Language Identification for Text Documents

Arkajyoti Misra
Priyank Mathur
Emrah Budur

Problem

Problem definition

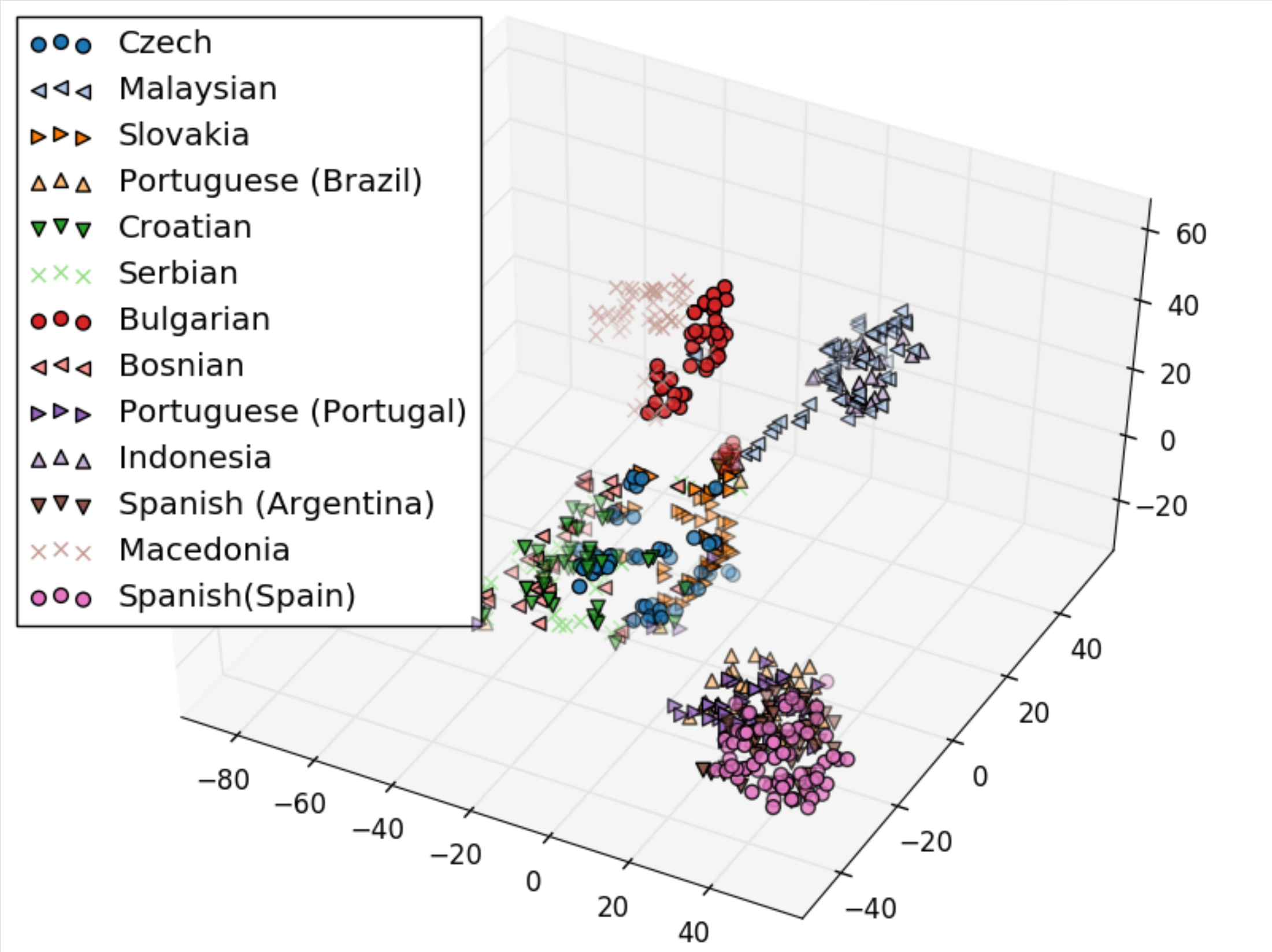
Identifying the language of short text documents with no prior knowledge, i.e. grammar rules.

Application areas

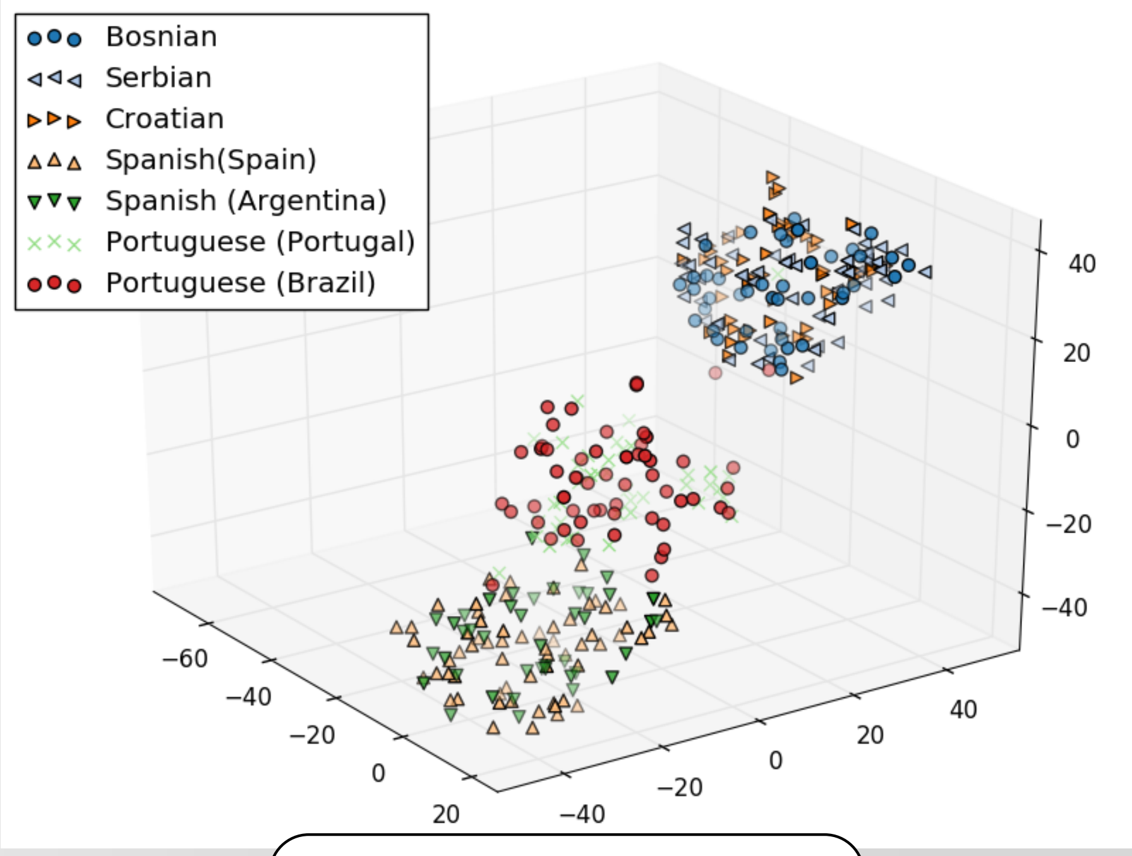
- One-click machine translation in social media
- Improving search relevancy

Data description

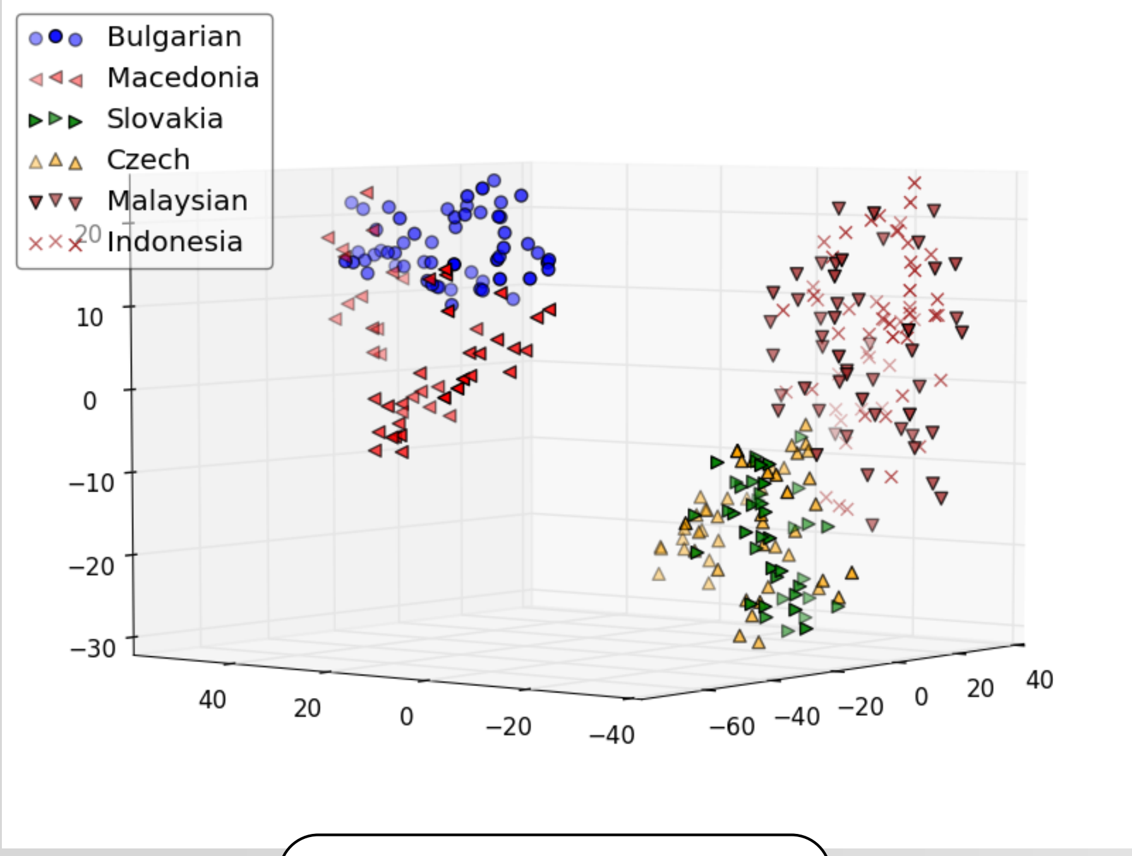
- Source: DSL 2015 competition
- 14 world languages with some very closely related ones (European vs Argentine Spanish)



Data Visualization (t-SNE plots)



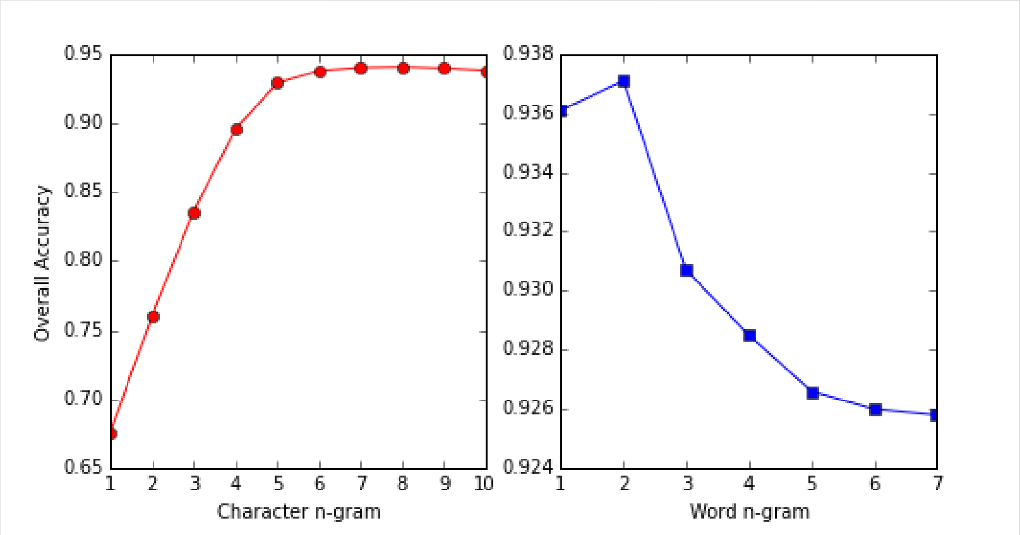
Clusters of highly similar languages



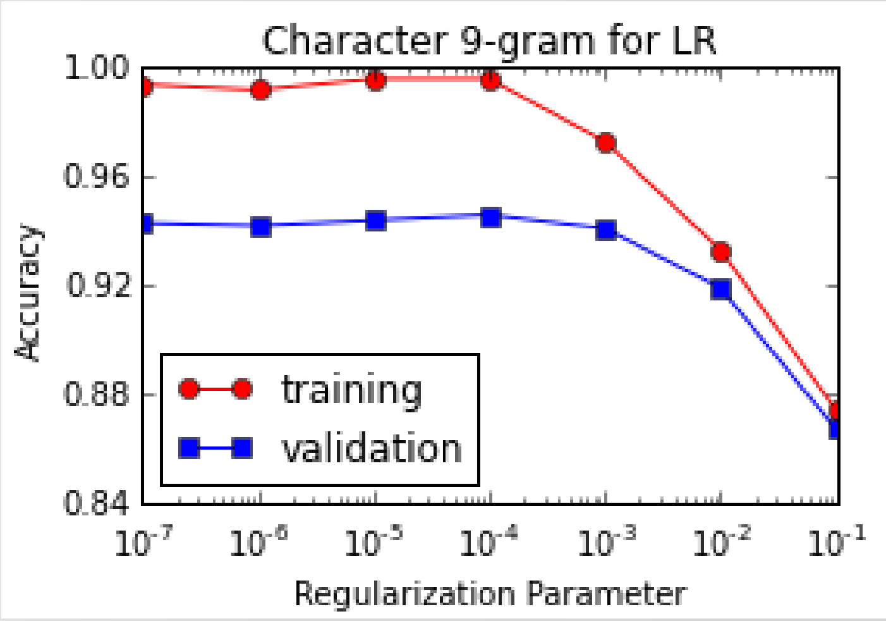
Clusters of less similar languages

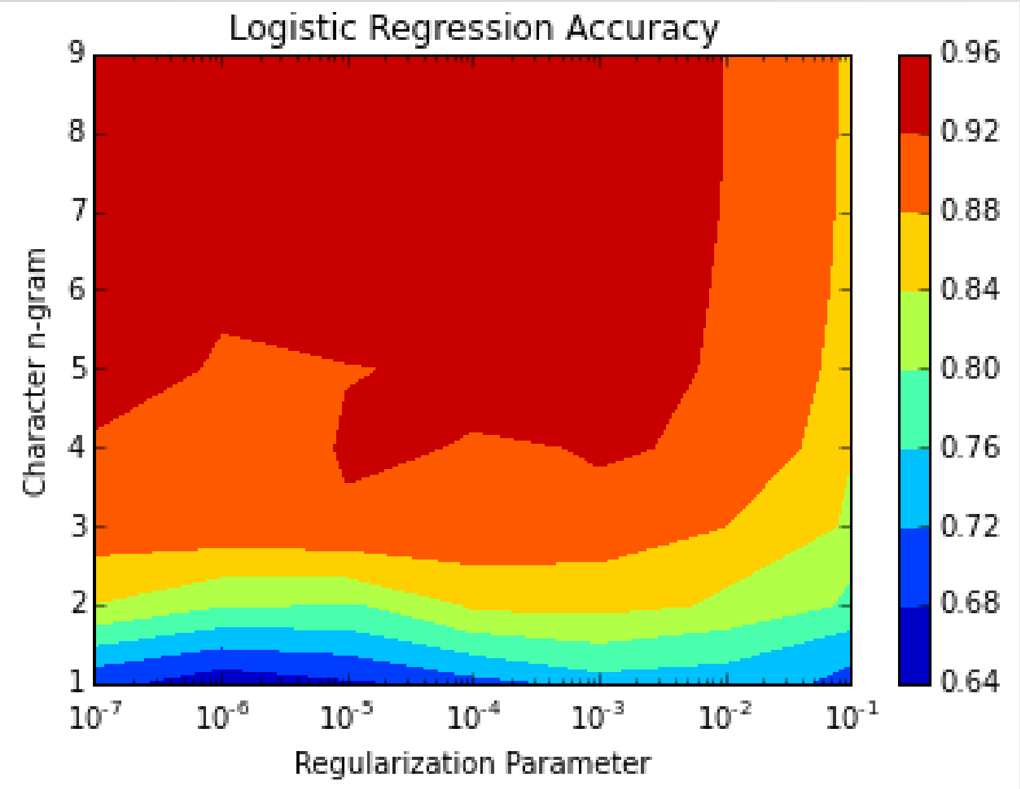
More interesting visualizations are available at: http://SeeYourLanguage.info

Baseline Results



- Multinomial Naive Bayes and Logistic Regression were used to setup a baseline score.
- Character n-grams delimited at word boundaries worked slightly better than word n-grams but at least 6 character n-grams were necessary to obtain good results.
- As accuracy of training set reached near 100% we incorporated more training data in the hard to distinguish group (bs, hr, sr) but that did not improve the performance.

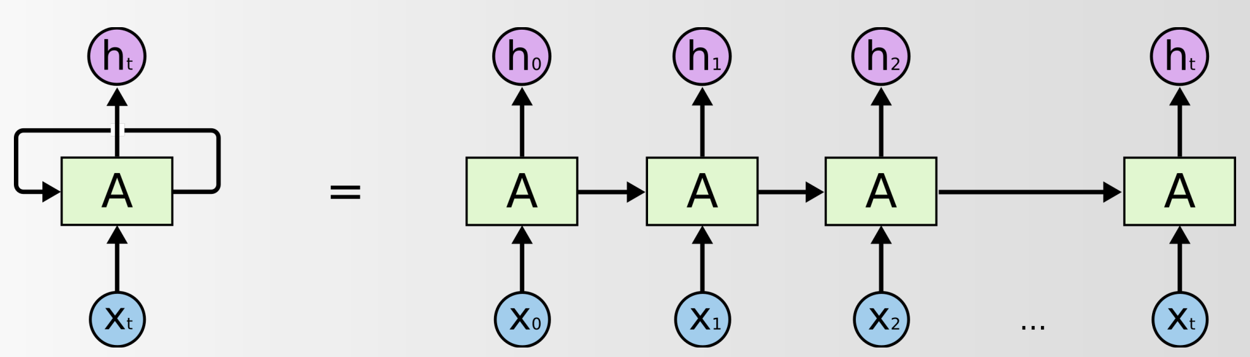




Recurrent Neural Networks (RNN)

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed cycle. This creates an internal state of the network which allows it to exhibit dynamic temporal behavior.

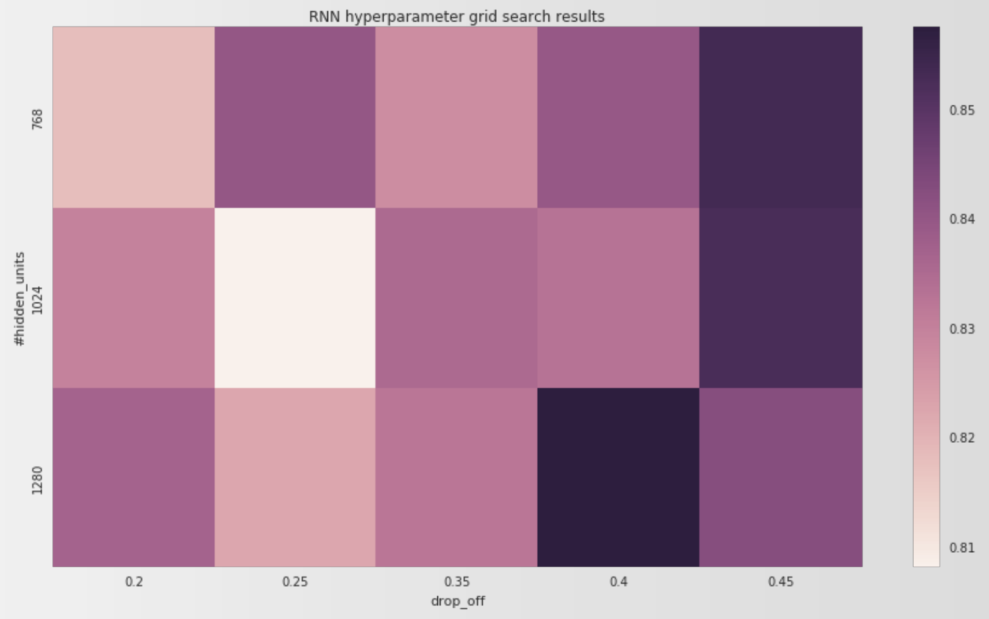
Since RNNs can use their internal memory to process arbitrary sequences of inputs, they hold great promise for learning general sequences, and have applications for text analysis, handwriting recognition and even machine translation.



RNN training

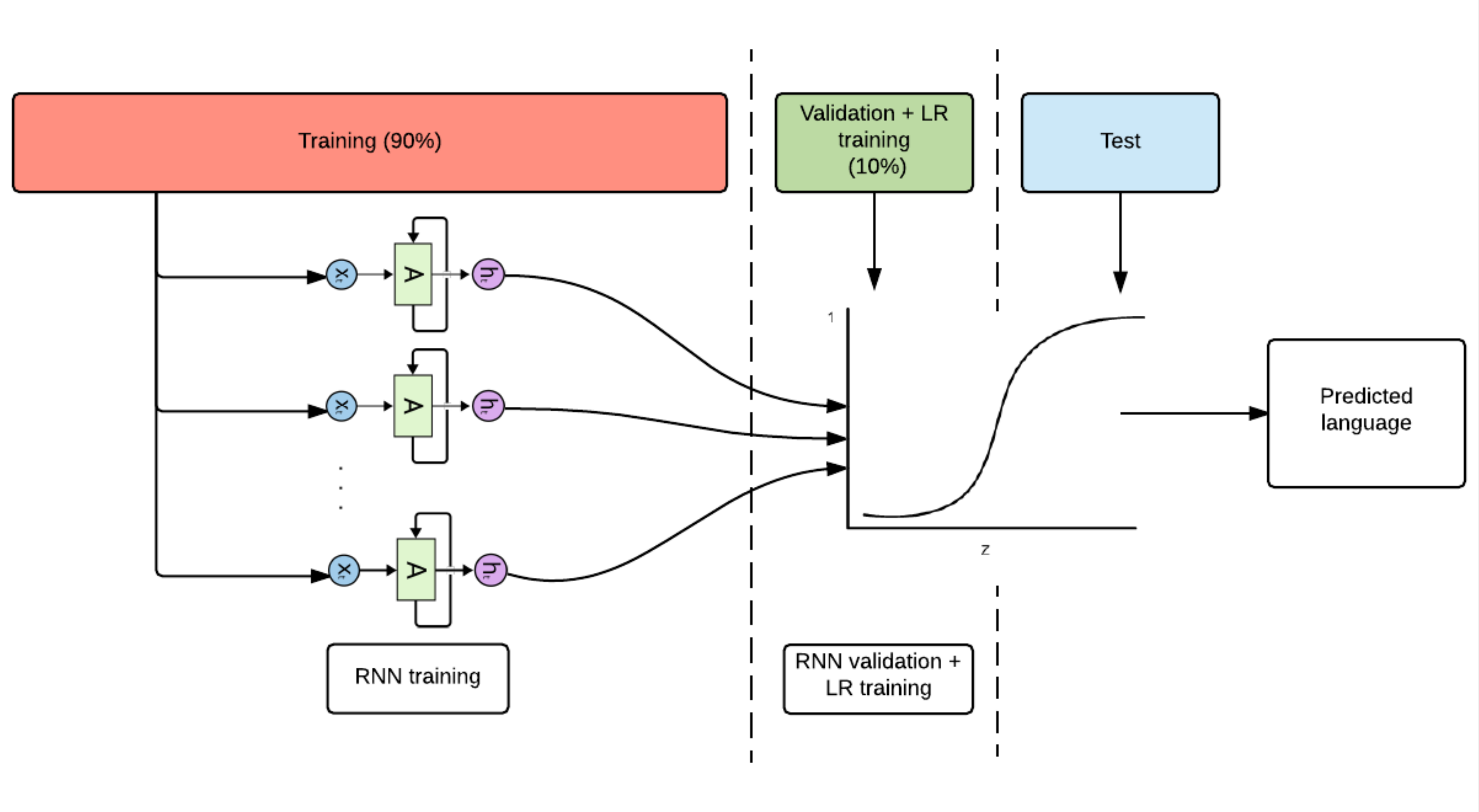
Hyper-parameters tuning

- Involved finding the best values for hidden layer size, drop off probability and number of training epochs.
- Tuning was performed in 2 steps:
  a. Varied one hyper-parameter while keeping others constant.
  b. Performed grid search on the best values found in the step above.



The training data was divided into training and validation sets. Multiple RNNs, each on a different feature set, were trained on the training data set.

An ensemble of the RNN models was created using a Logistic Regression method. It was trained on the validation data set to obtain the best overall accuracy on the test data set.



Made possible by

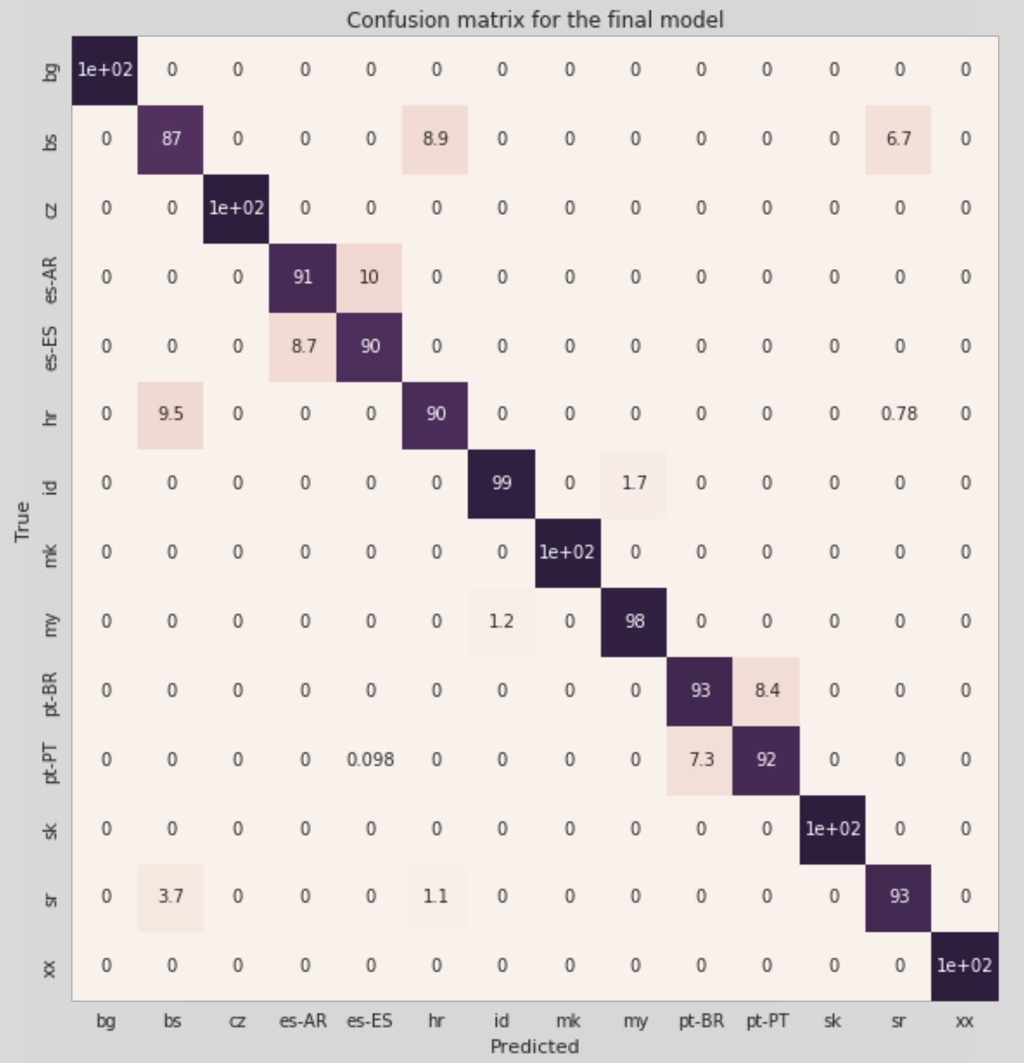
aws educate   Microsoft Azure

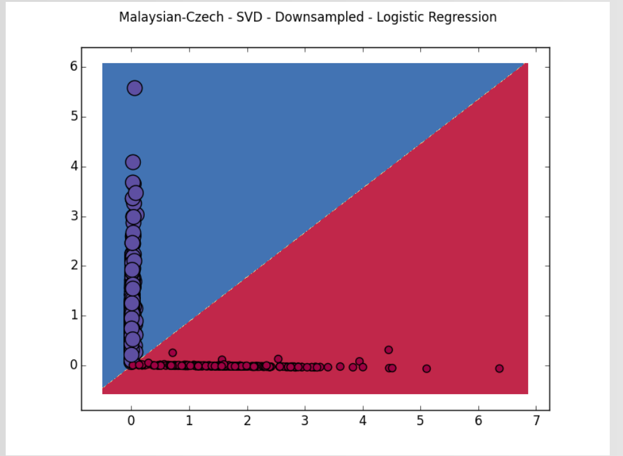
Results

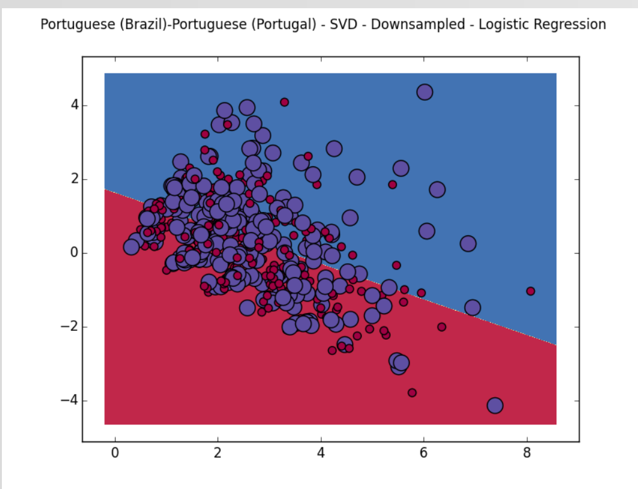
Model	Accuracy on Test Data
MNB (word bigram)	0.9359
MNB (char 8-gram)	0.9409
LR (char 9-gram)	0.9425

Model	Accuracy on Validation Data	Accuracy on Test Data
RNN (char 2-gram)	0.9200	0.9213
RNN (char 3-gram)	0.9328	0.9338
RNN (char 4-gram)	0.9377	0.9347
RNN (char 5-gram)	0.9347	0.9316
RNN (word unigram)	0.9351	0.9330
Ensemble of RNNs (LR)	0.9533	0.9512

Error analysis

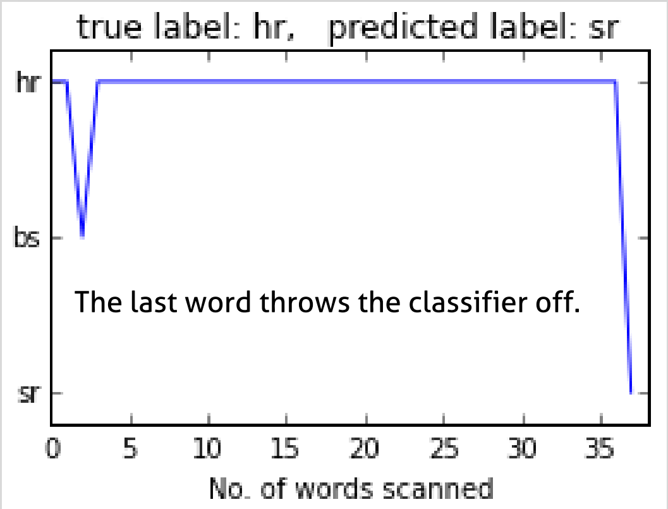




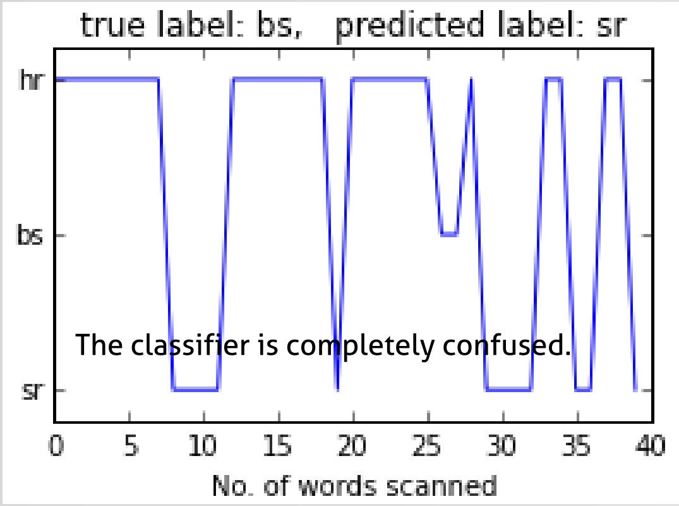


Failure modes in classifying the (bs, hr, sr) language group

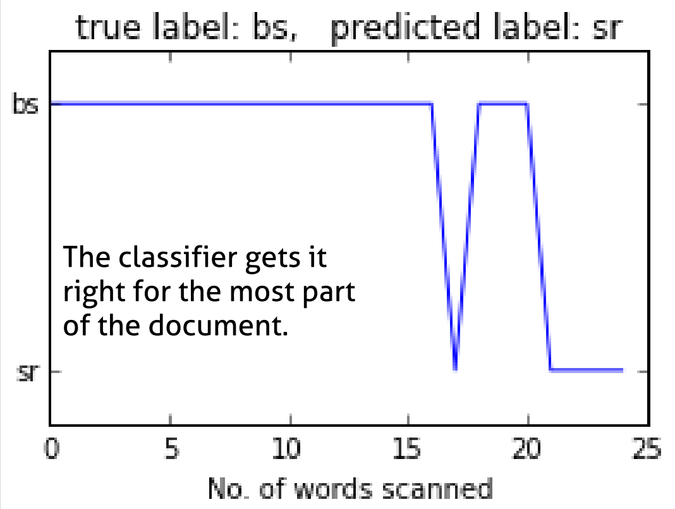
The LR classifier performance was re-evaluated for the (bs, hr, cr) language group by feeding an increasing fraction of the document to the classifier.



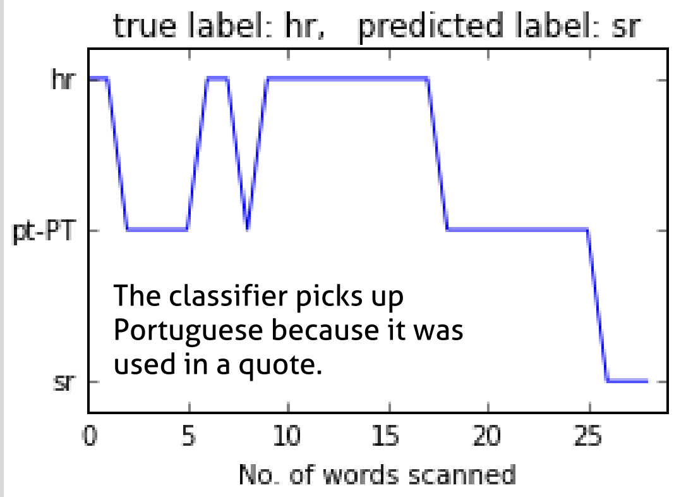
The last word throws the classifier off.



The classifier is completely confused.



The classifier gets it right for the most part of the document.



The classifier picks up Portuguese because it was used in a quote.

- Making predictions from fractions of a document and taking a median prediction could possibly improve the classification.
- Removing quotation from documents will improve the robustness of the classifier.