Primary objective of this project is to develop classification models for identifying author of a blog based on a set of stylometric features/predictors. For this project we have used 10,000 blogs from 100 authors, where each author is identified by a unique id, which we call Author\_id. Although 555 stylometric features are available for each blog, we have chosen only 18 of the available features for all the statistical analysis and experiments in this project, due to the computational and algorithmic complexity associated with using the entire feature set. Fifteen of the eighteen predictors are part of speech tags (POS) and remaining three are the syntactic features. Out of 10,000 different blogs used in this project 75% are randomly chosen for training and remaining 25% for validation. A brief description as well as the abbreviation used for each predictor are listed in Table 1. So, in a nutshell given a vector of 18 predictor values corresponding to a blog our goal is to predict the id of that blog’s Author.

Table 1. Description of the predictors used in this project.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Example** |
| ADJ | Frequency of adjective | big, old, green, incomprehensible, first |
| ADP | Frequency of adposition | in, to, during |
| ADV | Frequency of adverb | very, tomorrow, down, where, there |
| CCONJ | Frequency of coordinating conjunction | and, or, but |
| DET | Frequency of determiner | a, an, the |
| INTJ | Frequency of interjection | psst, ouch, bravo, hello |
| NOUN | Frequency of noun | girl, cat, tree, air, beauty |
| NUM | Frequency of numeral | 1, 2017, one, IV, MMXIV |
| PART | Frequency of particle | ’s, not, |
| PRON | Frequency of pronoun | I, you, he, she, themselves, somebody |
| PROPN | Frequency of proper noun | Mary, John, London, NATO, HBO |
| PUNCT | Frequency of punctuation | ., (, ), ? |
| SYM | Frequency of symbol | $, %, §, ©, +, −, ×, ÷, =, :), 😝 |
| VERB | Frequency of verb | run, runs, running, eat, ate, eating |
| SPACE | Frequency of space |  |
| average\_word\_length | Average word length | — |
| digits\_percentage | Percentage of digits used | — |
| total\_words | Total count of words | — |

Scatterplot between each pair of features are presented in Figure 1 as initial observation and the plots are used for preliminary inspection of any possible correlation or dependency among the features. It may be noted that a very strong correlation is present between many of the predictors. For example, ADJ is highly correlated with the ADP,ADV,CCONJ,DET,NOUN,PART,PRON,PUNCT,VERB and total\_words predictors. Similar trend can be observed for other predictors as well. Clearly this observation suggests the need for Principle Component Analysis(PCA) to filter out the variables which approximate most of the complexity in the dataset.

**Principle Component Analysis (PCA):** Scatterplots already confirmed that many predictors are highly correlated, hence we have performed PCA on the feature set consisting of 18 predictors to determine which of these predictors capture most of the variance in the dataset. Specifically, observations in the training dataset are used to find the principle components(PC) and summary of the PCA object is shown below. We have obtained 18 principal components which we call PC1-18 , and out of them, PC1 and PC2 explains 64% and 9.3% ,i.e. 73.3% of the total variability in the training dataset.

|  |
| --- |
|  |

Hence, we know that by analyzing the position of an observation based on PC1 and PC2 we can get an accurate view on where it stands in relation to other samples. The loading plot of the PCA presenting the

|  |
| --- |
| Figure 1. Scatterplots between all pairs of predictors. |

predictors contributing to the PC1 and PC2 respectively is shown in Figure 2. Loading plot shows that all predictors are significant for assigning blogs to their respective authors.

|  |
| --- |
| Figure 2. Loading plot of the PCA model. |

In the following sections we discuss about different classification models that we have developed using the training datasets and present their performance on the test dataset.

**Linear Support Vector Machines (SVM):** 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. Figure 3. shows that none of the features stand out in terms of importance across the dataset.

|  |
| --- |
| Figure 3. Importance of each feature as computed from the SVM models with 100 authors. |

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. Table 2. shows the results for 10-fold cross-validation(CV) for each experiment.

Table 2. Accuracy of the SVM models with 2,5,10,50 and 100 different authors obtained using 10-fold cross validation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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.