Sentiment Analysis Using Naive Bayes

Perform Text Classification on the data. The tweets have been pulled from Twitter related to coronavirus and manual tagging has been done then.  
The names and usernames have been given codes to avoid any privacy concerns.

You have to only use "OriginalTweet" and "sentiment" column for the assignment. You can either drop or ignore other columns for this assignment.

You might use some of the References given below:

1. [https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html (Links to an external site.)](https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html)
2. [https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html (Links to an external site.)](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)
3. [https://towardsdatascience.com/ml-pipelines-with-grid-search-in-scikit-learn-2539d6b53cfb (Links to an external site.)](https://towardsdatascience.com/ml-pipelines-with-grid-search-in-scikit-learn-2539d6b53cfb)

Dataset: [Corona\_NLP\_train.csv](https://bits-pilani.instructure.com/courses/972/files/176396?wrap=1)[download](https://bits-pilani.instructure.com/courses/972/files/176396/download?download_frd=1)

Steps to be performed are as follows:

1) Import required libraries - 2 Mark

2) Read dataset and perfom Text processing for the tweets ( Remove Stop words , special characters and convert the text to lowercase ) - 3 Mark

3) Using the train\_test\_split function of Sklearn, Split train and test dataset - 1 Mark

4) Create pipeline and define parameters for GridSearch ( You might Refer the code below ) - 1 Mark

text\_clf = Pipeline([('vect', CountVectorizer()),  
                     ('tfidf', TfidfTransformer()),  
                     ('clf', MultinomialNB())])

tuned\_parameters = {  
    'vect\_\_ngram\_range': [(1, 1), (1, 2), (2, 2)],  
    'tfidf\_\_use\_idf': (True, False),  
    'tfidf\_\_norm': ('l1', 'l2'),  
    'clf\_\_alpha': [1, 1e-1, 1e-2]  
}

5) Perform classification (using GridSearch) - 3 Mark

6) Print the confusion matrix, accuracy, F1 score on the test dataset - 2 Mark