| 『제1회 KRX 금융 빅데이터 활용 아이디어 경진대회』제안서 |
| --- |

| **󰊱 아이디어 명** | AI 댓글 감성분석 점수를 활용한 국내 주식시장 심리지표 | **󰊲 구분** | **개인 □** | **팀 ~~□~~** |
| --- | --- | --- | --- | --- |
| **󰊳 팀 명** | jykl | **󰊴 팀 구분**  (󰊲 ‘팀’ 체크 시) | **일반 ~~□~~** | **법인 □** |
| **󰊵 법인 명**  (󰊴 ‘법인’ 체크 시) |  | **󰊶 법인등록번호**  (󰊴 ‘법인‘ 체크 시) |  | |
| **아이디어 개요** | | | | |
| CNN 공포-탐욕 지수를 벤치마킹한 **국내 주식 시장 특화형 공포-탐욕 지수 모델 개발**   * 주식 시장 참여자들의 심리 상태를 분석하고 이를 점수화(0~1)하고 지수화하여 투자자들의 의사결정에 도움을 줄 수 있는 국내 주식 시장 투자 심리지수 제공 | | | | |
| **아이디어 제안 배경 및 목적** | | | | |
| 1. **시장의 비이성적인 행동 흐름 포착**  * 행동재무학(Behavioral Finance)에서는 자본 시장 참여자들이 이성적인 판단에 의한 합리적인 의사결정을 하는 것이 아닌 비이성적인 판단에 의해 **비합리적인 의사결정**을 하는 경향을 보인다고 주장함. * 특히 참여자들의 **기대 심리에 의해 판단의 오류**가 발생하게 되는데, 이러한 판단의 오류가 지속적으로 발생할 경우 주식 시장의 과열 혹은 침체가 지속될 수 있음. * 따라서, 주식 시장이 비합리적인 의사결정에 의해 움직이고 있는지를 파악하고자 함.  1. **투자자들의 정확한 심리 파악**  * 주식시장의 투자 심리지수는 주식 투자 전략이나 기업재무의 연구에서 유용한 변수로 이용되고 있음. * 하지만 투자자들의 심리는 주관적이기 때문에 정확한 심리 파악이 어려움. * 그러나, 최근 AI 기술, 특히 자연어 처리 분야의 발전으로 **BERT**같은 **대용량 데이터**를 미리 **사전에 학습시킨 모델**을 이용하여 사람이 작성한 문서를 분석하는 기법의 정확도가 크게 상승하였음. * 따라서, 이러한 딥러닝 모델의 **감성 분석**(Sentiment Analysis)을 이용하여 투자자들의 심리를 정확하게 파악하여 심리 지수에 반영함.  1. **불완전한 투자자들에게 위험 신호 제공**  * 합리적인 투자자들은 모든 정보를 이용하여 시장에 참여하지만, 경제 혹은 기업에 관련된 정보에 대해 무지한 일반 대중적인 투자자, 소위 개미 투자자들은 비합리적인 의사결정을 할 가능성이 높음. * 이러한 일반 개인 투자자들에게 **비정상적인 시장 흐름 여부를 인식**시켜 비합리적인 의사결정을 막는데 도움을 줌.  1. **정형 데이터와 비정형 데이터 조합**  * 기존의 투자 심리지수는 **거래량** 혹은 **VIX** 지수 등 정형 데이터로만 구성되어 있어 시장의 심리를 숫자를 통해 간접적으로 파악하였음. * 반면에, **텍스트**같은 비정형 데이터를 통해 시장 참여자들의 심리를 분석하고 집계하여 시장의 심리를 파악할 수 있음. * 따라서, 전통적으로 사용되어 왔던 투자 심리지수에 감성 분석한 비정형 데이터를 반영하여 최근 **기술 발전 트렌드를 반영한 투자 심리지수**로 재구성하려 함. * 또한, **대외 거시경제**를 반영할 수 있는 환율 데이터를 추가함으로써 거시적인 환경도 반영할 수 있도록 구성함. | | | | |
| **활용 데이터** | | | | |
| 1. 비정형 데이터 2. 유튜브 댓글 데이터 : 삼프로TV\_경제의신과함께 3. 종목토론실 댓글 데이터 : KTOP30 기업   대상 기업 목록   * '삼성전자', 'NAVER', '삼성SDI', 'LG화학', '카카오', '유한양행', 'SK텔레콤', 'POSCO홀딩스', '현대모비스', '삼성화재', '셀트리온', '현대차', '롯데케미칼', 'SK이노베이션', '아모레퍼시픽', '삼성전기', '삼성물산', 'CJ ENM', '이마트', 'SK하이닉스', 'LG전자', '한국조선해양', '기아', '넷마블', '삼성생명', 'KB금융', '현대건설', '신한지주', 'LG디스플레이', '미래에셋증권'  |  | From | To | # of Data | | --- | --- | --- | --- | | Raw Data | 2020-06-01 | - | 3,181,878 | | Train Data | 2020-06-01 | - | 32,013 | | Test Data | 2020-06-01 | - | 8,000 | | Prediction Target | 2022-06-01 | 2022-06-30 | 141,255 |  * 유튜브 Raw Data : 437,574개 * KTOP30 Raw Data : 2,744,304개 * 학습기에 들어갈 약 4만 개 데이터 중에서 3만개는 KTOP30, 1만 개는 유튜브 댓글에서 랜덤추출. * KTOP30 에서 랜덤추출된 3만 개 데이터는 KTOP30 기업 비중에 따라 추출. * KTOP30 비중 기준일자 : 2022-07-15   2. Train Data 레이블링 기준   1. 공포탐욕 사전 구성방법 :   [1] KNU 감성사전 긍정 : **6,223** 부정 : **6,451**을 탐욕/공포 사전으로 이용  [2] Raw Data corpus 로 학습된 soynlp 명사분류기를 이용하여 토크나이징한 후, 사용된 갯수를 나열하여, 100개 이상으로 나오는 단어 위주로 확인하여 공포/탐욕 단어 기준으로 직접 추가.  [3] 중복이 있다면 제거하여 사용.   1. 토크나이징 댓글 중에 탐욕사전에 있다면 +1 공포사전에 있다면 -1 을 해주고 양수이면 1로 레이블링 음수이면 -1로 레이블링.  * 만약 0 이라면 Train Data 기준에서 누락.     3. 정형 데이터   1. KTOP30 일일 거래회전율 : **(30개 종목의 총 거래량의 자연로그 값 / 30개 종목의 총 상장 주식 수 평균의 자연로그 값)**  * 각 값의 크기가 매우 크기 때문에 로그스케일링 이용.  1. 환율 : max값을 1500, min값을 900으로 min-max-scaling을 취한 후, 그래프를 뒤집기 위해 **1 - (min-max-scaling 환율값)** 을 취함. | | | | |
| **상품·서비스·콘텐츠 세부내용** | | | | |
| 1. 제공 서비스: 국내 주식시장에 대한 개인투자자들의 **감정 집합 해석 서비스**      1. 서비스 상세  * **댓글 데이터**를 활용하여 내포된 공포/탐욕 정도 수치화 * **거래회전율, 환율** 등 경제 지표와 결합 * 5가지 상태로 구분: 극단적 탐욕-탐욕-중립-공포-극단적 공포 * 국내 주식시장에 특화된 공포-탐욕 지수 제공   3. 서비스 제공 화면 예시 | | | | |
| **기존 상품·서비스·콘텐츠와의 차별성** | | | | |
| 1. 기존에 익숙한 정형 데이터 위주의 분석에서 벗어나 **텍스트 데이터**를 활용한 **감성분석**을 통한 국내 주식 시장 흐름을 분석 2. 주식시장에 내재된 개인투자자들의 감정 집합(공포-탐욕 지수)을 점수로 변환하여 서비스로 제공 3. 기존의 투자자의 기대 심리에 영향을 미칠 수 있는 긍정적/부정적 뉴스를 분석하는 것보다 댓글 및 SNS를 분석함으로써 비합리적인 의사결정을 하는 투자자들의 심리를 더 잘 반영할 수 있음 | | | | |
| **시장성 및 사업화 가능성** | | | | |
| * 국내 개인투자자들의 움직임을 대변할 수 있을 것이라고 기대 * 특히, 개인투자자들의 투자 의사결정에 참고 지표로 활용될 수 있을 것이라고 기대 * 타 지표만으로는 설명하기 힘들었던 시장 상황에 대한 부가 설명 지표로 활용될 수 있을 것이라고 기대 * 전문 투자자 혹은 기업에게 심리 지수를 제공함으로써 수익(이용료)을 거둘 것으로 기대 | | | | |
| **기타(아이디어 강점 등 자유 기재)** | | | | |
| **아이디어 강점**   1. 비정형 데이터 모델 2. koBERT 사전학습 언어모델 사용 : 한국어에 대한 학습이 이루어진 언어모델로, 한국어 예측에 있어서 좋은 성능을 보임 3. 약 4만개의 대용량 데이터를 이용하여 예측하였으며, 향후 더 큰 규모의 데이터를 훈련 데이터로 사용하여 예측 성능을 더욱 높일 수 있음 4. 2020년 이후 국내 주식 시장의 과열 시기와 2022년 침체 시기의 데이터를 학습하였기 때문에 양질의 데이터로 학습하여 예측 성능이 더욱 개선됨     2. 신조어 반영  soynlp 라이브러리를 이용하여 주식과 관련된 신조어(예 : 가즈아, 존버 등)를 학습시켜 단어집에 포함시켰으며, 감정을 나타내는 신조어를 새로 분류하여 국내 주식 관련 댓글 및 SNS에 적합한 토크나이저를 갖춤.  3. 정형 데이터로 보완  기존의 경제 이론에서 사용되던 심리 지표(거래회전율)와 거시적인 경제 상황을 반영(환율)하는 지표를 이용해 비교적 최근 기법인 비정형 데이터를 이용해 심리를 분석하는 방법론을 보완하여 더욱 타당성을 갖춤. | | | | |

※ 법인을 대표하여 팀을 구성하는 참가팀은 󰊴의 ‘법인’에 체크(**팀원 구성은 동일 법인 소속직원으로 제한**)