Ethics for NLP: SS 2024 **(Due: 08 May 2024, 11:59pm)**

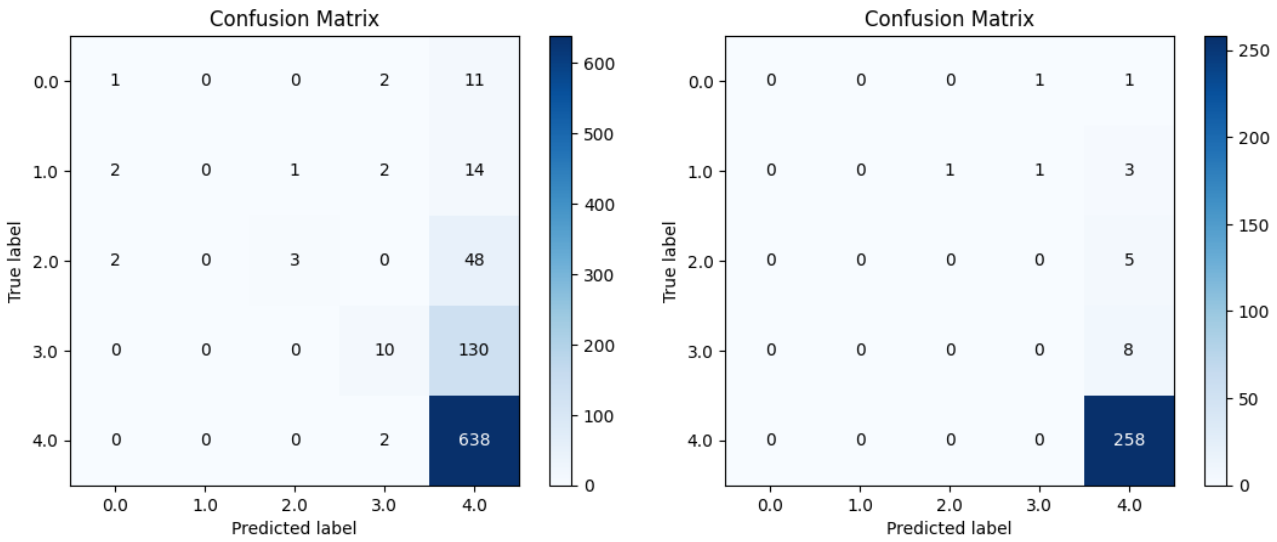
Homework 0

Name: Julian Steiner *Matriculation No.:* 2669944

Problem 1

(1.4)

In Figure 1a the confusion matrix calculated on the validation set is shown. In Figure 1b the confusion matrix calculated on the test set is shown. The classifier reached 75% accuracy on the validation set and on the test set even 95%. But the accuracy metric is misleading here. The classifier classifies most texts with the label 4, i.e. a rating of 5. Only a few examples of other ratings are correctly classified. For example in the validation set only 10 ratings of the label 3 are correctly classified, the remaining examples for the label 3 are wrongly predicted with the label 4. This allows a high accuracy to be calculated despite incorrect predictions. If the dataset contains many examples with label 4 and the classifier predicts label 4 with a higher probability, many will still be predicted correctly. We can see this from the calculated confusion matrix on the test set. In total, this dataset contains 258 label 4 samples, which are all predicted correctly. On the other hand, the dataset contains only 17 samples with other labels, none of which are predicted correctly.

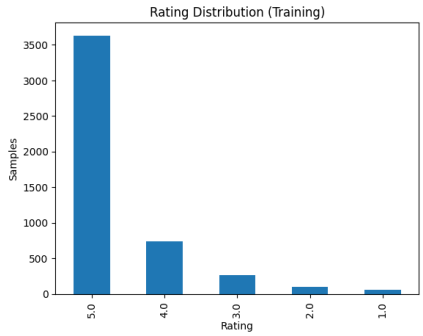


(a) Confusion matrix on validation set.

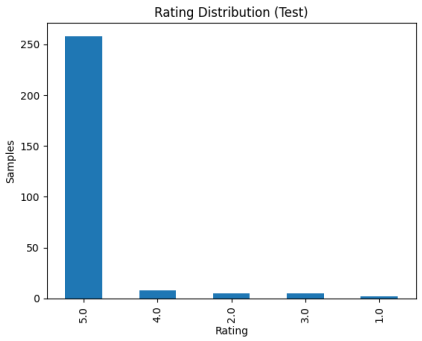
(b) Confusion matrix on test set.

Figure 1: Confusion matrix on different datasets.

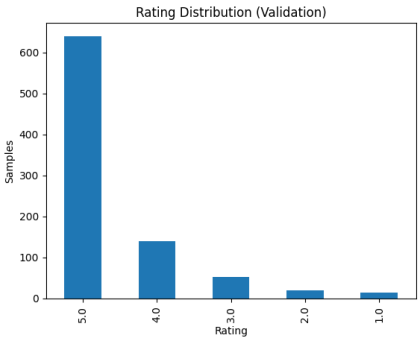
One possible reason may be that the training, validation and test data predominantly contain examples with a rating of 5, i.e. label 4. The distribution of the ratings of each dataset is shown in Figure X. But there can be many other reasons.



(a) Confusion matrix on validation set.



(b) Confusion matrix on test set.



(c) Confusion matrix on test set.

Figure 2: Confusion matrix on different datasets.

Problem 2

(2.1)

- (1):
- (2):
- (3):
- (4):
- (5):
- (6):

(2.2)