# Machine Learning Project Appendix

Kristóf Balázs, Stefanos Kypritidis

## List of Figures



## List of Tables

# Appendix

## Plots

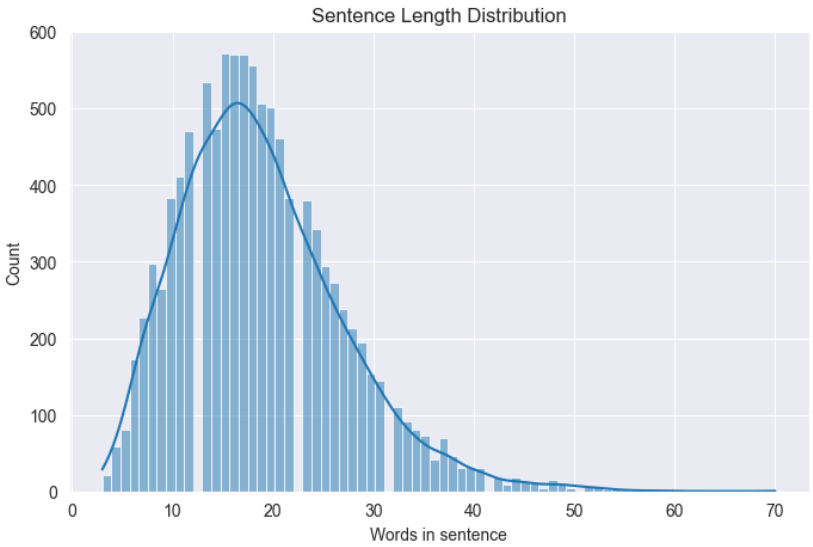


Figure 1: Sentence Length Distribution

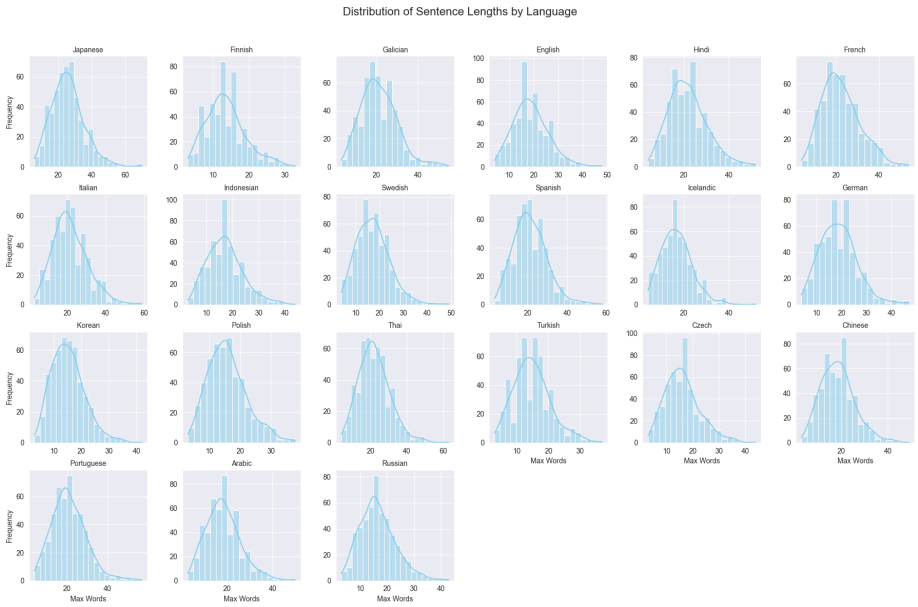


Figure 2: Sentence Length Distribution by Language

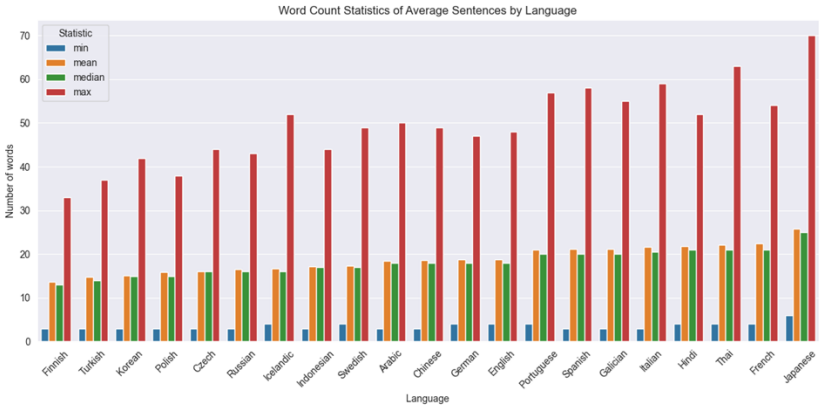


Figure 3: Sentence Word Count Statistics by Language

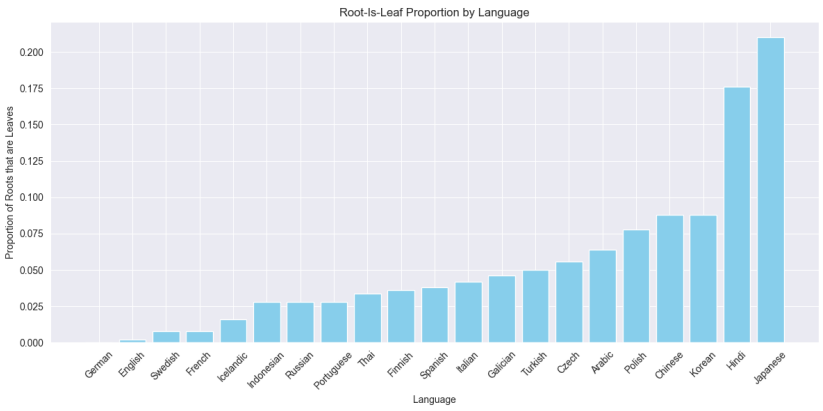


Figure 4: Proportion of Root Nodes That Are Also Leaf Nodes by Language

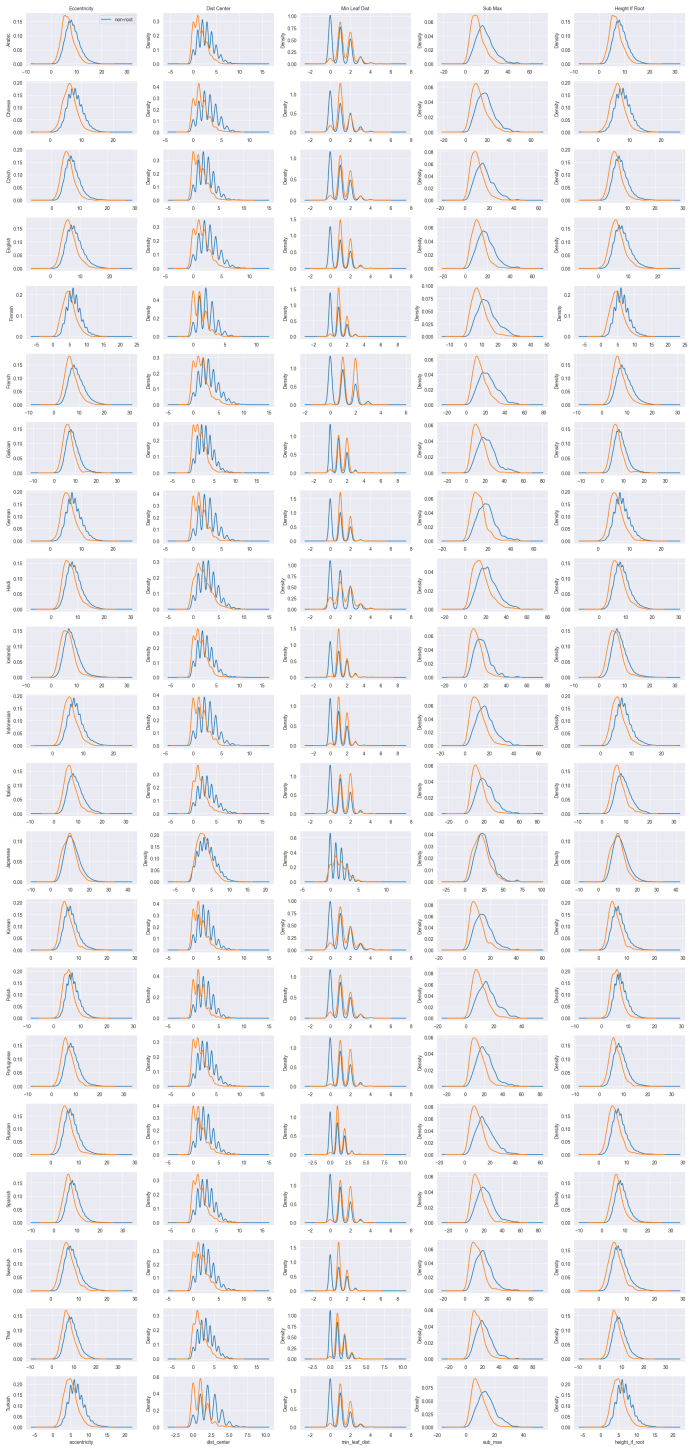


Figure 5: Distribution of Node-Level Features by Root Status and Language

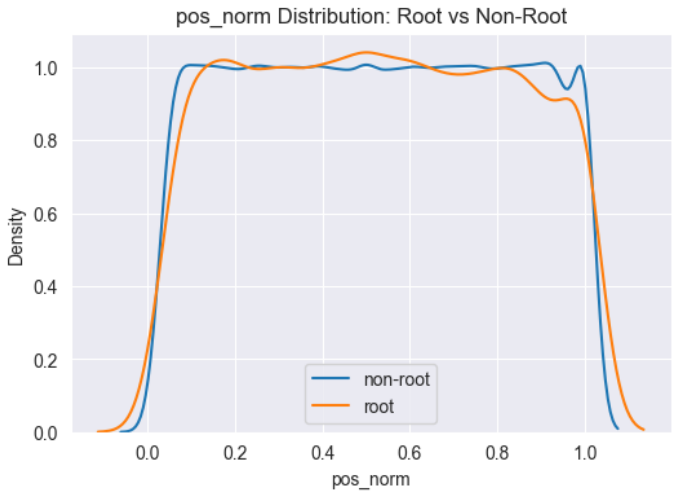


Figure 6: Density Distribution of normalized position for root and non-root

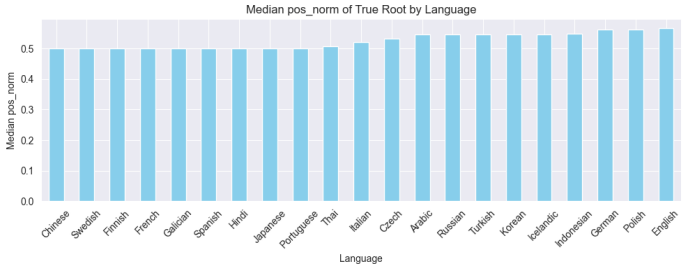


Figure 7: Median Normalized Root Position by Language

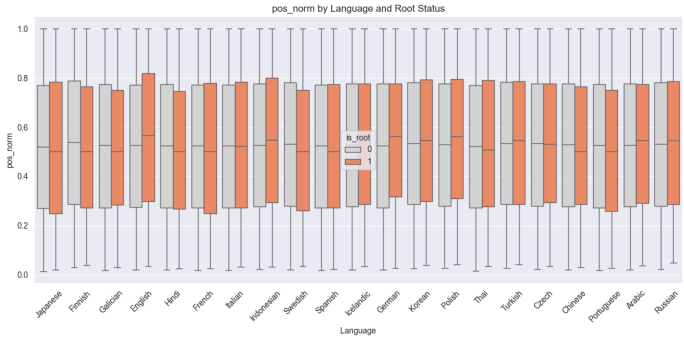


Figure 8: Boxplots of normalized position for Roots vs Non-Roots Across Languages

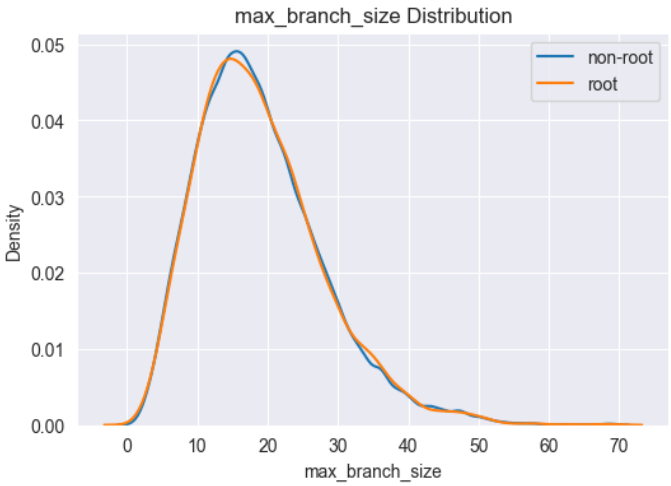


Figure 9: Probability density of max_branch_size for root and non-root nodes

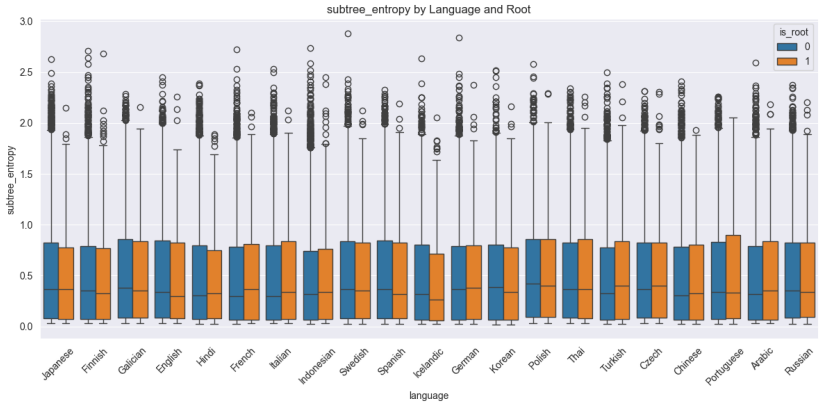


Figure 10: Boxplots of subtree_entropy for each language and for root and non root nodes

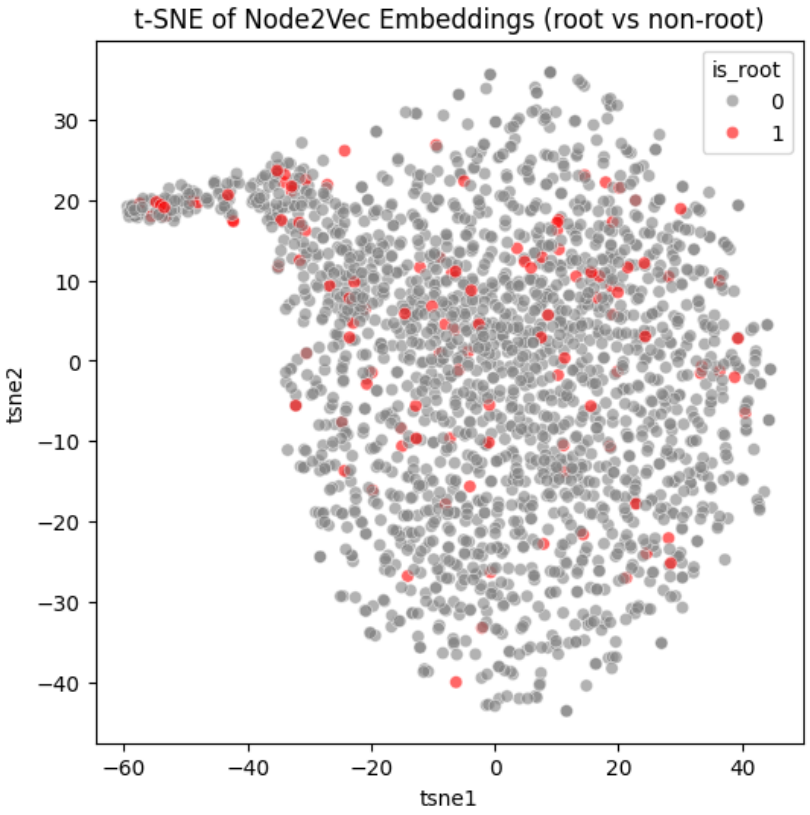


Figure 11: t-SNE visualization on a sample of embeddings for "root" and "non-root" nodes in 2D space

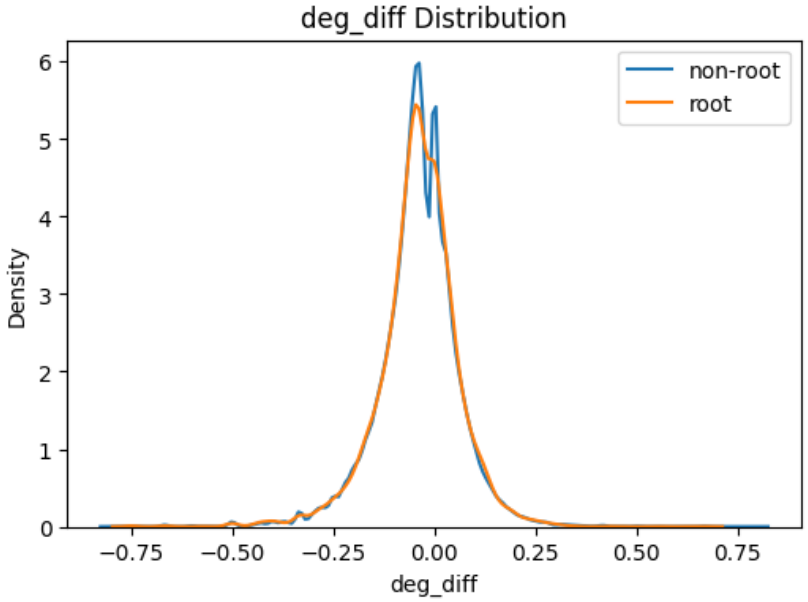


Figure 12: Kernel Density Estimate of degree difference variable between root and non-root nodes

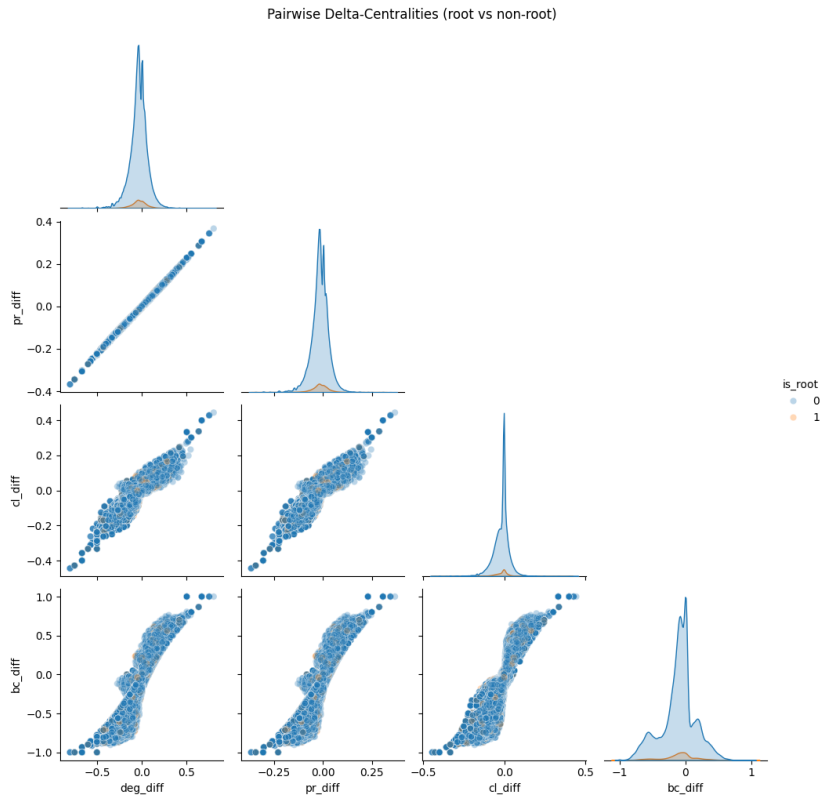


Figure 13: Pairwise relationships between centrality deltas (differences between a node's centrality and the average of its neighbors) for root vs non-root nodes

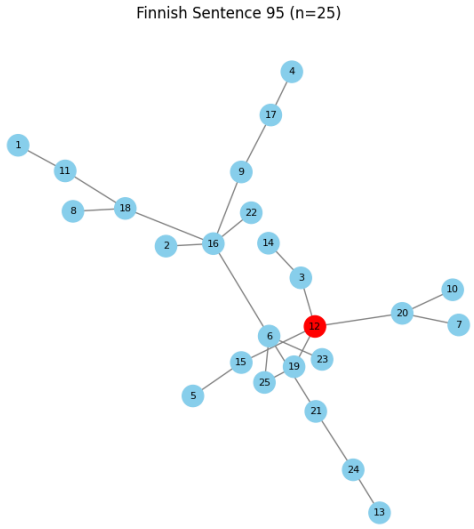


Figure 14: Tree visualization example

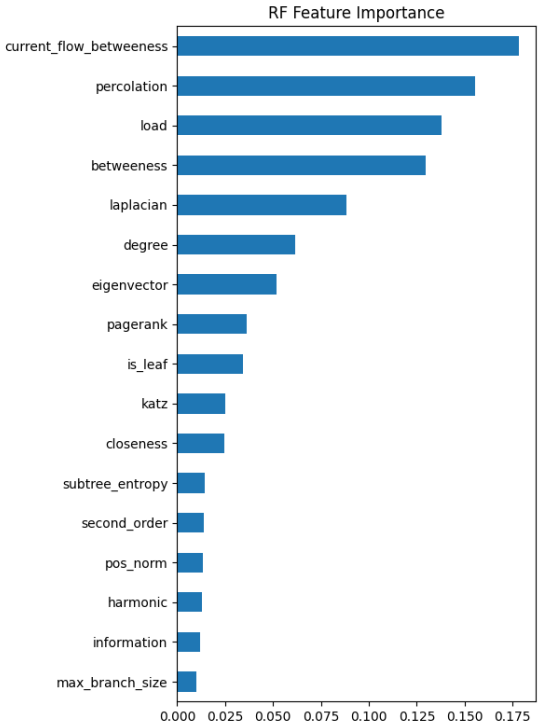


Figure 15: Random Forest feature importance

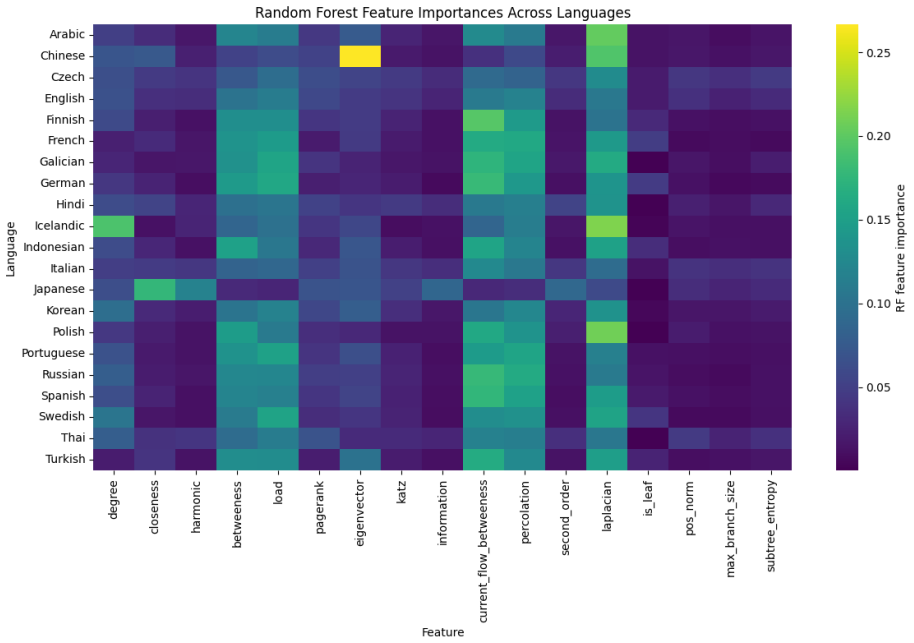


Figure 16: Random Forest feature importances by language

Table 1: Best Logistic Regression Parameters per Language

| **Language** | **Val. Score** | **PCA Comp.** | **Warm Start** | **Solver** | **Penalty** | **Max Iter** | **L1 Ratio** | **Intercept** | **C** |
|---|---|---|---|---|---|---|---|---|---|
| Japanese | 0.09 | 13 | True | lbfgs | l2 | 100 | 0 | True | 1 |
| Finnish | 0.38 | 13 | True | lbfgs | l2 | 500 | 0 | True | 10 |
| Galician | 0.31 | 13 | True | lbfgs | l2 | 500 | 0 | True | 100 |
| English | 0.28 | 9 | True | liblinear | l1 | 100 | 0 | True | 100 |
| Hindi | 0.26 | 9 | True | liblinear | l1 | 100 | 0 | True | 100 |
| French | 0.31 | 9 | True | lbfgs | None | 100 | 0 | True | 0.1 |
| Italian | 0.26 | 13 | True | lbfgs | None | 500 | 0 | True | 0.1 |
| Indonesian | 0.25 | 9 | True | lbfgs | None | 500 | 0 | False | 0.1 |
| Swedish | 0.30 | 11 | True | liblinear | l1 | 100 | 0 | True | 10 |
| Spanish | 0.30 | 9 | True | newton-cg | None | 100 | 0 | False | 0.1 |
| Icelandic | 0.34 | 7 | True | liblinear | l2 | 100 | 0 | False | 1 |
| German | 0.32 | 9 | True | lbfgs | None | 500 | 0 | True | 0.1 |
| Korean | 0.34 | 9 | True | newton-cholesky | l2 | 100 | 0 | False | 0.1 |
| Polish | 0.35 | 7 | True | newton-cg | l2 | 100 | 0 | True | 0.1 |
| Thai | 0.30 | 9 | True | lbfgs | l2 | 100 | 0 | True | 0.1 |
| Turkish | 0.34 | 11 | True | lbfgs | None | 500 | 0 | True | 0.1 |
| Czech | 0.34 | 9 | True | newton-cholesky | l2 | 100 | 0 | True | 100 |
| Chinese | 0.26 | 9 | True | lbfgs | None | 1000 | 0 | True | 0.1 |
| Portuguese | 0.31 | 9 | True | newton-cholesky | None | 100 | 0 | False | 0.1 |
| Arabic | 0.31 | 7 | True | lbfgs | l2 | 100 | 0 | True | 1 |
| Russian | 0.30 | 7 | True | newton-cholesky | l2 | 100 | 0 | True | 0.1 |

Table 2: Best Random Forest Parameters per Language

| Language | est_oob_root_sc | val_root_sc | param_n_estimators | param_max_depth | param_min_samples_leaf | param_max_features |
|---|---|---|---|---|---|---|
| Arabic | 0.298 | 0.37 | 100 | 5 | 10 | sqrt |
| Chinese | 0.260 | 0.30 | 100 | 5 | 10 | sqrt |
| Czech | 0.298 | 0.67 | 100 | 20 | 10 | sqrt |
| English | 0.278 | 0.60 | 200 | 10 | 10 | 0.5 |
| Finnish | 0.334 | 0.40 | 200 | 5 | 5 | sqrt |
| French | 0.256 | 0.32 | 100 | 5 | 20 | sqrt |
| Galician | 0.294 | 0.42 | 200 | 5 | 20 | 0.5 |
| German | 0.288 | 0.38 | 100 | 5 | 20 | sqrt |
| Hindi | 0.212 | 0.30 | 200 | 5 | 20 | sqrt |
| Icelandic | 0.334 | 0.37 | 100 | 5 | 20 | 0.5 |
| Indonesian | 0.298 | 0.44 | 300 | 10 | 20 | sqrt |
| Italian | 0.260 | 0.67 | 200 | 20 | 10 | sqrt |
| Japanese | 0.076 | 0.28 | 100 | 5 | 20 | sqrt |
| Korean | 0.294 | 0.36 | 300 | 5 | 20 | sqrt |
| Polish | 0.302 | 0.40 | 300 | 5 | 20 | 0.5 |
| Portuguese | 0.300 | 0.31 | 300 | 5 | 10 | sqrt |
| Russian | 0.342 | 0.38 | 200 | 5 | 10 | sqrt |
| Spanish | 0.314 | 0.36 | 200 | 5 | 20 | sqrt |
| Swedish | 0.316 | 0.40 | 100 | 5 | 20 | sqrt |
| Thai | 0.230 | 0.49 | 100 | 20 | 20 | 0.5 |
| Turkish | 0.294 | 0.42 | 100 | 5 | 5 | sqrt |

Table 3: Best XGBoost Parameters per Language

| Language | Val. Acc. | Round | Max Depth | Eta | Subsamp. | ColSamp ByTree | Gamma | MinCh Wght | Reg Alpha | Reg Lambda | Scale PosWt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Arabic | 0.570 | 1 | 4 | 0.01 | 1.0 | 1.0 | 0 | 1 | 0 | 100 | 17.48 |
| Chinese | 0.340 | 6 | 4 | 0.01 | 0.7 | 1.0 | 0 | 10 | 0.1 | 1 | 17.55 |
| Czech | 0.610 | 1 | 4 | 0.01 | 0.7 | 0.7 | 0 | 1 | 0 | 100 | 15.09 |
| English | 0.680 | 1 | 4 | 0.01 | 0.7 | 1.0 | 0 | 1 | 0 | 100 | 17.72 |
| Finnish | 0.570 | 7 | 4 | 0.10 | 0.7 | 1.0 | 0 | 1 | 0 | 100 | 12.58 |
| French | 0.480 | 1 | 4 | 0.01 | 1.0 | 0.7 | 0 | 1 | 0 | 100 | 21.41 |
| Galician | 0.550 | 1 | 4 | 0.01 | 0.7 | 1.0 | 0 | 1 | 0 | 100 | 20.10 |
| German | 0.650 | 1 | 4 | 0.01 | 0.7 | 1.0 | 0 | 1 | 0 | 100 | 17.70 |
| Hindi | 0.320 | 17 | 4 | 0.10 | 0.7 | 1.0 | 0 | 1 | 0 | 100 | 20.66 |
| Icelandic | 0.540 | 0 | 4 | 0.01 | 1.0 | 1.0 | 0 | 1 | 0 | 100 | 15.73 |
| Indonesian | 0.560 | 1 | 4 | 0.01 | 0.7 | 0.7 | 0 | 5 | 0 | 1 | 16.10 |
| Italian | 0.550 | 1 | 4 | 0.01 | 1.0 | 0.7 | 0 | 1 | 0 | 100 | 20.60 |
| Japanese | 0.150 | 1 | 4 | 0.05 | 0.7 | 0.7 | 0 | 5 | 0 | 1 | 24.77 |
| Korean | 0.370 | 1 | 4 | 0.10 | 0.7 | 1.0 | 0 | 10 | 0 | 100 | 14.02 |
| Polish | 0.650 | 1 | 4 | 0.01 | 1.0 | 0.7 | 0 | 1 | 0 | 100 | 14.83 |
| Portuguese | 0.510 | 1 | 4 | 0.01 | 0.7 | 0.7 | 0 | 1 | 0 | 100 | 19.82 |
| Russian | 0.670 | 1 | 4 | 0.01 | 0.7 | 1.0 | 0 | 1 | 0 | 100 | 15.41 |
| Spanish | 0.500 | 1 | 4 | 0.01 | 0.7 | 0.7 | 0 | 5 | 0 | 100 | 20.09 |
| Swedish | 0.650 | 8 | 6 | 0.05 | 0.7 | 0.7 | 0 | 10 | 0 | 100 | 16.15 |
| Thai | 0.620 | 1 | 6 | 0.01 | 1.0 | 0.7 | 0 | 5 | 0 | 100 | 21.08 |
| Turkish | 0.480 | 22 | 8 | 0.01 | 1.0 | 1.0 | 0 | 1 | 0 | 100 | 13.82 |