Process of Fall Detection in Clinic Note

Phase 1. Pre-Preprocessing

0. 라이브러리 설치 및 파일 불러오기

- 라이브러리 설치 및 불러오기
- 파일 불러오기 (data, dictionary 등)
    - dictionary 중복 여부 확인
    - 원본 데이터 알아보기
        - 결측값 여부 확인
        - 낙상 데이터와 비낙상 데이터의 개수, 비율 확인

1. 데이터 전처리

- 1. 개행 문자 제거
- 2. 단어 정규화
- 3. 검사 측정 구문 제거
- 4. 대문자를 소문자로 변환
- 5. 약물명 번역
- 6. En2Ko 번역
- 7. 약어 번역
- 8. 불용어 삭제
- 9. 남은 영어, 숫자, 특수문자 삭제

2. 단어 토큰화

- 동사, 명사, 부사의 형태소 분석을 통해 토큰화
- Term Frequency 추출

Phase 2. (필요 시) Word2Vec 모델 학습

Phase 3. Text Classification

- 모델 학습 전 imblearn 라이브러리를 활용한 오버샘플링
- sklearn 라이브러리를 활용한 train, val, test 데이터 분할 (train_test_split)

A. LSTM으로 낙상 데이터 이진 분류하기

- 아직 진행 x

B. FastText를 통한 Text Classification

- 1. fasttext 라이브러리 설치 및 불러오기
- 2. FastText 학습을 위한 .txt 파일 생성
- 3. FastText 모델 학습 및 추론

2022. 04. 11. 분석 결과

결론 먼저 확인하고 싶으면 클릭

Data Info.	label 0 (Non-Fall)	label 1 (Fall)	Total
Original Data	9838 (52.638%)	8852 (47.362%)	18690
Resampled Data	9838 (50%)	9838 (50%)	19676 (100%)
Resampled Training Data	5901 (50.013%)	5901 (49.987%)	11805 (59.997%)
Resampled Validation Data	1002 (50.863%)	968 (49.137%)	1970 (10.012%)
Resampled Testing Data	2935 (49.712%)	2969 (50.288%)	5904 (30.006%)

2022년 04월 11일 분석에서는 Original Data를 오버샘플링 기법으로 데이터를 늘려 Resampled Data를 생성하였다. 또한, 그 Resampled Data를 60:10:30의 비율로 각각 Traing Data, Validation Data, Testing Data로 구성하였다.

위 표는 앞서 설명한 5가지 종류의 데이터들에 대한 label의 비율과 데이터의 비율을 나타낸 표이다.

Original Data에서 비낙상:낙상의 비율이 52:47이었지만, 오버샘플링을 한 데이터들은 비율이 50:50으로 고정된 것을 확인할 수 있다. 비율이 52:47에서 50:50으로 맞춰준 이 과정이 효과가 없고 비효율적으로 느껴질 수 있지만, 추후 Original Data의 비율이 imbalance 할 때의 상황을 대비한 것이다.

UsedData	Raw Data	Processed Data	Tokenized Data
Precision	[**0.99383562** 0.98894102]	[0.99147049 0.99024554]	[0.99046646 **0.99123694**]
Recall	[0.98875639 **0.99393735**]	[0.99011925 0.99157966]	[**0.9911414** 0.99056922]
F1 score	[**0.9912895** **0.99143289**]	[0.99079441 0.99091215]	[0.99080381 0.99090296]
Accuracy Score	**0.9913617886178862**	0.9908536585365854	0.9908536585365854

위 표는 데이터의 분석 정도(깊이)에 따른 성능을 확인할 수 있는 표이다.

Raw Data는 말그대로 원본 데이터를 의미하며, 아무런 전처리과정을 거치지 않은 데이터이다.
Processed Data는 Raw Data에서 전처리 단계를 거친 데이터를 의미한다.
Tokenized Data는 Processed Data를 토큰화시킨 데이터이다. 즉, 전처리와 토큰화 과정이 모두 거쳐진 데이터이다.

위 표의 Precision과 Recall, F1 Score의 값이 대괄호([ ])로 둘러쌓여진 것을 볼 수 있는데 이는 순서대로 Label 0(비낙상)에 대한 값과 Label 1(낙상)에 대한 값이다.

Raw Data, Processed Data, Tokenized Data 모두 (Accuracy Score 기준) 0.99이상의 성능을 보여주었다. 아무것도 처리하지 않은 data인 Raw Data의 F1 Score가 가장 높았을 뿐만 아니라 거의 모든 평가 지표에서 작지만 우세한 것을 확인할 수 있다.

결론

> 1. label 비율이 53:47이었던 원본 데이터를 50:50으로 오버샘플링 하였음.
>
> 2. 오버샘플링한 데이터를 60:10:30의 비율의 Traing Data, Validation Data, Testing Data로 구성하였음.
>
> 3. Raw Data, Processed Data, Tokenized Data 중 근사한 차이지만 가장 성능이 우세한 것은 Raw Data임.