

# 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.

A. *Naive Bayes*

B. *Decision Trees*

C. *Extremely Randomized Trees*

### V. DISCUSSION

asdf

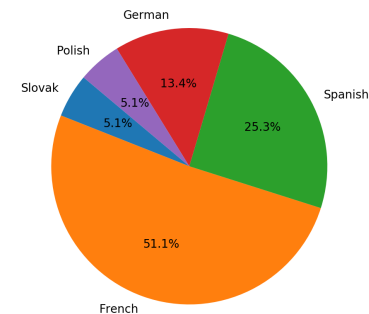


Fig. 1. Training set language distribution.

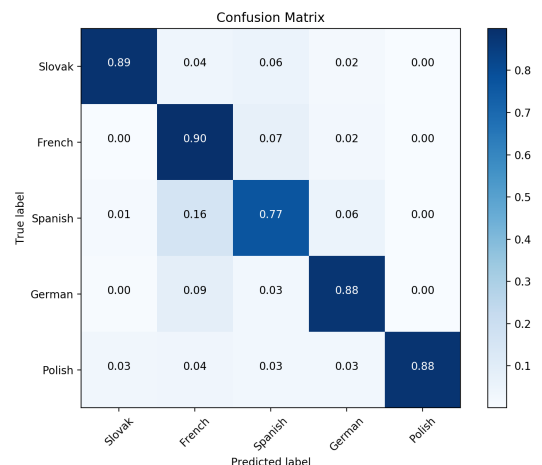


Fig. 2. Naive Bayes confusion matrix.

### 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.

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 L<sup>A</sup>T<sub>E</sub>X*, 3rd ed. Harlow, England: Addison-Wesley, 1999.