Nutchapol Winmoon 6231320321

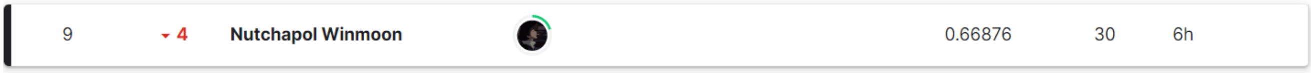
Take Home Exam Report
2110572 2022/2 NLP SYS

Methodology for best model

1. Text Preprocessing: including the followings steps:
    - Remove html tags
    - Remove punctuation and numbers
    - Remove stop words in both Thai and English using stop words list from PyThaiNLP (Thai) and NLTK (English)
    - Remove location words using Name-entity recognition model (pretrained) from PyThaiNLP (Thai) and NLTK (English)
2. Train-validation split: 90/10 split with stratification
3. Augmentation:
    - adding label name from “occupation_mapper.csv” to train set
4. Text Encoder: Universal Sentence Encoder (USE)
    - Using pre-trained model from https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3
    - Convert each texts into 512-sized embedded vectors
    - Why choosing this encoder:
        - Multilingual model: This encoder supports both Thai and English so that it can represent texts with similar meaning in different languages in similar vectors
        - High quality embeddings: This encoder was trained on a very large corpus of text data
5. Classifier: Logistic Regression with hyperparameters as followed:
    - class_weight = “balanced”: To deal with class imbalance, assigns a higher weight to the minority class and a lower weight to the majority class
    - Why choosing this classifier:
        - Support multiclass classification
        - More efficient than other choices of classifier
        - Other classifiers, such as linear SVC and deep learning, were also experimented with and gave similar performance to one another. However, logistic regression achieved the highest score.

Result

Private score: 0.66876
Ranking: 9



9	▾ 4	Nutchapol Winmoon		0.66876	30	6h