Documentation for code

The entire code consists of predominantly the following steps:

- Pre-processing of the Dataset
- Converting dataset to BOW or TFIDF matrix
- Feature Engineering
- Model Making
- Evaluation of the Model

Pre-processing of both train and test datasets:

Prior to pre-processing, Missing Values, Class Imbalance etc were checked.

1. Removing punctuation
2. Converting to lower-case characters
3. Removing Stop Words
4. Lemmatization

Converting both train and test datasets into numpy matrix:

For this, the below four methods were tested.

1. Bag of Words Model *(Using CountVectorizer of sklearn)*
2. Bag of Words with N-Gram (2,2) Model *(Using CountVectorizer of sklearn with n_gram=(2,2))*
3. TF-IDF Model *(Using TfidVectorizer of sklearn)*
4. TF-IDF Model with N-Gram(2,2) Model *(Using TfidVectorizer of sklearn with n_gram=(2,2))* (Chosen)

The functions for 1, 2, and 4 are available in the code and can be used in place of the chosen TF-IDF Model with N-grams, if needed.

Feature Engineering of both train and test datasets:

The following things were applied to the matrix obtained from previous step to choose the best features:

1. Clustering *(K-Means was used here)*
2. LDA
3. NMF
4. SVD-Truncated

Model Making:

The final feature-selected matrix was used to choose and fit the model.

Two Algorithms were chosen either for their simplicity or other constraints.

1. Logistic Regression *(LogisticRegression of sklearn was used)*
2. Random Forest *(RandomForestClassifier of sklearnw was used)*

Both the algorithms were implemented only after using Random Search for choosing best possible hyperparameters from a specified grid of hyperparameter values.

Also, Cross-Validation(5-fold) was used to make sure that the model would not overfit.

After this, the model was used to predict the test values.

Evaluation of the Model:

The Model was evaluated based on Accuracy along with Miss-Rate(with the help of confusion matrix generated).

Finally, the results were stored in the file "*final_submission.csv*".

The following is a list of functions and their usage:

Function	Usage
1. perform_pre_processing(args)	Performs the above mentioned pre-processing of data and contains 4 function calls, each of which performs a specific task.
2. make_bag_of_words(args)	Creates the first BOW Model and converts the given dataset into a numpy matrix.
3. make_bag_of_words_N_grams(args)	Creates the second BOW Model with N-Gram (2,2) and converts the given dataset into a numpy matrix.
4. perform_feature_engineering(args)	This function performs all the tasks mentioned under FeatureEngineering.
5. make_tf_idf_model(args)	Creates the 3 or 4th TF-IDF model depending upon what range of n_gram is passed to the function. If n_gram=(1,1), it creates Model Number 3 Else: it creates TF-IDF with (2,2) n_gram range model.
6. make_model(args)	This function initializes a model, chooses hyperparameters using RandomizedSearch CV which finds best possible value of hyperparameters and while fitting the model performs cross-validation to prevent overfitting.
7. eval_test_label(args)	This function evaluates the performance of the model on test data.

Also, there are comments everywhere in the code in order to facilitate understanding of the code.

Please feel free to ask any more questions if needed.

Thank you!