Steps for Sentiment Analysis using Unigrams, Bigrams & combination of both training on Naïve Bayes

1. **For Unigrams** - Creating list of list of all entries/sentences (positive and negative) by splitting all words in the sentences, removing special all characters and stop words. (line 33-51)
2. Creating a list of positive and negative unigrams. (line 54-55)
3. Calculating the frequency distribution of all words/unigrams and then the frequency distribution within positive and negative labels. (line 59-71)
4. Creating dictionary of unigrams along with scores based on chi-squared test. (line 74-78)
5. Sub setting positive and negative unigrams with scores into separate positive and negative unigrams dictionaries. (line 81-82)
6. Creating functions **(find\_best\_pos\_unigrams and find\_best\_neg\_unigrams)** to find best positive and negative unigrams from the dictionaries formed in point 5 based on scores. Combining the individual list of positive and negative best unigrams into one single list – total\_best\_unigrams. (line 87-99)
7. Ran the model by iterating best number of positive and negative unigrams from 800-1200. Got best results with best number of positive and negative unigrams = 1200.
8. Creating a function **(best\_uni\_features)** to find whether unigrams in total\_best\_unigrams occur in list of list of all entries. (line 102-103)
9. Creating list of list of all entries/sentences (positive and negative) and passing **best\_uni\_features** function to check whether unigrams in total\_best\_unigrams occur in the sentences. This creates a list of list entries only with unigrams in total\_best\_unigrams occurring in them. (line 106-126)
10. Splitting the lists formed in step 9 into train and test features to be passed to the model. (line 129-132)
11. Training a Naive Bayes Classifier on the training set. (line 135)
12. Initiating reference sets and test sets. (line 138-139)
13. Putting correctly labeled sentences in reference sets and the predictively labeled version in test sets. (line 142-145)
14. Printing metrics to see how well the model works. (line 148-153)
15. **For Bigrams** - Creating list of list of all entries/sentences (positive and negative) by splitting all words in the sentences, removing special all characters and stop words and there after creating lists for positive and negative bigrams. (line 159-193)
16. Calculating the frequency distribution of all bigrams and then the frequency distribution within positive and negative labels. (line 196-208)
17. Creating dictionary of bigrams along with scores based on chi-squared test. (line 211-215)
18. Sub setting positive and negative bigrams with scores into separate positive and negative bigrams dictionaries. (line 218-219)
19. Creating functions **(find\_best\_pos\_bigrams and find\_best\_neg\_ bigrams)** to find best positive and negative bigramsfrom the dictionaries formed in point 18 based on scores. Combining the individual list of positive and negative best bigramsinto one single list – total\_best\_bigrams. (line 225-243)
20. Ran the model by iterating best number of positive and negative bigramsfrom 4000-9000. Got best results with best number of positive and negative bigrams= 6000, 7000 and 8000.
21. Creating functions **(best\_** **bigrams\_features6, bigrams\_features7 and bigrams\_features8)** to find whether bigrams in total\_best\_bigrams occur in list of list of all entries. (line 246-253)
22. Creating list of list of all entries/sentences (positive and negative) and there after creating lists for positive and negative bigrams and passing respective **best\_** **bigrams\_features** functions to check whether bigrams in total\_best\_bigrams occur in the sentences. This creates a list of list entries only with bigrams in total\_best\_bigrams bigrams occurring in them. (line 256-287, 318-349 and 380-411)
23. Splitting the lists formed in step 22 into train and test features to be passed to the model. (line 290-293, 352-355 and 414-417)
24. Training a Naive Bayes Classifier on the training set. (line 296, 358 and 420)
25. Initiating reference sets and test sets. (line 299-300, 361-362 and 423-424)
26. Putting correctly labeled sentences in reference sets and the predictively labeled version in test sets. (line 303-306, 365-368 and 427-430)
27. Printing metrics to see how well the models works. (line 309-314, 371-376 and 433-438)
28. Combining top 1200 positive and negative unigrams (from point 6) and top 6000, 7000 and 8000 positive and negative bigrams respectively (from point 19) into separate lists. (line 441-453)
29. We get 3 separate lists with top 1200 unigrams and top 6000, 7000 and 8000 bigrams respectively.
30. Creating functions **(best\_uni\_bi\_features6, best\_uni\_bi\_features7 and best\_uni\_bi\_features8)** to find whether unigrams and bigrams in total\_best\_ uni\_bi occur in list of list of all entries. (line 456-463)
31. Creating list of list of all positive and negative unigrams and bigrams and passing respective **best\_uni\_bi\_features** functions to check whether unigrams and bigrams total\_best\_ uni\_bi occur in the sentences. This creates a list of list entries only with unigrams and bigrams in total\_best\_ uni\_bi unigrams and bigrams occurring in them. (line 466-494, 526-554 and 586-614)
32. Splitting the lists formed in step 31 into train and test features to be passed to the model. (line 497-500, 557-560 and 617-620)
33. Training a Naive Bayes Classifier on the training set. (line 504, 564 and 624)
34. Initiating reference sets and test sets. (line 507-508, 567-568 and 627-628)
35. Putting correctly labeled sentences in reference sets and the predictively labeled version in test sets. (line 426-430, 571-574 and 631-634)
36. Printing metrics to see how well the models works. (line 517-522, 577-582 and 637-642)
37. Creating respective (3 lists with 1200 unigrams and 6000, 7000 & 8000 bigram comb.), total list with test features (unigrams and bigrams) along with positive and negative probability calculated using Naïve Bayes classifier. (line 645-658, 666-679 and 687-700)
38. Adding a new label (‘pos’ and ‘neg’) column in the lists created in point 37 based on positive or negative probability is greater. (line 660-664, 681-685 and 702-706)
39. Creating respective (3 lists with 1200 unigrams and 6000, 7000 & 8000 bigram comb.), final list with iterating threshold value from 0.1 to 0.9 to ascertain the assignment of positive or negative label based on whether the positive and negative probability is greater than the threshold value.
40. Calculating True Positive(tp), True Negative(tn), False Positive(fp) and False Negative(fn) for the lists created in point 39. Thereafter calculating respective True Positive Rate and False Positive Rate with respect to the threshold values. The final lists would include 3 columns (threshold value, True Positive Rate and False Positive Rate). (line 708-774)
41. Taking only the True Positive Rate and False Positive Rate columns from the lists created in point 40 plotting respective ROC curves and finding the area under the curve. We will go with the model with the maximum area under the curve. (line 776-803)