1. Reading the dataset

The necessary packages are imported and the train dataset is read using pandas

2. Cleaning the text

Non English letters are substituted with 'space'.

The words are converted to lower case and stemmed using the 'Porter Stemmer'.

The Corpus is created by joining the preprocessed words.

3. Creating the Bag of Words model

The sparse columns of word in the corpus are created to denote whether a particular word is present in the tweet or not.

4. Fitting RF to the Training set

The Random forest model with 120 estimators is fitted for the training dataset.

5. Cleaning and creating bag of words for test dataset

The above steps 2 and 3 are applied to the test dataset.

6. Predicting the Test set results

The NB classifier model is used to predict the classes of test dataset tweets.

7. Writing the predictions to a csv file

The Result is written to a file 'test_predictions.csv'.