Arabic Dialect Classification

Data fetching

break the large into several requests (number of IDs // 1000 = 458 requests) plus one more request for the remainder (197 IDs).

Data exploration & preprocessing

1 Imbalanced Data

The dataset is imbalance

- EG = 12.6%
- AE = 5.74%
- YE = 2.17%

2 Preprocessing Pipeline

- Remove username
- Remove URL from text
- Remove punctuation, emoji
- Remove \n, \t ,,, etc
- Remove Discretization
- Remove Arabic Stop Words

3 Oversampling

technique to balance the imbalanced nature of the dataset.

Machine Learning Models

For large datasets consider using:-

- linear support vector machine (SVM): score is 0.722
- Logistic Regression : score is 0.703
- SGD Classifier: score is 0.657

- Most misclassifications happen for each dialect, with dialects for share border countries. So we can make groups of share border countries.

Deep Learning Models

- Tokenize sentences into subwords or word pieces for the BERT model Given a vocabulary generated by the Word piece algorithm.

- Padding to the maximum number of words through all texts (94). We can add 6 more

- BRET score is 0.731

The Eiffel Tower

...is where I'd like to go next!