



# Similar Language Detection

## *Project Documentation*

---

### Team Name- Bolts

Nishanth Reddy

Sravya Varma

Poojitha Nandigam

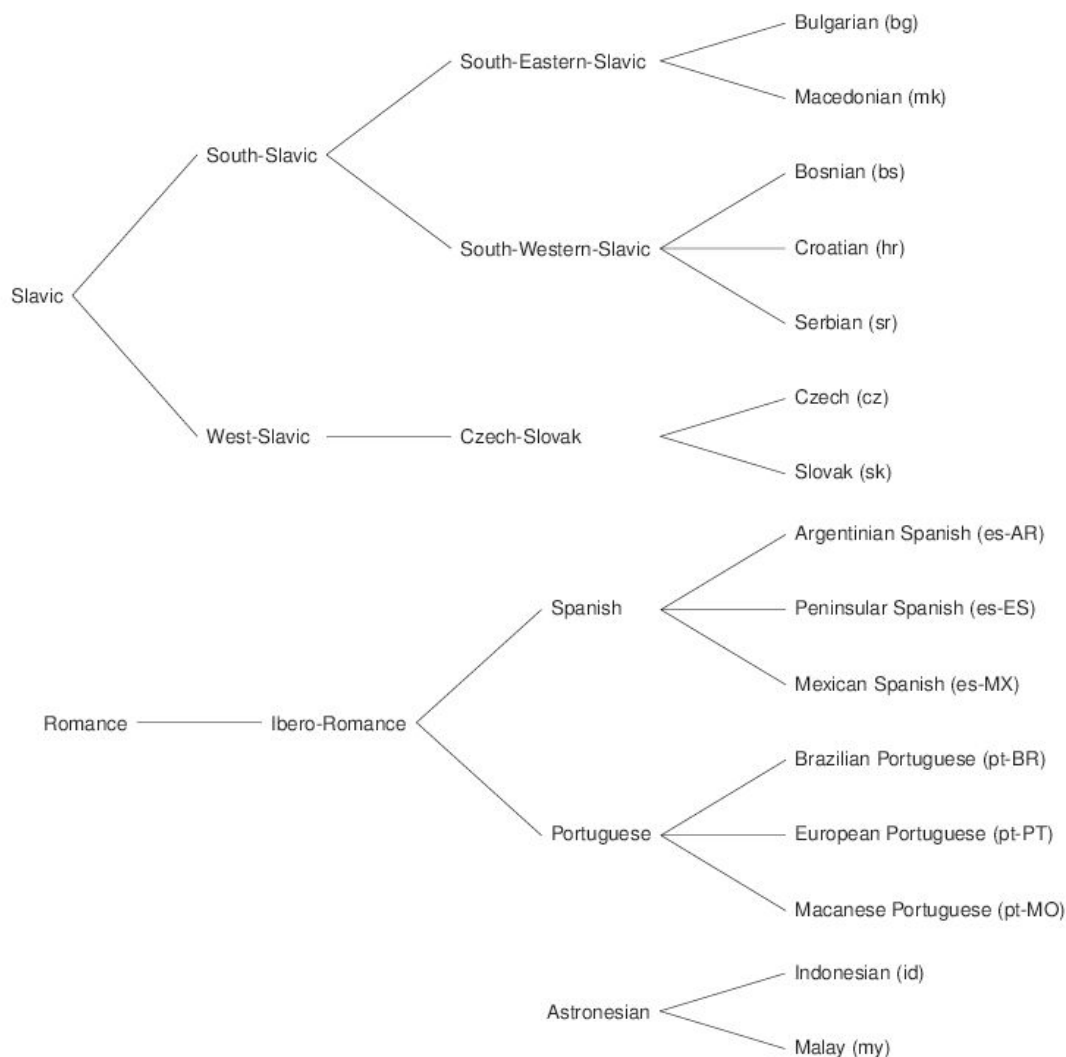
### Project Definition

The discrimination of similar languages can be defined as the subtask of the Language identification problem. Language Identification is a fundamental task in the area of natural language processing.

Unlike well-separated languages, multilingualism, dialects of language can seriously degrade the quality of Language Identification. Discrimination of Similar Languages, noisy data, non well-formatted text, short sentences, mixed language are other examples of challenging problems in this field.

- 
- Datasets from the DSL-Shared Task 2015 includes 13 different languages grouped into 6 categories
  - South-Eastern Slavic (ses): Bulgarian (bg), Macedonian (mk)
  - South-Western Slavic (sws): Bosnian (bs), Croatian (hr), Serbian (sr)
  - West Slavic (ws): Czech (cz), Slovak (sk)
  - Spanish(es): Argentine Spanish (esar), Peninsular Spanish (eses)

- Portuguese (pt): Brazilian Portuguese (ptbr), European Portuguese (ptpt)
- Austronesian (aus): Indonesian (id), Malay (my)

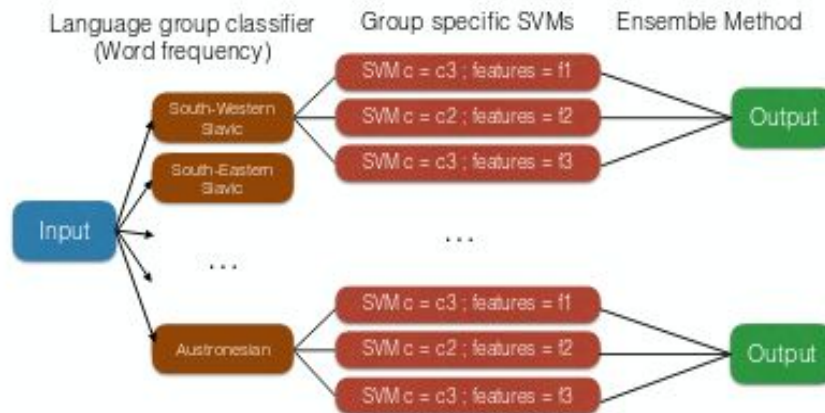


We will get the corpus from

<https://github.com/Simdiva/DSL-Task/tree/master/data/DSLCC-v2.0>

The corpus contains 20,000 instances per language (18,000 training + 2,000 development). Each instance is an excerpt extracted from journalistic texts containing 20 to 100 tokens and tagged with the country of origin of the text.

Language identification task isn't trivial for above group of languages.



## Standard tools and techniques required for the project

### Methods

The hierarchical method employs a simple word-frequency method that first identifies the group a test belongs to. It then uses an ensemble of SVM models to determine the final output.

#### 1. The Word-Frequency Method :

The first block of our pipeline is a classifier to distinguish the language group of an input. From the training set, for each language

$l \in 1 \dots 13$ , we define  $x^l \in R^{1000}$  s.t.  $x_i^l$  is the frequency of the  $i$ th common word of the  $l$ th language.

Given an unclassified sentence  $s$ , we define  $x^l(s), l = 1 \dots 13$  s.t.  $x_i^l(s) = 1$  if the  $i$ th word of the  $l$ th language is present in the sentence. The classifier is then  $h(s) := \operatorname{argmax}_{l \in 1 \dots 13} \langle x^l(s), x^l \rangle$ . To obtain the group, we then just have to return the group the language  $h(s)$  is part of.

#### 2. Multiclass SVMs :

The SVM models are trained with examples within the language group and can have different number and combination of features across groups.

Let's replace ourselves in the context of the course but with  $k$  classes this time.

Let  $(x^{(i)}, y^{(i)}), i = 1 \dots m$  with  $x^{(i)} \in R^n$  but  $y^{(i)} \in \{1 \dots k\}$ . Crammer and Singer (2002) proposed the following multi-class approach by solving the following optimization problem:

$$\begin{aligned} & \underset{w_l, \xi_i, l=1 \dots k, i=1 \dots m}{\text{minimize}} && \frac{1}{2} \sum_{l=1}^k w_l^T w_l + C \sum_{i=1}^m \xi_i \\ & \text{subject to} && w_{y^{(i)}}^T x^{(i)} - w_l^T x^{(i)} \leq \delta_{i,l} - \xi_i, \quad i = 1, \dots, m. \end{aligned}$$

Where  $\delta_{i,l} = 1$  if  $i = l$ ,  $\delta_{i,l} = 0$  if  $i \neq l$ .

With the following decision function:

$$h_{w_1, \dots, w_k}(x) = \underset{l \in \{1, \dots, k\}}{\operatorname{argmax}} w_l^T x$$

We can derive the dual problem:

$$\begin{aligned} & \underset{\alpha_i^l, l=1 \dots k, i=1 \dots m}{\text{minimize}} && \sum_{1 \leq i, j \leq m, 1 \leq l \leq k} \alpha_i^l \alpha_j^l \langle x^{(i)}, x^{(j)} \rangle + \sum_{1 \leq i \leq m, 1 \leq l \leq k} \delta_{i,l} \alpha_i^l \\ & \text{subject to} && \sum_{l=1}^k \alpha_i^l, \quad i = 1, \dots, m. \\ & && \alpha_i^l \leq \delta_{i,l} C, \forall i \in \{1, \dots, m\}, \forall l \in \{1, \dots, k\} \end{aligned}$$

Where  $w_l = \sum_{i=1}^m \alpha_i^l x^{(i)}, l = 1 \dots k$

### 3. Feature Scoring Using tf-idf:

We defined the tf-idf score of a feature  $t$  derived from training set  $D$  to be

$$\text{tfidf}(t, D) = \text{tf}(t, D) \times \log \frac{N}{\text{df}(t, D)} \quad (1)$$

where  $\text{tf}(t, D)$  is the total number of times  $t$  appeared in the training set,  $\text{df}(t, D)$  the number of examples that contain  $t$ , and  $N$  the number of training examples. Since  $N/\text{df}(t, D)$  is always greater or equal to 1, this definition of tf-idf is always non-negative. It is zero if the feature appears in every training example; intuitively, this means the feature is common to all languages in the training set and therefore cannot help distinguish them from one another.

### 4. Ensemble Methods :

For each language groups, we trained pSVM using different hyperparameters (features and C constant). Each one of these SVM i, outputs k weight vectors.

We have  $(w_{i,l})_{1 \leq i \leq p, 1 \leq l \leq k}$  weight vectors at our disposition.

After experimenting with several ensemble methods (majority vote, boosting, and mean confidence) we settled on mean confidence, which can be written as follows:

$$h_{(w_{i,l})_{1 \leq i \leq p, 1 \leq l \leq k}}(x) = \arg \max_{1 \leq l \leq k} \left\langle \sum_{t=1}^p w_{t,l}, x \right\rangle$$

Classifiers in Ensemble SVM			
SVM	Features	# of features	C
1	word 1-gram, 2-gram	25000	0.1
2	word 1-gram, 2-gram, char 3-gram	29791	0.05
3	word 1-gram, 2-gram, char 4-gram	44453	0.01
4	word 1-gram, 2-gram, char 5-gram	63826	0.005
5	word 1-gram, 2-gram, char 6-gram	53805	0.001


## Scope

DSL Shared Task of 2015 made available the required corpus. The corpus contains 20,000 instances per language (18,000 training + 2,000 development). Each instance is an excerpt extracted from journalistic texts containing 20 to 100 tokens and tagged with the country of origin of the text.

Now, word frequency method is applied to group the data into 6.

To differentiate between the similar languages present in the group, Multiclass SVM is used.

To avoid curse of dimensionality, tf-idf scoring is used for feature scoring, and only high ranked features are considered for ensemble.



Ensemble combines various SVMs with different hyperparameters. Ensemble method used is Mean Confidence.