Home Work - 4 Assignment

- what challenges do you notice in this data?

Ans.

The data given is a very raw data. Based on data, models cannot be directly built. Models can be built on the the data only after converting the data into a transformed form which has utilized by the model to produce good accuracy or prediction on the data. Based on the special characters, kind of sentence or the type of sentence, in the data, data has to be transformed. By calculating mean and variance of each of the features in our data, we can get a better transformed data. The data presented does not have proper sentence structures, many special characters are

utilized. It is difficult to decipher the sentences based on sentence breaks, special characters, structure of the sentences.

2. What data processing did you do?

I have applied fit - transform on the training data so the data can be scaled and scaling parameters can be learnt of the data. The transform method transforms all the features using the respective mean & variance. The fit part defines what features it will base future transformations on.

fit_transform utilizes CountVectorizer object and gets the vector of token counts.

fit_transform determines which tokens it will count and how they correspond to entries in the count vector.

- What models did you try and how did you represent the processed data for this/these models?

First, I used Naive Bayesian Classifier that expected data transformed by fit_transform to be converted to numpy array to fit the model.

For the NaiveBayesian Classifier, I got accuracy on test data to be 91 to 92%.

For the prediction, also similar transformation of the sparse matrix to numpy array.

Secondly, I used SVM classifier which does not require any transformation or changed representation of processed data for the model to run.

SVM classifier was able to give high accuracy of 97-1.-

- what steps did you take to avoid over-fitting?

I used train_test_split to divide the training data into training data and a testing data and checked the accuracy of the model on testing data.
This split is usually a 60-70% training data, 40-1. test data split or 50 - 30% training to test data split.

This will avoid overfitting of data.

on the model.

- Do you trust your model?

I do not trust my model even though it gives high accuracy as the size of the training data is less.

The Training data has 4000 traing samples. which might not build a model that gives high accuracy of large test data. If the test data has different kind of variation and different aspects that the existing model did not capture then it might not give good results.

- What could be improved on? If you had more time / more interest : what would you do next to build a better model?

If I had more time, I would preprocess the data using more accurate preprocessing techniques in order to get better model. Also, I would use a larger dataset to build a better model. Also, I would optimize parameters for the model example the 'C' parameter or the right kernel that has to be used.