**Model Evaluation**

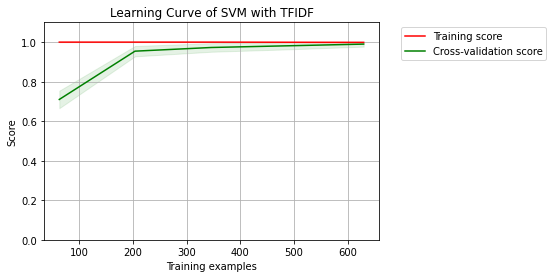
* 1. **Cross-Validation**

*Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample to estimate the skill of a machine learning model on unseen data.*

*Applying ten cross validations to the champion model (SVM based on TF\_IDF) gives the following accuracies 1, 0.98 ,0.98, 0.98, 1, 0.95 ,1 ,0.98 ,1 ,1with average cross validation accuracy 0.982 and 100% accuracy in training set, and testing accuracy of 0.976, and this small difference indicate that the model fit the data correctly avoiding Overfitting and Underfitting*

**1.2 Learning curve**

*Another way to get an estimate of model’s generalization performance is to look at the Learning Curve plot the model performance on the training set and the validation set as a function of training set size (or the training iterations), Ideal Learning Curve Model that generalizes to new data Testing and training learning curves converge at similar values Smaller the gap, the better our model generalizes.*



*If the training and cross-validation scores converge together as more data is added then the model will probably not benefit from more data. If the training score is much greater than the validation score then the model probably requires more training examples in order to generalize more effectively [1].*

* 1. **Bias and Variability**

*The curves are plotted with the mean scores, however variability during cross-validation is shown with the shaded areas that represent a standard deviation above and below the mean for all cross-validations. If the model suffers from error due to bias, then there will likely be more variability around the training score curve. If the model suffers from error due to variance, then there will be more variability around the cross validated score [1]. SVM Based on TF\_IDF model have low average bias and variability as shown in the previous figure, by using mlxtend.evaluate.bias\_variance\_decomp module we get average bias 0.043 and average variance: 0.015 and that guarantee avoiding underfitting and overfitting as the model perform well on the training set it also generalizes well according to the cross-validation metrics*

1. **- Error Analysis of Champion Model**

## *SVM based on TF\_IDF misclassifies only 4 documents out of 300 documents in the testing set. By giving this document a close look, we found that the model misclassifies Two documents of label C as E and B, One document of label E as C, and one document of label B as D.*

## 

## *These documents have common words in its actual label book and the predicted label book , So the model classify them to the most class that have a higher number of words in common with that document , as the model had trained to calculate the TF\_IDF for each word so it doesn’t understand that the sequence of some words or sentences is related to a specific class , So if a document of class c has one word with high TF\_IDF value ( as Marker Word) which indicates that this word belongs to Class C and at the same time it has several words which have TF\_IDF values a little bit lower than the Marker Word and this words refers class B then it assigned this document to Class B , another case that if specific words have high TF value in the document and TF value in any book rather than the actual class then the model will label it to the other class.*

## *For example, the next document is misclassified as B book while its actual label is C*

## *“ way deal child fancy believe every man woman since certain gentler sex cut small figure game mentioned volume actually existed time come every form learning however preposterous may seem made unlaborious possible would student knowledge string fact arranged sorted distilled set compact form ready rapid assimilation little fear student may wish future become master subject delve original source search fact date surely pirate taking broadest sense much entitled biographical dictionary clergyman race horse artist ferro concrete assured medical men directory lawyer list peer peerage book record name particular musician schoolmaster stockbroker saint bookmaker dare say average adjuster almanac peer* “

## *By plotting the word cloud for this document and the two Labels it indicates that two words (Man and Woman) have high TF value in the document and Label B and low TF in Label C, while it has the same IDF value for Both B and C so the model will label it as B while its actual class is C*

## *Document that misclassified*

## 

## *predicted class (B) Actual Class (C)*

## 

## 

## *In the end, the text classification problem depends directly on your feature engineering and how you prepare and clean the data and the model that is used.*

## *References*

## https://www.scikit-yb.org/en/latest/api/model\_selection/learning\_curve.html