**Linear Support Vector Machines (SVMs)**

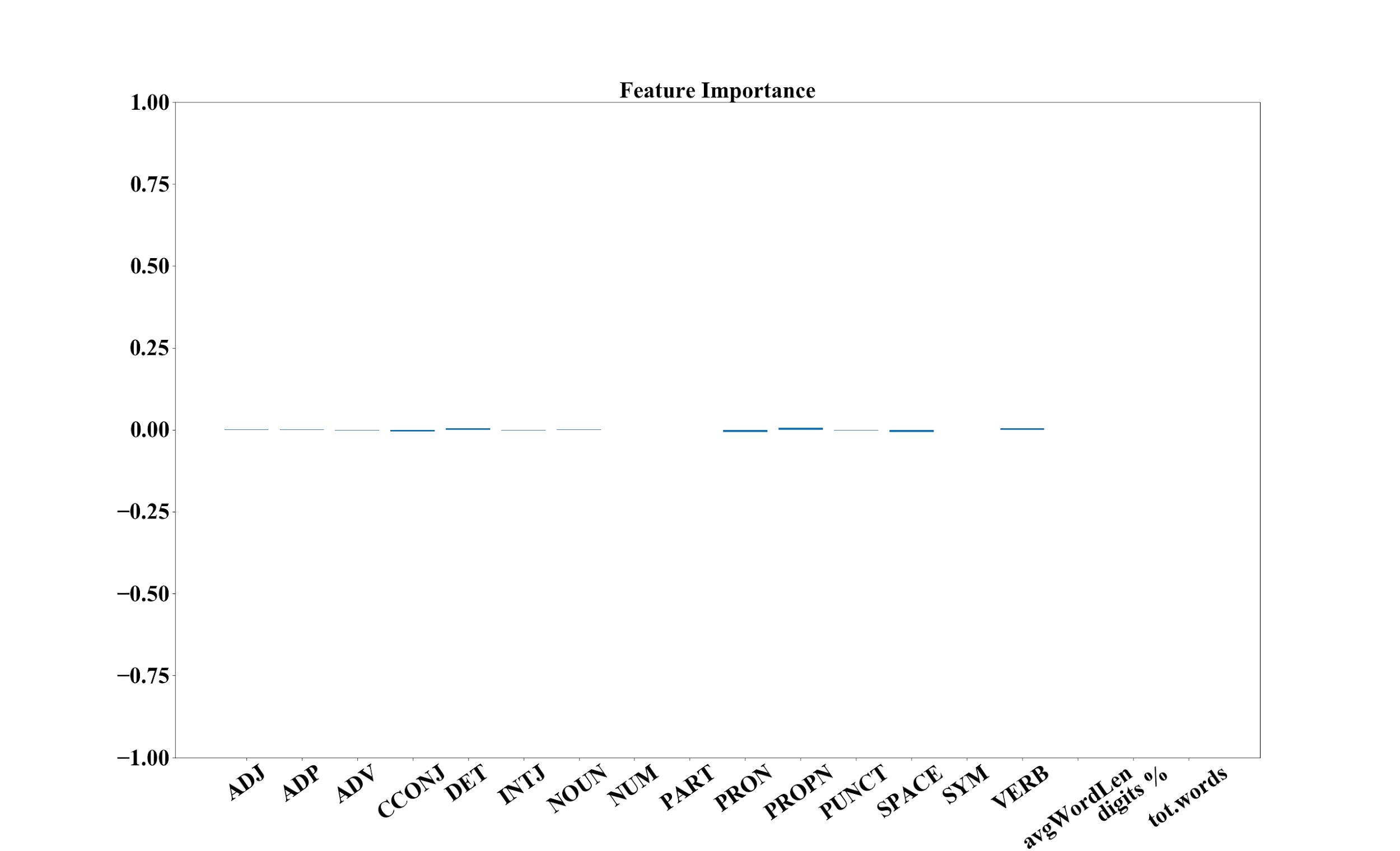
Support Vector Machines (SVM) is a supervised machine learning algorithm that can be used for both classification and regression but is more commonly used for classification. SVMs have different types of kernels that can be used for training. We experiment with a ‘linear’ kernel since it is the only one that can be interpreted.

In linear SVMs, the main idea is to find hyperplanes that are optimized to get a good divide between different classes. So for instance, if our dataset has two classes, SVM will give us one hyperplane. Similarly, if we have 3 classes, SVM will give us 3 different hyperplanes dividing all possible pairs of classes and so on.

We run two tests on our dataset using SVMs.

1. Check which feature has the highest impact on classification across the dataset.
2. How accurately does SVM classify documents?

For the first test, we train Linear SVM models for all the possible binary pairs of authors. Since we have 100 authors in total, we get 4950 different combinations. After training these models, we extract the weights assigned to each of the above-mentioned features and average them across all models. These averages would show us the features that played an important role in differentiating between authors across the whole dataset.



The higher the length of the bar (per feature), the more the average importance. This plot shows that none of the features stand out in terms of importance across the dataset.

For the second test, we perform author attribution considering different sets of authors i.e, 2, 5, 10, 50, 100. The authors are chosen randomly from the list of authors. The following table shows the results for 10-fold cross-validation for each experiment.

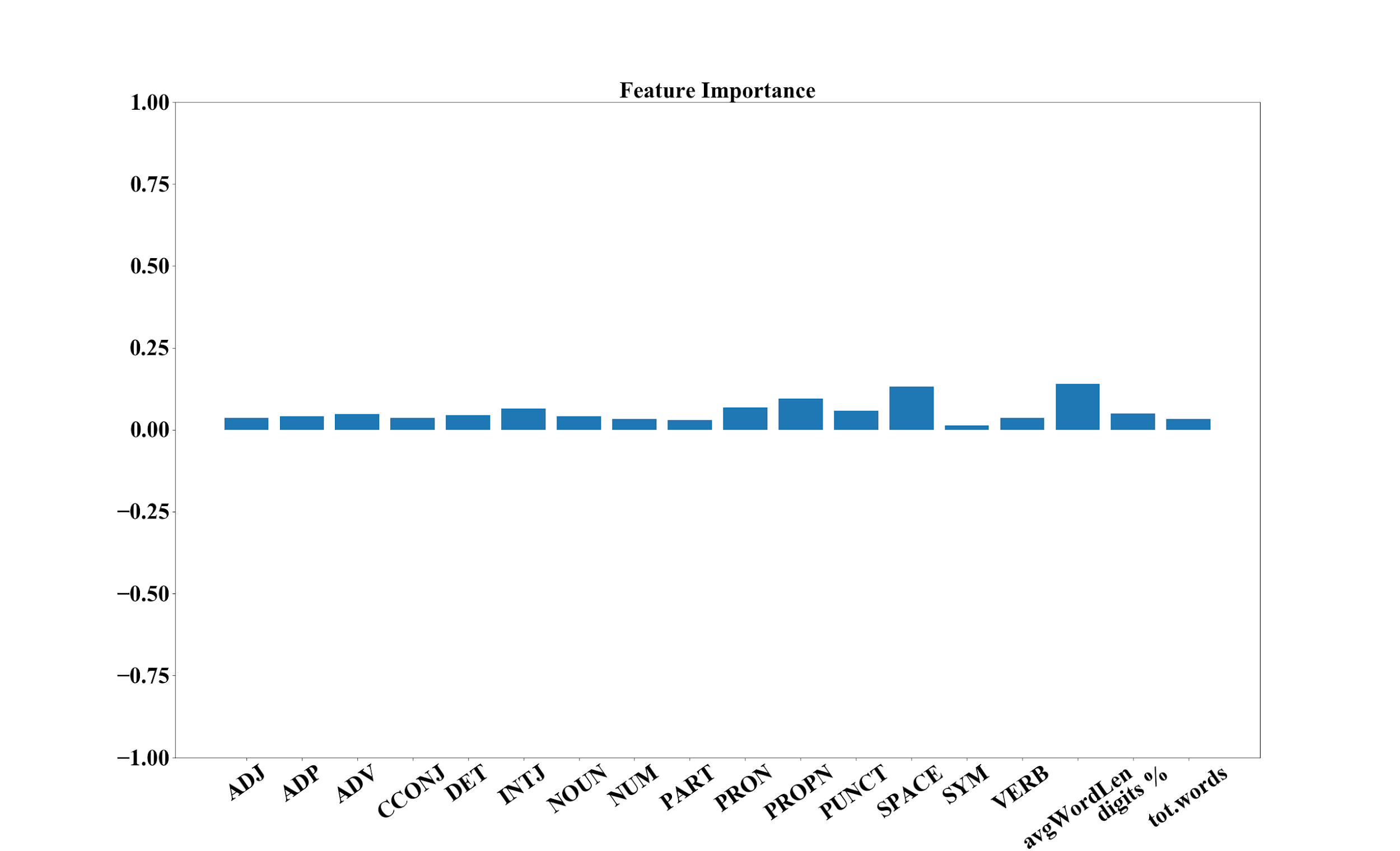
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **For SVM** | 2 Authors | 5 Authors | 10 Authors | 50 Authors | 100 Authors |
| 10-fold CV results | 0.90 | 0.38 | 0.05 | 0.02 | 0.001 |

This table shows an exponential decrease in accuracy when the number of authors is increased. This makes sense because as the number of authors increases, it gets hard for the model to differentiate between authors using this limited feature set.

**Random Forest Classifier (RFC)**

We also experimented with Random Forest Classifier (RFC). RFCs are an ensemble model that functions using a lot of different Decision trees (DTs). Same as SVM, we also ran two tests on RFC as well.

The following plot shows the result of the first test.



This plot shows some variation in terms of bar height. As per this plot, SPACE and average word length features are the most important ones for distinguishing between different authors. Other prominent features include PROPN, PRON, PUNCT and INTJ.

The following table shows the results for the second test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **For RFC** | 2 Authors | 5 Authors | 10 Authors | 50 Authors | 100 Authors |
| 10-fold CV results | 0.85 | 0.30 | 0.04 | 0.01 | 0.004 |

This also shows the same trend as SVMs. The accuracy decreases almost exponentially as the number of authors is increased.