기존에 nlp 분야에서 활용해온 데이터 증강 기법으로는 세 가지가 있다.

첫째, 문장을 다른 언어로 번역하고 다시 기존의 언어로 번역하여 새로운 데이터를 얻는 방법

둘째, 문장에 노이즈를 가볍게 주는 방법(data noising as smoothing)

셋째, 문장 내 단어 중 유의어로 교체하는 방법

2019년 EMNLP에서 발표된 EDA: Easy Data Augmentaion Techniques for Boosting Performance on Text Classification Tasks 논문에서 텍스트 분류 작업의 성능을 부스팅 해줄 쉬운 데이터 증강 방법 네 가지를 소개했다.

SR (Synonym Replacement) : 문장 내 stop word가 아닌 랜덤으로 n개의 단어들을 선택해 임의로 선택한 동의어들 중 하나로 바꾸는 방법

RI (Random Insertion) : 문장 내 stop word를 제외한 나머지 단어들 중에서, 랜덤으로 선택된 단어의동의어를 임의로 정하고 해당 동의어를 문장 내 임의의 자리에 삽입

RS (Random Swap) : 무작위로 문장 내 두 단어의 위치를 바꿈

RD (Random Deletion) : 확률 p를 정해 문장 내 임의의 단어들을 랜덤하게 삭제

논문에서 해당 EDA 방법이 CNN 및 RNN 모델에서 성능을 향상시킨다는 것을 보여주었고,

특히 데이터가 적은 경우에 더 강력하다고 설명하고 있다.

한계점으로는 데이터가 충분할 때 성능적으로 이득을 취하기가 제한적 일 수 있다는 것이다. 또한 pre-trained model을 활용하는 경우 상당한 개선 효과를 거두지 못할 수 있다고 설명하고 있다.

따라서 본 프로젝트에서 open-source python library인 nlpaug를 활용하여 data augmentation을 시도하였다. 그 중 사전학습모델 BERT를 활용하여 문맥의 흐름에 맞는 임의의 단어를 삽입하는 방법을 통해 train\_negative 데이터를 9만개 추가로 생성하였다.

Data augmentation을 통해 생성된 9만개와 기존 17만개를 합쳐 총 26만개의 negtive를 데이터를 학습에 활용하였으나, 기존의 높은 성능에 있어 크게 개선 효과를 보지 못하였다.