[월간 데이콘] 법원 판결 예측 AI 경진대회

데이터

- train.csv
  - ID : 사건 샘플 ID
  - first_party : 사건의 첫 번째 당사자
  - second_party : 사건의 두 번째 당사자
  - facts : 사건 내용
  - first_party_winner : 첫 번째 당사자의 승소 여부 (0 : 패배, 1 : 승리)
- test.csv
  - ID : 사건 샘플 ID
  - first_party : 사건의 첫 번째 당사자
  - second_party : 사건의 두 번째 당사자
  - facts : 사건 내용
- **sample_submission.csv**
  - ID : 사건 샘플 ID
  - first_party_winner : 예측한 첫 번째 당사자의 승소 여부 (0 : 패배, 1 : 승리)

코드흐름

(1) EDA

- Will 이름을 가진 사람들을 WillIn으로 변경 → 변경을 하지 않으면 불용어처리에서 지워짐
- United States 계열 이름들을 전부 USA로 통일

- A., S. 과 같이 한글자 대문자들 제거
- Co., Bd., Mt. 약어들을 Company, Building, Mount로 변경

(2) 데이터 증강

- Spacy 언어 모델을 이용해서 명사구를 추출하고 뽑은 명사구를 제외하고 SynonymAug 모델을 활영해서 동의어 데이터 증강을 진행
- 영어 호칭 제거

(3) 모델 학습

- Kfold을 사용하여 교차 검증 진행

(4) 후처리

- 최종 예측 결과를 분석하여 성능을 개선하고, 예측 확률에 따라 결과를 보정
- 불확실성이 큰 예측에 대해서 추가적인 처리를 통해 보다 안정적인 결과 도출

배울점

- 증강 전 전처리에서 data를 시작 전 확실하게 분석을 한 다음 오류가 날 수 있는 부분을 사전에 정리하는 것이 인상적이었다 .
- Kfold 검증을 활용해 다양한 데이터 샘플에 대해 모델을 테스트하여 성능을 안정적으로 평가한 것이 인상적이었다.