



#### **Contents**

### Project 1

- 1. 뉴스데이터를 활용한 ETF 구성 개요
  - ETF 테마 설명
  - 데이터 출처 및 전처리
- 2. 모델 소개 및 결과
  - Simple Version
  - Develop Version
  - Develop Version + Sentiment Analysis
  - 섹터 결과
- 3. ETF 구성
  - 최종 산업 분류
  - ETF 구성 방법
- 4. 한계 시사점 및 추가 과제



#### 1-1. 뉴스데이터를 활용한 ETF 구성 개요

- '22년 5월 10일 윤석열 정부 공식 출범
  - 대통령 취임에 따라 윤석열 정부의 정책 방향에 많은 관심 집중
  - 대통령 취임을 전후한 5월 중 다량의 기사를 통해 양질의 비정형 데이터를 확보하기에 적합
- 공약과 정책방향이 담긴 뉴스 데이터 키워드 분석을 통해 특정 산업 및 섹터에 대한 新정부의 관심도 파악
  - 윤석열 정부의 산업(섹터) 관심도가 곧 대한민국의 미래 산업 방향을 반영
  - 이를 통해 대한민국의 미래 유망산업을 추구하는 ETF 구성
- ※ 단, 단순 테마주(동문, 가족, 친인척 관련주)가 아닌 新정부의 관심도를 반영한 주식만을 대상으로 함 (제외 예시: 대영포장) 원석열 관련주로 편있된 이유

-권택환 대표이사가 윤석열 검찰총장과 서울대학교 동문

-육복희 사외이사가 윤석열 검찰총장과 서울대학교 법대 동문

-신호철 사외이사가 윤석열 검찰총장과 서울대학교 법대 동문

# KAIST 개요

데이터 수집 기간 : 2022/05/01(일) ~ 2022/05/ 31(화)

데이터 출처 : 네이버 정치, 경제 뉴스

총 수집된 데이터 개수 : 총 271,102 개= 145,627(정치) + 125,475개 (경제)



#### 1-1. 산업(테마) 단어 크<del>롤</del>링 및 전처리

- 데이터 출처 : 네이버 증시 테마
- 사용된 산업(테마) 단어: 257개 단어

```
수정 전 단어 : 2차전지(LFP/리튬인산철) --> 수정 후 단어 : 2차전지
수정 전 단어 : 2차전지(생산) --> 수정 후 단어 : 2차전지
수정 전 단어 : 2차전지(소재/부품) --> 수정 후 단어 : 2차전지
수정 전 단어 : 2차전지(장비) --> 수정 후 단어 : 2차전지
수정 전 단어 : 2차전지(전교체) --> 수정 후 단어 : 2차전지
수정 전 단어 : 3D 낸드(NAND) --> 수정 후 단어 : 3D 낸드
수정 전 단어 : 5G(5세대 이동통신) --> 수정 후 단어 : 5G
수정 전 단어 : GTX(수도권 광역급행철도) --> 수정 후 단어 : GTX
수정 전 단어 : K-뉴딜지수(2차전지) --> 수정 후 단어 : K-뉴딜지수
수정 전 단어 : K-뉴딜지수(게임) --> 수정 후 단어 : K-뉴딜지수
수정 전 단어 : K-뉴딜지수(비이오) --> 수정 후 단어 : K-뉴딜지수
```

#### |테마별 시세

테마명▲	전일대비 최근3일		전일대비 등락현황			주도주	
네마당=	선글네미족	등락률(평균)-	상승	보합	하락		<u> </u>
스마트카(SMART CAR)	+3.69%	-0.80%	11	1	2	↑ 모트렉스	▲ 인포뱅크
마리화나(대마)	+3,40%	+1.77%	7	4	3	↑ 마이더스A…	▲ 아이큐어
2022 상반기 신규상장	+2.59%	-0.07%	17	8	14	↑ 청담글로벌	▲ 공구우먼
N(네트워크 <del>통</del> 합)	+1.78%	-1.01%	2	1	6	▲ 오파스넷	▲ 콤텍시스템
자율주행차	+1.76%	-0.38%	32	7	15	↑ 모트렉스	▲ 라닉스
보톡스(보툴리 <u>눔톡</u> 신)	+1,60%	+0.40%	2	1	9	↑ 청담글로벌	▲ 휴온스글로
SNS(소셜네트워크서…	+1.46%	-0.66%	6	0	4	▲ 인포뱅크	▲ 이루온
건설 대표주	+1.41%	-0.67%	5	0	2	▲ DL	▲ 대우건설
비료	+1.38%	-1.26%	10	0	2	▲ 효성오앤비	▲ 체시스
밥솥	+1.30%	+0.05%	4	0	1	▲ 윌렁스	▲ PN풍년
소매유통	+1.13%	+0.08%	3	6	14	<b>↑</b> 청담글로벌	▲ 아이메스이
화학섬유	+1.11%	-0.27%	2	1	5	▲ 티케이케미	▲ 태광산업
RAD(NFC 등)	+0.91%	-1.31%	6	0	11	↑ 소니드	▲ 에이스테크
공작기계	+0.84%	-0.50%	7	4	6	▲ 스맥	▲ SNT <del>중공</del>
리비안(RIVAN) 관련주	+0.77%	-0.74%	13	0	9	▲ 씨아이에스	▲ 디와이



#### 1-2. 모델 소개 및 결과

- Simple Version
  - 단어 "윤석열" And "산업 단어"가 포함된 기사를 1차적으로 걸러내어서 뉴스 기사 본문 단위로 "윤석열"과 산업 단어가 매칭되는 것을 확인 후 Count하는 TF 방식
- Develop Version
  - Simple version의 한계점이 존재해 본문이 아닌 기사 문장 단위로 매칭
- Develop Version + Sentiment Analysis
  - 매칭된 문장을 Sentiment 분석을 통해 문장의 긍정, 부정 분석

## KAIST Model 방법론

#### Simple Version

- Step1: 경제 및 정치 기사가 포함된 OpenSearch에 저장된 데이터의 본문("body")에 단어 " 윤석열 " 및 "산업 단어" 가 포함된 기사들만 일별로 1차적으로 수집.
- Step2: 1차적으로 수집된 기사에서 "산업 단어"가 포함된 기사를 산업 단어 별로 Counting.
- 한계점: 제 8 전국동시선거 및 다양한 행사 인해 여러 정치인들의 이름이 언급되고 산업 단어가 언급될 때, 윤석열 대통령과 매칭 시키는 것에 대한 의문점이 존재한다. 또한 긍정, 부정에 대한 분석이 불가능

- [6·1 지방선거 후보 정책팸플릿] 원주시장 5월 26일
  - △구자열="원주 주력이자 미래 경쟁력이 우수한 디지털 <mark>헬스케어와 의료기기 산업</mark>을 더욱 육성해야 한다. 특히 '국제인증지원센터'를 신설해 이들 산업의 해외 수출 판로 개척을 적극 지원해야 한다. 미국 공공조달의 경우 12조원대 의료기기 및 14조원대 보건 물품시장이 포진돼 있다. 성공 진입하면 의료기기 2조원대 수출 시대를 열며 원주 자립 경제의 든든한 틀이 마련될 것으로 기대된다. 여기에 정부 '성공불융좌'제도를 지자체 차원에서 추가 실시하는 등 다양한 지원책으로 지속가능한 경제 기반인 백년기업 탄생 토대를 만들겠다."
- △원강수="원주 경제의 지속가능한 발전을 확보하기 위해서는 반드시 대기업을 유치해야 한다. <mark>윤석열</mark> 정부가 4차산업 혁명시대에 발맞춰 국정과제로 반도체 클러스트 단지를 육성한다는 계획을 밝혔다.

출처: 강원도민일보(http://www.kado.net)

# KAIST Model 방법론

#### Develop Version

- Step1: 경제 및 정치 기사가 포함된 OpenSearch에 저장된 데이터의 본문("body")에 단어 " 윤석열 " 및 "산업 단어" 가 포함된 기사들만 일별로 1차적으로 수집.
- Step2: "윤석열" 및 산업 단어를 포함된 문장을 구분하여 Count하는 TF방식
- 한계점: 긍정, 부정에 대한 분석이 불가능
- 예시) 2022년 5월 25일 경기신문 경기신문 (https://www.kgnews.co.kr)
- 기사 제목 :
- 김동연 "윤석열 정부, GTX 연장 기준안 5월내 공개해야"
- <u>본 문 :</u>
- 김 후보는 25일 오전 경기도의회 브리핑룸에서 기자회견을 열고 "최근 저와 민주당, 당내 시장·군수 후보들이 모여 <mark>윤석열</mark> 정부의 GTX 공약 파기를 규탄하고 항의하며 동시에 GTX 연장과 신설을 반드시 해내겠다고 약속드렸다"며 이같이 밝혔다.

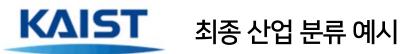
### KAIST Model 방법론

#### Develop Version + Sentiment Analysis

- Step1: Develop Version Proceed
- Step2: 문장("윤석열" + "산업 단어")에 대해서 Sentiment Analysis 실행 후, Negative sentiment 제거.
- 수업 때 배웠던 Hugging Face 사용 pipeline('sentiment-analysis', model ='snunlp/KR-FinBert-SC'
  - Positive 분류

'이어 "<mark>윤석열</mark> 당선인이 후보자 시절에 약속한 공약은 최우선 국정과제로 추진한다는 원칙을 갖고 있다"며"새 정부가 출범하면 경남지역 공약과제인 <mark>우주산업</mark> 클러스터, 항공우주청 설립 등에 대한 추진계획을 구 체적으로 검토하여 추진해 나갈 계획"이라고 덧붙였다.'

- Negative 분류
- '(수원=연합뉴스) 최찬흥 기자 = 더불어민주당은 6일 GTX-A 노선 건설 현장을 함께 방문한 <mark>윤석열</mark> 대통령 당선인과 국민의힘 김은혜 경기지사 후보를 공직선거법 위반 혐의로 경기도선거관리위원회에 고 발했다.'



no	theme	modified theme	count
335	재난/안전(지진 등)	안전	4847
324	일자리(취업)	일자리	1687
108	백신/진단시약/방역(신종플루, AI 등)	백신	1283
250	철도	철도	1239
317	SNS(소셜네트워크서비스)	SNS	1099
048	영화	영화	924
237	GTX(수도권 광역급행철도)	GTX	880
227	전기차	전기차	573
514	전기차(충전소/충전기)	전기차	573
232	미디어(방송/신문)	미디어	335
206	농업	농업	279
152	온실가스(탄소배출권)	온실가스	204
307	시스템반도체	시스템반도체	185
205	원자력발전	원자력발전	183
331	가상화폐(비트코인 등)	가상화폐	132
377	치매	치매	125
387	블록체인	블록체인	112
105	사료	사료	108
480	메타버스(Metaverse)	메타버스	93
204	비료	비료	88
060	육계	육계	88





no	theme	modifie -	catego	count
307	시스템반도체	시스템반의	2	185
242	탄소나노튜브(CNT)	탄소나노튀	2	7
370	3D 낸드(NAND)	3D 낸드	2	2

no	▼ theme	modifie -	catego(	count -
108	백신/진단시약/방역(신종플루, AI 등	) 백신	4	1283
377	치매	치매	4	125
175	의료기기	의료기기	4	55
241	바이오시밀러(복제 바이오의약품)	바이오시달	4	35
487	건강기능식품	건강기능식	4	16
493	mRNA(메신저 리보핵산)	mRNA	4	11
311	고령화 사회(노인복지)	고령화 사	4	6
389	면역항암제	면역항암제	4	4
470	마이크로바이옴	마이크로터	4	3
066	줄기세포	줄기세포	4	3
174	치아 치료(임플란트 등)	치아 치료	4	3
172	제약업체	제약업체	4	1



no	theme	modified theme	category	count
335	재난/안전(지진 등)	안전	1	4847
108	백신/진단시약/방역(신종플루, AI 등)	백신	4	1283
227	전기차	전기차	11	573
307	시스템반도체	시스템반도체	2	185
205	원자력발전	원자력발전	13	183
377	치매	치매	4	125
175	의료기기	의료기기	4	55
241	바이오시밀러(복제 바이오의약품)	바이오시밀러	4	35
487	건강기능식품	건강기능식품	4	16

20	- theme	- modifie-	catego 🗷	count
27	전기차	전기차	11	573
514	전기차(충전소/충전기)	전기차	11	573
362	자율주행차	자율주행치	11	86
386	수소차(연료전지/부품/충전소 등)	수소차	11	41
167	타이어	타이어	11	25
332	스마트카(SMART CAR)	스마트카	11	3

no	▼ theme	▼ modit	fie 🔻 ca	teaoi₹	count
205	원자력발전	원자력	격발진	13	183
079	자원개발	사원/	H발	13	48
184	수자원(양적/질적 개선)	수자원	김	13	21
187	핵융합에너지	핵융합	할에너	13	3
191	태양광에너지	태양평	랑에년	13	2
323	셰일가스(Shale Gas)	셰일기	<b>가스</b>	13	1
188	풍력에너지	풍력0	베너지	13	1



#### 최종 산업 분류

재난/안전(안보)	<ul><li> 경제 안보</li><li> 재난, 건축</li><li> 통신, 보안</li></ul>	개별주식 :22개	
바이오	<ul><li> 백신/진단시약/방역</li><li>&gt; 의료기기</li><li>&gt; 건강기능식품</li></ul>	개별주식 :189개	_
전기차	<ul><li>전기차</li><li>수소전기차</li><li>2차 전지</li></ul>	개별주식 :84개	
시스템반도체	<ul><li>&gt; 시스템반도체</li><li>&gt; 자율주행</li><li>&gt; IT 장비업체</li></ul>	개별주식 :38개	
원자력	<ul><li>원자력</li><li>에너지</li><li>전력설비정비</li></ul>	개별주식 :43개 21	



# 1-3. ETF 구성

Industry Weight = count proportional weight + Time Series Rank

Log\_Weight = log( count / count\_sum)

	no	modified theme	category	count	weight	log_count	log_weight	num_stock
0	335	재난/안전(안보)	1	4847	0.663791	8.486115	0.260395	23
1	108	바이오	4	1514	0.207340	7.322510	0.224690	39
2	227	전기차	11	573	0.078472	6.350886	0.194876	190
3	307	시스템반도체	2	185	0.025336	5.220356	0.160186	85
4	205	원자력	13	183	0.025062	5.209486	0.159852	44

Weights within industry = equal weight

Final stock weight = Industry weight X 1/total number of stock

	재난/안전(안보)	바이오	전기차	시스템반도체	원자력
0	삼일씨엔에스	HLB제약	우수AMS	하나마이크론	휴센텍
1	삼영엠텍	대응제약	에디슨EV	에이디테크놀로지	두산에너빌리티
	광림	보령	우신시스템	아나패스	삼영엠텍
	KT서브마린	SK케미칼	LS ELECTRIC	SK하이닉스	DL이앤씨
	한국종합기술	일동제약	만도	가온칩스	한전산업
185	None	광동제약	None	None	None
186	None	엠아이텍	None	None	None
187	None	파멥신	None	None	None
188	None	노바렉스	None	None	None
189	None	네이처셀	None	None	None
190 ro	ws × 5 columns				

섹터	재난/안전(안보)	바이오	전기차	시스템반도체	원자력
Weight(%)	20.82	21.47	20.16	18.86	18.69



#### 1-4. 한계 시사점 및 추가 과제

- 단순 TF 방법론을 사용을 해서 문맥 파악에 한계, TFIDF & BERT 모델의 접목?
- <mark>윤석열</mark> 대통령 부부가 나란히 걷는 사진이 SNS에서 화제가 됐고, 특히 이날 일정에는 10명 남짓한 최소한의 경호인력 만 동행해 눈길을 끌었습니다.(positive)
- Sentiment analysis의 모델 Error
- <mark>윤석열</mark> 당선인은 2일 1기 신도시인 안양 평촌에 있는 한 노후 아파트를 방문해 재건축활성화와 관련한 공약사항을 이행할 것이라고 재차 강조했다. 윤 당선인이 이날 방문한 단지는 안양시 동안구 평촌 동 초원7단지 부영아파트(초원부영)로 작년 리모델링컨설팅 시범사업단지로 선정된 곳이다. 다만 입주자대표회의 차원에서는 리모델링보다 <mark>재건축사업</mark> 추진을 위한 지원을 건의했고 윤 당선인도 이에 화 답했다.
- Keyword 별 과거 Trend 반영 못함
- 대표성 있는 산업별 Keyword 선정의 어려움 Ex: 건설, 안전



#### [뉴스투데이 카드뉴스] 윤석열 정부 출범 후 'BASE(배터리· 원자력 발전·반도체·전기차)'가 뜬다

김현주 입력: 2022.06.09 09:10 | 수정: 2022.06.09 09:23

삼성SDI, 미국에 첨단 배터리 공장 설립해 본격 공략 두산에너빌리티, 한·미 원전협력 최고 수혜 기업...600조원 시장 정조준 삼성전자, 첨단 3nm 반도체 초격차 기술력 공개 현대차, 전기차 및 로보틱스·UAM 등에 105억달러 투자



[뉴스투데이=김현주 기자 / 그래픽: 강선우] '윤석열 정부에서 급부상한 BASE(배터리·원자력 발전·반도체·전기차)산업에 주목하라'

조 바이든 미국 대통령은 지난 달 20일부터 22일까지 한국을 방문해 윤석열 대통령 및 재계 총수들과 만난 후 배터리('B'at tery), <mark>원자력 발전('A'tomic plant)</mark>, <mark>반도체('S'emiconductor)</mark>, <mark>전기차('E'lectric car)</mark> 분야에서 전방위적인 협력을 요청했다.

재계는 세계 최대 미국 시장을 공략하면서 미래 먹거리 사업으로 급부상한 BASE에 주목하고 있다.



#### ETF 포트폴리오 개별 주식(Appendix)

	재난/안전(안보)	바이오	전기차	시스템반도체	원자력
0	삼일씨엔에스	HLB제약	우수AMS	하나마이크론	휴센텍
1	삼영엠텍	대웅제약	에디슨EV	에이디테크놀로지	두산에너빌리티
2	광림	보령	우신시스템	아나패스	삼영엠텍
3	KT서브마린	SK케미칼	LS ELECTRIC	SK하이닉스	DL이앤씨
4	한국종합기술	일동제약	만도	가온칩스	한전산업
5	파라텍	뷰노	코다코	윈팩	한국전력
6	서전기전	레이언스	스카이이앤엠	앤씨앤	한전KPS
7	오텍	휴마시스	아모그린텍	다믈멀티미디어	L
8	HRS	유바이오로직스	한국앤컴퍼니	디아이	비엠티
9	동양파일	퓨쳐켐	한온시스템	아이앤씨	태광
10	유엔젤	셀리버리	삼화전기	지니틱스	DL
11	삼영이엔씨	LG생활건강	캠시스	DB하이텍	대우건설
12	코리아에스이	신풍제약	디아이씨	리노공업	현대건설
13	KH 건설	한국콜마홀딩스	SKC	아이텍	한전기술
14	인콘	레고켐바이오	에스트래픽	아진엑스텍	LS ELECTRIC
15	한창	네오펙트	인팩	네패스	대한전선
16	동아지질	차백신연구소	우리산업	피에스케이	서전기전
17	이엔플러스	피제이전자	피에스텍	삼성전자	이엠코리아
18	희림	퀀타매트릭스	이지트로닉스	엘비세미콘	광명전기
19	대창스틸	인트론바이오	덕양산업	어보브반도체	HRS
20	한컴라이프케어	엔케이맥스	현대모비스	에이티세미콘	성광벤드
21	리노스	일진홀딩스	씨아이에스	두산테스나	에스앤더블류
22	현대에버다임	프로스테믹스	엠에스오토텍	에이디칩스	디티앤씨
23		SK바이오사이언스	엠플러스	엘비루셈	우진
24		KT&G	코스모신소재	한미반도체	금화피에스시
25		세운메디칼	지엠비코리아	SFA반도체	강원에너지
26		프롬바이오	동아화성	아이에이	일진파워
27		제노레이	삼성SDI	오디텍	GS건설
28		체시스	상아프론테크	라닉스	우리기술
29		국전약품	한일단조	고영	디케이락
30		엘앤케이바이오	엘앤에프	텔레칩스	한양이엔지
31		프레스티지바이오파마	신흥에스이씨	시그네틱스	하이록코리아
32		서흥	솔루스첨단소재	알파홀딩스	보성파워텍
33		대한뉴팜	삼기	네패스아크	에너토크
34		씨젠	한국단자	티엘아이	대창스틸
35	1	이수앱지스	휴맥스	코아시아	태웅
36		알테오젠	성창오토텍	LX세미콘	한신기계
37		녹십자홀딩스	센트랄모텍	테크윙	비에이치아이
38		초록뱀헬스케어	보성파워텍	에스앤에스텍	티플랙스
39		셀트리온	화신		효성중공업
40		뉴트리	KEC		조광ILI
41		인바디	알파홀딩스		오르비텍
42		에이프로젠제약	엘아이에스		SG&G
43	1	진양제약	폴라리스세원		에이프로젠 MED



#### 산업 단어 별 Time Series(Appendix)

▼	우주항공산업 🔻	셰일가스 🔻	그린에너지 🔻	NFT 🔻	바이오헬스 🔻	가상화폐 🔻	게임 🔻	블록체인 🔻	K-뉴딜지수 ▼	캐릭터상품 🔻	메타버스 🔻
2022-05-01 00:00:00	5	0	0	0	2	0	12	0	0	0	1
2022-05-02 00:00:00	7	0	0	3	2	2	31	9	0	0	3
2022-05-03 00:00:00	4	0	0	28	7	6	33	18	0	0	12
2022-05-04 00:00:00	0	0	0	13	4	1	19	4	0	0	2
2022-05-05 00:00:00	0	0	0	5	0	0	28	5	0	0	1
2022-05-06 00:00:00	0	0	0	1	0	2	25	4	0	0	0
2022-05-07 00:00:00	0	0	0	1	1	2	11	2	1	0	0
2022-05-08 00:00:00	1	0	1	1	5	1	12	2	1	0	2
2022-05-09 00:00:00	0	0	0	0	2	0	37	2	0	0	2
2022-05-10 00:00:00	1	0	0	4	4	5	68	4	0	0	2
2022-05-11 00:00:00	0	0	0	5	1	4	15	0	0	0	4
2022-05-12 00:00:00	0	0	1	1	1	8	15	1	0	0	4
2022-05-13 00:00:00	0	0	1	2	2	5	6	5	0	0	0
2022-05-14 00:00:00	0	0	0	0	3	3	2	0	0	0	2
2022-05-15 00:00:00	0	0	1	0	0	3	4	2	0	0	6
2022-05-16 00:00:00	0	0	1	4	1	8	12	2	0	0	5
2022-05-17 00:00:00	0	0	0	3	0	17	15	4	0	0	3
2022-05-18 00:00:00	1	0	0	4	0	4	11	4	0	0	3
2022-05-19 00:00:00	0	1	0	3	0	5	13	10	0	0	8
2022-05-20 00:00:00	2	0	0	1	0	7	15	4	0	0	3
2022-05-21 00:00:00	0	0	0	1	0	33	6	32	0	0	1
2022-05-22 00:00:00	0	0	0	0	0	4	4	3	0	0	2
2022-05-23 00:00:00	0	0	1	2	14	3	53	5	0	0	6
2022-05-24 00:00:00	0	0	0	1	1	8	18	5	0	0	2
2022-05-25 00:00:00	1	0	0	1	2	3	65	3	0	0	8
2022-05-26 00:00:00	0	0	2	3	20	3	20	3	0	0	9
2022-05-27 00:00:00	0	0	1	1	1	0	7	0	0	0	1
2022-05-28 00:00:00	0	0	0	1	0	1	11	1	0	0	1
2022-05-29 00:00:00	0	0	0	0	2	0	10	1	0	0	2
2022-05-30 00:00:00	0	0	1	0	0	0	11	0	0	0	1
2022-05-31 00:00:00	0	0	0	0	0	0	0	0	0	0	0



# 감사합니다.