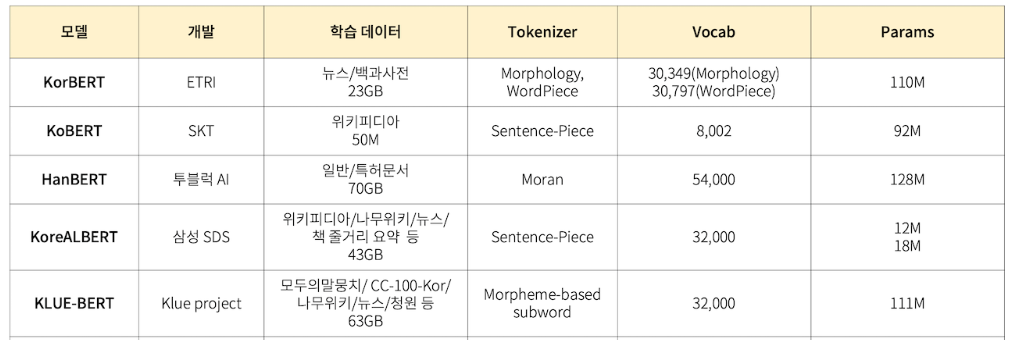
* 과제 동기(왜?)
  + 행동경제학 🡪 인간의 경제적 선택에는 다양한 심리적 요인이 작용하고, 제한된 합리성으로 의사결정시 비합리적인 선택을 하기도 한다고 전제
  + 이러한 행동경제학의 전제는 주식 가격의 결정에도 적용된다. 즉 주가가 실물경제 펀더멘털의 영향을 받을 뿐 아니라 투자자들의 투자심리에 의해서도 상당 부분 영향을 받는다는 것인데, 뉴스에 의한 주가영향도를 연구해서 투자를 위한 지표로 활용하기 위해 뉴스 지수 개발
  + 뉴스 ref) **홍콩·싱가포르 ‘숏쟁이’들 에코프로 때문에 줄퇴사... 그래도 계속친다, 공매도**
  + 개인 투자자들은 기관투자자에 비해서 투자에 관한 정보가 부족, 그들을 위한 투자보조지표 개발
  + Ref ) 처음 만나는 행동경제학, 신문기사(utm\_source=naver&utm\_medium=original&utm\_campaign=biz), 논문(다수)
* NLP( Natural Language Processing, 자연어 처리 )
  + 우리가 평상시에 사용하는 언어를 자연어(뉴스기사,게시판 글, 영화리뷰)라고 하는데, 인간처럼 인공지능이 텍스트와 데이터를 처리하고 해석하도록 하는 머신러닝
  + 자연어를 긍부정으로 감성분류 하거나
  + 문장의 단어 연관도와 순서를 학습해서 다음에 올 단어 예측
  + 사람들의 질문을 이해하고 답변(ex. Chat GPT`)
  + 이번 과제에서는 뉴스라는 자연어를 활용하여 긍정, 부정 분류하는 머신러닝 기법
* Word cloud
  + 데이터에서 얻어진 단어들을 중요도나 빈도수를 고려하여 시각적으로 표현함
  + 긍정뉴스만, 부정뉴스만을 모아서 워드클라우드 생성
  + 빈도수가 많을수록 단어 크기를 크게 표현함
* KLUE BERT(Korean Language Understanding Evaluation )



* + 한국어 전용 pretrained bert모델
  + KLUE : 한국어 언어모델의 공정한 평가를 위한 목적을 달성하기 위한 데이터 집합체(KLUE Project)
* BERT(Bidirectional Encoder Representations from Transformers)
  + 어휘의 양방향 문맥정보와 문장 간의 선후관계를 학습하여 단어를 문맥을 반영한 벡터로 표현하는 모델
  + 나는 사과를 먹었다/ 친구에게 사과했다. 사과의 문맥적 의미 파악
* F1-Score
  + Negative를 negative라고 정확히 예측하는 지 확인하는 지표
  + 100명중에 암환자 1명 있다고 가정, 99명의 정상인을 모두 맞췄다고 성능 좋은 모델이라 말할 수 없다. 1명의 암환자를 맞췄는지 성능을 나타냄
* K-means clustering 유사도 활용한 복제
  + 비슷한 특성을 가진 데이터끼리 클러스터, 즉 무리로 묶는 군집화 기법
  + 클러스터의 중심과 데이터들의 거리의 평균을 최소화하게 여러 번 군집에 할당하고, 중심점을 갱신하면서 반복학습, 더 이상 갱신 안 될때까지
* XGBooster
  + XG : Extreme Gradient 의 약자이다. 해석하면 극한 변화도(경사도, Gradient), 변화도는 예측률을 높이기 위해 방정식을 조정하는 정도, 변화도 조정성능을 극한으로 끌어올림
  + Booster : 여러 개의 약한 의사결정나무(Decision Tree)를 조합해서 사용하는 앙상블(Ensemble) 기법 중 하나/약한 예측 모형들의 학습 에러에 가중치를 두고, 순차적으로 다음 학습 모델에 반영하여 강한 예측모형을 만든다.
* Optuna
  + 모델 튜닝 프로세스
  + 모델의 최적의 하이퍼파라미터제시
* H2o AUTOML
  + AUTOML : Automated Machine Learning 머신러닝 모델 개발작업을 자동화하는 프로세스, 여러 머신러닝 중 최적의 모델 선정, 최적의 모델튜닝 제시
* LSTM(Long Short – Term Memory)
  + 딥러닝 모델, 주로 시계열이나 자연어 처리에 이용 되는 모델
  + 입력 데이터가 길어지면 후반부 데이터의 특성은 학습에 반영되지 않는 기울기 소실 문제가 발생, 이 부분을 개선시켜 장기, 단기 데이터 모두 중요도 있게 학습시킴
* Bi – LSTM
  + 양방향 LSTM, 뒤에서 앞으로(역방향) 실행되는 다른 LSTM을 추가, 성능 향상