프로젝트 진행 상황

: kf-deBERTa 파인튜닝을 완료함

YG엔터테인먼트의 해당 기간 뉴스를 크롤링 하기 위해 여러 방법을 생각해봄

크롤링은 beautiful soup을 활용해 진행.

1, 네이버 파이낸스 페이지에서 크롤링-> 기간선택을 제공하지도 않고 최근 1년간의 기사만 제공함

2, 네이버 포털에서 yg를 검색하여 3년간 기간 설정을 해두고 크롤링

->우리가 설정해둔 날짜 이외의 날짜도 크롤링 되는데, 하루에만 쏟아지는 기사가 최소 몇십개이기 때문에 성능 뛰어난 GPU가 아니면 잘 크롤링이 되지 않고 끊겼음

3, 네이버 포털에서 yg를 검색하고, 주요 사건 타겟 날짜를 지정하여 61번의 크롤링을 진행, 엑셀파일로 만들어 병합

4, 해당 프로젝트의 목적: 사건이 주가에 미치는 영향을 보는 것일지, 뉴스의 표현이 주가에 미치는 영향을 보는 것일지 중 후자에 가까움. 당연히 사건은 주가에 영향을 미치는데, 사람들이 이를 접하는 뉴스의 표현 방식과 포장 방식에 따라서 주식시장에서의 반응을 보고 싶었던 측면이 존재.

5, 파인튜닝 해둔 kf-deBERTa 모델에 해당 크롤링 결과물을 투입하여 sentimental analysis를 진행하고 부정,중립,긍정 결과를 마찬가지로 라벨링

6, 회귀분석을 앞둔 상황에서, 고려할 것들이 존재

첫번째로, 야후 파이낸스에서 크롤링 해온 주가데이터를 어떻게 가공하여 수치데이터로 활용할 것인가? 당일 시가와 당일 최대가를 놓고 상승률을 구해야 할지, 다음날 최대가로 변화률을 구해야 할지, 당일과 다음날의 각각 평균주가의 비교로 변화률을 구해야 할지 고민이 있다. 두 번째로, 회귀분석 시 dummy variable을 활용해야 하는데, 이때 긍정 중립 부정 에 대해서 각각 2,1,0으로 라벨링을 한 것이 적합할지에 대한 고민이다. 2,1,0보다는 1,0,-1로 그 의미를 살려 allocate code 방식으로 할당을 활용하는 것이 좋을 것 같다는 생각도 든다. 아직 회귀직선을 그리기 전 단계라 좀 더 생각을 해보아야 한다.

따라서 회귀직선은 allocated code를 활용하여 라벨링된 뉴스의 수치형 데이터와, 주가의 변동률(상승 및 하락률) 사이의 관계를 입증하게 될 것임. 라벨링 된 감정분석 결과들이 각 숫자들에 중복되어 있을 것이므로 Lack of fit test 등을 R로 수행하여 단순선형회귀모형에 적합한지에 대해서 파악하고 그려볼 것이고 그렇지 않다고 한다면 상관계수를 구해보는 정도로 효과를 설정해볼 것 같음.

이 프로젝트가 지니는 의미에 대해 생각해봄. 분석적 역량을 키우기 위해 데이터와 소통하는 경험을 최대한 늘려보고 자기주도적으로 고민해보고 싶었기 때문임. 사실 이 프로젝트는 ‘사건’이 아니라 ‘날짜’에 의미가 더 가중되어야 함. 사건이 주가에 미치는 영향이 아니라, 해당 날짜에 등장한 뉴스들에 드러난 어조와 감정이 주가에 미치는 영향을 보는 프로젝트이기 때문임. 그러나 사건이 뉴스의 소재가 되는 것은 부정할 수 없기 때문에, 사건의 성격도 물론 배제할 수 없을 것이라고 생각함. .