An Investigation of Multiple Language Classifiers COMP 551 - Applied Machine Learning

Thomas Page
Department of Mechanical Engineering
McGill University
Montreal, Canada
thomas.page@mail.mcgill.ca
260771672

David Tamrazov
School of Computer Science
McGill University
Montreal, Canada
david.tamrazov@mail.mcgill.ca
260561439

Alexander Wong
School of Computer Science
McGill University
Montreal, Canada
alexander.wong4@mail.mcgill.ca
260602944

Notorious Language Classifiers
https://github.com/a22wong/notorious-language-classifier

October 23, 2017

I. Introduction

As a result of the first project, corpora in various languages are available to act as training sets for many language related machine learning tasks. The goal of this project is to create classifiers that can determine the language of a sample of text from corpora information developed in the first project. Based on a review of the methods discussed in COMP 551, we construct four classifiers: Naive Bayes, Linear Discriminant Analysis (LDA), Decision Trees, and Extremely Randomized Decision Trees.

II. Related Work

asdf

III. Problem Representation

The compiled training dataset contains 276516 labeled language samples from five different languages: Slovak, French, Spanish, German, and Polish. The test set contains 118507 unlabeled samples of zero to ten individual characters. This presented the challenge of classifying seemingly nonsensical data; a common task in practical machine learning.

Fig. 1 represents the amount each language is present in the training set.

IV. Testing and Validation

The most effective way to represent the performance metrics of each individual classifier is through their respective confusion matrix. We chose to evaluate our classifiers using a normalized confusion matrix as it highlights important outputs like the true positives and exactly how each class is correctly or incorrectly classified.

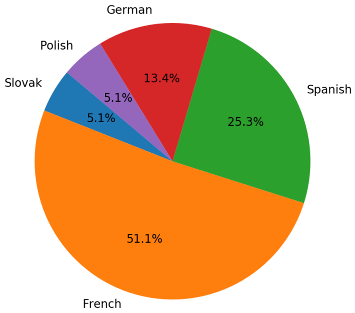

Fig. 1. Training set language distribution.

A. Naive Bayes

Fig. 2 shows the confusion matrix for the Naive Bayes classifier on the training set. Slovak, French, German, and Polish are all classified 90% correctly. Spanish is the only classifier with a significant difference at 77% correct. It is clear from the matrix that most of the incorrect Spanish classifications are classified at French. This is due in part to both languages sharing many of the same symbols, and because French accounts for over double the training set samples, a Spanish sample is more likely to be classified as French than vice versa. This hypothesis is supported by the confusion matrix, as 16% of Spanish samples are classified as French, but only 7% of French samples are classified as Spanish.

B. Decision Trees

C. Extremely Randomized Trees

V. Discussion

asdf

VI. Statement of Contributions

Thomas was responsible for fully implementing the Naive Bayes classifier. He also wrote the outline of the report and any sections pertaining to Naive Bayes.

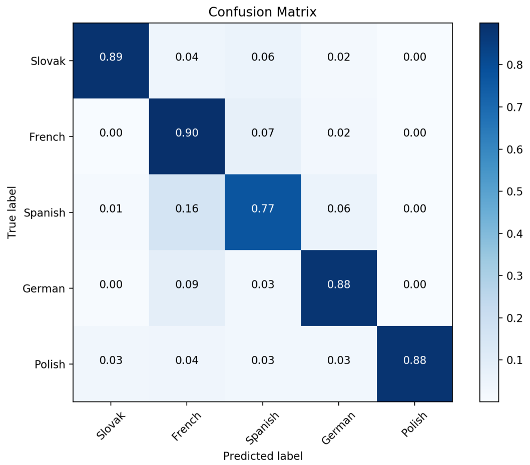

Fig. 2. Naive Bayes confusion matrix.

Alex was responsible for fully implementing the Decision Trees classifier. For the report, he wrote any sections pertaining to Decision Trees.

David was responsible for the LDA and Extremely Randomized Trees classifiers. For the report, he wrote any sections pertaining to LDA or Extremely Randomized Trees.

We hereby state that all the work presented in this report is that of the authors.

References

[1] H. Kopka and P. W. Daly, *A Guide to* $\LaTeX$, 3rd ed. Harlow, England: Addison-Wesley, 1999.