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?

Solution 1.

System 1 **Gold**	**1 (Positive)**	**0 (Neutral)**
1 (Positive)	1	0
0 (Neutral)	3	17

System 2 **Gold**	**1 (Positive)**	**0 (Neutral)**
1 (Positive)	1	0
0 (Neutral)	2	18

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*.

System 1:

$$\begin{aligned} \mathit{Accuracy} &= \frac{1+17}{1+3+17+0} &= 0.86 \\ \mathit{Precision} &= \frac{1}{1+3} &= 0.25 \\ \mathit{Recall} &= \frac{1}{1+0} &= 1.0 \\ F_1\text{-}\mathit{Measure} &= \frac{2 \times 0.25 \times 1.0}{0.25 + 1.0} &= 0.4 \end{aligned}$$

System 2:

$$\begin{aligned}
\mathit{Accuracy} &= \frac{1+18}{1+2+18+0} &= 0.9\\
\mathit{Precision} &= \frac{1}{1+2} &= 0.33\\
\mathit{Recall} &= \frac{1}{1+0} &= 1.0\\
F_1\text{-}\mathit{Measure} &= \frac{2\times 0.33\times 1.0}{0.33+1.0} &= 0.5
\end{aligned}$$

	Accuracy	Precision	Recall	F$_1$-Measure
System 1	0.86	0.25	1.0	0.4
System 2	0.9	0.33	1.0	0.5

System 2 performs better than System 1, since it achieves higher scores for all measures (except for recall, for which the scores are identical, i.e. the performance of the two systems is equal).

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.

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.

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?

Solution 2.

Majority BL	Positive	Neutral
Gold		
Positive	0	1
Neutral	0	20

Random BL	Positive	Neutral
Gold		
Positive	1	0
Neutral	12	8

Majority Baseline:

$$
\begin{aligned}
\mathit{Accuracy} &= \frac{0+20}{0+0+20+1} &&= 0.95\\
\mathit{Precision} &= \frac{0}{0+0} &&= NA\\
\mathit{Recall} &= \frac{0}{0+1} &&= 0.0\\
F_1\text{-}\mathit{Measure} &= NA &&= NA
\end{aligned}
$$

Random-Baseline:

$$
\begin{aligned}
\mathit{Accuracy} &= \frac{1+8}{1+12+8+0} &&= 0.43\\
\mathit{Precision} &= \frac{1}{1+12} &&= 0.08\\
\mathit{Recall} &= \frac{1}{1+0} &&= 1.0\\
F_1\text{-}\mathit{Measure} &= \frac{2\times 0.08\times 1.0}{0.08+1.0} &&= 0.15
\end{aligned}
$$

	Accuracy	Precision	Recall	F_1-Measure
System 1	0.86	0.25	1.0	0.4
System 2	0.9	0.33	1.0	0.5
Majority BL	0.95	NA	0.0	NA
Random BL	0.43	0.08	1.0	0.15

The accuracy for datasets in which one class is much more frequent than other classes can be misleading. This becomes clear looking at these results, where the majority baseline achieves the highest accuracy simply by labeling all tokens as *0*. Since a baseline should be as strong as possible while being as simple to implement as possible, a majority baseline is nontheless a great tool to fairly judge your system's performance when your dataset is unbalanced. The majority baseline has the disadvantage that precision cannot be computed for the non-majority class (*1*) since it never labeles a token as *1*.

The random baseline is suitable for datasets in which all classes are more or less equally distributed, since it can achieve the best results for such datasets.

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.

Solution 3.

System 1 Gold	1 (Positive)	0 (Neutral)	-1 (Negative)
1 (Positive)	1	0	0
0 (Neutral)	3	14	2
-1 (Negative)	0	1	0

System 2 Gold	1 (Positive)	0 (Neutral)	-1 (Negative)
1 (Positive)	1	0	0
0 (Neutral)	2	16	1
-1 (Negative)	0	0	1

System 1:

$$Precision_{pos} = \frac{1}{1+3+0} = 0.25$$

$$Precision_{neut} = \frac{14}{0+14+1} = 0.93$$

$$Precision_{neg} = \frac{0}{0+2+0} = 0.0$$

$$Recall_{pos} = \frac{1}{1+0+0} = 1.0$$

$$Recall_{neut} = \frac{14}{3+14+2} = 0.74$$

$$Recall_{neg} = \frac{0}{0+1+0} = 0.0$$

$$F_1\text{-}Measure_{pos} = \frac{2*0.25*1.0}{0.25+1.0} = 0.4$$

$$F_1\text{-}Measure_{neut} = \frac{2*0.93*0.74}{0.93+0.74} = 0.82$$

$$F_1\text{-}Measure_{neg} = \frac{2*0.0*0.0}{0.0+0.0} = NA$$

$$Macro\text{-}Average\text{-}Precision = \frac{0.25+0.93+0.0}{3} = 0.39$$

$$Macro\text{-}Average\text{-}Recall = \frac{1.0+0.74+0.0}{3} = 0.58$$

$$Macro\text{-}Average\text{-}F_1\text{-}Measure = NA = NA$$

$$or$$

$$= \frac{0.82+0.4}{2} = 0.61$$

$$Micro\text{-}Average\text{-}Precision = \frac{0.25*1+0.93*19+0.0*1}{21} = 0.85$$

$$Micro\text{-}Average\text{-}Recall = \frac{1.0*1+0.74*19+0.0*1}{21} = 0.72$$

$$Micro\text{-}Average\text{-}F_1\text{-}Measure = NA = NA$$

$$or$$

$$= \frac{0.4*1+0.82*19}{20} = 0.8$$

System 2:

$$\begin{aligned}
\textit{Precision}_{pos} &= \frac{1}{1+2+0} &&= 0.33 \\
\textit{Precision}_{neut} &= \frac{16}{0+16+0} &&= 1.0 \\
\textit{Precision}_{neg} &= \frac{1}{0+1+1} &&= 0.5 \\
\textit{Recall}_{pos} &= \frac{1}{1+0+0} &&= 1.0 \\
\textit{Recall}_{neut} &= \frac{16}{2+16+1} &&= 0.84 \\
\textit{Recall}_{neg} &= \frac{1}{0+0+1} &&= 1.0 \\
F_1\textit{-Measure}_{pos} &= \frac{2*0.33*1.0}{0.33+1.0} &&= 0.5 \\
F_1\textit{-Measure}_{neut} &= \frac{2*1.0*0.84}{1.0+0.84} &&= 0.91 \\
F_1\textit{-Measure}_{neg} &= \frac{2*0.5*1.0}{0.5+1.0} &&= 0.67 \\
\textit{Macro-Average-Precision} &= \frac{0.33+1.0+0.5}{3} &&= 0.61 \\
\textit{Macro-Average-Recall} &= \frac{1.0+0.84+1.0}{3} &&= 0.95 \\
\textit{Macro-Average-}F_1\textit{-Measure} &= \frac{0.5+0.91+0.67}{3} &&= 0.69 \\
\textit{Micro-Average-Precision} &= \frac{0.33*1+1.0*19+0.5*1}{21} &&= 0.94 \\
\textit{Micro-Average-Recall} &= \frac{1.0*1+0.84*19+1.0*1}{21} &&= 0.86 \\
\textit{Micro-Average-}F_1\textit{-Measure} &= \frac{0.5*1+0.91*19+0.67*1}{21} &&= 0.88
\end{aligned}$$

	System 1	System 2
Macro-Average Precision	0.39	0.61
Macro-Average Recall	0.58	0.95
Macro-Average F$_1$-Measure	NA/0.61	0.69
Micro-Average Precision	0.85	0.94
Micro-Average Recall	0.72	0.86
Micro-Average F$_1$-Measure	NA/0.8	0.88

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.

Solution 4.

```
import pandas as pd
import sklearn.metrics
import numpy as np
df = pd.read_csv("evaluation.csv")

for system in ["System1", "System2", "MajorityBaseline", "RandomBaseline"]:
    print(system)
    accuracy = sklearn.metrics.accuracy_score(df["Gold"], df[system])
    precision = sklearn.metrics.precision_score(df["Gold"], df[system], average = None,
                                            zero_division = np.nan)
    recall = sklearn.metrics.recall_score(df["Gold"], df[system], average = None,
                                            zero_division = np.nan)
    f1 = sklearn.metrics.f1_score(df["Gold"], df[system], average = None, zero_division
                                            = np.nan)
    print(f"Accuracy: {accuracy}")
    print(f"Precision: {precision}")
    print(f"Recall: {recall}")
    print(f"F1 Score: {f1}")
    print(sklearn.metrics.classification_report(df["Gold"], df[system], zero_division =
                                            np.nan))
```