LANGUAGE RECOGNITION

ANDREA CADOLI

1837028

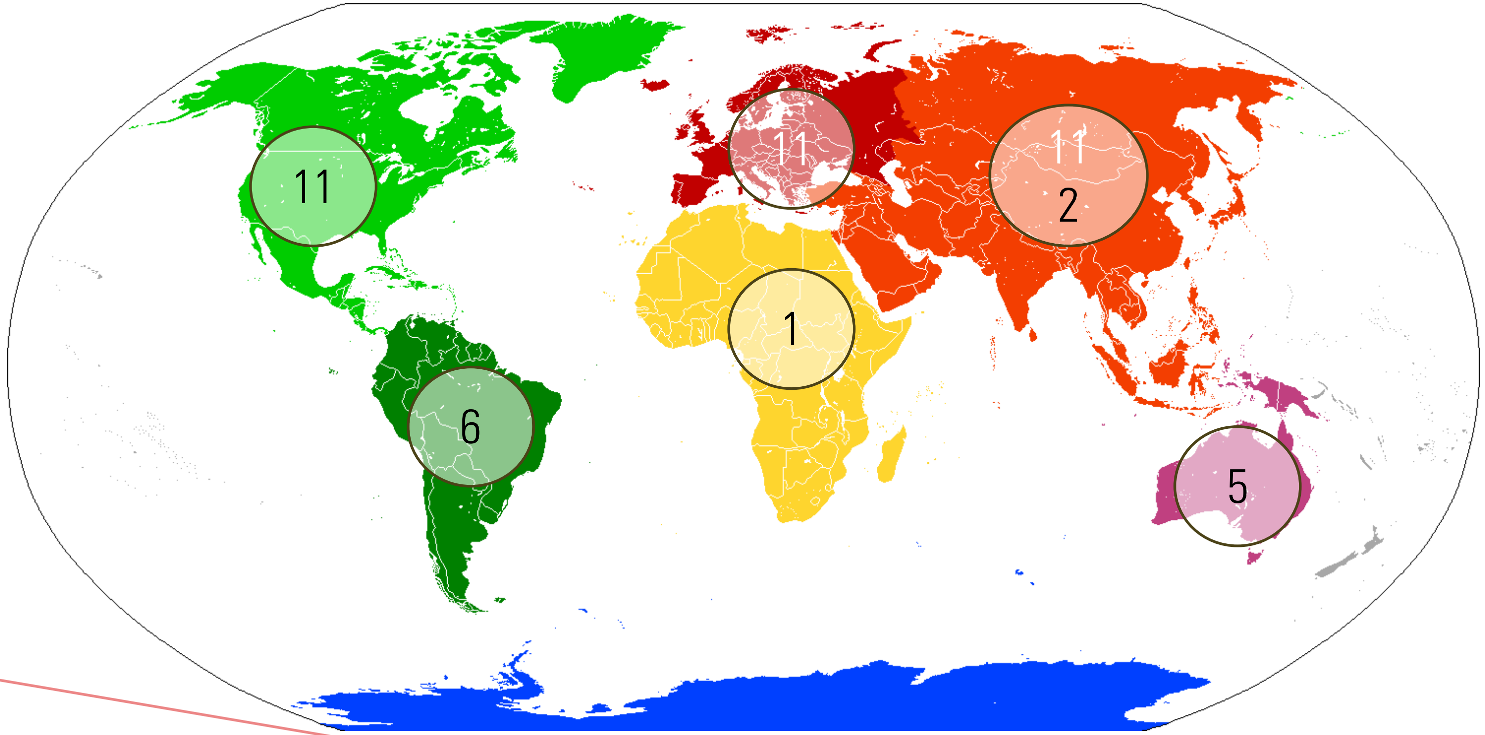
CORPUS

High-Resources Languages:

- (Part of) WiLI-2018
- 22 languages
- single csv file (1000 samples for each language)

Low-Resources Languages:

- Bible(s) from ScriptureEarth
- 25 languages
- multiple SFM files for each language



PREPROCESSING

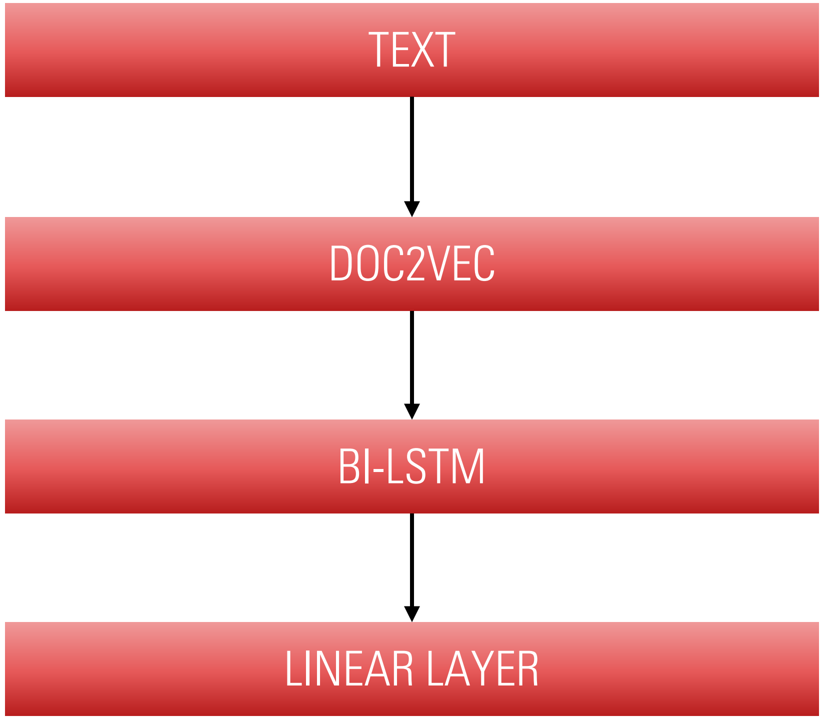
High-Resource data:

- remove punctuation
- remove numbers
- remove parenthesis
- remove extra whitespaces
- remove url addresses
- remove HTML tags

Low-Resource data:

- divide the documents into sentences (1000 for each language)
- remove punctuation
- remove numbers
- remove parenthesis
- remove extra white spaces
- remove ‘bible’ charachters (eg. ‘\v’)
- combine all the sentences in one single csv file

MODEL



TEXT

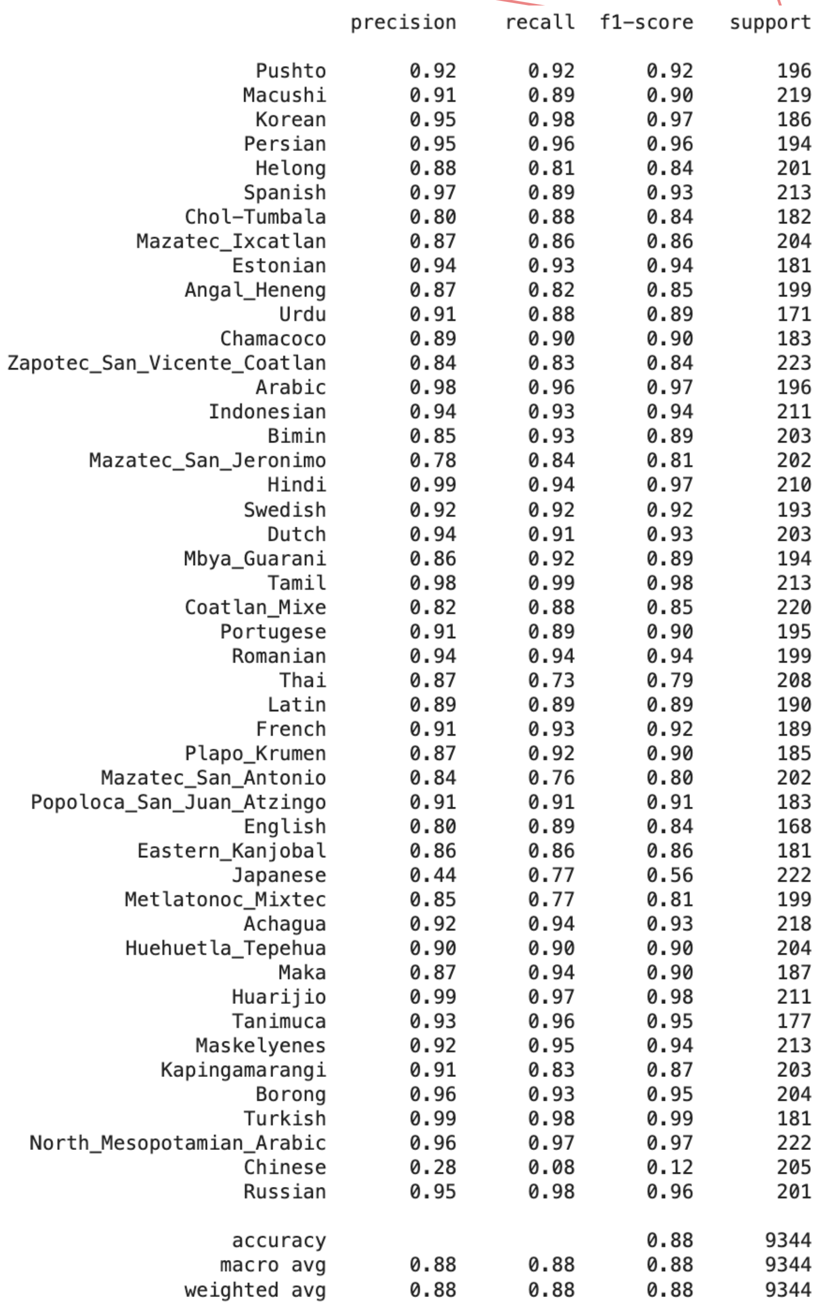
DOC2VEC

BI-LSTM

LINEAR LAYER

RESULTS

	ACCURACY		LOSS	
	TRAIN	VAL	TRAIN	VAL
LSTM (H-R dataset)	0.914	0.897	0.177	0.374
BI-LSTM (H-R dataset)	0.965	0.922	0.144	0.298
BI-LSTM (complete dataset)	0.928	0.886	0.206	0.418
BI-LSTM (complete dataset – {CH, JAP})	0.953	0.914	0.152	0.336

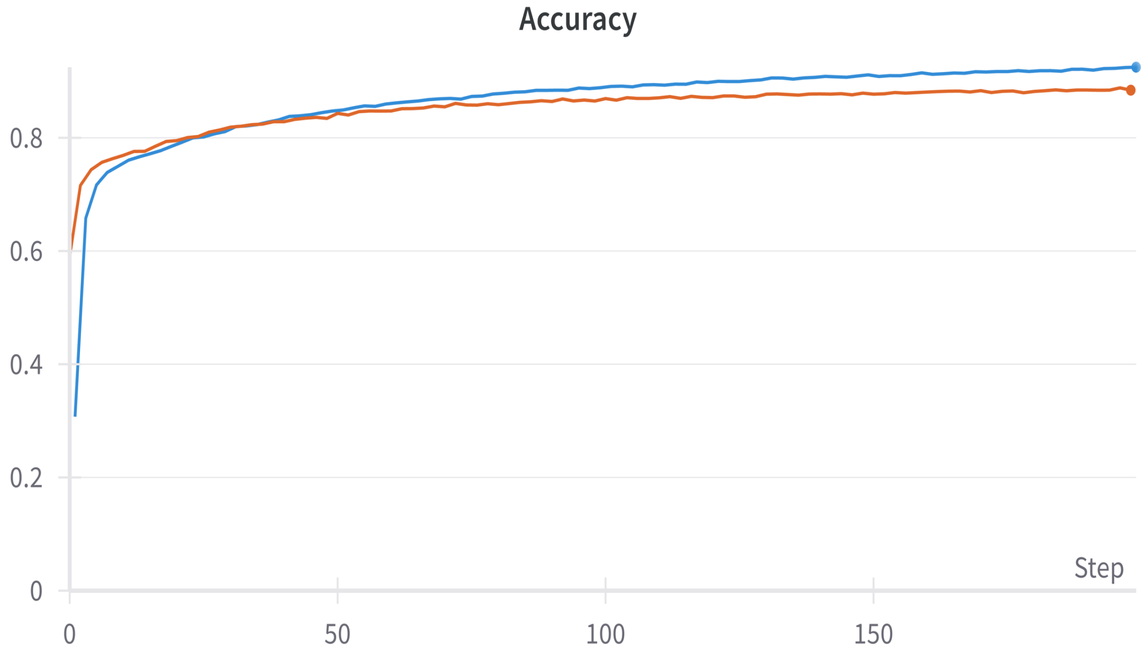


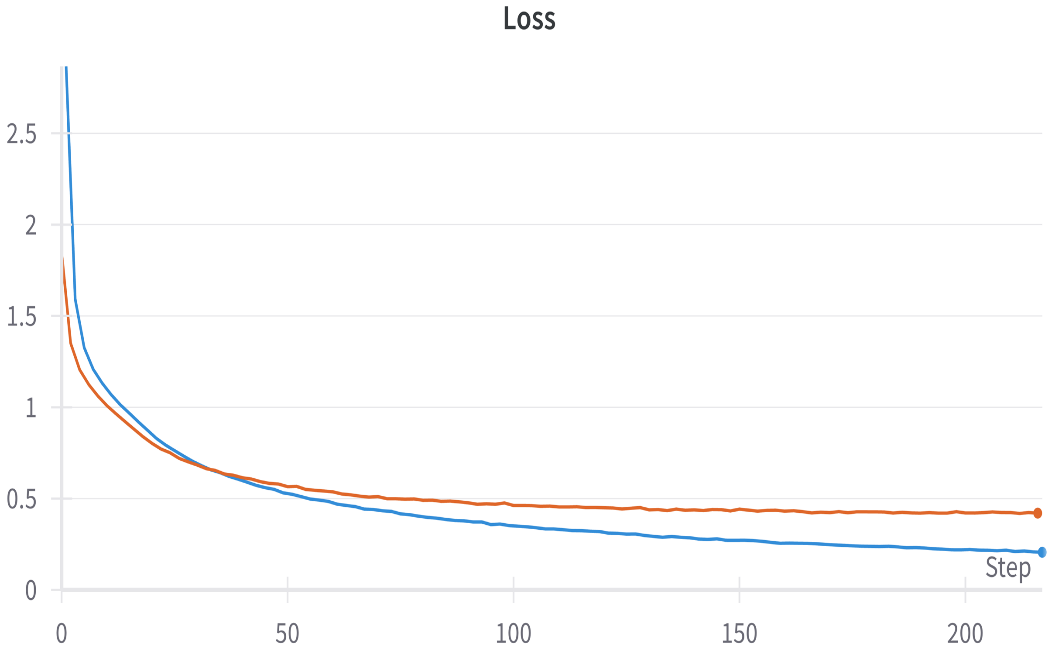
```
                            precision    recall  f1-score   support

                    Pushto       0.92      0.92      0.92       196
                   Macushi       0.91      0.89      0.90       219
                    Korean       0.95      0.98      0.97       186
                   Persian       0.95      0.96      0.96       194
                    Helong       0.88      0.81      0.84       201
                   Spanish       0.97      0.89      0.93       213
              Chol-Tumbala       0.80      0.88      0.84       182
          Mazatec_Ixcatlan       0.87      0.86      0.86       204
                  Estonian       0.94      0.93      0.94       181
              Angal_Heneng       0.87      0.82      0.85       199
                      Urdu       0.91      0.88      0.89       171
                 Chamacoco       0.89      0.90      0.90       183
Zapotec_San_Vicente_Coatlan       0.84      0.83      0.84       223
                    Arabic       0.98      0.96      0.97       196
                Indonesian       0.94      0.93      0.94       211
                     Bimin       0.85      0.93      0.89       203
     Mazatec_San_Jeronimo       0.78      0.84      0.81       202
                     Hindi       0.99      0.94      0.97       210
                   Swedish       0.92      0.92      0.92       193
                     Dutch       0.94      0.91      0.93       203
              Mbya_Guarani       0.86      0.92      0.89       194
                     Tamil       0.98      0.99      0.98       213
               Coatlan_Mixe       0.82      0.88      0.85       220
                Portugese       0.91      0.89      0.90       195
                  Romanian       0.94      0.94      0.94       199
                      Thai       0.87      0.73      0.79       208
                     Latin       0.89      0.89      0.89       190
                    French       0.91      0.93      0.92       189
              Plapo_Krumen       0.87      0.92      0.90       185
       Mazatec_San_Antonio       0.84      0.76      0.80       202
  Popoloca_San_Juan_Atzingo       0.91      0.91      0.91       183
                   English       0.80      0.89      0.84       168
           Eastern_Kanjobal       0.86      0.86      0.86       181
                  Japanese       0.44      0.77      0.56       222
          Metlatonoc_Mixtec       0.85      0.77      0.81       199
                   Achagua       0.92      0.94      0.93       218
          Huehuetla_Tepehua       0.90      0.90      0.90       204
                      Maka       0.87      0.94      0.90       187
                  Huarijio       0.99      0.97      0.98       211
                  Tanimuca       0.93      0.96      0.95       177
               Maskelyenes       0.92      0.95      0.94       213
            Kapingamarangi       0.91      0.83      0.87       203
                    Borong       0.96      0.93      0.95       204
                   Turkish       0.99      0.98      0.99       181
  North_Mesopotamian_Arabic       0.96      0.97      0.97       222
                   Chinese       0.28      0.08      0.12       205
                   Russian       0.95      0.98      0.96       201

                  accuracy                           0.88      9344
                 macro avg       0.88      0.88      0.88      9344
              weighted avg       0.88      0.88      0.88      9344
```

PLOTS


Accuracy
0.8
0.6
0.4
0.2
0
0
50
100
150
Step


Loss
2.5
2
1.5
1
0.5
0
0
50
100
150
200
Step