Machine Learning for NLP

Part 1

The dataset contains a number of languages with 1000 sentences per language, which means all classes are balanced. These languages range across a number of families including not latin languages such as Arabic, Japanese etc and prove to be a challenge.

Furthermore as the initial train test split is 50:50 we modify this to include greater training data i.e 80:20 training and testing respectively. We furthermore incorporate some linguistic features into our training dataset such as average word length and sentence length to extend the feature space.

Once we have the dataset ready we use count vectoriser and tf-idf to create a feature vector from the sentences. We then use logistic regression to predict the language. Using grid search we then search for the optimal hyper-parameters. We experiment with the following:

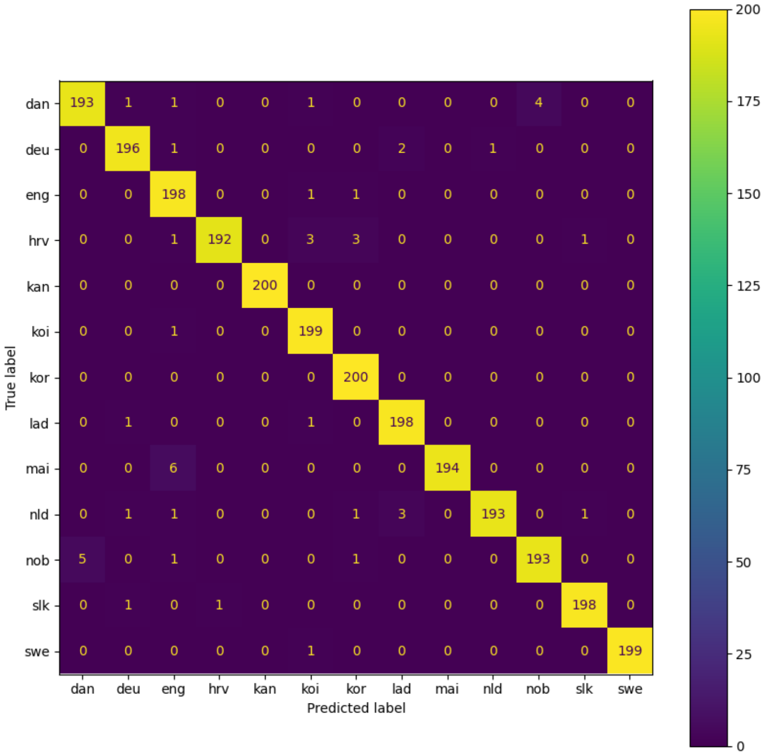
- L2 and L!
- LBFGS and Newton-CFGS solvers
- Different n-gram vectors (1,1) and (1,2)

Using the best pipeline we then look at the best features for logistic regression for two languages: . These hyper parameters are:

- L2 penalty
- Newton CFG solver
- Ngram range of (1,1)

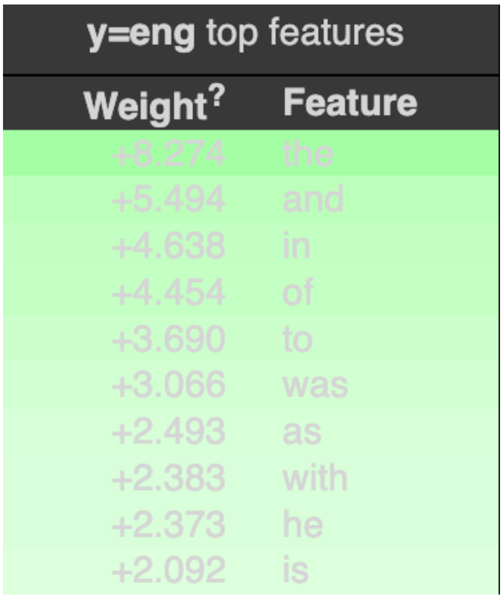
Based on the confusion matrix we get from the best model we see our model perform very well with an accuracy of 0.98. Furthermore we see that we have very high precision and recall as well.

Furthermore we also take a look at the logistic regression weights to understand the importance of features. We notice that certain words are very important to detect a language and the linguistic features are not considered



Confusion Matrix

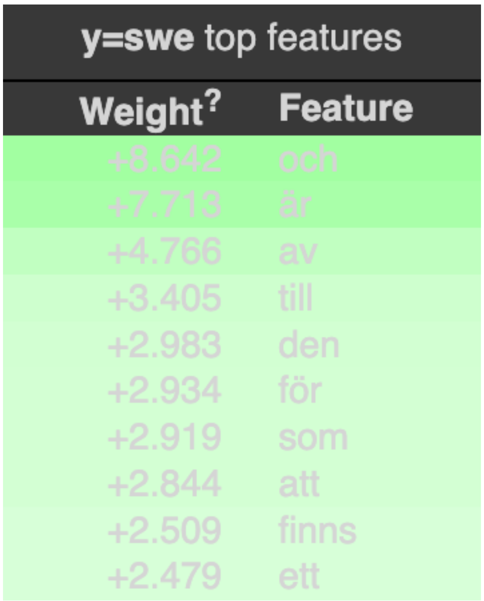
that important. Specifically the vectoriser output is considered very important while the extended features are not considered that important



y=eng top features

Weight?	Feature
+8.274	the
+5.494	and
+4.638	in
+4.454	of
+3.690	to
+3.066	was
+2.493	as
+2.383	with
+2.373	he
+2.092	is

Top features for English



y=swe top features

Weight?	Feature
+8.642	och
+7.713	är
+4.766	av
+3.405	till
+2.983	den
+2.934	för
+2.919	som
+2.844	att
+2.509	finns
+2.479	ett

Top features for Swedish

Our ablation study showed that just focusing on a few words is sufficient for classifying a language. Specifically for our best languages, limiting the maximum number of features did not reduce our accuracy.

Part 2

The dataset and its processing remains as before but without any additional linguistic features. We experiment with different architectures i.e number of layers and nodes per layer, activation functions, vectorisers and early stoping. Due to computational and time constraints we do not perform grid search but instead opt for a greedy approach i.e choose the best option from a single hyper-parameter and then modify the next one.

- We find that ReLu works better than Tanh and Sigmoid activation functions.
- Furthermore we keep our network with three layers of 500 x 250, 250 x 100 and finally 100 x 26. Increasing the number of layers increased the time but did not improve the performance significantly, while reduced layers decreased the accuracy.
- We try a range of learning rates and decide on 0.1 which yields the lowest validation accuracy.
- We also find that early stopping is better for this approach as training for longer yields to overfitting.

Using neural networks and an input feature vector of size 500 we get an accuracy of 97.2 %.