Case 3. Patient Drug Review

Neural Networks for Machine Learning Applications, Spring 2022

Type

Team work, 15-20 hours

Aim

The aim of this assignment is to learn to use neural networks in machine-learning problems involving text data.

Task

Your task is to use the methods of text processing and to experiment with recurrent and/or convolutional (1D) neural networks to create a classifier for a collection of patient drug reviews extracted from Drugs.com. Drugs.com is a comprehensive source of drug information online. The original dataset is explained in Grässer et al (2018) article.

The task is to build a classifier to address the following question:

- Can you predict the rating of the drug based on the review?

Following the article mentioned above, simplify the rating labels down to just three different classes: “negative” (rating < 5), “neutral” (rating 5 or 6) and “positive” (rating > 6). After this, you should build a multiclass classifier to predict these three classes (negative, neutral, positive) from the content of the written review.

To assess the performance of your classifier, you should evaluate your model with test data, and present the full confusion matrix together with the classification report and the most relevant metrics (accuracy + Cohen’s kappa) of this evaluation, with appropriate interpretations (compare your results to the original article, Grässer et al. Table 2, p. 124).

Download the KUC Hackathon Winter 2018 dataset from the Hackathon’s homepage: https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018 or use the Kaggle environment to build your Notebook. The dataset is already split into training and test sets. Use pandas to read the data.

Return

Upload your Notebook in OMA.

Evaluation

The following categories are used for evaluation:

- Organisation
  - The code is sequential and the code cells (parts of scripts) are in proper order
  - The document follows a clear structure
- Clarity
  - The document (and embedded code) is clear, polished, and easy to understand
  - The code follows good coding practices and contains sufficient comments
  - The document parts support the code
- Contents
  - The background and data preprocessing are well explained
  - The models are validated
  - The results are reasonable
  - The conclusions are clearly stated and in a line with the results

max. 15 points. Late submission reduces the maximum achievable points.

Materials

- Deep Learning with Python.
  - Chapter 6. Deep learning for text and sequences.
- Deep learning for text
- GitHub - fchollet/deep-learning-with-python-notebooks samples
  - See the first edition folder
  - 6.1. One-hot encoding of words or characters
  - 6.1. Using word embeddings
  - 6.2. Understanding recurrent neural networks
- Tutorials | TensorFlow Core
  - Beginner: Load and Preprocessing data > Text
  - Advanced: Text