
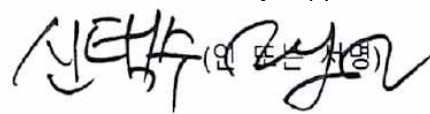




Capstone Design 최종보고서					
1. 신청과제					
수행기간	2021 학년도 1 학기		교과목명	[비대면]의사결정과정보기술(캡스톤디자인)	
과제명	딥러닝과 텍스트 마이닝을 이용한 ETF 매매				
팀명	엠프		신청예산	총	600000 원
지도교수(학과)	경영학과		지도교수(성명)	신택수	
2. 참여학생(최소 2인 이상)					
구분	역할	성명	전공	학년	학번
팀장	팀장, 총괄	하지민	경영학	4	2015231035
팀원1	팀원, 보조	이종혁	경영학	4	2018232012
팀원2					
팀원3					
팀원4					
팀원5					
3. 과제타입(택1)					
<input checked="" type="checkbox"/> 일반형	<input type="checkbox"/> 융합형	<input type="checkbox"/> L2M 인재양성형	<input type="checkbox"/> 기업연계형	<input type="checkbox"/> 지역사회기여형	
4. 결과물 종류(택1)					
유형	<input type="checkbox"/> 시제품 및 결과모형		<input type="checkbox"/> 하드웨어	<input type="checkbox"/> 기타()	
무형	<input type="checkbox"/> 인쇄물 및 영상		<input type="checkbox"/> 설계도면	<input checked="" type="checkbox"/> 보고서(조사, 분석결과 등)	
	<input type="checkbox"/> 소프트웨어(어플리케이션, 웹페이지 등)		<input type="checkbox"/> 기타 ()		
5. 참여업체(기업연계형만)					
기업(관)명		사업자등록번호		담당자명	
<p>위와 같이 Capstone Design 과제 최종보고서를 제출합니다.</p> <p>2021 년 6 월 30 일</p> <p>(대표학생) 하지민 (인 )</p> <p>(지도교수) 신택수 (인 )</p>					
	연세대학교 원주LINC+사업단장 귀하				 <small>사회맞춤형 산학협력 선도대학 육성사업 Leaders in Industry-university Cooperation</small>

Capstone Design 과제요약서

1. 예산집행결과 (홈페이지 확인 후 작성)

지원금 집행내역 (단위 : 원)				
지출항목	신청 예산	집행 액	잔액	비고
물품 · 제작비	300000	0	300000	
회의비	120000	34000	86000	
교통비	180000	0	180000	
멘토비	0	0	0	
합 계	600000원	34000원	566000원	

2. 수행 과제

2.1 과제(작품)수행 개요

최근 ‘주식’, ‘코인’ 등은 대화 주제로 자주 등장하고 많은 사람들 사이에서 가장 뜨거운 주제가 아닐 수 없다. 지난 해 코로나19로 인해 전세계 금융시장이 흔들리고 증시 폭락이 예상되는 가운데 한국 시장에서 외국인 투자자들의 대규모 매도에 맞서 개인투자자들이 대규모 매수한 ‘동학개미운동’이라 불리는 신조어가 등장하기도 했다. 한국거래소 정보데이터시스템에서 KOSPI 주가지수 월별 추이를 보면 2020년 10월 이후 거래량이 급증한 것을 알 수 있다. 10월 거래량 15,705,577에서 11월에는 24,377,975로 급증했고 5월 거래량 감소 이전까지 비슷한 수준으로 많은 거래량을 유지했다.



〈그림 1〉 KOSPI 월별 추이

주식투자에 대한 관심이 높아지고 너도나도 투자를 하는 이유에는 여러 가지 이유가 있겠지만 가장 큰 이유로는 부의 상승에 대한 낮은 기대가 가장 큰 이유로 꼽히고 있다. 미국 중앙은행 기준 금리는 20년 3월 0.25%로 낮춘 이후로 지난 4월까지 동결 상태이다. 한국 중앙은행 기준금리도 20년 5월 이후 0.50%로 동결되어 있다. 저금리로 인해 은행 예금으로는 자산을 키울 수 없다고 판단한 개인투자자들이 주식과 코인 시장에 뛰어들었고 재테크로 활용하기 위해 많은 관심을 쏟고 있다. 이런 상황 속에서 ETF가 하나의 투자 수단으로 관심을 받고 있다.

ETF(상장지수펀드)는 특정 주가 지수의 성과를 추종하는 인덱스펀드로 거래소에서 일반 주식처럼 매매가 가능하다. 그리고 시장대표, 업종섹터, 스타일지수, 해외지수 등으로 ETF를 구분한다. 1주만 매수해도 분산투자 효과를 가지고 있으며 높은 접근성과 투명성, 그리고 낮은 투자비용이라는 특성을 가지고 있어 인기가 높은 투자수단으로 자리 잡고 있다.

23일 자산운용업계에 따르면 올 초부터 이날까지 총 15개 ETF가 출시됐다. 지난해 같은 기간 출시된 ETF는 1개에 그친 것과 비교하면 올 들어 경쟁적으로 ETF 출시가 이어지고 있는 것으로 분석된다. 이에 따라 현재 ETF 전체 수는 476개로 연내 500개를 넘어설 것으로 예상된다. 지난해 출시한 ETF 47개 중 40개는 하반기에 출시됐다. 또한, 증시에 관심이 있는 투자자라면 직관적으로 투자할 수 있는 상품들이 쏟아지면서 올해 출시한 15개 ETF에는 17일 현재까지 총 4169억2500만원(순자산 총액)이 몰려들었다. 현재 ETF 전체 순자산 규모는 57조8590억원으로 연초 52조3145억원 대비 10.59% 정도 늘었다. 주간 수익률은 -2.29%(TIGER미국필라델피아반도체나스닥)부터 4.09%(HANARO Fn전기&수소차) 정도로 집계됐다.¹⁾

여러 종류의 ETF 중에서 올 들어선 인덱스형 ETF보다 테마형 ETF에 대한 투자가 급격히 늘었다. 지난 1월부터 이달 8일까지 순매수 상위 5개 ETF를 보면 TIGER 차이나전기차SOLACTIVE, KODEX 200선물인버스2X, TIGER KRX 2차전지 K-뉴딜, KODEX 2차전지산업, KODEX 미국 FANG플러스 등으로 나타났다.²⁾

종가 데이터나 KOSPI 지수와 연계되는 인덱스형 ETF와는 달리 테마형 ETF는 뉴스 데이터, 원자재 데이터, 구성 종목 등의 다양한 컨텍스트 데이터와 쉽게 연계 된다. 우리는 이러한 비정형의 데이터들 까지 활용하여 테마형 ETF에 대한 추정을 시도하고자 할 것이다. 기존에는 이러한 테마형 ETF에 대한 예측 모델 연구 사례가 없다.

기존에 종가 예측을 위해서는 기본적인 회기 분석 같은 통계 기법이나 기계 학습이 사용될 수 있었다. 최근에는 딥러닝 기법을 이용하여 이러한 예측을 시도하는 경우가 많다. 특히, 테마형 ETF의 경우에는 비정형 데이터들이 추가적으로 입력으로 주어지는 이러한 경우에는 기존의 통계 기법을 이용하기 어려우므로, 최신의 딥러닝 기술들을 활용하여 분석하고자 한다. 그 중 우리는 딥러닝 모델 중에서 시퀀스 데이터 특성에 적합한 LSTM 모델을 사용하여 연구를 진행했다.

1) 황준호, "ETF 전성시대.. 연내 500개 돌파", <아시아경제>, 2021.04.23

2) 한국거래소, 개인투자자 국내 ETF 순매수 상위, 2021년(1월1일 ~ 4월8일)

3. 과제(작품)의 개발과정 및 역할

3.1 과제(작품)의 개발과정

일정	내용
1~3주차	주제 선정 / 문제 정의
4주차	관련 내용 학습 / 기획 - 관련 논문 자료 조사
5주차	관련 내용 학습 / 기획 - 분석 방법론 정의(예측 모델 개발)
6주차	관련 내용 학습 / 기획 - 분석 방법론 정의(예측 모델 개발)
7주차	관련 내용 학습 / 기획 - 추가 데이터 확보 및 시각화
8주차	중간 프로젝트 보고서 제출 (중간시험기간)
9주차	중간 프로젝트 보고서 발표
10주차	데이터 분석 수행 - sns 언급량 및 나스닥 증감량 관련 데이터 수집 및 전처리
11주차	데이터 분석 수행 - 기술통계 분석/ 가설 확인
12주차	데이터 분석 수행 - 예측 모형 개발
13주차	데이터 분석 수행 - 모형 훈련 및 검증을 통한 정확도 적합 문제 해결
14주차	데이터 분석 수행 - 고도화 및 보고서 제출을 위한 데이터 정리
15주차	최종 프로젝트 보고서 발표 및 보고서 제출
16주차	수정 최종 보고서 제출 (기말고사 기간)

기술적 지표와 추가 보조 지표는 파이썬 라이브러리인 pykrx를 활용해서 가져온다. 뉴스 데이터 및 재무제표 데이터는 각각 네이버 뉴스, 네이버 금융을 통해 크롤링한다. 그리고 원자재 가격 데이터는 한국자원정보서비스(KOMIS)에서 엑셀 파일로 다운로드하여 사용한다.

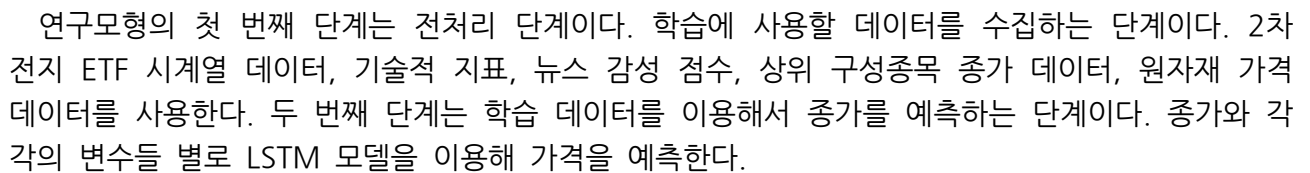
데이터 분석 모델은 딥러닝 LSTM 모델을 사용하고자 한다. LSTM은 RNN의 단점을 극복한 것으로 장기 기억과 단기 기억을 모두 유지할 수 있다. LSTM은 RNN의 일종으로 ‘게이트’ 구조를 도입해 과거의 정보를 기억할지 말지 판단하면서 필요한 정보만을 다음 시각에서 이어받을 수 있다. 은닉층 대신에 LSTM 블록이하는 회로 구조를 가지는데 기억 셀, 입력 게이트, 출력 게이트, 망각 게이트로 구성되어 있다.

2차전지 종가 데이터를 종속변수로 활용하고 각각의 독립변수들을 학습시켜 예측률을 비교하는 형태로 연구를 진행하였다.

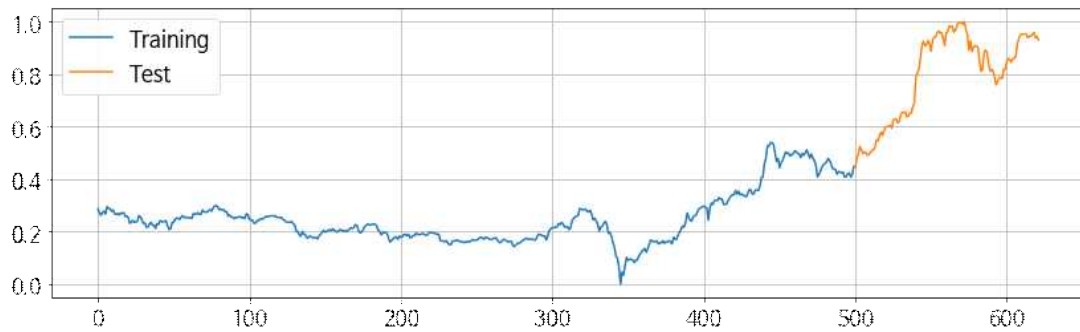
3.2 과제(작품)의 팀원간 역할

구분	성명	전공	역할	참여도(%)
대표학생	하지민	경제학과	프로젝트 개요 및 목적 재정립, 최종 보고서 실증분석 결과, 결론 (실증 결과 요약 정리), 나머지 파트 수정, colab 코드 작성	60%
팀원	이종혁	경영학과	문헌 고찰 정리, 분석기법 개요, 연구모형 작성, 참고문헌 정리, 발표자료 작성	40%
팀원				
팀원				

4.1 과제(작품)의 개발 결과



데이터의 시차는 5일로 하였고, Test와 Train 데이터는 75:25 비율로 나누어 진행하였다. 그리고 최적화 함수는 adam을 사용하였고, loss는 mean_absolute_error를 사용하였다. fit 과정에서 epoch는 100회, batch_size는 32로 설정하였다.



〈그림 8〉 Test, Train 데이터 비율

〈표 5〉 모델 별 점수 2

모델	점수(mae)
감성 점수	0.064980
종가만 사용	0.093815
기술적 지표	0.123307
원자재 가격	0.124834
NOHLCV	0.134652
상위 3 종목 증가	0.182619
전부 사용	0.203677

분석 결과 감성점수를 사용하여 예측하였을 때 가장 높은 점수가 나왔고, 상위 3개 종목 종가를 사용하여 예측하였을 때 가장 낮은 점수가 나왔다. 감성 점수 제외한 나머지 모두 종가만을 사용해 예측했을 때 보다 낮은 점수가 나왔다.

4.3 과제(작품)의 활용방안 및 기대효과

본 연구는 기존에 테마형 ETF에 대한 예측 모델 연구 사례가 없다는 점에 착안해 테마형 ETF를 분석하는데 사용되는 여러 변수들을 이용해 각 변수별 예측률을 비교하였다. 2018년 9월부터 2021년 4월 30일까지 약 2년간의 주가 데이터를 이용하여 실증분석을 진행하였다.

분석대상으로는 국내 2차 전지 테마 관련 ETF중 TIGER 2차전지테마, KODEX 2차전지산업을 사용했고 변수로는 기술적 지표 뉴스 데이터, 원자재 가격, 뉴스 감성 점수, nohlc, 상위 3개 구성 종목 증가 데이터를 이용하여 종속변수를 예측하고 예측력을 비교하였다. 연구 결과 감성 점수가 가장 예측력이 좋았고 차례로 종가만 사용, 기술적 지표, 원자재 가격, nohlc, 상위 3개 종목 증가 순이다.

테마형 ETF는 선행 연구에서의 인덱스형 ETF에 비해 기술적 지표나 거시적 지표의 주가 예측력이 상대적으로 약하게 나타났다. 그렇지만 프로젝트 목적과 개요에서 예상했듯이 테마형 ETF가 훨씬 뉴스 소식 같은 비정형 데이터에 민감하게 반응한다는 것을 알 수 있었다. 실제 점수로 보았을 때 도 정형 데이터 보다 두 배 이상 낮은 loss를 보이고 있는 것을 확인할 수 있었다.

주식의 기술적 분석은 많은 연구를 통해 그 효과성이 검증되었다. 하지만 시장은 수요와 공급에 의한 영향과 더불어 투자자의 심리적인 요인, 회사의 변화, 관련 정치적 사회적 이슈 등에도 크게 영향을 받는다. 기술적 분석은 이러한 부분을 담아내기 못한다는 한계점이 있다.

감성분석, 뉴스 기사를 통한 주가 예측 연구도 상당한 예측력을 보여주고 있다. 하지만 모든 뉴스를 분석하는 것은 불가능하고 어떤 뉴스가 관련이 있는지, 주식에 어떻게 영향을 주는지 정확하게 파악하기 어려울 뿐더러 중요한 정보만 있는 것이 아니라 잡음이 존재하기 때문에 분석의 한계가 있다.

이번 연구를 통해 ETF의 가격을 예측하는데 기술적 분석, 뉴스 기사를 통한 감성 분석, 괴리율 등이 유효성 있고 상당한 예측력을 보여준다는 것을 알 수 있었다. 연구결과를 바탕으로 주식 매매전략으로 활용하거나 기술적 분석과 감성 분석 등 각 변수들을 적절히 사용하면 예측의 정확도를 향상시킬 수 있을 것으로 기대된다.

주식에 영향을 주는 뉴스는 수없이 많다. 또 어떤 종목인지에 따라 경기 불황이 호재가 될수도 있고 악재가 될수도 있다. 따라서 개별 종목에 대한 감성사전을 활용하면 예측력을 높일 수 있으나 ETF는 여러 지수를 종합하여 만든 상품이기 때문에 이것을 모두 반영하여 예측하는 것은 매우 어렵다. 이 연구에서는 ETF에 특화된 감성사전을 적용하지 않고 일반적인 감성사전을 사용했다는 점에서 한계가 있다. 또한 예측결과는 나왔으나 다양한 데이터 중 어떤 데이터가 큰 영향을 주는지에 대해서는 알아내기 힘들었다. 데이터 간의 상관 관계와 인과관계를 확실하게 찾아내지 못하였다. 또한 시장에 영향을 주는 많은 요소들 중 극히 일부만을 다루었으며 개별 종목별로 예측력이 좋은 모델과 보조 지표가 존재하나 이를 모두 반영할 수 없다는 점이 한계점으로 남는다.

주식 시장에 모든 변수를 반영하여 분석한다는 것은 매우 어려운 일이 될 것이다. 하지만 좋은 예측력을 가진 분석방법들을 혼합하여 분석하거나 특정 변수에 가중치를 두는 등의 방법으로 예측력을 높일 수 있다. 감성분석의 경우도 일반 감성사전을 사용하였지만 상당한 예측력을 보여준 것을 보면 해당 종목, 테마에 적합한 감성사전을 도출하여 이를 통해 분석한다면 개선될 것으로 보인다.

해당 연구는 테마형 ETF를 분석하는 여러 지표들을 LSTM을 통해 분석하여 각 변수별 예측력을 비교하여 감성 점수가 상당히 중요한 변수로 작용한다는 것을 검증했다는 측면에 성과가 있다. 하지만 각 변수의 예측력만을 보여줄 뿐 BLSTM, LightBGM 등 더 발전된 예측 모델들을 활용하지 못한 측면에서 기존 분석과 크게 다르지 않다는 비판을 받을 수 있다.

다만 테마형 ETF에 대한 기존 연구가 없었고 변수별 예측력을 비교함으로써 주요 변수로 작용할 수 있는 변수들을 분석했다는 점에서 의의가 있다. 주식만큼이나 투자 수단으로써 관심을 받고 있는 ETF에 사용할 수 있는 예측 모델의 지속적인 개발과 더 많은 변수들을 반영할 수 있는 지속적인 개선이 필요하다.