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

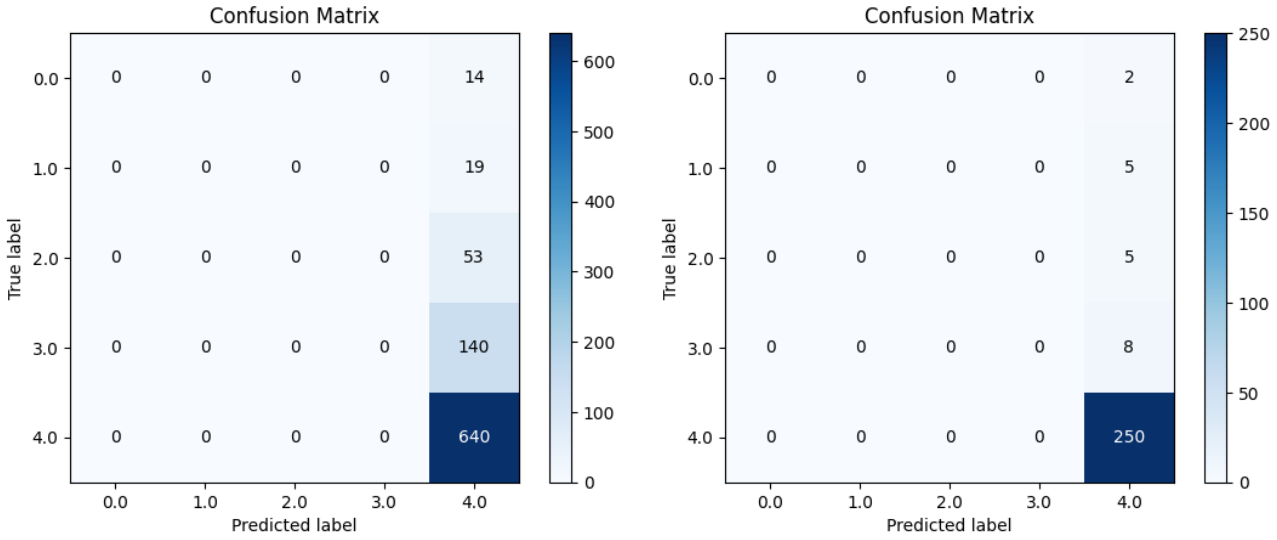
Homework 0

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 74% accuracy on the validation set and on the test set even 93%. But the accuracy metric is misleading here. The classifier classifies all texts with the label 4, i.e. a rating of 5. Despite incorrect predictions, a high accuracy can be calculated. If the dataset, on which the accuracy are calculated, contains many examples with label 4 and the classifier predicts label 4 with a higher probability (or only label 4), many will still be predicted correctly. We can see this from the calculated confusion matrix on the test set. In total, this dataset contains 250 label 4 samples, which are all predicted correctly. On the other hand, the dataset contains only 20 samples with other labels, none of which are predicted correctly, all categorized as label 4. This ends in a high accuracy.

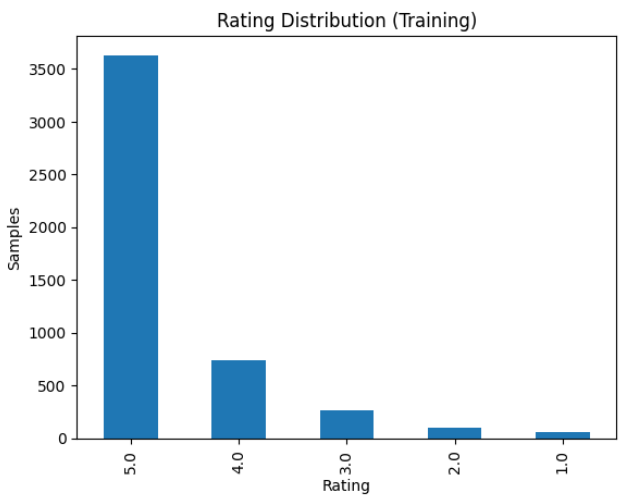


(a) Confusion matrix on validation set.

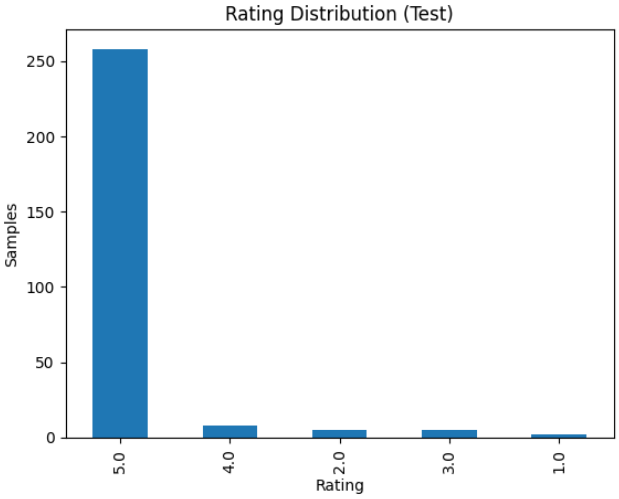
(b) Confusion matrix on test set.

Figure 1: Confusion matrix on different datasets.

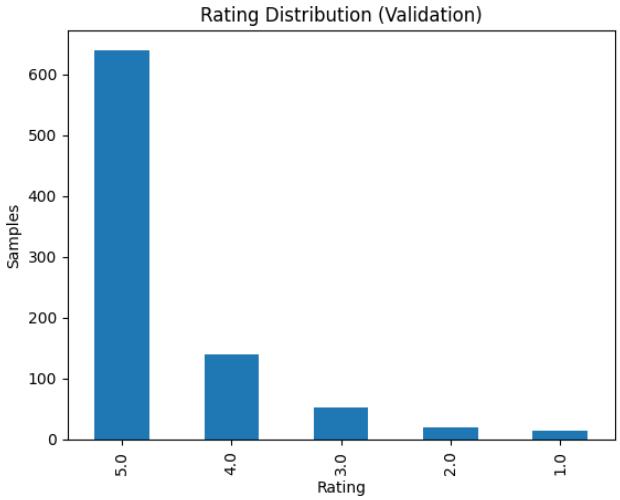
One reason may be that the training, validation and test data predominantly contain examples with a rating of 5, i.e. label 4. The distribution of the ratings of each dataset is shown in Figure 2. Note: This distributions don't show the preprocessed data and also includes empty texts, which are filtered out for training the model. Another note: Rating is not the label. The label of a respective rating is always the rating minus one. In addition, there can be many other reasons, i.e. the model architecture.



(a) Rating distribution of training set.



(b) Rating distribution of test set.



(c) Rating distribution of validation set.

Figure 2: Rating distribution of different datasets.

Problem 2

(2.1)

- (1): Benefit from such a classifier would the teacher or the professor. He or she would no longer have to evaluate the essays or use the classifier as support for grading the essays. But students could also benefit. Students could get immediate feedback on their submission without having to wait for the long correction period.
- (2): The students could be harmed by such a system. For example, incorrect grading could unfairly punish or reward students. In addition the teachers could also be harmed. The teacher could rely too heavily on the classifier and may overlook important things in the student essays.
- (3): The data is not representative. The data is only from one course of one year. The diversity in the data set could be missing. For example, from other courses, other universities and also from several years.
- (4): Ethical concerns regarding data collection may arise from the explicit consent of the participants with whom the classifier was trained. Have the students consented to the use of their essays for the task of training a classifier?

- (5): Neural networks are black boxes, so to speak. It is not possible to understand exactly what criteria the model used to arrive at its prediction. Therefore, I would say that neural networks, as long as they do not make their prediction comprehensible, are not a suitable modeling approach for the task. On the other hand, neural networks are the best approach for modeling NLP at the moment. Of course with different architectures than we used in the exercise. This raises the question of whether it is generally a good idea to develop models for such purposes.
- (6): A possible dual use for such a classifier could be, for example, that the classifier classifies the essay according to how closely the opinion of the essay matches that of the teacher. This could limit the diversity of opinion within courses and lead to unfair and subjective assessments.

(2.2)

I think, there are several issues with project A that need to be clarified by the ethics board. First, in the research project idea is mentioned that the tool tries "to find other papers related to each sentence - even if they are not cited in the main paper". How can it be ensured that for the respective sentence, it is a correct classification of the found paper, even if it was not cited. In the case of misclassification, the author of the paper could be falsely suspected of plagiarism because he or she did not cite the found paper. As a result, the author could be falsely harmed. Secondly, are the researchers allowed to use the publicly accessible research papers for their study? Thirdly, the researcher "will record the participants using the tool via screen recording, and analyize the recordings to improve the usability of the tool". Is this collection of data the absolute necessary minimum for this study?

On the one hand, the participants in Project B are lied to. The participants are informed that the study is being conducted by the psychology department, although it is being conducted by computer science department. In addition, the participants do not know that they are interacting with a chatbot. The board must clarify whether these lies are necessary to achieve the objectives of the research. Another point that should be examined by the ethics board is whether the participant could be harmed if the chatbot is placed on the "politness" level rude.