Sprachverarbeitung: Übung

SoSe 24

Janis Pagel
Department for Digital Humanities, University of Cologne

2024-05-14

Please submit your solutions as a single PDF file on Ilias. For this exercise, you need to do calculations and document your approach/steps to solution. You do do not need to submit any Python code this time. You can either solve the exercise by hand on a sheet of paper, scan it and submit as a PDF file 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 (positive, negative, 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	neutral	neutral	neutral
quick	neutral	positive	neutral
brown	neutral	neutral	neutral
fox	neutral	negative	neutral
jumped	neutral	neutral	neutral
over	neutral	neutral	neutral
the	neutral	neutral	neutral
lazy	negative	neutral	negative
dog	neutral	negative	positive
.	neutral	neutral	neutral
Mary	neutral	positive	neutral
ate	neutral	neutral	negative
her	neutral	neutral	neutral
apple	neutral	neutral	neutral
.	neutral	neutral	neutral
This	neutral	neutral	neutral
made	neutral	neutral	neutral
me	neutral	neutral	positive
very	neutral	positive	neutral
happy	positive	positive	positive
.	neutral	neutral	neutral

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 *positive* sentiment. To evaluate this, you replace all occurrances of *negative* with *neutral* 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 *positive* and *neutral* for each system. Afterwards, calculate accuracy, precision, recall and F1 score for both systems. Which system is better in classifying positive sentiment?

Solution 1.

System 1 / **Gold**	**Positive**	**Neutral**
Positive	1	0
Neutral	3	17

System 2 / **Gold**	**Positive**	**Neutral**
Positive	1	0
Neutral	2	18

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

Table 2: Only *positive* and *neutral*.

System 1:

$$Accuracy = \frac{1+17}{1+3+17+0} = 0.86$$

$$Precision = \frac{1}{1+3} = 0.25$$

$$Recall = \frac{1}{1+0} = 1.0$$

$$F_1\text{-}Measure = \frac{2 \times 0.25 \times 1.0}{0.25+1.0} = 0.4$$

System 2:

$$Accuracy = \frac{1+18}{1+2+18+0} = 0.9$$

$$Precision = \frac{1}{1+2} = 0.33$$

$$Recall = \frac{1}{1+0} = 1.0$$

$$F_1\text{-}Measure = \frac{2 \times 0.33 \times 1.0}{0.33 + 1.0} = 0.5$$

	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 *positive* 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	neutral	neutral	positive
quick	neutral	neutral	positive
brown	neutral	neutral	neutral
fox	neutral	neutral	positive
jumped	neutral	neutral	positive
over	neutral	neutral	neutral
the	neutral	neutral	positive
lazy	neutral	neutral	neutral
dog	neutral	neutral	neutral
.	neutral	neutral	neutral
Mary	neutral	neutral	positive
ate	neutral	neutral	positive
her	neutral	neutral	positive
apple	neutral	neutral	neutral
.	neutral	neutral	neutral
This	neutral	neutral	positive
made	neutral	neutral	positive
me	neutral	neutral	positive
very	neutral	neutral	positive
happy	positive	neutral	positive
.	neutral	neutral	neutral

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 / Gold	Positive	Neutral
Positive	0	1
Neutral	0	20

Random BL / Gold	Positive	Neutral
Positive	1	0
Neutral	12	8

Majority Baseline:

$$Accuracy = \frac{0+20}{0+0+20+1} = 0.95$$

$$Precision = \frac{0}{0+0} = NA$$

$$Recall = \frac{0}{0+1} = 0.0$$

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

Random-Baseline:

$$Accuracy = \frac{1+8}{1+12+8+0} = 0.43$$

$$Precision = \frac{1}{1+12} = 0.08$$

$$Recall = \frac{1}{1+0} = 1.0$$

$$F_1\text{-}Measure = \frac{2 \times 0.08 \times 1.0}{0.08+1.0} = 0.15$$

	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 *neutral*. 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 (*positive*) since it never labeles a token as *positive*.

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 *positive*, *neutral* and *negative* 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**	**Positive**	**Neutral**	**Negative**
Positive	1	0	0
Neutral	3	14	2
Negative	0	1	0

System 2 / **Gold**	**Positive**	**Neutral**	**Negative**
Positive	1	0	0
Neutral	2	16	1
Negative	0	0	1

System 1:

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

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

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

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

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

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

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

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

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

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

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

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

$$or$$

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

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

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

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

$$or$$

$$= \frac{0.4*1+0.82*19}{\frac{9}{20}} = 0.8$$

System 2:

$$Precision_{pos} = \frac{1}{1+2+0} = 0.33$$

$$Precision_{neut} = \frac{16}{0+16+0} = 1.0$$

$$Precision_{neg} = \frac{1}{0+1+1} = 0.5$$

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

$$Recall_{neut} = \frac{16}{2+16+1} = 0.84$$

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

$$F_1\text{-}Measure_{pos} = \frac{2*0.33*1.0}{0.33+1.0} = 0.5$$

$$F_1\text{-}Measure_{neut} = \frac{2*1.0*0.84}{1.0+0.84} = 0.91$$

$$F_1\text{-}Measure_{neg} = \frac{2*0.5*1.0}{0.5+1.0} = 0.67$$

$$Macro\text{-}Average\text{-}Precision = \frac{0.33+1.0+0.5}{3} = 0.61$$

$$Macro\text{-}Average\text{-}Recall = \frac{1.0+0.84+1.0}{3} = 0.95$$

$$Macro\text{-}Average\text{-}F_1\text{-}Measure = \frac{0.5+0.91+0.67}{3} = 0.69$$

$$Micro\text{-}Average\text{-}Precision = \frac{0.33*1+1.0*19+0.5*1}{21} = 0.94$$

$$Micro\text{-}Average\text{-}Recall = \frac{1.0*1+0.84*19+1.0*1}{21} = 0.86$$

$$Micro\text{-}Average\text{-}F_1\text{-}Measure = \frac{0.5*1+0.91*19+0.67*1}{21} = 0.88$$

	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	0.71	0.94
Micro-Average Recall	0.30	0.86
Micro-Average F_1-Measure	NA/0.8	0.88