IDENTIFICATION OF ITALIAN LANGUAGE DIALECTS

Sravya Sangaraju (ssangara@gmu.edu) & Anisha Yidala (ayidala@gmu.edu)

Task:

- To build a classification model that identifies 11 regional languages and dialects of Italy

Dataset:

- Train dataset is a raw Wikipedia dump with 265016 Wikipedia articles
- Dev dataset has 6800 annotated sentences that cover only 7 out of 11 dialects evaluated

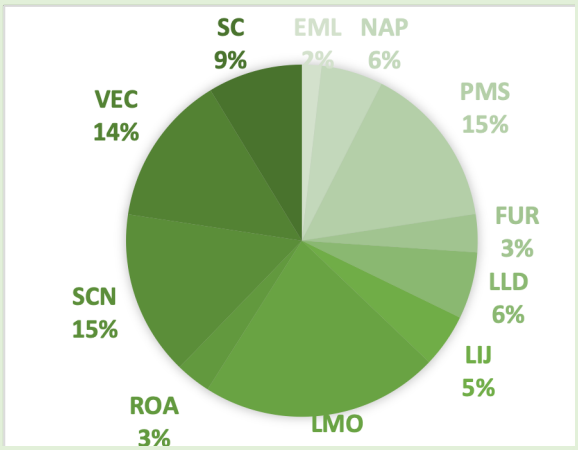
Generating Train Data:

- Using WikiExtractor to generate dataset
- Removing irrelevant information and duplicate data
- Converting documents to sentences
- Cleaning sentences

Implementation of Model & Results:

- Finetuning a BERT model that was pretrained on Italian Language corpora
- The accuracy of the model we implemented is 0.73
- The model was observed to perform diversly on various dialects

Distribution of 11 Dialects across Train data



Pre-processing steps	No. of documents
Original documents	265016
Remove length <50	245193
Remove duplicates	219178
Sentence split	562495
Sentence cleaning	388781

Dialect	Precision	Recall	F1	Support
EML	0.99	0.71	0.83	825
NAP	0.85	0.77	0.81	2026
FUR	0.98	0.95	0.96	1323
LLD	0.98	0.56	0.71	2200
LIJ	0.87	0.89	0.88	2282
LMO	0.39	0.98	0.56	689
ROA_TARA	0.57	0.09	0.15	603
VEC	0.53	0.59	0.56	1139