Sprachverarbeitung: Übung

SoSe 24

Janis Pagel
Department for Digital Humanities, University of Cologne

13 May 2025

Please submit your solutions as a ZIP file on Ilias, containing a single PDF file with the calculations for Exercise 1–3, and a Python file or Notebook for Exercise 4. You can either solve the exercise for which you need to do calculations by hand on a sheet of paper and scan it or use the capabilities to write mathematical equations of tools like MS Word / LibreOffice / LaTeX, etc. to write down your calculations digitally.

Imagine the following situation:
You have a dataset with gold annotations of sentiment ratings (1 (positive), -1 (negative), 0 (neutral)) of a small text. You have implemented two different machine learning systems (System 1 and System 2) which are able to automatically classify tokens regarding their sentiment. You run your two systems on the dataset in order to check how good they are in classifying sentiment. The results are shown in table 1:

Token	Gold	System 1	System 2
The	0	0	0
quick	0	1	0
brown	0	0	0
fox	0	-1	0
jumped	0	0	0
over	0	0	0
the	0	0	0
lazy	-1	0	-1
dog	0	-1	1
.	0	0	0
Mary	0	1	0
ate	0	0	-1
her	0	0	0
apple	0	0	0
.	0	0	0
This	0	0	0
made	0	0	0
me	0	0	1
very	0	1	0
happy	1	1	1
.	0	0	0

Table 1: Gold data and results of the two systems..

Exercise 1.

As a first step, you are interested in how well your systems are able to classify only *1* (positive) sentiment. To evaluate this, you replace all occurrances of *-1* with *0* and get table 2.

Using this table, identify the number of all true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for the sentiment classes *1* and *0* for each system. Afterwards, calculate accuracy, precision, recall and F1 score for both systems. Which system is better in classifying 1 (positive) sentiment?

Token	**Gold**	**System 1**	**System 2**
The	0	0	0
quick	0	1	0
brown	0	0	0
fox	0	0	0
jumped	0	0	0
over	0	0	0
the	0	0	0
lazy	0	0	0
dog	0	0	1
.	0	0	0
Mary	0	1	0
ate	0	0	0
her	0	0	0
apple	0	0	0
.	0	0	0
This	0	0	0
made	0	0	0
me	0	0	1
very	0	1	0
happy	1	1	1
.	0	0	0

Table 2: Only *1* and *0*.

Exercise 2.

You decide to compare the performance of your systems regarding classifying the *1* class with two baselines:

1. A Majority Baseline (all tokens are labeled with the most frequently occurring class)
2. A Random Baseline (all tokens are labeled randomly)

You receive the results in table 3.

Using this table, identify the number of TP, TN, FP and FN for the two baselines and calculate accuracy, precision, recall and F1 measure. Compare the results for the baselines with the results for System 1 and System 2. Based on this comparison, can you explain how accuracy can sometimes be misleading in judging the performance of different systems? When should you compare your results with a majority baseline, when with a random baseline?

Token	Gold	Majority BL	Random BL
The	0	0	1
quick	0	0	1
brown	0	0	0
fox	0	0	1
jumped	0	0	1
over	0	0	0
the	0	0	1
lazy	0	0	0
dog	0	0	0
.	0	0	0
Mary	0	0	1
ate	0	0	1
her	0	0	1
apple	0	0	0
.	0	0	0
This	0	0	1
made	0	0	1
me	0	0	1
very	0	0	1
happy	1	0	1
.	0	0	0

Table 3: Baselines.

Exercise 3.

Now, use the original results from table 1 and identify the number of TP, TN, FP and FN for the *1*, *0* and *-1* classes for System 1 and System 2 and calculate macro-average precision, macro-average recall and macro-average F1 score as well as micro-average precision, micro-average recall and micro-average F1 score.

Exercise 4.

Load the CSV file from `https://lehre.idh.uni-koeln.de/site/assets/files/5615/evaluation.csv` into a pandas DataFrame and calculate accuracy, precision, recall, F1 score and the macro and micro averages using `sklearn.metrics` for System 1, System 2, the Majority Baseline and the Random Baseline.