Machine Learning Project

Name: Umama Nasir Abbasi

Roll No: 23100265

Phase 1

Q1: What data cleaning did you end up doing in Task 2 to create your features in task 3 and why?

I used regex functions to clear out

1. Emojis
2. Stop words
3. Punctuations
4. Links
5. Mention to other users
6. Hashtags
7. Numbers
8. Extra spaces
9. **i** character

I removed this data because all of them would have caused problems during classification and none of them are important features for the authorship attribution.

I also converted all the data into lower case so that it is standardized.

Q2: In relation to the problem of authorship attribution, which classifiers do you feel would be most appropriate and why? Give a thorough comparison across all the classifiers you have studied in class

KNN: Not appropriate because high features can lead to curse of dimensionality.

Linear Regression: not appropriate because authorship attribution is a classification problem, and this is a regression method.

Logistic Regression: not feasible because of high number of features and since the complexity is equal of O(M*N), there will be a very high computational cost.

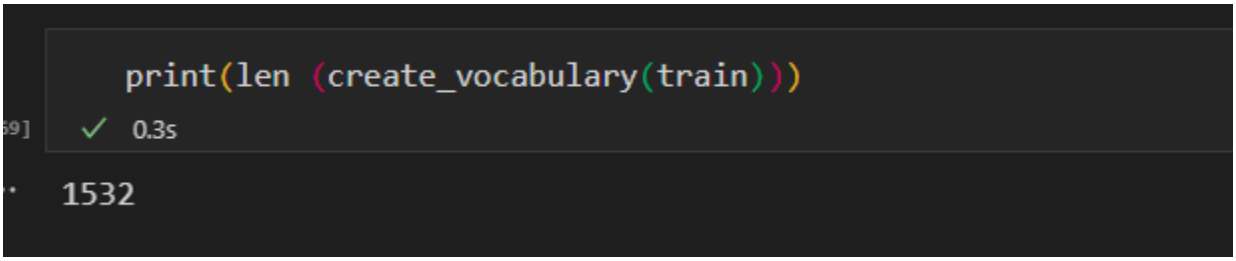
Perceptron: Perceptron require data to be linearly separable. Currently, our data is not linearly separable, so perceptron are not appropriate

SVM: Can work with SVM due to the use of kernel trick, which makes the data linearly separable.

Neural Networks: Can work. Neural networks learn the kernel which SVM applies to make the data linearly separable. Thus, they are also better than SVM and the best option for authorship attribution for now.

Q3: What is the ambient dimensionality of your solution and how would you determine the intrinsic dimensionality? Report both in your answer.

The ambient dimensionality is the total number of features in the dataset which is equal to 1532.



The intrinsic dimensionality is the actual number of dimension or features we end up working with. For this case, they can the most frequently occurring words within the training dataset.