Evaluation: Performance Evaluation Part 2

COMP61332: Text Mining
Week 5
Riza Batista-Navarro

Scoring

Applies to automated systems and IAA

Which metric/measure?

Accuracy: used when all classes are equally important

Precision, recall and F-score: widely used

Scoring

Precision: fraction of annotated items that are correct

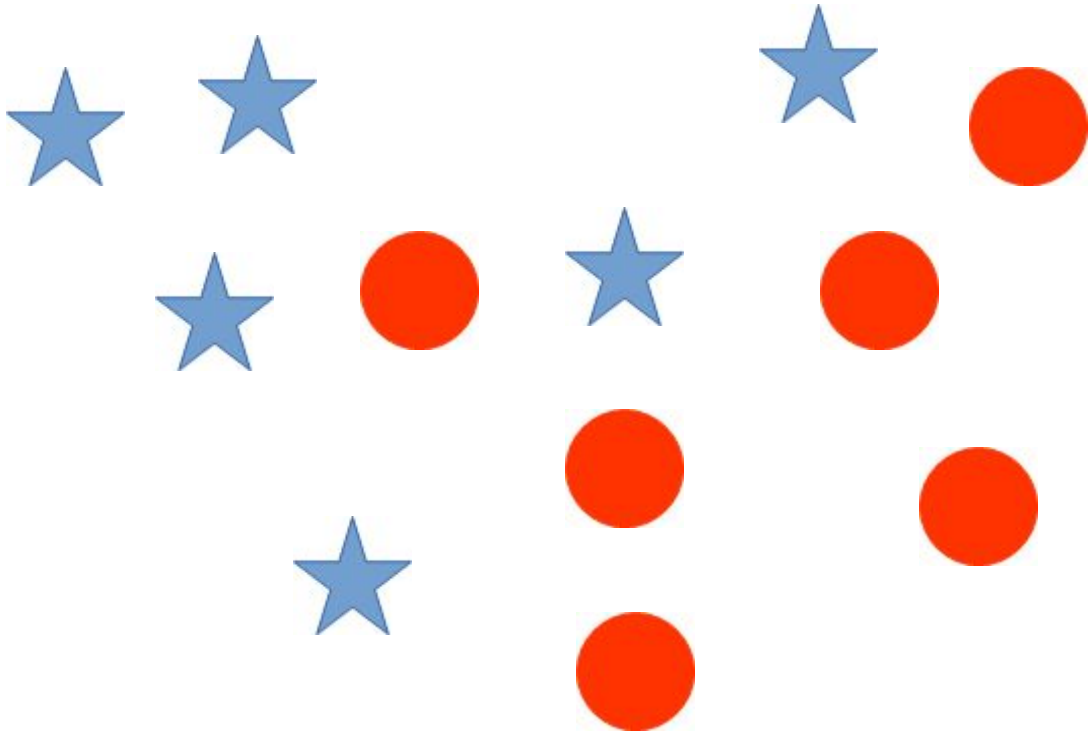
Recall: fraction of correct items that are annotated

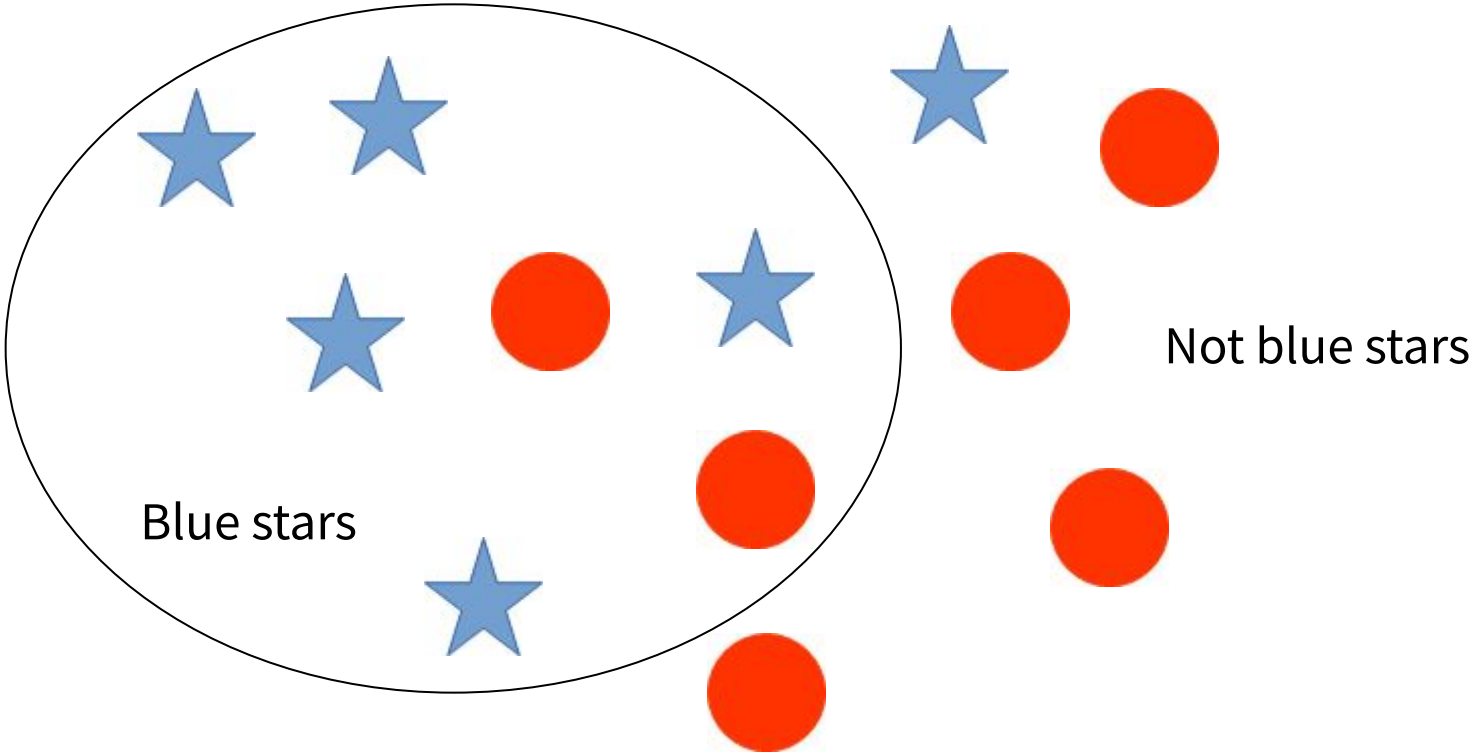
Confusion matrix:

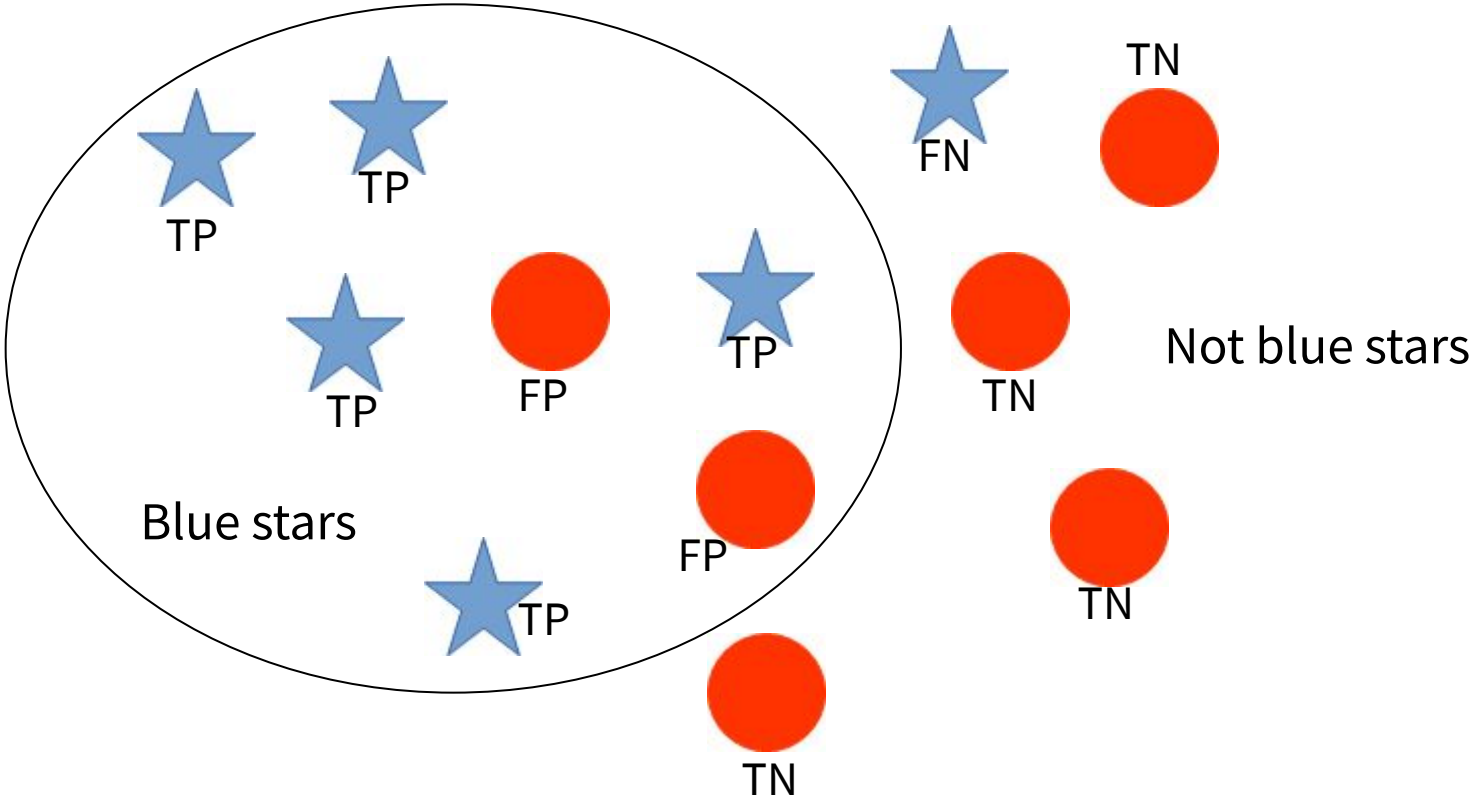
	Correct	Not correct
Annotated	True positive (TP)	False positive (FP)
Not annotated	False negative (FN)	True negative (TN)

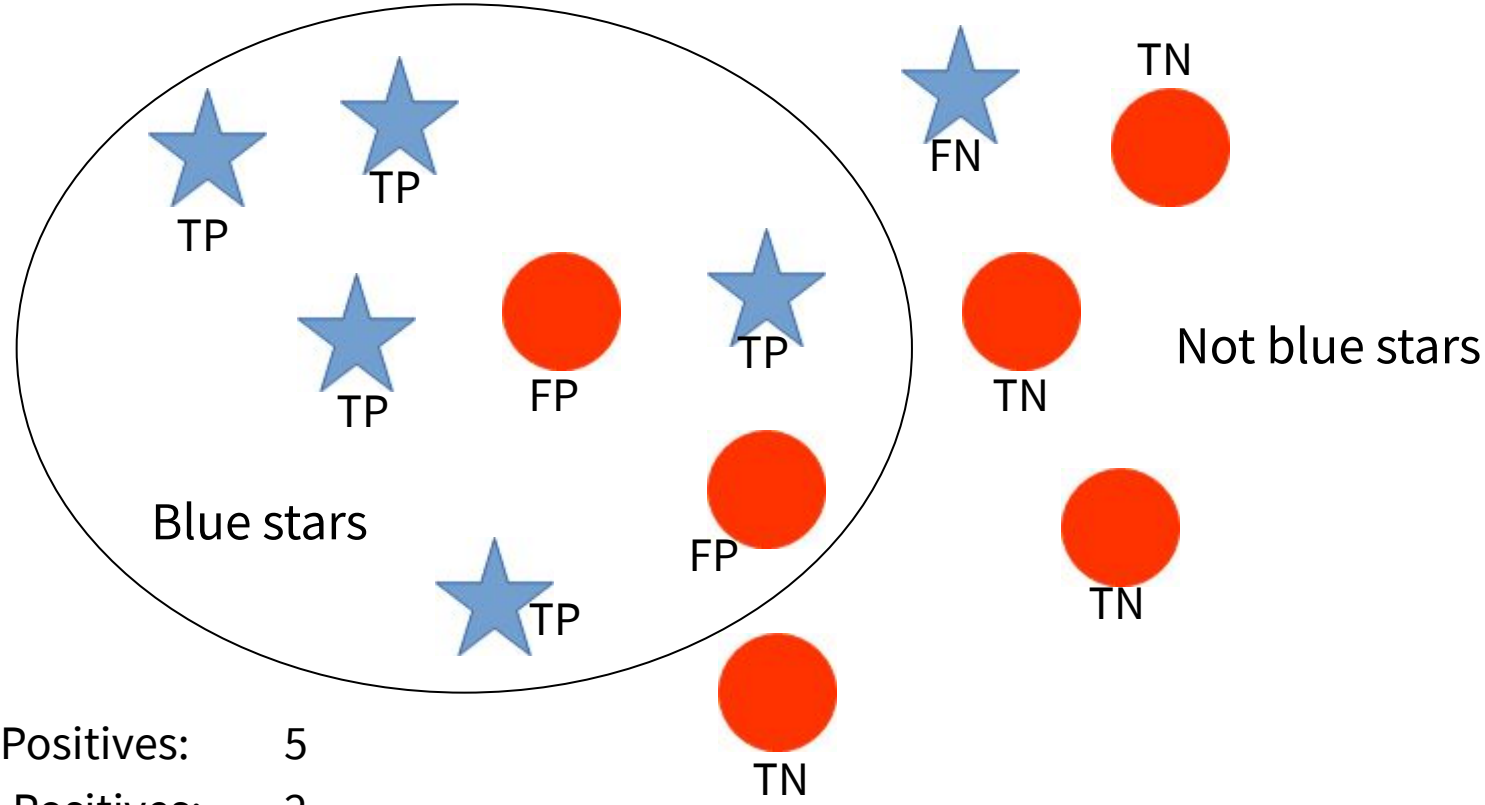
Precision = TP / (TP+FP)

Recall = TP / (TP + FN)




Blue stars
Not blue stars


TP
TP
TP
TP
FP
TP
Blue stars
FP
TP
FN
TN
TN
Not blue stars
TN
TN

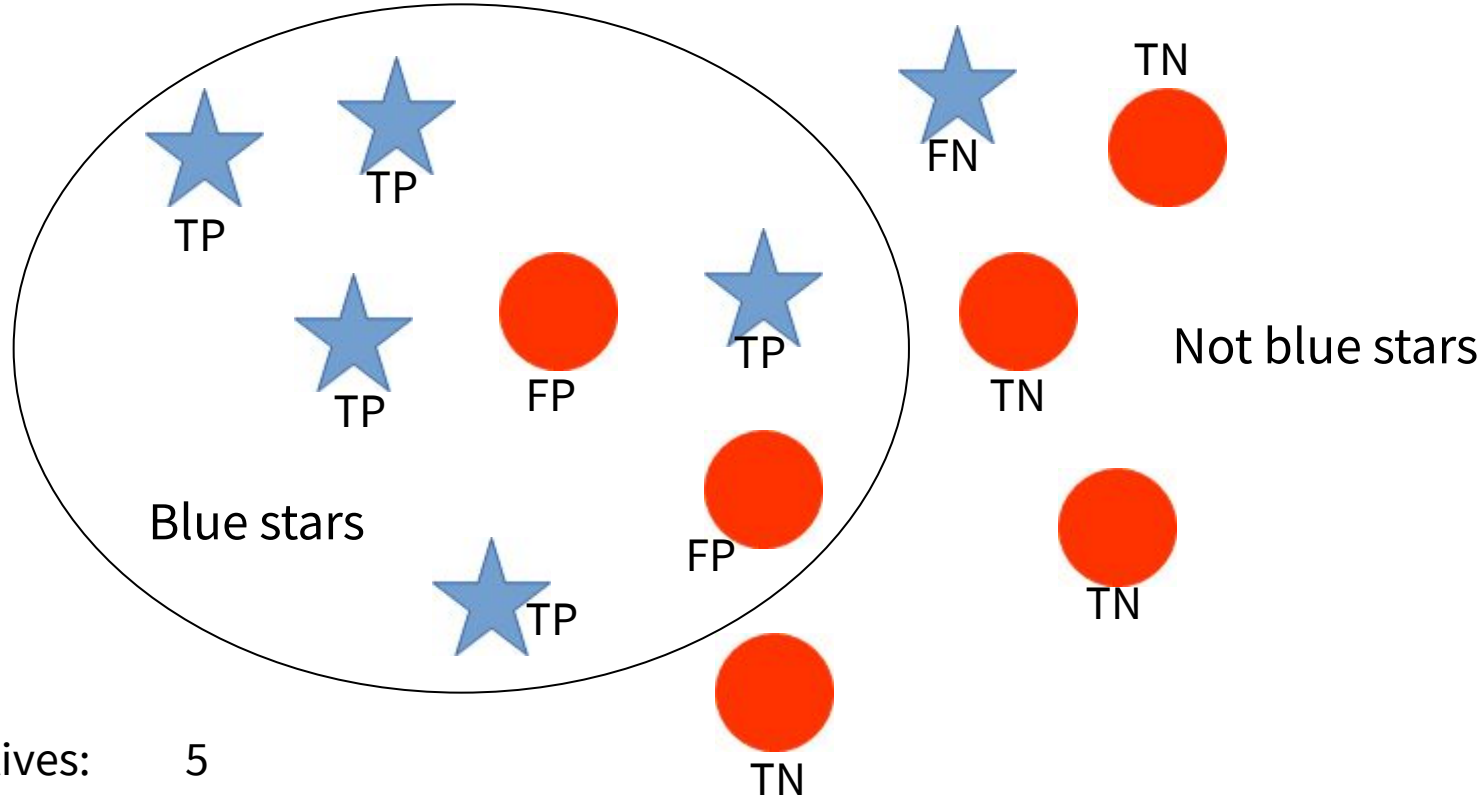


True Positives: 5

False Positives: 2

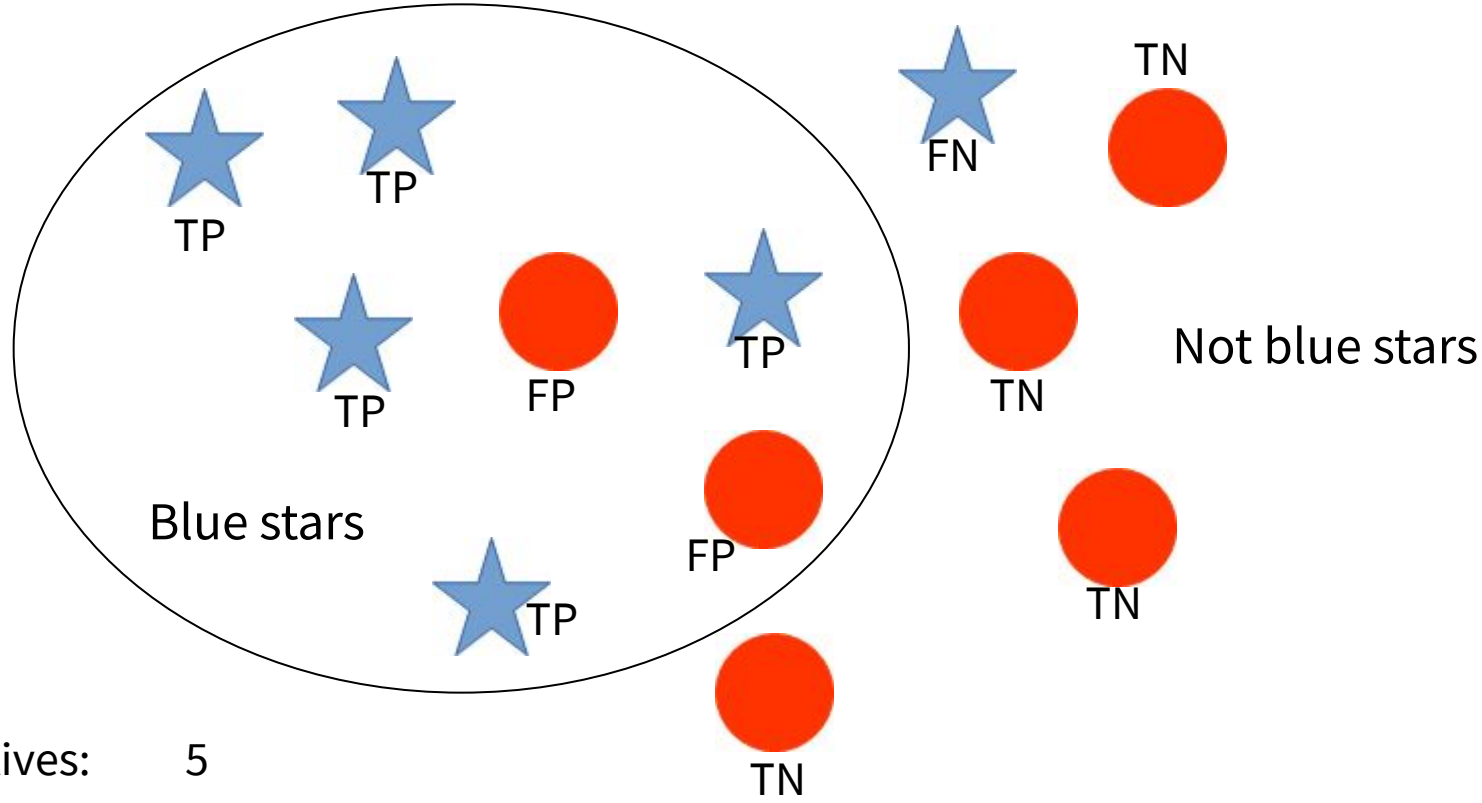
False Negatives: 1

True Negatives: 4



True Positives: 5
False Positives: 2
False Negatives: 1
True Negatives: 4

Precision = TP / (TP+FP) = 5 / (5 + 2) = 5 / 7 = 0.714



True Positives: 5
False Positives: 2
False Negatives: 1
True Negatives: 4

Precision = TP / (TP+FP) = 5 / (5 + 2) = 5 / 7 = 0.714
Recall = TP / (TP + FN) = 5 / (5 + 1) = 5 / 6 = 0.833

F-score (a.k.a. F-measure, F1-score)

Weighted harmonic mean

$$F_\beta = (\beta^2 + 1)PR/\beta^2P+R$$

Usually balanced F1 measure is used, where β=1

$$F1 = 2PR/(P+R)$$

Harmonic mean is a more conservative average (truer picture)

F-score vs Arithmetic Mean

P = 0.714

R = 0.833

Arithmetic Mean: (0.714 + 0.833) / 2 = **0.774**

F-score: 2 * 0.714 * 0.833 / (0.714 + 0.833) = **0.769**

F-score vs Arithmetic Mean

P = 1

R = 0.15

Arithmetic Mean: (1 + 0.15) / 2 = **0.575**

F-score: 2 * 1 * 0.15 / (1 + 0.15) = **0.261**

What about Accuracy?

Measure of all correctly identified cases

$Acc = (TP + TN) / (TP + TN + FP + FN))$

Suitable if all classes are equally important. What if not?

Example: Hate vs Neutral

Use F-score

Multiple Categories: Micro vs Macro-averaging

Category	TPs	FPs	FNs	Precision	Recall
Person	78	5	33	0.94	0.70
Location	20	3	2	0.87	0.91

How do we report combined performance for Person and Location?
Option 1: **Macro-averaging** -- Simply get the average

$$P_{\text{Person+Location}} = (0.94+0.87)/2 = 0.91$$

$$R_{\text{Person+Location}} = (0.70+0.91)/2 = 0.81$$

$$F1_{\text{Person+Location}} = (2*0.91*0.81)/(0.91+0.81) = 0.86$$

Multiple Categories: Micro vs Macro-averaging

Category	TPs	FPs	FNs	Precision	Recall
Person	78	5	33	0.94	0.70
Location	20	3	2	0.87	0.91

How do we report combined performance for Person and Location?
Option 2: **Micro-averaging** -- Pool together the TPs, FPs and FNs

$$P_{Person+Location} = (78+20)/((78+20)+(5+3)) = 0.92$$

$$R_{Person+Location} = (78+20)/((78+20)+(33+2)) = 0.74$$

$$F1_{Person+Location} = (2*0.92*0.74)/(0.92+0.74) = 0.82$$

Which is better?

Category	TPs	FPs	FNs	Precision	Recall
Person	78	5	33	0.94	0.70
Location	20	3	2	0.87	0.91

Macro-averaging does not consider class imbalance; micro-averaging is less sensitive to imbalance

Weighted average:
average weighted by the number of true instances for each class

Olympic judging

Alternative method used in some performance evaluations

Committee of judges determines whether some system proposal is relevant or how close it is to desired response

No automatic scoring (not reproducible)