Problemset 4

Your role is that of a data scientist working for a company that provides sentiment analysis services to restaurants. The company has provided you with a dataset of reviews (`restaurant-reviews.csv`) for different restaurants located in Kiel. The dataset contains the following columns:

- `name`: Name of the restaurant
- `restaurant_url`: URL of the restaurant on the tripadvisor website
- `title`: Title of the review
- `text`: Text of the review
- `rating`: Rating of the review (1-5)

Your task is to carry out a machine learning experiment to predict the rating of a review on a scale from 1 - 5 based only on the text of the review. The focus will be on the selection of text preprocessing and feature engineering strategies. In the exercises, not all details are specified. Make reasonable assumptions and decisions, and document them in your submission.

Exercise 1

The first part of your experiment is about a **systematic analysis of the effectiveness of different text preprocessing techniques**. In this part of the experiment, vectorize the texts using Bag of Words and choose a baseline machine learning algorithm with default hyperparameters. Vary ONLY your text preprocessing strategy (e.g. lowercasing, stopword removal, stemming, lemmatization, or some combination of these). Keep the rest of your setup fixed. Evaluate the performance of your models on a holdout set. Provide a table with the evaluation results. The table should also show the number of features used in each case. Briefly summarize your main findings.

You do NOT need to do an exhaustive grid search of all possible combinations of preprocessing strategies. Instead, focus on 5 strategies, including: 1) no cleaning, 2) selected single preprocessing measures and 3) selected combinations of measures that you consider most promising.

Exercise 2

The second part of your experiment is about a **systematic analysis of the effectiveness of different feature engineering techniques**. Pick the best preprocessing strategy from Exercise 1 and vary ONLY your feature engineering strategy: bag of words, TF-IDF, bag of 2-grams. Keep the rest of your setup fixed. Evaluate the performance of your models on a holdout set. Provide a table with the evaluation results. Briefly summarize your main findings.

Exercise 3

There are different ways how you can further improve the performance of your model: e.g. using different machine learning algorithms, tuning hyperparameters of a given algorithm, or also different strategies for handling class imbalances. In the final part of your experiment, you are free to choose one of these aspects and evaluate its effectiveness. Briefly summarize your main findings.