뉴스 데이터 감성 분석을 위한 Library
- Beautiful Soup 4: 뉴스 데이터를 웹에서 크롤링하기 위한 Library
- Google Trans: NLTK에 접한 데이터로 뉴스 데이터를 전처리하기 위한 번역 기능이 내재된 Library

- 어려웠던 점과 그 해결과정 등 기술

1. Transformer 구현의 어려움
    a. 문제점
        i. Transformer의 Input이 단일 차원 시계열(주가 데이터)이 아니라 다차원 시계열(주가 데이터, 뉴스 감성 분석 데이터, 기술 지표 데이터)이었다는 점에 구현의 어려움
    b. 해결 방안
        i. 설계: 데이터 Input을 행렬로 결정
            1. 딥러닝 모델은 각각의 데이터에 대한 가중치를 스스로 학습하므로, 주가 데이터 시계열과 함께, 뉴스 감성 점수와 기술 지표를 시계열로 전처리하여 행렬로 제공



ii. 구현

1. 감성 점수, 기술 지표를 각각 계산
2. 감성 점수, 기술 지표를 주가 시계열 데이터와 통합
3. Transformer Input의 차원을 수정

2. 10년치 뉴스 데이터 수집의 어려움

a. 문제점

i. 데이터 양이 방대함

ii. 크롤링 -> 번역 -> 감성 분석 파이프라인이 하루 당 10-20초 가량 소요되었음

b. 해결 방안

i. MSA를 도입과 서버(AWS EC2) 증진을 통한 병렬 처리

1. 크롤링/번역/감성 분석이 각각 5-10초가 소요되는 점을 감안하여 병렬 처리 가능한 파이프라인 구축함
2. AWS EC2에서 크롤링은 미허용이고 Google Trans, NLTK는 사용이 가능한 점을 고려하여 AWS EC2에서 번역/감성 분석 task를 수행

ii. 다중 프로세스 설정하여 병렬 처리

1. Network I/O가 대부분이라는 점에서 병렬로 실행하여 효율성을 증진
2. 종목별로 수집하도록 수정
3. 3개부터는 http request 403 (FORBIDDEN, 클라이언트가 콘텐츠에 접근할 권리를 가지고 있지 않다는 의미) 이 자주

발생되어 2개로 결정

4. 프로젝트의 역할 분담

4.1 개별 임무 분담

번호	학과	학번	학년	이름	담당업무
1	컴퓨터소프트웨어	2019080737	4	김나현	ML 모델링, 기술 지표 데이터 수집/전처리, ML 실험, Web Frontend
2	컴퓨터소프트웨어	2019011449	4	이수빈	ML 모델링, 뉴스 감성 데이터 수집/전처리, ML 실험, Web Backend

4.2 개발 일정

구분	추진내용	수행기간											
		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
계획	팀 빌딩, 교수님과 연구 방향성 미팅												
분석	유관 논문 서칭 (Sota 모델 및 주제의 현황)												
	주식 종목과 사용할 데이터 선정												

설계	Transformer 모델링												
	Ensemble(Transformer, LSTM, SVM) 모델링												
	텍스트 데이터 파이프라인 구성												
개발	ML 모델 구현												
	기술 지표 데이터 수집/전처리												
	뉴스 감성 점수 데이터 수집/전처리												
테스트	ML 실험 및 웹 페이지 개발												
논문작성	-												
종료	평가												

5. 결론 및 기대효과

- 프로젝트를 진행하면서 얻은 경험과 느낀 점

프로젝트를 1년에 걸쳐 진행하며 다수의 선행 연구 논문을 통해 주식 가격 예측 분야의 역사와 최신 동향을 탐구해보았다. 이 경험을 통해 학술적 지식을 쌓는 중요성을 깨닫게 되었다. 앞으로도 연구 논문을 통해 컴퓨터 과학 분야에서 지식을 쌓아 전문적으로 성장하고자 한다.

이 프로젝트에서 딥러닝 모델 개발과 데이터 수집 과정에서 발생한 문제들을 해결하는 과정을 경험하면서 많은 성장을 이루었다. 팀원과 함께 협업하고 같이 문제를 해결해나가는 과정을 통하여 협업 능력도 기를 수 있었다.

- 활용방안 및 기대효과 등 기술

이 연구에서는 Transformer, LSTM, SVM 모델을 결합한 앙상블 모델을 사용하여 한국 주식 시장의 주가 예측을 시도했다. 이는 이전에 어떤 논문에서도 시도되지 않은 새로운 접근 방식이다. 연구의 핵심은 뉴스 데이터와 기술적 지표를 통합하여 주가의 1일 단위의 변동을 예측하는 것이다.

그러나, 이 연구는 한국 시장에서의 주가 예측에 있어 AI 기반 모델의 사용이 어렵다는 결론으로 귀결되었다. 이는 한국 주식 시장의 독특한 특성과 뉴스 데이터의 복잡성, 기술 지표의 한계 때문에 정확한 예측이 어려웠기 때문이다. 연구는 AI 모델이 높은 변동성과 예측 불가능성이 높은 시장에서는 제한된 효과를 보일 수 있음을 시사한다. 이러한 발견은 향후 주가 예측 모델을 개발하는데 있어 중요한 시사점을 제공한다.

- 차후 계획 기술

팀원 각각의 계획은 다음과 같다.

1. 기술적 역량 강화: 이 프로젝트를 통해 얻은 경험을 바탕으로, 특히 딥러닝과 데이터 과학 분야에서의 전문 지식과 기술을 더욱 발전시키고자 한다. 이를 위해 관련 Coursera, K-Mooc와 같은 온라인 코스와 Workshop에 참여하고, 최신 연구와 기술 동향을 지속적으로 학습할 계획이다.
2. 프로그래밍 언어 다양화: 현재 사용하는 프로그래밍 언어 외에도 Scala, Go와 같은 다른 언어들을 학습하여, 다양한 프로그래밍 언어에 능숙해지는 것을 목표로 한다. 최근에는 Scala, Go 언어가 AI/ML 데이터의 수집/전처리를 위해 사용되는 경우 많으므로, 이는

새로운 프로젝트와 환경에서의 유연한 대응 능력을 향상시키는 데 도움이 될 것이다.

3. 개인 프로젝트 개발: 새로운 아이디어와 기술을 적용해 볼 수 있는 개인 프로젝트를 계획하고 있다. 이를 통해 실제 문제 해결에 필요한 기술과 도구를 활용하는 능력을 기르고자 한다.